National College of Ireland

Data Warehousing and Business Intelligence Project

on

Analysis on S&P 500 Index

Smit Jain

x18135340

https://www.youtube.com/watch?v=_nqUg4cVk54

MSc Data Analytics – 2019/20

Submitted to: Sean Heeney

National College of Ireland
Project Submission Sheet – 2019/2020
School of Computing

Student Name:	Smit Jain
Student ID:	x18135340
Programme:	MSc Data Analytics
Year:	2019/20
Module:	Data Warehousing and Business Intelligence
Lecturer:	Sean Heeney
Submission Due Date:	12/4/2019
Project Title:	Analysis on S&P500 Index

I hereby certify that the information contained in this (my submission) is information pertaining to my own individual work that I conducted for this project. All information other than my own contribution is fully and appropriately referenced and listed in the relevant bibliography section. I assert that I have not referred to any work(s) other than those listed. I also include my TurnItIn report with this submission.

ALL materials used must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is an act of plagiarism and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	Smit Jain
Date:	April 8, 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table 1: Mark sheet – do not edit

Criteria	Mark Awarded	Comment(s)
Objectives	of 5	
Related Work	of 10	
Data	of 25	
ETL	of 20	
Application	of 30	
Video	of 10	
Presentation	of 10	
Total	of 100	

Project Check List

This section capture the core requirements that the project entails represented as a check list for convenience.

- ☒ Used LaTeX template
- ☐ Three Business Requirements listed in introduction
- ☐ At least one structured data source
- ☐ At least one unstructured data source
- ☐ At least three sources of data
- ☐ Described all sources of data
- ☐ All sources of data are less than one year old, i.e. released after 17/09/2017
- ☐ Inserted and discussed star schema
- ☐ Completed logical data map
- ☐ Discussed the high level ETL strategy
- ☐ Provided 3 BI queries
- ☐ Detailed the sources of data used in each query
- ☐ Discussed the implications of results in each query
- ☐ Reviewed at least 5-10 appropriate papers on topic of your DWBI project

Analysis on S&P 500 Index

Smit Jain
18135340

April 12, 2019

Abstract

S&P 500 Index is an Index of 505 stocks which is maintained by Standard & Poors Dow Jones indices. These stocks are issued by 500 companies which are a part of American Stock Exchange. There are 5 component companies which are having 2 share classes of stock, that's the reason why there are 505 stocks in the S&P 500 Index.
There is a variety of analysis that can be done over stock market. But for any individual, before investing in any stock, the major factor on deciding which sector to invest in, is driven by the earnings of that particular sector. Every company has a specific weight index, according to which, more valuable companies are having higher index.
To analyze this, I have made a data warehousing and Business Intelligence system from the different datasets (structured and un-structured) available online which will help the investors in making decision for which sector they should invest in.

1 Project Introduction

When people buy shares of any company which are listed in the stock market, they are like gamblers trying their luck in Vegas. Most of the people in United States of America invest in stock markets. America's Stock market is huge and there are companies ranging from fortune 500 to pink sheet or penny stock companies. Most of the people are interested in earning profit and hence they tend to invest more towards the more reputed and firmly established companies. The Standard and Poor's is such an index which maintains a list of 500 companies that are the best performing companies in the American market. If we are able to find out which sector's share earnings are the highest, and weight of the companies in different industry, it will give a rough picture of where and how much to invest. This project aims to find the behavioral pattern/trend of the S&P 500 index companies performance, which will help anyone investing in the American stock market to analyze the market first and then invest. The performance of the S&P 500 index for last 5 years has been obtained from Statista. Considering the performance of past 5 years, other datasets have been found.

(Req-1) To find the sector wise share earnings.

(Req-2) To find the industry wise weighing index of different countries.

Source	Type	Brief Summary
Statista	Structured	Annual development of S&P index.
Kaggle	Structured	Historical data for all S&P 500 companies.
Investing	Structured	Fundamental details of S&P 500 companies.
Slickcharts	Unstructured	Weight of components of S&P 500 index.
Wikipedia	Unstructured	List of S&P 500 companies

Table 2: Summary of datasets used.

(Req-3) To find each companies performance over the last 5 years in comparison to the annual stock index.

The sequence of the project report, detailed further, is as follows:

- Data source details
- Research related to this subject
- Data Model
- ETL process
- Application
- Conclusion and future work
- References
- Appendix

2 Data Sources

I have used a total of 5 datasets for this project, out of which 3 are structured and 2 are unstructured. The table (2) represents a short summary of the datasets used.

