Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

IS71061A: Data Science Final Project

Data-driven social science

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

Table of contents

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

1. Introduction

Data science is a new interdisciplinary field; it is a data-driven approach to gain knowledge using statistical methods applied with technology. Although many concepts within data science are very old, it is just now that – thanks to processing power and the ever growing data available in the world – widespread research is possible.

Data science can be thought as a collection of tools which can be applied to any field, business, physics, natural sciences, medicine, social science, marketing etc. In this paper social science is intended in its broader sense, it is intended as the counterpart of physical and natural science. Therefore, it is not limited to sociology, political science or criminology but it includes fields such as business, economics, finance, psychology and marketing.

This paper will describe the problem with traditional social science research, explore the ways in which it can be improved both theoretically and using some examples. Moreover, it will explore the new ways that data-driven social research is already making a much bigger impact than it used to thanks to the new technologies and the new extensive data available in fields like marketing.

Finally, this project will provide 5 first hand examples of the data-driven approach to research using methodologies that should be included in the social sciences.

2. Problems with social science

For a long time now social sciences, especially psychology, have used statistics to allow researchers to cast their inquiry into the framework of statistical inference but even though this has proven to be effective in some aspects it has not been helpful in scientific inference where you are trying to model what is going on and not just finding if one variable affect another (Taagepera,2008). In other words, only finding correlations is not a great scientific result. A great scientific result would be understanding the nature of the independent factors and the causations of dependent variables on independent ones.

As Duncan Luce said, social researchers need to have quantitative and technical skills that allow them to be accurate, scientific and able to find the mathematical structure in the data (Taagepera,2008). Even though social sciences have evolved a lot in the last 100 years they have plateaued because of the clueless, simplistic and ritualized use of statistical data analysis, with an excessive focus on linear regression and its logit and probit versions (Taagepera,2008). There is pressure from social science editors and scholars to publish papers without more advanced statistical techniques but they are content with mindless hypotheses testing; as a result, there is very little constructive substantive theories or effective scientific probability models and the social sciences are not as scientific as they could be (Taagepera,2008). This is relevant as there is a large number of articles that are not

even concerned to what kinds of data is used and which data manipulations are allowed (Tart, 2009). For example, Geoffry Lotus said that all social scientists do is hypothesis testing with t-, F- and p- values which is substantiated by the fact that there are so many false positives in social science that even when models predict well there is always the question of whether that is due to chance (because models are known to be fuzzy and imprecise) (Taagepera,2008).

For the most part social research has a descriptive approach rather than a quantitatively explanatory one. They tend to answer the question “what is?” by using data collection and statistical analysis which leads to empirical data; this is a descriptive model. Social research rarely goes as far asking “How should it be on logical grounds?” because that requires creating logically consistent and quantitative models; this is an explanatory or predictive model. In other words, social science is stuck in empiricism and does not provide explanations, only description (Taagepera,2008). This is proven by the fact that a social scientist almost never knows how an experiment will play out like it is possible in physics or chemistry.

Descriptive models are based in empiricism, their output is non-falsifiable, have very limited scope and cannot predict while explanatory models require a lot of logical consideration rather than blind data analysis and they are falsifiable upon testing; their purpose is finding broad models that can predict. Descriptive models are “postdictive”, they predict existing data; social scientists tend to use many predictors thinking that it will make the model more accurate but all they are doing is overfitting the model to the training set describing all the noise in the data and creating idiosyncratic insights that are not generalized to future data (Schwab, 2011). Hence why it is empirical but it does not work in principle and no broader theories can be extrapolated from that. An explanatory model seeks the reasons behind the interaction among variables, and therefore predictive theories can be extracted (Taagepera,2008).

Neglecting the proper application of the scientific method in the social sciences has led these to be incomplete (most research stops and gets published before reaching a useful conclusion) and unimpressive by outsiders, especially the sociopolitical decision makers; in fact, politicians do not really consider political science. In the past this was the case for the physical sciences, doctors did not pay attention to biology nor did engineers to physics. This is because science becomes useful in practice only when it is in an advanced state (Taagepera,2008).

3. Possible Solutions

This trend in social science can change and it does not need to start from zero, all the descriptive research done until now can be used to start predictive, science based models. There is a need to expand statistical techniques, for example by applying the modern machine learning and big data techniques to social science. A shift in paradigm by the researchers (Schwab, 2011) and an increase in scientific standards is required by social

science publications. In fact, most of the publications are dead on arrival, they are just printed but never used for anything; the findings of research are almost never used as bricks to build up to a theory, which is what happens in the physical sciences (Taagepera,2008).

Another way to solve the problem is to rethink the way hypotheses are formulated. In social sciences often researcher only try to disprove the null hypothesis, but this is only the first step, they should strive for directional hypothesis, then empirically based quantitative hypothesis and finally logically based quantitative hypothesis.

Significance test only tells there is an impact/correlation $\left(\frac{dx}{dy} \neq 0\right)$, while the directional hypothesis predicts that increase in x will increase y$\left(\frac{dx}{dy} > 0\right)$ or decrease y $\left(\frac{dx}{dy} < 0\right)$. Then you need the intensity $\left(\frac{dx}{dy} = b\right)$ and after that knowing the full function $(y = f(x)$)which is usually non-linear over a wide range of x. The more detailed the hypothesis or model is the more predictive it is but also the harder to satisfy and the more disprovable (Taagepera,2008).

Although every specific problems requires different creative approaches there are common mistakes that happen in social research. For example, when x and y can conceptually range from $-\infty$ to ∞ then it is ok to use linear regression, y as a linear function of x ($function: y = a + bx$). When x and y cannot take negative values, the linear regression should be carried out on their logarithms, which is fitting y to a fixed exponent function of x $\left(function: \log y = \log A + k \log x\,, where\ k\ \in R,\ A > 0\ and\ equals\ \frac{y_0}{x_0^k}\ and\ (x_0 y_0)\ is\ a\ point\ \ in\ the\ I\ quadrant\right)$. If y is always positive and $x \in R$ then the regression should be taken on the natural logarithm of y, which is fitting y to an exponential function of x $\left(function: \ln y = \ln A + Kx\,, where\ k \in R,\ A > 0\ and\ equals\ \frac{y_0}{x_0^k}\ and\ (x_0 y_0)\ is\ a\ point\ \ in\ the\ I\ or\ II\ quadrant\right)$. Lastly when $x \in R$ and y cannot take negative or positive values beyond some value $y = C$, this is the case when x is time and y is a percentage , then the linear regression should be done on the logarithm of $\frac{y}{(C-y)}$. This means fitting y to a simple logistic function of x $\left(function: \ln \frac{y}{C-y} = -\ln \frac{y_0}{C-y_0} - Kx_0 + Kx\,, where\ k \in R, y = y_0\ \ when\ x = a\ given\ value\ of\ x_0\right)$ (Taagepera,2008).

Another method to avoid mistakes is to take the geometric mean instead of the arithmetic one when all the values of a distribution are positive, especially when the smallest and largest values have several order of magnitude of difference since this will be a much more representative piece of statistics and closer to the median. Moreover, it is advisable to take the lognormal distribution (normal distribution of $\log x$) rather than normal distribution when all the x are positive because the normal curve includes negative values; if you fit the telephones per capita in the countries of the United Nations with a normal distribution the σ (standard deviation) exceeds the μ (arithmetic mean), which implies that $\frac{1}{2e} = 18\%$ (from

the normal distribution equation) of the member countries contain a negative number of telephones (this kind of nonsense has been published in social science papers). The reason behind this is that if x is only positive the mean, median and mode cannot be equal and high values will have a huge impact, so a lognormal distribution will give high values a smaller impact, same goes for the geometric mean. However, when $\frac{\sigma}{\mu} < \frac{1}{2}$ a normal distribution can be used even when all the values are positive, for example in adult women heights (Taagepera,2008).

One of the main problems of social science is the fact that there are very few sequential pathways between variables, finding these pathways is crucial because they allow description and prediction of social phenomena. Taagepera describes this approach as interlocking models and they are common in physical sciences. This models are simply a set of equations where the variables appear in other equations and therefore the variable and equations are linked and interlocked, whereas in social sciences additive models are predominant (Nicolai, 2009). On the same note, additive models should often be substituted with models that relate the variables multiplicatively because as Nicolai (2009) puts it "an outcome often depends on the factor in the shortest supply". Multiplicative models have the potential to be superior for many sociopolitical issues (Nicolai, 2009).

There are many other possible quantitatively predictive logical models applicable to social science but they are outside the scope of the paper; for example, growth rates and communication channels among many.

Historically each discipline went from description models to quantitative predictive approaches, by looking at the index of quantitative formalism (amount of equation, tables and graphs), even physics and chemistry publications had a very low index 300 years ago. It is not sure whether social sciences can ultimately reach physical sciences in predictive power because the social, psychological and political component is hard to account for (Tart, 2009) (Taagepera,2008); however, the physical world must have looked as complex as social science look to us today in Galileo's times and yet he found the law of fall despite contradicting evidence such as the fact that feathers float. As Colomer puts it "Physics does not predict the future in an unconditional sense. It merely says that if certain conditions are fulfilled, then certain outcomes can be expected" (Taagepera,2008) and this is what social science should strive for. Even in cosmology and sub-atomic physics the normal physics laws are not taken for granted (Taagepera,2008). What can be done now is make social sciences the most predictive as possible; in fields like electoral and party studies, marketing, business and sales advances have been made at least to the point where they are useful to practitioners.

According to Robert Adcock, social science should drop the aspiration to be exactly like natural science, instead of focusing on "epistemic" more theoretical knowledge it should focus on "phronetic" knowledge which is a more practical kind of knowledge which would make social science more usable (Tart, 2009). A major issue is that researchers that wants to be more scientific face challenges and risk getting the publication rejected as there is resistance and backwardness that favors oversimplification of models (Schwab, 2011). This

is why Taagepera puts his faith in individual researchers that act as agents of change toward new editorial policies (Shwab, 2011).

4. The web, computational social science and data-driven marketing

With the new technologies and the incredible amount of data that is generated every day, social science, in its broader sense, can benefit - and has already in some ways - from those to achieve more informed and more useful models. Web data is providing new understanding for social science and it is proving to be useful for business/sales related models (Ackland, 2013).

The intense digitalization and the digital footprints of internet users allows computational techniques to collect data cheaply and effortlessly; in the past it was hard to find the data for research, now the world has a lot more data than it can be analyzed. Computational Social Sciences (CSS) is the modern version of social science that is more computer and data based. It is more systematic and scientific than traditional social science. It is a new era where the data is already available and it is up to the social researchers to apply methods and find theories (Egger, Stuetzer and Welker, 2018).

The web can help in both quantitative and qualitative research; data science can mostly help out with quantitative research. The main data sources are online surveys, the digital trace that user leave by using the internet, simply downloading datasets that have been taken by others (secondary data) and quantitative web content analysis (Ackland, 2013).

The digital trace refers to search history, usage of social media (friends, likes, posts, comments etc.), buying products online etc.; for example, with Facebook you can select the characteristics of the people your ad or your survey is supposed to reach with the tool Facebook Ads Manager, this is a form of sampling or targeted advertisement. Nowadays, older segments of population are more likely to use technology therefore online social research is less likely to have sampling errors due to coverage bias compared to the first years of the internet. An example of web content analysis is Facebook's A.I. that tries to detect suicidal tendencies from users' posts or a 2010 research from Ackland et al. that tried to use keywords to assess how nanotechnology organizations were handling the possible social issues related to nanotechnology (Ackland, 2013).

Social medias are particularly apt for automated data analysis compared to less structured web pages. But obtaining large datasets from commercial social media websites is difficult therefore researchers usually develop an app that allows Facebook users to give consent for their data to be used for research purposes like Aral and Walker did in 2010 when researching social influence in Facebook (Ackland, 2013). In microblogs such as Twitter, APIs can be used to research tweets on a particular topic (with a specific hashtag) and create the network of Twitter users who tweeted a hashtag or a text string. The APIs of Twitter are open and many types of data can be programmatically retrieved, for example by using the Python tool tweepy.

Researchers can collect data in three main methods: using raw data, using APIs and exploit user interfaces. Raw data can be log files that records access to websites or a freely available database such as Wikipedia. Some providers offer APIs to their platforms for data access. They are mainly designed to allow websites to be linked to other applications or for advertising management. They can be used for commercial and scientific purposes but there are terms of services to be respected. APIs allow for incredible data retrieval in formats such as XML and JSON with some limitations from the platform; for example, not all data are available because of privacy and because sometime applications want to keep some data for themselves and there are limitations to the number of requests within a specific time. APIs need a tool like Facepager or a script written in a programming language. Finally, user interfaces such as the one on social media that are openly available on a web browser display user information, thanks to the HTTP request-response cycle and HTML or XML knowledge it is relatively easy to web-scrape data in a targeted manner; this is used for those cases where APIs is not possible, for example for blogs (Jünger, 2018). Hyperlink networks also are an excellent way to conduct web social research; the web can be thought as a large electronic library of hyperlinked documents and crawling the web for those can return great insights such as which organization is affiliated with what. Some of the tools for that are Issue Crawler, SocSciBot or using HTML knowledge and python's package BeautifulSoup.

Web data, like any other type of social data can help formulate models in marketing research, for example a research from Aral and Walker in 2010 studied online marketing within Facebook and found that passive broadcasts were more useful than active personalized messages in order to get people to use an app (Ackland, 2013).

Rating systems such as Netflix "stars" and recommender systems such as the Amazon suggestions when buying products are all data-driven approaches to marketing and although not perfect they increase sales or at least provide useful services because they can be tailored to single individual users (Ackland, 2013). In this highly connected world, big-data analytics allows the creation of more personalized services and products. Data driven marketing is now fundamental, it enhances and personalize user experience and campaigns with data-driven ads, it also allows marketers to understand their target audience and have a lot more information on potential customers (Kotler Kartajaya and Setiawan, 2017).

The use of web analytics is also useful in assessing digital marketing efforts with levels of monitoring impossible with the traditional marketing methods. Digital channels allow customers to be increasingly interacting with companies. Web analytics allows for things like: Tracking clicks and where they came from to gain insight on website traffic, calculating profits from and costs of specific digital marketing channels and finding the pathways that customer use to buy certain products/services (Joel and Heikki 2015). Some of the tools are even free, such as Google Analytics.

As seen computer based social sciences (in its broadest sense) is an exploding field and it is up to researchers to harness the power of all the world's data. Berkelaar and Francisco-

Revilla (2018) propose a new workflow for the new data and computer based social research

1. Examining motivation: what are we curious about
2. What evidence is available and what could/should be found
3. Creating appropriate algorithms in line with the theoretical aspect

The workflow starts from point 1 and then it becomes iterative and could go in any direction, this is much more flexible and modern than the classic and rigid social sciences workflow of finding null hypothesis, gathering data, rejecting null hypothesis.

5. First hand example 1: Webscraping

5.1 Introduction

In this example I will fetch data from the internet and tidy it up with the goal to make it ready for analysis. The output will be clean tabular data gathered from a webpage. Then some descriptive analysis will be run on the output data.

The webpage in question is a New York Times article:
https://www.nytimes.com/interactive/2017/06/23/opinion/trumps-lies.html

This page has collected a multitude of quotations from United States of America`s president Donald Trump that are deemed to be lies from the authors David Leonhardt and Stuart A. Thompson

Using python 3.5, its package BeautifulSoup and html knowledge it is possible to fetch the relevant data from the page.

5.2 Procedure

This task was carried out with jupyter notebook using Python 3.5.

First, the package “requests” and “BeautifulSoup” were loaded, then the webpage was loaded into python and parsed. At this point the data is very unclean and impossible to use for any analysis.

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

This is a sample of what the data looks like:

```
[<span class="short-desc"><strong>Jan. 21 </strong>"I wasn't a fan o
f Iraq. I didn't want to go into Iraq." <span class="short-truth"><a
href="https://www.buzzfeed.com/andrewkaczynski/in-2002-donald-trump-
said-he-supported-invading-iraq-on-the" target="_blank">(He was for
an invasion before he was against it.)</a></span></span>,
 <span class="short-desc"><strong>Jan. 21 </strong>"A reporter for T
ime magazine — and I have been on their cover 14 or 15 times. I thin
k we have the all-time record in the history of Time magazine." <spa
n class="short-truth"><a href="http://nation.time.com/2013/11/06/10-
things-you-didnt-know-about-time/" target="_blank">(Trump was on the
cover 11 times and Nixon appeared 55 times.)</a></span></span>,
```

Fig 1

There are 180 sentences deemed as lie on the webpage.

BeautifulSoup allows to search for tags, strings, nested tags and nested strings - among other commands - from the html code of the webpage.

It is then possible to create a loop that cycle through the whole parsed html code and extract the:

- date of the lie
- the lie
- the reason why it is a lie (provided by the author)
- and a link to an article that debunks the lie

This loop is available in the appendices to the thesis as "example1code.ipynb".

Once the information is gathered it is put into a dataframe for easy access and possible analysis.

This is how the data looks after the procedure:

	date	lie	reason	hyperlink
0	Jan. 21, 2017	I wasn't a fan of Iraq. I didn't want to go in...	He was for an invasion before he was against it.	https://www.buzzfeed.com/andrewkaczynski/in-20...
1	Jan. 21, 2017	A reporter for Time magazine — and I have been...	Trump was on the cover 11 times and Nixon appe...	http://nation.time.com/2013/11/06/10-things-yo...
2	Jan. 23, 2017	Between 3 million and 5 million illegal votes ...	There's no evidence of illegal voting.	https://www.nytimes.com/2017/01/23/us/politics...
3	Jan. 25, 2017	Now, the audience was the biggest ever. But th...	Official aerial photos show Obama's 2009 inaug...	https://www.nytimes.com/2017/01/21/us/politics...
4	Jan. 25, 2017	Take a look at the Pew reports (which show vot...	The report never mentioned voter fraud.	https://www.nytimes.com/2017/01/24/us/politics...

Date: 17/09/2018
Fig 2

The dataframe is then exported as csv, which is available in the appendices as “example1data.csv”.

On this dataset it is possible to do qualitative analysis which is out of the scope of this paper; however, further possible quantitative analysis can be done; fore example, understanding the distribution of the dates of the lies, which can be helpful in spotting trends based on the events at the time of the lie, or to count how many times a specific source appears in the hyperlinks provided by the two authors, which is helpful to understand if there is bias in their sources.

Fig 3 is a bar plot showing how many lies are in each month; February and October are the months with the most lies.

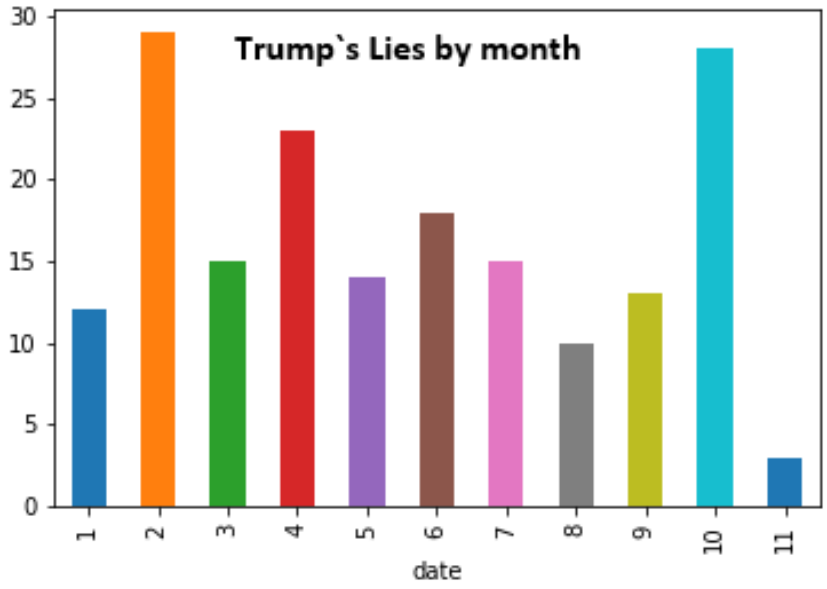

Fig 3

In order to find out which are the more prevalent sources the hyperlinks column was cleaned up so that only the name of the source is visible and not the entire link.

Fig 4 shows the distribution of the sources.

Distribution of Hyperlink Sources

nytimes.com	57	transcripts.cnn.com	1
washingtonpost.com	39	warontherocks.com	1
politifact.com	31	markets.on.nytimes.com	1
factcheck.org	10	chicagotribune.com	1
content.govdelivery.com	5	pbs.org	1
cnn.com	4	businessinsider.com	1
usatoday.com	3	nation.time.com	1
time.com	3	thehill.com	1
buzzfeed.com	2	heritage.org	1
money.cnn.com	2	palmbeachpost.com	1
realclearpolitics.com	2	opensecrets.org	1
bea.gov	2	pewresearch.org	1
snopes.com	2	talkingpointsmemo.com	1
dnainfo.com	1	public.tableau.com	1
washingtonmonthly.com	1	mdjonline.com	1
nbcnews.com	1		

Fig 4

It is possible to see that most of the hyperlinks used as evidence for Trump`s lies come from the New York Times itself, which could mean that the authors may have some bias as they are both New York Times` writers.

Moreover, the New York Times and the Washington Post, the two main sources, are both generally considered center-left wing news outlets according to mediabiasfactcheck.com. While politifact.com and factcheck.org are considered unbiased by mediabiasfactcheck.com, CNN, Usa Today, Time and BuzzFeed are all considered left to center-left wing. This could imply a bias against Trump who is a right wing politician or it could mean that the right wing news outlet are more than willing to close an eye on Trump`s lies.

6. First hand example 2: Recommender System

6.1 Introduction

This example uses the freely available dataset from MovieLens, a movie recommendation service.

The tasks accomplished are descriptive statistics on the two datasets, clustering on the feature space and most importantly the development of two recommender systems, one

Content-based filtering and the other User-Based Collaborative Filtering (UBCF). The analysis has required in multiple occasions to pre-process the data and change its structure.

Recommender systems are becoming more and more relevant. They fully embrace the new type of data-driven marketing or services; the core of recommender system is personalization. Examples of recommender systems are the suggested videos in YouTube, the suggested items in Amazon or the movies in Netflix. There is too much content or items to be watched/bought, recommender systems attempt to give users the possibility of being exposed to mainly products and services that are of potential interest to them.

