**Big Data Analytics for Competitive Advantage**

ITCS/DSBA/ITIS-6100

Technical Project Report

By

**Group 4**

Bhavya Devineni

Kalyan Purushotham Karra

Nevetha Balasubramanian

Praveen Kumar Govindaraj

Raghunath Viswanath Nallapeta

Sridevi Rachakulla

Sushma Venkatesh Reddy

Vishwanarayan Sriganesh

**Process Flow:**

1. We collected the Election data from year 1952 and Patent data of the last eight years from below mentioned respective sources. extracting data for the initial clean up and loading process.

* Election Data - <http://www.opensecrets.org/>
* Patent Data of Fortune 500 companies:

/users/wzadrozn/dsba-6100/patentData2000\_2015/patBiblio2000\_1h2015 on the DSBA cluster

http://www.uspto.gov/ (US patent data)

https://bulkdata.uspto.gov/data2/patent/officialgazette/ (Year wise)

* K-10 Reports - http://www.sec.gov/Archives/edgar/data/
* CSR Data - www.compustat.com

1. Extraction and downloading of the data from the sources is carried on in UNIX using shell scripting, FileZilla, Winscp tools and basic UNIX commands (wget, grep, cat, cp etc.)
2. Conversion of the html files and xml files into text files was done using J2EE JAVA programming language
3. For pattern Checking, Redundancy removal and formatting, we have used scripting on AIX based UNIX platform.
4. Converted xml files are loaded into the Hadoop cluster and we have applied Map reduce Algorithm over the data for further classification, grouping and counting.
5. The output of the Map-Reduce functions is fetched into the HIVE tool for the purpose of formatting, and bringing the data into a table and easily readable format using HQL queries and other querying languages
6. Topic Modelling using Mallet has been done for Clustering analysis of the K-10 reports for all the fortune 500 companies to know the area of focus of the organizations. Importance was given to find keywords, “Analytics and Innovation” from reports.
7. Regression on revenue, patent and election data using SPSS.
8. Clustering on CSR data using R and excel.
9. Topic Modeling on K10 reports using Mallet.
10. Visualization using R and Tableau.

**Observations:**

* IBM being the company with highest patent count, 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 with the highest revenue is Walmart and known for its sales and customer service. From the word cloud it's clear that it focuses more on service, products and customers which makes it to the top of revenue. Huge Revenue Capital makes Walmart feature top amongst all the Private employers in U.S and Mexico and the top most in Canada
* We have selected these two companies for our analysis to know how top performing companies get affected with the elections
* As per the Analysis and observations made from this phase of the project, Presidential elections for the years 1952 to 2012 does not have significant effect or impact or influence towards the fortune 500’s Revenue and Patent count.

**Appendix:**

**Data Downloading Steps:**

1.Login to dsba-hadoop.uncc.edu  
2.gotothelink" https://bulkdata.uspto.gov/data2/patent/officialgazette/2010/ "  
3. select Patent Official Gazettes (JAN 2013 - DEC 2013) ( year as per what is allocated to you)  
4. You will end up in the web page : " https://bulkdata.uspto.gov/data2/patent/officialgazette/2013/ "  
5. Now type the command " cd /users/<your user name>  
6. Now type the command " wget “https://bulkdata.uspto.gov/data2/patent/officialgazette/2013/  
e-OG20130101\_1386-1.zip”  
7. Repeat this step for all the zip files present in the web page according to the year dedicated to you.

**Steps to follow while parsing patent data xml**

* Convert xml file to txt file containing only required tags like sponsor, patent id, date and country using java.output.txt

* Load output file to hdfs (/users/rnallape/Data).

* Converting all characters to lowercase: tr '[:upper:]' '[:lower:]' <input.txt> <output.txt>

* Create table with required fields and load data into it using the below mentioned link (here Raghu\_table): <http://doc.mapr.com/display/MapR/Getting+Started+with+Hive>. Then follow the queries steps:

1. CREATE TABLE raghu\_table(viewTime INT, userid BIGINT, url STRING, referrer STRING, ip STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY '|';

2. LOAD DATA LOCAL INPATH '/home/mapr/java.output.txt' INTO TABLE raghu\_table;

* Then create other table with same fields as above table to store the distinct records (here

patent\_table).

