**ITCS/DSBA – 6100 Big Data Analytics for Competitive Advantage**

**Project Phase III – Technical Document**

The project is intended to find insights from patent filed by the *Fortune 500* companies in USPTO, EPO and JPO and how it influence the Presidential Elections in the country.

**Data Extraction and Wrangling**

To achieve our project objective, we require two main datasets.

* **Fortune 500 companies’ information** that includes Patent Published, Revenue, Corporate and Social Responsibility and Headquarter Location from 2005 – 2015.
* **Presidential Election information** from 1952 – 2012 and Election Sponsorship information.

The unstructured data was available from USPTO and EPO in XML format and it was parsed using standard DOM Parser by running a MapReduce job. Since the size of the data was too large to handle by a standalone system, it was placed within Hadoop Distributed File System.

The Presidential Election information is collected from various sources from 1952 to 2012 along with the companies that funded the elections towards a specific political party. Also the Revenue and Competiveness information of these Fortune 500 companies were also gathered to do a deeper analysis.

**Data Aggregation**

The total Patent count collected from USPTO and EPO was aggregated for over 10 years. From this we found that the patents published by a company in a year and its previous year has high correlation. Since there was no drastic increment in each year’s annual revenue, this data is averaged for the period of 10 years. The companies are segregated into 12 major categories in order to perform specific analysis on each category.

The KLD data is divided into Corporate Responsibilities and Social Responsibilities, in each category we considered the positive and negative influence it had on the total CSR score. Based on the data aggregated above, we found a relationship between the location of the company, the company profile and its responsibilities.

**Data Analysis for Competitive Advantage**

Looking at the descriptive statistics of the aggregated data, IBM has published the highest number of patent published. On further analyzing it, IBM is focused on growing in its cloud computing, cognitive technology and data analytics. It spends 6.17% of their net revenue in R&D to develop their business and sales. Comparing with its competitors on Davenport and Harris analytical competitor continuum, they maintain a full score of 5 on a scale of 5.

Wal-Mart Corporation has the highest net revenue and the reason is they focus more on customer behavior analysis and social analytics. However, their CSR score is very poor when compared to its competitors such as Target Corporation and Dollar General. The reason being, Wal-Mart Corporation tends to focus more on customer behavior and satisfaction than innovation. They also maintain a full score on Davenport and Harris analytical competitor continuum, as they use predictive analysis to compete in the market.

When considering revenue and election results, during Democratic Period, Financial Sector including Commercial Banks and Engineering sector seems to have an increase in revenue. While in Republican’s period only Financial Sector enjoy more benefits. It is found that the successful contributions may have an increment in Patent Count or Revenue and this also depends on the type of industry the company belongs to.

**Conclusion**

After 2016 election whoever is elected as the President, Finance Sector will have an increase in revenue. However, Engineering sector can do better if Democrats come into power.

**APPENDIX:**

1. **Data Extraction and Data Wrangling**

Data pertaining to Fortune 500 companies that includes Patent Count, Revenue, Social Responsibility Score, Corporate Responsibility Score and Headquarter Locations were collected.

* **XML Parsing of US and European Patent Data using Hadoop MapReduce in Java:**

**Main Method:**

public static void main**(**String**[]** args**)** **throws** Exception **{**

//Define a new configuration file

Configuration conf **=** **new** Configuration**();**

conf**.**set**(**"xmlinput.start"**,** "<us-patent-grant"**);**

conf**.**set**(**"xmlinput.end"**,** "</us-patent-grant>"**);**

// Keep properties file in classpath when executing in eclipse

// Keep properties file in local path when executing in HDFS

File file **=** **new** File**(**"fortuneData.properties"**);**

Job job **=** Job**.**getInstance**(**conf**);**

//Distributed Cache

job**.**addCacheFile**(**file**.**toURI**());**

job**.**setJarByClass**(**XMLParsing**.**class**);**

job**.**setOutputKeyClass**(**Text**.**class**);**

job**.**setOutputValueClass**(**IntWritable**.**class**);**

job**.**setMapperClass**(**Map**.**class**);**

job**.**setReducerClass**(**Reduce**.**class**);**

job**.**setInputFormatClass**(**XmlInputFormat1**.**class**);**

job**.**setOutputFormatClass**(**TextOutputFormat**.**class**);**

FileInputFormat**.**addInputPath**(**job**,** **new** Path**(**args**[**0**]));**

FileOutputFormat**.**setOutputPath**(**job**,** **new** Path**(**args**[**1**]));**

job**.**waitForCompletion**(true);**

**}**

**}**

**Custom Class:**

public void initialize**(**InputSplit split**,** TaskAttemptContext context**)**

**throws** IOException**,** InterruptedException **{**

Configuration conf **=** context**.**getConfiguration**();**

**this.**startTag **=** conf**.**get**(**"xmlinput.start"**).**getBytes**(**"utf-8"**);**

**this.**endTag **=** conf**.**get**(**"xmlinput.end"**).**getBytes**(**"utf-8"**);** FileSplit fileSplit **=** **(**FileSplit**)** split**;**