Recommender systems are mainly either Content-based Filtering or Collaborative Filtering. The project attempts to create one for both methods with the ultimate goal of suggesting interesting movies to users.

6.2 Data

The dataset can be downloaded from the following link: https://grouplens.org/datasets/movielens/ . It is the MovieLens Latest Datasets, ml-latest.zip (size: 224 MB). Or it can be found in in my home directory on the DSM cluster in the PROJECT.DATA folder with the name “ml-latest.zip”.

The dataset contains different csv files that are connected, the project only uses the files movies.csv and ratings.csv. The first contains information about different movies with their movieId, the second contains a lot of observations about user ratings. The column movieId is consistent between the two csv files.

A detailed explanation of the datasets with variable description and the movielens data collection methodology is provided as an appendix.

6.3 Analysis

The project is entirely developed in R and the whole code is contained in one R file which has plenty of comments that describe each passage (“example2code.R”).

The two datasets described above were loaded. The observations for the ratings dataset were heavily reduced because the amount of data was unmanageable by RStudio.

Some descriptive statistics were carried out, for example the following is the histogram of how many times different ratings are selected by users.

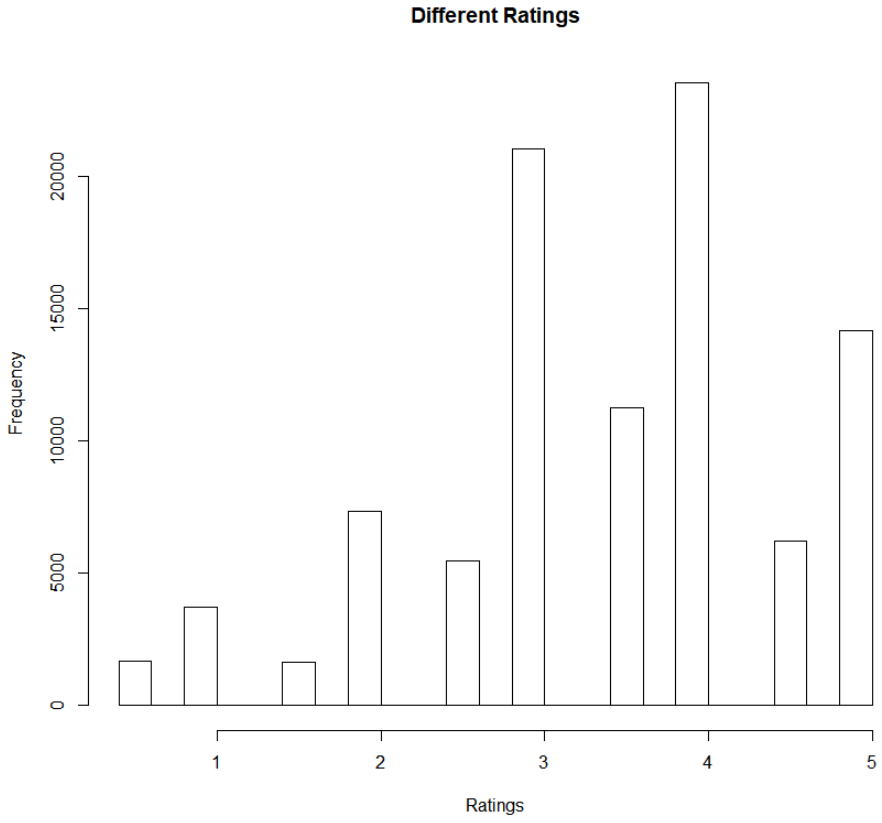

Fig 5

The ratings mean across all users (and movies) is 3.44 which is 0.94 points away from the median 2.5. It can also be noted that most people do not give half stars.

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

This is the histogram of users' average scores:

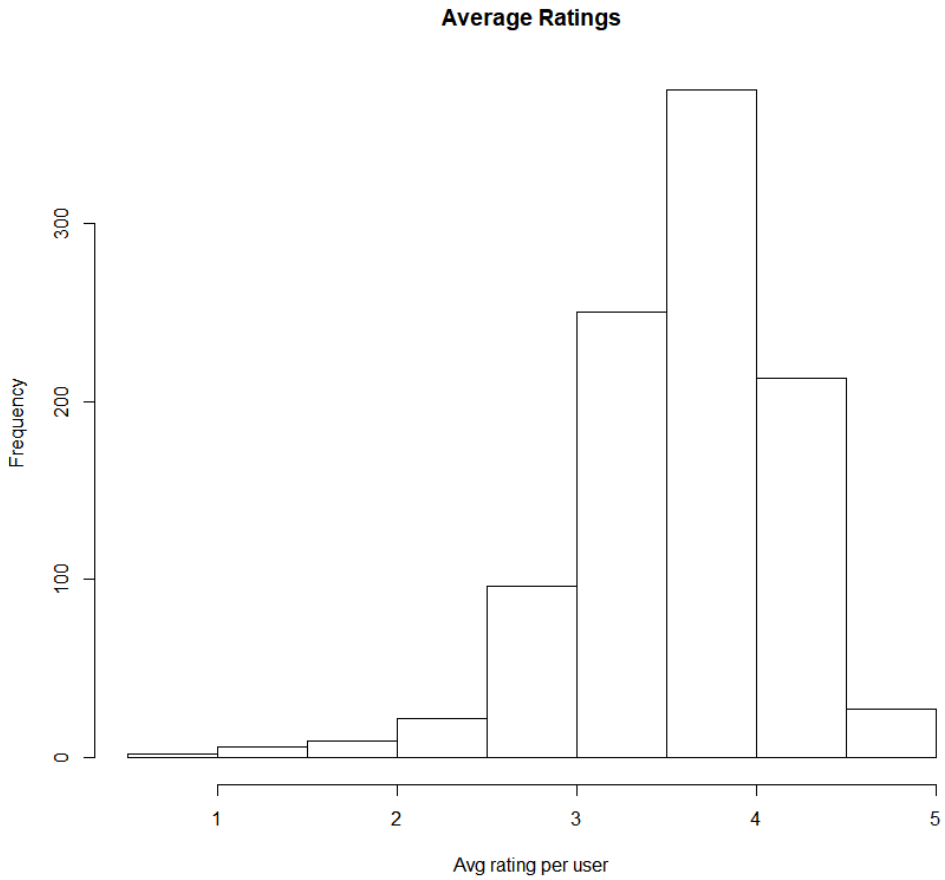

Fig 6

and its density curve (now that it is an average, it is a continuous variable):

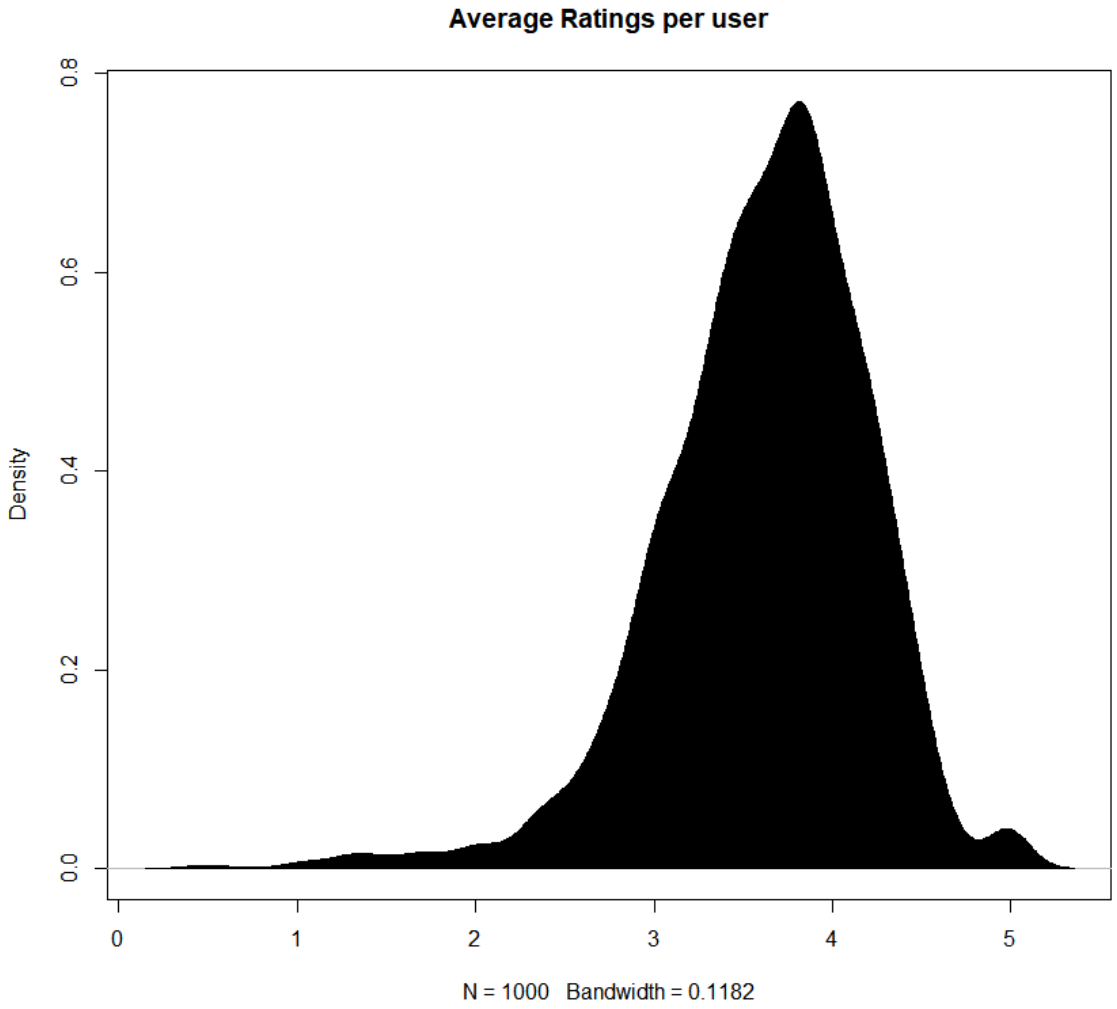

Fig 7

We can see how users are more inclined to give higher ratings on average.

The next descriptive statistic is the top and bottom movies by average user ratings.

Top movies:

movieId	avgratings	
48	49	5
78	80	5
168	190	5
326	363	5
413	467	5
492	549	5

Fig 8

Bottom movies:

```
     movieId avgratings
358      397        0.5
596      702        0.5
642      777        0.5
1969    2464        0.5
2333    2909        0.5
2541    3165        0.5
```

Fig 9

The movie Ids correspond to a movie title, the code in the appendices allow you to input the movie id and retrieve the movie title and its genres.

A list of the most common genres assigned in movies was computed:

```
Drama       19806
Comedy      13001
Thriller     6761
Romance      6069
Action       5775
Horror       4448
Crime        4247
Documentary  4122
Adventure    3368
Sci-Fi       2847
Mystery      2274
Fantasy      2211
Children     2181
Animation    1941
War          1544
Musical      1079
Western      1028
Film-Noir     360
IMAX          197
```

Fig 10

Date: 17/09/2018

This is the bar plot for the movies genres above:

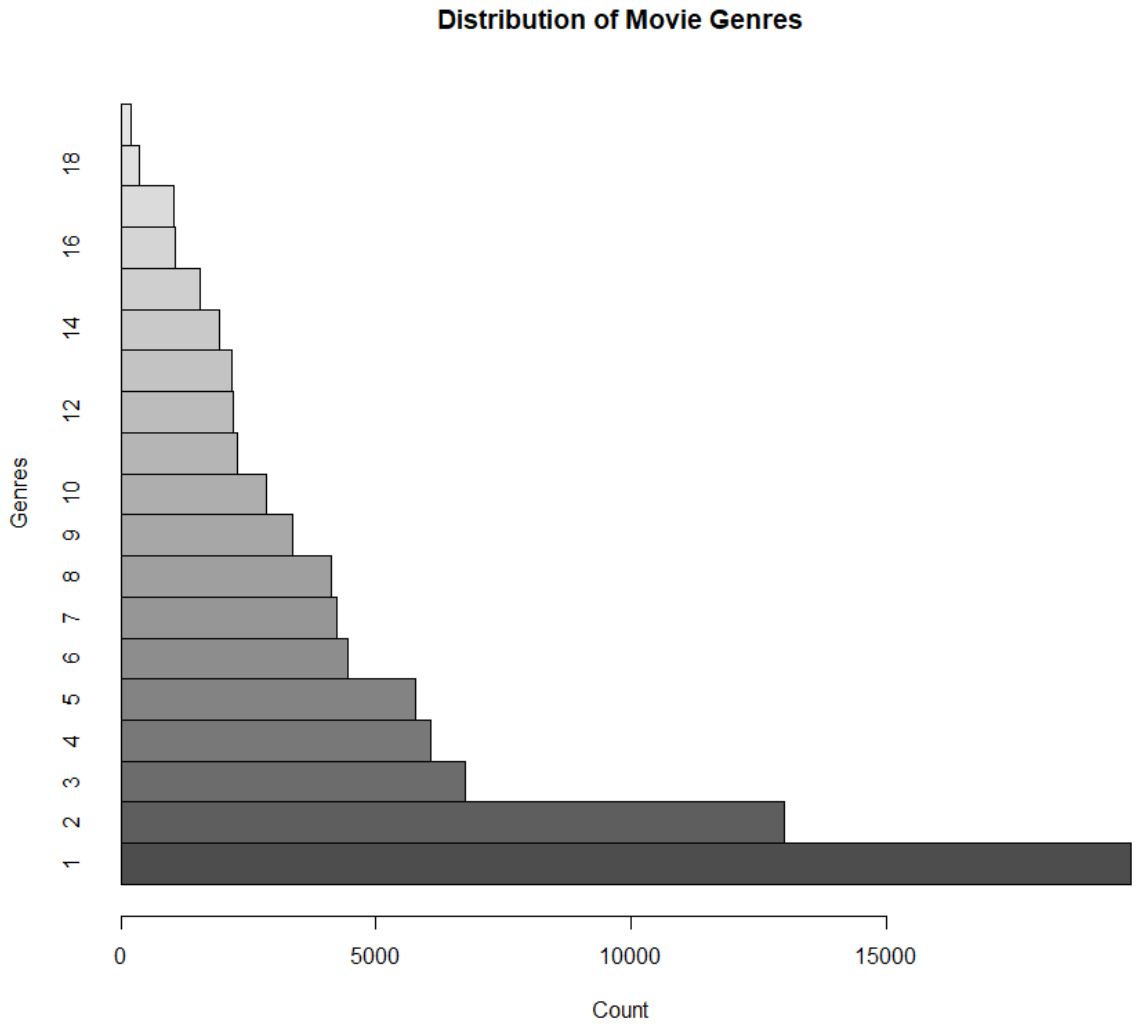

Fig 11

In order to not overcrowd the labels, the genres have a number assigned instead of their name, the table above has the same order of the bar chart; therefore, it is easy to check which number corresponds to which genre.

The next piece of statistics is ranking all the individual genres by the average of all users' ratings. This was obtained by finding the average rating for each movie, subsequently finding the average rating for each genre across all the movies with that assigned genre (movies can have more than one genre assigned), again, a descripted procedure is found in the code.

Date: 17/09/2018

These are the average ratings for each genre:

```
1  Genre        Average ratings
2  Documentary   3.401878
3  Film-Noir     3.274961
4  IMAX          3.271057
5  War           3.229247
6  Drama         3.185497
7  Musical       3.180637
8  Romance       3.155419
9  Animation     3.101921
10 Crime         3.084989
11 Mystery       3.046006
12 Comedy        3.045956
13 Western       3.026998
14 Fantasy       3.011245
15 Adventure     3.005380
16 Children      2.994169
17 Thriller      2.916095
18 Action        2.907481
19 Sci-Fi        2.770549
20 Horror        2.569902
```

Fig 12

It can be observed that differences are present but not sharp.
An interesting aspect to notice is that some of the most widely available/most watched movies genres such as Thriller, Action and Horror (Fig 10) are found at the bottom of the above table while less available/less watched movie genres such as Documentary and Film-Noir are found at the bottom of Fig 10.

Date: 17/09/2018

Below is the bar chart of the movie genre ratings.

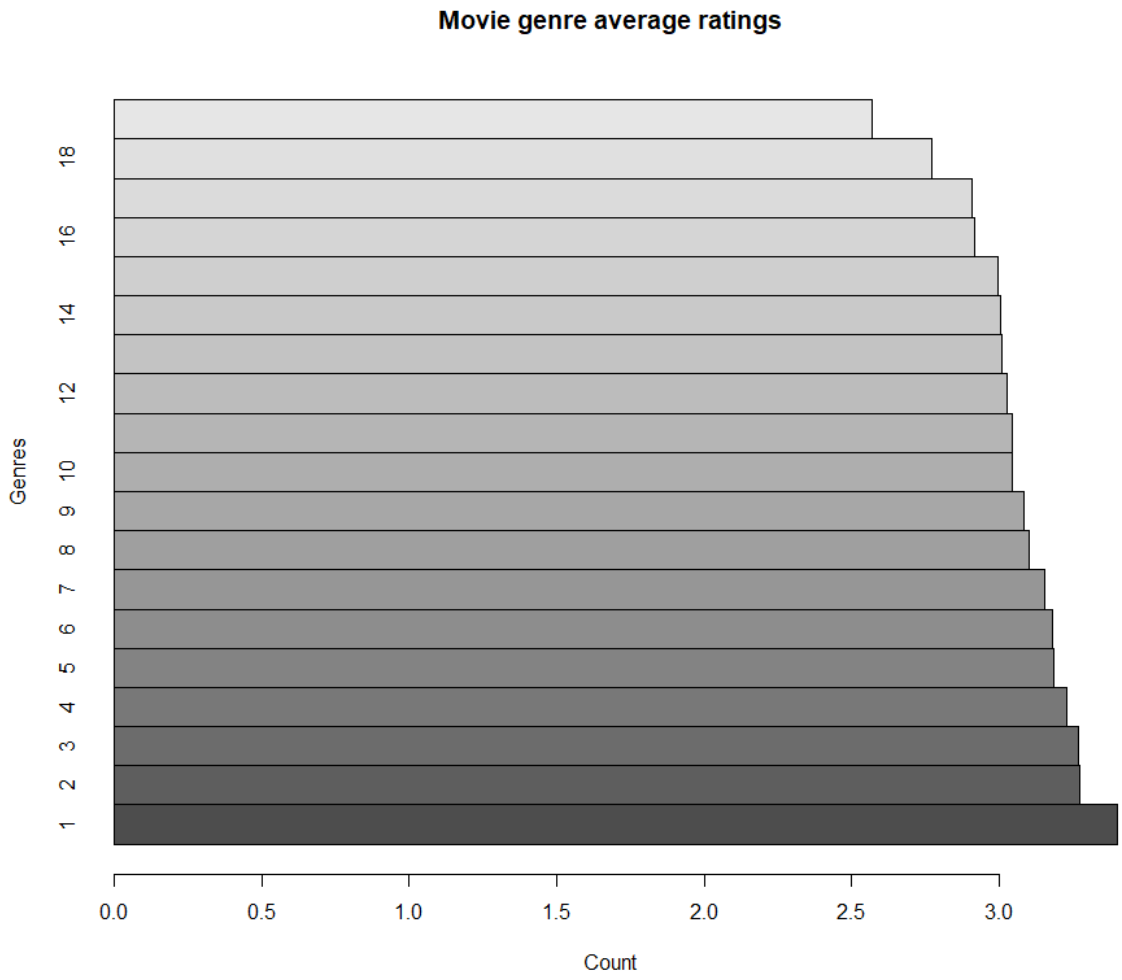

Fig 13

As discussed above the differences between average genre ratings are not too big. The numbers on the y-axis in Fig 13 are different genres and are the same numbers assigned in Fig 12.

6.4 Clustering

Now I will perform two different k-means clustering, one by clustering the movie Ids with similar ratings and another by clustering the genres with similar ratings.

The clustering is done on 10,000 ratings for computation purposes, this is its visualization:

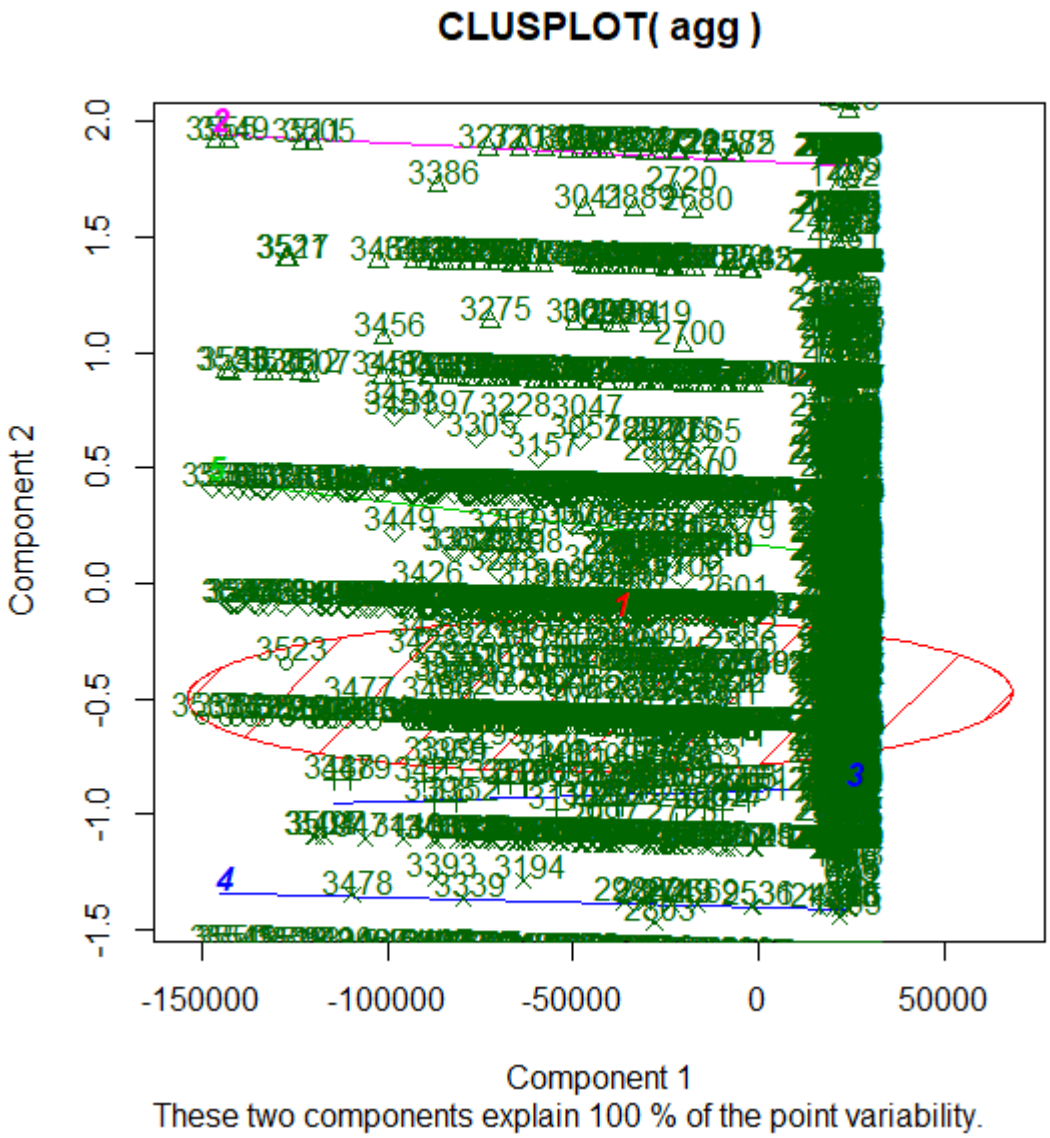

Fig 14

It is possible to observe that the only meaningful cluster is cluster 1. The following is a breakdown of the cluster centres:

```
1 3.914785
2 1.840065
3 4.258913
4 4.738287
5 3.188843
```

Fig 15

It is worth mentioning that depending on the sampling of the observations (moviesId) it is not always possible to find relatively defined clusters.