* insert overwrite table patent\_table select distinct country, pid, from\_unixtime(unix\_timestamp(pdate,'yyyyMMdd'),'yyyy-MM-dd'),regexp\_replace(company,',','') from raghu\_table;
* insert overwrite table Patent\_Table4 select country, pid, pdate, regexp\_replace(regexp\_replace(sponsor,"&#x26\;","")," ll","") from patent\_table;

Query to select count of patents sponsored by each company:

Select count(pid), sponsor from patent\_table group by sponsor;

To load multiple data files into one table:

* Similar steps to be followed to produce final table named after “Patent\_Table4”

**Java Code Used for Parsing:**

import [java.io](http://java.io/).\*;

public class xmlParser {

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

{

int flag =1,date=1,country=1,docnum=1;

BufferedReader br=null;

br=new BufferedReader(new FileReader("input2.xml")); //input xml file

FileWriter out= new FileWriter("output2.txt"); // output xml file

String line;

String line2 = null;

while((line =br.readLine())!=null)

{

if(line.contains("<?xml version"))

{

flag =1;

date =1;

country =1;

docnum=1;

line2 =null;

}

if(line.contains("<country>")) //fetching country tag in input xml file

{

if(country ==1)

{

line2 = (line.substring(9,line.length()-10)+"|");

country =0;

}

}

if(line.contains("<doc-number>")) //fetching doc-number tag in input xml file

{

if(docnum ==1)

{

line2 = line2 + ((line.substring(12,line.length()-13)+"|"));

docnum =0;

}

}

if(line.contains("<date>")) //fetching date tag in input xml file

{

if(flag==1)

{

line2 = line2 + (line.substring(6,14)+"|");

flag =0;

date =0;

}

}

if(line.contains("<orgname>") && date ==0) //fetching organization name tag in input xml file

{

out.write(line2 + line.substring(9,line.length()-11) +"|"+"\n");

}

}

br.close();

out.close();

}

}

**List of R commands used for clustering and scatter plot**-

library("cluster", lib.loc="C:/DSBA/R/R-3.2.3/library")

library("ggplot2", lib.loc="C:/DSBA/R/R-3.2.3/library")

install.packages("hclust")

z<- agnes(distance,method="average")

plot(z,ask=TRUE)

cluster<- cutree(z,k=4)

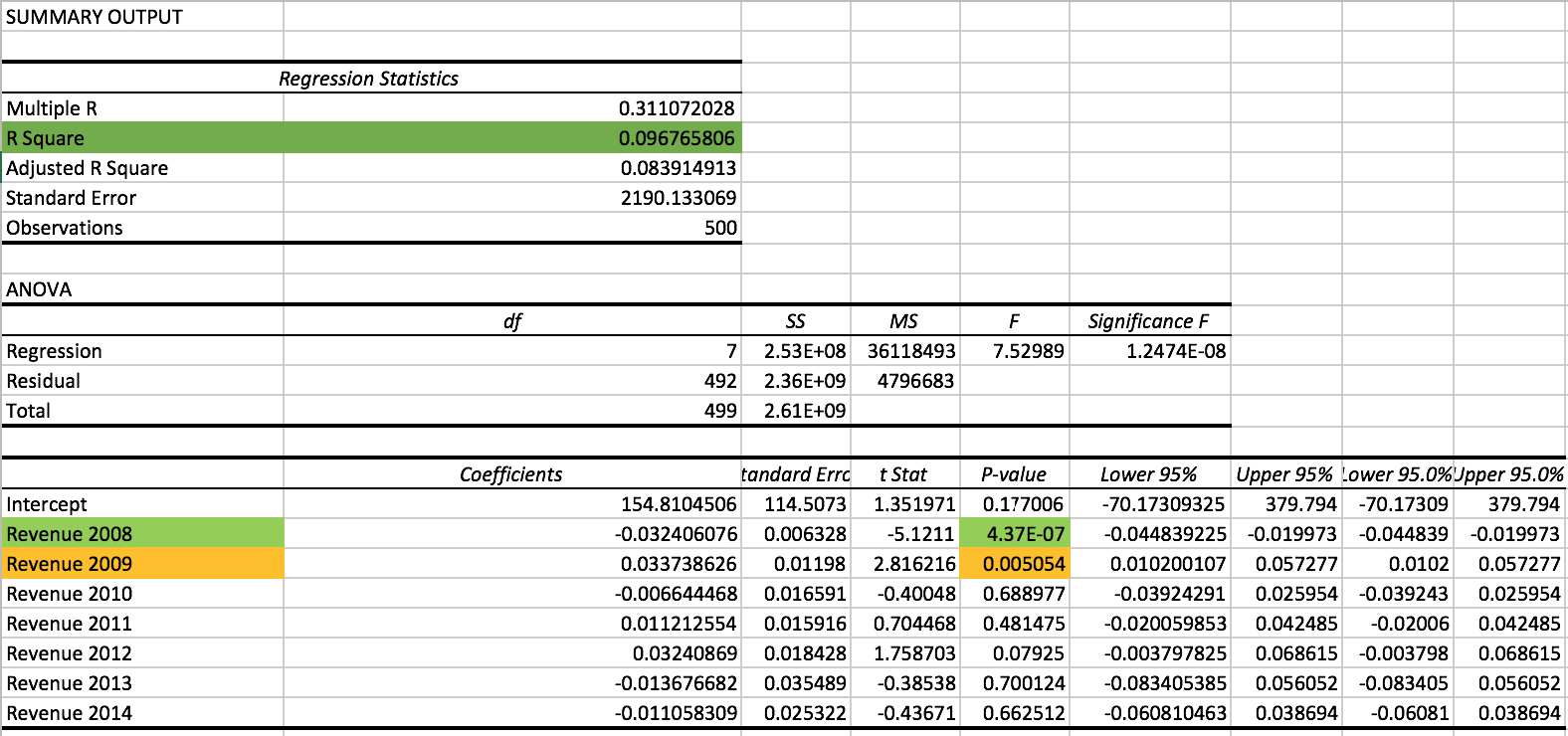
xy <- data.frame(cmdscale(dist(csr,method="euclidean")), factor(cluster))

