Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea triennale in Informatica per il Management

Tecniche di analisi statistica sull'attività di lobbying negli Stati Uniti usando il software R

Relatore:
Chiar.mo Prof.
Antonio Messina

Presentata da:
Alessandro Morsiani

Correlatore:
Chiar.mo Prof.
Edoardo Vincenzo Eugenio Mollona

III
2013/2014

Dedicato a
Giuseppe Daniela Gabriele e Giulia

Indice

Elenco delle figure

Abstract

La tesi fornisce una panoramica delle principali metodologie di analisi dei dati utilizzando il software open source R e l'ambiente di sviluppo integrato (IDE) RStudio.

Viene effettuata un'analisi descrittiva e quantitativa dei dati GICS, tassonomia industriale che cataloga le principali aziende per il processo di gestione e ricerca degli asset di investimento. Sono stati studiati i principali settori del mercato USA considerando il fatturato, le spese per il lobbying e tre indici che misurano il grado di collegamento fra industrie. Su questi dati si sono svolte delle analisi quantitative e si sono tentati alcuni modelli nell'ambito della regressione lineare semplice e multipla.

Tale studio ha il compito di verificare eventuali interazioni fra le variabili o pattern di comportamento strategico societario durante il periodo 2007 - 2012, anni di rinnovo e miglioramento delle regolamentazioni in materia di *lobbying* negli Stati Uniti.

Più nello specifico vengono presi in esame tre settori: IT, Health Care e Industrial dove viene studiato l'andamento del reddito medio e la spesa media in attività di lobbying dei settori.

I risultati ottenuti mostrano l'utilità dei pacchetti di R per l'analisi dei dati: vengono individuati alcuni andamenti che, se confermati da ulteriori e necessarie analisi, potrebbero essere interessanti per capire non solo i meccanismi impliciti nell'attività di lobbying ma anche comportamenti anomali legati a questa attività.

Introduzione

Esiste un fenomeno nascosto che si protrae nelle democrazie pluralistiche che più di ogni altro crea scalpore e disagio. Un fenomeno, oppure si potrebbe dire un comportamento umano, una "*forma mentis*" che si concretizza nell'azione determinata di gruppi di interesse nell'influenzare le decisioni pubbliche. Ed è in vertiginoso aumento.

Stiamo parlando del *lobbying*, una realtà esistita in tutte le epoche e in tutti i regimi che rappresenta la dinamica degli interessi particolari che cercano di influire sulla funzione pubblica governativa per trarre vantaggi o non subire danni per la propria azienda o categoria professionale.
Tale parola, associata molto spesso a corruzione e malaffare, pone interrogativi sulle regole essenziali di democrazia che esistono per evitare caos e per garantire certezza per tutti coloro che giocano la stessa partita. In genere a non volere le regole sono coloro che si sentono più forti e capaci di imporre le proprie volontà sugli altri o coloro che preferiscono l'opacità, l'agire nell'ombra perché lì il confine tra lecito e illecito diventa impercettibile o facilmente valicabile. La democrazia è incompatibile con l'opacità ed è basata sulla trasparenza.

Lo sanno bene negli Stati Uniti dove pochi anni fa, nel 2009, il Presidente Obama ha firmato due ordini esecutivi e tre memorandum presidenziali per garantire che la sua amministrazione sarebbe stata più aperta, trasparente e responsabile.

Molte domande sorgono spontanee; quanto costa l'investimento in attività di lobbying e soprattutto, ha senso? Le aziende ne traggono beneficio? E se sì, quali indici economici ne sono interessati?

La tesi vuole rispondere a queste domande cercando di capire, con il software statistico R, se la dinamica dell'attività di lobbying ha un effetto positivo di crescita sulle aziende americane ed in particolare nei settori più tecnologici.

Il primo capitolo offre una generale panoramica della legislazione vigente negli Stati Uniti in materia di Lobbying, con un focus particolare all'aspetto economico/finanziario.

Il secondo capitolo si occupa della classificazione GICS, modello primario statunitense di catalogazione aziendale, progettato per soddisfare le esigenze della comunità degli investitori e utilizzato in questa tesi per l'estrapolazione di dati finanziari.

Il terzo capitolo fornisce un'introduzione al software R, linguaggio e ambiente per il calcolo statistico e la grafica; tramite questo strumento e l'ambiente di sviluppo integrato (IDE) associato R Studio, si è condotta un'analisi dei dati globale e settoriale. Il capitolo include la gestione dei dati in R.

Il quarto capitolo analizza i dati GICS con la funzione di lettura da file in formato 'csv' e successiva eliminazione delle aziende che hanno valori incompleti (NA).

Il quinto capitolo descrive l'analisi quantitativa e la regressione di alcuni parametri relativi ai settori GICS nel 2012: nella prima parte si hanno funzioni classiche che studiano media, varianza, deviazione standard, massimo e minino dell'investimento in lobbying; nella seconda parte si presentano le funzioni che permettono di investigare le correlazioni tra variabili e le tecniche di regressione lineare, multipla e polinomiale.

Il sesto capitolo illustra il rapporto fra il reddito e l'investimento in attività di lobbying nei settori dell'*Information Tecnology*, *Health Care* e *Industrials*. Tali settori sono stati scelti soprattutto dalla quantità di dati a disposizione rispetto ad altri settori.

Capitolo 1

La Legislazione sul lobbying negli Stati Uniti

Negli Stati Uniti l'attività di lobbying trova formale riconoscimento costituzionale nel Primo emendamento, in base al quale ogni cittadino americano può presentare petizioni al decisore pubblico.

Un primo tentativo di regolamentazione *Ad hoc* dell'attività di lobbying risale al 1946, tentativo infruttuoso a causa dell'ambiguità del linguaggio utilizzato e della debolezza delle previsioni. La normativa attualmente in vigore è stata approvata nel 1995: si tratta del *Lobbying Disclosure Act* [1]. Tale legislazione [1] ha permesso una maggiore responsabilità per le pratiche di lobbismo federale negli Stati Uniti definendo una serie di disposizioni che cercano di mantenere un certo grado di trasparenza.

Il primo aspetto da notare è che il decreto legge individua la figura del lobbista ancorandola ad un criterio quantitativo: esso infatti stabilisce che:

«The term «lobbyist» means any individual who is employed or retained by a client for financial or other compensation for services that include more than one lobbying contact, other than an individual whose lobbying activities constitute less than 20 percent of the time engaged in the services provided by such individual to that client over a 3-month period » (Section 3, 2 USC 1602).

In altre parole, si definisce *lobbista* ai fini della legge chi, in un periodo di tre mesi, dedichi almeno il 20% del tempo impiegato ad attività di lobbying per un cliente (e queste attività includano più di un contatto). Per periodi di tempo inferiori non si hanno controllo legislativi o obblighi di qualsiasi natura.

La soglia del 20% è stata criticata come arbitraria e difficile da misurare e pertanto abbastanza agevolmente aggirabile ma essa rimane ad oggi il criterio cui fare riferimento. Il provvedimento inoltre risponde alla necessità di rendere noti su chi sono e come operano i lobbisti e, di conseguenza, come è influenzato il processo decisionale pubblico sia a livello legislativo sia a livello esecutivo, dai massimi vertici agli impiegati che lavorano in tali organi. Partendo da questi presupposti, la legge detta un'ampia definizione di lobbying: ogni "contatto lobbistico" (comunicazione orale o scritta) e ogni attività svolta a sostegno di tale contatto, comprese la preparazione, la programmazione, la ricerca e ogni altro lavoro preparatorio, destinato ad essere utilizzato in contatti, oltre al lavoro di coordinamento con attività simili svolte da altri

[1] PUBLIC LAW 104-65—DEC. 19, 1995 - http://www.gpo.gov/fdsys/pkg/STATUTE-109/pdf/STATUTE-109-Pg691.pdf.

soggetti [2] .

Negli anni il testo è stato integrato e arricchito di nuove disposizioni; il 2 Agosto 2007 vi è stata una modifica sostanziale tramite la "*Honest Leadership e Open Government Act*" [2] la quale rafforza obblighi di informativa pubblica relativi alle attività di lobbying e ai finanziamenti, pone più restrizioni per i membri del Congresso e il loro personale, e prevede la pubblicazione obbligatoria delle fatture di spesa. L'Atto, approvato dal Senato Americano si proponeva quindi di garantire un maggior livello di trasparenza all'attività di lobbying, rendendo più stringenti le norme in vigore[3]:

- venne imposto l'obbligo della presentazione di report delle attività svolte a scadenza trimestrale (nella legge del 1995 erano semestrali); si richiedeva che tutte le voci di spesa diretta del Congresso, agevolazioni fiscali limitate e benefici tariffari limitati siano identificati nelle fatture; inoltre le risoluzioni, i rapporti di conferenze e dichiarazioni dei dirigenti devono essere pubblicate su Internet almeno 48 ore prima di un voto.

- fu istituito il divieto assoluto di fare regali ai parlamentari o ai loro collaboratori (pranzi, cene, biglietti per eventi sportivi, viaggi della durata superiore a un giorno); i parlamentari dovevano comunque obbligatoriamente dichiarare ogni eventuale donazione ricevuta superiore ai quindicimila dollari;

- venne circoscritto il fenomeno del *revolving door* cioè la partecipazione ad attività lobbistiche inerenti la precedente funzione pubblica di deputati e senatori, rispettivamente prima di uno o due anni.

- si proibiva ai Membri del Senato di prendere decisioni in base alla sola pressione politica determinata da gruppi di potere privati. I soggetti che violano questa disposizione riceveranno una multa e la reclusione fino a 15 anni.

- si aumentava la pena per i membri del Congresso, Senior Staff e gli alti funzionari esecutivi per la falsificazione o mancata segnalazione di informativa finanziaria (da diecimila a cinquantamila dollari) e si stabilivano sanzioni penali fino a un anno di reclusione.

- si richiedeva che l'informativa dei lobbisti, sia nei rapporti con il Senato che con la Camera, venga archiviata elettronicamente in una banca dati su Internet pubblica e reperibile.

[2] Pub.L. 110–81, 121 Stat. 735, firmata da George W. Bush il 15 settembre 2007
[3] Maggiori informazioni su: http://www.fec.gov/law/feca/s1legislation.pdf

1.1 Attuali normative - 2009/2014

Il testo legislativo è stato arricchito grazie alle modifiche operate dal Presidente Barack Obama a partire dal 2009 dopo le sue dichiarazioni in campagna elettorale: «I intend to tell the corporate lobbyists that their days of setting the agenda in Washington are over».

Una volta ottenuto l'incarico, Il Presidente Americano operò prima con l'executive order n. 13490 del 21 gennaio 2009, su "*Ethics Committments by Executive Branch Personnel*", con cui ha fissato rigidi obblighi di trasparenza per i membri del governo, i loro collaboratori e i dipendenti pubblici federali. [3].

In seguito con l' "*Ensuring Responsible Spending of Recovery Act*" [4], stabilisce che i lobbisti non possano parlare con i funzionari pubblici governativi riguardo a programmi finanziati o finanziabili del *RecoveryAct* e che i contatti debbano avvenire solo ed esclusivamente in forma scritta.

Il 18 giugno 2010, in un altro memorandum (*Memorandum For The Heads Of Executive Departments And Agencies*), Obama ha posto ulteriori regole: i lobbisti registrati non possono ricoprire ruoli all'interno di commissioni o comitati consultivi del governo, cercando così di limitare il *revolving door*[4] [5]

Si cerca inoltre di rendere chiaro all'elettore l'etica e l'informazione sull'attività di lobbying creando una banca dati internet centralizzata dove poter visionare i rapporti fra lobbisti e senatori (in particolare sui finanziamenti per le campagne elettorali) in un formato ricercabile, ordinabile e scaricabile. Essi utilizzano inoltre il potere della presidenza per lottare a favore di un organismo di controllo indipendente che sorvegli le indagini sulle violazioni etiche del Congresso in modo che il cittadino possa essere certo che le denunce vengano prese in esame.[6]

Le ultime revisioni sono state presentate in materia di registrazione e certificazione presentando un rapporto semestrale di taluni contributi insieme con la certificazione dei contatti avuti con il Senato in maniera dettagliata[7], utilizzando il nuovo portale elettronico attivo dal Gennaio 2015[8]. Le persone o aziende specializzate pagate per fare lobbying devono registrarsi con la segreteria del Senato e l'impiegato della Camera dei Rappresentanti entro 45 giorni dal contatto primario con un legislatore o dopo essere stati assunti come consulenti per quel determinato impiego.[6]

Sono esenti dalla resistrazione i lobbisti che guadagnano meno di $ 3000 per cliente per ogni trimestre fiscale o le cui spese di lobbying totale siano inferiori a 11500 $ a trimestre; stessa norma per chi non supera il 20 % dell'orario di lavoro per attività di lobbying sempre a trimestre. In generale le organizzazione senza scopo di lucro sono esenti dalla registrazione se assumono una società esterna per l'attività lobbyistica.

[4] Università commerciale Luigi Bocconi - RULES Research Unit Law and Economics Studies Paper No. 2014 – 13 - pagina 17.

[5] Informazioni più dettagliate al sito: http://www.whitehouse.gov/the_press_office/Ethics-Commitments-By-Executive-Branch-Personnel/

[6] Maggiori informazioni su: http://change.gov/agenda/ethics_agenda/

[7] Informazioni più dettagliate al sito: http://lobbyingdisclosure.house.gov/index.html

[8] Il *Lobbying Disclosure Filing* è un sistema elettronico che permette di registrare le imprese di lobbying e la registrazione delle relazioni LD-1 e LD-2.

La legge sulla lobbying è un campo in continua evoluzione; l'*American Bar Association*[9] ha pubblicato un libro di linee guida nel 2009 di oltre 800 pagine. Le leggi sono spesso piuttosto specifiche e, quando non sono osservate, possono portare a guai seri. Ad esempio, in mancanza della presentazione di un rapporto trimestrale o depositato in modo non corretto, può portare a multe fino a $ 200.000 e la reclusione fino a cinque anni.

I lobbisti a volte sostengono norme che richiedono una maggiore trasparenza e comunicazione:
«La nostra professione è a un punto critico in cui possiamo o abbracciare i cambiamenti costruttivi delle riforme da parte del Congresso o cercare scappatoie e continuare a scivolare nella categoria dei venditori di fumo.»(Lobbyist Gerald SJ Cassidy, 2007)

1.2 L'ombra dell'Insider game

Una crescente preoccupazione espressa dai critici del fenomeno del lobbying è che la politica di Washington sia dominata dalle élite e che questo potrebbe far nascere il cosiddetto *Insider game* ovvero l'esclusione di cittadini regolari, che non possono permettersi investimenti cospicui in attività di lobbying, favorendo le imprese più ricche e radicate. Si teme quindi che coloro i quali hanno più soldi e maggiori contatti politici possano esercitare più influenza sugli altri.[5]
I critici suggeriscono che quando una potente coalizione combatte una meno potente, che non segue le regole di network di settore o sotto finanziata, il risultato può essere visto come ingiusto e potenzialmente dannoso per l'intera società.

«In molte aree, la posta in gioco è tra le grandi aziende, ed è difficile sostenere che una soluzione sia migliore di un'altra per quanto riguarda l'interesse del consumatore.

La soluzione giusta dovrebbe essere sulla base della tecnologia e dell'impronta che tale impresa può avere nella nostra società.»(Lobbyist Gerald Cassidy, 2007)[7]

[9] Associazione volontaria di avvocati e studenti di legge - informazioni su http://www.americanbar.org/aba.html

Capitolo 2

La Classificazione GICS

Come modello di classificazione delle imprese si è scelto di prendere in considerazione il *Global Industry Classification Standard's* (*GICS*), tassonomia industriale organizzata in raggruppamenti basati su processi analoghi di produzione, prodotti o comportamenti simili nei mercati finanziari.

Sviluppata nel 1999 da *Standard & Poors*[10] e da *MSCI*[11], GICS, è una metodologia che mira a rafforzare il processo di gestione della ricerca degli asset di investimento per professionisti della finanza di tutto il mondo. È il risultato di numerose discussioni con proprietari di attività, gestori di portafoglio e analisti di tutto il mondo ed è progettata per rispondere alle esigenze della comunità finanziaria globale con definizioni precise e complete.

La struttura GICS è composta da 10 settori, 24 gruppi industriali, 68 industrie e 154 sub-industrie[12] e presenta quattro caratteristiche:

- Universalità: la struttura si applica alle società di livello globale
- Affidabilità: la struttura riflette correttamente lo stato attuale delle industrie nell'universo di partecipazione
- Evoluzione: revisioni annuali sono condotte da Standard & Poor's e MSCI per garantire che la struttura resti pienamente rappresentativa dei mercati globali di oggi
- Flessibilità: la struttura offre quattro livelli di analisi, che vanno dal settore più generale alla più specializzata sub-industria

[10] *Standard & Poor's* è leader mondiale nella fornitura di informazioni finanziarie indipendenti e analisi degli investimenti; pioniere nello sviluppo di indici azionari e per la creazione di prodotti investibili e commerciabili correlati. Tra gli indici più noti dell'azienda vi sono l'*S & P Global* 1200, primo indice in tempo reale dell'azionario globale e l'*S & P* 500, l'indice con le migliori aziende di portafoglio degli Stati Uniti. *Standard & Poor's* fornisce anche informazioni aziendali finanziarie, valutazioni aziendali e analisi del valore, servizi di analisi e rating su più di 220.000 titoli e fondi in tutto il mondo. Con più di 8.500 dipendenti in 23 paesi, *Standard & Poor's* è parte integrante dell'architettura finanziaria mondiale. Ulteriori informazioni sono disponibili all'indirizzo www.standardandpoors.com.

[11] *MSCI* Morgan Stanley Capital International, è un fornitore leader di indici azionari a reddito fisso e indici di hedge fund e dei relativi prodotti e servizi.

[12] La classificazione è costantemente aggiornata da *S&P Dow Jones Indices* e *MSCI*. In Novembre 2014 hanno proposto di introdurre un nuovo settore per immobili, nonché l'aggiunta di una nuovo sotto-settore per il rame. Le modifiche saranno probabilmente operative dal 2016. http://www.msci.com/resources/pdfs/GICS_Consultation2015.pdf

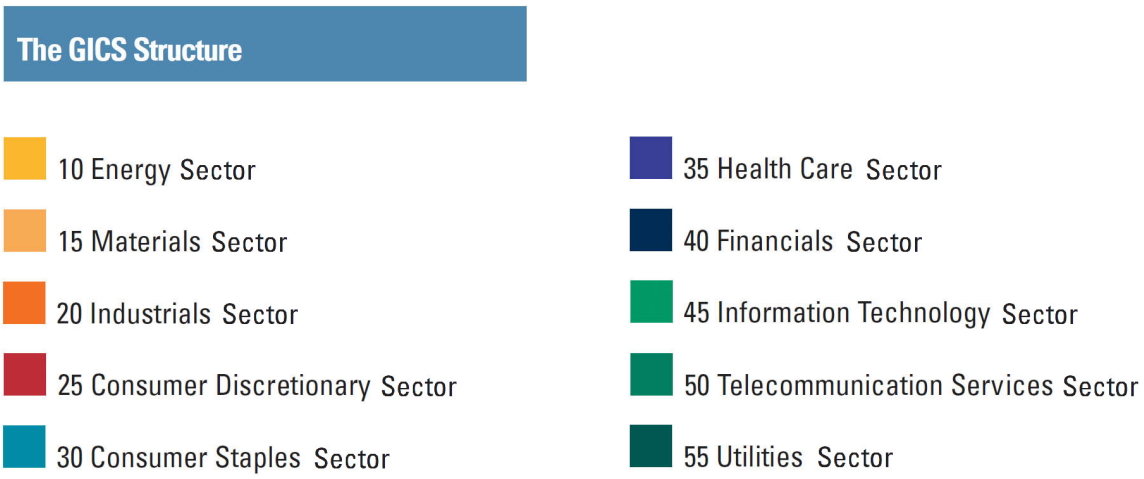


Figura 2.1: Settori GICS

La metodologia GICS prevede all'interno dei dieci settori elencati in **Figura 2.1** e identificati da due cifre, le seguenti sottocategorie:

- Gruppi industriali, identificati da quatto cifre e costituiti da 24 gruppi;
- Industrie, identificate da sei cifre, che formano un insieme di 67 elementi;
- Sottoindustrie, identificate da otto cifre, che formano un insieme di 156 sottoindustrie;

è stata ampiamente accettata come un quadro di analisi settoriale per la ricerca di investimenti, gestione del portafoglio e di asset allocation. Il suo approccio universale per le industrie di tutto il mondo ha contribuito alla trasparenza e l'efficienza del processo di investimento sostenendo la tendenza dell'investimento basato sul settore.

La catalogazione[13] prevede che ogni impresa sia parte di un $Sub - Industry$ corrispondente ad un $Industry$ corrispondente a un $Industry\ Group$ di un preciso $Sector$. Dal momento che la classificazione è strettamente gerarchica, la società può appartenere a un solo gruppo per ciascuno dei quattro livelli.

La classificazione si propone di fornire una visione precisa, completa e a lungo termine dell'universo di investimento globale da un punto di vista industriale. Per la maggior parte delle aziende, i ricavi riflettono le attività delle imprese più precisamente dei guadagni, in quanto sono meno volatili rispetto ad essi. Inoltre, mentre molte

[13] GICS_MAPBOOK_electronic_07-11 e GICS_map2014

aziende non forniscono gli utili disaggregati, nei ricavi si ha una generale disponibilà. Tuttavia, le valutazioni aziendali sono più da vicino relative agli utili rispetto ai ricavi e, di conseguenza, gli utili restano di considerevole importanza.[8]

2.1 I settori e la loro composizione

Di seguito vengono illustrate le caratteristiche che raggruppano i settori[14]. Nella parte sinistra colorata vi sarà il codice del gruppo industriale e a destra le attività necessarie alle imprese per far parte della classificazione nel settore specifico.
Da sottolineare che negli anni, dal 2007 al 2012, molte imprese sono entrare o uscite da uno specifico settore in base alla loro situazione economica positiva o negativa o ad un cambio strategico di produzione industriale. In generale, un codice GICS, cambia ogni volta che c'è una grande azione aziendale che ridefinisce la linea principale di una società.

Standard & Poor's e MSCI rivedono la struttura GICS su base annua. Spesso questo viene fatto attraverso una consultazione aperta. L'obiettivo è quello di garantire che la struttura GICS continui a rappresentare con precisione i mercati azionari globali, e in tal modo, consentire ai proprietari di asset, asset manager e specialisti di ricerca di investimenti, di effettuare un confronto globale con il mondo industriale.

L'analisi si concentra su tutti i settori nell'anno 2012 mentre nell'ultimo capitolo vengono presi in esame solo le industrie che fanno perte dell' *Information Tecnology Sector*, *Health Care Sector* e *Industrials Sector*.

Nelle tabelle che seguono sono riportate le industrie con il loro codice e una breve descrizione delle sottoindustrie relative; vengono esaminate per esteso le industrie sopra citate, oggetto dell'analisi nel **Capitolo 6**.

Industry	Energy
Energy Equipment & Services 101010	Appaltatori di perforazione o proprietari di impianti di trivellazione, che contraggono i loro servizi per i pozzi di perforazione. I fabbricanti di apparecchiature, compresi gli impianti di perforazione e attrezzature e fornitori di forniture e di servizi alle imprese coinvolte nella foratura, la valutazione e il completamento dei pozzi di petrolio e di gas.
Oil, Gas & Consumable Fuels 101020	Compagnie petrolifere integrate impegnate nell'esplorazione e produzione di petrolio e di gas, raffinazione, commercializzazione e trasporto, o prodotti chimici. Imprese impegnate nell'esplorazione e produzione di petrolio e gas non classificati altrove

[14] GICS_MAPBOOK_electronic_07-11 e GICS_map2014

Industry	Materials
Chemicals 151010	Le aziende che producono principalmente prodotti chimici industriali e prodotti chimici di base. Incluso ma non limitato alle materie plastiche, fibre sintetiche, film, vernici a base di materie prime e di pigmenti, esplosivi e prodotti petrolchimici.
Construction Materials 151020	I produttori di materiali da costruzione, tra cui sabbia, argilla, gesso, calce, inerti, cemento, cemento e mattoni. Inclusi i materiali semilavorati da costruzione che sono classificati nel sub-settore Building Products.
Containers & Packaging 151030	Produttori di contenitori metallici, vetro o plastica e contenitori di cartone e imballaggi.
Metals & Mining 151040	I produttori di alluminio e prodotti correlati, comprese le società che estraggono o lavorano bauxite e le aziende che riciclano alluminio per la produzione di prodotti finiti o semilavorati.