The other clustering is carried out on genres with similar ratings.
Breakdown of the cluster centres:

```
1 3.258422
2 3.401878
3 3.141693
4 2.670225
5 2.994166
```

Fig 16

This is the visualization of the clustering:

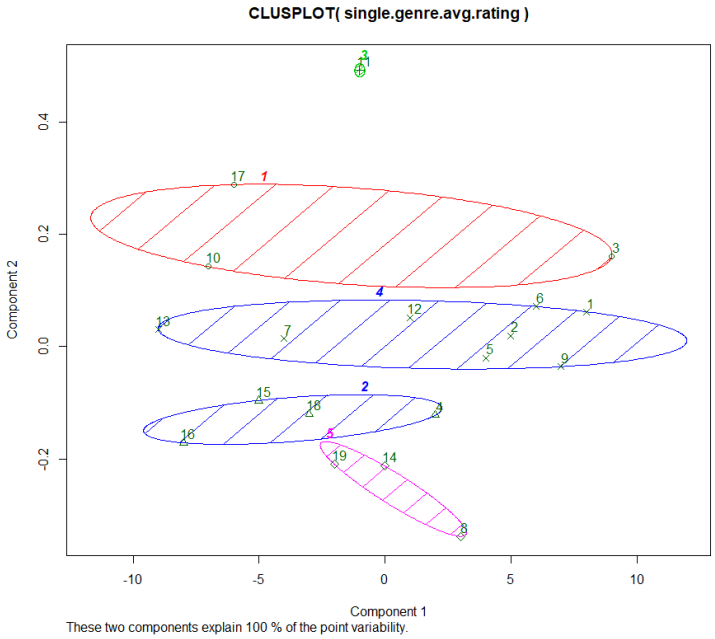

Fig 17

It can be seen that genres by rating did produce clusters but they do not seem that well defined.

6.5 Recommender Systems

Now I will describe the recommender systems developed.

For the Content-Based filtering on genres the ratings dataset was simplified to a binary rating scale, the ratings below 3 were considered dislikes while above considered likes. Also the movies without any ratings were removed as long as all the movies that did not have a genre attributed to them.

Various dataframes were obtained by processing the data in the two original datasets, these dataframes were used to compute the recommender systems methods. The comments in the code explain it step-by-step.

Once those essential matrices were obtained a user matrix with columns as users' ids and rows as genres was found. This matrix was calculated by the dot product of two previous matrices and each cells stands for the general inclination of a user id toward a genre. Subsequently, the values were turned into binary. The Jaccard distance between the user matrix and the movies was calculated. The result of the computation is that for a given user it is possible to retrieve the recommended movies. This method favors movies that have fewer genres attached to them, because it is easier to find similarity between users. Content Based approaches like this consists in knowing the items users have showed interest in, looking for similar items and suggesting them; therefore, similar items are suggested based on items' features. This method does not require much user data but the suggestions will be of the same type, providing possibly tedious recommendations to users who will be presented with substitutes rather than alternatives. It also requires the features of the items to be relatively distributed, for example, if most of the items share the same features values there is little prediction to be made.

For the Collaborative Filtering the package recommenderlab was used which made the procedure easier. This approach clusters users on the basis of their behavior history and seeks to recommend movies that a similar user watched and liked. The UBFC method from recommenderlab was used and the result of the computation is a list of recommendations for a specific user (which can be easily changed in the code).

Collaborative Filtering systems like this look at different users' preferences and recommend to one user what other users have found interesting. This simulates how people in real life suggest items to one another, these methods also provide complementary and alternative items rather than just a substitute. These systems are generally more complex and use a lot of user data which needs to be tracked/collected, maintained. Moreover, the computations can be extremely long depending on the method.

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

Recommenderlab allows to test the Collaborative Filtering recommender system just developed by 5-fold cross validation. this is the output confusion matrix:

```
     TP    FP     FN       TN  precision      recall         TPR
1  0.020 0.065 12.970 11930.94 0.23529412 0.002564864 0.002564864
3  0.040 0.215 12.950 11930.80 0.15686275 0.005102123 0.005102123
5  0.050 0.375 12.940 11930.64 0.11764706 0.007408203 0.007408203
10 0.055 0.795 12.935 11930.22 0.06470588 0.007478869 0.007478869
15 0.080 1.195 12.910 11929.82 0.06274510 0.008200003 0.008200003
20 0.090 1.610 12.900 11929.40 0.05294118 0.009870601 0.009870601
             FPR
1  5.446137e-06
3  1.801824e-05
5  3.143356e-05
10 6.664609e-05
15 1.001761e-04
20 1.349707e-04
```

Fig 18

this assesses performance of the top 1,3,5,10,15,20 recommenders. Sensitivity (or recall/True Positive Rate) and specificity are both low which is to be expected in such a task where the model is trying to guess which movie will be liked by a user.

For the first recommender:

- sensitivity (or recall or true positive rate) = $\frac{TP}{TP+FN}$ = probability of a positive test with a positive entry = 0.002564864
- precision =$\frac{TP}{TP+FP}$ = fraction of relevant instances among the retrieved instances = 0.23529412
- F-measure = $2\,\frac{precision \cdot sensitivity}{precision+sensitivity}$ = 0.0050744134 . This is the harmonic mean of precision and recall; it has maximum 1 and minimum 0.

7. First Hand example 3: Multivariate Analysis

7.1 Introduction and Objective

The dataset used is the European Social Survey (ESS) round 6 from 2012. The survey is taken in 36 European countries and it is a collection of social statistics used to monitor and interpret public attitudes and values. The Sampling method used was partly repetitive cross section. Some of the columns had been removed because of too many missing values.

The objective of the project is to investigate some of the questions in the survey that are part of how people see themselves, namely the master question H inside of the ESS round 6.

Conclusions will be drawn if any is found from a multiple correspondence analysis on the data and its multiple visualizations.

7.2 Data preparation and description

The software used for the analysis is R (version 3.4.3) used through RStudio.

The main library used is FactoMiner and factoextra.

The dataset contains 52177 observations and 50 variables, each observation represent a person and the variables are:

```
> names(ESS3)
 [1] "cntry"   "gndr"    "agea"    "eisced"  "ipcrtiv" "imprich" "ip
eqopt"
 [8] "ipshabt" "impsafe" "impdiff" "ipfrule" "ipudrst" "ipmodst" "ip
gdtim"
[15] "impfree" "iphlppl" "ipsuces" "ipstrgv" "ipadvnt" "ipbhprp" "ip
rspot"
[22] "iplylfr" "impenv"  "imptrad" "impfun"  "wkvlorg" "optftr"  "ps
tvms"
[29] "flrms"   "fltdpr"  "flteeff" "slprl"   "wrhpp"   "fltlnl"  "en
jlf"
[36] "fltsd"   "cldgng"  "enrglot" "fltanx"  "fltpcfl" "dclvlf"  "lc
hshcp"
[43] "accdng"  "wrbknrm" "pplahlp" "trtrsp"  "dngval"  "nhpftr"  "lf
wrs"
[50] "flclpla"
```

Fig 19

All the columns were converted to factor, beside agea which is converted to integer.

Now I will describe every variable below and their possible values because the variable names are not self-explanatory. The variable description are taken from: http://www.europeansocialsurvey.org/docs/round6/survey/ESS6_appendix_a7_e02_2.pdf .

Date: 17/09/2018

I will provide here a summary list with the definition for each variable in order to quickly understand the meaning of the variables (the definitions are taken from the above link):

- contextual variables

cntry = country

gndr = gender

agea = age

eisced = education level

- First set of questions:

For the variables below the following question applies = “Now I will briefly describe some people. Please listen to each description and tell me how much each person is or is not like you. Use this card for your answer”.

1 Very much like me

2 Like me

3 Somewhat like me

4 A little like me

5 Not like me

6 Not like me at all

7 Refusal

8 Don't know

9 No answer

ipcrtiv = Important to think new ideas and being creative

imprich = Important to be rich, have money and expensive things

ipeqopt = Important that people are treated equally and have equal opportunities

ipshabt = Important to show abilities and be admired

impsafe = Important to live in secure and safe surroundings

impdiff = Important to try new and different things in life

Date: 17/09/2018

ipfrule = Important to do what is told and follow rules

ipudrst = Important to understand different people

ipmodst = Important to be humble and modest, not draw attention

ipgdtim = Important to have a good time

impfree = Important to make own decisions and be free

iphlppl = Important to help people and care for others well-being

ipsuces = Important to be successful and that people recognize achievements

ipstrgv = Important that government is strong and ensures safety

ipadvnt = Important to seek adventures and have an exciting life

ipbhprp = Important to behave properly

iprspot = Important to get respect from others

iplylfr = Important to be loyal to friends and devote to people close

impenv = Important to care for nature and environment

imptrad = Important to follow traditions and customs

impfun = Important to seek fun and things that give pleasure

- Second set of questions:

The next variables have different value encodings, therefore they are difficult to group, I will only provide their definition**.**

wkvlorg = Involved in work for voluntary or charitable organizations, how often past 12 months

optftr= Always optimistic about my future

pstvms = In general feel very positive about myself

flrms = At times feel as if I am a failure

fltdpr = Felt depressed, how often past week

flteeff = Felt everything did as effort, how often past week

slprl = Sleep was restless, how often past week

wrhpp = Were happy, how often past week

fltlnl = Felt lonely, how often past week

enjlf = Enjoyed life, how often past week

fltsd = Felt sad, how often past week

cldgng = Could not get going, how often past week

enrglot = Had lot of energy, how often past week

fltanx = Felt anxious, how often past week

fltpcfl = Felt calm and peaceful, how often past week

dclvlf = Free to decide how to live my life

lchshcp = Little chance to show how capable I am

accdng = Feel accomplishment from what I do

wrbknrm = When things go wrong in my life it takes a long time to get back to normal

pplahlp = Feel people in local area help one another

trtrsp = Feel people treat you with respect

dngval = Feel what I do in life is valuable and worthwhile

nhpftr = Hard to be hopeful about the future of the world

lfwrs = For most people in country life is getting worse

flclpla = Feel close to the people in local area

Here I will explain the contextual variables more in depth:

cntry= country

Possible values =

AT Austria

BE Belgium

BG Bulgaria

CH Switzerland

CY Cyprus

CZ Czech Republic

Date: 17/09/2018

DE Germany

DK Denmark

EE Estonia

ES Spain

FI Finland

FR France

GB United Kingdom

GR Greece

HR Croatia

HU Hungary

IE Ireland

IL Israel

IS Iceland

IT Italy

LT Lithuania

LU Luxembourg

NL Netherlands

NO Norway

PL Poland

PT Portugal

RU Russia

SE Sweden

SI Slovenia

SK Slovakia

TR Turkey

UA Ukraine

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

Table showing how many observations per each variable:

```
cntry
  BE   BG   CH   CY   CZ   DE   DK   EE   ES   FI   FR   GB   HU   I
E   IL
1869 2260 1493 1116 2009 2958 1650 2380 1889 2197 1968 2286 2014 262
8 2508
  IS   IT   LT   NL   NO   PL   PT   RU   SE   SI   SK   UA
 752  960 2109 1845 1624 1898 2151 2484 1847 1257 1847 2178
```

Fig 20

gndr = gender

possible values= 1 for male, 2 for female,9 for no answer

Distribution of the gndr column:

```
gndr
    1     2     9
23762 28398    17
```

Fig 21

Agea= age, 999 is for no answer

Visualizing age as a histogram since a table is not much use

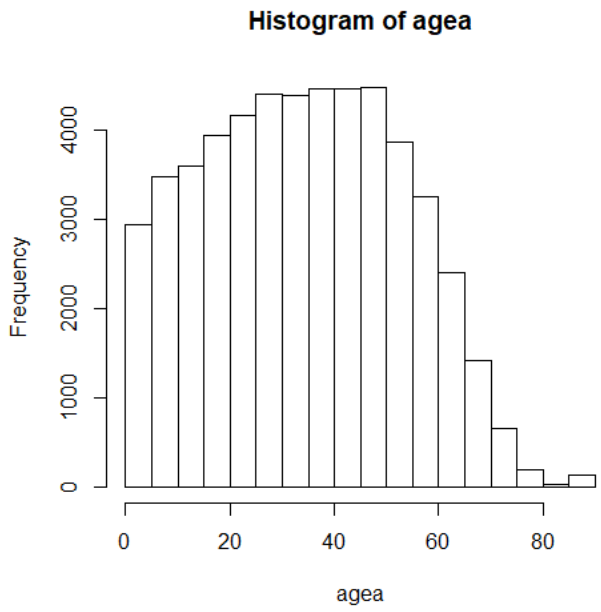

Fig 22

Date: 17/09/2018

eisced= highest level of education, 1 for less than lower secondary and 7 for higher tertiary education (>= MA level)

```
eisced
    1     7
45548  6629
```

Fig 23

The rest of the questions all follow their own value coding therefore they cannot be easily grouped.

Detailed description of the questions can be found at this address
http://www.europeansocialsurvey.org/docs/round6/survey/ESS6_appendix_a7_e02_2.pdf

This is an overview of the dataset which can be helpful in understanding the next steps:

```
     cntry          gndr           agea        eisced       ipcrtiv     
 DE     : 2958   1:23762   Min.   : 15.00   1:45548   2      :16635  
 IE     : 2628   2:28398   1st Qu.: 33.00   7: 6629   3      :13081  
 IL     : 2508   9:   17   Median : 49.00             1      :10279  
 RU     : 2484             Mean   : 50.61             4      : 6401  
 EE     : 2380             3rd Qu.: 63.00             5      : 3856  
 GB     : 2286             Max.   :999.00             6      : 1118  
 (Other):36933                                        (Other):  807  
    imprich         ipeqopt         ipshabt         impsafe      
 5      :15639   2      :20701   2      :13975   2      :18498  
 4      :11163   1      :17456   3      :13138   1      :15676  
 3      : 9936   3      : 8360   4      : 8573   3      : 9540  
 6      : 6384   4      : 3231   5      : 7295   4      : 4567  
 2      : 5975   5      : 1379   1      : 6515   5      : 2713  
 1      : 2422   8      :  455   6      : 1907   6      :  552  
 (Other):  658   (Other):  595   (Other):  774   (Other):  631  
    impdiff         ipfrule         ipudrst         ipmodst      
 2      :13860   2      :14435   2      :21182   2      :17566  
 3      :12666   3      :12267   3      :12329   3      :12565  
 4      : 9050   4      : 8386   1      :10090   1      : 9050  
 1      : 7504   5      : 7763   4      : 5304   4      : 7127  
 5      : 6643   1      : 6038   5      : 1959   5      : 4243  
 6      : 1702   6      : 2306   8      :  536   6      :  878  
 (Other):  752   (Other):  982   (Other):  777   (Other):  748  
    ipgdtim         impfree         iphlppl         ipsuces      
 2      :14598   2      :19933   2      :21765   2      :13831  
 3      :12757   1      :15439   1      :13495   3      :13510  
 4      : 8834   3      : 9916   3      :10940   4      : 9151  
 1      : 7061   4      : 4152   4      : 4032   5      : 6952  
 5      : 6193   5      : 1650   5      : 1043   1      : 6122  
 6      : 1993   8      :  428   8      :  406   6      : 1749  
 (Other):  741   (Other):  659   (Other):  496   (Other):  862  
    ipstrgv         ipadvnt         ipbhprp         iprspot      
 2      :18946   5      :13488   2      :18769   2      :13925  
 1      :15657   4      :10347   3      :12016   3      :12954  
 3      : 9665   3      : 9891   1      : 9456   4      : 8905  
```

```
 4      : 4411   2      : 7341   4      : 6551   5      : 7441
 5      : 2050   6      : 6856   5      : 3765   1      : 6314
 8      :  652   1      : 3444   6      :  775   6      : 1761
 (Other):  796   (Other):  810   (Other):  845   (Other):  877
    iplylfr          impenv          imptrad          impfun
 2      :22222   2      :19689   2      :16339   2      :13614
 1      :18410   1      :16466   3      :11338   3      :12599
 3      : 7472   3      : 9965   1      :10940   4      : 9499
 4      : 2498   4      : 3793   4      : 6767   1      : 6548
 5      :  684   5      : 1223   5      : 4442   5      : 6480
 8      :  402   8      :  466   6      : 1640   6      : 2655
 (Other):  489   (Other):  575   (Other):  711   (Other):  782
```

```
wkvlorg          optftr          pstvms          flrms       fltdpr
 6      :32870   2      :25061   2      :29998   4      :19763   1:3
0817
 5      : 6376   3      :10593   1      : 9519   5      :11887   2:1
6708
 2      : 3511   1      : 9654   3      : 8822   2      : 9461   3:
3207
 1      : 3386   4      : 5376   4      : 2931   3      : 9441   4:
1091
 4      : 3065   5      : 1143   5      :  564   1      : 1125   7:
40
 3      : 2477   8      :  293   8      :  281   8      :  414   8:
281
 (Other):  492   (Other):   57   (Other):   62   (Other):   86   9:
33
 flteeff   slprl     wrhpp     fltlnl    enjlf     fltsd     cldgng
 1:23642   1:23131   1: 2611   1:34069   1: 3078   1:25648   1:26075
 2:19979   2:19762   2:12577   2:12859   2:13241   2:21476   2:19694
 3: 6055   3: 6367   3:24020   3: 3371   3:22134   3: 3477   3: 4259
 4: 2071   4: 2647   4:12278   4: 1511   4:13101   4: 1179   4: 1355
 7:   24   7:   28   7:   32   7:   29   7:   36   7:   26   7:   25
 8:  359   8:  198   8:  599   8:  298   8:  532   8:  328   8:  709
 9:   47   9:   44   9:   60   9:   40   9:   55   9:   43   9:   60
 enrglot   fltanx    fltpcfl       dclvlf         lchshcp
 1: 5188   1:25180   1: 3414   2      :24759   4      :18399
 2:16782   2:20457   2:14141   1      :15996   2      :12954
 3:21335   3: 4622   3:23920   3      : 7255   3      :12644
 4: 8337   4: 1453   4:10192   4      : 3281   5      : 4655
 7:   20   7:   30   7:   24   5      :  634   1      : 2846
 8:  476   8:  395   8:  437   8      :  211   8      :  618
 9:   39   9:   40   9:   49   (Other):   41   (Other):   61
     accdng          wrbknrm          pplahlp          trtrsp
 2      :28706   4      :21369   4      :12169   5      :18609
 3      :10029   3      :11973   5      :11107   4      :12747
 1      : 8015   2      :10337   3      :10982   6      :10088
 4      : 4261   5      : 5542   6      : 6021   3      : 6869
 5      :  681   1      : 2302   2      : 5798   2      : 1988
 8      :  423   8      :  586   1      : 2796   8      :  783
```

```
 (Other):   62   (Other):   68   (Other): 3304   (Other): 1093
    dngval          nhpftr          lfwrs          flclpla
 2      :30195   2      :18736   2      :20799   2      :24522
 1      :10610   3      :13278   1      :12611   3      :12436
 3      : 8210   4      :10969   3      :10065   1      : 7213
 4      : 2034   1      : 6374   4      : 7016   4      : 6034
 5      :  554   5      : 1897   5      :  888   5      : 1348
 8      :  521   8      :  874   8      :  745   8      :  569
 (Other):   53   (Other):   49   (Other):   53   (Other):   55
```

Fig 24

7.3 Multiple Correspondence Analysis

I am going to perform an MCA on the contextual variables and the first set of questions.

Creating new dataset with the columns used for this analysis.

As during the analysis, the graphs were incredibly overcrowded I am removing some variables. Namely cntry from the contextual variables and all the variables after ipfrule.

Variables used:

```
gndr" ,   "agea"  ,  "eisced" , "ipcrtiv" ,"imprich", "ipeqopt", "ip
shabt" ,"impsafe", "impdiff","ipfrule")]
```

Fig 25

These variables have information about how much respondents identify with the specific question asked (each variable is a question).

The variables cntry, gndr, agea and eisced are considered supplementary.