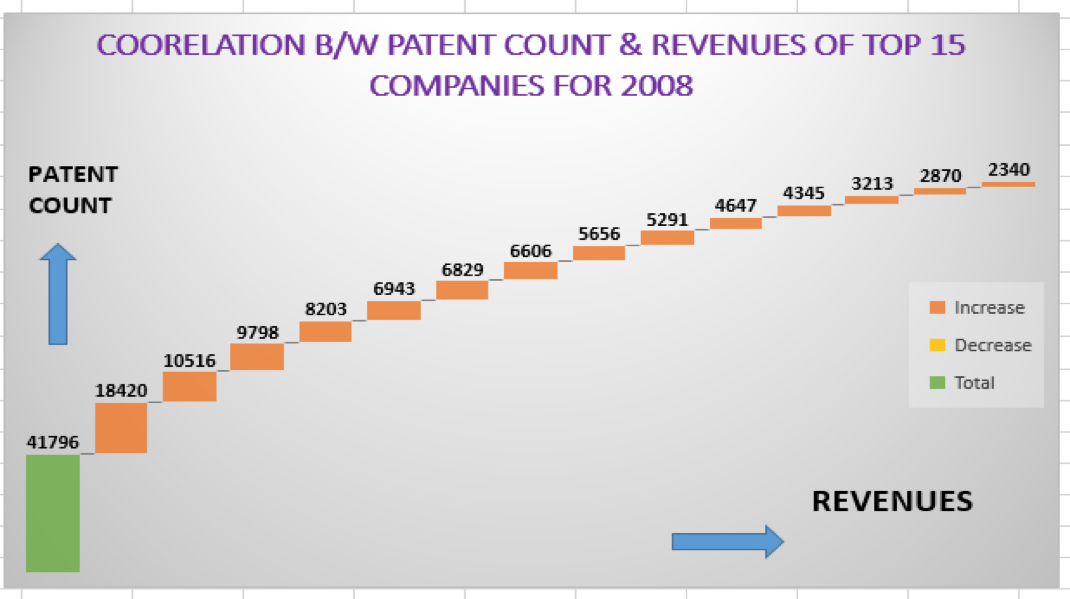
xy$factor.cluster. <- factor(xy$factor.cluster.,levels = c(1,2,3,4), labels=c("Top Contributers","Least Contributers","Moderate Contrubuters","Low Contributers"))

names(xy) <- c("companies", "CSRweigtage", "clusters")