Industry	Industrials
Aerospace & Defense 201010	Produttori di componenti aerospaziali civili, militari e di difesa, le parti o prodotti. Include l'elettronica per la difesa e il materiale spaziale
Building Products 201020	I produttori di componenti per l'edilizia e prodotti per la casa e le attrezzature. Esclude legname.
Construction & Engineering 201030	Imprese impegnate nella costruzione prevalentemente non residenziale. Include le società di ingegneria civile e appaltatori di grandi dimensioni
Electrical Equipment 201040	Le aziende che producono cavi elettrici e fili, componenti elettrici o attrezzature di equipaggiamento per apparecchiature di potenza generatrice e di altre macchine elettriche pesanti, comprese le turbine della power-industry.
Industrial Conglomerates 201050	Società industriale diversificata con attività commerciali in tre o più settori, nessuno dei quali contribuiscono la maggioranza dei ricavi.
Machinery 201060	Produttori di autocarri pesanti, macchinari di laminazione, movimento terra e costruzioni, macchine agricole pesanti e i produttori di parti correlate. Include la cantieristica non militare. Vi sono inoltre i produttori di macchine e componenti industriali.
Trading Companies & Distributors 201070	Società commerciali e altri distributori di attrezzature industriali e di prodotti.

Commercial Services & Supplies 202010	Le aziende che forniscono servizi di stampa commerciali. Include stampanti che servono in primo luogo il settore dei media. I fornitori di servizi di elaborazione dati elettronici commerciali. Società che forniscono principalmente servizi commerciali, industriali e professionali per le imprese ei governi non classificate altrove. Include servizi commerciali di pulizia, servizi di consulenza, istituti penitenziari, servizi di ristorazione e di catering, documenti e comunicazioni servizi, servizi di attrezzature di riparazione, i servizi di sicurezza e di allarme, di stoccaggio e deposito, e servizi di noleggio uniformi. Le aziende che forniscono servizi di sostegno alle imprese in materia di gestione del capitale umano. Include agenzie di lavoro, formazione dei dipendenti, servizi paghe e benefici di assistenza, servizi di supporto di riposo e agenzie di lavoro temporaneo. Le aziende che forniscono servizi ambientali e servizi di manutenzione. Include la gestione dei rifiuti, gestione delle strutture e servizi di controllo dell'inquinamento.
Professional Services 202020	Le aziende che forniscono servizi di sostegno alle imprese in materia di gestione del capitale umano. Società che forniscono principalmente servizi di ricerca e servizi di consulenza alle imprese. Include aziende coinvolte in servizi di consulenza di gestione, progettazione architettonica, informazioni commerciali o di ricerca scientifica,marketing e servizi di test e certificazione.
Air Freight & Logistics 203010	Le aziende che forniscono trasporto aereo trasporto, corriere e logistica, compreso il pacchetto e la consegna della posta e gli agenti dog
Airlines 203020	Società che forniscono principalmente trasporto aereo di passeggeri
Marine 203030	Le aziende che forniscono beni e di passeggeri del trasporto marittimo.
Road & Rail 203040	Società che forniscono principalmente trasporto merci e trasporto ferroviario di passeggeri.
Transportation Infrastructure 203050	Gli operatori aereoportuali e società che forniscono servizi correlati.

Consumer Discretionary

Industry	
Auto Components 251010	Produttori di parti ed accessori per auto e moto. Escluse le società classificate nel Tires & Rubber
Automobiles 251020	Le aziende che producono principalmente autovetture e autocarri leggeri. Sono escluse le imprese che producono principalmente motocicli e produttori e autocarri pesanti classificate nella Construction & Farm.
Hotels, Restaurants & Leisure 253010	Proprietari e gestori di casinò e strutture di gioco. Sono incluse le società che forniscono servizi di lotterie e scommesse.Proprietari e gestori di ristoranti, bar, pub, fast-food o strutture take-out, catering
Media 254010	Le aziende che forniscono pubblicità e marketing.Proprietari e gestori di sistemi televisivi o radiofonici e aziende che si impegnano nella produzione e vendita di prodotti e servizi d'intrattenimento, comprese le società impegnate nella produzione,distribuzione e proiezione di film e show televisivi, produttori e distributori di musica, teatri, cinema e squadre sportive
Distributors 255010	Distributori e grossisti di merce generale non classificati altrove. Include distributori di veicoli.
Internet & Catalog Retail 255020	Le aziende che forniscono servizi di vendita al dettaglio prevalentemente su Internet, non classificati altrove.

Consumer Staples

Industry	
Food & Staples Retailing 301010	Proprietari e gestori di sostanze alimentari principalmente negozi e farmacie.Proprietari e gestori di ipermercati e super centri che vendono cibo e una vasta gamma di prodotti in fiocco di consumatori.
Tobacco 302030	Produttori di sigarette e altri prodotti del tabacco.
Household Products 303010	I produttori di prodotti per la casa non durevoli, tra cui detersivi, saponi, e altri prodotti.
Personal Products 303020	I fabbricanti di prodotti personali e di bellezza, tra cui cosmetici e profumi

Industry	Health Care
Health Care Equipment & Supplies 351010	Produttori di apparecchiature di assistenza sanitaria e di dispositivi. Include strumenti medici, sistemi di drug delivery, cardiovascolari e ortopedici, e attrezzature diagnostiche.
Health Care Providers & Services 351020	I fornitori di servizi di assistenza sanitaria dei pazienti. Include centri dialisi, servizi di analisi di laboratorio. Comprende anche le aziende che forniscono servizi di sostegno alle imprese per gli operatori sanitari.
Health Care Technology 351030	Le aziende che forniscono servizi informatici in primo luogo agli operatori sanitari. Sono incluse le società che forniscono applicazioni, sistemi e / o software di elaborazione dati, strumenti basati su Internet, e servizi di consulenza IT a medici, ospedali o imprese che operano principalmente nel settore della sanità.
Biotechnology 352010	Aziende impegnate principalmente nella ricerca, sviluppo, produzione e / o commercializzazione di prodotti a base di analisi genetiche e l'ingegneria genetica. Include le società specializzate in terapie a base di proteine per il trattamento di malattie umane.
Pharmaceuticals 352020	Le aziende impegnate nella ricerca, sviluppo e produzione di prodotti farmaceutici. Include farmaci veterinari.
Life Sciences Tools & Services 352030	Le aziende che permettono la scoperta dei prodotti farmaceutici lo sviluppo e la continuità della produzione, fornendo strumenti di analisi e dimateriali di consumo, forniture, servizi di sperimentazione clinica e di servizi di ricerca a contratto. Comprende le imprese a servizio soprattutto farmaceutico e le industrie biotecnologiche.

Pharmaceuticals 352020	Le aziende impegnate nella ricerca, sviluppo e produzione di prodotti farmaceutici. Include farmaci veterinari.
Life Sciences Tools & Services 352030	Le aziende che permettono la scoperta delle molecole, lo sviluppo e la continuità della produzione, fornendo strumenti di analisi, strumenti, materiali di consumo e forniture, servizi di sperimentazione clinica e di servizi di ricerca a contratto.

Industry	Financials
Commercial Banks 401010	Le banche commerciali le cui imprese sono derivati principalmente da operazioni di prestiti al settore bancario per le piccole e medie corporate lending.
Diversified Financial Services 402010	I fornitori di servizi di credito al consumo, tra cui crediti personali, carte di credito, società di locazione finanziaria, erogatori di mutui ipotecari, servizi di moneta e banchi di pegno.
Capital Markets 402030	Le istituzioni finanziarie impegnate principalmente nella gestione degli investimenti e / o servizi a pagamento di custodia e di titoli correlati. Include società operanti nei fondi comuni, fondi chiusi e fondi di investimento.
Insurance 403010	Intermediazione assicurativa delle imprese.

Real Estate Investment Trusts (REITs) 404020	Aziende o trust impegnati nell'acquisizione, sviluppo, proprietà, leasing, gestione e funzionamento di proprietà industriale. Include società operative capannoni industriali e le proprietà di distribuzione.
Real Estate Management & Development 404030	Imprese impegnate in una gamma diversificata di attività immobiliari, tra cui lo sviluppo e vendita immobiliare, gestione immobiliare.

Industry — Information Technology

Internet Software & Services 451010	Aziende in via di sviluppo e commercializzazione di software Internet e / o che forniscono servizi Internet, tra cui banche dati on-line e servizi interattivi, servizi di registrazione di indirizzi web, costruzione di database e servizi di progettazione internet.
IT Services 451020	I fornitori di information technology e servizi di integrazione dei sistemi non classificati nella elaborazione dei dati e servizi in outsourcing o Internet Software & Servizi per sub-industrie. Include consulenza tecnologica di informazioni e servizi di gestione delle informazioni.
Software 451030	Aziende impegnate nello sviluppo e nella produzione di software progettati per applicazioni specializzate per il mercato business e consumer. Include enterprise e di soluzioni tecniche.
Communications Equipment 452010	I produttori di apparecchiature di rete di computer e di prodotti, tra cui LAN, WAN e router.
Computers & Peripherals 452020	I produttori di personal computer, server, mainframe e workstation. Comprende i produttori di Automatic Teller Machines
Electronic Equipment, Instruments & Components 452030	I produttori di apparecchiature elettroniche e strumenti, tra cui analisi, test di elettronica e strumenti di misura, prodotti scanner / codici a barre, laser, schermi, point-of-sales macchine e apparecchiature del sistema di sicurezza
Office Electronics 452040	Produttori di apparecchiature elettroniche per ufficio, tra cui fotocopiatrici e fax.
Semiconductors & Semiconductor Equipment 452050	Produttori di attrezzature per semiconduttori.
Semiconductors & Semiconductor Equipment 453010	Produttori di attrezzature per semiconduttori, tra cui i produttori di materie prime e attrezzature utilizzate nel settore dell'energia solare.

Industry	Telecommunication Services
Diversified Telecommunication Services 501010	I fornitori di servizi di comunicazione e di trasmissione dati ad alta densità in primo luogo attraverso una rete di cavi in fibra ottica ad alta larghezza di banda. Gli operatori di reti di telecomunicazione su rete fissa in primo luogo e società che forniscono comunicazioni wireless e fissa.
Wireless Telecommunication Services 501020	I fornitori di servizi di telecomunicazione principalmente cellulari o hardware wireless, compresi i servizi di paging.

Industry	Utilities
Electric Utilities 551010	Le aziende che producono o distribuiscono elettricità. Include sia nucleari e impianti non nucleari.
Gas Utilities 551020	Le aziende il cui statuto principale è quello di distribuire e trasmettere il gas naturale compresi gli impianti. Sono escluse le società coinvolte principalmente in gas di esplorazione o di produzione classificate nel settore Oil & Gas Exploration & Production sub-industria.
Multi-Utilities 551030	Le società di servizi con attività diversificate in modo significativo, oltre alle utility per l'energia nucleare vi e` quella elettrica, del gas e dell'acqua.
Independent Power Producers & Energy Traders 551050	Le aziende indipendenti che operano nella produzione di energia, Gas & Power Marketing e specialisti commerciali. Comprende i produttori di energia solare ed eolica, utilizzati per generare elettricità. Include anche aziende che generano energia elettrica e / o di potenza attraverso fonti di energia alternative.

Capitolo 3

Il software R e la gestione dei dati

3.1 Introduzione a R

R è un linguaggio di programmazione e suite integrata di servizi software per la manipolazione dei dati, il calcolo e la visualizzazione grafica.
Viene utilizzato in questa tesi perchè offre: [9]

- Codice sorgente open-source operativo su Unix, Windows e Macintosh.
- È un linguaggio performante, facile da imparare e con molte funzioni statistiche incorporate.
- Una gestione efficace dei dati
- Una facilità di salvataggio
- Un tool per i calcoli su array, in particolare matrici
- Una grande e coerente, raccolta integrata di strumenti intermedi per l'analisi dei dati
- Strutture grafiche complesse e la visualizzazione direttamente sul computer

R è una implementazione del linguaggio di programmazione S[15]. È stato quindi inizialmente sviluppato secondo un paradigma funzionale, includendo in seguito anche elementi del paradigma orientato agli oggetti.[16][10]

R viene scritto da Ross Ihaka [17] e Robert Gentleman[18], del Dipartimento di Statistica di Auckland, Nuova Zelanda. In seguito, un numeroso gruppo di persone comincia a dare il suo contributo dando vita nel 1997 all'*R Core Team* che tuttora si occupa dei codici sorgenti di R. Quando viene progettato, R è sviluppato per sistemi Unix. Ora può essere istallato su macchine con diverse architetture e con diversi sistemi

[15] S è stato creato da John Chambers e Allan Wilks nei laboratori Bell; lo scopo del linguaggio, come espresso da John Chambers, è quello di "trasformare le idee in software, rapidamente e fedelmente."
[16] Maggiori informazioni su http://en.wikipedia.org/wiki/Object-oriented_programming
[17] Informazioni più dettagliate: https://www.stat.auckland.ac.nz/~ihaka/
[18] Informazioni più dettagliate: http://www.gene.com/scientists/our-scientists/robert-gentleman

operativi.[11]

Il linguaggio R è ampiamente utilizzato tra statistici ed esperti di data mining per lo sviluppo di software statistico e l'analisi dei dati. I sondaggi e i pareri di molti studiosi mostrano che la popolarità di R è in notevolmente aumento negli ultimi anni. [12][13]

R e le sue librerie implementano un'ampia gamma di tecniche statistiche e grafiche, test statistici lineari e non lineari, analisi di serie, classificazione, clustering, e altri.

R è facilmente estendibile attraverso funzioni e pacchetti inviati dagli utenti per specifiche aree di studio; la comunità di R è nota per i suoi contributi attivi in questi termini. Molte delle funzioni standard di R sono scritte in R per sé e questo facilita le scelte metodologiche e algoritmiche degli utenti. Per le attività computazionalmente intensive, R si integra con i linguaggi C , C ++ e Fortran . Altro punto di forza di R è la grafica disponibile attraverso i pacchetti aggiuntivi.

Per le analisi effettuate è stato utilizzato RStudio, ambiente di sviluppo integrato (IDE) di R. Esso comprende una console, un editor per evidenziare la sintassi e l'esecuzione di codice diretta, nonché strumenti per la stampa, la storia, il debug e la gestione del workspace.

L'edizione open source di RStudio permette di:

- accedere localmente ad RStudio (esiste anche la versione RStudio Server)
- evidenziare la sintassi, completamento del codice e indentazione intelligente
- esecuzione del codice R direttamente dall'editor
- scrivere velocemente le definizioni di funzione
- gestire facilmente più directory di lavoro con i progetti
- avere un *help* integrato completo, funzionale e ricco di esempi
- usufruire di un debugger interattivo per la diagnostica e la risoluzione rapida degli errori
- utilizzare un'ampia gamma di librerie

3.1.1 Nuove applicazioni

Ci sono diverse applicazioni industriali di analisi dei dati che utilizzano piattaforme basate sul linguaggio R:

- Settore Finanziario: Le banche, le società di intermediazione, i commercianti, le compagnie di assicurazione e gli hedge fund utilizzano l'analisi dei dati per guidare le loro decisioni fondamentali e creare efficienze operative per massimizzare il potenziale di guadagno di capitale e per soddisfare i requisiti di riserva minimi normativi.
  È il caso della *ANZ Bank*[19], la quarta più grande banca Australiana utilizza R per l'analisi del rischio di credito; anche la *Bank of America*[20] usa R per i report interni.

[19] Informazioni più dettagliate: http://blog.revolutionanalytics.com/2011/08/how-anz-uses-r-for-credit-risk-analysis.html

[20] Informazioni più dettagliate: http://blog.revolutionanalytics.com/2014/06/bank-of-america-uses-r-for-reporting.html

- Settore Scientifico: Le aziende farmaceutiche, gli ospedali e le agenzie governative stanno utilizzando i dati e le analisi statistiche per la ricerca, contribuendo ai costanti progressi della medicina.

- Marketing: I dati dei consumatori sono diventati fondamentali per le aziende che effettuano targeting e acquisizione di nuovi clienti, fidelizzazione dei clienti esistenti e la realizzazione di un maggior valore nelle loro relazioni con i clienti nel tempo.

- Ambito Accademico: R è utilizzato nelle aule in tutto il mondo per insegnare la programmazione statistica agli scienziati di domani. Ricercatori, professori e studenti spingono lo sviluppo delle statistiche moderne per fornire gli algoritmi migliori e i pacchetti che a loro volta aiuteranno le imprese ad ottenere maggiore ritorno sul loro investimento.

- Settore IT: Molte aziende famose dell'industria tecnologica informatica utilizzano R; Facebook ad esempio applicando gli strumenti di R all'analisi degli *status updates*[21] e per i *Facebook's Social Network Graphs*[22].
  Google lo utilizza per calcolare il ROI delle campagne pubblicitarie[23], per le previsioni di attività economiche[24] e sulle valutazioni di come rendere più efficace la pubblicità online[25].
  La Microsoft usa R per la *Xbox matchmaking service*[26] e come motore statistico per il Framework *Azure ML*[27]. Mozilla Firefox usa R per le Web activity[28].
  Twitter inserisce il linguaggio nella *Twitter's Data Science toolbox*[29] e per il monitoraggio della *user experience* del sito [30].

[21] Informazioni più dettagliate: http://blog.revolutionanalytics.com/2010/12/analysis-of-facebook-status-updates.html
[22] Informazioni più dettagliate: http://blog.revolutionanalytics.com/2010/12/facebooks-social-network-graph.html
[23] Informazioni più dettagliate: http://blog.revolutionanalytics.com/2014/09/google-uses-r-to-calculate-roi-on-advertising-campaigns.html
[24] Informazioni più dettagliate: http://blog.revolutionanalytics.com/2009/09/google-uses-r-to-predict-economic-activity.html
[25] Informazioni più dettagliate: http://blog.revolutionanalytics.com/2011/08/google-r-effective-ads.html
[26] Informazioni più dettagliate: http://blog.revolutionanalytics.com/2014/05/microsoft-uses-r-for-xbox-matchmaking.html
[27] Informazioni più dettagliate: http://blog.revolutionanalytics.com/2014/11/r-on-azure-ml.html
[28] Informazioni più dettagliate: http://blog.revolutionanalytics.com/2010/08/whats-for-lunch-private-browsing.html
[29] Informazioni più dettagliate: http://blog.revolutionanalytics.com/2012/03/r-twitter-and-mcdonalds.html
[30] http://blog.revolutionanalytics.com/2014/11/breakout-detection.html

3.2 La gestione dei dati in R

Ogni oggetto in R contiene una serie di attributi per descrivere la natura delle informazioni che contiene. Due delle caratteristiche più importanti della gestione dei dati in R sono le modalità (numerica, carattere, logica) e la classe (vettori, matrici, array, liste o data frame).

È importante comprendere le differenze tra i diversi tipi di dati che R supporta e, quando sorgono ostacoli, spesso è perchè i dati non sono in una modalità corretta o appartengono a una classe errata per la gestione di una particolare operazione.
Uno dei compiti più impegnativi in analisi dei dati è la preparazione dei dati.

R offre varie strutture per lo svolgimento di questa funzione compresi molti metodi per l'importazione dei dati da tastiera e da fonti esterne.

Se ne illustrano qui le principali strutture dati: [14][15]

- Vettori: i numeri con cui siamo abituati a ragionare sono in realtà un caso particolare di una famiglia più grande, i vettori. Un vettore di dimensione n può essere considerato come una sequenza ordinata di n numeri; ad esempio, (2,5,9.5,-3) rappresenta un vettore di dimensione 4 in cui il primo elemento è 2 ed il quarto è -3. Esistono molti modi per costruire un vettore, il più comune è utilizzare la funzione c():

```
> x<-c(2,5,9.5,-3) #costruisci un vettore
> x[2] #seleziona il suo secondo elemento
```

- Matrici: tabella ordinata di numeri in cui ciascun elemento è univocamente individuato da una coppia di valori interi che costituiscono l'indice di riga e colonna; le matrici hanno la caratteristica di avere un'unica tipologia di elemento che può essere numerico o di stringa; un'applicazione in r potrebbe essere la seguente:

```
> x<-matrix(1:10,ncol=5) #costruisci una matrice
> x
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    3    5    7    9
[2,]    2    4    6    8   10
```

- Array: costituiscono una estensione delle matrici; in un array (multidimensionale) ogni suo elemento è individuato da un vettore di indici (si ricordi che in vettori e matrici gli elementi sono individuati da uno e due indici rispettivamente). Ad esempio, in un array tridimensionale, ogni elemento è caratterizzato da una terna ($i1$, $i2$, $i3$). Sebbene in statistica applicata gli array possano trovare numerose applicazioni, in un approccio di gestione dei dati tali elementi possono essere tralasciati. Soltanto per completezza qualche esempio di codice per la gestione degli array è riportato di seguito:

```
> a<-array(1:24, dim=c(3,4,1,2)) #crea un array
> dim(a) #la sua dimensione
[1] 3 4 2
> a[,,2]
     [,1] [,2] [,3] [,4]
[1,]   13   16   19   22
[2,]   14   17   20   23
[3,]   15   18   21   24
```

- Liste: In R una lista è una raccolta di oggetti, anche differenti tra loro, compreso altre liste. Questo non è vero per i vettori o le matrici i cui elementi sono tipicamente tutti numeri o tutti caratteri. Una lista può essere creata con il comando list() ed il numero degli oggetti che la costituiscono definisce la dimensione della lista, mentre le sue componenti sono individuate con semplici [] o doppie [[ ]] parentesi quadre:

```
> a<-array(1:24, dim=c(3,4,1,2)) #crea un array
> dim(a) #la sua dimensione
[1] 3 4 2
> a[,,2]
     [,1] [,2] [,3] [,4]
[1,]   13   16   19   22
[2,]   14   17   20   23
[3,]   15   18   21   24
```

- Dataframe: è forse l'oggetto più importante di tutto l'ambiente R, almeno in una sua ottica di gestione e analisi dei dati. In questa tesi è la struttura maggiormente utilizzata.
  Il dataframe rappresenta la matrice dei dati in cui ad ogni riga corrisponde una osservazione e ad ogni colonna una variabile, e nell'ambiente viene trattato come una lista. Ogni elemento di tale lista rappresenta una variabile statistica, per cui $length()$ restituisce il numero delle variabili, mentre $names()$ i rispettivi nomi; è anche possibile aggiungere/modificare i nomi di riga (attraverso $row.names()$), ma probabilmente per un dataframe questo potrebbe non essere molto utile. Tra le diverse opzioni disponibili, è possibile costruire un data.frame direttamente con la funzione $data.frame()$:

```
> #crea un dataframe con una variabile 'quantitativa' ed una
> #'qualitativa':
> X<-data.frame(a=1:4, sesso=c("M","F","F","M"))
> X
  a sesso
1 1     M
2 2     F
3 3     F
4 4     M
> dim(X) #la 'dimensione' (numero dei casi e di variabili)
[1] 4 2
> X$eta<-c(2.5,3,5,6.2) #aggiungi una variabile di nome eta
  a sesso eta
1 1     M 2.5
2 2     F 3.0
3 3     F 5.0
4 4     M 6.2
```

L'ultimo dataframe creato sopra è definito da 4 casi e 3 variabili (due numeriche ed una 'carattere') che possono essere selezionate in tre modi diversi, utilizzando sia il nome (ad es., X$sesso o X[,sesso]), sia il numero di colonna che occupa (ad es., X[,2]); il risultato sarà comunque un vettore.
La funzione $data.frame()$ descritta precedentemente non esaurisce le diverse possibilità per affrontare il problema dell'inserimento e gestione di dati. Ad esempio la funzione $as.data.frame()$ può essere utilizzata per forzare una matrice di dati ad un dataframe, oppure i dati potrebbero essere inseriti più facilmente attraverso un foglio elettronico: in questo caso $X < -data.frame()$ crea un dataframe che è possibile aprire con $fix(X)$ per l'inserimento dei dati direttamente nelle celle.
Il dataframe risulta essere particolarmente importante perchè consente una stratificazione diversificata degli elementi nella costruzione di un grafico[17]. Ogni carattere estetico[31]può provenire da un set di dati diversi e hanno una diversa mappatura estetica, che ci permette di creare grafici che non potevano essere generati utilizzando qplot()[32], che consente solo un singolo set di dati e un singolo insieme di mappe estetiche, comportando quindi un utilizzo esclusivo di soli vettori e matrici.
In questa tesi si utilizza il pacchetto *ggplot*2 per la costruzione di grafici composti da singoli livelli[33] che hanno le seguenti caratteristiche:

- mappatura(opzionale): Una serie di mappature estetiche, specificate attraverso la funzione $aes()$;
- dati (opzionale): il dataframe oggetto di analisi;
- parametri per il *geom* o *stat*: cioè la tipologia geometrica delle figure (istogrammi,punti o linee) e la loro larghezza di banda per una maggiore chiarezza espositiva;
- posizione (opzionale): scelta di un metodo per la regolazione degli oggetti sovrapposti.