Carrying out the MCA

```
> MCA1=MCA(data.MCA.1, ncp = 5, ind.sup = NULL, quanti.sup = 2 ,qual
i.sup = c(1,3), excl=NULL, graph = TRUE)
> var<-get_mca_var(MCA1)
```

Fig 26

Let us look at the explained variance and eigenvalues for the top 25 dimensions

```
        eigenvalue variance.percent cumulative.variance.percent
Dim.1  0.886533476       11.0816685                     11.08167
Dim.2  0.856229820       10.7028727                     21.78454
Dim.3  0.548897189        6.8612149                     28.64576
Dim.4  0.338774897        4.2346862                     32.88044
Dim.5  0.295407365        3.6925921                     36.57303
Dim.6  0.225203694        2.8150462                     39.38808
Dim.7  0.210233809        2.6279226                     42.01600
Dim.8  0.175279361        2.1909920                     44.20700
Dim.9  0.163366135        2.0420767                     46.24907
Dim.10 0.154970666        1.9371333                     48.18621
Dim.11 0.150549280        1.8818660                     50.06807
```

Dim.12	0.146902999	1.8362875	51.90436
Dim.13	0.143534140	1.7941767	53.69854
Dim.14	0.141176712	1.7647089	55.46324
Dim.15	0.140304158	1.7538020	57.21705
Dim.16	0.139007482	1.7375935	58.95464
Dim.17	0.137593860	1.7199232	60.67456
Dim.18	0.136605826	1.7075728	62.38214
Dim.19	0.135898353	1.6987294	64.08087
Dim.20	0.134202268	1.6775284	65.75839
Dim.21	0.133692751	1.6711594	67.42955
Dim.22	0.131123167	1.6390396	69.06859
Dim.23	0.130540685	1.6317586	70.70035
Dim.24	0.128172264	1.6021533	72.30250
Dim.25	0.126784406	1.5848051	73.88731
Dim.26	0.125466284	1.5683286	75.45564
Dim.27	0.122243261	1.5280408	76.98368
Dim.28	0.120047387	1.5005923	78.48427
Dim.29	0.118091948	1.4761494	79.96042
Dim.30	0.116685664	1.4585708	81.41899
Dim.31	0.112665782	1.4083223	82.82731
Dim.32	0.110994321	1.3874290	84.21474
Dim.33	0.109269474	1.3658684	85.58061
Dim.34	0.103955147	1.2994393	86.88005
Dim.35	0.098927226	1.2365903	88.11664
Dim.36	0.094536944	1.1817118	89.29835
Dim.37	0.086369645	1.0796206	90.37797
Dim.38	0.085353458	1.0669182	91.44489
Dim.39	0.080287442	1.0035930	92.44848
Dim.40	0.076890449	0.9611306	93.40961
Dim.41	0.073689729	0.9211216	94.33074
Dim.42	0.072202471	0.9025309	95.23327
Dim.43	0.063372486	0.7921561	96.02542
Dim.44	0.061078150	0.7634769	96.78890
Dim.45	0.036333402	0.4541675	97.24307
Dim.46	0.029487358	0.3685920	97.61166
Dim.47	0.027193009	0.3399126	97.95157
Dim.48	0.025982364	0.3247795	98.27635
Dim.49	0.023668786	0.2958598	98.57221
Dim.50	0.022059516	0.2757439	98.84796
Dim.51	0.021119130	0.2639891	99.11194
Dim.52	0.017813331	0.2226666	99.33461
Dim.53	0.017129136	0.2141142	99.54873
Dim.54	0.014444535	0.1805567	99.72928
Dim.55	0.012314046	0.1539256	99.88321
Dim.56	0.009343357	0.1167920	100.00000

Fig 27

The cumulative variance percent reaches 100 % at the 56th dimension. It seems like there are no overwhelmingly dominant axes. There are no dimensions that explain a lot of variance. The top 10 dimensions explain 48.2 % of the variance, and the top 5 dimensions 36.7 % while top 2 they both explain roughly 11 % each.

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

Now I will plot the explained variance.

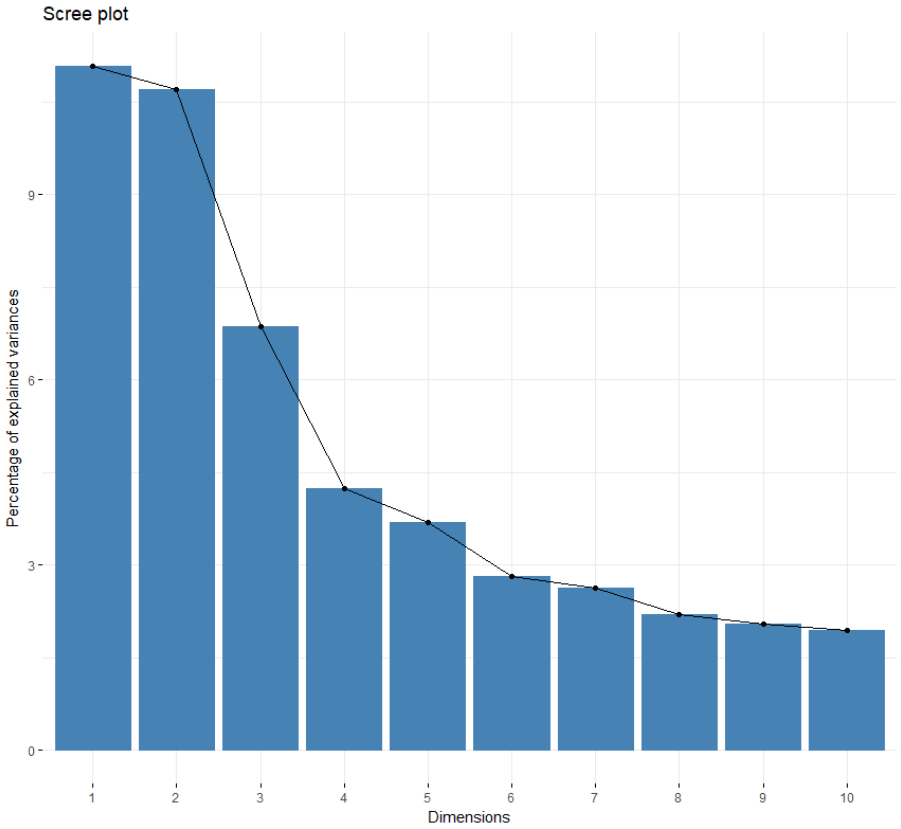

Fig 28

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

The following charts is the squared correlation of coefficient of variables with regards to the first 2 dimensions.

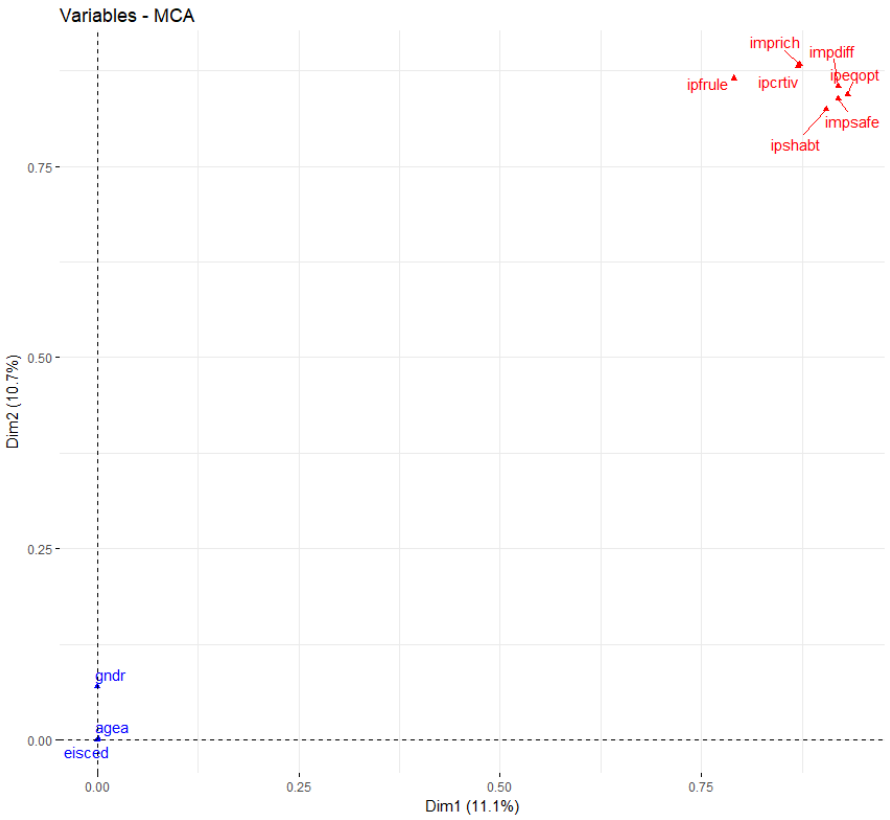

Fig 29

It is possible to see how there is a very low correlation between the supplementary variables in blue and the first two dimensions. While all the variables selected from the first question have a high correlation.

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

The next visualization is the squared cosine. It shows the quality of representation of variable categories on the first 2 axes which are the ones that explain more variance before it starts declining too much.

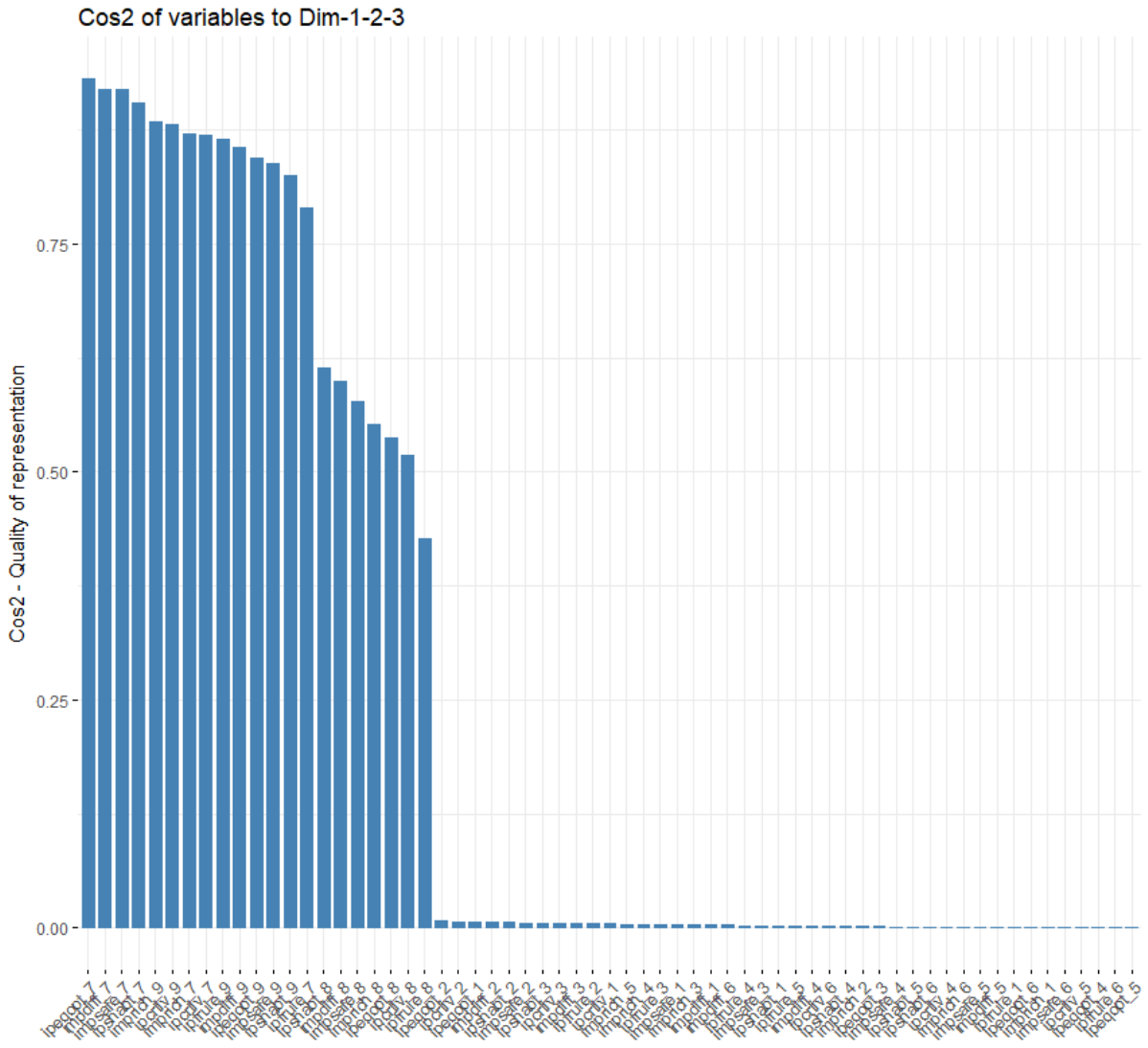

Fig 30

What can be understood from this is that the coefficients for some variables are drastically higher than the rest of the variables.

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

Another way of visualizing it:

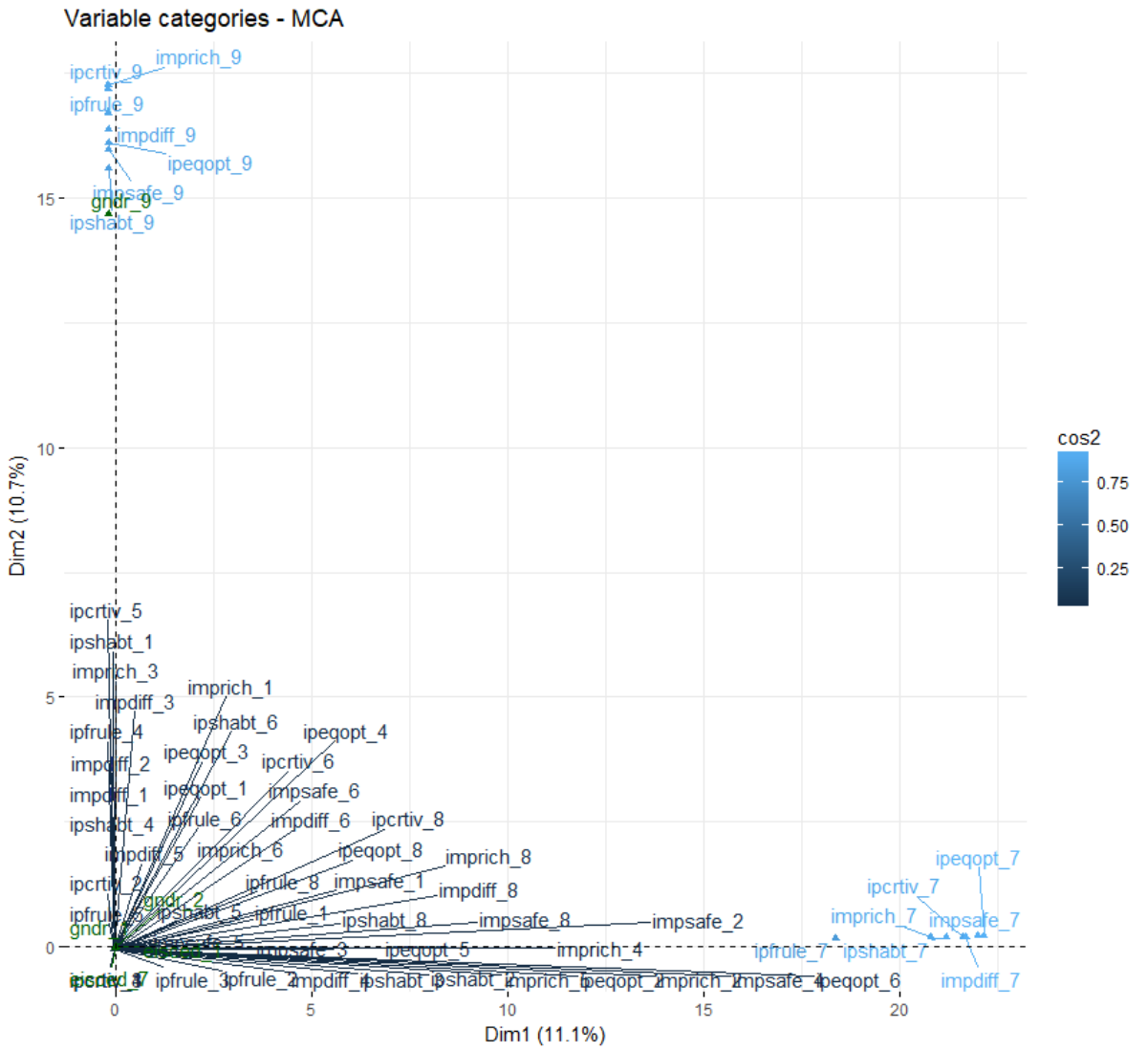

Fig 31

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

Now I will compute the point clouds of individuals.

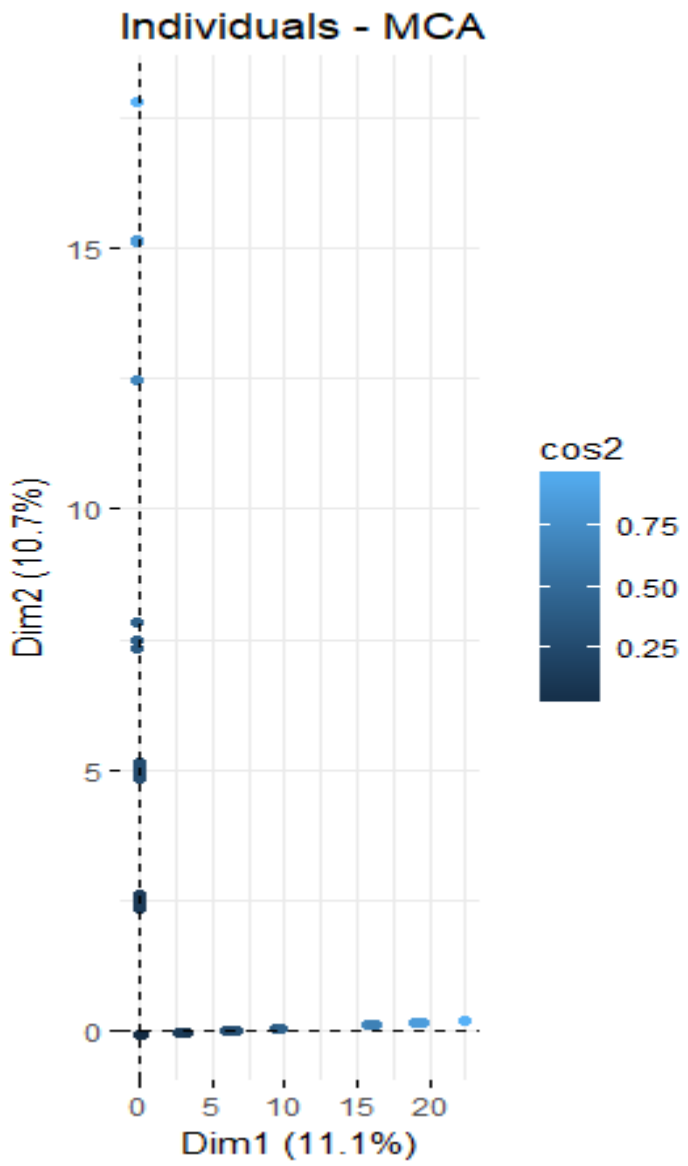

Fig 32

As we can see the points follow very closely the axis, with lower squared cosine values being closer to the origin and higher squared cosine values being further. The points are also fairly spaced between each other, the furthest ones could be considered outliers. Specific groups do not seem evident.

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

I am going to explore the coordinates of the variable categories on dimension 1 and 2

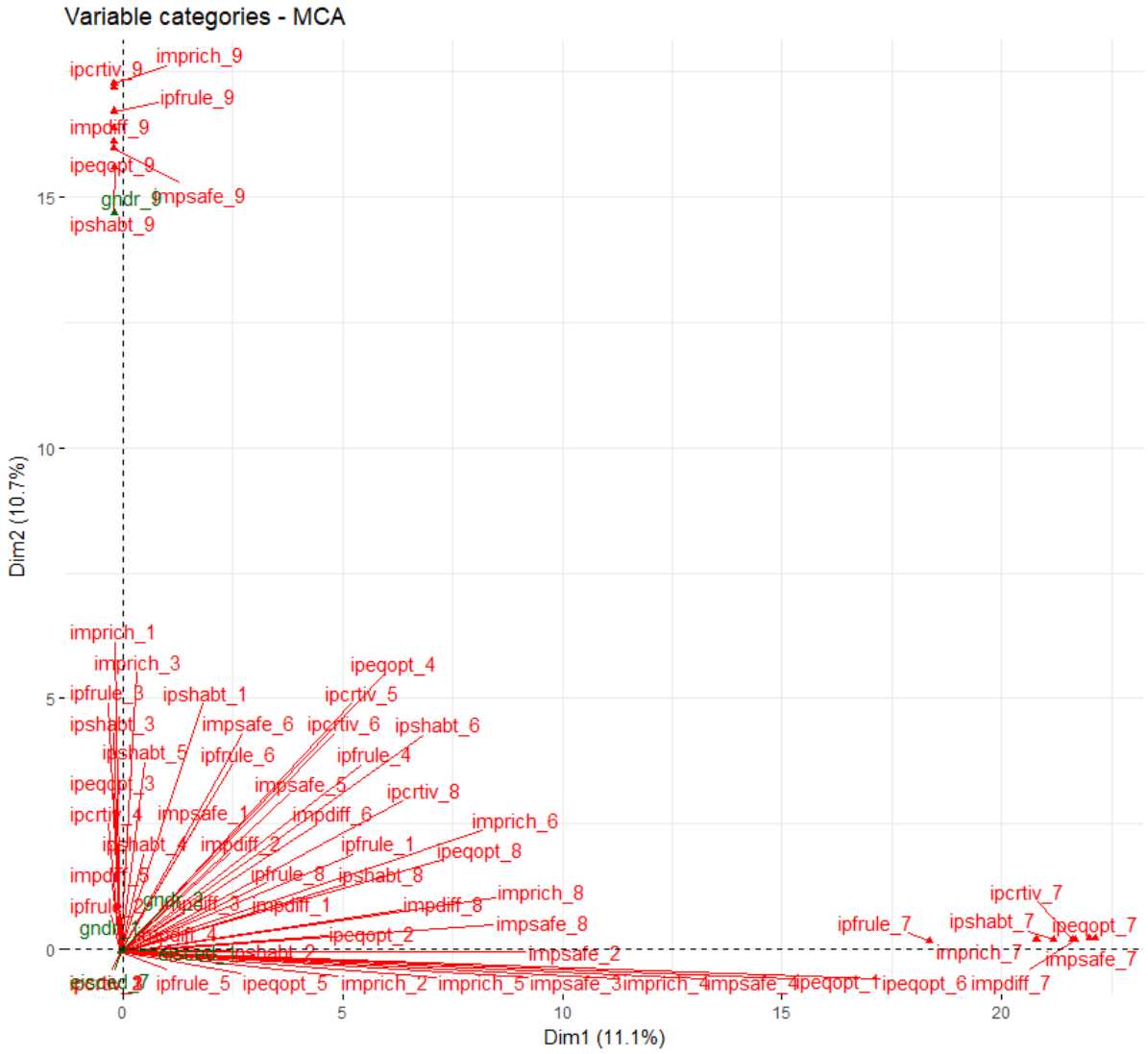

Fig 33

It is possible to see how all the response 9 (no answer) and all the response 7 (refusal) have their own cluster, while all the other responses are squished close to the origin.