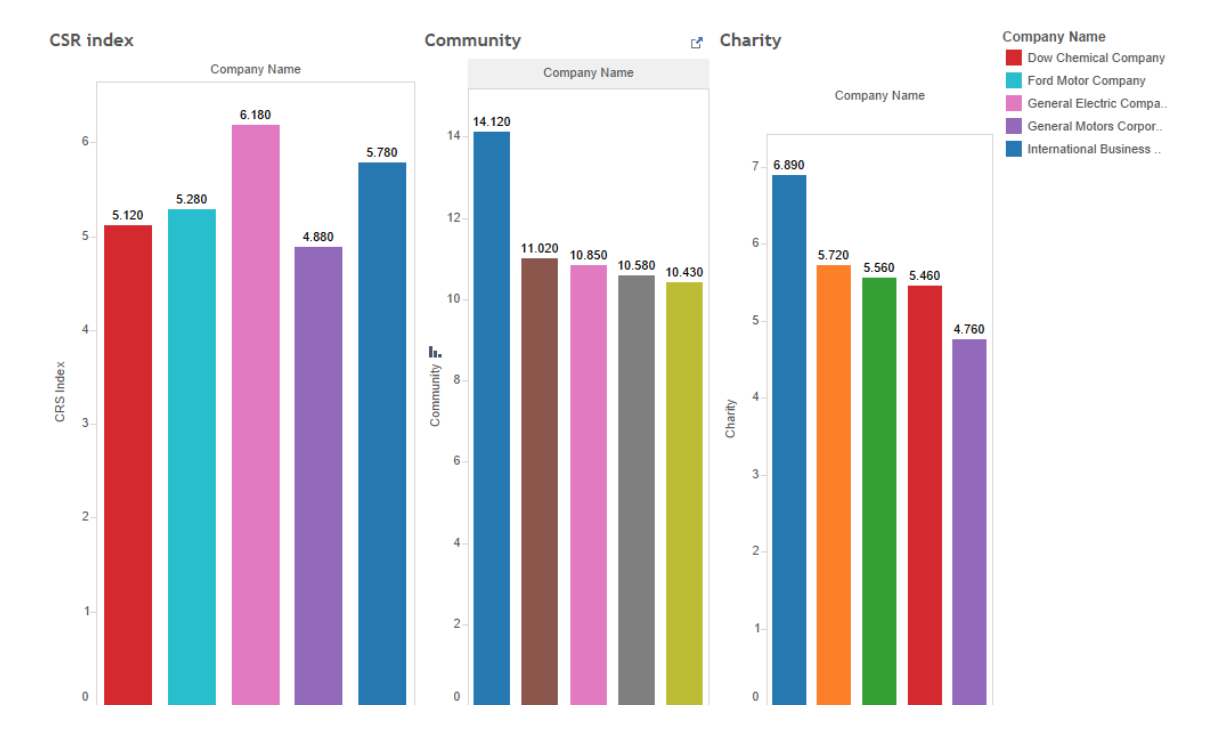
ggplot(xy, aes(companies, CSRweigtage)) + geom\_point(aes(colour=clusters), size=3)

