CMPE 272 Sec 02 – Enterprise SW Platforms – Extra Assignment

Name: Shounak Gujarathi
SJSUID: 010728939

Topic: Use 2012 presidential donation datasets to find out who donated to presidential candidates most and if there is any correlation.

- ➔ The dataset consisted of CSV files containing donations given to 3 candidates, viz. Buddy Roemer, Gary Jhonson and Jill Stein.
- ➔ I've done complete analysis on data related to Buddy Roemer and Gary Jhonson

Technologies used:

- ➔ Apache Spark Service for Bluemix, IPython Notebook
- ➔ Libraries used Pandas, numpy, matplotlib, urllib2

Steps Taken:

- ➔ I converted the CSV data from CSV format to RDD (Relation Data Structure) format native to apache spark.
- ➔ My next aim was to convert this data to Apache spark data frame. But for this, the data needed to be formatted and filtered.

```
dataFile = sc.textFile("swift://notebooks.spark/JohnsonSubmission1.csv,swift://notebooks.spark/JohnsonSubmission2.csv")
dataFile.count()

dataFile2 = sc.textFile("swift://notebooks.spark/JohnsonSubmission5-6.csv,swift://notebooks.spark/JohnsonSubmission7.csv")
dataFile2.count()

dataFile3 = sc.textFile("swift://notebooks.spark/RoemerSubmission1.csv")
dataFile3.count()
```

```
#fields

fields2 = [StructField(field_name, StringType(), True) for field_name in fields_temp2]
#fields2

fields3 = [StructField(field_name, StringType(), True) for field_name in fields_temp3]
fields3
```

```
[StructField(Prefix,StringType,true),
 StructField(FirstName,StringType,true),
 StructField(MI,StringType,true),
 StructField(LastName,StringType,true),
 StructField(Suffix,StringType,true),
 StructField(Address1,StringType,true),
 StructField(Address2,StringType,true),
 StructField(City,StringType,true),
 StructField(State,StringType,true),
 StructField(Zip5,StringType,true),
 StructField(Employer,StringType,true),
 StructField(Occupation,StringType,true),
 StructField(DonationDate,StringType,true),
 StructField(Amount,StringType,true),
 StructField(AmountSubmittedForMatching,StringType,true),
 StructField(TotalThisSubmission,StringType,true),
 StructField(TotalAllSubmission,StringType,true),
 StructField(,StringType,true)]
```

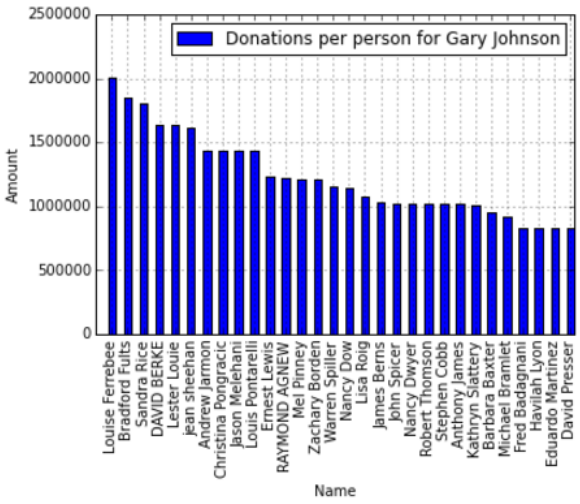
```
data_temp = dataNoHeader.map(lambda k: k.split(",")).map(lambda p: (p[0], p[1], p[2], p[3], p[4], p[5], p[6], p[7], p[8], p[9], p[10].replace(" ","").replace("","0"

data_temp2 = dataNoHeader2.map(lambda k: k.split(",")).map(lambda p: (p[0], p[1], p[2], p[3], p[4], p[5], p[6], p[7], p[8], p[9], p[10], p[11], p[12].replace(" ","

data_temp3 = dataNoHeader3.map(lambda k: k.split(",")).map(lambda p: (p[0], p[1], p[2], p[3], p[4], p[5], p[6], p[7], p[8], p[9], p[10], p[11], p[12], p[13].replace
```

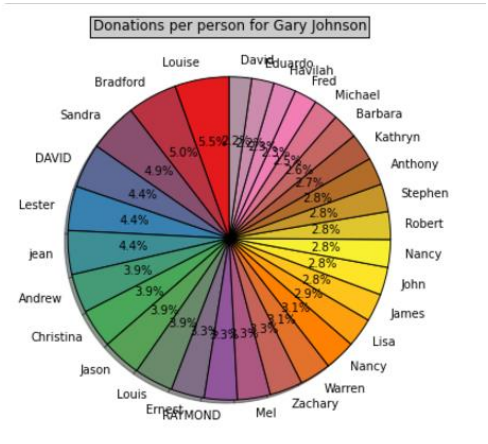
```
data_df = sqlContext.createDataFrame(data_temp, schema)
data_df.show();

data_df2 = sqlContext.createDataFrame(data_temp2, schema2)
data_df2.show();

data_df3 = sqlContext.createDataFrame(data_temp3, schema3)
data_df3.show();
```

- ➔ This was done using a series of functions, where headers were mapped, renamed and the data was made consistent. This eventually gave me 3 different files, data_df and data_df2 for Gary Jhonson and data_df_roemer for Buddy Roemer
- ➔ This is now structured data that I can make use of. I created temporary data tables from this data, to then do sql analysis on it.