**this.**start **=** fileSplit**.**getStart**();**

**this.**end **=** **(this.**start **+** fileSplit**.**getLength**());**

Path file **=** fileSplit**.**getPath**();**

FileSystem fs **=** file**.**getFileSystem**(**conf**);**

**this.**fsin **=** fs**.**open**(**fileSplit**.**getPath**());**

**this.**fsin**.**seek**(this.**start**);**

**}**

**Mapper Class: Initializing Map Class:**

public static class Map **extends**

Mapper**<**LongWritable**,** Text**,** Text**,** IntWritable**>** **{**

private static final IntWritable one **=** **new** IntWritable**(**1**);**

private Text word **=** **new** Text**();**

//MultipleOutputs<Text, IntWritable> mos;

static Set**<**Object**>** alist **=** **new** HashSet**<**Object**>();**}

**Map Method to define Map Functions:**

@SuppressWarnings**({** "unchecked"**,** "rawtypes" **})**

protected void map**(**LongWritable key**,** Text value**,** Mapper**.**Context context**)**

**throws** IOException**,** InterruptedException **{**

String document **=** value**.**toString**();**

document **=** document**.**replace**(**"&#x26;"**,** "dsba"**);**

Boolean flag **=** **false;**

**try** **{**

XMLStreamReader reader **=** XMLInputFactory**.**newInstance**().**createXMLStreamReader**(** **new** ByteArrayInputStream**(**document**.**getBytes**()));**

String currentElement **=** ""**;**

String propertyName **=** ""**;**

**while** **(**reader**.**hasNext**())** **{**

int code **=** reader**.**next**();**

**switch** **(**code**)** **{**

**case** 1**:**

**if** **(**currentElement**.**equalsIgnoreCase**(**"assignees"**))** **{**

flag **=** **true;**

**}else** **if(**currentElement**.**equalsIgnoreCase**(**"orgname"**)** **&&** flag**==true){**

propertyName **=** reader**.**getText**();**

propertyName **=** propertyName**.**trim**();**

propertyName **=** propertyName**.**toLowerCase**();**

**if(**propertyName**.**contains**(**"dsba"**)){**

propertyName **=** propertyName**.**replace**(**"dsba"**,** "&"**);**

**}**

Iterator iterator **=** alist**.**iterator**();**

**while(**iterator**.**hasNext**()){**

String temp **=** **(**String**)**iterator**.**next**();**

temp **=** temp**.**trim**();** **if(**propertyName**.**matches**(**temp**)){** **this.**word**.**set**(**properties**.**getProperty**(**temp**));**

context**.**write**(this.**word**,** one**);**

**}}**flag **=** **false;}break;**

**}}**

reader**.**close**();**

**}** **catch** **(**Exception e**)** **{**

**throw** **new** IOException**(**e**);}}**

**Reducer Class:**

//Reducer begins //

public static class Reduce **extends** Reducer**<**Text**,** IntWritable**,** Text**,** IntWritable**>** **{**

public void reduce**(**Text key**,** Iterable**<**IntWritable**>** values**,** Reducer**<**Text**,** IntWritable**,** Text**,** IntWritable**>.**Context context**)**

**throws** IOException**,** InterruptedException **{**

//Same as word count

int sum **=** 0**;**

**for** **(**IntWritable val **:** values**)** **{**

sum **+=** val**.**get**();**

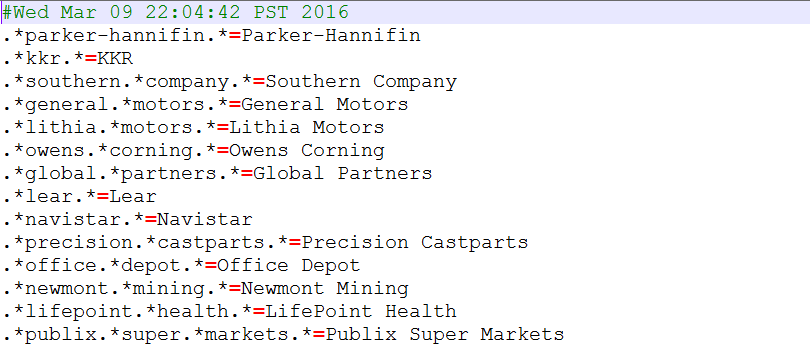
**}**

context**.**write**(**key**,** **new** IntWritable**(**sum**));**

**}**

**}**

* **Sample Properties File:**



* **Mallet Implementation:**

1.       Set the environment variable and give the path to the mallet directory

2.       Make text files for different companies and save them in the mallet folder as follows:

C drive->mallet->sample-data->web->create folder->save company.txt

3.       Type : bin\mallet import-dir --input pathway\to\the\directory\with\the\files --output tutorial.mallet --keep-sequence --remove-stopwords