**ANOVA Regression:**

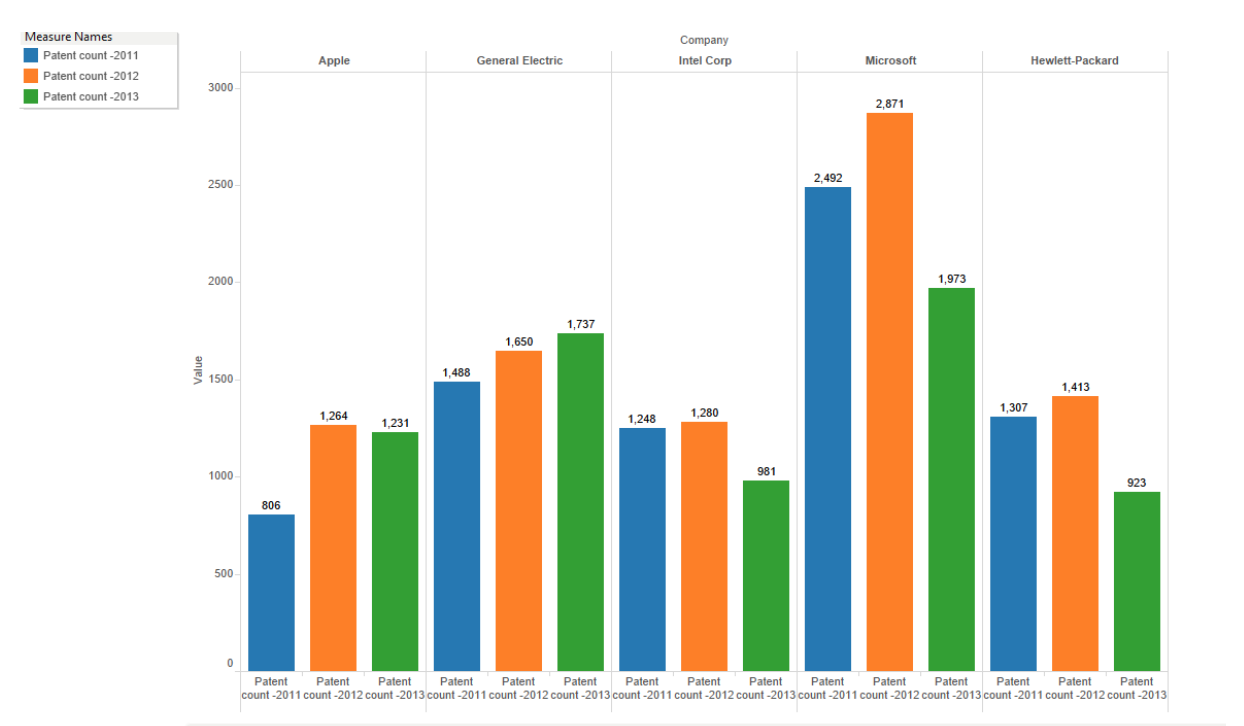




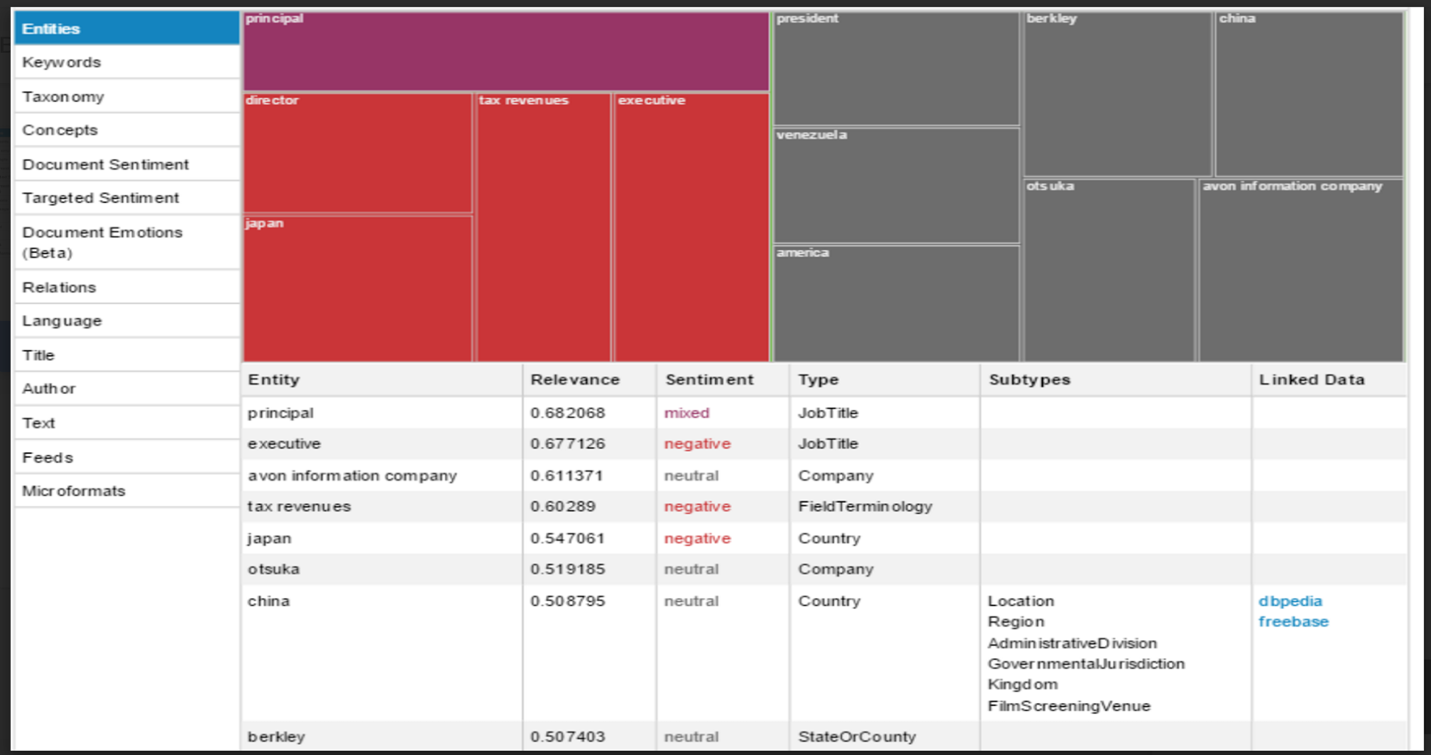
**Corporate Social