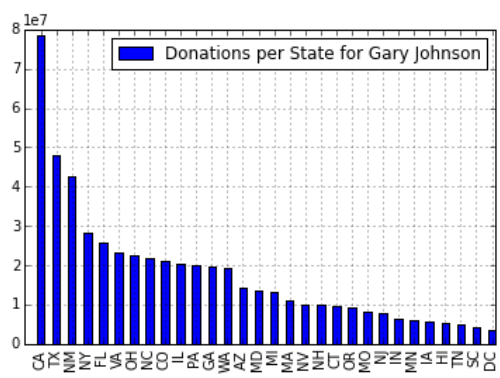
```
filter_df=sqlContext.sql("SELECT Null as Prefix, First_Name, Middle_Name, Last_Name, Null as Suffix, Address, City, State, Postal_Code, Employer, Occupation, Donat
filter_df.registerTempTable("Filter_Data")

filter_df_Roemer=sqlContext.sql("SELECT * FROM Elec_Data_Roemer where Total_All_Submissions REGEXP '[^0-9]' and Total_This_Submission REGEXP '[^0-9]' and Amount_Sul
filter_df_Roemer.registerTempTable("Filter_Data_Roemer")
```

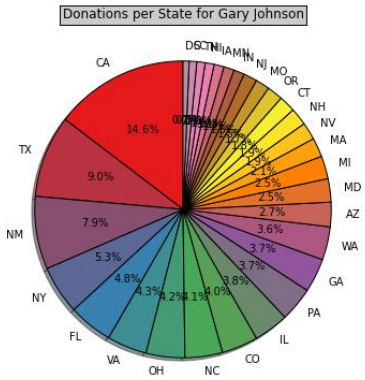
- ➔ For Jhonson the data table is Filter_data and for Roemer its Filter_data_roemer
- ➔ Then based on these data tables I was able to perform analytics based on different attributes, like, names, zip, city, state and occupation. Here are some of the graphs that were generated.



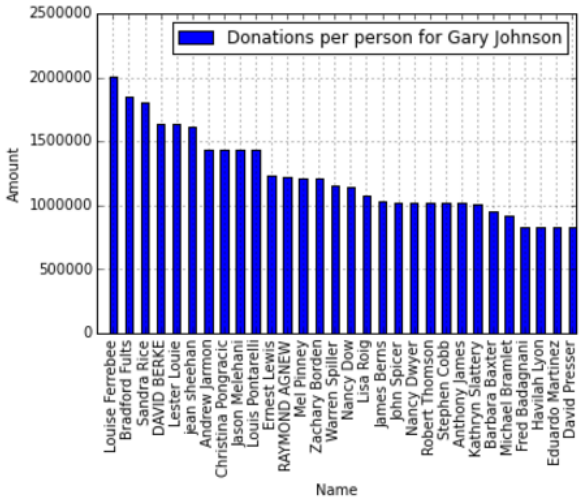


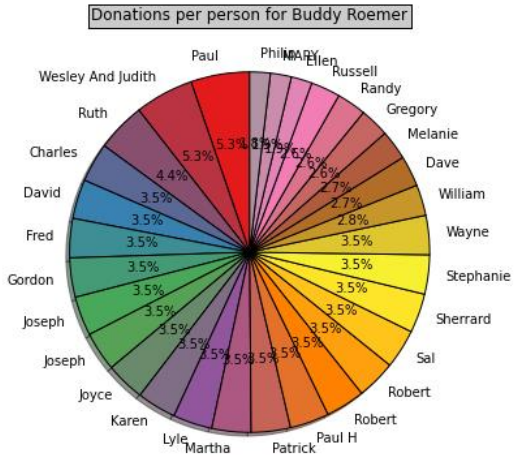
Amount donated to Gary Jhonson by individuals