This runs the mallet program and removes all the stop words from the text files

4.       At the command prompt in the MALLET directory, type:

bin\mallet train-topics  --input tutorial.mallet

This command opens your tutorial.mallet file, and runs the topic model routine on it using only the default settings. As it iterates through the routine, trying to find the best division of words into topics, your command prompt window will fill with output from each run

5.       This command is used to save the output in the mallet directory

bin\mallet train-topics  --input tutorial.mallet --num-topics 20 --output-state topic-state.gz --output-topic-keys company\_keys.txt --output-doc-topics company\_compostion.txt

The output is a series of paragraphs containing the keywords

* **Alchemy API Implementation in Java:**

class KeywordTest **{**

public static void main**(**String**[]** args**)** **throws** IOException**,** SAXException**,**

ParserConfigurationException**,** XPathExpressionException **{**

// Create an AlchemyAPI object.

AlchemyAPI alchemyObj **=** AlchemyAPI**.**GetInstanceFromFile**(**"api\_key.txt"**);**

String htmlDoc **=** getFileContents**(**"Microsoft.html"**);**

Document doc **=** alchemyObj**.**HTMLGetRankedKeywords**(**htmlDoc**,** "https://en.wikipedia.org/wiki/Microsoft"**);**

System**.**out**.**println**(**getStringFromDocument**(**doc**));**

**}**

private static String getFileContents**(**String filename**)**

**throws** IOException**,** FileNotFoundException

**{**

File file **=** **new** File**(**filename**);**

StringBuilder contents **=** **new** StringBuilder**();**

BufferedReader input **=** **new** BufferedReader**(new** FileReader**(**file**));**

**try** **{**

String line **=** **null;**

**while** **((**line **=** input**.**readLine**())** **!=** **null)** **{**

contents**.**append**(**line**);**

contents**.**append**(**System**.**getProperty**(**"line.separator"**));**

**}**

**}** **finally** **{**

input**.**close**();**

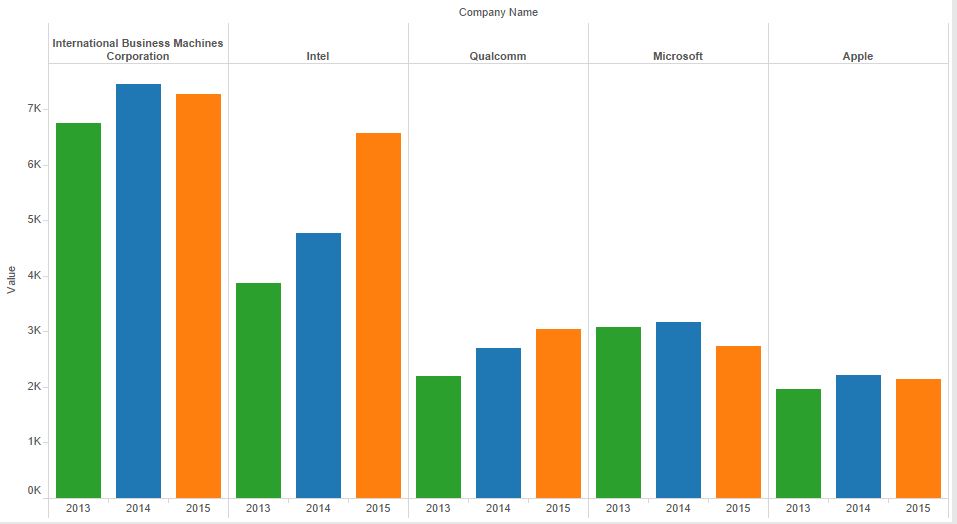
**}**

**return** contents**.**toString**();**

**}**

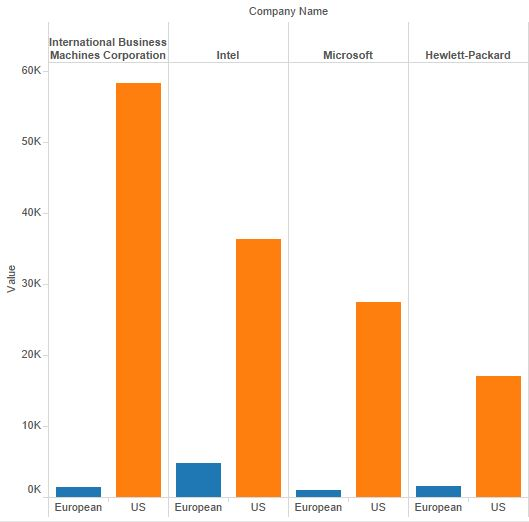
1. **Data Aggregation:**

* **Patent Published by Fortune 500 Companies**
  + - * Aggregate of the total number patents filed by each company is taken for over ten years. Figure I shows that number of patent filed by each company and it is seen that it is not consistent every year.



***Figure I* -­** U*S Patent 2013 - 2015*

* + The Patent published in EPO & USPTO differs significantly for example, Figure II shows the number of US patents published by IBM & Intel in Europe are 1367 and 4747 respectively. Therefore, the sum of patents file in both the patent office is considered for analysis.