Responsibility:**



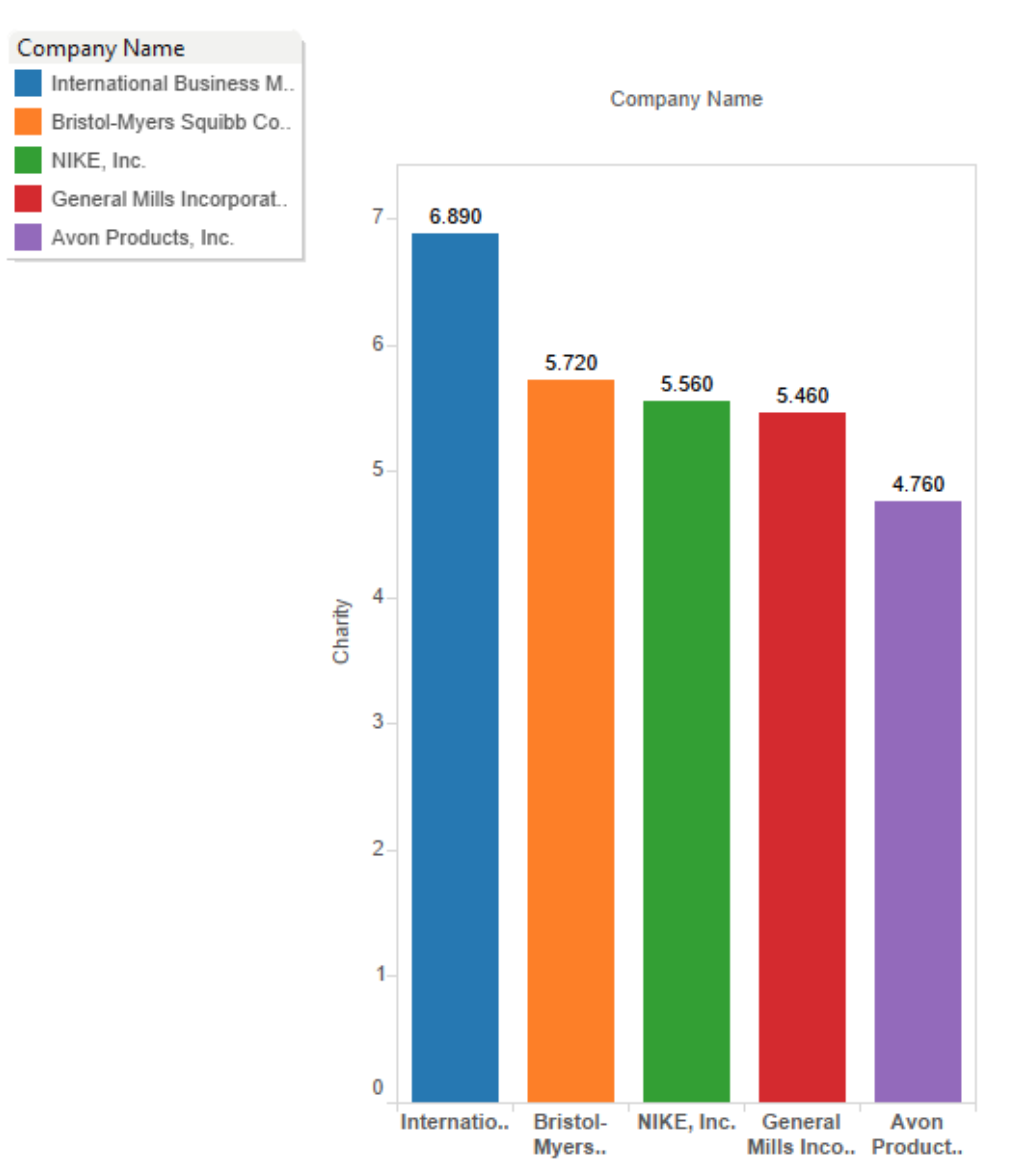
**Patent count of top 5 companies before and after election 2012:**