2.1 Source 1: Kaggle

This data source is having all the historical data of previous five years related to the S&P 500 components. The dataset was published in March 2019. The data present is having each stocks open and close amount of each day & volume traded for each stock. There are a total of 6,19,040 records present in the table. The cleaning of the dataset has been done using R. The R-script for the cleaning of this dataset has been included(8) in the appendix section of the report. The URL to download the dataset can be found below:

`https://www.kaggle.com/florentbaptist/sp-500`

The attributes used from this dataset are as follows:

- Date
- Open

- Close
- Name
- Volume

2.2 Source 2: Statista

This data source has the annual index point of the S&P 500 index. The dataset is having annual index starting from 1986-2018. The dataset was published on January 2019. The dataset downloaded from statista was in excel format and that too with two sheets. The data cleaning has been done using R to remove extra rows and bring the data in a standard format. The R-code has been attached(8) within the appendix section of the report. The dataset has been downloaded from the following source:

url `https://www.statista.com/statistics/261713/changes-of-the-sundp-500-\ \during-the-us-election-years-since-1928`

The attributes used from this dataset are as follows:

- Performance year
- Index Points

2.3 Source 3: Investing.com

This dataset contains all the fundamental details related to a stock market company of S&P 500 index. The revenue generated by each company has been mentioned here in the table. This field will help us in determining the sector wise earnings. This dataset has been cleaned with the help of R. The R-script to clean the code can be found in the appendix attached(8) in the end of report. The URL to download the dataset can be found below:

`https://www.investing.com/indices/investing.com-us-500-components`

The attributes used from this dataset are as follows:

- Name
- Average volume
- Market cap
- Revenue

2.4 Source 4: Slickcharts.com

This dataset has been obtained by web-scraping using R. The data that will be useful from this dataset is the weight of each company in the total index weight. The dataset has been cleaned up with the help of R. The R-script is attached(8) in the index section

of the project report. The URL to the website is:

`https://www.slickcharts.com/sp500`

The attributes used from this dataset are as follows:

- Company name
- Symbol
- Stock weight

2.5 Source 5: Wikipedia.com

This dataset has been created by web-scraping using R. The data related to sector and sub-sector of each company can be found here. For cleaning the data, there were many columns that are not required for any computations. Hence, I have removed all the unnecessary columns using R. The R-script for cleaning the data has been attached(8) in the index section of the project report. The URL to the website is:

`https://en.wikipedia.org/wiki/List_of_S%26P_500_companies`

The attributes used from this dataset are as follows:

- Company name
- Symbol
- Sector
- Sub-industry
- CIK

3 Related Work

Stock market is a very hot topic for any researcher to do his/her research work on. There are numerous people who do research on trend analysis of stock market and write their reports on. According to the type of analysis done in the paper Kominek (2004) the impact of stock market on economy growth is significant. This paper by the economist, provides a case-study-based insight into the direction of causality between stock market development and growth. According to the analysis done Abera W. H. (n.d.) in the paper, the analysis has been done to know about the impact of stock market performance on firms growth. Multiple factors like technological innovation, public policy and market landscape, the nature of competitions, firms (cost) structures, firms strategies, stock market performances managerial skills and other indigenous and exogenous factors have an explicit and implicit influence on firms growth. From the article by RAFA CZYZYCKI czyzycki (2014) - Brand of the investment fund companies and risk of an investment at a capital market for an individual investor. Uncertainty and risk are the elements, which are inseparably connected with investment operations at the capital market. Amanulla (n.d.)

conducted a study to examine the Indian stock market efficiency by using Ravallion co integration and error correction market integration approaches. The article by Argiddi & Kale (2014) gives a brief overview on stock market analysis and defines a research domain to understand the intelligence of stock market. This refers as stock market intelligence, which is to develop data mining techniques. Another study Wu & Zhang (2018) has been done related to the trade size and liquidity of stock market. The authors have develop a 3D measure for liquidity where they have studied the relationship of liquidity with the order flow in S&P 500 index. The analysis concludes that the trade size during the time of high liquidity is also high. As the limit order book's depth increases, the resilience also increases, the bid-ask spread decreases, and the trade size increases.In the paper by Grantham (2019), the importance of sectors has been analyzed over time for all the listed equities, where the sectors and the constituent components of the S&P 500 index have been taken in prime focus. Any firms individual performance has always been strongly correlated to the sector under which it is categorized. It has been concluded in the analysis by Edwards & Bruna Pipino (2018) that the sectoral effects are almost half of variance as calculated in the daily returns. In another research done by Silverblatt (2018) Information Technology(IT) was the leading sector in the S&P 500 index also it was the only sector in the index which had maximum weight. This gave me major motivation in determining the sector wise weight so that the investors can wisely decide which sector to invest in. Another study done by Soe & Guarino (2010) takes the assumption that the rankings of the companies reflect the variability of their growth and stability in terms of earnings and dividends. The change in sector, style, and size of high and low quality stocks are signified by the change in S&P 500 High Quality Rankings Index and the change in S&P 500 Low Quality Rankings Index, and also the indices with exposure to the S&P 500 constituents having above average and having below average rankings, respectively. The quality can also be considered same as sector, style and size exposures.