***Figure II* -** *US vs EU Patent Count*

* + The average of the US and European patent count is not considered because many companies may not file patent every year which would result in a lower average when compared to the others.
* **Corporate Social Responsibility Data** 
  + The CSR data needs to be separated into Corporate Data and Social Data. Also most of the data were Boolean entries and some of the null and NR values were replaced with 0. Also some range conversion to convert all the scores into positive and negative entries. Finally, the data is consolidated as Social Score as one and the Corporate Score into five categories i.e. General, Economic behavior, Employee Benefits, Investor Specific and Innovation.
* **Fortune 500 Headquarter Location**
  + The category information and the headquarter location for each company is extracted from the Fortune website and the data is cleaned using the below formula. There were around 30 categories in the data which was reclassified into 12 categories.

**Formula to extract last word from a string**

=RIGHT(A2,LEN(A2)-FIND("\*",SUBSTITUTE(A2," ","\*",LEN(A2)-LEN(SUBSTITUTE(A2," ","")))))

**Consolidated score from scores of Positive & Negative Attributes in CSR**

=IF(S3=0,R3,IF(R3=0,0,IF(R3>=S3,AVERAGE(R3:S3),R3/2)))

**Convert a number scale to another scale by maintaining its ratio**

We used this formula when we had columns with values ranging from 0 to 37 and we had to bring it to the scale of o to 5.

OldRange = (OldMax - OldMin)

NewRange = (NewMax - NewMin)

NewValue = (((OldValue - OldMin) \* NewRange) / OldRange) + NewMin

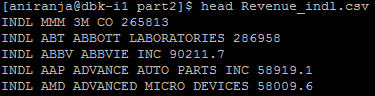
* **Segregate revenue for INDL & FS & aggregate for the period 2005-2015**

cat Revenue\_10.csv | grep "COMPANY NAME" | grep FS | awk -F "," '{ SUM += $5 } END { print $2,$3,$4,SUM }' >> Revenur\_fs.csv

cat Revenue\_10.csv | grep "COMPANY NAME" | grep INDL | awk -F "," '{ SUM += $5 } END { print $2,$3,$4,SUM }' >> Revenur\_indl.csv

Revenur\_10.csv has all the revenue information of fortune 500 for both the industry types FS & INDL from 2005 to 2014.

**Sample output**



* **VB Macro to extract location of all fortune 500 companies**

Sub Button1\_Click()

Dim rangeOfRow As Range

Dim rangeCopy As Range

'Range("B1:B4").Copy

RowCount = Range("A" & Rows.Count).End(xlUp).Row

Set rangeOfRow = Range("A1:A" & RowCount)

' For Each c In rangeOfRow.Cells

' If IsNumeric(c.Value) Then

' addressVal = Cells(c.Row, c.Column + 1).Address

' Range(addressVal).PasteSpecial

' End If

' Next

For Each c In rangeOfRow.Cells

'MsgBox Cells(c.Row, c.Column + 1).Value

'MsgBox Cells(c.Row, c.Column + 1).Value = 0

If Cells(c.Row, c.Column + 1).Value = 1 And Cells(c.Row + 1, c.Column + 1).Value = 1 Then