[31] i caratteri estetici sono elementi di visualizzazione dei livelli grafici che consentono un'illustrazione chiara della diversità e variabilità dei parametri
[32] qplot() è la funzione standard in R per la generazione di un grafico
[33] I livelli sono oggetti normali di R, memorizzati come variabili, il che rende facile scrivere codice pulito che riduce la duplicazione. Ad esempio, un insieme di grafici può utilizzare diversi dati poi ridotti allo stesso livello. Se in seguito si decide di cambiare quel livello, si ha solo la necessità di farlo in un unico luogo.

Capitolo 4

I dati

I dati esaminati in questa Tesi riguardano 619 imprese americane e costituiscono una parte dei dati che sono stati raccolti da Daniele Incicco per la sua Tesi di Laurea Magistrale. La selezione di partenza faceva riferimento alle classifiche annuali stilate dalla rivista *Fortune*[31], la quale selezionava annualmente le 500 imprese societarie degli Stati Uniti sulla base del loro fatturato. Successivamente i dati finanziari delle aziende ottenute sono stati raccolti dalla banca dati Osiris[32] e i dati relativi alle spese di lobbying sono stati reperiti dalla sezione *Open Data* di Opensecrets[33] [34].

Ogni anno analizzato viene rappresentato da una tabella che riporta:

- il nome della ditta;
- l'investimento in attività di lobbying, in dollari USA;
- l'investimento in ricerca e sviluppo, in dollari USA;
- il fatturato, in dollari USA;
- l'investimento in lobbying, espresso in percentuale del fatturato;
- l'investimento in ricerca e sviluppo, espresso in percentuale del fatturato;
- il ROE (Return On Equity), ovvero la capacità di remunerare il capitale di rischio che i soci o il proprietario hanno utilizzato; corrisponde quindi al rapporto tra il reddito netto conseguito e il capitale netto;
- i codici che rappresentano misure di relazioni fra le reti di aziende secondo la definizione che ne è stata data da Hanneman;[16]
- i codici dei settori di attività industriale secondo il sistema GICS;

[31] *Fortune* è una rivista globale di business, pubblicata dalla Time Inc.

[32] Osiris è una banca dati con i dettagli finanziari delle principali aziende quotate e non di tutto il mondo di proprità di Bureau van Dijk

[33] Opensecrets.org è uno dei principali siti americani che si occupa di tenere traccia dei flussi di denaro generati dai finanziamenti, di vario genere, alla politica statunitense, e in che modo questi danno forma alle policy. Maggiori informazioni su http://www.opensecrets.org/

[34] D. Incicco, Dinamica, dimensione e strutture di rete dell'attività di lobbying negli Stati Uniti, Sessione III, A.A. 2012/2013 - https://it.linkedin.com/in/danieleincicco

4.1 Funzione di lettura del file

I dati sono presenti in file nel formato csv (comma-separated values), basato su file di testo utilizzato per l'importazione ed esportazione (ad esempio da fogli elettronici o database) di una tabella di dati[29].

Viene illustrata in questo capitolo la metodologia di acquisizione dei dati utilizzando il file relativo all'anno 2012. A tale scopo si crea una funzione chiamata 'leggiAnno' che prende in input una variabile 'mioPercorso' all'interno della quale avremo il path del file: ∼/Projects/RStudio/GICS/Dati/DB_GICS_2012.

Si può utilizzare sia la funzione 'scan' che la funzione 'read.delim', che di default legge un file in formato tabella e crea un frame di dati da esso, specificando:

- 'dec' il carattere utilizzato nel file per i punti decimali, in questo caso la virgola;
- 'header' un valore logico che indica se il file contiene i nomi delle variabili sulla prima riga. In questo caso TRUE perchè disponiamo delle intestazioni;
- 'sep' il carattere separatore di campo. I valori su ogni riga del file possono essere separati da spazi, tabulazione o ritorni a capo; in questo caso da tabulazione.
- 'stringAsFactor' che serve per la conversione in fattori; in questo caso FALSE;

```
leggiAnno <- function(mioPercorso)
{
  azienda <- read.delim(mioPercorso, dec=",",
                        header=TRUE, sep ="\t",
                        stringsAsFactor=FALSE)
  identifica <- c(1) # nome azieda
  all <- data.frame(azienda[identifica])
  # 4: Revenue, 5: Lobby (%), 7: ROE (%),
  #8: nBtw2012, 9: Clo2012,
  # 10: nDgr2012, 11: E-I2012,
  #14: GICS Industry, 15: GICS Sub-Industry,
  # 16: GICS Code Description
  indice <- c(4,5,7:10,14:16)
  all <- cbind(all, azienda[indice])
  return (all)
}
```

Figura 4.1: Funzione di lettura dal file DB_GICS_2012

Dopo l'acquisizione dei dati e il salvataggio nella variabile 'azienda' inserisco in un data frame i nomi delle aziende passando il parametro 'identifica' specificato come prima colonna del data frame 'azienda' con la funzione 'c' che restituisce un vettore:

[29] Informazioni più dettagliate su http://it.wikipedia.org/wiki/Comma-separated_values

```
identifica <- c(1) # nome azieda
all <- data.frame(azienda[identifica])
```

Sempre utilizzando la funzione 'c' seleziono altre 4 colonne fondamentali che interessano l'analisi dei dati Gics:

- 4 - Revenue: Si intende la quantità di denaro che una società riceve effettivamente nel corso di un determinato periodo, compresi gli sconti e le detrazioni per la merce restituita. È la "top line" o "reddito lordo", da cui vengono sottratti i costi per determinare il reddito netto. Tali ricavi sono calcolati moltiplicando il prezzo al quale beni o servizi sono venduti per il numero di unità o quantità venduta.
- 5 - Lobby %: Percentuale di investimento in attività di lobbying calcolata sul fatturato annuo
- 7:10 - Con la funzione ':' si genera una sequenza di valori, dalla colonna 7 alla 11, inclusi gli estremi:
  - 7 - Roe: Return On Equity %
  - 8 - nBtw2012: Betweeness, ovvero la misura del ruolo di connettore con altre imprese, assumendo una funzione di broker: maggiore la betweenness, maggiore è probabilmente il potere posseduto all'interno del settore.
  - 9 - Clo2012: Closeness, cioè la 'prossimità' misurata come la distanza (vicinanza) dalle altre aziende: minore è la distanza (espressa p.e. in termini di lunghezza dei percorsi geodetici), maggiore può essere il potere derivante dall'essere un "punto di riferimento" per gli altri attori, per poterli raggiungere facilmente.
  - 10 - nDgr2012: Degree, cioè il 'grado' di un nodo che rappresenta il numero complessivo dei legami che esso possiede, senza tener conto della loro direzione. Così espressa, questa misura è tipica di una rete a legami vincolati, ovvero non orientati.
- 14:16
  - 14 - GICS Industry: Codice a 6 cifre che indica il gruppo industriale (ad esempio: 201010)
  - 15 - GICS Sub-Industry: Codice a 8 cifre che indica il sotto gruppo industriale (ad esempio: 20101010)
  - 16 - GICS Code Description: Breve descrizione del gruppo industriale al quale l'impresa appartiene (ad esempio: Aerospace & Defense)

Si esegue successivamente la funzione 'cbind' che unisce la prima colonna precedentemente selezionata, con i nomi delle aziende, salvata nella variabile 'all' e gli ultimi parametri salvati in 'indice'; la funzione utilizza il 'return' per fornire il risultato finale:

```
all <- cbind(all, azienda[indice])
return (all)
```

Figura 4.2: Campione di dati GICS_2012 letti con la funzione 'leggiAnno'

Una volta identificato il percorso del file DB_GICS_2012 all'interno dei progetti R si passa all'effettivo utilizzo della funzione:

```
percorso <- "~/Projects/RStudio/GICS/Dati/DB_GICS_2012.csv"
aziendeLette <- leggiAnno(percorso)
```

Il risultato è visibile cliccando nella grid relativa al risultato della Global Environment in alto a destra:

	Company	Revenue..USD.	Lobby....	ROE....	nBtw20....	Clo20....	nDgr20....	E.I20....	GICS.Industry	GICS.Sub.Industry	GICS.Code.Description
1	3COM CORP	1294879000	0.0000	-22.99	NA	NA	NA	NA	452010	45201020	Communications Equipment
2	3M COMPANY	25269000000	0.0070	35.02	0.4036	0.8469	4.1588	0.7068	201050	20105010	Industrial Conglomerates
3	ABBOTT LABORATORIES	29527552000	0.0168	27.92	0.2312	0.7987	4.5683	0.7153	351010	35101010	Health Care Equipment
4	ADELPHIA COMMUNICATIONS CORP	NA	NA	NA	NA	NA	NA	NA	254010	25401020	Broadcasting
5	ADVANCE AUTO PARTS, INC.	5142255000	0.0000	22.06	NA	NA	NA	NA	255040	25504050	Automotive Retail
6	ADVANCED MICRO DEVICES INC	5808000000	0.0112	NA	0.1066	0.7242	2.1518	0.7277	453010	45301020	Semiconductors
7	ADVANCEPCS	NA	NA	NA	NA	NA	NA	NA	351020	35102000	Health Care Providers & Services
8	AECOM TECHNOLOGY CORPORATION	5194682000	0.0033	10.33	0.0001	0.5573	0.3078	0.6512	201030	20103010	Construction & Engineering
9	AES CORPORATION	15197000000	0.0027	33.63	0.0046	0.5730	0.4248	0.6699	551050	55105010	Independent Power Producers & Energy Traders
10	AETNA INC	27384100000	0.0074	NA	0.1927	0.7883	4.8453	0.7091	351020	35102030	Managed Health Care
11	AFFILIATED COMPUTER SERVICES INC	6160550000	0.0124	14.25	0.0001	0.5641	0.4002	0.9130	451020	45102020	Data Processing & Outsourced Services
12	AFLAC INCORPORATED	14947000000	0.0239	NA	0.0984	0.7115	1.9640	0.6444	403010	40301020	Life & Health Insurance
13	AGCO CORP	8273100000	0.0004	19.16	NA	NA	NA	NA	201060	20106010	Construction & Farm Machinery & Heavy Trucks
14	AGILENT TECHNOLOGIES INC	5774000000	0.0016	27.08	0.0621	0.7344	2.0686	0.7427	352030	35203010	Life Sciences Tools & Services
15	AIR PRODUCTS & CHEMICALS INC	10414500000	0.0173	18.08	0.0379	0.6381	0.7665	0.8443	151010	15101040	Industrial Gases
16	AIRBORNE INC	NA	NA	NA	NA	NA	NA	NA	203010	20301010	Air Freight & Logistics
17	AIRGAS INC	4361479000	0.0000	16.61	NA	NA	NA	NA	151010	15101040	Industrial Gases
18	AK STEEL HOLDING CORPORATION	7644300000	0.0031	0.41	NA	NA	NA	NA	151040	15104050	Steel
19	ALBERTSON'S LLC	NA	NA	NA	NA	NA	NA	NA	301010	30101000	Food & Staples Retailing

Tale risultato è osservabile anche tramite la funzione 'str' che ritorna la struttura dell'oggetto acquisito. Come si nota, all'interno della tabella sono presenti valori NA[30] : 'Not Available', cioè valori che non è stato possibile reperire e quindi non sono presenti nel data base originale. Per avere un'idea più chiara di quanti NA sono presenti all'interno del data frame si può utilizzare la funzione 'summary' che ritorna un sommario dei dati delle colonne considerate:

[30]Informazioni più dettagliate su http://www.statmethods.net/input/missingdata.html

```
##    Company          Revenue..USD.         Lobby....         ROE....
##  Length:619         Min.   :4.620e+05   Min.   :0.0000   Min.   :-97.680
##  Class :character   1st Qu.:6.702e+09   1st Qu.:0.0012   1st Qu.:  9.742
##  Mode  :character   Median :1.144e+10   Median :0.0052   Median : 17.540
##                     Mean   :2.477e+10   Mean   :0.0108   Mean   : 20.097
##                     3rd Qu.:2.355e+10   3rd Qu.:0.0162   3rd Qu.: 24.405
##                     Max.   :4.515e+11   Max.   :0.0714   Max.   :359.570
##                     NA's   :443         NA's   :444      NA's   :449
##     nBtw2012          Clo2012          nDgr2012         E.I2012
##  Min.   :0.00000   Min.   :0.2500   Min.   :0.0000   Min.   :-1.0000
##  1st Qu.:0.00705   1st Qu.:0.5353   1st Qu.:0.1726   1st Qu.: 0.6240
##  Median :0.08080   Median :0.6647   Median :0.6468   Median : 0.7201
##  Mean   :0.15466   Mean   :0.6319   Mean   :0.9661   Mean   : 0.6375
##  3rd Qu.:0.22252   3rd Qu.:0.7279   3rd Qu.:1.3900   3rd Qu.: 0.7984
##  Max.   :1.58820   Max.   :0.8289   Max.   :6.1568   Max.   : 1.0000
##  NA's   :279       NA's   :279      NA's   :279      NA's   :279
```

Dal risultato ottenuto si ha che le percentuali di dati mancanti sono circa 71% per il revenue misurato in USD, circa il 72% per attività Lobbying, sempre il 72% per il Roe, 45% per l'nBtw, Clo e nDgr.

Inoltre risulta che gli intervalli in cui sono presenti i dati varia notevolmente nelle colonne considerate, con un ampio intervallo per le variabili revenue e Roe.

4.2 Funzione di eliminazione degli NA

Per quanto riguarda i dati mancanti, si possono scegliere diverse strategie[18]. In questo caso, per esempio, assegnare il valore zero al dato per lobbying o per il revenue potrebbe essere significativo ai fini dell'analisi dei dati; analogamente per l'indice ROE. Un'altra procedura potrebbe essere quella di discretizzare i dati lavorando per intervalli. Un'altra ancora potrebbe essere quella di verificare se i dati per l'impresa siano presenti in altri parametri e, in caso affermativo, procedere a considerare i valori selezionando un data frame con colonne più complete.

Tuttavia non abbiamo nessun modo per sostenere, con un qualche fondamento, una qualunque di queste ipotesi. Come prima metodologia di lavoro si decide di escludere dal campione le aziende che non hanno valori riservandoci di investigare le cause della mancanza di dati. Procediamo quindi a individuare quali sono i dati mancanti e a settare correttamente i valori inserendoli in un data frame.
Si utilizza la funzione 'delNA':

```
#funzione che elimina tutte le aziende con valori NA
delNA <- function(aziende)
{
  size <- dim(aziende)
  valoriNA <- is.na(aziende[, 2:size[2]])
  del <- apply(valoriNA, 1, any)
  Aziende <-aziende[!del,1]
  colonne <- aziende[!del, 2:size[2]]
  aziendeAll <- cbind(Aziende, colonne)
  intervallo <- range(aziendeAll[,2])
  aziendeAll[,2] <-
    (aziendeAll[,2]-intervallo[1])/(intervallo[2]-intervallo[1])
  colnames(aziendeAll)<-c("Aziende","Rev2012","Lobby2012",
                          "ROE2012","nBtw2012","Clo2012","nDgr2012"
                          "E-I2012","GICSIndustry","GICSSubIndustry
                          "GICSCodeDescription")
  return(aziendeAll)
}
```

Figura 4.3: Funzione di eliminazione degli NA e standardizzazione del Revenue

All'interno della funzione si inizia con il recupero della dimensione del data frame 'aziende' con la funzione 'dim'. Essa restituirà un valore che verrà passato alla funzione 'is.na' che fornisce una metodologia sicura per la gestione degli NA:

```
size <- dim(aziende)
valoriNA <- is.na(aziende[, 2:size[2]])
```

Tale funzione assegna ad ogni cella del data frame un FALSE se il dato è diverso da NA, e TRUE, se manca l'elemento nella cella.

Vogliamo mantenere solo le righe che contengono tutti gli elementi quindi si filtra valoriNA per ottenere tutti FALSE. Poichè la funzione 'any', applicata ad un vettore, ritorna TRUE se esiste almeno un TRUE all'interno di esso (e quindi un NA), applichiamo questa funzione ad ogni riga della matrice e salviamo nel vettore 'del':

```
del <- apply(valoriNA, 1, any)
```

Si selezionano le righe con le aziende che non contengono NA quindi si applica la negazione di 'del' sia nella colonna delle aziende che nella matrice con i valori corrispettivi:

```
Aziende <-aziende[!del,1]
colonne <- aziende[!del, 2:size[2]]
```

Viene creato in seguito un nuovo dataframe con i nomi e i valori delle aziende di cui si ha Revenue, investimento in Lobby e R&D, ROE, indici vicinanza, settori e sottosettori utilizzando la funzione 'cbind' che combina in righe e colonne gli elementi passati:

```
aziendeAll <- cbind(Aziende, colonne)
```

Le righe successive descrivono come viene standardizzato il Revenue.

Questo valore di solito si esprime in miliardi di dollari USA (USD) e non può essere confrontato con parametri percentuali. Deve quindi subire un processo di standardizzazione che in questo caso avverrà portandolo ad un numero fra 0 e 1 con la seguente formula:

$$\frac{Rev - Rev_{min}}{Rev_{max} - Rev_{min}}$$

dove *Rev* rappresenta il valore del *Revenue* dell'azienda che vogliamo studiare, il Rev_{min} il *Revenue* minimo del data set e Rev_{max} il massimo.

Quindi usando la funzione 'range' che restituisce il minimo e il massimo di un vettore e in particolare, in posizione 1 il minimo e in 2 il massimo, considerando sempre la seconda colonna cioè il Revenue ([,2]) si avrà:

```
intervallo <- range(aziendeAll[,2])
aziendeAll[,2] <-
  (aziendeAll[,2]-intervallo[1])/(intervallo[2]-intervallo[1])
```

Infine la funzione ritorna i valori delle aziende correttamente salvate per l'analisi dei dati:

```
return(aziendeAll)
```

Dopo aver effettuato la lettura dei dati si può ora utilizzare la funzione 'delNA':

```
aziendeLette <- leggiAnno(percorso)
aziendeSel <- delNA(aziendeLette)
```

	row.names	Aziende	Rev2012	Lobby2012	ROE2012	nBtw2012	Clo2012
1	2	3M COMPANY	0.063026291	0.0198	25.29	0.5451	0.7483
2	3	ABBOTT LABORATORIES	0.044335932	0.0198	22.32	0.3541	0.7434
3	8	AECOM TECHNOLOGY CORPORATION	0.014831789	0.0029	-2.70	0.0004	0.5022
4	13	AGCO CORP	0.018707694	0.0075	15.07	0.0119	0.5407
5	14	AGILENT TECHNOLOGIES INC	0.011808929	0.0007	22.25	0.1804	0.6726
6	15	AIR PRODUCTS & CHEMICALS INC	0.017928744	0.0052	18.02	0.0242	0.5512
7	20	ALCOA INC	0.049238540	0.0133	1.45	0.0004	0.5322
8	21	ALLEGHENY ENERGY INC	0.006939435	0.0026	6.39	0.0308	0.6360
9	23	ALLERGAN INC	0.009116717	0.0312	18.82	0.0040	0.5045
10	24	ALLIANT TECHSYSTEMS INC	0.006262148	0.0358	18.09	0.1191	0.7048
11	29	ALTRIA GROUP, INC.	0.051278701	0.0429	131.94	0.3757	0.7567
12	39	AMERISOURCEBERGEN CORP	0.173225174	0.0016	29.27	0.0837	0.6835

Figura 4.4: Campione di dati GICS_2012 dopo la standardizzazione e filtro degli NA

Capitolo 5

Analisi quantitativa e la regressione sui dati GICS 2012

5.1 Statistica Descrittiva

Come prima analisi dei dati globali nell'anno 2012 definisco una funzione che stima media, varianza, deviazione standard, massimo e minimo effettuando un'analisi preliminare descrittiva per vedere il comportamento e il range di azione dei parametri. Utilizzo la funzione 'statisticaFile':

```
#funzione che calcola la media per ogni colonna
#In: nome file con percorso, colonna iniziale e finale
#Out: data.frame con valori ottenuti
statisticaFile <- function(aziende, cIni, cFin)
    {
     dim <- length(aziende[,cIni:cFin])
     media <- apply(aziende[,cIni:cFin], 2, mean, na.rm=TRUE)
     varianza <- apply(aziende[,cIni:cFin], 2, var, na.rm=TRUE)
     devStd <- sqrt(varianza)
     minimo <- apply(aziende[,cIni:cFin], 2, min, na.rm = TRUE)
     massimo <- apply(aziende[,cIni:cFin], 2, max, na.rm = TRUE)
     statDescr <- data.frame(media, varianza, devStd, minimo,
                                + massimo)
     colnames <- c("media", "varianza", "dev. std", "minimo",
                    "massimo")
     return (statDescr)
     }

descrStat <- statisticaFile(aziende, 1, 5)
descrStat
```

Figura 5.1: Funzione che calcola media, varianza, deviazione standard, massimo e minimo nell'anno 2012

Si ottengono inizialmente le colonne in esame dal data frame 'aziende' passato come parametro:

```
dim <- length(aziende[,cIni:cFin])
```

Successivamente si ha il calcolo della media, varianza, deviazione standard, minimo e massimo utilizzando la funzione 'apply' che restituisce un vettore o una matrice applicando una funzione ad un array passato. Si applica alla matrice 'aziende' dalla colonna iniziale a quella finale considerate, specificando:

- un 'margin' cioè se la funzione 'apply' si applica a righe o colonne (1 o 2, in questo caso 2)
- il terzo parametro è la funzione 'mean', 'var', 'min' e 'max' che di default R mette a disposizione (la deviazione standard viene calcolata come radice quadrata della varianza con 'sqrt')
- esclusione di eventuali valori NA con la funzione 'na.rm = TRUE'

```
media <- apply(aziende[,cIni:cFin], 2, mean, na.rm=TRUE)
varianza <- apply(aziende[,cIni:cFin], 2, var, na.rm=TRUE)
devStd <- sqrt(varianza)
minimo <- apply(aziende[,cIni:cFin], 2, min, na.rm = TRUE)
massimo <- apply(aziende[,cIni:cFin], 2, max, na.rm = TRUE)
```

Si raccolgono i dati statistici descrittivi in un data frame e si applicano le intestazioni alle colonne:

```
statDescr <- data.frame(media, varianza, devStd, minimo,
                                + massimo)
colnames <- c("media", "varianza", "dev. std", "minimo",
                "massimo")
```

Si applica infine il 'return' per avere il risultato.
Utilizziamo quindi la funzione passando come data frame le 'aziende' precedentemente filtrate dagli NA (capitolo 3 - sezione 3.3) selezionando le colonne dalla 1 alla 5 comprese che corrispondono a Rev2012, Lobby2012, nBtw2012 ,Clo2012, nDgr2012; si esegue successivamente il risultato lanciando la funzione creata:

```
descrStat <- statisticaFile(aziende, 1, 5)
descrStat
```

I risultati sono i seguenti:

```
##                media    varianza     devStd minimo X.massimo
## Rev2012   0.06521447 0.013293480 0.11529736   0.00    1.0000
## Lobby2012 0.01445645 0.000218202 0.01477166   0.00    0.0714
## nBtw2012  0.16676532 0.053494963 0.23128978   0.00    1.2175
## Clo2012   0.62454919 0.015814679 0.12575643   0.25    0.8289
## nDgr2012  0.84349032 0.625039509 0.79059440   0.00    4.0474
```

5.2 Grafici di densità per la Lobby%

Passiamo ora alla distribuzione di densità per la lobby % cioè si illustra come vengono distribuiti i dati globali nell'anno 2012 tramite i seguenti comandi usando la libreria *ggplot*2 creata nel 2005 da Hadley Wickham[17] che permette di generare grafici decisamente più utili per l'analisi rispetto a quelli che si possono ottenere con il pacchetto grafico contenuto nella distribuzione base di R.