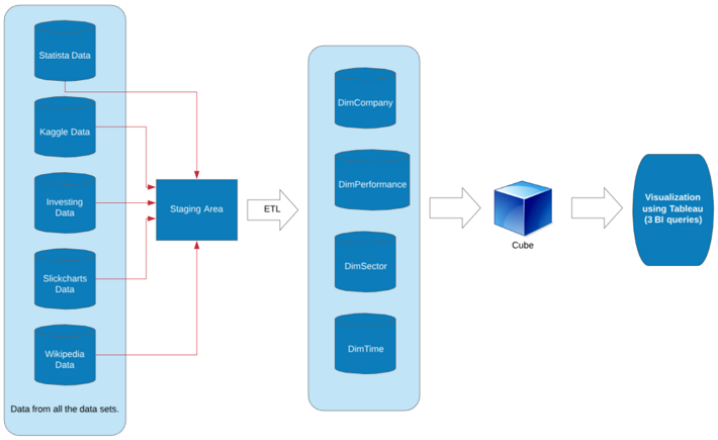

Figure 1: Design architecture of Data Warehouse

In this project, I've applied Kimball & Ross (2013) methodology - bottom-up because it consumes less space, makes the execution simple and it is quicker if compared with Inmon's methodology. In future, if there's a need to do changes in fact-table, at that point Inmon's methodology should be implemented. Be that as it may, for our situation, we will take the fact table as it is.

4 Data Model

To design this data warehouse, I have used the most widely used star schema in which all the dimension tables are connected directly to the centrally placed fact table using the foreign key attributes. The Fact table has only the numeric measures and the dimensions are all stored in the dimension tables. All the dimension tables have one primary key each, which are present in the fact table as foreign key. This helps link each dimension to the fact which is centrally placed in the star schema. The structure for this star schema can be found below (1).

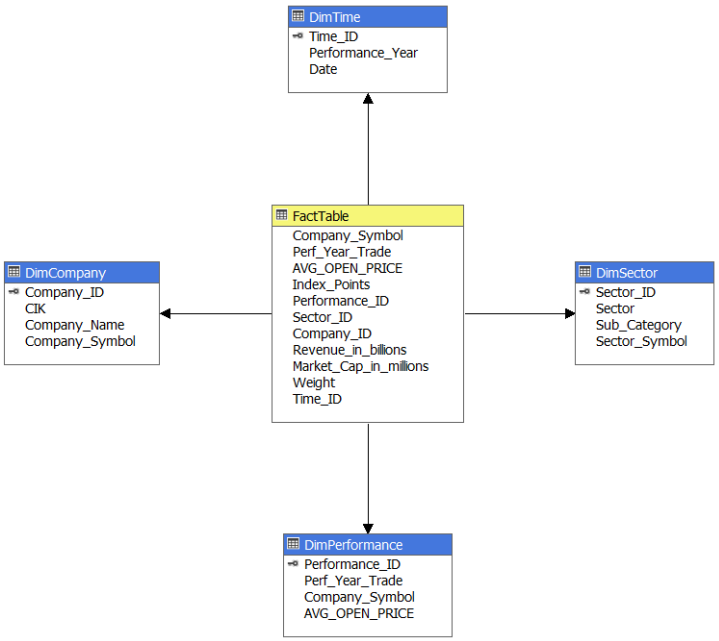

Figure 2: Data Model: Star Schema

The data sets used in the project can be molded into 4 meaningful dimensions which are as named Dimension Company (DimCompany), Dimension Performance (DimPerformance), Dimension Sector (DimSector), Dimension Time (DimTime). Explanation and composition of each of the four dimensions has been given below.

4.1 Dimension 1: Dimension Company

This dimension contains all the data related to the company. To populate this dimension, we have used one dataset which gives information related to all the S&P 500 companies. All the datasets have a common column which is the companys stock ticker symbol. The cleaning has been done using R for the raw datasets and further the file has been imported to SSIS as a flat file source and a table has been created in the SQL database. Further DimCompany which contains following columns:

- Company_ID: Primary Key
- CIK: unique identifier of each company
- Company_Name: Name of the company

- Company_Symbol: Stock ticker of the company

4.2 Dimension 2: Dimension Performance

This dimension contains all the data related to the performance of the S&P Index over the years. To populate this dimension, we have used one dataset only. The dataset needed some cleaning, which was taken care of with the help of R. The dimension Dim-Performance contains following columns:

- Performance_ID: Primary Key
- Perf_Year_Trade: Year of trade
- Company_Symbol: Symbol of the company

4.3 Dimension 3: Dimension Sector

This dimension contains all the data related to the sectors and sub-industry present in the American stock market. To populate this dimension, we have used one data set extracted from Wikipedia. The dataset needed a lot of cleaning, which included removing extra columns, changing column names and removing unwanted characters from the columns. All of this has been taken care of with the help of R. The dimension contains following columns:

- Sector_ID: Primary key
- Sector: Sectors in which all the companies fall into
- Sub_Category: Industry details of all the companies
- Sector_Symbol: Symbol of each company

4.4 Dimension 4: Dimension Time

This dimension contains all the time related data. The data has been populated in this table using two datasets, and date column represents the exact number of days on which the American stock market was active. Following are the columns included:

- Time_ID: Primary key
- Performance_Year: performance year
- Date: date of each day stock performance

5 Logical Data Map

The logical data map, i.e. each row of the data source used is presented in a tabular form in this section. This step is important and necessary to formulate before ETL process.

Table 3: Logical Data Map describing all transformations, sources and destinations for all components of the data model illustrated in Figure 2

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
NA	Company_ID	DimCompany	Company ID	Dimension	Auto-incremental field.
5	CIK	DimCompany	Central Index Key	Dimension	Used the column data as it is.
4	Company_Name	DimCompany	Name of the company	Dimension	Changed the column name.
4	Company_Symbol	DimCompany	Company symbol	Dimension	Changed the column name.
NA	Performance_ID	DimPerformance	Performa-nce ID	Dimension	Auto-incremental field.
1	Perf_Year_Trade	DimPerformance	Year of Index	Dimension	Corrected the format.
NA	Sector_ID	Dim Sector	Sector ID	Dimension	Auto-incremental field.
5	Sector	DimSector	Sector	Dimension	Used the column data as it is.
5	Sub_Category	DimSector	Sub Category	Dimension	Used the column data as it is.
5	Sector_Symbol	DimSector	Ticker Symbol	Dimension	Used the column data as it is.
NA	Time_ID	DimTime	Time ID	Dimension	Auto-incremental field.
1	Performance_Year	DimTime	Performance year	Dimension	Used the column data as it is.

Continued on next page

Table 3 – *Continued from previous page*

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
2	Date	DimTime	Date	Dimension	Date of the stock performance.
2	AVG_OPEN-_PRICE	FactTable	average of open price	Fact	took average of the open price and stored.
1	Index_Points	FactTable	Index Price per year	Fact	Used the column data as it is.
3	Revenue_in_Billions	FactTable	Revenue in Billions	Fact	Converted the data type to numeric.
3	Market_Cap-_in_Millions	FactTable	Market cap in millions	Fact	Converted the data type to numeric.
4	Weight	FactTable	stock weight	Fact	Converted the data type to numeric.

6 ETL Process

Extraction:
For extraction of data, Ive used 5 different sources of data, out of which 3 are structured data sources and 2 are un-structured data sources. The structured data set has been downloaded directly from the websites and the unstructured data sources have been extracted using web scraping.

Cleaning:
The data needs to be cleaned before processing it further. Ive cleaned the data here with the help of R. The cleaning of data and converting it into flat files has been done using automation. The execute process task in SSIS module has been implemented to process this task. The details of cleaning the data has been described below:

- Dataset 1: Statista (Structured) This data set has been downloaded from statista.com, which is available in CSV format. The data present in this has been cleaned by deleting the extra rows and the format of the year has been changed.

- Dataset 2: Kaggle (Structured) This data set has been downloaded from Kaggle.com, which is available in csv format. Ive added another column to this data set and changed the column name using R. Apart from these modifications, all the data from this data set has been used without any modifications.

- Dataset 3: Investing (Structured) This data set has been downloaded from Investing.com. The raw data set contains a lot of values quoted in billions and millions and thousands. We need to convert those values to actual numbers so that we can perform calculations based on that. This has been taken care of with the help of R.

- Dataset 4: Slickcharts (Unstructured) This data set has been scraped from the website slickcharts.com. The raw data set contains few characters in the price column and there is an extra column which is not required, both of these have been cleaned through R.

- Dataset 5: Wikipedia (Unstructured) This data set has been scrapped from the website Wikipedia.com. The raw data set has columns which are not required and a few characters in the columns which have been cleaned using R.

Transform:
The cleaning has been done using automation and the outputs of all the sources have been saved in csv format. The tables are loaded into SSIS as flat file sources and then using OLE DB destination these tables are pushed to staging area. The dimensions are created next using the data available in the staging area tables which is then populated using the insert queries. To populate the fact table, Ive used a combination of SQL query and lookup to pull out the numeric values.