Cells(c.Row, c.Column + 4).Value = Cells(c.Row, c.Column).Value

If Cells(c.Row + 1, c.Column + 1).Value = 1 Then

Cells(c.Row, c.Column + 5).Value = Cells(c.Row + 1, c.Column).Value

End If

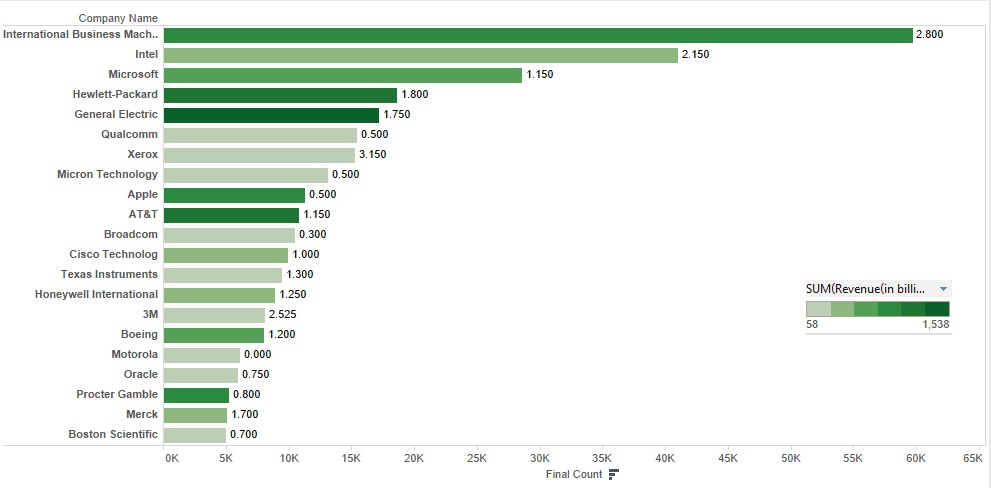
End If

Next

End Sub

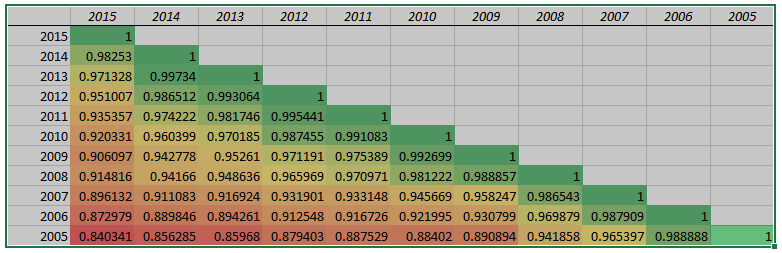
1. **Data Analysis**

* **Patent Published and Revenue**
  + The number of patents and the revenue does not seem to have a direct relationship, however there are some companies like General Electric which has a good patent count and revenue. This information can be better explained in *Figure III.* The length of the bar shows the total revenue of fortune 500 and the value next to the bar is its Corporate Responsibility Score.



***Figure III –*** *Patent and Revenue Relation*

* + There is a high correlation between the category information and the headquarter location for each company where it is extracted from the Fortune website.



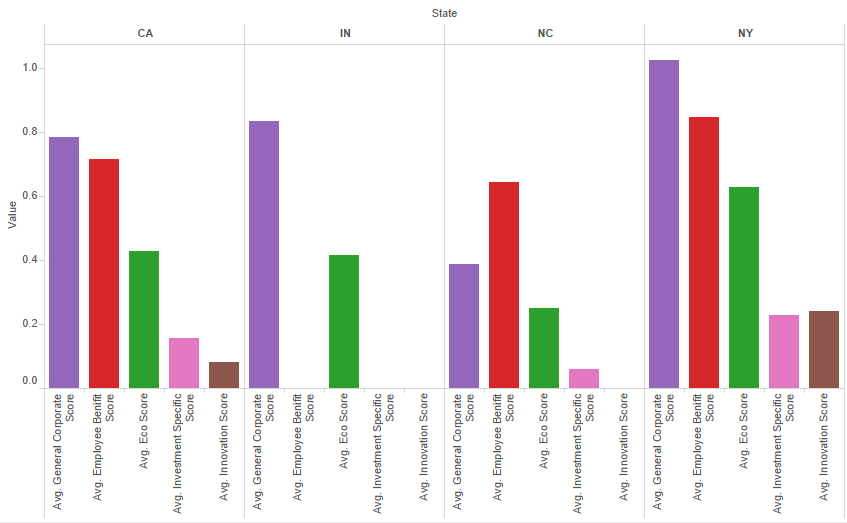
­­­***Figure IV –*** *Correlation Table*

* **Business Entity, Revenue and Headquarter Location**
  + From our analysis we found that, New York and California holds more number of companies from four major industries which also contributes high in revenue. Visualizing their Corporate Score, New York stands high in responsible behavior when compared to other states. The results are summarized in the below pivot table.

|  |  |  |
| --- | --- | --- |
| **Company Type** | **State** | **Revenue (Average)** |
| Computer and Information Services | NY | $ 984.57 |
| Electrical and Electronics | NY | $ 425.90 |
| Engineering and Machinery | CA | $ 1,055.88 |
| Finance and Commercial Banks | NC | $ 2,288.38 |
| Food and Retail Consumer Products | CA | $ 630.23 |
| Health Care, Insurance and Pharmaceuticals | IN | $ 1,224.01 |

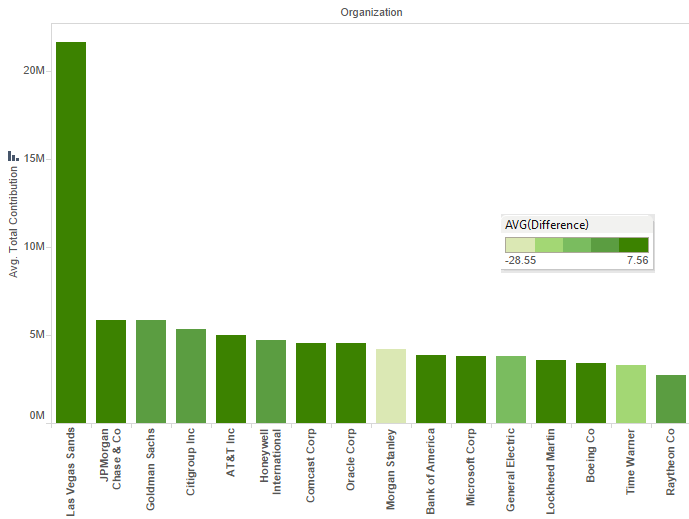
***Figure V*** *– Pivot Table*

* + *Figure VI* shows the corporate responsibilities of the companies in the state mentioned in *Figure V*. The graph indicates that New York has taken serious steps in improving their corporate responsibilities. It also shows that companies in North Carolina and Indiana has to do some serious work in maintaining their corporate responsibilities.