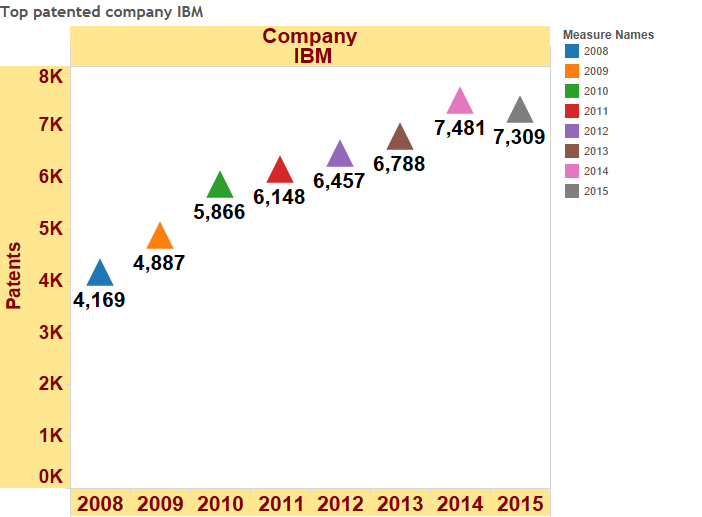
**Mallet output on data analyzed:**



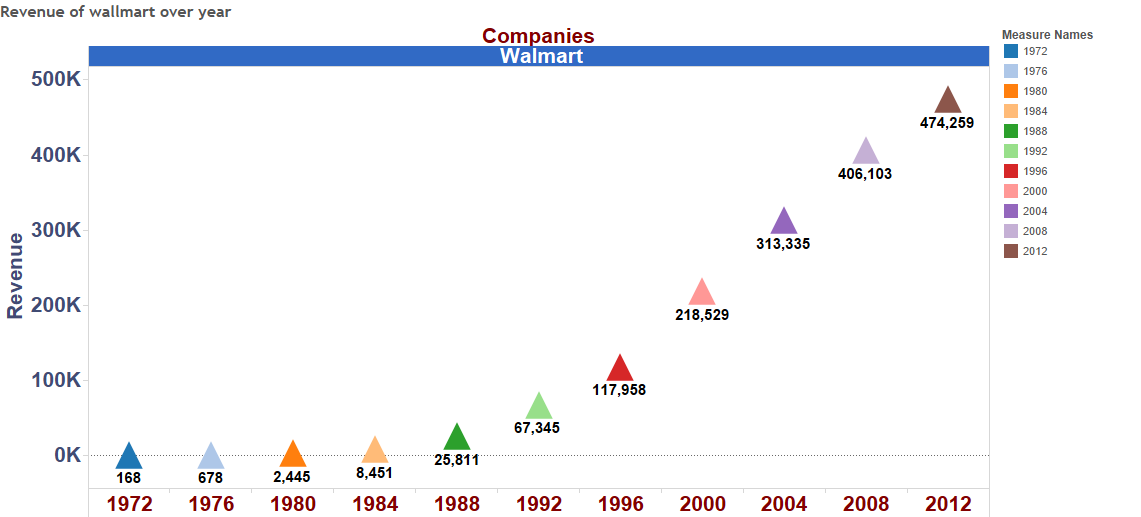
**Top 5 fortune 500 companies that has contributed to CSR:**