Load:
For loading the data into SQL databases, SSIS has been used. The raw data sets have

been pulled up in staging area first with the help of R automation and then the data has been populated into dimension and fact tables.

Degree of Automation:
All of the ETL process has been automated. All that happens on a single click is:

- Extraction of the data sources.
- Cleaning of the data sources using R.
- Transformation of data to staging area.
- Loading of the data into dimension and fact tables.
- Truncation of the data present already in the dimension and fact tables.
- Automated cube processing and deployment.

7 Application

The visualization has been done using tableau desktop version. The BI queries have been built and run on tableau. The three non- trivial BI queries that I have made are as follows:

7.1 BI Query 1: Which sector weighs more in the S&P 500 Index.

For making this BI query, the sources of data that are used are dataset 3 (Investing.com) and dataset 5 (Wikipedia.com). This query has been executed considering two major factors- Sector (Companys sector) and Revenue generated by the company.
The general conclusion that can be drawn is that the weight of each company is a direct implication of the size of the company. In the Figure 3, it can be seen that different sectors have different companies and the company's weight shows how big the company is. The weight of any organization is regularly utilized benchmark for any organization's performance on the portfolio.

In tableau, If you select any of the given sectors in the bubble chart, the companies falling under that sector will be reflected on the right side with their respective weights.

7.2 BI Query 2: Which industry is earning more in the S&P 500 index.

For making this BI query, the sources of data that are used are dataset 4 (Slickcharts.com) and dataset 5 (Wikipedia.com). This query has been executed considering two major factors- Industry (Companys industry) and weight of each company.
The general conclusion that can be drawn is that the earnings are not concentrated towards any specific sector. It is distributed in all the sectors. Hence, this will facilitate the investor to make his decision wisely.

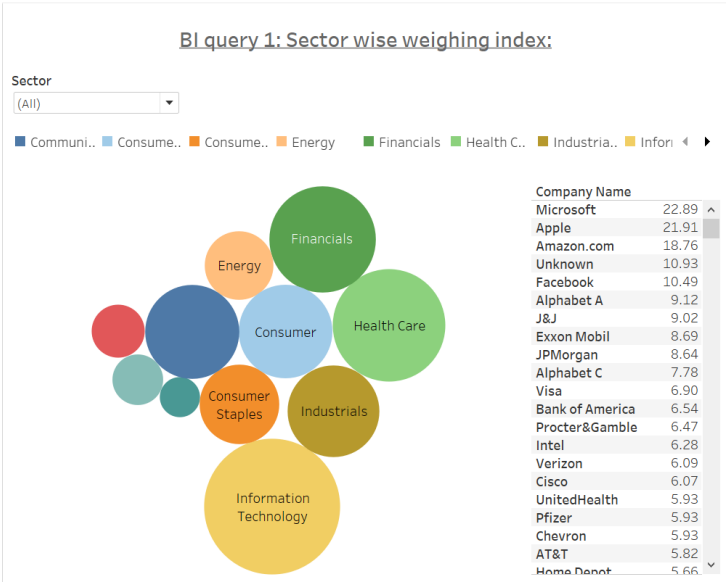

Figure 3: Design architecture of Data Warehouse

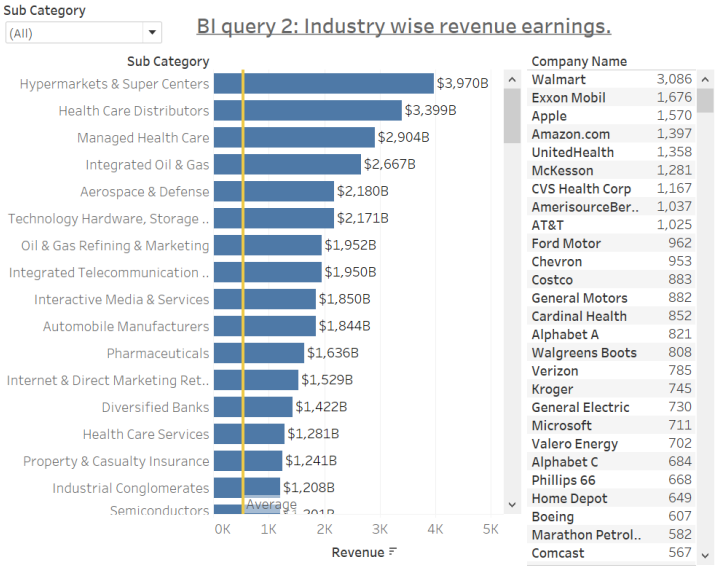

Figure 4: Results for BI Query 2

In tableau, the average revenue of all companies is shown by the yellow line (Figure 4). Here, we can easily identify the companies which are making revenue more than the average are definitely performing better. Hence, it will help the investor to make the decision wisely.