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

7.4 Clustering

The next analysis to carry out is the hierarchical clustering of the individuals with 4 clusters.

As the dataset very large for this task, I am randomly slicing the dataset to only contain 10 % of the observations.

therefore, reducing the entries in the dataframe, the new dataframe has the following dimensions:

```
> dim(data.MCA.2)
[1] 5217   10
```

Fig 34

Now computing the 4-clusters hierarchical clustering and its plot:

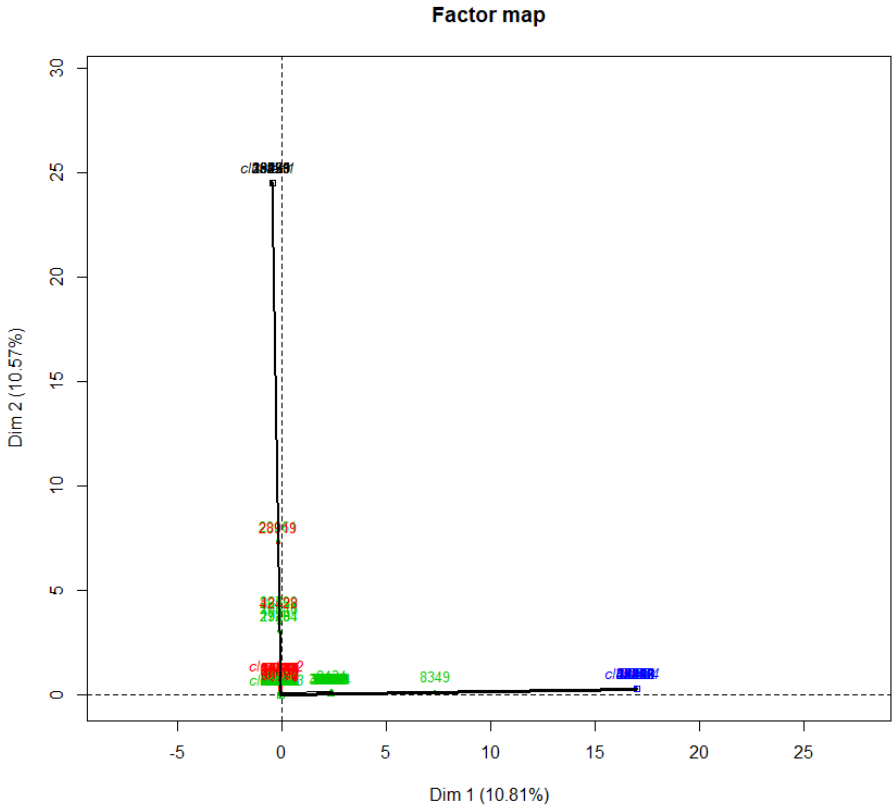

Fig 35

Clusters very sharply separated beside a few outsiders.

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

Trying again with three clusters because the green and red groups do not seem well separated.

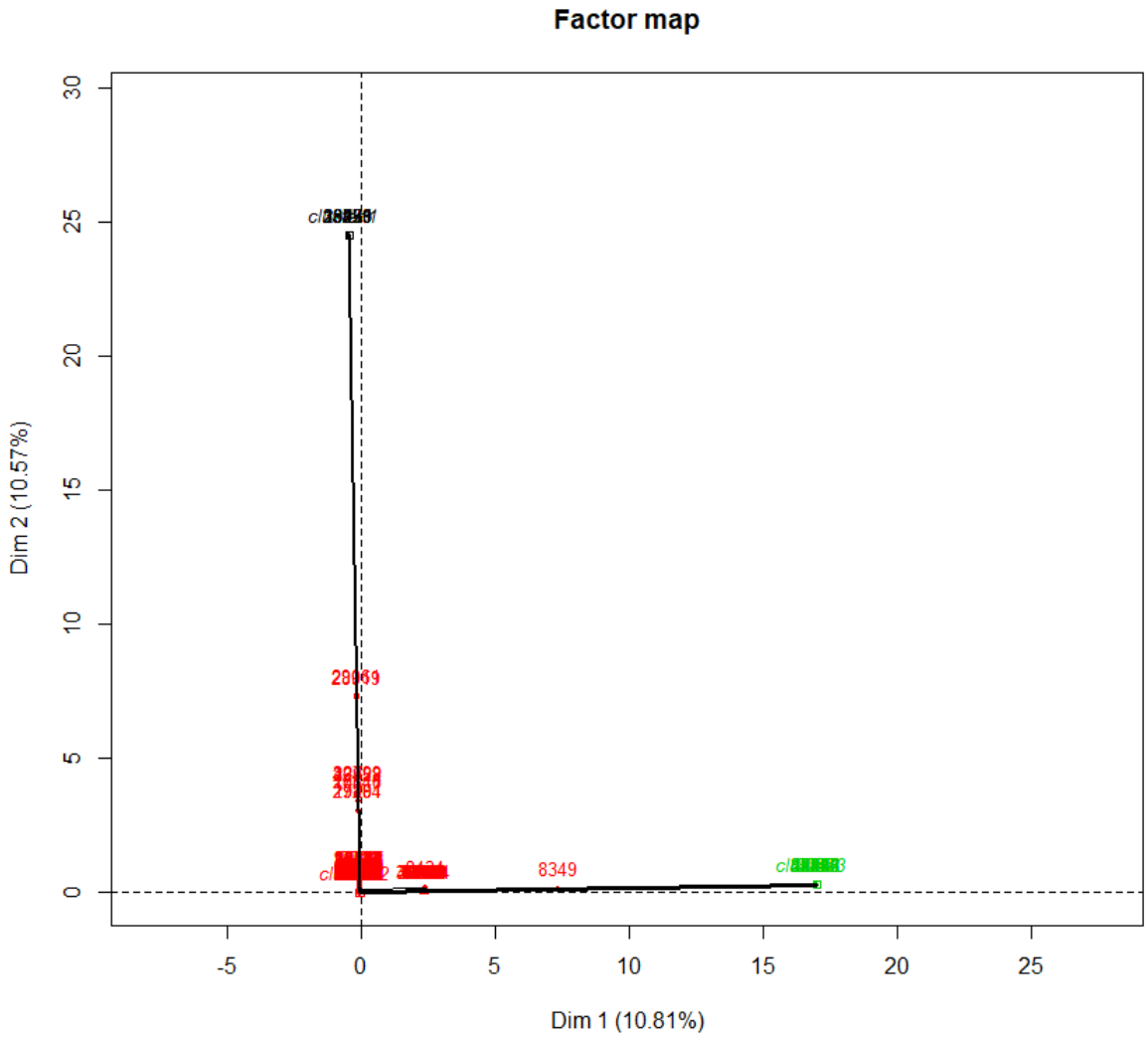

Fig 36

As expected the three clusters are very well defined now and this confirms the clear separations of the group. The red cluster has a few outsiders but green and black are very concentrated.

7.5 Observations

From Fig 33 it is possible to conclude that people that didn't want to respond to some questions also didn't want to respond to other questions; which suggests that response rate is more about the person and personality rather than the question itself. We see two clusters, 'no answer' and 'refusal'. More qualitative research is necessary to investigate the reasons behind this trend. This observation is reflected in other plots, especially Fig 31. Moreover, the clear separation of the data is also discernible from the two clustering visualizations.

Another observation could be the fact that 'no answer' in the contextual variable gender is connected with no answer with the other questions. There may be some social science insight in this trend, which could be investigated with more quantitative research based on in-depth qualitative interviews or focus groups.

One more possible reason for the non-response is the Hawthorne effect: since we are dealing with obtrusive data (the researcher had contact with the respondent) the behavior of the respondent could be altered. Maybe some of the question alienated the subject into not responding multiple questions.

A limitation of the sampling method of this research could be the fact that repeated cross-sectional sampling measures different respondents at different time frames, maybe a longitudinal study with the same people could be examined as well to see if the trends are the same.

Maybe directly asking questions is not the best way to gather insightful data. The new technological and constantly connected era we are living in allows for analyzing web data and the 'digital trace' from social media, Internet of Things and general internet usage; giving the possibility of new ways for social research to be carried out anonymously and seamlessly.

We could go as far as saying that an in-person survey lacks the objectivity of quantitative data and it lacks the depth of qualitative data.

Nevertheless, the ESS is a massive repeated survey that helps with insights into the perspective of normal European citizens on socio-political aspects of Europe.

8. First hand example 4: Retrieving and analyzing financial data

8.1 Introduction

This task is performed with R version 3.5.0. and the main package used is quantmod. The website from which the data is fetched is https://finance.yahoo.com/sector/technology .

The goal of the task is to programmatically retrieve financial data from the web and quantitatively analyze S&P500 companies in terms of volatility and growth but also in terms of correlation.

The procedure will include the financial terminology description required to understand the task.

I will first focus on a single stock, with its retrieval, analysis and visualization then on a multitude of stocks for a complete analysis. The full code is available in the appendices as "example4code.R"

8.2 Analysing NVIDIA stock prices

Using the ticker symbol for the company chosen it is possible to quickly retrieve a table with the main trading days' measures from 03/01/2007 until the day this command is issued. A ticker symbol (or stock symbol) is a unique abbreviation given to identify a stock on a particular stock market. For this example, I will examine NVIDIA, the graphic cards manufacturer, its ticker symbol is NVDA.

The result is a OHLC table (Fig 37), which stands for open-high-low-close with the date of the trading day. These measures, display the price movements of the stock and are used to calculate a variety of technical indicators. Indicators are mathematical calculations that are based on several stock measures such as price and volume. Some indicators are MCDS, moving average, RSI, Bollinger Bands and stochastics; these are used by traders to inform their investment decisions.

Date: 17/09/2018

row.names	NVDA.Open	NVDA.High	NVDA.Low	NVDA.Close
2007-01-03	24.71333	25.01333	23.19333	24.05333
2007-01-04	23.96667	24.05333	23.35333	23.94
2007-01-05	23.37333	23.46667	22.28	22.44
2007-01-08	22.52	23.04	22.13333	22.60667
2007-01-09	22.64	22.79333	22.14	22.16667
2007-01-10	21.93333	23.46667	21.6	23.26
2007-01-11	23.26	23.44	22.79333	23.17333
2007-01-12	22.82667	23.58	22.72	23.48667
2007-01-16	23.66667	23.68	23.26	23.52667
2007-01-17	23.2	23.34	22.94667	23.03333

Fig 37

With the table it is possible to check which day had the highest point of price:

```
  Date       NVDA.Open NVDA.High NVDA.Low NVDA.Close NVDA.Volume

2018-09-04     280.15    285.22       279       283.7     9793000
           NVDA.Adjusted
                   283.7
```

Fig 38

The highest point of price at this moment is September 4, 2018.

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

Now I will chart the time series for the high point of price and add several indicators as add-ons.

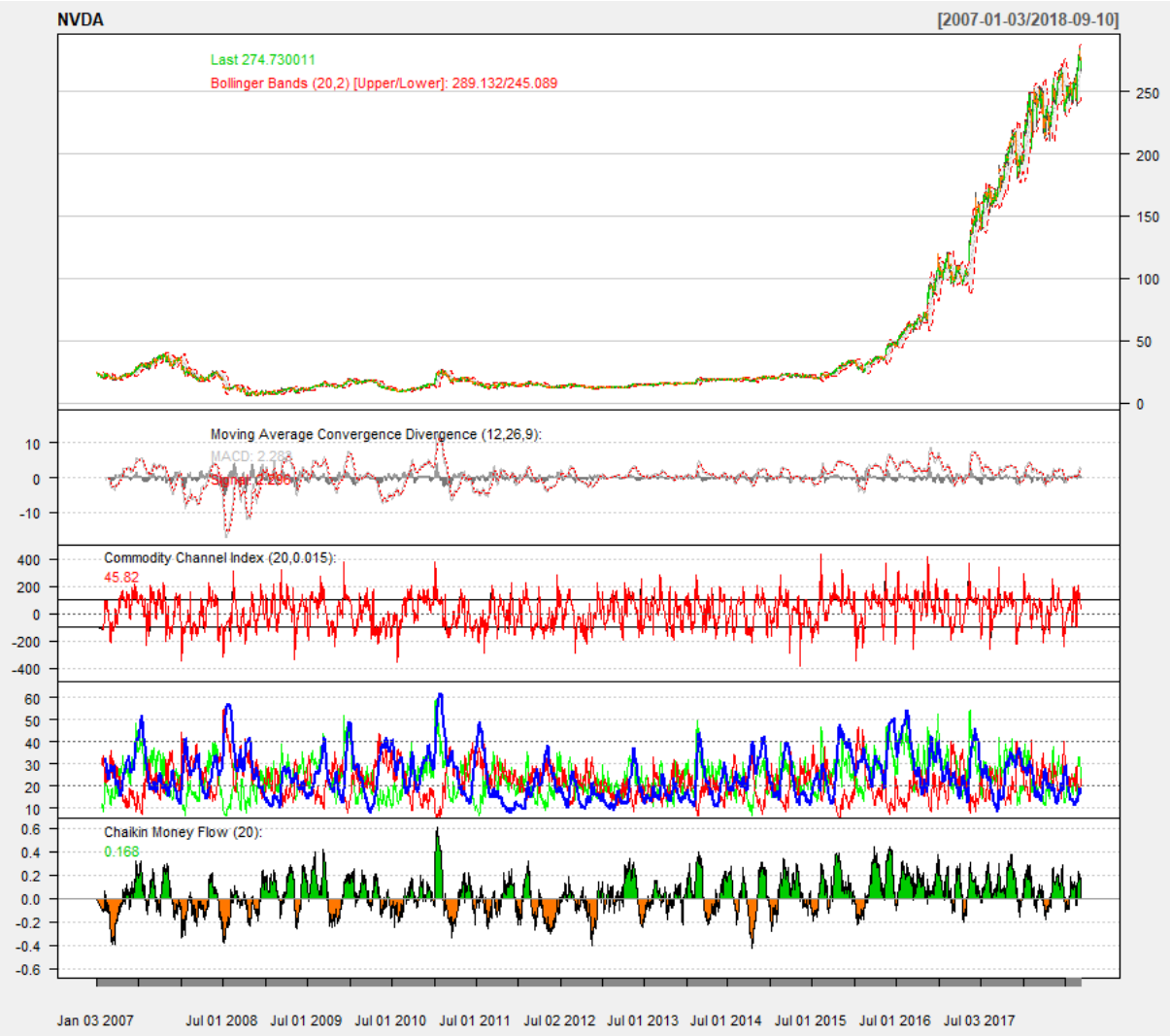

Fig 39

In Fig 39 from top to bottom, the first box is a time series plotting the high point of price from 03/01/2007 to 10/09/2018. The code will update itself to new trading days automatically depending on when the code is run. In the first box the indicator Bollinger Bands in red represents two standard deviations away from the moving average positively and negatively, the more contracted and close to each other the bands are, the less volatile is the stock. For the past two years the Bollinger bands are more far apart (Fig 39).

In the second box there is the MACD (Moving Average Convergence Divergence), this indicator shows the relationship between two moving averages of prices. This measure is calculated by subtracting the 26 days EMA (exponential moving average) from the 12 day

EMA. The grey line is the signal line (9 day EMA), when the MACD goes below this line it means that it is time to sell while if it goes above, it is time to buy.

The third box contains the CCI (Commodity Channel Index), an indicator that communicates if a stock is overbought (going over the 100 mark) or oversold (going below the -100 mark).

The fourth box represents the Directional Movement Indicator (DMI) which assesses price direction and strength without relying on price itself which is more susceptible to volatility. It discerns strong and weak trends so that the trader can enter during a period of trend.

The fifth and last box shows the Chaikin Money Flow, it measures the amount of money flow volume over a period. The crosses over the zero line represent changes on money flow, a positive value means that there is buying pressure while a negative value means there is selling pressure.

The next chart will show the time series of NVIDIA`s closing price:

Fig 40

Over the years, both closing and high price skyrocketed starting from 2016.

Date: 17/09/2018

Fig 41 displays a histogram of the frequency of closing prices.

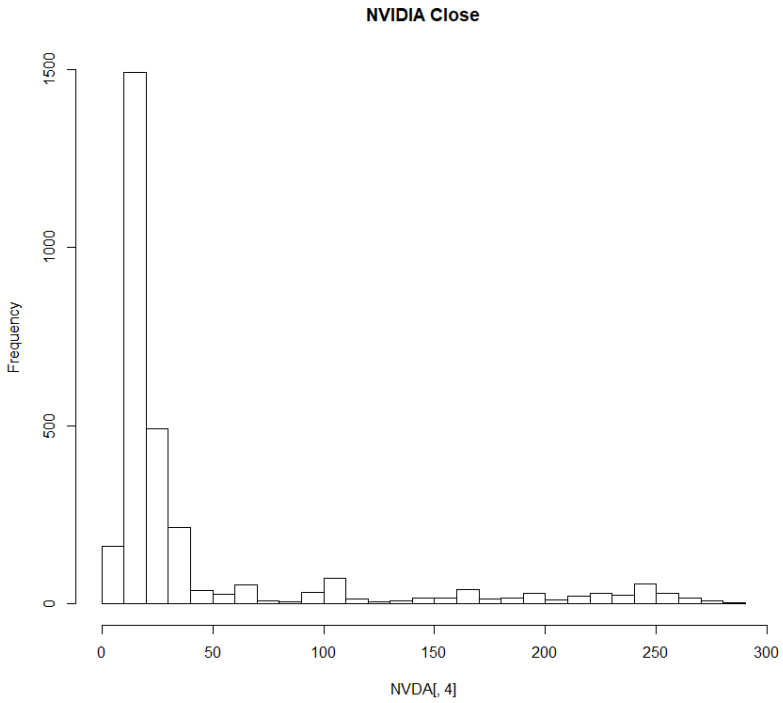

Fig 41

Now I will analyze the financial returns, which are the money earned or lost on an investment; it is represented by the change in monetary value of an investment over time. Stocks returns are approximately normally distributed and uncorrelated, therefore, based on the stock`s previous returns it is possible to model the behavior of stock prices within a confidence interval.

Date: 17/09/2018

Once applied the logarithm (log-returns are used so that it reduces the variation of the time series) it is possible to see that the daily returns are roughly normally distributed:

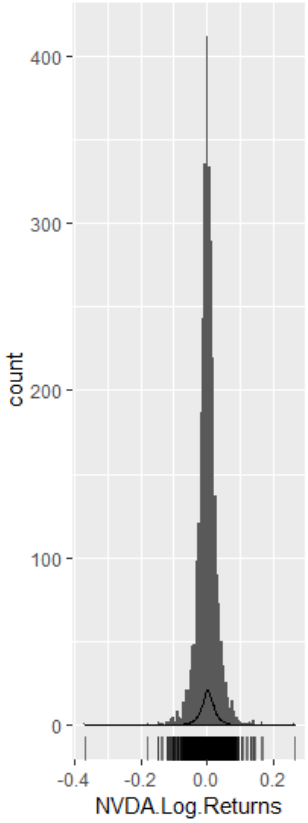

Fig 42

Now it is possible to observe the distribution of the log returns:

```
        0.5%         2.5%          25%          50%          75%
-0.103073831 -0.059936309 -0.012823797  0.001059895  0.014972548

        97.5%        99.5%
 0.058746324  0.092884093
```

Fig 43

This means that the 95% confidence interval is between -0.059936309 and 0.058746324.

While the mean and standard deviation of the daily log returns are:

```
> mean_log_returns
[1] 0.0008538634
> sd_log_returns
[1] 0.02975096
```

Fig 44

These are logarithmic measures, the actual measure for the mean is:

```
> mean_log_returns
[1] 1.000854
```

Fig 45

This means that the average daily return is 0.0854% larger than the price of the previous day, which is good since this is calculated daily and is therefore exponential.

With the measures calculated (mean log returns and returns standard deviation) it is now possible to compare stocks and make an initial screen of different stocks because the mean represents the average growth which is the potential reward of an investment, while the standard deviation represents the volatility which is the risk of an investment.

With these two measures it is also possible to apply a random walk, a process that simulates the prices for the hypothetical next trading days.

This random walk will be over 500 trading days, considering that a year is 252 trading days, this random walk will predict the next 2 years for the NVIDIA stock prices (Fig 46).

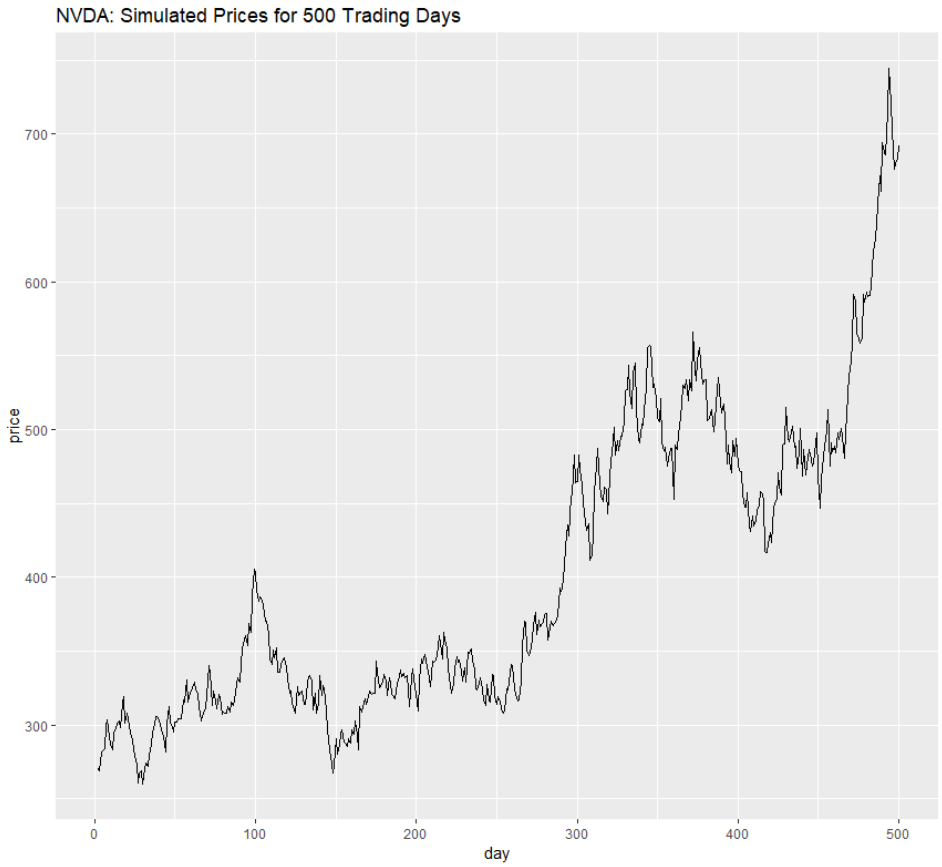

Fig 46

Date: 17/09/2018

The random walk simulation is not very trustworthy because confidence intervals are found by simulating many iterations of this. Therefore, I am going to perform a Monte Carlo Simulation which repeatedly performs random walk simulations a number of times.