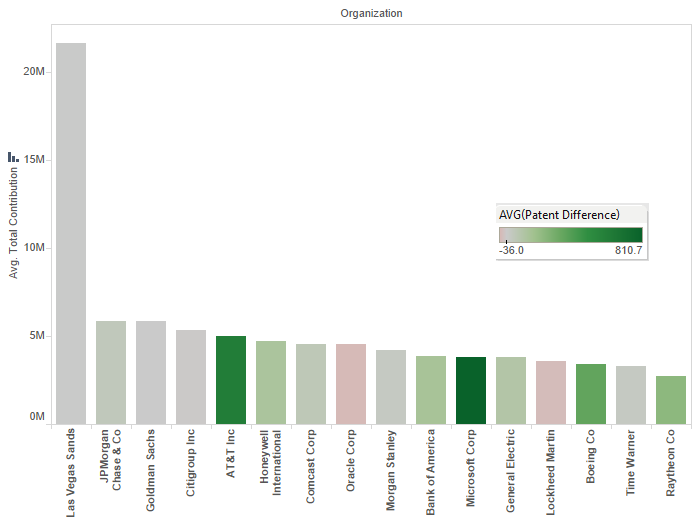
***Figure VI –*** *Average Corporate Scores on various states*

* **Revenue and Elections**
  + The contribution of the organizations towards the election has some effect on the revenue after the election. But there are some contributors like Las Vegas Sands which does not have a shift in revenue after the election as shown in *Figure VII.*

****

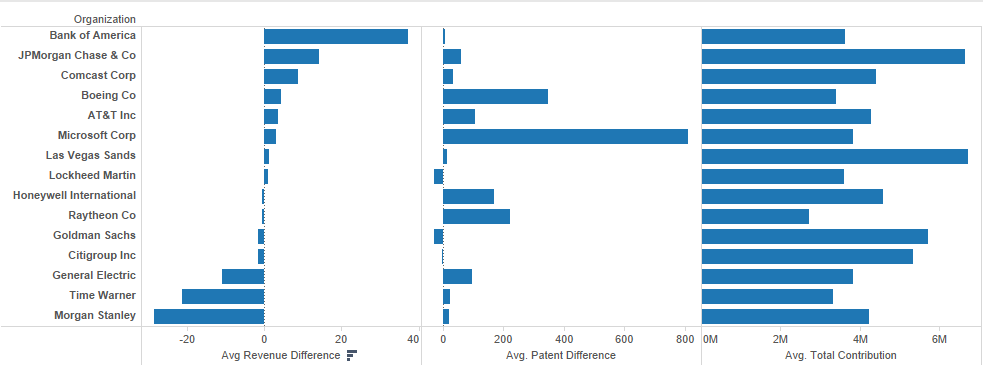
***Figure VII –*** *Contributions and After Effects in Revenue*

* + There is also some effect in patents published by the companies after successful contribution to the elections. *Figure VII* shows the change in patent count after election by many organizations that have contributed for the elections campaign.

****

***Figure VIII –*** *Contributions and After Effects in Patent Count*

* + Companies that have contributed for the successful party has a considerable increase in revenue or patent count or both. This can be better explained in *Figure IX;* in the graph you can see that Microsoft has had an increase in patent count after successful contributions. Similarly Bank of America has an increase in annual revenue as a result of successful contributions.

****

***Figure IX*** *– Outcome after Successful Contributions*

* **Relationship among CSR Scores and Patent Count**

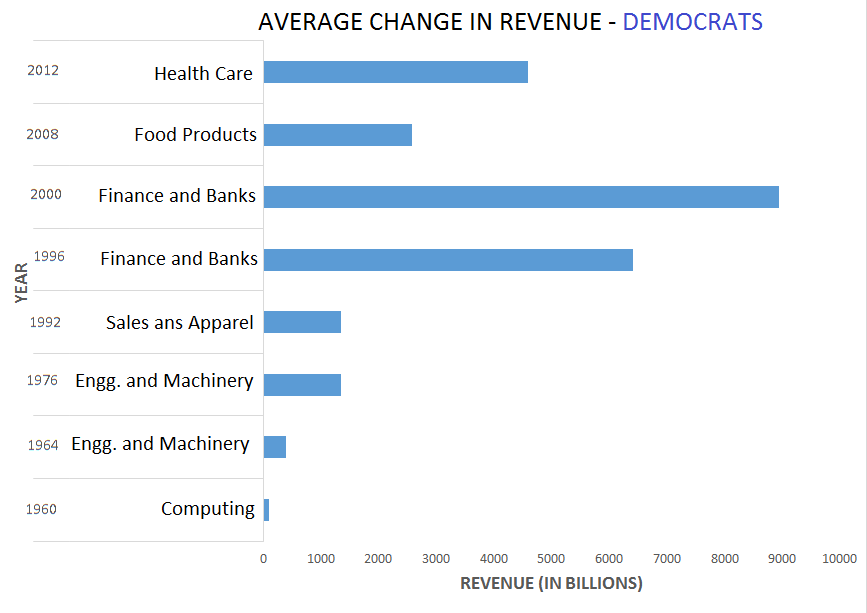
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
| 1 | (Constant) | 79.875 | 22.298 |  | 3.582 | .000 |
| Final Count | .004 | .004 | .037 | .819 | .413 |
| Social Score | -73.092 | 28.952 | -.119 | -2.525 | .012 |
| Corporate Score | 257.583 | 91.181 | .228 | 2.825 | .005 |
| Employee cons | 205.130 | 47.554 | .312 | 4.314 | .000 |
| a. Dependent Variable: 2005-2015 | | | | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summary** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .497a | .247 | .240 | 356.6439 |
| a. Predictors: (Constant), Employee cons, Social Score, Final Count, Corporate Score | | | | |