7.3 BI Query 3: Performance of each company compared to index price of last 5 years.

For making this BI query, the sources of data that are used are dataset 1 (Statista) and dataset 2 (Kaggle). This query has been executed considering two major factors -Performance of S&P 500 index of last 5 years and each stocks individual performance over previous 5 years.

The general conclusion that can be drawn is that the performance of overall S&P 500 index is dependent on each companys individual performance in stock market.

From the visualization done in tableau, it can be seen that if you want to see the performance for any specific year (example 2017) and for the top N companies (for example 10), the result will be reflected in the heat map and the name of the companies will be displayed with the contribution percentage displayed right next to it. (Figure 5)

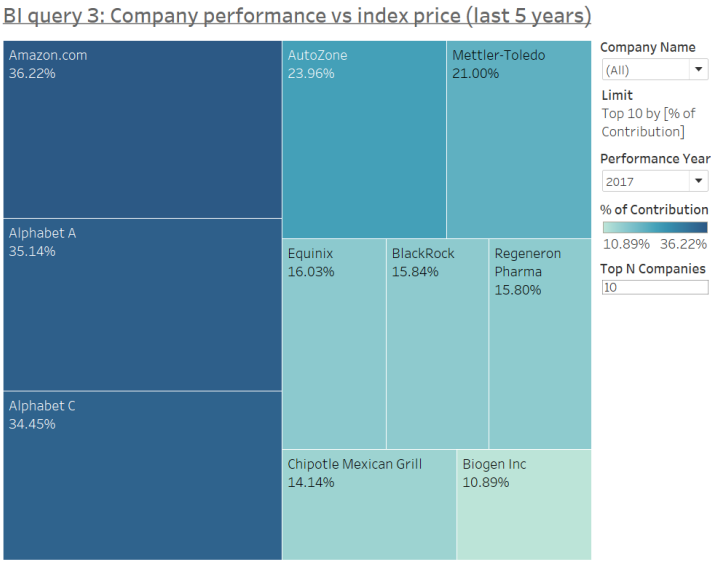

Figure 5: Results for BI Query 3

Depending on the requirement, the results in the heat map can be altered by just selecting the year from the drop-down list and the number of companies. If the details specific to any specific company is required, then it can be selected from the drop-down field under the label company name.

7.4 Discussion

A detailed discussion / summarisation of the findings from the 3 queries. Note that this discussion will have a lot more detail than the discussion in the following section (Conclusion). You should relate your main findings to the literature that you reviewed in Section 3, i.e. those with a similar topic to your data warehousing project (but which are not necessarily data warehousing projects), and compare and contrast your findings with theirs.

The S&P 500 index is an index with 505 companies listed in it. It is very difficult for an investor to go to the portfolio of each company and check its details and the performance.

To solve this difficulty, I've designed a data warehouse where all these questions can be answered. The Business Intelligence queries that are implemented in tableau and answers all the questions for any investor.

From the above mentioned BI queries, one can easily identify which sector and which industry is best performing. Also, before investing into any particular company, one can actually check the company's previous performance and accordingly come to a decision that the company is worth investing or not.

The first BI query is used to determine which sector's weight in the S&P 500 index is the highest. After that, if we select any specific sector, the companies which belong to that sector are displayed in a list with their respective weights which is regularly utilized benchmark for any organization's performance on the portfolio. Here, by looking at the bubble chart, we can easily judge which sector is having maximum weight. Suppose, we find out that the sector 'Information Technology' is having the maximum weight. So we can choose any company from the list of companies under the Information Technology sector. Moving on to the second BI query, it determines the highest revenue generated by any industry and also shows the industries which are generating revenue more than the average revenue rate in the market. Here, we can easily find out by analysing the bar chart. The yellow vertical line in the graph denotes the average revenue rate of all the industries. The industry with maximum revenue is displayed at the top and if we select it by clicking on it, the companies with their respective revenues will be displayed on the right side of the graph. Hence, for investing the two major factors have been decided from the first two BI queries. The question still remains, which company to invest into. This question can be answered by the third BI query. Here, the performance of each stock's previous five years is calculated with respect to its contribution to the stock index of any particular year. Suppose that the user decides to buy the shares of the company Accenture. So, before investing, he can check Accenture's previous years performance directly by selecting Accenture from the drop down under the label company name. If the investor wants to check the performance of top 10 companies for the year 2017, he can enter the number of companies as 10 and select the performance year to 2017 and the results will be reflected in the heat map.