This Monte Carlo simulation is computed for one year of trading (252 days) and the random walk is repeated 300 times (Fig 47).

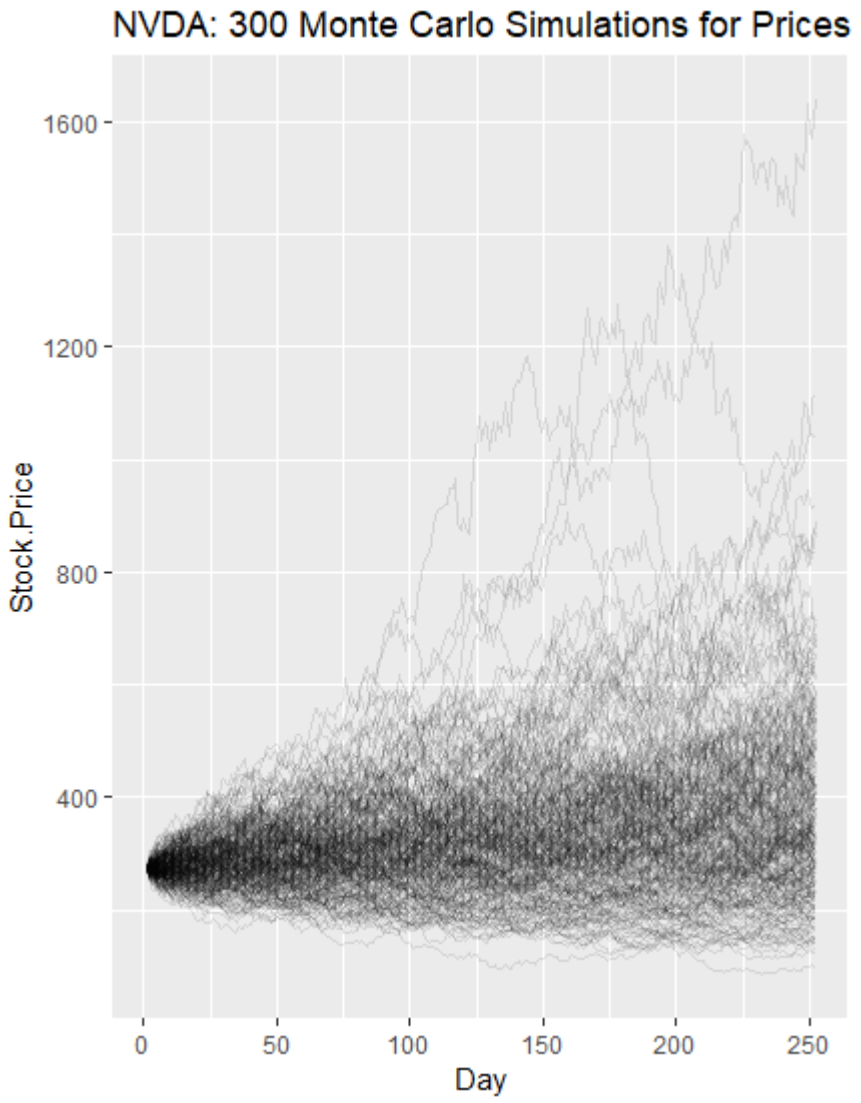

Fig 47

Every line represents a random walk.

The power of this process is to extract the confidence intervals for the stock price.

```
   0.5%    2.5%     25%     50%     75%   97.5%   99.5%
 128.09  146.91  258.52  337.74  457.43  779.45 1074.68
```

Fig 48

The 95% confidence interval is between 146.91 $ and 779.45 $, the median is the most likely outcome, and it is 337.74 $.

One of the metrics to understand if this simulation is trustworthy is checking if the simulated growth is close to the historical growth, intended as CARG (Compound Annual Growth Rate). The CARG represents the mean annual growth rate of a stock (or an investment) over a specified period which can be one year or more. It represents the rate at which a stock would have grown if it grew steadily, which never happens in reality; therefore, it is a representational figure only. However, this stock has been increasing its price over the last two years very steeply; thus making this figure less valid.

Fig 49 shows the historical and simulated CARG.

```
> CARG.H # historical CARG
[1] 0.2400771
> CARG.sim # simulated
[1] 11.34091
```

Fig 49

As anticipated the figures are rather far apart, which may be due to the rapid recent growth.

This task was performed using the adjusted closing price rate, which is necessary when analyzing historical returns as it takes into account corporate actions that took place before the next trading day.

8.3 S&P500 analysis

S&P500 stands for Standard and Poor`s 500, it is an index for the American stock market based on the market capitalizations of 500 companies, all very large, that have common stocks listed on the two largest stock exchange in the world, the NYSE (New York Stock Exchange a.k.a. Wall Street) and the NASDAQ. Market capitalization is the share price multiplied by the number of shares outstanding, which are all the shares owned by stockholders, company officials and public investors; however, it does not include the shares repurchased by a company.

First I retrieve the 500 companies by webscraping from a table in Wikipedia (https://en.wikipedia.org/wiki/List_of_S%26P_500_companies).

Date: 17/09/2018

The table below (Fig 50) shows the retrieved dataset, with ticker symbols, the security which is the name for the company, the sector of the company and the sub-sector.

ticker.symbol	security	gics.sector	gics.sub.industry
MMM	3M Company	Industrials	Industrial Conglomerates
ABT	Abbott Laboratories	Health Care	Health Care Equipment
ABBV	AbbVie Inc.	Health Care	Pharmaceuticals
ABMD	ABIOMED Inc	Health Care	Health Care Equipment
ACN	Accenture plc	Information Technology	IT Consulting & Other Services
ATVI	Activision Blizzard	Information Technology	Home Entertainment Software
ADBE	Adobe Systems Inc	Information Technology	Application Software
AMD	Advanced Micro Devices Inc	Information Technology	Semiconductors
AAP	Advance Auto Parts	Consumer Discretionary	Automotive Retail
AES	AES Corp	Utilities	Independent Power Producers &

Fig 50 These are only the first 10 elements

This is the breakdown for unique elements in each column, there are 505 companies, 11 main industries and 123 sub-industries. The industries and sub-industries are selected by the GICS (Global Industry Classification Standard).

```
    ticker.symbol     security    gics.sector   gics.sub.industry
               505          505             11                 123
```

Fig 51

Fig 52 shows the distribution of the companies in the different main sectors.

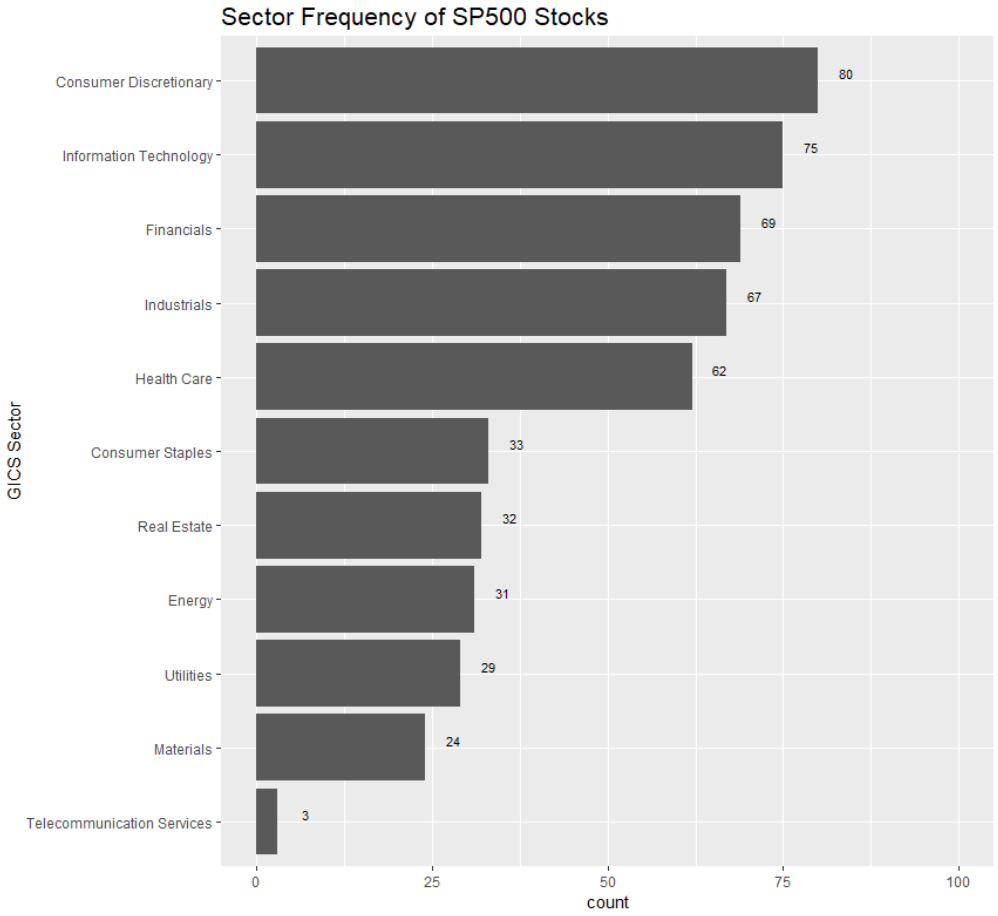

Fig 52

Knowing which company is in which industry is useful in diversification strategies where risk is limited by investing stocks that are not much return correlated, which means that the success of some company is not tied to the success of another company.

The next step is creating two functions, one that takes a ticker and returns the prices of a stock, the second takes stock prices and returns the log returns. More information on these two functions available in the code.

Date: 17/09/2018

Thanks to these two functions it is possible to build a nested dataframe where some cells contain another dataframe. The result is a new dataframe where each row corresponds to a company and where two columns are the stock prices and the log returns for the entire history of that company until the day the code is run, these columns are SP and lg.return, each cell from those columns is a dataframe (Fig 53). It is important to note that 10 companies were removed from the analysis because they were returning a HTTP 400 error, so the data was impossible to download (namely: BRK.B, BF.B, APTV, BKNG, BHF, DXC, JEF, DGX, UA, PPL) .

	ticker.symbol	security	gics.sector	gics.sub.industry	SP	lg.return
	<chr>	<chr>	<chr>	<chr>	<list>	<list>
1	MMM	3M Company	Industrials	Industrial Conglomer~	<tibbl~	<tibble [~
2	ABT	Abbott Lab~	Health Care	Health Care Equipment	<tibbl~	<tibble [~
3	ABBV	AbbVie Inc.	Health Care	Pharmaceuticals	<tibbl~	<tibble [~
4	ABMD	ABIOMED Inc	Health Care	Health Care Equipment	<tibbl~	<tibble [~
5	ACN	Accenture ~	Information ~	IT Consulting & Othe~	<tibbl~	<tibble [~
6	ATVI	Activision~	Information ~	Home Entertainment S~	<tibbl~	<tibble [~
7	ADBE	Adobe Syst~	Information ~	Application Software	<tibbl~	<tibble [~
8	AMD	Advanced M~	Information ~	Semiconductors	<tibbl~	<tibble [~
9	AAP	Advance Au~	Consumer Dis~	Automotive Retail	<tibbl~	<tibble [~
10	AES	AES Corp	Utilities	Independent Power Pr~	<tibbl~	<tibble [~

Fig 53 These are only the first 10 elements

For example, if we access the stock prices for the first element of the dataframe, the company MMM the result will be a dataframe with this company`s prices history (Fig 54).

	Date	Open	High	Low	Close	Volume	Adjusted
	<date>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	2007-01-03	77.5	78.8	77.4	78.3	3781500	57.8
2	2007-01-04	78.4	78.4	77.4	77.9	2968400	57.6
3	2007-01-05	77.9	77.9	77.0	77.4	2765200	57.2
4	2007-01-08	77.4	78.0	77.0	77.6	2434500	57.3
5	2007-01-09	78	78.2	77.4	77.7	1896800	57.4
6	2007-01-10	77.3	78.0	77.0	77.8	1787500	57.5
7	2007-01-11	78.1	79.0	77.9	78.7	2372500	58.1
8	2007-01-12	78.4	79.5	78.2	79.4	2582200	58.6
9	2007-01-16	79.5	79.6	78.9	79.6	2526600	58.8
10	2007-01-17	79.3	79.5	78.8	78.9	2711300	58.3

Fig 54 These are only the first 10 elements

Thanks to this nested dataframe it is possible to create a scatter plot (Fig 55) with the volatility on the x-axis and the reward on the y-axis, while the number of trade days is given by the size and color. From RStudio this plot is interactive, it can be zoomed in and out, panned, and hovered in order to understand which bubble is which company. This is an excellent feature to check out each stock individually for its risk and reward.

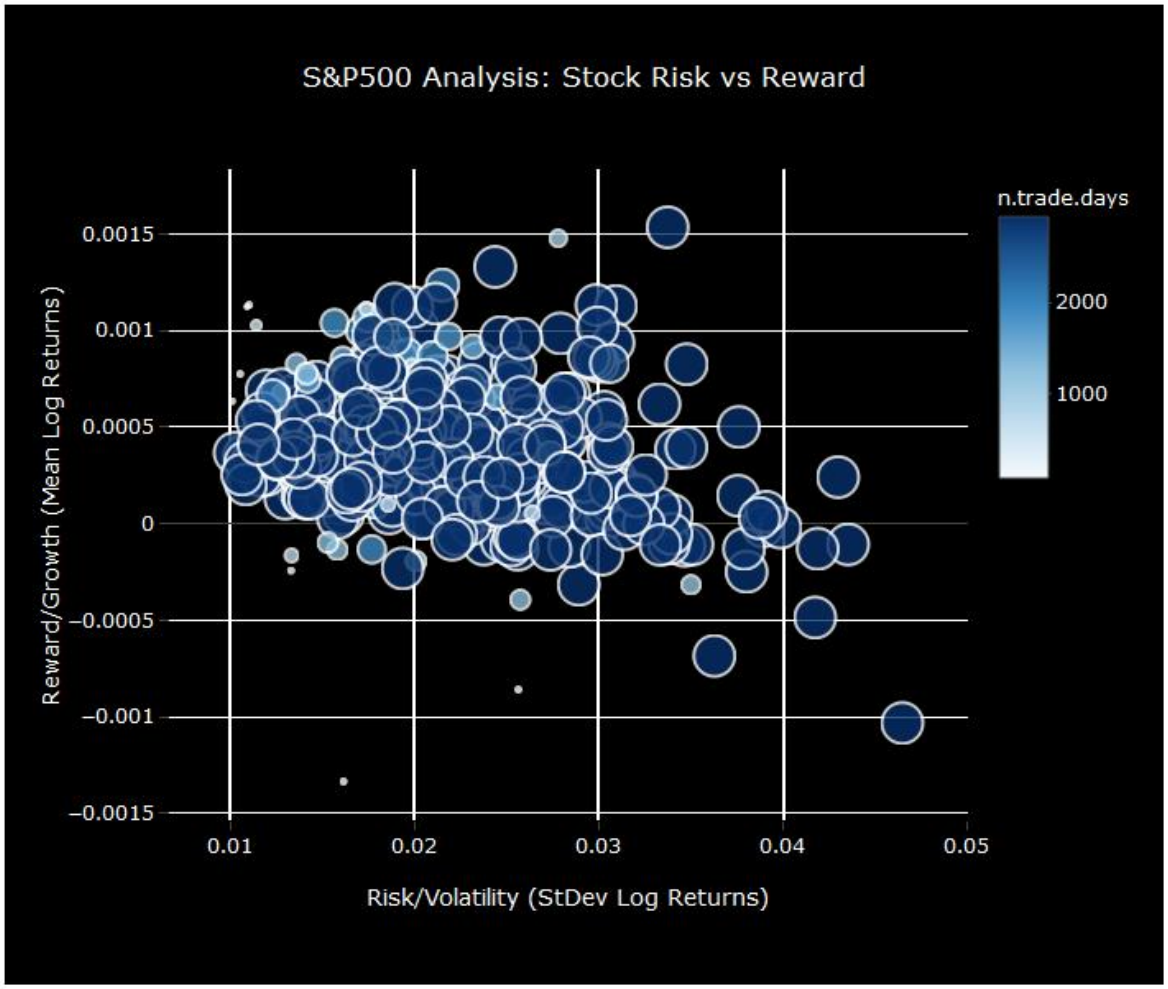

Fig 55

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

From the dataframe it is possible to isolate the top companies for high mean log returns and low volatility (Fig 56). These are the best stocks according to these statistics, it is a way to screen stocks in terms of their potential.

	ticker.symbol	mean.log.returns	sd.log.returns	n.trade.days
	<chr>	<dbl>	<dbl>	<dbl>
1	ANET	0.00148	0.0278	1074
2	AMZN	0.00133	0.0244	2943
3	AVGO	0.00124	0.0215	2290
4	MA	0.00114	0.0212	2943
5	TDG	0.00114	0.0189	2943
6	ALGN	0.00113	0.0299	2943
7	AEE	0.00113	0.0110	166
8	ABMD	0.00113	0.0309	2943
9	EVRG	0.00112	0.0109	75
10	AAPL	0.00112	0.0199	2943
11	PYPL	0.00111	0.0174	803
12	FLT	0.00107	0.0174	1947
13	HII	0.00104	0.0157	1881
14	FTV	0.00103	0.0114	551
15	REGN	0.00102	0.0299	2943
16	AOS	0.00101	0.0202	2943

Fig 56

Let us take the second entry, Amazon, and plot it (Fig 57).

Date: 17/09/2018
Fig 57

These are not the only things to take into account when analyzing stocks and choosing a portfolio, there is for example qualitative analysis and analyzing dividends among other metrics.

When selecting a portfolio, it is not advisable to only select the best performance but also diversify the stocks so that a few companies going down will not affect the investments extremely negatively. In order to do this, it is necessary to do a correlation assessment and choose stocks with low correlation.

First I retrieve a table with the top 30 companies in terms of mean log returns and standard deviation, with companies having over 1000 trading days (Fig 58).

	ticker.symbol	ranking	mean.log.returns	sd.log.returns	lg.return
	<chr>	<int>	<dbl>	<dbl>	<list>
1	ANET	1	0.00148	0.0278	<tibble [1,074 x ~
2	AMZN	2	0.00133	0.0244	<tibble [2,943 x ~
3	AVGO	3	0.00124	0.0215	<tibble [2,290 x ~
4	MA	4	0.00114	0.0212	<tibble [2,943 x ~
5	TDG	5	0.00114	0.0189	<tibble [2,943 x ~
6	ALGN	6	0.00113	0.0299	<tibble [2,943 x ~
7	ABMD	7	0.00113	0.0309	<tibble [2,943 x ~
8	AAPL	8	0.00112	0.0199	<tibble [2,943 x ~
9	FLT	9	0.00107	0.0174	<tibble [1,947 x ~
10	HII	10	0.00104	0.0157	<tibble [1,881 x ~

... with 20 more rows

Fig 58 These are only the first 10 elements

The next step is to obtain a dataframe where the top thirty companies are both the column and the rows and each cell corresponds to the correlation between the company on a given row and a company on a given column (Fig 59).

	AAPL	ABMD	ALGN	AMZN	ANET	AOS	AVGO
AAPL	1.0000000	0.2674052	0.3587554	0.3884646	0.2468660	0.3963468	0.4862424
ABMD	0.2674052	1.0000000	0.3383890	0.2409951	0.1888290	0.2645946	0.2360271
ALGN	0.3587554	0.3383890	1.0000000	0.2922803	0.3037096	0.3861574	0.3364555
AMZN	0.3884646	0.2409951	0.2922803	1.0000000	0.2886500	0.3575059	0.3115264
ANET	0.2468660	0.1888290	0.3037096	0.2886500	1.0000000	0.2746489	0.3156349
AOS	0.3963468	0.2645946	0.3861574	0.3575059	0.2746489	1.0000000	0.4032649
AVGO	0.4862424	0.2360271	0.3364555	0.3115264	0.3156349	0.4032649	1.0000000

Fig 59 These are only the first 7 elements

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

This is the plot (Fig 60) of the correlations. Bluer colors mean higher correlations, the squares collect groups of highly correlated companies, redder colors determine negative correlation but there are no negative correlated companies. On the diagonal the correlation is always perfect because it is correlating a row and a column with the same company.