* + Patent count is not statistically significant with the Revenue.
  + Social Scores have a negative impact in the Revenue, therefore companies might tend to avoid being socially responsible.
  + Corporate Score, Innovation and Employee Benefit Score has a positive coefficient in the regression with the Revenue.
* **Cluster Analysis**
  + Clustering of Social and Corporate Scores along with the business entity of the fortune 500 companies gave fruitful result.
  + Based on the cluster output, these companies were grouped into Responsible, Swing and Irresponsible groups.

|  |  |  |  |
| --- | --- | --- | --- |
| **CLUSTER** | **IRRESPONSIBLE** | **SWING** | **RESPONSIBLE** |
| **PROFILE** | **1** | **2** | **3** |
| Apparel and Sales | 146.7504917 | 167.5288 | 517.9270295 |
| Chemicals and Drug | 92.50213215 | 306.8911 | 221.2582626 |
| Computing | 101.1762586 | 557.9086 | 537.9694708 |
| Engineering and Machinery | 96.52497925 | 421.0325 | 250.4423682 |
| Finance and Banks | 140.8683099 |  | 1061.119667 |
| Food Products | 160.5847444 | 131.4233 | 299.5309695 |
| HealthCare | 245.1959155 | 497.347 | 454.4258627 |
| HouseHold | 70.6413415 |  |  |
| Miscellaneous | 95.14471968 | 153.4511 | 178.931519 |
| Oil, Minerals and Energy | 179.8369298 | 205.4271 | 630.6325339 |
| Real Estate | 85.86601957 |  |  |
| Transportation | 110.469451 | 306.7242 | 296.1088636 |

* **Steps followed in finding Insights from Election**
  + Got revenue for Fortune 500 companies for a time frame of almost 60 years. The data is cleaned in excel and this is combined with the profile and names of the Fortune 500 list. Then we managed to segregate the period when democrats were on power and similarly for republicans.
  + The revenue before and after election are taken for one year before and after the election, because most of the dealings takes place immediately after election. Then we made a pivot table of each profile during each term before and after the election.

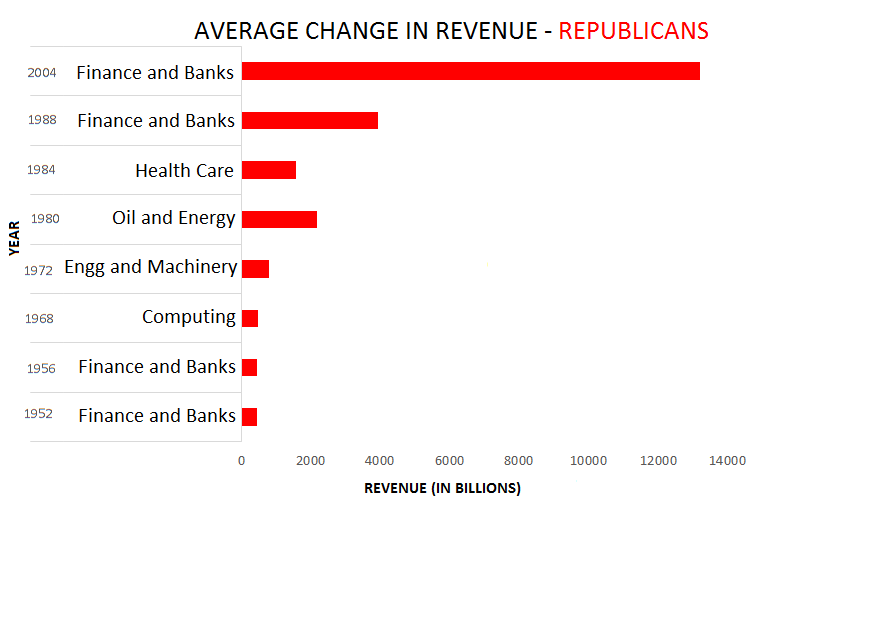


***Figure X –*** *Shift in Revenue for Each Business Entities During Democrats*

* + Then the profile with the highest shift in revenue for each term is found using the formula:

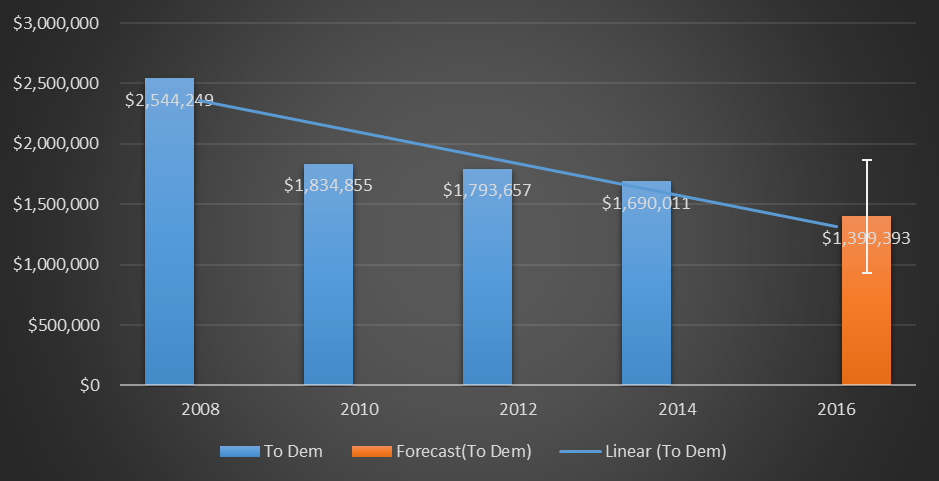
=INDEX (B$1:M$1, MATCH (MAX (B3:M3), B3:M3,))

* After finding the type of firm with the highest shift, we plotted a clustered graph on the obtained results to visualize the findings. The graphs which show these details are:



***Figure XI –*** *Shift in Revenue for Each Business Entities During Republicans*

* **Forecast of Contribution to Democrats by Fortune 500 for 2016**
  + The forecasted contribution to Democrats by the fortune 500 companies who funded them is predicted to be ***$1,399,393*** for the year ***2016***.

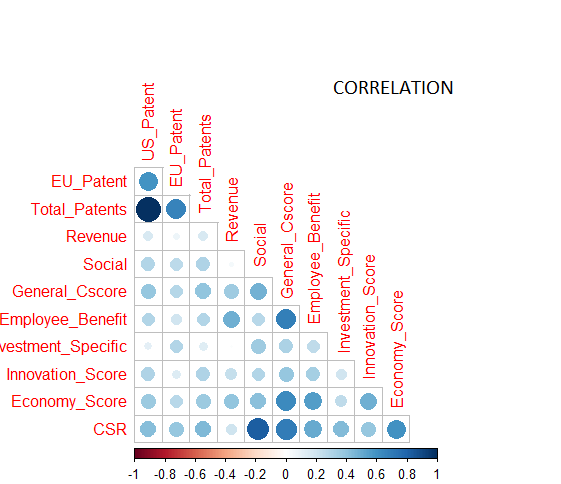
****

***Figure XII*** – Funding to Democrats Forecast for 2016

* **Forecast of Contribution to Republican by Fortune 500 for 2016**
  + The forecasted contribution to Republican by the fortune 500 companies who funded them is predicted to be ***$5,591,845*** for the year ***2016***.

***Figure XII*** – Funding to Republican Forecast for 2016

* **Correlation Between all the Attributes:**
  + There is a correlation among variables within the corporate score. However, there is not much influence on the Patent Count and Revenue when considering all the Fortune 500 companies.



***Figure XIII*** – Correlation between all the Entities

1. **Business Insights & Strategies**

* **International Business Machines**
  + It can be seen from *Figure XIV* that though IBM stands first in patent count, Microsoft Corporation spends more on R&D when compared.
  + IBM is known for providing cognitive solutions using its product - IBM Watson. Now, IBM is focused on growing in its cloud computing, data analytics, security and mobile businesses.
  + The company’s next chapter is ushering in an entirely new era of human-organization-computer interaction-embodied in Cognitive Solutions and the Cloud Platform
  + As per the most recent Annual Report, Hybrid is the fastest growing segment in the cloud market and IBM is the global leader in Hybrid cloud for the enterprise.
* **Wal-Mart Corporation**
* Though Walmart has the highest revenue, it maintains a poor Corporate and Social Responsibility score, when compared with its competitors Target and Dollar General.



* + Information Technology has long been regarded as a core strength that enabled Wal-Mart to reduce costs and improve operational efficiency.
  + The realization of the powers of predictive analysis led to the creation of Walmart Labs
  + When considering their business strategy, Walmart believes in the power of social analytics in analyzing their customer behavior. The development of projects like Social Genome and ShoppyCat predictive analysis tools to provide more customized services to its customers shows that Walmart is betting on social media analytics for competitive advantage