La libreria gestisce tantissimi parametri in grado di rendere estremamente significativo il risultato finale. La sua logica di funzionamento però non è proprio intuitiva, soprattutto perché è veramente diversa da quella a cui un utente R è comunemente abituato.

La libreria ggplot2 è un'implementazione della cosiddetta "Grammar of Graphics" di Wilkinson[34] . Tale "grammatica" consiste in uno schema generale da applicare alla visualizzazione dei dati, che permette di organizzare un grafico attraverso la combinazione di componenti semantiche distinte, come oggetti geometrici, scale e coordinate. Essa si adatta all'integrazione di strati grafici all'interno di R. In sostanza, la grammatica ci racconta che un grafico statistico è una mappatura dai dati, tramite attributi estetici (colore, forma, dimensione) a oggetti geometrici (punti, linee, bar). La trama può contenere anche trasformazioni statistiche dei dati e viene disegnata su uno specifico sistema di coordinate.

Questo rende ggplot2 molto potente perché non limitato ad una serie di elementi grafici pre-specificati, ma è possibile creare una nuova grafica appositamente per la nostra analisi.

```
require("gridExtra")
require("ggplot2")
y <- "densit"
x <- "lobbying"
g0 <- ggplot(aziende)
b <- "blue"
x <- "Lobby2012"
g2 <- g0 + geom_density(aes(x = Lobby2012), fill = b, col = b)
g2 <- g2 + ggtitle("Lobby%") + ylab(y) + xlab(x)
grid.arrange(g2, ncol = 1)
```

Si applica tale metodologia ai nostri dati visualizzando la distribuzione della Lobby % globale per l'anno 2012 lavorando un oggetto *ggplot* al quale viene passato il data frame:

```
g0 <- ggplot(aziende)
```

A tale oggetto viene calcolata la *geom_density*, che è funzione di densità di default in R, e in aggiunta le caratteristiche estetiche come il colore blu.

[34]Leland Wilkinson è uno statistico e scienziato informatico di Software Tableau. È Professore di Informatica presso l'Università dell'Illinois a Chicago. Wilkinson ha sviluppato il pacchetto statistico SYSTAT nei primi anni 1980 venduto a SPSS nel 1995 - Informazioni più dettagliate su http://en.wikipedia.org/wiki/Leland_Wilkinson

Viene specificato il parametro analizzato con *ggtitle* e passato l'oggetto finale alla funzione *grid.arrange* che organizza i dati degli oggetti *ggplot* specificando inoltre un'unica colonna di visualizzazione con $ncol = 1$ per evitare disambiguità e imprecisioni negli intervalli:

```
g2 <- g0 + geom_density(aes(x = Lobby2012), fill = b, col = b)
g2 <- g2 + ggtitle("Lobby%") + ylab(y) + xlab(x)
grid.arrange(g2, ncol = 1)
```

Questo è il risultato:

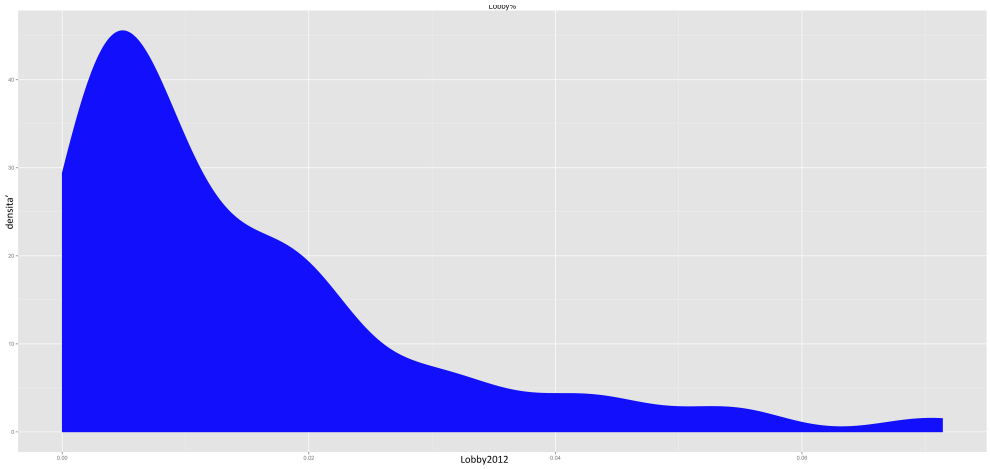


Figura 5.2: Funzione di densità continua in un unico intervallo per l'indice Lobby nell'anno 2012

Per avere una panoramica del comportamento grafico dell'investimento in attività di lobbying nei settori Gics si è adottata una visualizzazione della distribuzione decomposta di istogrammi utilizzando la seguente funzione:

```
ggplot(aziende, aes(x=Rev2012,
                            fill=Lobby2012)) +
  geom_histogram(binwidth=2) + labs(x="Lobby2012") +
  facet_wrap(~GICSIndustry)
```

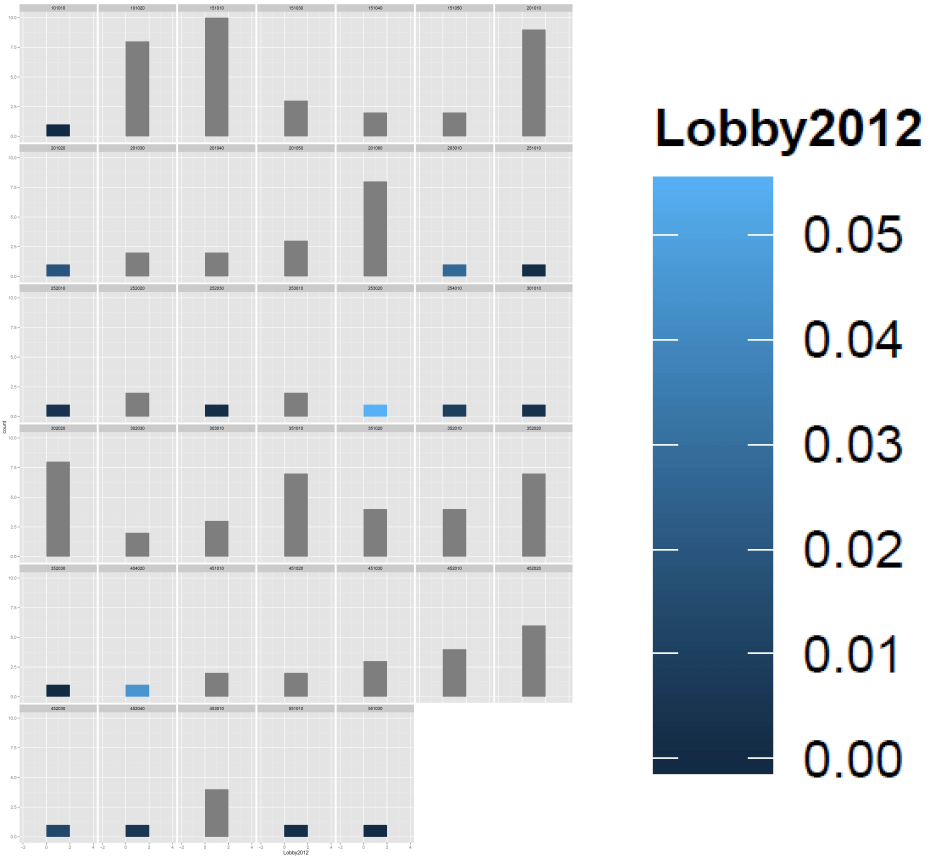
Figura 5.3: Funzione di densità in un unico intervallo per l'indice Lobby nell'anno 2012

Si utilizza sempre la funzione 'ggplot' dove si prende in considerazione il data frame 'aziende' e le caratteristiche estetiche dell'oggetto di analisi con la funzione 'aes', in

particolare:

- `x=Rev2012`: che identifica la variabile di distribuzione, il Rev2012
- `fill=Lobby2012`: con il quale si indicano le sfumature di colore a seconda della variazione della 'Lobby2012'
- `geom_histogram(binwidth=2) + labs(x="Lobby2012")` : dove si indica la tipologia di grafico, in questo caso un istogramma con larghezza della colonna 2 unità ('binwidth'), intestazione della x con la funzione 'labs' e intestazione di ogni grafico con il codice GICS corrispondente

Il risultato è il seguente:



Il grafico, a sinistra mostra il livello di Revenue standardizzato tra 0 e 1 e colorato a seconda della media del gruppo industriale (GICSIndustry) per l'investimento in attività di lobbying nell'anno 2012. Non risulta agevole visualizzare il dettaglio globale delle industrie per la dimensione del grafico in output. Da segnalare che, in questo caso è stato utilizzato un `facet_wrap(~GICSIndustry)` cioè si effettua l'analisi e si visualizzano gli istogrammi per ogni industria, il quale potrebbe essere una buona tecnica per vedere la distribuzione globale dei dati ma non risulterebbe agevole la comprensione in quanto il *GICSIndustry* risulta essere un numero.

Si può quindi sostituire con `facet_wrap(~GICSCodeDescription)` che fornisce un numero molto elevato di dati ma analizza la distribuzione delle sotto industrie in maniera più chiara. In dettaglio i gruppi industriali che hanno un alto livello di investimento in attività di lobbying sul fatturato annuo non risultano poi essere quelli dove si riscontra un revenue standardizzato elevato:

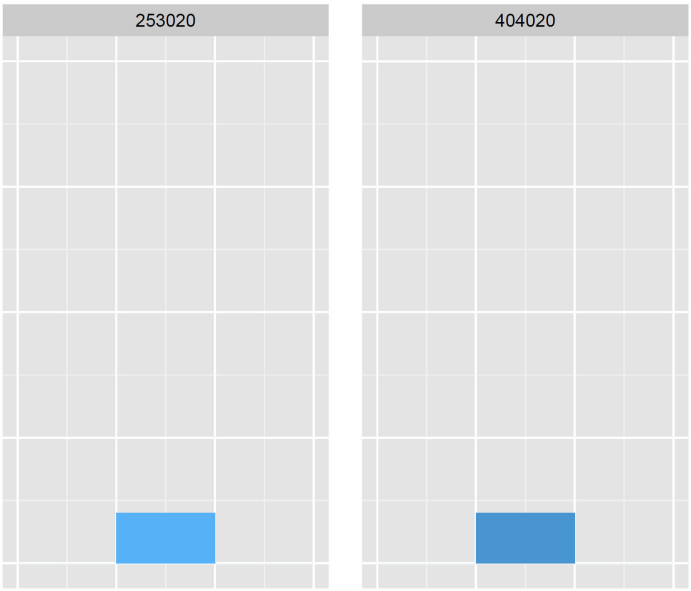


Figura 5.4: Gruppi industriali con il più alto livello di investimento in attività di lobbying nell'anno 2012

Le industrie corrispondenti sono:

- 253020: *Diversified Consumer Services*[35], sono le aziende che forniscono servizi di istruzione, online o attraverso metodi di insegnamento tradizionali. Include università private, insegnamento per corrispondenza, i fornitori di corsi di formazione, materiale didattico e formazione tecnica. Sono escluse le società di formazione per i dipendenti classificati nel sub-settore Risorse Umane e Servizi per l'impiego.
  Il gruppo include anche le aziende che forniscono servizi di consumo come servizi residenziali, sicurezza domestica, servizi legali, personali, di ristrutturazione e interior design, le aste di consumo, i servizi per matrimoni e funerali.
- 404020: *Real Estate Investment Trusts*[36], investimenti immobiliari che comprendono aziende impegnate nell'acquisizione, sviluppo, proprietà, leasing, gestione e funzionamento di proprietà industriale. Include società operative in

[35]Standard & Poor's, MSCI (2013) - Gics Mapbook electronic 0711 - pagina 14
[36]Standard & Poor's, MSCI (2013) - Gics Mapbook electronic 0711 - pagina 22

capannoni industriali e le proprietà di distribuzione.
In più attività che hanno origine con l'acquisto o gestione di mutui residenziali e / o commerciali. Vi sono inoltre gruppi di investimento in titoli garantiti da ipoteche e altre attività legati ai mutui. Si intendono sia uffici abitazioni plurifamiliari, appartamenti, case prefabbricate e proprietà abitative degli studenti.
Sono comprese anche le società che operano e investono nella sanità, tempo libero, hotel / resort e le proprietà di stoccaggio. Esse generano la maggior parte dei loro ricavi e proventi da operazioni di noleggio e di leasing immobiliare.

5.3 Correlazione degli indici GICS 2012

Dopo avere effettuato un'analisi distributiva dei dati con il particolare focus sull'andamento dell'attività di lobbying all'interno dei dati GICS, sarebbe importante, se non fondamentale, la costruzione di un modello di previsioni di questi dati.

Non è lo scopo della tesi, tuttavia vengono qui illustrate le basi per poter effettuare tale modello.

La prima metodologia affrontata è il calcolo della correlazione fra le variabili cioè capire se esistono relazioni fra i dati. Nel nostro caso, per esempio, fra tutte le possibili coppie dei cinque dati su Rev2012, Lobby2012, nBtw2012 ,Clo2012, nDgr2012; Si ricorda che la correlazione è definita come:

$$c_{xy} = \left(\frac{1}{N-1}\right) \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x s_y},$$

dove, al numeratore, appare la somma dei prodotti delle differenze dei valori delle due variabili dalle loro medie e, al denominatore il prodotto delle deviazioni standard delle due variabili. La correlazione è compresa fra 1 e -1. Valori alti positivi implicano che le due variabili sono strettamente correlate, valori negativi bassi implicano una correlazione inversa fra le due variabili, mentre valori prossimi allo zero implicano la non esistenza di relazioni fra i dati. Forse è utile sottolineare che l'esistenza di una correlazione non comporta una relazione di causalità fra le due variabili. Per calcolare la correlazione fra due variabili si può usare la funzione 'cor' considerando sempre il data frame di origine 'aziende':

```
cor(aziende[, 1:5])
```

```
##              Rev2012  Lobby2012  nBtw2012   Clo2012  nDgr2012
## Rev2012    1.0000000 -0.1862429 0.4942748 0.2937936 0.5219704
## Lobby2012 -0.1862429  1.0000000 0.1999754 0.3123138 0.2771024
## nBtw2012   0.4942748  0.1999754 1.0000000 0.6219391 0.7522848
## Clo2012    0.2937936  0.3123138 0.6219391 1.0000000 0.7941585
## nDgr2012   0.5219704  0.2771024 0.7522848 0.7941585 1.0000000
```

Per avere una maggiore chiarezza si organizzano i dati in colonne con la funzione 'melt':

```
correlazioni <- cor(aziende[,1:5])
organizzaDati <- melt(correlazioni, varnames = c("x", "y"))
organizzaDati <- organizzaDati[order(organizzaDati$value),]
organizzaDati
```

```
##            x         y      value
## 2  Lobby2012   Rev2012 -0.1862429
## 6    Rev2012 Lobby2012 -0.1862429
## 8   nBtw2012 Lobby2012  0.1999754
## 12 Lobby2012  nBtw2012  0.1999754
## 10  nDgr2012 Lobby2012  0.2771024
## 22 Lobby2012  nDgr2012  0.2771024
## 4    Clo2012   Rev2012  0.2937936
## 16   Rev2012   Clo2012  0.2937936
## 9    Clo2012 Lobby2012  0.3123138
## 17 Lobby2012   Clo2012  0.3123138
## 3   nBtw2012   Rev2012  0.4942748
## 11   Rev2012  nBtw2012  0.4942748
## 5   nDgr2012   Rev2012  0.5219704
## 21   Rev2012  nDgr2012  0.5219704
## 14   Clo2012  nBtw2012  0.6219391
## 18  nBtw2012   Clo2012  0.6219391
## 15  nDgr2012  nBtw2012  0.7522848
## 23  nBtw2012  nDgr2012  0.7522848
## 20  nDgr2012   Clo2012  0.7941585
## 24   Clo2012  nDgr2012  0.7941585
```

Come si può notare dai risultati le variabili *Lobby*2012 e *Rev*2012 sono correlate negativamente quindi ciascun valore della prima variabile non corrisponda con una "certa regolarità" al valore della seconda; ripeto non si tratta necessariamente di un rapporto di causa-effetto, ma semplicemente si ha che la tendenza della variabile *Lobby*2012 non varia direttamente rispetto alla *Rev*2012.

Al contrario si vede che la variabile *Clo*2012 sembra più correlata con la variabile *Lobby*2012 rispetto alla *Rev*2012.

Mentre la $nBtw2012$ e l'$nDgr2012$ sono più correlati con la *Rev*2012 rispetto alla *Lobby*2012. Per avere una visione di insieme ancora più intuitiva si può ricorrere all'utilizzo del pacchetto *ggplot*2 (come nel paragrafo 4.2) e utilizzare una scala di colori:

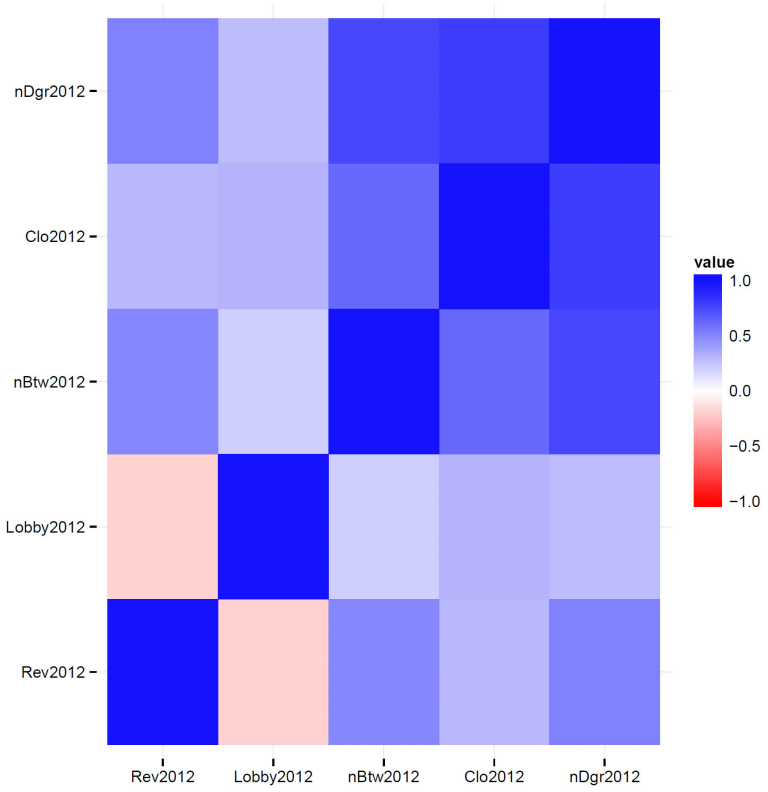


Figura 5.5: Grafico a colori di correlazione fra variabili Gics anno 2012

5.4 Tecniche di regressione sui valori GICS 2012

La regressione è una tecnica che formalizza e risolve il problema di una relazione funzionale tra variabili misurate sulla base di dati campionari estratti da un'ipotetica popolazione infinita. Originariamente Galton[37] utilizzava il termine come sinonimo di correlazione, tuttavia oggi in statistica l'analisi della regressione è associata alla risoluzione del modello lineare.

Per la loro versatilità, le tecniche della regressione lineare trovano impiego nel campo delle scienze applicate: chimica, geologia, biologia, fisica, ingegneria, medicina, nonché nelle scienze sociali: economia, linguistica, psicologia e sociologia.

5.4.1 Regressione Lineare

Più formalmente la regressione lineare[20] è un metodo statistico utilizzato per modellare la relazione lineare tra due variabili quantitative a fini esplicativi o di previsione. Se si può ipotizzare l'esistenza di una dipendenza lineare ad esempio di Y da X, si può affermare che le osservazioni della variabile Y si possono ottenere, a meno di un errore (o residuo), da una funzione lineare delle osservazioni della variabile X. Per ciascuna osservazione avremo quindi:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

dove la variabile Y viene detta variabile risposta (o variabile dipendente) la variabile X viene detta variabile esplicativa (o variabile indipendente o regressore) e ε è l'errore di approssimazione. Il β_0 corrisponde all'intercetta di regressione, β_1 il coefficiente angolare. Possiamo stimare questi parametri utilizzando il metodo dei minimi quadrati:

$$(\hat{\beta_0}, \hat{\beta_1}) = \text{argmin} \sum_{i=1}^{n} (y - \beta_0 - \beta_1 x_i)^2$$

Stimati i parametri otteniamo la retta di regressione lineare:

$$f(x) = \hat{\beta_0} + \hat{\beta_1} x_i$$

dalla quale può essere effettuata la previsione. Gli "smoothed values" sono definiti da:

$$\hat{y_i} = \hat{\beta_0} + \hat{\beta_1} x_i$$

e i residui dalla seguente formula:

$$\hat{\varepsilon_i} = y_i - \hat{y_i}$$

L'analisi dei residui è essenziale ed è usata sia per l'individuazione degli outlier sia per il "fitting" globale del modello.

[37] Sir Francis Galton (Sparkbrook - Birmingham, 16 febbraio 1822 – Haslemere, 17 gennaio 1911) è stato un esploratore, antropologo e climatologo britannico e patrocinatore dell'eugenetica, termine da lui creato. Maggiori informazioni su: http://it.wikipedia.org/wiki/Francis_Galton

5.4.2 Regressione Multipla

La regressione lineare multipla ha la capacità esplicativa e di previsione della variabile quantitativa Y da "p" variabili $X_1, ..., X_p$. Tale modello è una generalizzazione della regressione lineare semplice.
Supponiamo che il modello abbia tale conformazione:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i,$$
$$i = 1, ..., n,$$

dove x_{ij} sono numeri conosciuti misurando le variabili esplicite dalla matrice X detta matrice sperimentale. I parametri β_j del modello sono sconosciuti e devono essere stimati. Il parametro β_0 (l'intercetta) corrisponde alla costante del modello. Gli ϵ_i sono variabili random sconosciute e rappresentano gli errori di misurazione.
Se scriviamo il modello in forma di matrice otteniamo:

$$Y = X\beta + \epsilon$$

$$\mathbb{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Da queste osservazioni stimiamo i parametri del modello utilizzando il metodo dei minimi quadrati:

$$\hat{\beta} = \underset{\beta_0,\cdots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} (\mathbb{Y} - \mathbb{X}\beta)'(\mathbb{Y} - \mathbb{X}\beta)$$

Se la matrice ha rango massimo e quindi che le variabili esplicite non sono collineari (vettori linearmente indipendenti), l'estimatore di minimo quadrato $\hat{\beta}$ di β è

$$\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}$$

Una volta che i parametri sono stati stimati, calcoliamo i valori di fit:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$$

e la previsione di nuovi valori. La differenza fra i valori osservati e i valori di fit è per definizione il valore dei residui:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

Si illustrano i modelli più significativi per verificare le possibili correlazioni fra le variabili sui dati globali GICS_2012: *Lobby*2012, *Rev*2012, $nBtw$2012 e $nDgr$2012:
Per primo si studia un modello per verificare la relazione fra l'investimento in attività di lobbing e il revenue:

```
modelA <- lm(aziende$Rev2012 ~ aziende$Lobby2012)
summary(modelA)

##
## Call:
## lm(formula = aziende$Rev2012 ~ aziende$Lobby2012)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08140 -0.05621 -0.03401  0.01964  0.91799
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.08623    0.01432   6.021 1.88e-08 ***
## aziende$Lobby2012 -1.45368    0.69430  -2.094   0.0384 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1137 on 122 degrees of freedom
## Multiple R-squared:  0.03469,Adjusted R-squared:  0.02677
## F-statistic: 4.384 on 1 and 122 DF,  p-value: 0.03835
```

Come si può notare, il $Pr(>|t|)$ (definito come $p-value$) del coefficiente per la variabile *Lobby*2012, 0.0384, che stima l'effetto della variabile *Lobby*2012 sul modello, non risulta essere significativo; usualmente sono considerati pessimi i valori del $p-value$ sopra lo 0.05.
Inoltre il coefficiente stimato per *Lobby*2012 incide negativamente sulla *Rev*2012 avendo un valore di -1.45. Quindi la variabile *Lobby*2012 non risulta essere correlata direttamente con il *Rev*2012.
D'altronde non ci si aspetta che l'attività di lobbying e la revenue, cioè il reddito, di un'impresa siano correlati sullo stesso anno; per vedere l'effetto dell'investimento si devono aspettare probabilmente più anni.