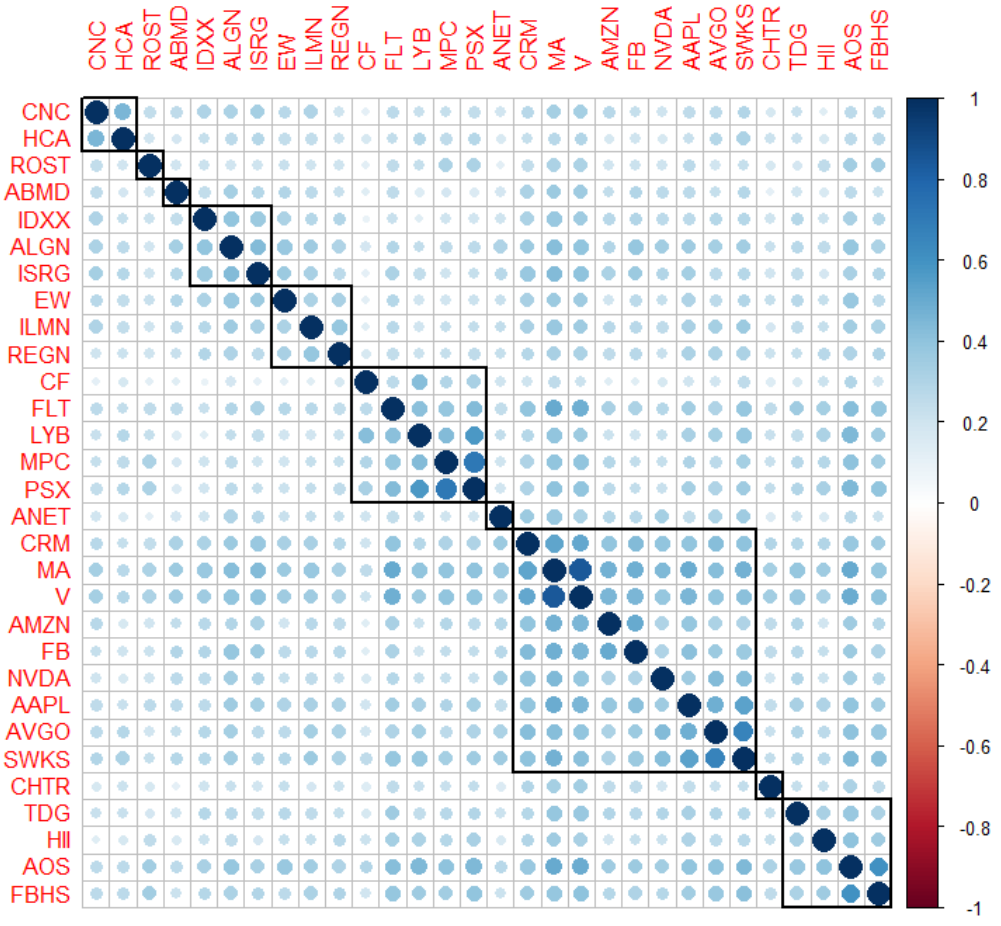

Fig 60

Many correlated companies have a good reason to be so, for example one is the supplier of another, or one is a service necessary in order to use another service, for example, Master Card (MA) is highly correlated with Amazon and Nvidia (NVDA) is correlated to Skyworks Solutions Inc (SWKS) because the latter is a semiconductor manufacturer.

9. First hand example 5: Using Twitter API

9.1 Introduction

Twitter is a great source of data as tweets are public and possible to extract programmatically, as long as a Twitter developer key is granted from the company. This has great applications for business, for example it is possible to have an understanding of how a brand is perceived or a new product is received.

This task will use python 3.5 and the package tweepy which allows to interface with Twitter API.

The goal is to extract a number of tweets with a specific keyword or hashtag, then run the package TextBlob; this library processes textual data and it provides a simple API for natural language processing (NLP). Thanks to TextBlob it will be possible to have an understanding of how many tweets are a positive, negative or neutral opinions.

9.2 Procedure

The first step is authenticating the app (in this case a python script) with the Twitter developer key codes.

Then select which keyword or hashtag and how many tweets to analyze.

At this point all the tweets are retrieved in a variable, selecting only English tweets.

After having obtained all the tweets in a variable, I looped all the tweets and applied TextBlob which gives a score between -1 and 0 (not included) for negative tweets, with -1 being extremely negative; a score of 0 for neutral tweets, and a score between 0 (not included) and 1 for positive tweets, with 1 being extremely positive.

The positive, negative and neutral tweets are then transformed into the percentage of the total number of tweets.

Finally, a pie chart is built with the ratio of the positive, negative and neutral tweets for the keyword and number of tweets selected.

Fig 61 displays 1000 tweets for #Cristiano (football player Cristiano Ronaldo)

Fig 62 displays 1000 tweets for #EA (videogame publisher Electronic Arts, famous for implementing questionable practices that split the community`s opinion i.e. heavy use of in-game micro-transaction)

Fig 63 displays 1000 tweets for #nvidia (videocard manufacturer used in example 4)

Fig 64 displays 1000 tweets for #LouisCK (famous comedian at the center of a controversy who has split the audience)

Full code available in the appendices as “example5code.ipynb”

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

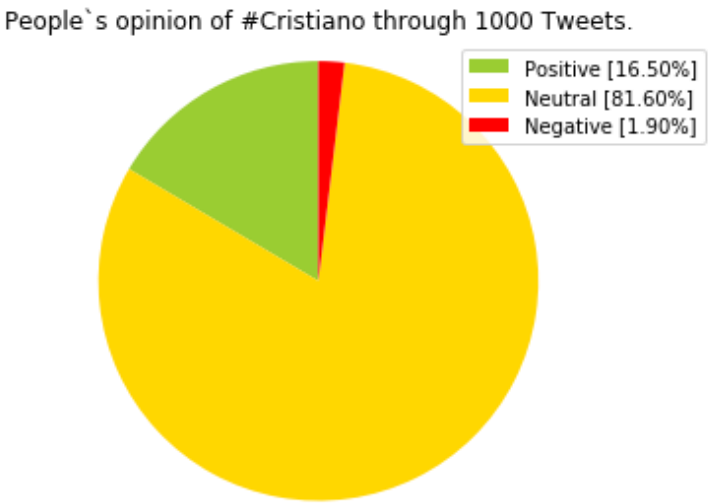

Fig 61

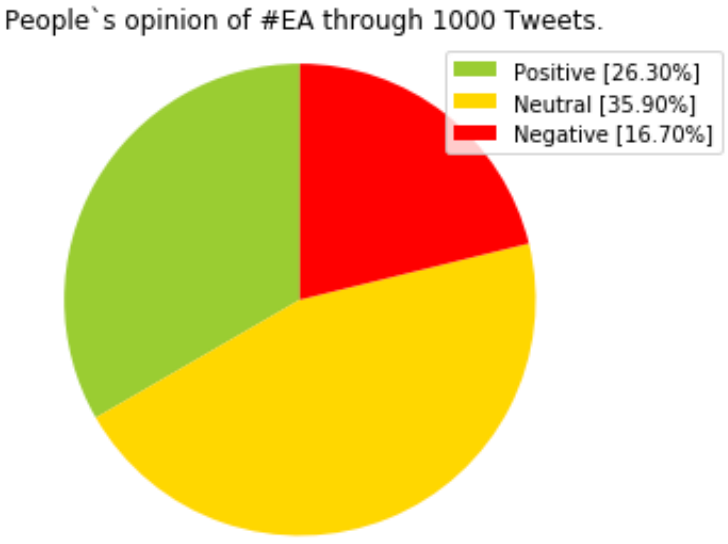

Fig 62

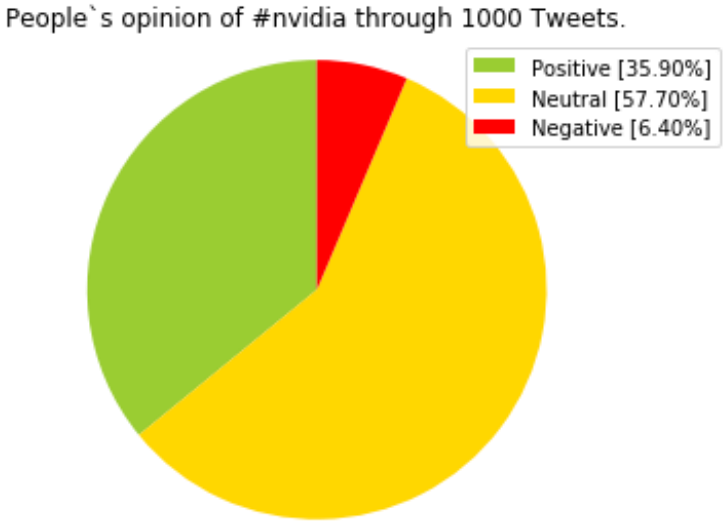
People`s opinion of #nvidia through 1000 Tweets.
Positive [35.90%]
Neutral [57.70%]
Negative [6.40%]

Fig 63

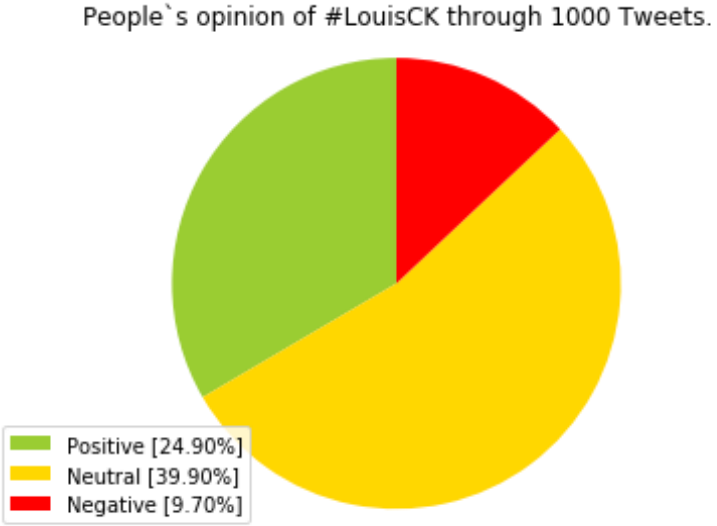
People`s opinion of #LouisCK through 1000 Tweets.
Positive [24.90%]
Neutral [39.90%]
Negative [9.70%]

Fig 64

10. First hand example 6: Loan approval prediction

10.1 Introduction

This task`s goal is to predict whether a customer will be able to pay a loan; therefore, on the basis of the algorithm a customer can get their loan approved or denied.

The data comes from the company Dream Housing Finance (provided in the appendices as "example6data.csv").

The task is carried out with R 3.5.0 and the main package used is "caret", the code is available in the appendices as "example6code.R.

The dataset consists of 614 observations and 13 variables. The variable description is as follows:

- ApplicantIncome = Applicant income
- CoapplicantIncome = Coapplicant income
- Credit_History = credit history meets guidelines
- Dependents = Number of dependents
- Education = Graduate/ Under Graduate
- Gender = Male/ Female
- Loan_Amount_Term = Term of loan in months
- Loan_ID = Unique Loan ID
- Loan_Status = Y/N
- LoanAmount = In thousands
- Married = Y/N
- Property_Area = Urban/ Semi Urban/ Rural
- Self_Employed = Y/N

This task will use four different machine learning algorithms and then compare them with the aid of the AUC (area under the curve) of the ROC curve (Receiver operating characteristic).

10.2 Analysis

The data is loaded into R, then it is checked for missing values. Since there are NAs in the dataframe a KNN (K-Nearest Neighbor) is applied and thus filling all the cells with missing values.

The response variable is Loan_Status, which determines if a customer (row) has their loan approved or denied.

The next step is to convert all the categorical variables to dummy numerical variables so to apply algorithms on the dataset.

Then the data is split into train and test; the test is 25% of the data and contains the same ratio of positive and negative values in the response variable compared to the remaining 75% training test.

Recursive Feature with crossvalidation is applied in order to perform feature selection. From this step only the top 6 variables are considered and used for analysis; the variables in question are Credit_History, LoanAmount, Loan_Amount_Term, ApplicantIncome, CoapplicantIncome, Property_Area.Semiurban.

At this point, four different models are applied to the training set, a random forest, a neural network, a linear regression and a GBM (Gradient Boosting Machine).

The algorithms are then parameter optimized with the use of tunegrid (for GBM) and tunelength (for the neural network and random forest) with the inclusion of 5-fold crossvalidation. These functions will try many combinations of parameters so that the algorithms are as tuned to the task as possible.

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

Fig 65 shows the accuracy of the GBM per number of iterations, every box and the color of the line, represents a different combination of parameters. The best combination in terms of accuracy was n.trees = 500, interaction.depth = 1, shrinkage = 0.01 and n.minobsinnode = 10.

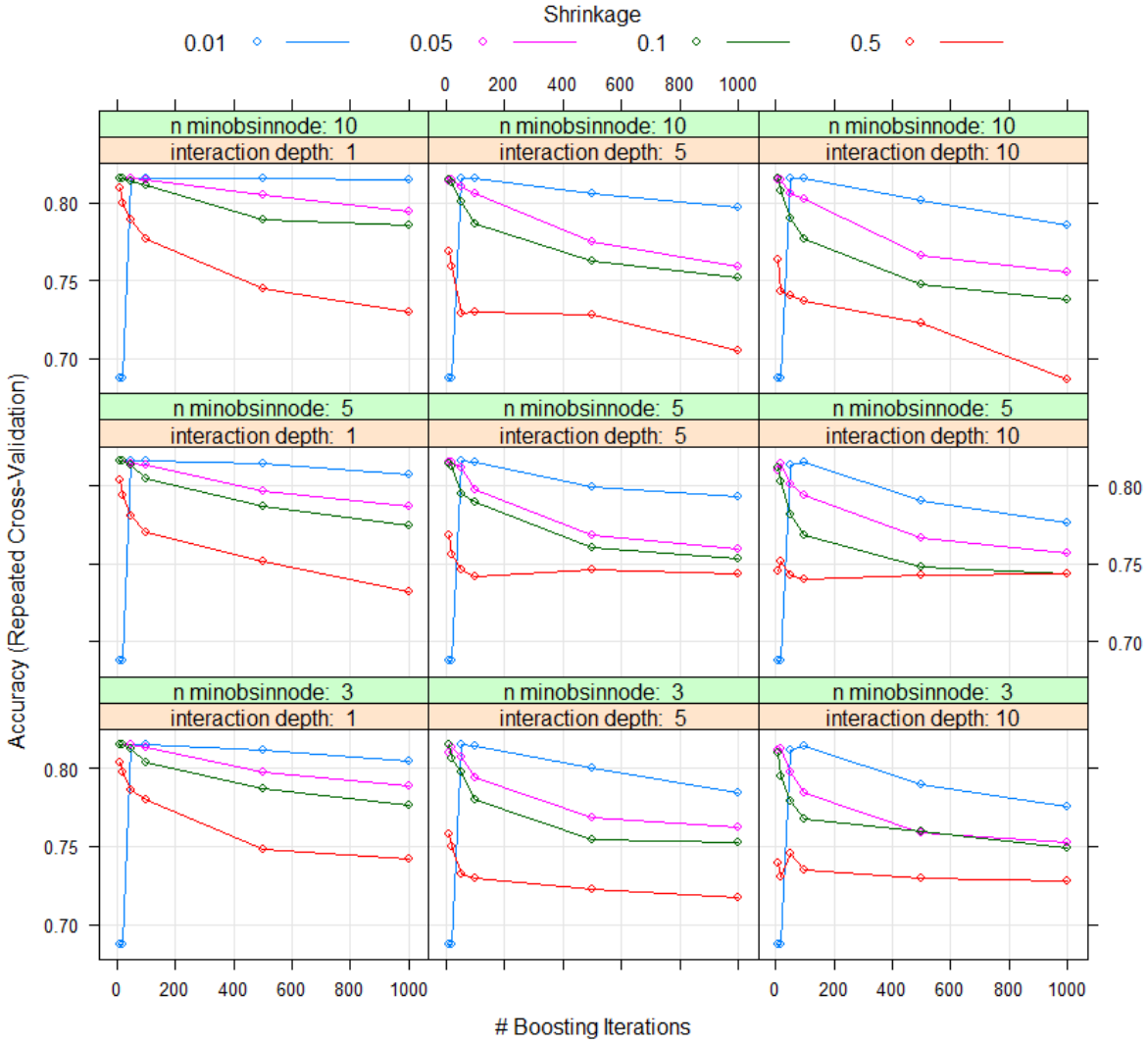

Fig 65

Date: 17/09/2018

For the remaining algorithms only a few parameters were optimized to make the process faster.

Fig 66 shows the best random forest, with mtry = 2.

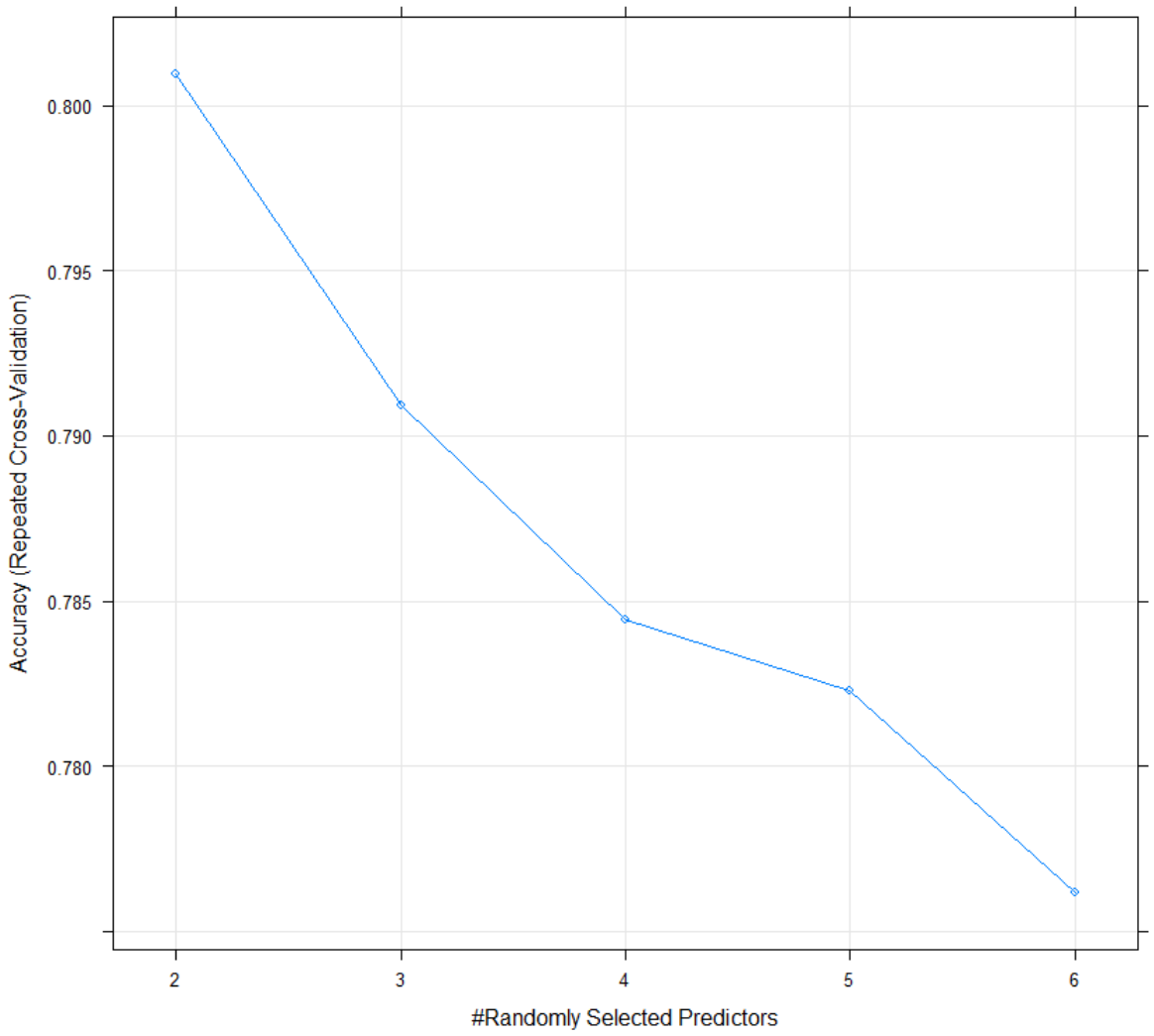

Fig 66

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

Fig 67 shows the accuracy of the neural network with different parameters. The final values used for the model were size = 1 and decay = 0.1.

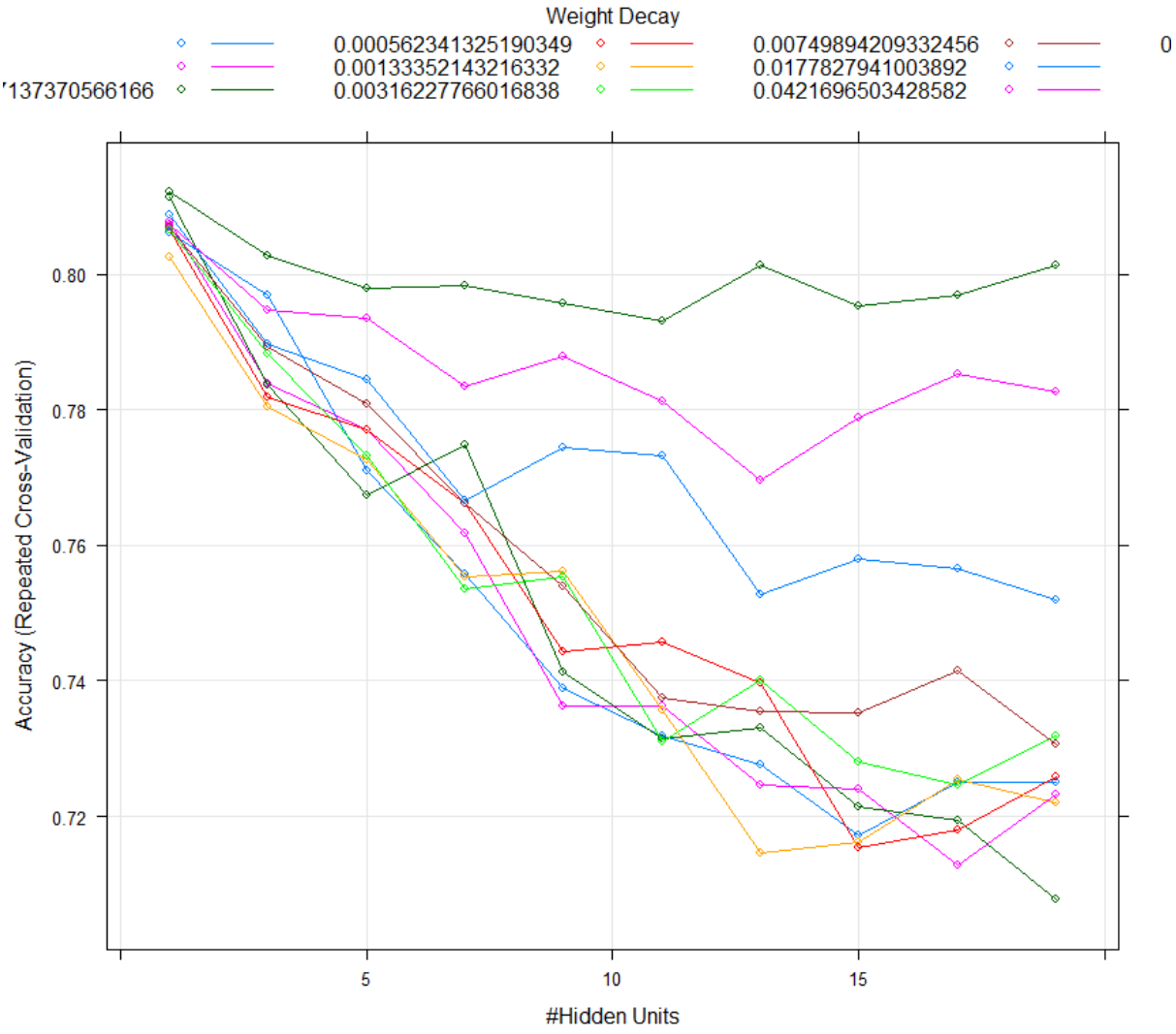

Fig 67

The most important variable for all the models is by far Credit_History, beside the neural network that finds both Credit_History and Property_area.Semiurban equally important.