8 Conclusion and Future Work

The analysis done above with the help of three BI queries clearly explains us that before investing capital in any stock, the investor needs to do the analysis which can result him in earning profits. The basic questions that are raised before investing are which sector to invest into? Which industry should be opted to invest in? and last but the most important, which stock to choose? All these questions were answered successfully above with the help of the BI queries.

If we take for an example, the sector with maximum weight is Information Technology. The companies with highest weight in it is Microsoft and then Apple. In the second BI query, we can check the highest revenue generated by the companies we selected in the first step. So Apple is having higher revenue generated in comparison to Microsoft. In the third BI query, we can check the company - Apple's previous year performance. Also, we can check other companies which were top contributors in the previous years and finally come to a decision.

As of now, we are able to do analysis over the companies previous year's performance. As a part of future work, I would like to add a functionality where we can predict the performance of the stocks for the upcoming year by doing regression analysis. By doing that, the investor can not only get an idea of the next years stock performance, but also be able to manage all the funds in different stocks without any hassle.

References

Abera W. H., Gouder S., S. M. R. B. . (n.d.), 'Impacts of stock market performance on firms growth: With reference to south africa', *Financial Markets* .

Amanulla, S Kamaiah, B. . (n.d.), 'Market integration as an alternative test of market efficiency: A case of indian stock market', *Artha Vijnana: Journal of The Gokhale Institute of Politics and Economics* .

Argiddi, R., A. S. & Kale, B. . (2014), 'An analysis on stock market intelligence and research approaches. international journal of application or innovation in engineering management (ijaiem)', *https://pdfs.semanticscholar.org/db2c/ac71321fdca6b3f9d65caafa5b75cf1f60ba.pdf* **3**(1), 297–300.

czyzycki, r. . (2014), 'Brand of the investment fund companies and risk of an investment at a capital market for an individual investor', *https://www.academia.edu/36639180/BRAND_OF_THE_INVESTMENT_FUND_COMPANIES_AND_RISK_OF_AN_INVESTMENT_AT_A_CAPITAL_MARKET_FOR_AN_INDIVIDUAL_INVESTOR* p. 22.

Edwards, T., L. C. P. H. & Bruna Pipino, F. . (2018), 'Global applications of s&p 500 sectors', *Index Investment Strategy* **1**(1), 1–31.

Grantham, J. . (2019), 'Sector effects in the s&p 500', *Finance Research Letters* **1**(1), 1–24.

Kimball, R. & Ross, M. . (2013), *The data warehouse toolkit*.

Kominek, Z. . (2004), 'Stock markets and industry growth: an eastern european perspective', *Applied Economics* **36**(10), 1025–1030.

Silverblatt, H. . (2018), 'S&p 500 2017: Global sales', *Research Core* **1**(1), 1–15.

Soe, A. & Guarino, D. . (2010), 'Is high quality always better?', *Research & Design* **1**(1), 1–14.

Wu, L., Y. X. F. Z. & Zhang, R. . (2018), 'Do investors choose trade-size according to liquidity, empirical evidence from the s&p 500 index future market', *Finance Research Letters* **28**(10), 275–280.

Appendix

R code 1

```
#Investing cleaning

#read file from local
investing <- read.csv(file = "C:\\Users\\MOLAP\\Downloads\\RawDatasets
\\Investing.csv", header = TRUE)

#remove extra columns
investing$P.E.Ratio <- NULL
investing$Beta <- NULL

#Change the column names
colnames(investing)[1] <- "Name"
colnames(investing)[3] <- "Average_Volume_in_K"
colnames(investing)[4] <- "Market_Cap_in_millions"
colnames(investing)[5] <- "Revenue_in_billions"

investing$Average_Volume_in_K <- NULL

#convert million and billion to number in Average_Volume column
#investing$Average_Volume_in_K <- gsub('B', 'e6',
investing$Average_Volume_in_K)
#investing$Average_Volume_in_K <- gsub('M', 'e3',
investing$Average_Volume_in_K)
#investing$Average_Volume_in_K <- gsub('K', '',
investing$Average_Volume_in_K)
#investing$Average_Volume_in_K <- format(as.numeric
(investing$Average_Volume_in_K), scientific = FALSE, big.mark = ",")
#investing$Average_Volume_in_K  <- gsub("[,]","" ,
investing$Average_Volume_in_K ,ignore.case = TRUE)
#investing$Average_Volume_in_K <- as.numeric
(investing$Average_Volume_in_K)


#convert million and billion to number in Revenue column
investing$Revenue_in_billions <- gsub('B',
'␣', investing$Revenue_in_billions)
#investing$Revenue_in_billions <- gsub('M',
'␣', investing$Revenue_in_billions)
investing$Revenue_in_billions[155] <- 0.947
investing$Revenue_in_billions[187] <- 0.915
investing$Revenue_in_billions[288] <- 0.960
investing$Revenue_in_billions <- format(as.numeric
(investing$Revenue_in_billions), scientific = FALSE, big.mark = ",")
investing$Revenue_in_billions <- as.numeric(investing$Revenue_in_billions)

#convert million and billion to number in Market_Cap_in_millions column
investing$Market_Cap_in_millions <- gsub('B', 'e3',
```

```
investing$Market_Cap_in_millions)
investing$Market_Cap_in_millions <- gsub('M', '␣',
investing$Market_Cap_in_millions)
investing$Market_Cap_in_millions <- format(as.numeric
(investing$Market_Cap_in_millions), scientific = FALSE, big.mark = ",")
investing$Market_Cap_in_millions  <- gsub("[,]","" ,
investing$Market_Cap_in_millions ,ignore.case = TRUE)
investing$Market_Cap_in_millions <- as.numeric
(investing$Market_Cap_in_millions)

write.csv(investing,"C:\\Users\\MOLAP\\Downloads\\RawDatasets\\
Investingclean.csv",row.names = FALSE)
```