Come secondo modello si prendono in considerazione la $Lobby2012$ con la $Clo2012$:

```
modelB <- lm(aziende$Lobby2012 ~ aziende$Clo2012)
summary(modelB)
```

```
##
## Call:
## lm(formula = aziende$Lobby2012 ~ aziende$Clo2012)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.019492 -0.009440 -0.003335  0.004491  0.054370
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.008455   0.006435  -1.314 0.191344
## aziende$Clo2012  0.036685   0.010103   3.631 0.000413 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01409 on 122 degrees of freedom
## Multiple R-squared:  0.09754,Adjusted R-squared:  0.09014
## F-statistic: 13.19 on 1 and 122 DF,  p-value: 0.0004135
```

Qui si può notare, sempre dal $Pr(> |t|)$ (uguale a 0.000413) che il coefficiente stimato per la variabile $Clo2012$ è abbastanza significativo e che quindi si ha una relazione positiva con la $Lobby2012$. Tuttavia il coefficiente stimato è basso (0.037), quindi all'aumentare di un'unità di $Clo2012$ la variabile $Lobby2012$ aumenta di soli 0.037; si ha poi un Multiple R-squared di 0.1 quindi solo il 10% dei dati sono descritti da questo modello.
Si prende anche in considerazione il $Rev2012$ con $nBtw2012$ e $nDgr2012$:

```
modelC <- lm(aziende$Rev2012 ~ aziende$nBtw2012 + aziende$nDgr2012)
summary(modelC)
```

```
##
## Call:
## lm(formula = aziende$Rev2012 ~ aziende$nBtw2012 + aziende$nDgr2012)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18676 -0.05374 -0.00222  0.01686  0.66672
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.003208   0.012880   0.249   0.8037
## aziende$nBtw2012 0.116686   0.057704   2.022   0.0454 *
## aziende$nDgr2012 0.050442   0.016881   2.988   0.0034 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09752 on 121 degrees of freedom
## Multiple R-squared:  0.2962,Adjusted R-squared:  0.2846
## F-statistic: 25.47 on 2 and 121 DF,  p-value: 5.879e-10
```

Non si hanno relazioni significative se non un aumento del Multiple R-squared. Con tale modello quindi si descrive quasi il 30% dei dati.

Capitolo 6

Dinamica del rapporto fra lobby e revenue nei settori GICS

Il capitolo finale si occupa di mostrare la correlazione fra le variabili *lobby* e *revenue* nel periodo 2007 - 2012 e l'influenza che questi valori hanno all'interno delle seguenti industrie:

- Information Tecnology: Aziende in via di sviluppo e commercializzazione di software Internet e/o che forniscono servizi di rete, tra cui banche dati on-line e servizi interattivi, di registrazione di indirizzi web, costruzione di database e servizi di progettazione wireless. Qui sono compresi:
  - i fornitori di information technology e servizi di consulenza tecnologica, di informazioni e servizi di gestione delle informazioni;
  - produttori di software di intrattenimento domestico e software didattico utilizzato principalmente in casa;
  - produttori di personal computer, server, mainframe e workstation. Comprende i produttori di *Automatic Teller Machines* (ATM), tecnologie di supporto dati e periferiche;
  - produttori di componenti e periferiche per computer elettronici. Include i componenti di storage dei dati, schede madri, audio e video carte, monitor, tastiere, stampanti e altre periferiche;
  - produttori di attrezzature per semiconduttori.
- Health Care: Sono i produttori di apparecchiature di assistenza sanitaria e di dispositivi medici, sistemi di drug delivery, dispositivi cardiovascolari e ortopedici e apparecchiature diagnostiche. Comprendono:
  - i produttori di forniture tecnologiche e servizi di assistenza sanitaria dei pazienti. Include centri dialisi, servizi di analisi di laboratorio e servizi di gestione per laboratori;
  - fornitori di servizi di supporto, servizi di agenzia di raccolta, servizi di personale e di vendita e servizi di marketing in outsourcing;

  - i proprietari e gli operatori di strutture sanitarie, tra cui ospedali, case di cura, centri di riabilitazione e di riposo e ospedali animali e operatori di Health Maintenance Organizations (HMO);
  - le sotto industrie che si occupano di di tecnologia e sanità in particolare servizi di informazione, assistenza sanitaria, società che forniscono applicazioni, sistemi e / o software di elaborazione dati, strumenti basati su Internet, e servizi di consulenza IT a medici, ospedali o imprese che operano principalmente nel settore della sanità.
  - le aziende impegnate principalmente nella ricerca, sviluppo, produzione e / o commercializzazione di prodotti farmaceutici,l'ingegneria genetica per lo studio e il trattamento di malattie umane e le industrie biotecnologiche.

- Industrials: Sono le industrie che si occupano costruzioni ingegneristiche, elettriche, aereospaziali e per il commercio. Comprendono:

  - sotto industrie che si occupano di componenti aereospaziali civili e militari;
  - produttori di componenti per l'edilizia;
  - appaltatori edilizi di grandi dimensioni;
  - costruzioni ingegneristiche civili come strade ponti o aereoporti;
  - produttori di autocarri pesanti, macchine di laminazione e macchine agricole
  - società che forniscono servizi commerciali di consulenza, ristorazione, riparazione, deposito;
  - imprese che si occupano di servizi ambientali e di manutenzione;
  - aziende che si forniscono trasporto aereo, marittimo e doganale;

6.1 Settore Information Technology

La metodologia di analisi sviluppata è la seguente:

- Selezione delle aziende che fanno parte del settore dell'IT utilizzando la funzione di *subset* che seleziona le imprese che hanno un codice GICSIndustry compreso fra 451010 e 453010 dal data frame di origine *aziende* nell'arco temporale fra il 2007 al 2012:

```
It2007 <- subset(aziende, (aziende$GICSIndustry > 451010 & aziende$GICSIndustry < 453010))
It2008 <- subset(aziende, (aziende$GICSIndustry > 451010 & aziende$GICSIndustry < 453010))
It2009 <- subset(aziende, (aziende$GICSIndustry > 451010 & aziende$GICSIndustry < 453010))
It2010 <- subset(aziende, (aziende$GICSIndustry > 451010 & aziende$GICSIndustry < 453010))
It2011 <- subset(aziende, (aziende$GICSIndustry > 451010 & aziende$GICSIndustry < 453010))
It2012 <- subset(aziende, (aziende$GICSIndustry > 451010 & aziende$GICSIndustry < 453010))
```

- Si calcolano le medie di settore per i parametri 'lobby' e 'revenue' per ogni anno utilizzando la funzione *statisticaFile* che prende in ingresso il dataframe (in questo caso saranno i 'subset' annuali) oggetto dell'analisi e le colonne iniziale e finale. Nella colonna 1 si avrà il *revenue* e nella colonna 2 il *lobby*%:

```
statisticaFile <- function(subset, cIni, cFin)
{
  media <- apply(subset[,cIni:cFin], 2, mean, na.rm=TRUE)
}

mediaRevLobby2007 <- statisticaFile(It2007,1,2)
mediaRevLobby2008 <- statisticaFile(It2008,1,2)
mediaRevLobby2009 <- statisticaFile(It2009,1,2)
mediaRevLobby2010 <- statisticaFile(It2010,1,2)
mediaRevLobby2011 <- statisticaFile(It2011,1,2)
mediaRevLobby2012 <- statisticaFile(It2012,1,2)
```

- Creazione del data frame con le colonne: Anni, MediaLobby, MediaRevenue.
  In *Anni* inserisco un vettore con gli anni (2007,2008,2009,2010,2011,2012);
  In *MediaLobby* inserisco un vettore con i valori delle medie per la 'lobby %' per ogni anno. In *MediaRevenue* inserisco un vettore con i valori delle medie per il 'revenue' per ogni anno. Unisco le 3 colonne con il metodo *cbind* e trasformo in data frame.

```
Anni <- c(2007,2008,2009,2010,2011,2012)
MediaLobby <- c(mediaRevLobby2007[2], mediaRevLobby2008[2],
                  mediaRevLobby2009[2], mediaRevLobby2010[2],
                  mediaRevLobby2011[2], mediaRevLobby2012[2])

MediaRevenue <-c(mediaRevLobby2007[1], mediaRevLobby2008[1],
                 mediaRevLobby2009[1], mediaRevLobby2010[1],
                 mediaRevLobby2011[1], mediaRevLobby2012[1])

ITmatrix <- cbind(Anni, MediaLobby,MediaRevenue)
ITdataframe <- as.data.frame(ITmatrix)
```

L'ITdataframe è il seguente:

	Anni	MediaLobby	MediaRevenue
1	2007	0.01558846	0.04882630
2	2008	0.01541304	0.05071884
3	2009	0.01652609	0.05796697
4	2010	0.01649091	0.06535400
5	2011	0.01405217	0.06248794
6	2012	0.01341765	0.08968873

- Illustrazione grafica che mostra la lobby% media su base annua. Tale grafico viene sviluppato con il pacchetto *ggplot*2 che mostra un grafico a barre colorate che ha sull'asse 'x' il tempo, cioè gli anni dal 2007 al 2012 e sulla 'y' i valori medi annuali della lobby%:
  Il codice sviluppato per l'illustrazione del grafico è il seguente:

```
y <- "% Lobbying"
x <- "Anni"
d <- ggplot(data=ITdataframe, aes(x=Anni, y=MediaLobby))
d + geom_bar(stat="identity", fill="forestgreen")
  + ggtitle("Lobby - periodo 2007-2012 - Settore IT") + ylab(y) + xlab(x)
```

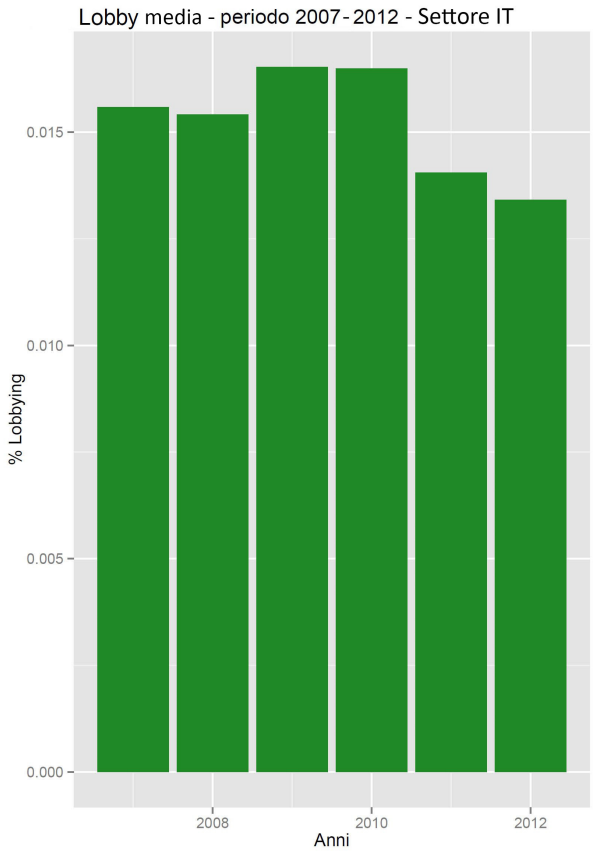


Figura 6.1: Grafico della lobby media nelle industrie del settore IT nel periodo 2007 - 2012

- Visualizzazione del grafico che mette in relazione il revenue su base annua e verifica dell'andamento utilizzando la stessa tecnica sopra descritta:

```
r <- "Reddito"
x <- "Anni"
d <- ggplot(data=ITdataframe, aes(x=Anni, y=MediaRevenue))
d + geom_bar(stat="identity", fill="darkred")
  + ggtitle("Reddito -  periodo 2007/2012 - Settore IT") + ylab(r) + xlab(x)
```

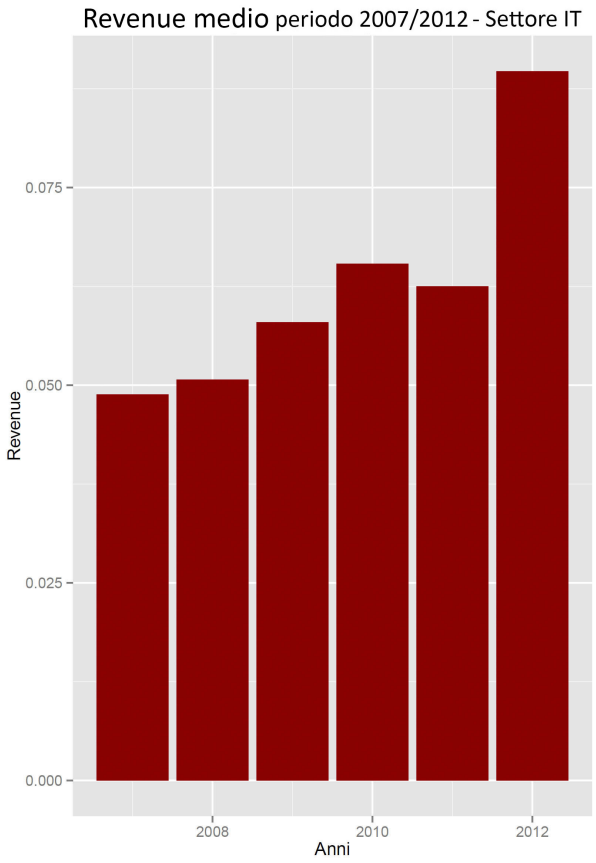


Figura 6.2: Grafico della revenue media nelle industrie del settore IT nel periodo 2007 - 2012

- Visualizzazione del grafico globale con la lobby%, il revenue e l'arco temporale; quest'ultimo è stato realizzato aggiungendo un carattere estetico all'oggetto ggplot che abbiamo analizzato permettendo una visualizzazione chiara dell'andamento. Più si va avanti negli anni e maggiore sembra essere il revenue medio standardizzato del settore e minore l'investimento medio in attività di lobbying:

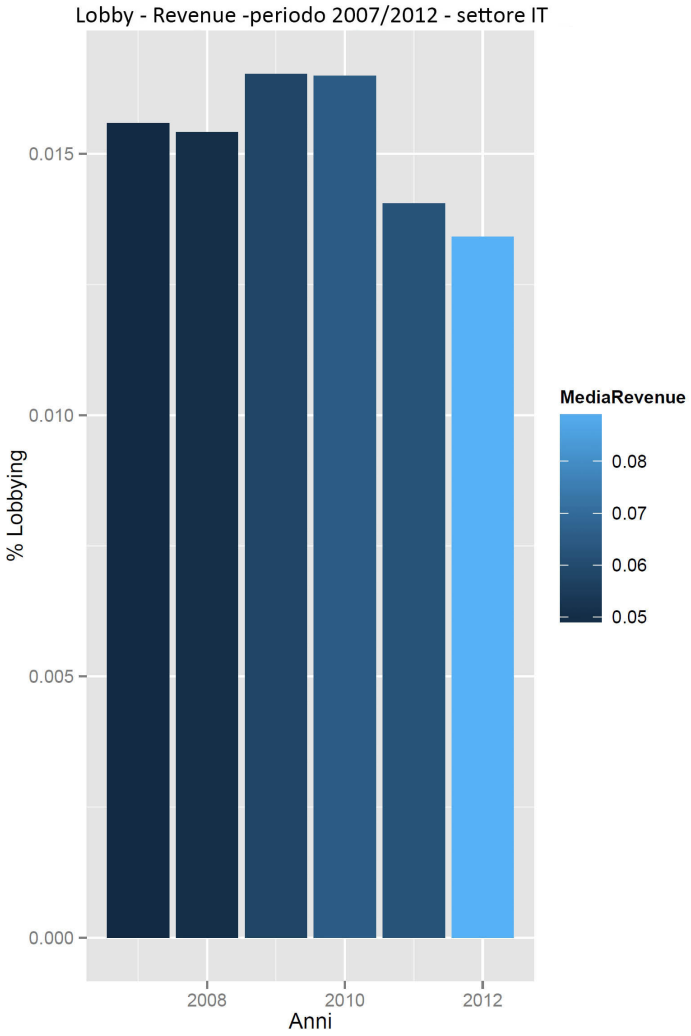


Figura 6.3: Grafico di relazione fra la lobby media e il revenue medio delle industrie nel periodo 2007 - 2012 sul settore IT

Analizzando i dati dal 2007 al 2012 si osserva che il picco di investimento in lobbying è fra il 2009 e il 2012, mentre il picco della revenue è ben evidente nel 2012.

I dati sembrano indicare che, come ci si aspetta, effetti dell'investimento in lobbying possano rilevarsi solo a distanza di tempo. Qui la scala di tempo potrebbe essere circa di 3 anni. Questa ipotesi andrebbe però verificata con un modello opportuno.

6.1.1 Ulteriori verifiche

Correlazione fra le variabili

A dimostrazione di questo comportamento effettuo l'analisi di correlazione fra le variabili revenue e lobby % di settore per ogni anno al fine di capire l'inversione di tendenza. Ci si aspetta che i primi anni la correlazione sia debolmente negativa e che nel periodo 2011/2012 risulti essere maggiormente negativa.

La metodologia adottata è la seguente:

- Calcolo della correlazione: con la funzione *subset* seleziono il settore dal data frame di partenza 'aziende'. In colonna '1' vi sono i valori standardizzati della revenue di ogni impresa, in colonna '2' i valori corrispondenti dell'investimento % in lobbying. Quindi se prendiamo in esame l'anno 2007 avremo che $it2007$ è il *subset* che applicato alla funzione *cor* nelle colonne 1 e 2 restituisce la correlazione fra le variabili:

```
correlazioni <- cor(It2007[,1:2])

                x          y      value
Lobby2007   Rev2007 -0.1713513
  Rev2007 Lobby2007 -0.1713513
```

  Tali valori sono stati organizzati in colonne 'x' , 'y' e value perchè verranno inseriti in un grafico colorato per permettere una chiarezza espositiva maggiore; a tal scopo utilizzo la funzione melt:

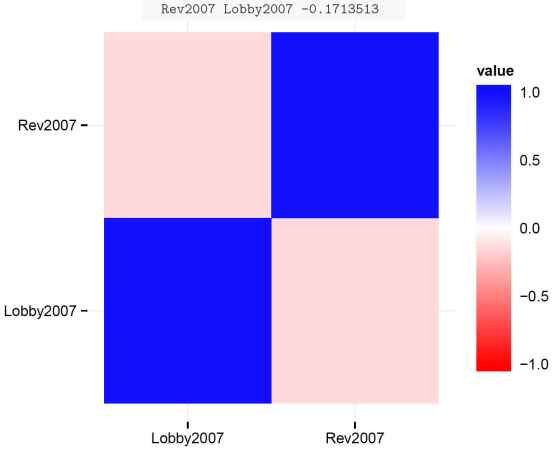
```
mescDati <- melt(correlazioni, varnames = c("x", "y"))
mescDati <- mescDati[order(mescDati$value),]
mescDati
```

- Grafico colorato per le correlazioni: si associano 2 colori, se la correlazione è tendente al rosso è negativa, al blu è positiva. Da notare ovviamente che le correlazioni fra se stessi sono pari a 1. Per effettuare questo grafico utilizzo un oggetto *ggplot* passando nelle caratteristiche estetiche i valori delle correlazioni con $aes(fill = value)$ visualizzati tramite un gradiente di colori rosso e blu con la funzione $scale\_fill\_gradient2$ e i valori limite delle correlazioni: 1 e -1:

```
ggplot(mescDati, aes(x=x, y=y)) + geom_tile(aes(fill=value)) +
  scale_fill_gradient2(low="red", mid="white", high="blue",
                       guide=guide_colorbar(ticks=FALSE, barheight=10),
                       limits=c(-1,1)) + theme_minimal() + labs(x=NULL, y=NULL)
```

Per l'anno 2007 il risultato è il seguente:

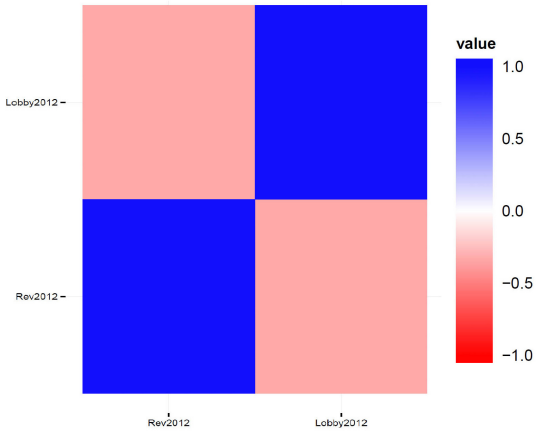
```
correlazioni <- cor(It2007[,1:2])

                x         y      value
Lobby2007   Rev2007 -0.1713513
  Rev2007 Lobby2007 -0.1713513
```



- Confronto la correlazione delle variabili con l'anno 2012:

2012

```
                x         y      value
Lobby2012   Rev2012 -0.3335436
  Rev2012 Lobby2012 -0.3335436
```



L'analisi nel tempo della correlazione fra investimento in lobbying e revenue non mostra variazioni significative, passando da un valore di -0.17 ad un valore di -0.33 nel 2012.

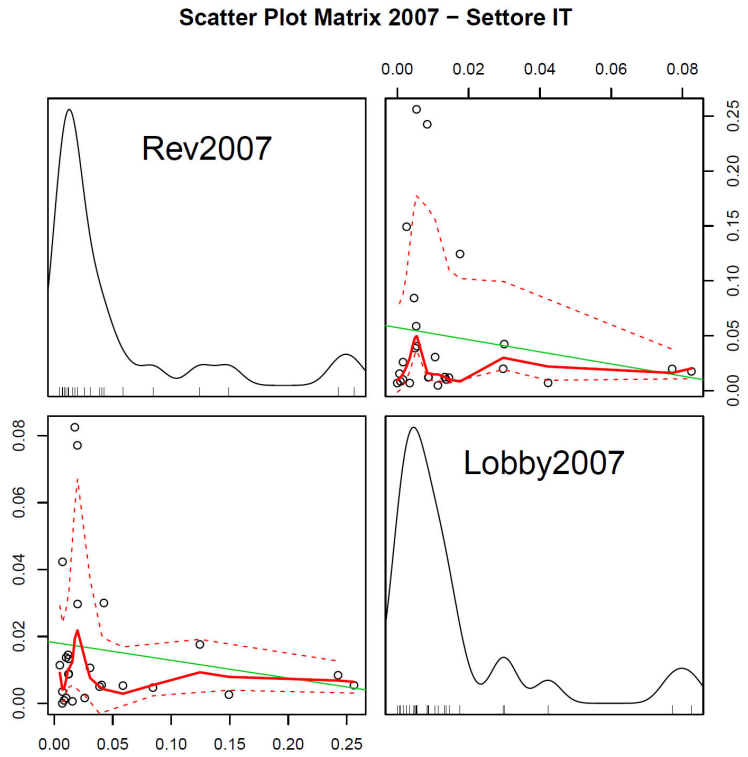
Regressione

Per la regressione si procede nel modo seguente:

- Verifica delle regressioni lineari e multiple per le variabili in esame nell'anno 2007 e 2012: è utile vedere la regressione lineare sulle variabili e come variano negli anni il p-value ($Pr(>|t|)$) dei coefficienti; Viene visualizzato inoltre lo $scatterplotMatrix$, grafico dove si evidenziano densità, distribuzione dei punti, una linea verde che è la retta di regressione lineare, una linea rossa che rappresenta la regressione non parametrica (qui si utilizza quella di default cioè la GAM) e l'intervallo fra le rette tratteggiate rosse che rappresenta lo *spread* dei dati.