Date: 17/09/2018

Below (Fig 68) is the breakdown of the variable importance for each model. They are in order from top to bottom, GBM, neural network, random forest and linear regression.

```
> summary(gbm)
                          rel.inf
Credit_History          75.2707146
LoanAmount               8.5572604
ApplicantIncome          7.5488203
Property_Area.Semiurban  4.4321237
CoapplicantIncome        3.6079885
Loan_Amount_Term         0.5830924


> varImp(object=nnet)

                         Overall
Property_Area.Semiurban  100.00
Credit_History            98.04
Loan_Amount_Term          39.63
CoapplicantIncome         13.84
ApplicantIncome            7.24
LoanAmount                 0.00


> varImp(object=rf)

                         Overall
Credit_History           100.00
ApplicantIncome           57.60
LoanAmount                52.77
CoapplicantIncome         35.30
Loan_Amount_Term           8.87
Property_Area.Semiurban    0.00


> varImp(object=glm)

                         Overall
Credit_History           100.000
Property_Area.Semiurban   33.729
CoapplicantIncome         12.683
LoanAmount                10.193
Loan_Amount_Term           9.918
ApplicantIncome            0.000
```

Fig 68

Gastone Riccardo Tolli
ID: 33506151

Date: 17/09/2018

Fig 69 and 70 shows the confusion matrix and the ROC curve (AUC = 0.678) for the GBM model.

```
> confusionMatrix(predictions.gbm,validation[,y.variable])
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  18   2
         1  30 103

               Accuracy : 0.7908
                 95% CI : (0.7178, 0.8523)
    No Information Rate : 0.6863
    P-Value [Acc > NIR] : 0.002665

                  Kappa : 0.4229
 Mcnemar's Test P-Value : 1.815e-06

            Sensitivity : 0.3750
            Specificity : 0.9810
         Pos Pred Value : 0.9000
         Neg Pred Value : 0.7744
             Prevalence : 0.3137
         Detection Rate : 0.1176
   Detection Prevalence : 0.1307
      Balanced Accuracy : 0.6780
```

Fig 69

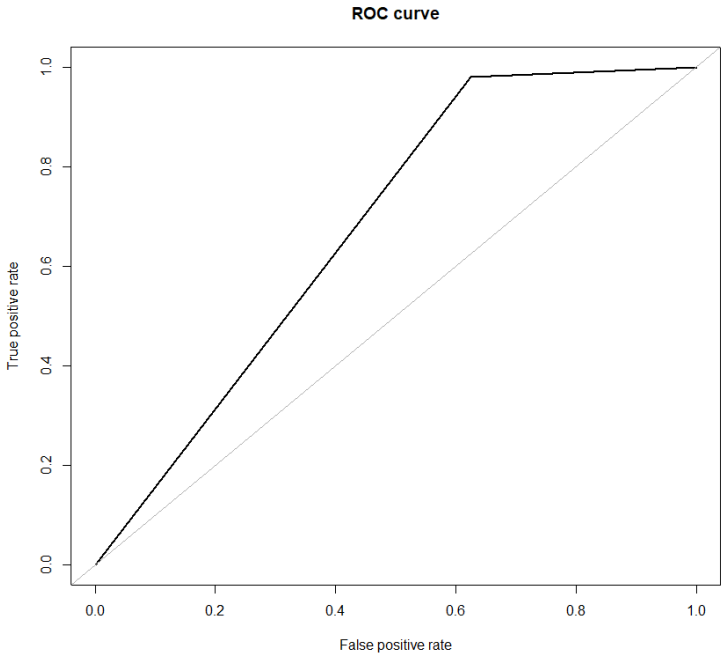

Fig 70

While Fig 71 and 72 show the confusion matrix and the ROC curve for the neural network (AUC = 0.673).

```
> confusionMatrix(predictions.nnet,validation[,y.variable])
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  18   3
         1  30 102

               Accuracy : 0.7843
                 95% CI : (0.7106, 0.8466)
    No Information Rate : 0.6863
    P-Value [Acc > NIR] : 0.00468

                  Kappa : 0.4089
 Mcnemar's Test P-Value : 6.011e-06

            Sensitivity : 0.3750
            Specificity : 0.9714
         Pos Pred Value : 0.8571
         Neg Pred Value : 0.7727
             Prevalence : 0.3137
         Detection Rate : 0.1176
   Detection Prevalence : 0.1373
      Balanced Accuracy : 0.6732
```

Fig 71

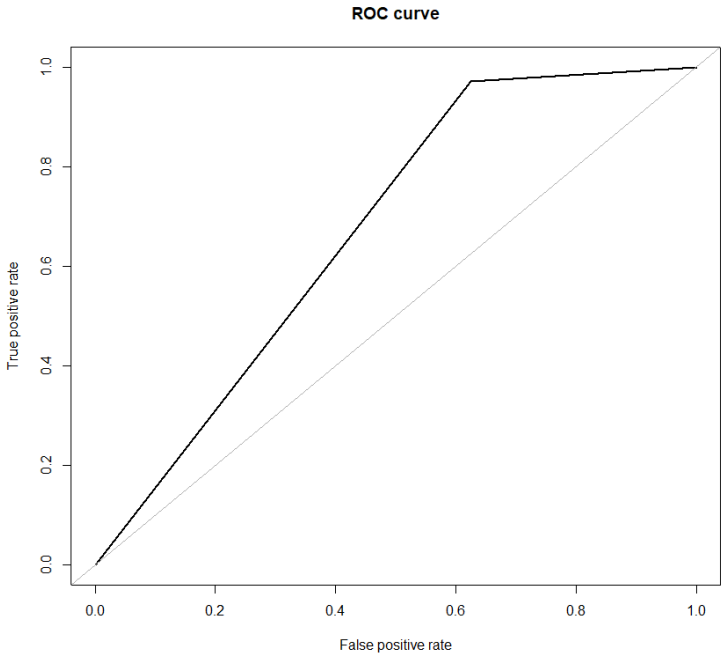

Fig 72

Date: 17/09/2018

The random forest has the following confusion matrix and ROC curve (AUC = 0.704).

```
> confusionMatrix(predictions.rf,validation[,y.variable])
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  21   3
         1  27 102

               Accuracy : 0.8039
                 95% CI : (0.7321, 0.8636)
    No Information Rate : 0.6863
    P-Value [Acc > NIR] : 0.0007737

                  Kappa : 0.4731
 Mcnemar's Test P-Value : 2.679e-05

            Sensitivity : 0.4375
            Specificity : 0.9714
         Pos Pred Value : 0.8750
         Neg Pred Value : 0.7907
             Prevalence : 0.3137
         Detection Rate : 0.1373
   Detection Prevalence : 0.1569
      Balanced Accuracy : 0.7045
```

Fig 73

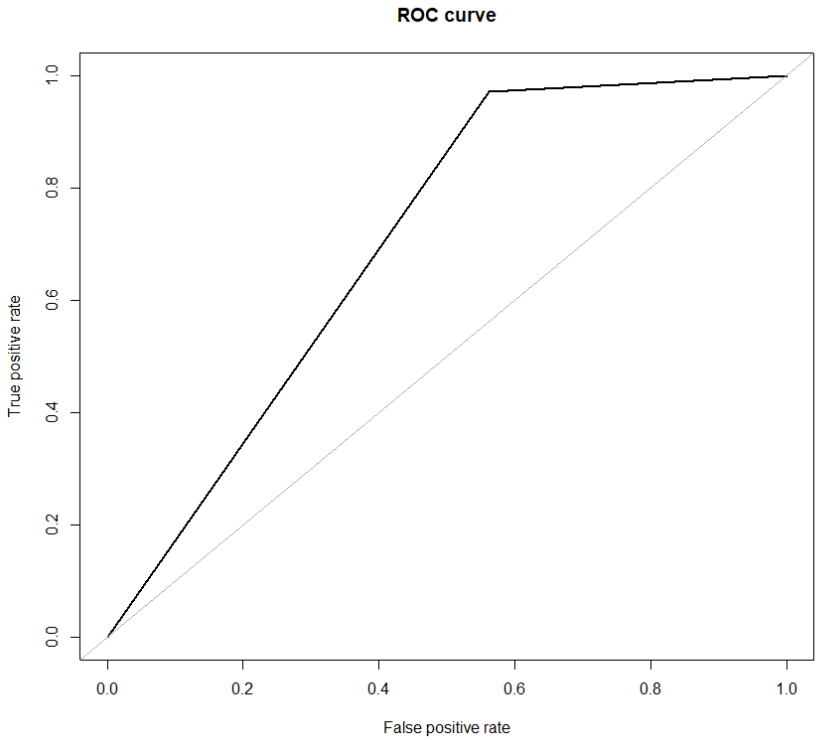

Fig 74

The last model, the linear regression has the following confusion matrix and ROC curve (AUC = 0.673).

```
> confusionMatrix(predictions.glm,validation[,y.variable])
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  18   3
         1  30 102

               Accuracy : 0.7843
                 95% CI : (0.7106, 0.8466)
    No Information Rate : 0.6863
    P-Value [Acc > NIR] : 0.00468

                  Kappa : 0.4089
 Mcnemar's Test P-Value : 6.011e-06

            Sensitivity : 0.3750
            Specificity : 0.9714
         Pos Pred Value : 0.8571
         Neg Pred Value : 0.7727
             Prevalence : 0.3137
         Detection Rate : 0.1176
   Detection Prevalence : 0.1373
      Balanced Accuracy : 0.6732
```

Fig 75

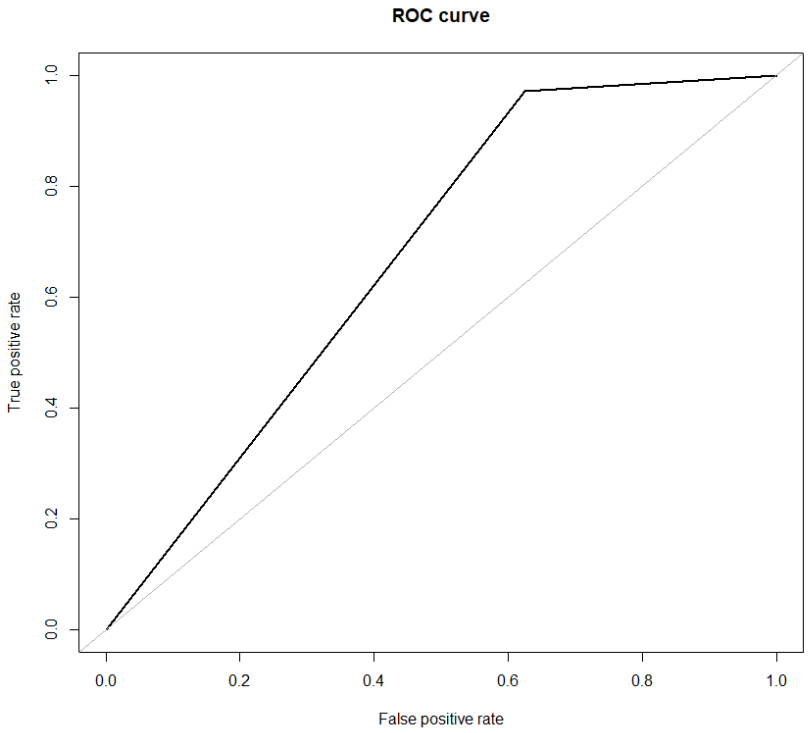

Fig 76

It is possible to observe that the models have very similar predictive capability; based on the AUC, the best model is the random forest.

11. Discussion of the examples

Example number one is of great potential importance in the social sciences as it allows to quickly retrieve data from a webpage by webscraping it and shows how to go from messy data to tidy data, ready for analysis. This is fundamental as it stops a potential researcher from "going with their guts" after having read a newspaper article and instead give them the opportunity to verify their hypothesis by allowing easy access and manipulation of data. In fact, in the analysis it was possible to find out that the source to why it was indeed a lie always came from a left wing news outlet and that most of the sources were the New York Times itself which may or may not suggest a bias. It is the role of the social scientist to consider the quantitative data insight so that more informed qualitative research can happen.

The second example is a more complex task, it uses an already available dataset, and after the necessary preprocessing, descriptive and exploratory statistics were displayed, which is a good way of getting a feel for the dataset; but more importantly, the task tries to explain extremely complex psychological mechanisms using data. Predicting users` likelihood of enjoying a specific movie is not only very useful for business but it also sheds light into people`s psychology and this knowledge can be expanded to other fields. Recommender systems are increasingly more popular for business applications, the same approach would be extremely useful in other aspects of social science and it is a first step towards making social science knowledge more phronetic and practical.

The third example is closer to that side of social science that struggles to be practical, the data analyzed is the European Social Survey, therefore is a sociological research as it deals with many people`s opinions. The methodologies used in the example are not the only ones possible as this type of data also needs a qualitative approach but the quantitative approach can find patterns that would be impossible to conceive due to the large amount of data and variables. The exploratory multivariate analysis allows to reduce the dimensionality and examine several variables at once with visualization so that possible patterns are easily recognizable; this approach is extremely useful in questionnaires and surveys.

Example number four deals with financial data. Finance can be seen as a mere interaction of number and technical indices but the reality is that finance is made up of human actors; which is the key difference from physical sciences and social sciences, the first studies objects that do not possess an own mind, the latter studies agents with a mind and thus less predictable. The task showed how to quickly extract massive amount of financial data and model the future stock prices for a company; then the task created useful visualization to inform a potential trader on which stock to buy. These methodologies, although applied to finance here, are general concepts that are applicable in other areas of social sciences.

The fifth example is useful for a wide array of possibilities, retrieving tweets can be extremely informative on the public opinion in this constantly connected society. This does not only have business utility but it is also applicable to gather information on how people, events, products are received by the public. The example uses an NPL package to analyze how positive, negative and neutral a collection of tweets is about a specific chosen topic. The possibility of doing this type of analysis, although preliminary, is a great data based (and thus more objective) addition to a researcher`s opinion on public perception which usually tends to be anecdotal.

The sixth and last example takes an existing dataset and tries to predict whether a customer should be granted a loan based on if the client will be able to pay it off. The highest prediction rate that was possible to achieve is an 80% accuracy with an AUC of 0.704. This task displayed that machine learning algorithms are a great addition to social science research/tasks as people`s behavior is only unpredictable to a certain extent and data analysis can provide incredible help on tasks that would be impossible by simple observation as the number of variables and cases are too large to keep track of.

12. Final Considerations

Science is a process where observations are made in the world in order to test hypothesis. Scientific interpretations that work, survive, while the ones that do not, die off. Eventually, the concepts that survive repeated application and massive amount of data become scientific theory (Parson and Wright, 2015). The concept of scientific theory is the maximum possible scientific echelon; it is as close as an absolute truth as humans can ever get. Reaching absolute truth is, philosophically speaking, impossible.

“Bad science” is the use of a bad experimental design or illogical interpretation of the findings. For example, not accounting for confounding variables that leads to test the hypothesis incorrectly and the inferences based on it are flawed, other problems can be a bad sample or sample size. A big problem arises when selective data is used, which is typical of people with an agenda; for example, intentionally leaving data outside the statistical process by classifying the data as outliers. Having an agenda beside unveiling the truth is extremely detrimental to the scientific process; economic or political agenda will sway research deceptively in order to achieve a desired goal, this is often the case with political propaganda or for various economic gains (Parsons and Wright, 2015), for example when, in the 1960s, the sugar industry paid Harvard researchers to blame fats for coronary heart disease so that they could continue doing business unhindered (Domonoske, 2016).

In actuality, “Bad Science” is technically not science at all, it is something that claims to be science but it is not actually following the scientific method by the word. This inevitably leads to personal opinions coloring the interpretation of what data really means (Parsons and Wright, 2015). Different scientific research instances often find very different results, this is true for some physical sciences too, for example medical research or nutritional and fitness research often has this problem (Vivalt, 2018). However, the problem is much more

pronounced with social sciences; this has led to a “replication crisis” where the outcome of research is not replicable consistently (Vivalt, 2018) and therefore it cannot be useful for practical tasks such as in business or policy-making contexts. One way to use social science more effectively is to read systematic reviews that synthetize results from many studies in order to not rely on a single study (Vivalt,2018).

The scientific community debates endlessly on controversial research because, as stated above, nothing is actually proven absolutely and it is impossibly hard to pinpoint a single interpretation as true and all other interpretations as false (Parsons and Wright, 2015). Therefore, what is regarded as undeniable true scientific theory (for example the law of gravity) is a scientific community consensus based on overwhelming data and logical conclusions, but it is still limited by our human condition. The role of the scientific community is to focus on the best available science while constantly looking for a better interpretation.

This paper is meant as a call for action to reshape social sciences towards a “best science available” approach that involves the large amount of data and the statistical and computational practices that are widely available today. Thus moving away from personal opinion papers that cause a very low citation incidence and moving towards a practical approach where the intention is rigorous research with reproducible results that yields useful knowledge.

Using a reductionism approach, especially a methodological reductionism one, reality can be reduced entirely to physics, for both physical phenomena and abstract phenomena, such as social or psychological phenomena. Everything that exists can be seen as a natural consequence of atoms interacting with one another (Ney, no date). Therefore, with an infinite and absolute knowledge of the world it would be possible to predict anything. Although infinite and absolute knowledge is evidently not possible with the current means, the role of science is to strive for it relentlessly.

Bibliography (Harvard referencing)

Ackland, R., 2013. Web social science: concepts, data and tools for social scientists in the digital age. London: Sage

Berkelaar, B.L. and Francisco-Revilla, L., 2018. Motivation, Evidence, and Computation: A Research Framework for Expanding Computational Social Science Participation and Design. In: M. Egger, C.M. Stuetzer and M. Welker eds. 2018. Computational Social Science in the Age of Big Data Concepts, Methodologies, Tools, and Applications. Köln, GERMANY: Herbert von Halem Verlag. Pp. 16-62.

Domonoske,C., 2016. 50 Years Ago, Sugar Industry Quietly Paid Scientists To Point Blame At Fat. *The two-way (NPR)*, [online] 13 September. Available at : https://www.npr.org/sections/thetwo-way/2016/09/13/493739074/50-years-ago-sugar-industry-quietly-paid-scientists-to-point-blame-at-fat [Accessed 09/09/2018]

Egger, M., Stuetzer, C.M. and Welker, M., eds. 2018. Computational Social Science in the Age of Big Data Concepts, Methodologies, Tools, and Applications. Köln, GERMANY: Herbert von Halem Verlag

Grolemund,G. and Wickham H., 2017. *R for Data Science*. [e-book] O`Reilly Media. Available at: http://r4ds.had.co.nz/index.html

Joel, J. and Heikki, K., 2015. The use of Web analytics for digital marketing performance measurement. Industrial Marketing Management. Vol.50, pp.117-127

Jünger, J., 2018. Mapping the Field of Automated Data Collection on the Web: Collection Approaches, Data Types, and Research Logic. In: M. Egger, C.M. Stuetzer and M. Welker eds. 2018. Computational Social Science in the Age of Big Data Concepts, Methodologies, Tools, and Applications. Köln, GERMANY: Herbert von Halem Verlag. Pp. 104-130

Kotler, P., Kartajaya, H. and Setiawan, I., 2017. Marketing 4.0 moving from traditional to digital. Hoboken, New Jersey: Wiley

Leonhardt, D. and Thompson, S.A., 2017. Trump`s Lies. *The New York Times*, [online] 14 December. Available at : https://www.nytimes.com/interactive/2017/06/23/opinion/trumps-lies.html [Accessed 16/09/2018]

Ney, A., no date. Reductionism. *Internet Encyclopedia of Philosophy*,[online]. Available at : https://www.iep.utm.edu/red-ism/ [Accessed 09/09/2018]

Nicolai, P., 2009. MAKING SOCIAL SCIENCES MORE SCIENTIFIC: THE NEED FOR PREDICTIVE MODELS - by Rein Taagepera. Public Administration, Vol.87(4), pp.977-980.

Parsons, E.C.M. and Wright, A. J., 2015. The good, the bad and the ugly science: examples from the marine science arena. *Frontiersin,*[online] 04 June. Available at : https://www.frontiersin.org/articles/10.3389/fmars.2015.00033/full [Accessed 09/09/2018]

Schwab, A., 2011. Making Social Sciences More Scientific: The Need for Predictive Models. Public Administration Review, Vol.71(5), pp.807-810

Taagepera, R., 2008. Making social sciences more scientific: the need for predictive models. Oxford: Oxford University Press.

Tart, I., 2009. Can Physics Save Social Sciences? Book Review:Making social sciences more scientific: the need for predictive models by Rein Taagepera, 2008, Oxford. Studies of Transition States and Societies, Vol.1(1), pp.92-9.

Date: 17/09/2018

Vivalt, E., 2018. How to Be a Smart Consumer of Social Science Research. *Harvard Business Review*, [online] 27 July. Available at : https://hbr.org/2018/07/how-to-be-a-smart-consumer-of-social-science-research [Accessed 09/09/2018]