R code 2

```
#wikipedia webscraping and cleaning

#webscraping
#install.packages("htmltab")
library(htmltab)
url1 ="https://en.wikipedia.org/wiki/List_of_S%26P_500_companies"
wikipedia=htmltab(doc=url1, which=1)

wikipedia$`SEC filings` <- NULL
wikipedia$`Headquarters Location`<- NULL
wikipedia$`Date first added`<- NULL

#Change the column names
colnames(wikipedia)[1] <- "Company_Name"
colnames(wikipedia)[3] <- "Sector"
colnames(wikipedia)[4] <- "Sub_Industry"

# change the data types
wikipedia$CIK <- as.numeric(wikipedia$CIK)
wikipedia$Company_Name <- as.character(wikipedia$Company_Name)

write.csv(wikipedia,"C:\\Users\\MOLAP\\Downloads\\RawDatasets\\
wikipediaClean.csv",row.names = FALSE)
```

R code 3

```
#Slickcharts webscraping and cleaning

#webscraping
library(htmltab)
url="https://www.slickcharts.com/sp500"
slickCharts=htmltab(doc=url, which=1)

#removing extra characters from price column
slickCharts$Price  <- gsub("[^0-9A-Za-z.,///'␣]","␣" ,
slickCharts$Price ,ignore.case = TRUE)
slickCharts$Price <- gsub("''","" , slickCharts$Price ,ignore.case = TRUE)
```

```
# convert weight data type to numeric
slickCharts$Weight <- as.numeric(slickCharts$Weight)

# removing extra columns
slickCharts$Change <- NULL

#change colname
colnames(slickCharts)[1]<-"Weight_ID"

write.csv(slickCharts,"C:\\Users\\MOLAP\\Downloads\\RawDatasets\\
slickchartsClean.csv",row.names = FALSE)
```

R code 4

```
#statista cleaning

#read the file from local
statista <- read.csv(file = "C:\\Users\\MOLAP\\Downloads\\RawDatasets
\\statista.csv", header = TRUE)

#delete the rows not required
statista <- statista[-c(1:16),]

#change the column names
colnames(statista)[1]<-"Year"
colnames(statista)[2]<-"Index_Points"

#convert the year in correct format
statista$Year <- as.character(statista$Year)
statista$Year <- (substr(statista$Year,2,length(statista$Year)))
statista$Year <- paste("20",statista$Year)
statista$Year <- gsub("␣", "", statista$Year)
statista$Year <- as.factor(statista$Year)

statista$Index_Points <- as.numeric(gsub(",","",statista$Index_Points))

write.csv(statista,"C:\\Users\\MOLAP\\Downloads\\RawDatasets\\
statistaClean.csv",row.names = FALSE)
```

R code 5

```
# Kaggle cleaning

# Read the file from local
all_stocks <- read.csv(file = "C:\\Users\\MOLAP\\Downloads\\RawDatasets
\\all_stocks_5yr.csv", header = TRUE)

# install packages
#install.packages("lubridate")
library(lubridate)
```

```
# remove the columns that are not required
all_stocks$high <- NULL
all_stocks$low <- NULL

#change the column names
colnames(all_stocks)[4]<-"Volume_Traded"
colnames(all_stocks)[2]<-"open_price"

all_stocks["year_trade"] <- year(dmy(all_stocks$date))

write.csv(all_stocks,"C:\\Users\\MOLAP\\Downloads\\RawDatasets\\
all_stocks_clean.csv",row.names = FALSE)
```