**Election Data analysis of U.S presidential election 2016 Using Machine Learning**

*A project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

**CERTIFIED SPECIALIST**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

**Team No.3**

**Members**

**Asma K**

**Dilip Rajendran Nambisan**

**Harikrishnan Ajith**

**Neethu John**

**Sruthy B**



**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**Feb 2022**

**List of Figures**

**List of Abbreviations**

**Table of Contents**

**Abstract**

Elections make a fundamental contribution to democratic governance. Election data analysis is a valuable tool for improving the quality of elections and help build public confidence in the honesty of electoral processes.

In this work, our aim is to build a ML model for finding the number of votes gained by political parties in region-wise elections and to predict the most popular political party. For that we are using the Kaggle datasets of the 2016 U.S presidential election by Ian Jeffries and Ben Hamner. From the data we are trying to predict the winning party in each state and to find the number of votes gained by the major political parties; Republic and Democratic for each county. Also attempting to explore more insights from the data. For better performance, we have merged the two datasets and the new dataset is used for analysis. Our process involves the following steps; Data understanding, Data preparation, Exploratory data analysis, modeling, model evaluation and model deployment. We are using Logistic Regression and Classification methods in Supervised learning techniques and using the best model we will go for prediction.

**1. Problem Definition**

**1.1 Overview**

United States Presidential Election of 2016 sets off intense debates as even though the Democrat candidate Hillary Clinton won popular votes of more than 2.8 million than Republican candidate Donald Trump, Clinton lost to Trump by 227 decisive electoral votes to Trump’s 304 votes and Trump became the 45th president of the U.S. Various socio-economic and cultural factors have played a vital role in the voting preference between the major parties democrat and republican.

In this study we are going to analyze these voting preferences of people across a range of demographic data based on the 2016 U.S presidential election. We are creating a machine learning model using supervised learning to predict the number of voters who are favoring democrat and republican to make the final winner.

**1.2 Problem Statement**

* To find how socio-economic factors affect voting behavior.
* To find count of the county wise winners per state.
* To predict the number of electors who are favoring Democrat and Republican to finally decide the President.

**2. Introduction**

**2.1 Data Understanding**

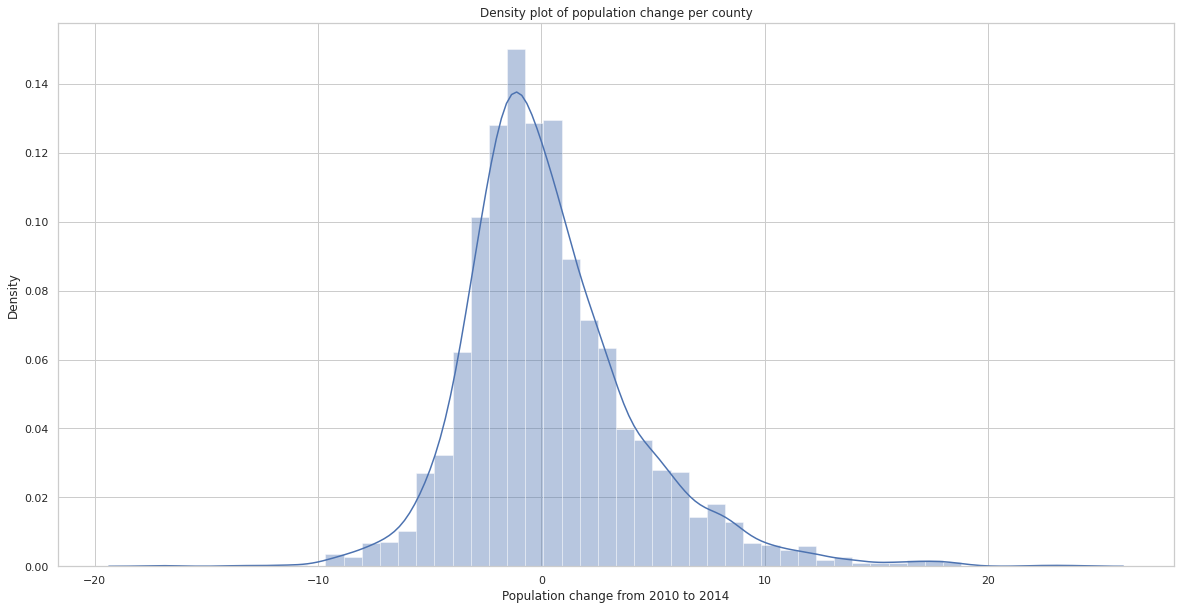
We are using Kaggle Datasets of the 2016 US presidential election by Ben Hamner (county\_facts.csv) and Ian Jeffries (primary\_results.csv). for better performance we had merged the two datasets and that is used for analysis. Our final dataset is "project\_data.csv". In this data, area wise total votes gained by each party (democratic and Republican) and demographic data such as age, race, education, population etc. are given. From this data we are going to predict the voting preference based on demographic data. Our data contains 2772 rows and 57 columns with object, integer, float data types. Demographic data are our features. The target column is going to be a calculated column, which will have the column name "winner". It will be a categorical column with the value "0" for counties where Democrats are more popular and "1" where republicans are more popular.