Amount donated per state for Gary Jhonson

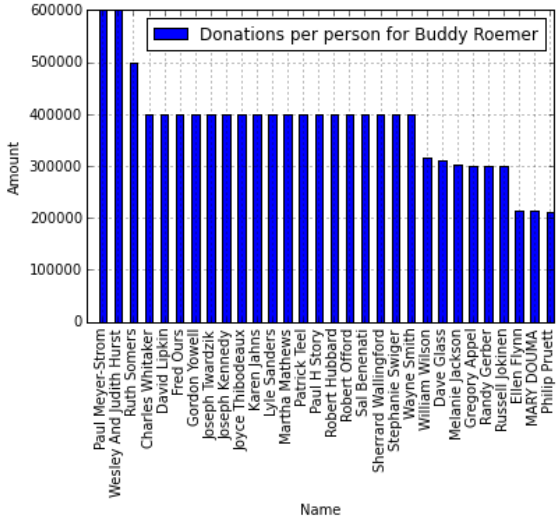


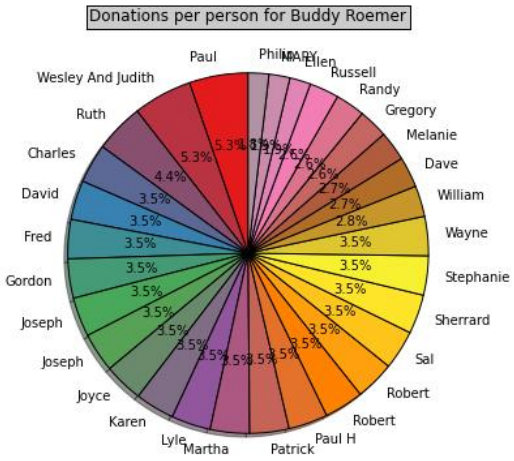


Amount donated by individuals to Buddy Roemer

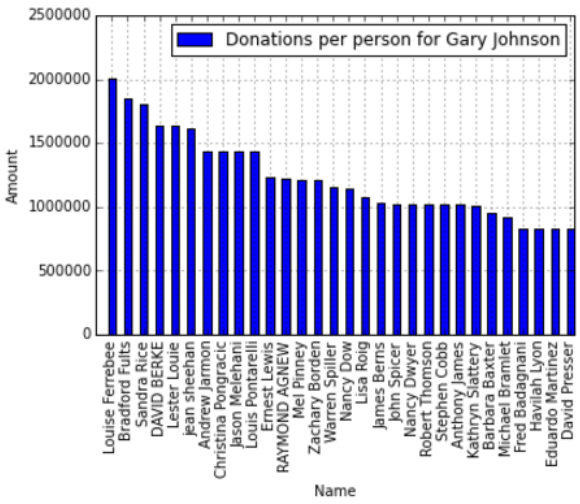
The data quantity was huge so I've taken a subset of the complete data based on the descending order of paid amount.

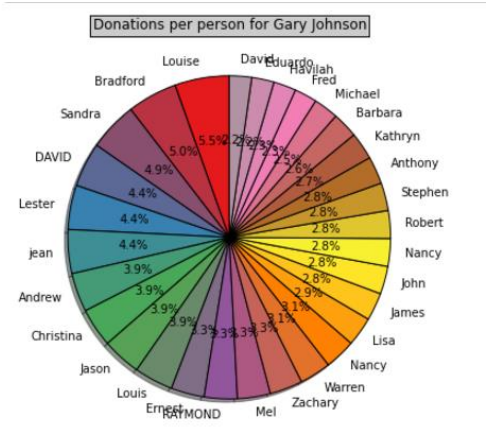
Findings: I'll list a few of the findings for this report





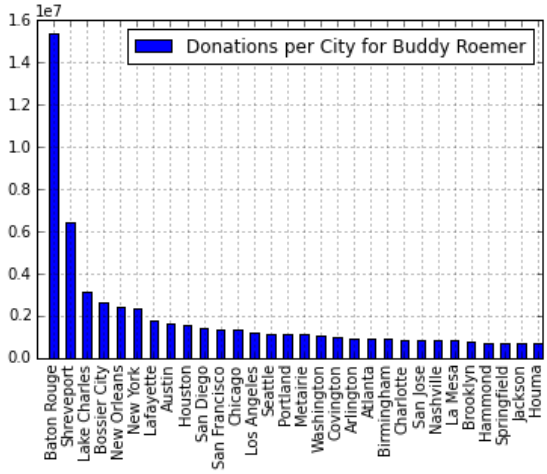
Amount donated by individuals to Buddy Roemer

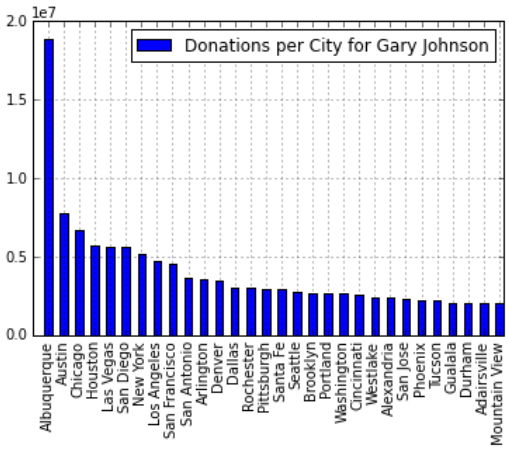




Amount donated to Gary Jhonson by individuals

Paul Meyer-Strrom was the Highest doner for Buddy Roemer with a donation of $600,000 while Loise Ferrebee was the Highest doner for Gary Jhonson with a donation of $2 million.





➔ Buddy Roemer was supported the most by Baton Rouge while Gary Jhonson was supported by Albuquerque