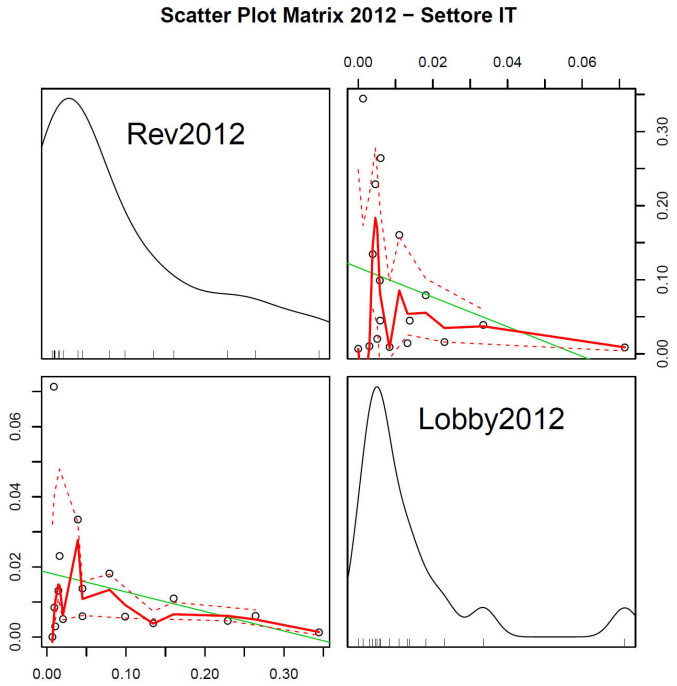
```
It2007 <- subset(aziende, (aziende$GICSIndustry > 451010 & aziende$GICSIndustry < 453010))
model <- lm(It2007$Rev2007 ~ It2007$Lobby2007)
summary(model)
car::scatterplotMatrix(It2007[1:2], spread=TRUE,
                       main="Scatter Plot Matrix 2007 - Settore IT")
```

```
## Call:
## lm(formula = It2007$Rev2007 ~ It2007$Lobby2007)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.050538 -0.040434 -0.024041  0.004766  0.201709
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.05744    0.01698   3.382  0.00246 **
## It2007$Lobby2007 -0.55250    0.64844  -0.852  0.40261
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06958 on 24 degrees of freedom
## Multiple R-squared:  0.02936,Adjusted R-squared:  -0.01108
## F-statistic: 0.726 on 1 and 24 DF,  p-value: 0.4026
```



Nell'anno 2012 la regressione e lo scatterplotmatrix sono i seguenti:

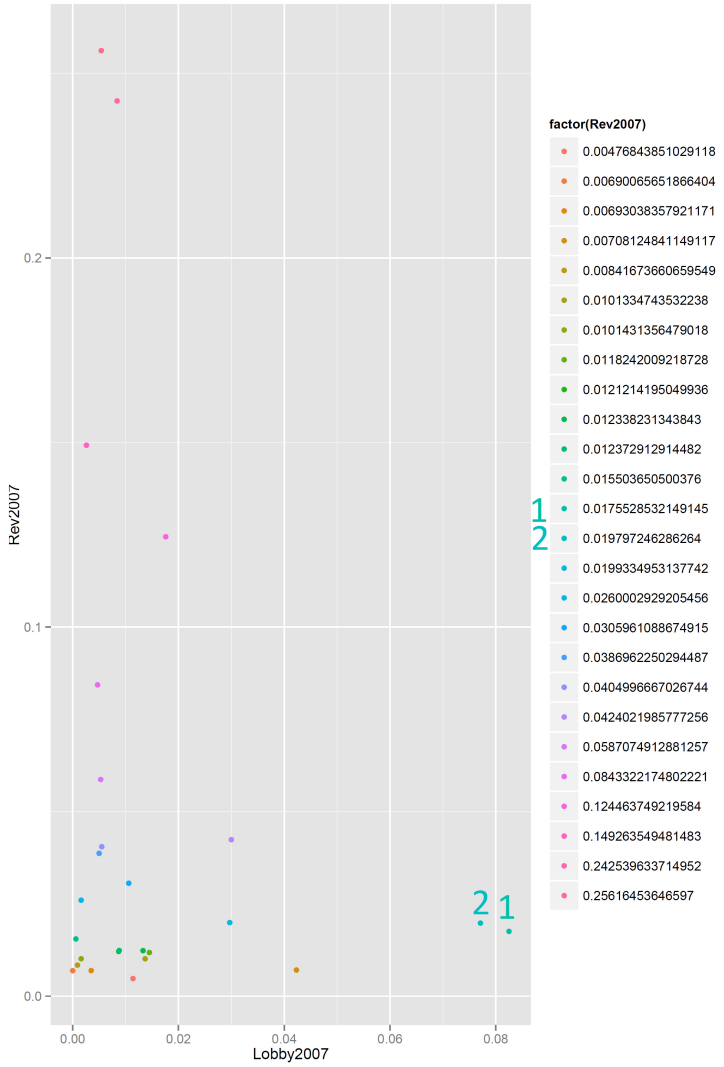
```
## Call:
## lm(formula = It2012$Rev2012 ~ It2012$Lobby2012)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10968 -0.07575 -0.01051  0.03482  0.23050
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.11649    0.03126   3.727  0.00202 **
## It2012$Lobby2012 -1.99759    1.45780  -1.370  0.19076
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1005 on 15 degrees of freedom
## Multiple R-squared:  0.1113,Adjusted R-squared:  0.052
## F-statistic: 1.878 on 1 and 15 DF,  p-value: 0.1908
## F-statistic: 2.721 on 1 and 21 DF,  p-value: 0.1139
```



Come si può notare la distribuzione della revenue nel 2007 risulta essere multimodale suggerendo probabilmente diversi raggruppamenti di investimento. Nel 2012 si osserva una distribuzione con un numero decisamente minore di picchi. Questo potrebbe suggerire, per esempio, un tentativo di razionalizzazione dei finanziamenti in lobbying. Questa ipotesi però richiederebbe un'approfondita analisi per sotto gruppi di industrie.

Scatterplot della distribuzione dei punti colorati per revenue - IT

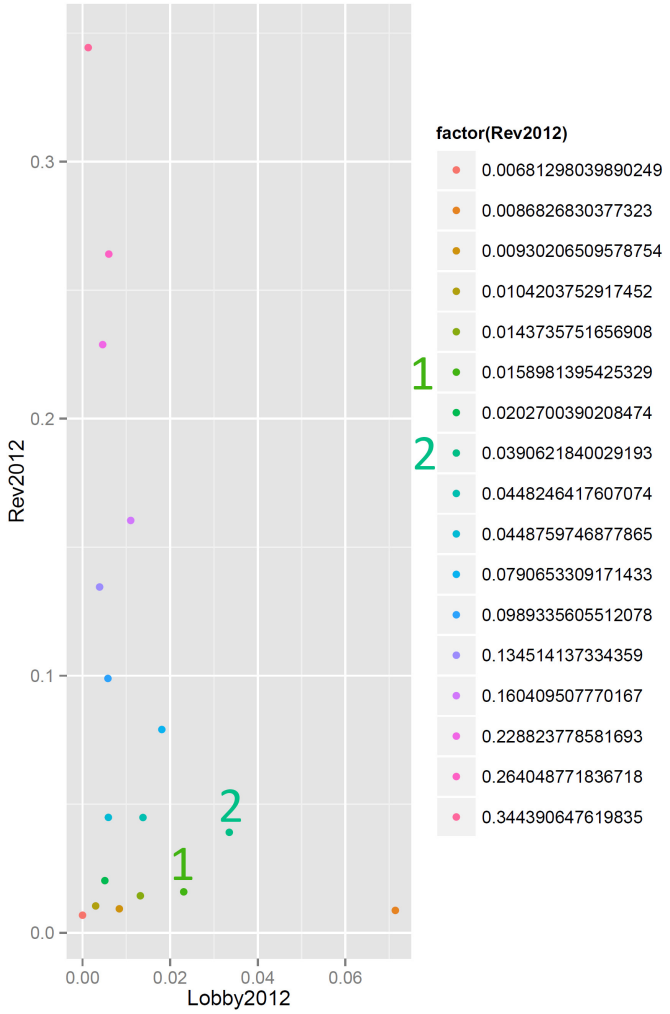
Si effettua in questa sezione una ricerca della dinamica delle imprese che hanno investito maggiormente in attività di lobbying nel 2007 per osservare la loro posizione nel 2012.



Nel grafico l'impresa 1 è la Motorola Solutions, Inc. mentre l'impresa 2 è la Qualcomm Inc. I dettagli per l'anno 2007 sono i seguenti:

	Revenue2007	Lobby2007	Roe2007	nBtw2007	Clo2007	nDgr2007
1 - Motorola Solutions Inc.	7965000000	0.0825	-0.32	0.5580	0.7745	5.7952
2 - Qualcomm Inc.	8871000000	0.0771	20.86	0.0720	0.6262	2.0207

Nel 2012 invece tali imprese risultano in queste posizioni:



I dettagli per il 2012 di queste imprese sono

	Revenue2007	Lobby2007	Roe2007	nBtw2007	Clo2007	nDgr2007
1 - Motorola Solutions Inc.	8698000000	0.0231	26.98	0.1515	0.7167	1.4848
2 - Qualcomm Inc.	19121000000	0.0335	18.22	0.7724	0.7451	2.4192

Come si può osservare il beneficio in termini di revenue è minimo per la ditta 1, mentre per l'impresa 2 raddoppia. Si ha un sostanziale decremento dell'investimento in attività di lobbying e una generale aggregazione al gruppo centrale; questo presuppone forse un qualche genere di accordo fra le imprese; questo risultato tuttavia dovrà essere approfondito da ulteriori verifiche. In ogni caso le differenti analisi evidenziano la complessità del problema.

6.2 Settore Health Care

La metodologia di analisi sviluppata è la medesima:

- Insieme delle aziende che fanno parte del settore *Health Care* utilizzando la funzione di *subset* che seleziona le imprese che hanno un codice GICSIndustry compreso fra 351010 e 352030 dal data frame di origine *aziende* nell'arco temporale fra il 2007 al 2012:

```
Hc2007 <- subset(aziende, (aziende$GICSIndustry > 351010 & aziende$GICSIndustry < 352030))
Hc2008 <- subset(aziende, (aziende$GICSIndustry > 351010 & aziende$GICSIndustry < 352030))
Hc2009 <- subset(aziende, (aziende$GICSIndustry > 351010 & aziende$GICSIndustry < 352030))
Hc2010 <- subset(aziende, (aziende$GICSIndustry > 351010 & aziende$GICSIndustry < 352030))
Hc2011 <- subset(aziende, (aziende$GICSIndustry > 351010 & aziende$GICSIndustry < 352030))
Hc2012 <- subset(aziende, (aziende$GICSIndustry > 351010 & aziende$GICSIndustry < 352030))
```

- Si effettuano le medie di settore per i parametri 'lobby' e 'revenue' per ogni anno utilizzando la funzione *statisticaFile* che prende in ingresso il dataframe (in questo caso saranno i 'subset' annuali) oggetto dell'analisi e le colonne iniziale e finale. Nella colonna 1 si avrà il *revenue* e nella colonna 2 il *lobby*%:

```
statisticaFile <- function(subset, cIni, cFin)
{
  media <- apply(subset[,cIni:cFin], 2, mean, na.rm=TRUE)
}


mediaRevLobby2007 <- statisticaFile(Hc2007,1,2)
mediaRevLobby2008 <- statisticaFile(Hc2008,1,2)
mediaRevLobby2009 <- statisticaFile(Hc2009,1,2)
mediaRevLobby2010 <- statisticaFile(Hc2010,1,2)
mediaRevLobby2011 <- statisticaFile(Hc2011,1,2)
mediaRevLobby2012 <- statisticaFile(Hc2012,1,2)
```

- Creazione del data frame con le colonne: Anni, MediaLobby, MediaRevenue.
  In *Anni* inserisco un vettore con gli anni (2007,2008,2009,2010,2011,2012);
  In $MediaLobby$ inserisco un vettore con i valori delle medie per la 'lobby %' per ogni anno. In $MediaRevenue$ inserisco un vettore con i valori delle medie per il 'revenue' per ogni anno. Unisco le 3 colonne con il metodo *cbind* e trasformo in data frame.

```
Anni <- c(2007,2008,2009,2010,2011,2012)
MediaLobby <- c(mediaRevLobby2007[2], mediaRevLobby2008[2],
                mediaRevLobby2009[2], mediaRevLobby2010[2],
                mediaRevLobby2011[2], mediaRevLobby2012[2])

MediaRevenue <-c(mediaRevLobby2007[1], mediaRevLobby2008[1],
                 mediaRevLobby2009[1], mediaRevLobby2010[1],
                 mediaRevLobby2011[1], mediaRevLobby2012[1])

HCmatrix <- cbind(Anni, MediaLobby,MediaRevenue)

HCdataframe <- as.data.frame(HCmatrix)
```

Il data frame $HCdataframe$ è il seguente:

	Anni	MediaLobby	MediaRevenue
1	2007	0.02266667	0.05814844
2	2008	0.02137600	0.05164723
3	2009	0.02314074	0.05827295
4	2010	0.02199655	0.05938707
5	2011	0.01765556	0.05759012
6	2012	0.02023333	0.08173967

- Visualizzazione del grafico che mette in relazione la lobby% su base annua e verifico l'andamento. Tale oggetto viene sviluppato con il pacchetto *ggplot*2 che mostra un diagramma a barre colorate che ha sull'asse 'x' il tempo, cioè gli anni dal 2007 al 2012 e sulla 'y' i valori medi annuali della lobby%:

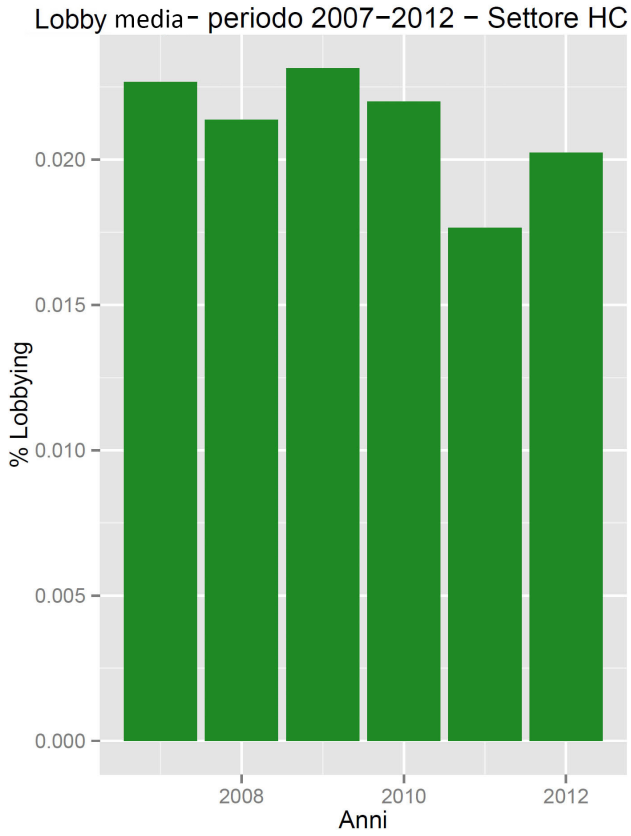


Figura 6.4: Grafico della lobby media nelle industrie del settore HC nel periodo 2007 - 2012

Tale grafico è stato realizzato con il seguente script:

```
y <- "% Lobbying"
x <- "Anni"
d <- ggplot(data=HCdataframe, aes(x=Anni, y=MediaLobby))
d + geom_bar(stat="identity", fill="forestgreen")
  + ggtitle("Lobby - periodo 2007-2012 - Settore HC")
  + ylab(y) + xlab(x)
```

- Visualizzazione del grafico che mette in relazione il revenue standardizzato medio di settore su base annua e verifico l'andamento utilizzando la stessa tecnica sopra descritta:

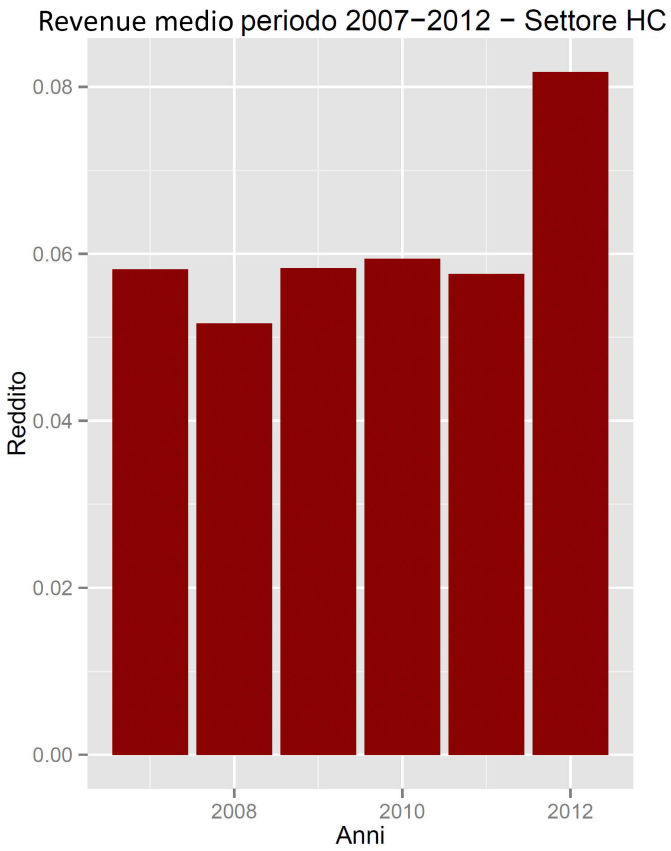


Figura 6.5: Grafico della revenue media nelle industrie del settore HC nel periodo 2007 - 2012

- Visualizzazione del grafico globale con la lobby%, il reddito e l'arco temporale; quest'ultimo è stato realizzato aggiungendo un carattere estetico all'oggetto ggplot che abbiamo analizzato permettendo una visualizzazione chiara dell'andamento. Più si va avanti negli anni e maggiore sembra essere il reddito medio standardizzato del settore e minore l'investimento medio in attività di lobbying:

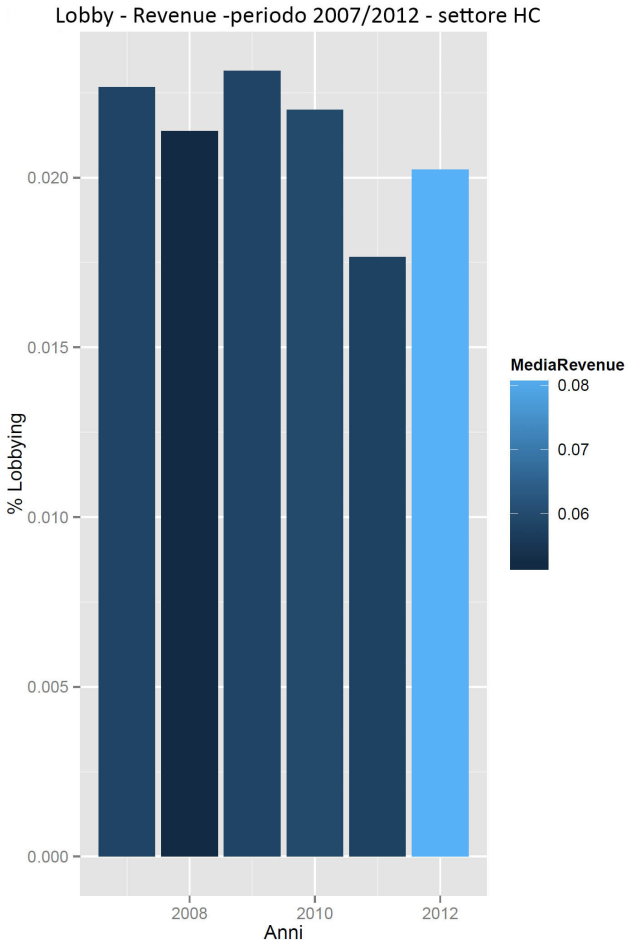


Figura 6.6: Grafico di relazione fra la lobby media e il revenue medio delle industrie nel periodo 2007 - 2012 sul settore HC

L'andamento dei grafici sembra riprodurre, anche se in modo attenuato, quanto visto per le industrie IT. A fronte di un investimento in lobbying più o meno costante dal 2007 al 2010, si evidenzia un picco nel 2012 per il revenue. Il ritardo fra l'investimento e la crescita del revenue sembra più breve ma richiede maggiori investimenti.

6.2.1 Ulteriori verifiche

Correlazione fra le variabili

A dimostrazione di questo comportamento effettuo l'analisi di correlazione fra le variabili revenue e lobby % di settore per ogni anno al fine di capire l'inversione di tendenza. Ci si aspetta che i primi anni la correlazione sia debolmente negativa e che nel periodo 2011/2012 risulti essere maggiormente negativa.
La metodologia adottata è la stessa usata per le industrie IT:

- Calcolo della correlazione: con la funzione *subset* seleziono il settore $Health\ Care$ dal data frame di partenza 'aziende'. In colonna '1' vi sono i valori standardizzati del reddito di ogni impresa, in colonna '2' i valori corrispondenti dell'investimento % in lobbying. Quindi se prendiamo in esame l'anno 2007 avremo che $Hc2007$ è il *subset* che applicato alla funzione *cor* nelle colonne 1 e 2 restituisce la correlazione fra le variabili:

```
cor(Hc2007[, 1:2])
```

```
##                Rev2007  Lobby2007
## Rev2007     1.0000000 -0.3064858
## Lobby2007 -0.3064858  1.0000000
```

  Tali valori sono stati organizzati in colonne 'x' , 'y' e value perchè verranno inseriti in un grafico colorato per permettere una chiarezza espositiva maggiore; a tal scopo utilizzo la funzione melt:

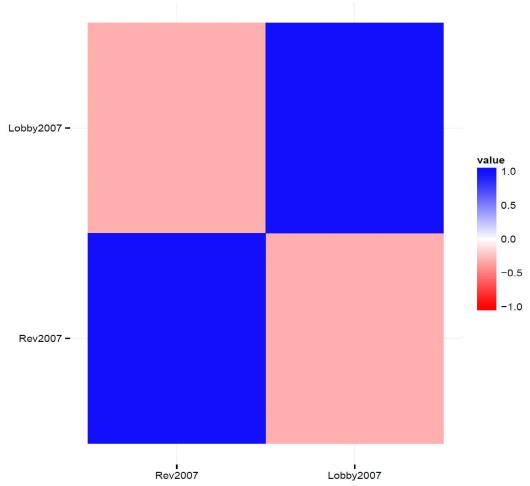
- Grafico colorato per le correlazioni: si associano 2 colori, se la correlazione è tendente al rosso è negativa, al blu è positiva. Da notare ovviamente che le correlazioni fra se stessi sono pari a 1. Per effettuare questo grafico utilizzo un oggetto *ggplot* passando nelle caratteristiche estetiche i valori delle correlazioni con $aes(fill = value)$ visualizzati tramite un gradiente di colori rosso e blu con la funzione $scale_fill_gradient2$ e i valori limite delle correlazioni: 1 e -1:

```
ggplot(mescDati, aes(x=x, y=y)) + geom_tile(aes(fill=value)) +
  scale_fill_gradient2(low="red", mid="white", high="blue",
                       guide=guide_colorbar(ticks=FALSE, barheight=10),
                       limits=c(-1,1)) + theme_minimal() + labs(x=NULL, y=NULL)
```

I risultati delle correlazioni numeriche e i grafici sono stati confrontati nel 2007 e nel 2012:

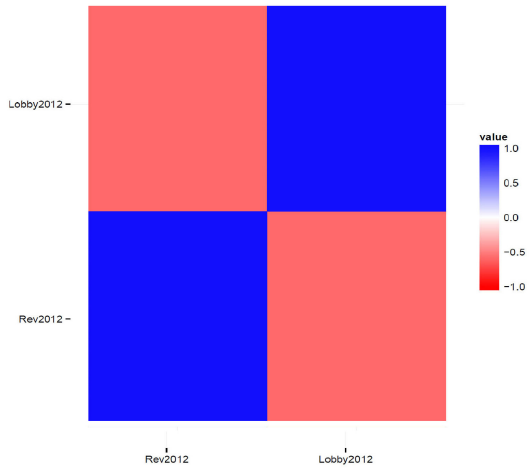
2007

```
cor(Hc2007[, 1:2])

##              Rev2007  Lobby2007
## Rev2007    1.0000000 -0.3064858
## Lobby2007 -0.3064858  1.0000000
```



2012

```
cor(Hc2012[, 1:2])

##              Rev2012  Lobby2012
## Rev2012    1.0000000 -0.5852458
## Lobby2012 -0.5852458  1.0000000
```



Per queste industrie la crescita della correlazione è più marcata passando da un valore di -0.30 nel 2007 ad un valore di -0.58 nel 2012.