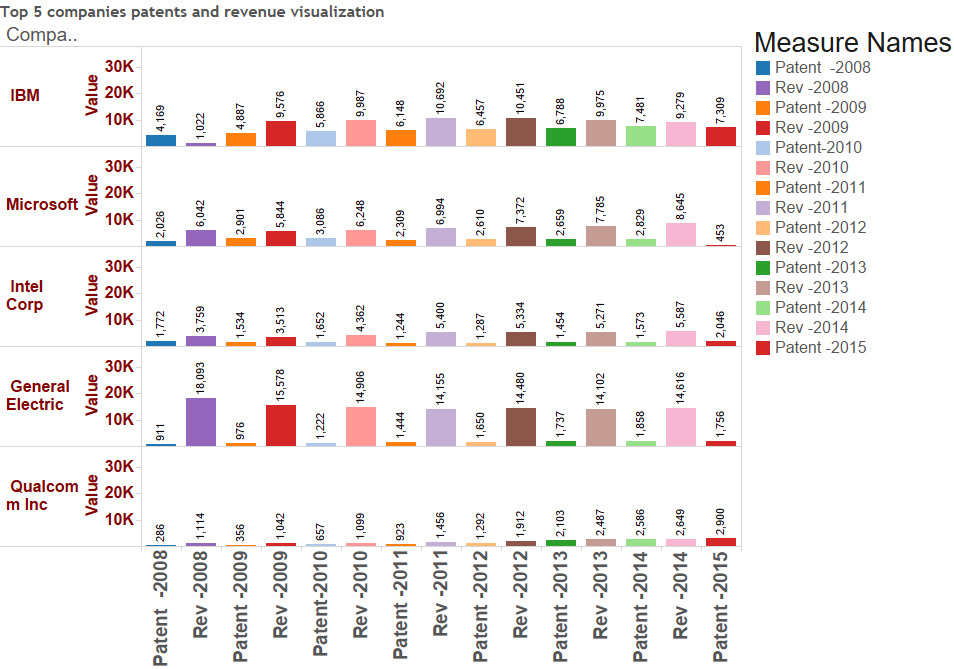
**IBM revenue over the years**

****

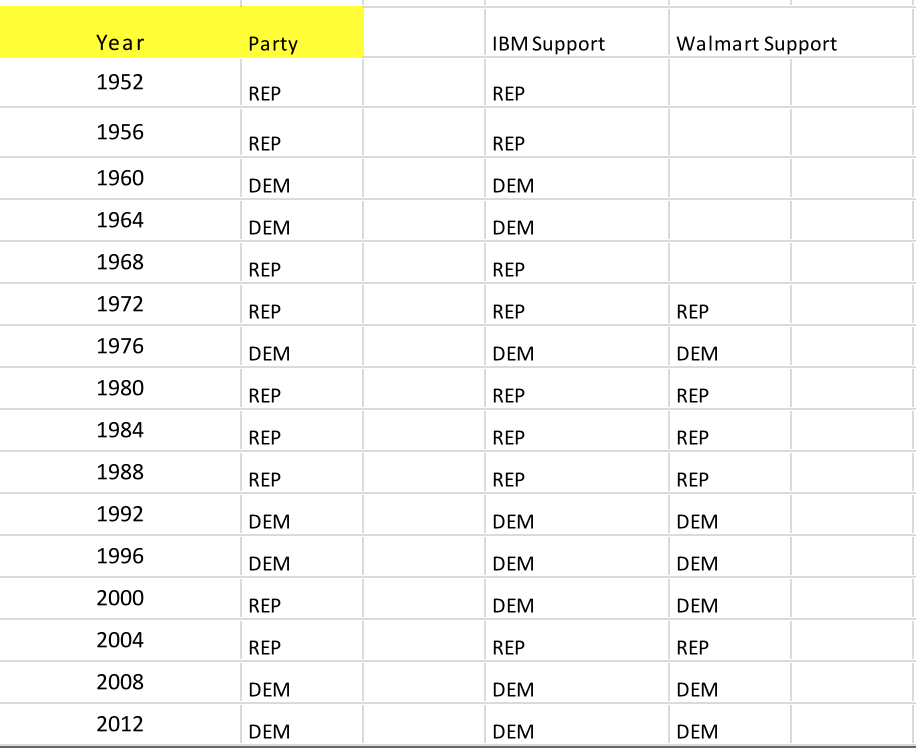
**Walmart revenue over the years**

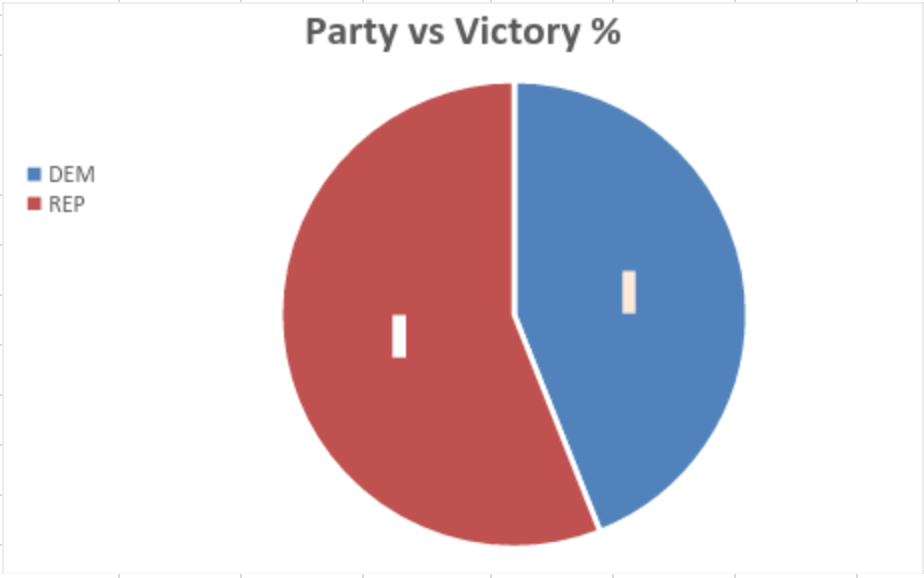
****

**Top 5 Companies Revenue and their patent count**

****

**Election Winners with support of IBM and Walmart**





Overall win percentage of Democratic and Republican party from 1952 to 2012.