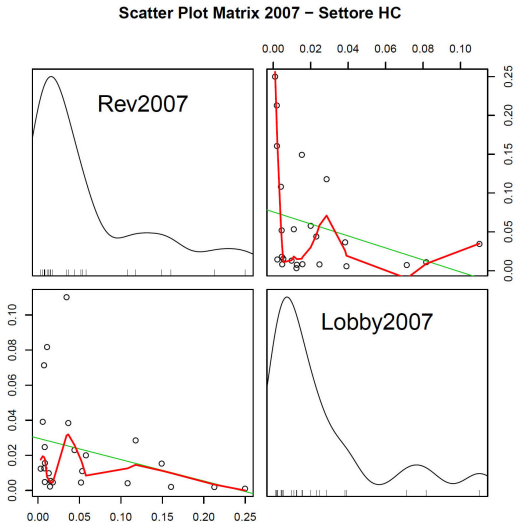
Regressione

Si effettua la verifica delle regressioni lineari e multiple per le variabili in esame nel periodo 2007 - 2012: è utile vedere la regressione lineare sulle variabili e come

variano negli anni il p-value ($Pr(>|t|)$) dei coefficienti; Viene visualizzato inoltre lo *scatterplotMatrix*, grafico dove si evidenziano densità, distribuzione dei punti, una linea verde che è la retta di regressione lineare, una linea rossa che rappresenta la regressione non parametrica (qui si utilizza quella di default cioè la GAM) e l'intervallo fra le rette tratteggiate rosse che rappresenta lo *spread* dei dati.

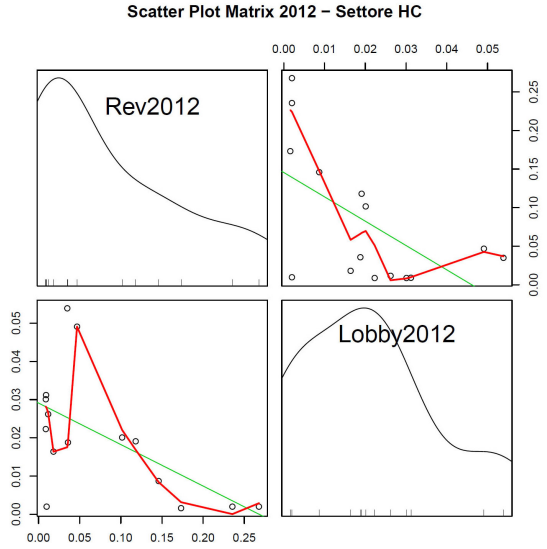
2007

```
#regressione lineare
model <- lm(Hc2007$Rev2007 ~ Hc2007$Lobby2007)
summary(model)

##
## Call:
## lm(formula = Hc2007$Rev2007 ~ Hc2007$Lobby2007)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.06377 -0.05520 -0.01388  0.03764  0.17486
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.07567    0.01821   4.156 0.000413 ***
## Hc2007$Lobby2007 -0.77316    0.51195  -1.510 0.145216
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06875 on 22 degrees of freedom
## Multiple R-squared:  0.09393,Adjusted R-squared:  0.05275
## F-statistic: 2.281 on 1 and 22 DF,  p-value: 0.1452
```



2012

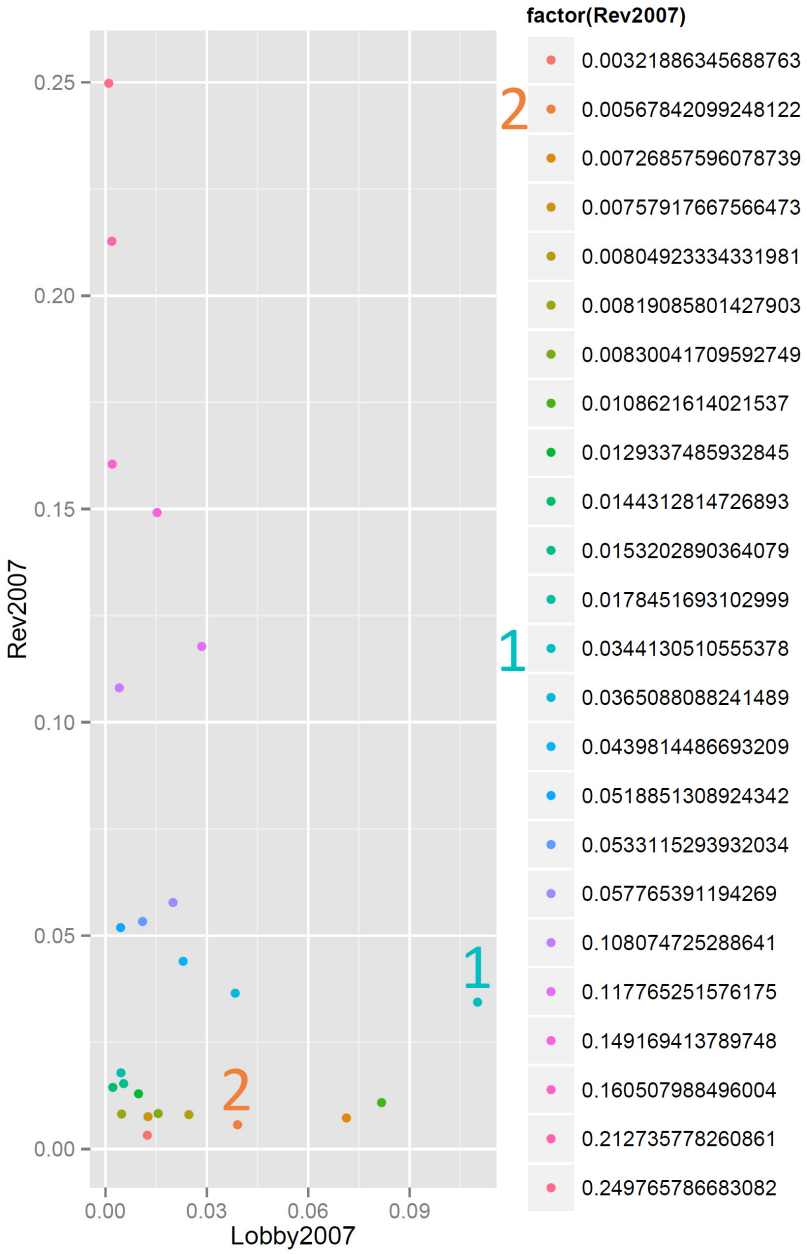
```
#regressione lineare
model <- lm(Hc2012$Rev2012 ~ Hc2012$Lobby2012)
summary(model)

##
## Call:
## lm(formula = Hc2012$Rev2012 ~ Hc2012$Lobby2012)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12947 -0.05089  0.01945  0.04442  0.12859
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.14557    0.03108   4.684 0.000428 ***
## Hc2012$Lobby2012 -3.15472    1.21226  -2.602 0.021908 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07394 on 13 degrees of freedom
## Multiple R-squared:  0.3425,Adjusted R-squared:  0.2919
## F-statistic: 6.772 on 1 and 13 DF,  p-value: 0.02191
```



Anche nel settore Health Care si può notare che la distribuzione della revenue nel 2007 risulta essere multimodale suggerendo probabilmente diversi raggruppamenti di investimento e reddito. Nel 2012 si osserva invece una maggiore uniformintà avendo un solo picco; come nell'analisi delle industrie IT, sembra suggerire una razionalizzazione degli investimenti.
Si osserva molto bene anche nella lobby. Nel 2007 si è in presenza di una trimodale, nel 2010 si passa ad una bimodale per arrivare all'anno 2012 dove si ha un unico picco e una distribuzione quasi normale; si potrebbe ipotizzare che dopo anni di strategie differenti di investimento, nel 2012 si trovi un accordo fra le imprese, favorendo quelle che hanno un alto revenue a sfavore delle imprese con meno reddito.

Scatterplot della distribuzione dei punti colorati per revenue - HC

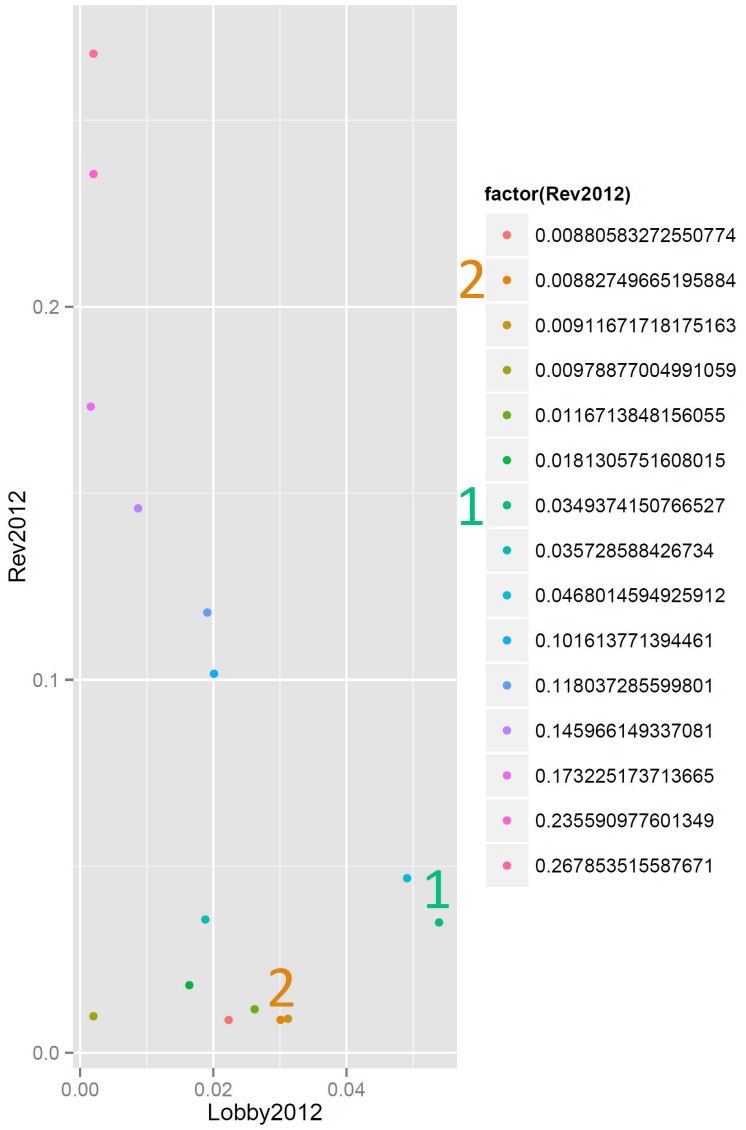
Si effettua in questa sezione una ricerca della dinamica delle imprese che hanno investito maggiormente in attività di lobbying nel 2007 per osservare la loro posizione nel 2012.



L'impresa numero 1 è la Amgen Incorporeted e la numero 2 è la Biogen Idec Inc. I dettagli per l'anno 2007 sono i seguenti:

	Revenue2007	Lobby2007	Roe2007	nBtw2007	Clo2007	nDgr2007
1 - A.I.	14771000000	0.1101	17.23	0.2126	0.6654	3.4695
2 - B.I.I.	3171617000	0.0391	11.51	0.1007	0.6115	2.6307

Nell'anno 2012 tali imprese le troviamo in queste posizioni:



I dettagli per l'anno 2012 sono i seguenti:

	Revenue2007	Lobby2007	Roe2007	nBtw2007	Clo2007	nDgr2007
1 - A.I.	17265000000	0.0539	22.80	0.2881	0.7533	1.4914
2 - B.I.I.	5516461000	0.0301	19.82	0.1931	0.6990	0.7136

6.3 Settore Industrials

Il settore industriale risulta essere importante per la quantità di aziende che forniscono i dati dal 2007 al 2012.
La metodologia di analisi sviluppata è la seguente:

- Selezione dell'insieme delle aziende che fanno parte del settore *Industrial* utilizzando la funzione di *subset* che seleziona le imprese che hanno un codice GICSIndustry compreso fra 201010 e 203050 dal data frame di origine *aziende* nell'arco temporale fra il 2007 al 2012:

```
In2007 <- subset(aziende, (aziende$GICSIndustry > 201010 & aziende$GICSIndustry < 203050))
In2008 <- subset(aziende, (aziende$GICSIndustry > 201010 & aziende$GICSIndustry < 203050))
In2009 <- subset(aziende, (aziende$GICSIndustry > 201010 & aziende$GICSIndustry < 203050))
In2010 <- subset(aziende, (aziende$GICSIndustry > 201010 & aziende$GICSIndustry < 203050))
In2011 <- subset(aziende, (aziende$GICSIndustry > 201010 & aziende$GICSIndustry < 203050))
In2012 <- subset(aziende, (aziende$GICSIndustry > 201010 & aziende$GICSIndustry < 203050))
```

- Effettuo le medie di settore per i parametri 'lobby' e 'reddito' per ogni anno utilizzando la funzione *statisticaFile* che prende in ingresso il dataframe (in questo caso saranno i 'subset' annuali) oggetto dell'analisi e le colonne iniziale e finale. Nella colonna 1 si avrà il *reddito* e nella colonna 2 il *lobby%*:

```
#calcola la media - colonna 1 il rev, 2 la lobby
statisticaFile <- function(subset, cIni, cFin)
{
  media <- apply(subset[,cIni:cFin], 2, mean, na.rm=TRUE)
}


mediaRevLobby2007 <- statisticaFile(In2007,1,2)
mediaRevLobby2008 <- statisticaFile(In2008,1,2)
mediaRevLobby2009 <- statisticaFile(In2009,1,2)
mediaRevLobby2010 <- statisticaFile(In2010,1,2)
mediaRevLobby2011 <- statisticaFile(In2011,1,2)
mediaRevLobby2012 <- statisticaFile(In2012,1,2)
```

- Creazione del data frame con le colonne: Anni, MediaLobby, MediaRevenue.
  In *Anni* inserisco un vettore con gli anni (2007,2008,2009,2010,2011,2012);
  In *MediaLobby* inserisco un vettore con i valori delle medie per la 'lobby %' per ogni anno cioè tutti i valori nella colonna 2 delle *medieRevLobby*. In *MediaRevenue* inserisco un vettore con i valori delle medie per il 'revenue' per ogni anno cioè tutti i valori nella colonna 1 delle *medieRevLobby*. Unisco le 3 colonne con il metodo *cbind* e trasformo in data frame.

```
Anni <- c(2007,2008,2009,2010,2011,2012)
MediaLobby <- c(mediaRevLobby2007[2], mediaRevLobby2008[2],
                mediaRevLobby2009[2], mediaRevLobby2010[2],
                mediaRevLobby2011[2], mediaRevLobby2012[2])
MediaRevenue <-c(mediaRevLobby2007[1], mediaRevLobby2008[1],
                 mediaRevLobby2009[1], mediaRevLobby2010[1],
                 mediaRevLobby2011[1], mediaRevLobby2012[1])

FImatrix <- cbind(Anni, MediaLobby,MediaRevenue)
FIdataframe <- as.data.frame(FImatrix)
```

Il risultato finale è $FIDataframe$ composto dalle seguenti colonne:

	Anni	MediaLobby	MediaRevenue
1	2007	0.012335294	0.04616453
2	2008	0.013096970	0.04212971
3	2009	0.017122222	0.03777400
4	2010	0.014073171	0.03740823
5	2011	0.011305405	0.03750781
6	2012	0.009764706	0.05707219

- Si studia ora il grafico che mette in relazione la lobby% media di settore dal 2007 al 2012 e verifico l'andamento. Tale oggetto viene sviluppato con il pacchetto *ggplot*2 che mostra un diagramma a barre colorate che ha sull'asse 'x' il tempo, cioè gli anni dal 2007 al 2012 e sulla 'y' i valori medi annuali della lobby%:

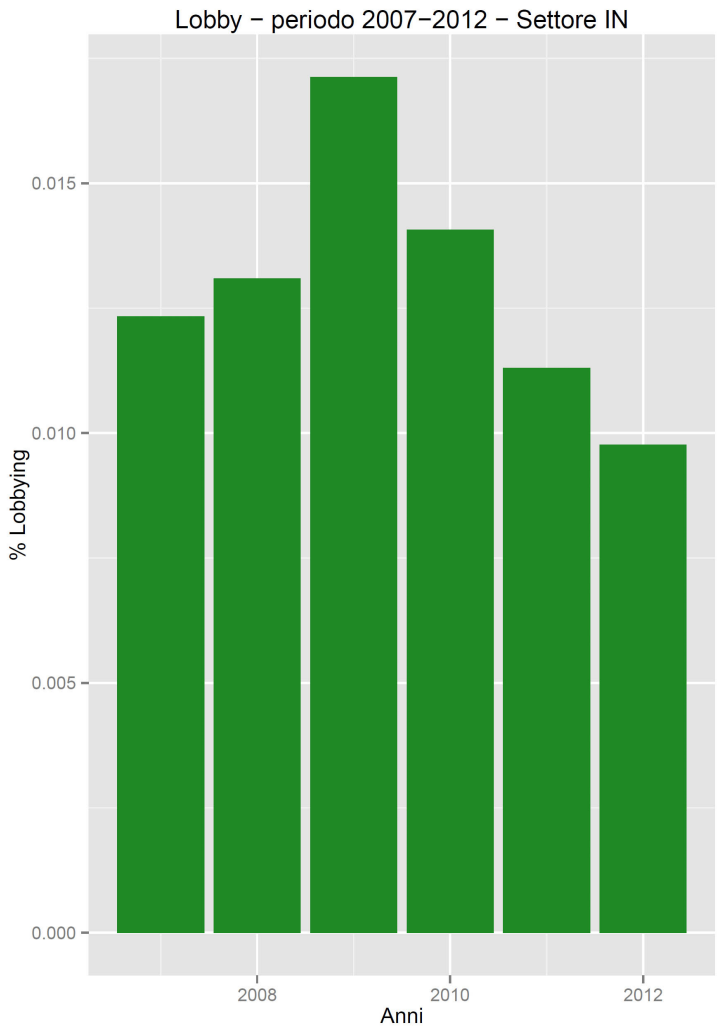


Figura 6.7: Grafico della lobby media nelle industrie del settore IN nel periodo 2007 - 2012

- Visualizzo il grafico che mette in relazione il revenue standardizzato medio di settore su base annua e verifico l'andamento utilizzando la stessa tecnica sopra descritta:

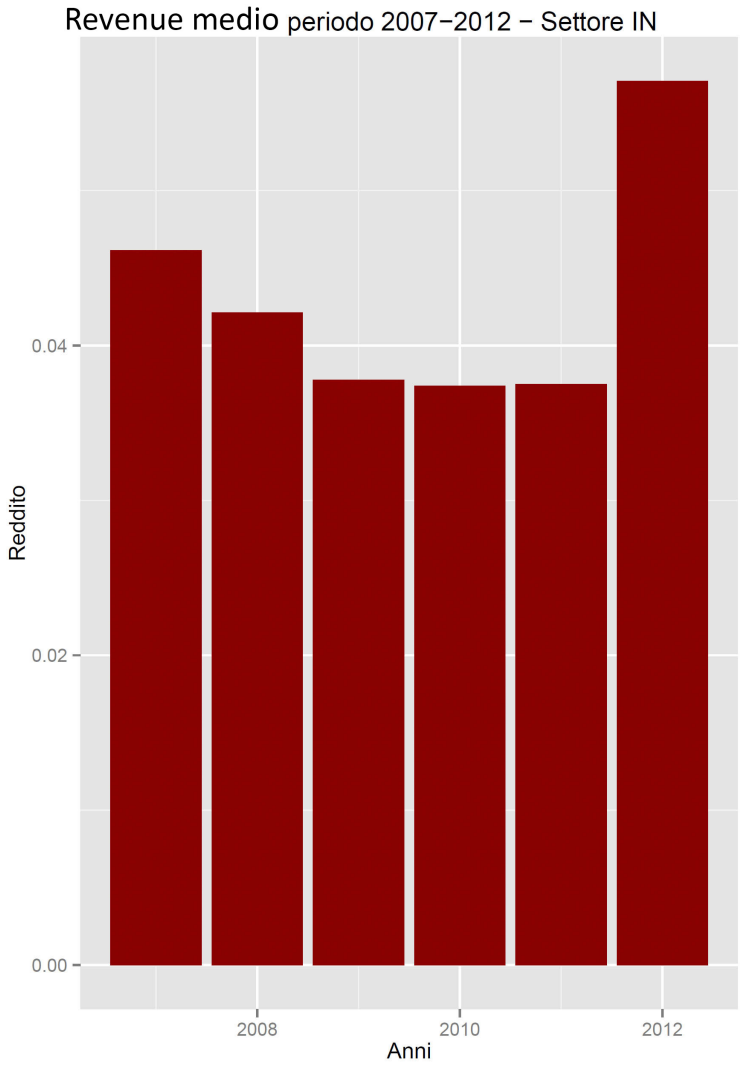


Figura 6.8: Grafico della revenue media nelle industrie del settore IN nel periodo 2007 - 2012

- Visualizzazione del grafico globale con la lobby%, il reddito e l'arco temporale; quest'ultimo diagramma è stato realizzato aggiungendo un carattere estetico all'oggetto ggplot che abbiamo analizzato permettendo una visualizzazione chiara dell'andamento. Come si può notare negli anni si ha un aumento medio del reddito a fronte di una diminuzione dell'investimento in attività di lobbying.

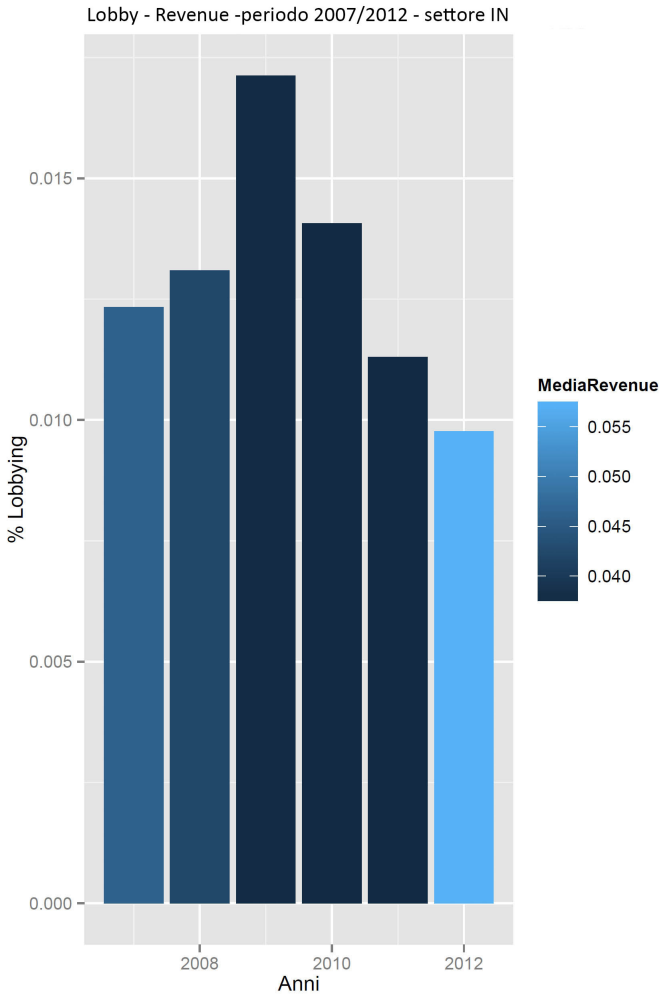


Figura 6.9: Grafico di relazione fra la lobby media e il revenue medio delle industrie nel periodo 2007 - 2012 sul settore IN

L'andamento dei grafici sembra riprodurre, in maniera decisa, quanto visto per le industrie IT e HC. A fronte di un picco di investimento in lobbying nel 2009, si evidenzia un picco nel 2012 per il revenue. Il ritardo fra l'investimento e la crescita del revenue sembra essere circa di 3 anni.

6.3.1 Ulteriori verifiche

Correlazione fra le variabili

A dimostrazione di questo comportamento effettuo l'analisi di correlazione fra le variabili revenue e lobby % di settore per ogni anno al fine di capire l'inversione di tendenza. Ci si aspetta che i primi anni la correlazione sia debolmente negativa e che nel periodo 2011/2012 risulti essere maggiormente negativa.
La metodologia adottata è la seguente:

- Si calcola la correlazione: con la funzione *subset* seleziono il settore *Industrials* dal data frame di partenza 'aziende'. In colonna '1' vi sono i valori standardizzati del reddito di ogni impresa, in colonna '2' i valori corrispondenti dell'investimento % in lobbying. Quindi se prendiamo in esame l'anno 2007 avremo che $In2007$ è il *subset* che applicato alla funzione *cor* nelle colonne 1 e 2 restituisce la correlazione fra le variabili:

```
In2007 <- subset(aziende, (aziende$GICSIndustry > 201010 &
                 aziende$GICSIndustry < 203050))

cor(In2007[, 1:2])

##               Rev2007   Lobby2007
## Rev2007    1.00000000 -0.02335588
## Lobby2007 -0.02335588  1.00000000
```

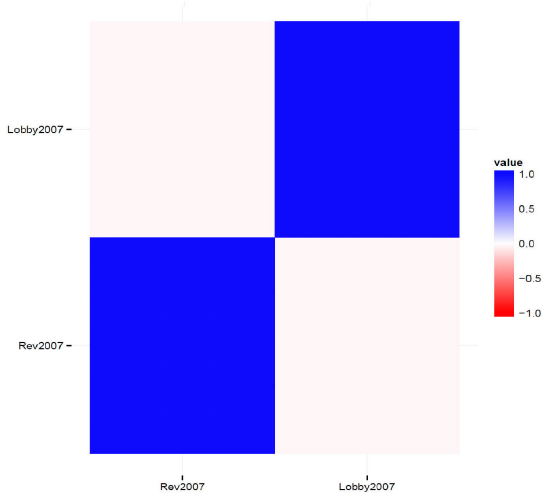
  Tali valori verranno calcolati per ogni anni e organizzati in colonne 'x' , 'y' e 'valore di correlazione' perchè verranno inseriti in un grafico colorato per permettere una chiarezza espositiva maggiore; a tal scopo utilizzo la funzione melt:

- Si visualizza il grafico colorato per le correlazioni: si associano 2 colori, se la correlazione è tendente al rosso è negativa, al blu è positiva. Da notare ovviamente che le correlazioni fra se stessi sono pari a 1 (quindi blu intenso). Per effettuare questo grafico utilizzo un oggetto *ggplot* passando nelle caratteristiche estetiche i valori delle correlazioni con $aes(fill = value)$ visualizzati tramite un gradiente di colori rosso e blu con la funzione $scale\_fill\_gradient2$ e i valori limite delle correlazioni: 1 e -1:

2007

```
cor(In2007[, 1:2])
```

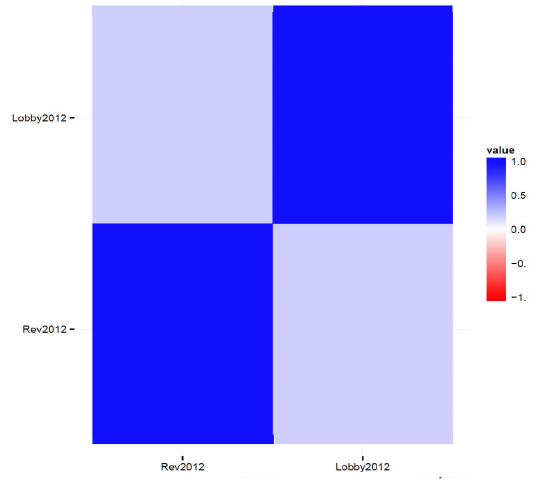
```
##                Rev2007   Lobby2007
## Rev2007    1.00000000 -0.02335588
## Lobby2007 -0.02335588  1.00000000
```



2012

```
cor(In2012[, 1:2])
```

```
##                Rev2012 Lobby2012
## Rev2012    1.0000000 0.1862831
## Lobby2012 0.1862831 1.0000000
```



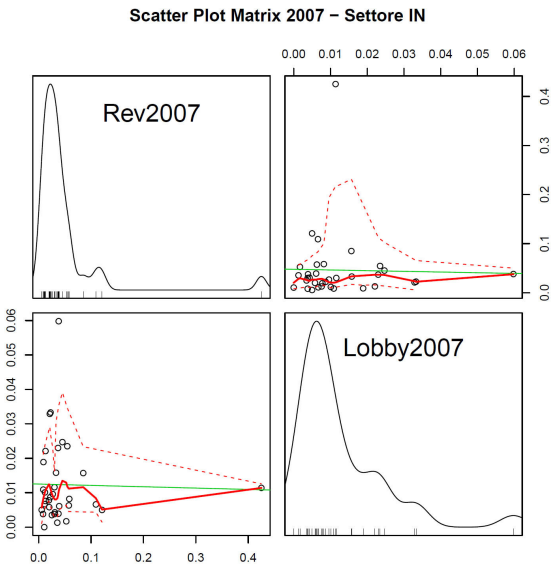
Per queste industrie la correlazione tra lobbying e revenue è pressochè nulla sia nell'anno 2007 che 2012.

Regressione

Si effettua inoltre la verifica delle regressioni lineari per le variabili in esame nell'anno 2007 e 2012: è utile vedere la regressione lineare sulle variabili e come variano negli anni il p-value ($Pr(>|t|)$) dei coefficienti; Viene visualizzato inoltre lo *scatterplotMatrix*, grafico dove si evidenziano densità, distribuzione dei punti, una linea verde che è la retta di regressione lineare, una linea rossa che rappresenta la regressione non parametrica (qui si utilizza quella di default cioè la GAM) e l'intervallo fra le rette tratteggiate rosse che rappresenta lo *spread* dei dati.

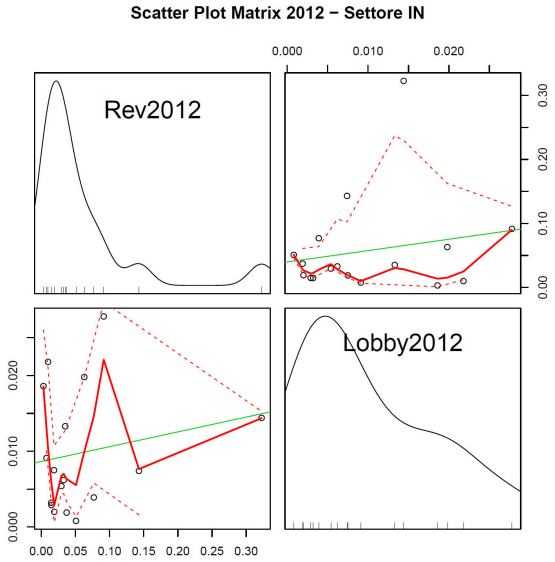
2007

```
#regressione lineare
model <- lm(In2007$Rev2007 ~ In2007$Lobby2007)
summary(model)

##
## Call:
## lm(formula = In2007$Rev2007 ~ In2007$Lobby2007)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.04150 -0.03066 -0.01686  0.00029  0.37882
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.04789    0.01813   2.642   0.0126 *
## In2007$Lobby2007 -0.13991    1.05865  -0.132   0.8957
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0733 on 32 degrees of freedom
## Multiple R-squared:  0.0005455,Adjusted R-squared:  -0.03069
## F-statistic: 0.01747 on 1 and 32 DF,  p-value: 0.8957
```



2012

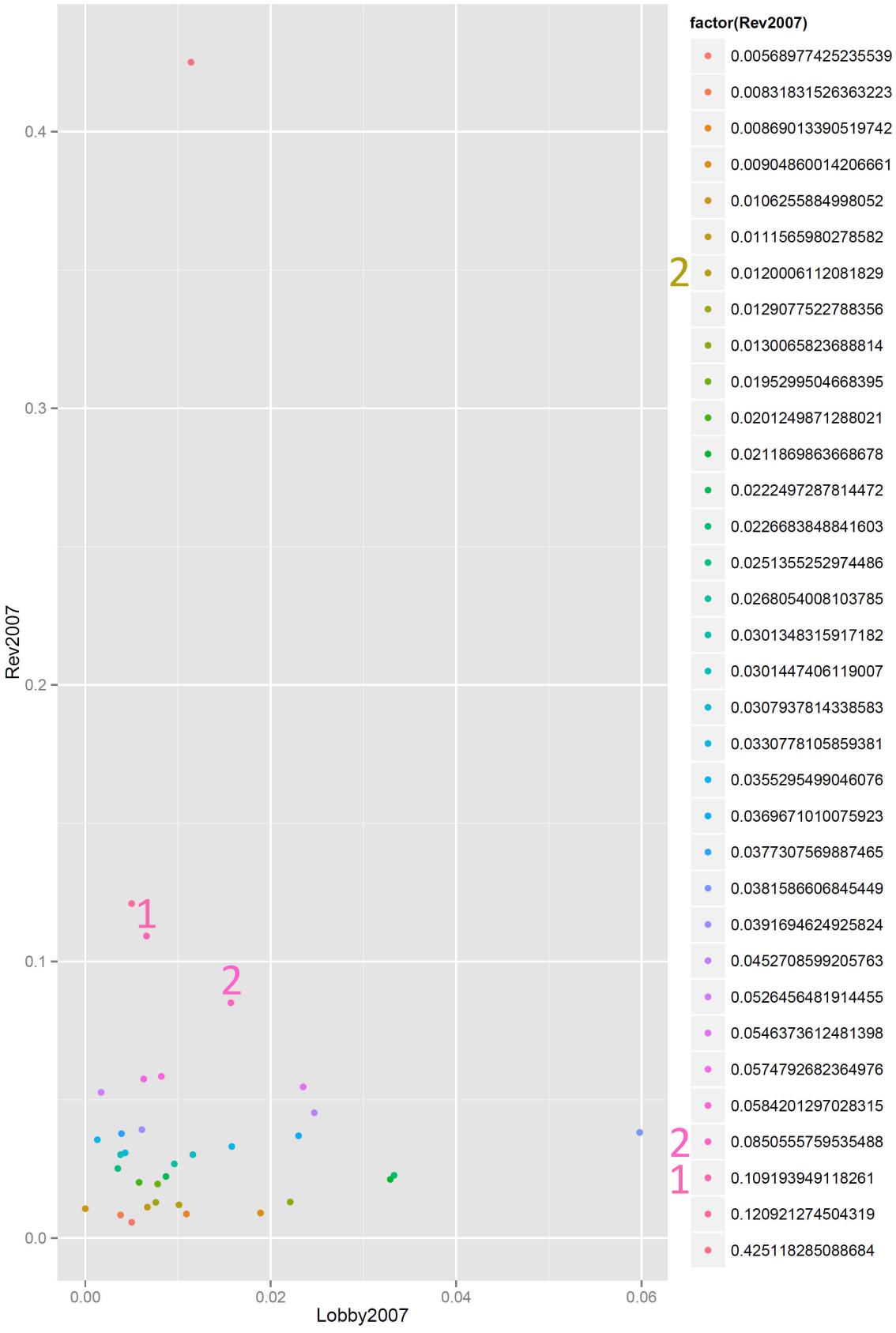
```
#regressione lineare
model <- lm(In2012$Rev2012 ~ In2012$Lobby2012)
summary(model)

##
## Call:
## lm(formula = In2012$Rev2012 ~ In2012$Lobby2012)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.069650 -0.030746 -0.019877  0.002299  0.257249
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.03972    0.03035   1.309    0.210
## In2012$Lobby2012   1.77700    2.41991   0.734    0.474
##
## Residual standard error: 0.07852 on 15 degrees of freedom
## Multiple R-squared:  0.0347,Adjusted R-squared:  -0.02965
## F-statistic: 0.5392 on 1 and 15 DF,  p-value: 0.4741
```



Nel settore Industrials si può notare che, la distribuzione della revenue rimane più o meno invariata. Si ha un leggero aumento mantenendo sempre una distribuzione trimodale. Si osserva invece che, per la lobby, si ha un passaggio da una multimodale a una distribuzione monomodale.

Scatterplot della distribuzione dei punti colorati per revenue - IN

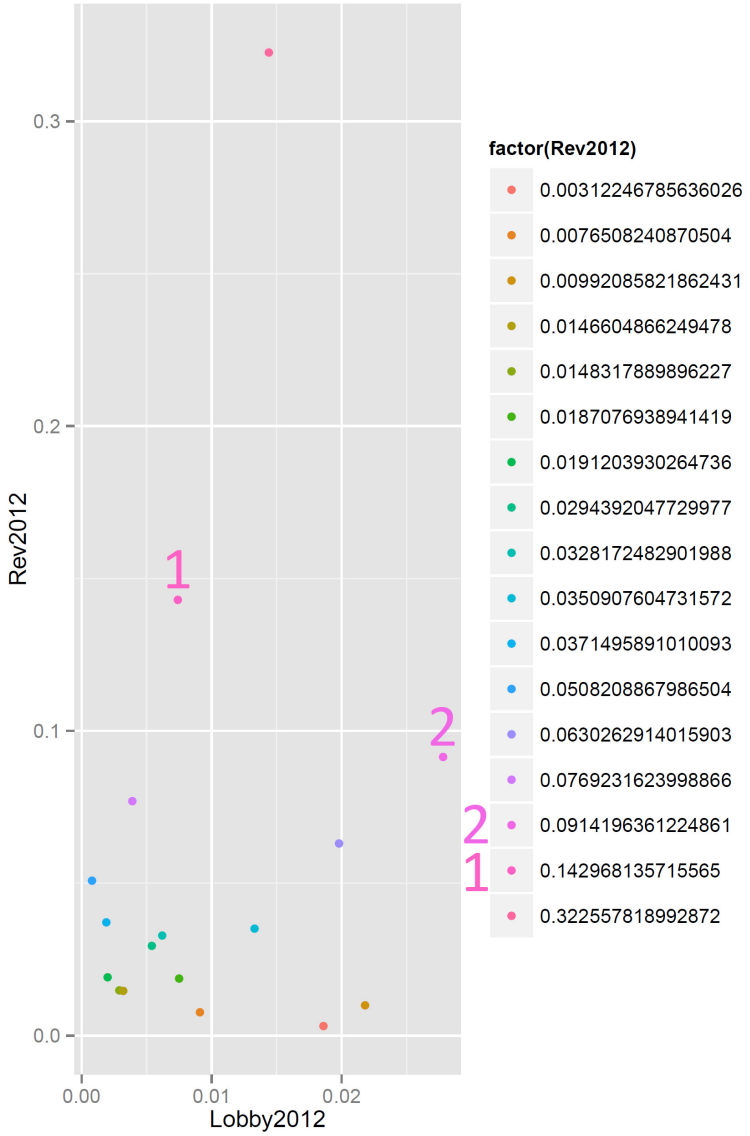
Si effettua in questa sezione una ricerca della dinamica delle imprese che hanno investito maggiormente in attività di lobbying nel 2007 per osservare la loro posizione futura nel 2012 nel settore Industrials.
Ci si aspetta che dato il loro alto investimento, esse possano assumere una posizione più alta nel grafico, corrispondente a maggiori valori di revenue e un decremento della spesa in attività di lobbying.


factor(Rev2007)
0.00568977425235539
0.00831831526363223
0.00869013390519742
0.00904860014206661
0.0106255884998052
0.0111565980278582
0.0120006112081829
0.0129077522788356
0.0130065823688814
0.0195299504668395
0.0201249871288021
0.0211869863668678
0.0222497287814472
0.0226683848841603
0.0251355252974486
0.0268054008103785
0.0301348315917182
0.0301447406119007
0.0307937814338583
0.0330778105859381
0.0355295499046076
0.0369671010075923
0.0377307569887465
0.0381586606845449
0.0391694624925824
0.0452708599205763
0.0526456481914455
0.0546373612481398
0.0574792682364976
0.0584201297028315
0.0850555759535488
0.109193949118261
0.120921274504319
0.425118285088684
Rev2007
0.0
0.1
0.2
0.3
0.4
Lobby2007
0.00
0.02
0.04
0.06
1
2
2
1
2

L'azienda con il numero 1 è la Caterpillar Inc, mentre la numero 2 è la Fedex Corp; i dettagli per il 2007 sono i seguenti:

	Revenue2007	Lobby2007	Roe2007	nBtw2007	Clo2007	nDgr2007
1 - Caterpillar Inc	44958000000	0.0066	39.86	0.7750	0.8153	6.3725
2 - Fedex Corp	35214000000	0.0157	15.93	0.1782	0.6814	4.5588

Nel 2012 tali variabili sono posizionate nel seguente modo:



Come si può notare nel 2012 vi è una sostanziale riduzione del numero delle imprese. Per quanto riguarda la Caterpillar Corp. (corrispondente al numero 1) nel 2007 aveva

un revenue standardizzato appena sopra lo 0.1. Nel 2012 ha migliorato la posizione avvicinandosi allo 0.15. L'azienda 2, la Fedex Corp., passa da un investimento in lobbying medio basso nel 2007, intorno allo 0.017, ad uno più alto, quasi a 0.025. Risulta essere nel 2012 l'impresa che investe di più in attività di lobbying presupponendo forse un mancato accordo fra essa e le altre industrie sviluppando quindi un comportamento solitario. Tale ipotesi andrebbe confermata da ulteriori analisi più specifiche.

Conclusioni

Nel corso della Tesi si sono messe a punto procedure che possono automatizzare l'analisi dei dati nel settore dell'attività di lobbying. I dati considerati sono quelli di industrie statunitensi classificate secondo i codici GICS (Global Industry Classification Standard).

I pacchetti a disposizione, basati sul linguaggio di programmazione R, si sono dimostrati estremamente utili per l'analisi di questi dati. Questo è stato reso possibile non solo dai pacchetti di statistica ma anche dai pacchetti di grafica, basati su una grammatica sviluppata da Wickham.

L'analisi ha mostrato alcuni andamenti interessanti come, per esempio, quello che illustra nei dati, che coprono un periodo di tempo che va dal 2007 al 2012, un ritardo temporale fra l'investimento in attività di lobbying di settori industriali distinti e il valore del revenue.

Questo valore oscilla fra circa due anni e mezzo e quattro anni. Se questo andamento sarà confermato dalle ulteriori e necessarie analisi che si intende svolgere in futuro, si potrebbero avere strumenti non solo per definire le relazioni fra investimento e ritorno dell'investimento in questo settore ma anche per rivelare comportamenti anomali.

Bibliografia

[1] President William J. Clinton on December 19 1995 - 104th United States Congress

[2] Lorenzo Cuocolo, Gianluca Sgueo - 2014 - Rules Research Unit Law and Economics Studies Paper No. 2014/13 - pagine 8 - 17 - Università commerciale Luigi Bocconi

[3] United States Government Accountability Office - Aprile 2009 - 2008 Lobbying Disclosure - Observations on Lobbyists' Compliance with Disclosure Requirements - GAO 09-487

[4] President Barack Obama - March 20, 2009 - Memorandum For The Heads Of Executive Departments And Agencies- Ensuring Responsible Spending of Recovery Act Funds - sezione 3

[5] Evangeline Marzec (2012-01-14) - Demand Media - What Is Corporate Lobbying? Chron.com

[6] Donald E. deKieffer (2007). The Citizen's Guide to Lobbying Congress: Revised and Updated . Chicago Review Press . ISBN 978-1-55652-718-0

[7] Robert G. Kaiser, Alice Crites (2007) - How lobbying became Washington's biggest business – Big money creates a new capital city. As lobbying booms, Washington and politics are transformed. . Citizen K Street (The Washington Post)

[8] Standard & Poor's, MSCI (2013) - Gics Mapbook electronic 0711 - pagina 3

[9] R Development Core Team (2013)- An Introduction to R - 1.1 The R enviroment - CRAN http://cran.r-project.org/

[10] Hill, Brandon (2012) - Evaluating the design of the R language: objects and functions for data analysis - ECOOP'12 Proceedings of the 26th European conference on Object-Oriented Programming .

[11] Angelo M. Mineo - Una guida all'utilizzo dell'Ambiente Statistico R - Prefazione - Dipartimento di Scienze Statistiche e Matematiche S.Vianelli - Università degli studi di Palermo

[12] David Smith (2012); R Tops Data Mining Software Poll, Java Developers Journal, May 31, 2012.

[13] Karl Rexer, Heather Allen, & Paul Gearan (2011); 2011 Data Miner Survey Summary, presented at Predictive Analytics World, Oct.

[14] R Development Core Team (2013)- An Introduction to R - 1.1 The R enviroment - CRAN http://cran.r-project.org/ dal capitolo 2 al capitolo 7

[15] Vito M. R. Muggeo, Giancarlo Ferrara (2005) - Il linguaggio R: concetti introduttivi ed esempi II edizione - capitolo 2

[16] R. A. Hanneman (2001) - Introduction to Social Network Methods

[17] Hadley Wickham (2009), Elegant Graphics for Data Analysis - Capitolo 4

[18] Gelman and Hill , & Paul Gearan (2006); Data Analysis Using Regression and Multilevel/Hierarchical Models - capitolo 25

[19] Hadley Wickham (2009), Elegant Graphics for Data Analysis - Capitolo 1

[20] Pierre-Andre Cornillon,Arnaud Guyader,Francois Husson,Nicolas Jegou,Julie Josse,Maela Kloareg,Eric Matzner-Lober,Laurent Rouvière (2012)- R for Statistics - capitolo 7

Hoc unum scio, me nihil scire.

— Socrate

Ringraziamenti

Determinanti per lo svolgimento di questa tesi sono stati gli incontri periodici avuti con il Relatore e il Correlatore ai quali sono grato per i loro insegnamenti.
Un ringraziamento particolare a Daniele Incicco per aver consentito all'utilizzo della base di dati per l'analisi svolta.

Vorrei esprimere la mia sincera gratitudine ai miei compagni di corso per avermi fornito testi appunti e dati indispensabili al raggiungimento di questo traguardo. In particolare ringrazio Rita Fragapane, Cristina Pileggi, Roberto Squillace, Lorenzo Guerzoni, Davide Maccarone, Nicolò Gambarelli, Andrea Giampietro, Andrea Betti, Fabrizio Camassa e Lorenzo Ligregni per i numerosi progetti portati a termine con loro.

Ho il desiderio inoltre di ringraziare con affetto i miei genitori e mio fratello Gabriele per il sostegno e il grande aiuto che mi hanno dato in questi tre anni.

Infine ma non per ultima ringrazio di cuore la mia inseparabile musa ispiratrice e compagna di vita Giulia Baldi. A lei più di ogni altra persona è dedicata questa opera.

A. M.