Under the Democratic and Republican columns, total votes gained on each party per county are given. Demographic data contains mainly population percent, age groups percentage, Education, Sex and ethnic groups percentage. There is some population data from 2010 to 2014. Also, there is some data related to Residence, income, sales etc. population per square mile and land area per square mile are also given.

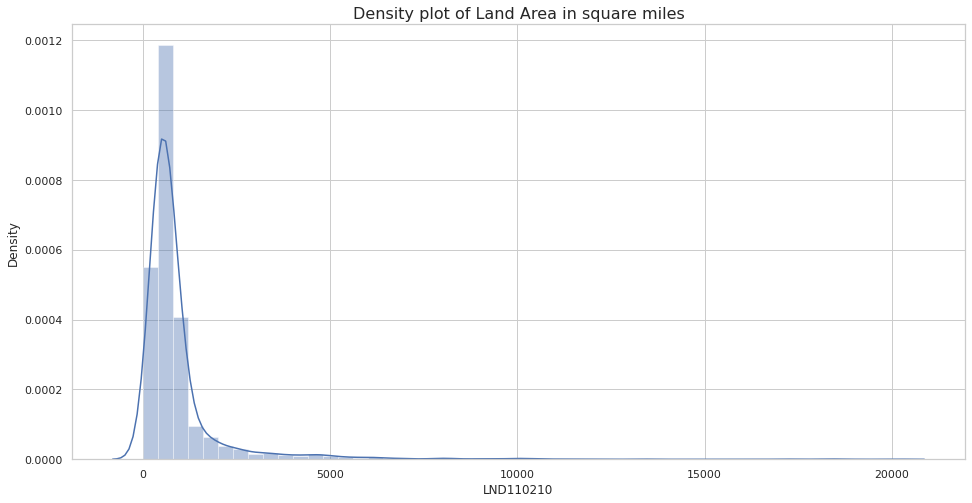
There are some missing values in the columns ‘Democratic’ and ‘Republic’ which we need to fill or drop. There are also some duplicate columns we need to remove. From the statistical data, it is clear that Republican votes are more than Democratic votes and whites population percentage is more compared to other races.

**2.2 Exploratory Data Analysis**

* + 1. **Univariate analysis**

1. Density plot of population change from 2010 to 2014 per county is plotted. It is shown below.

The graph is almost normally distributed and the population change ranges between -10 to 10 with mean zero and most of the counties show an increase in population from 2010 to 2014.

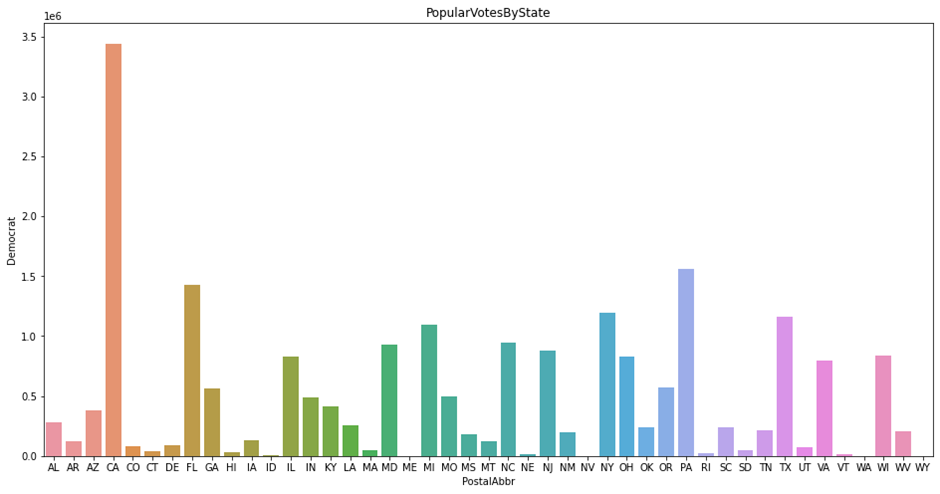
1. Density plot of Land area in square miles is shown below;

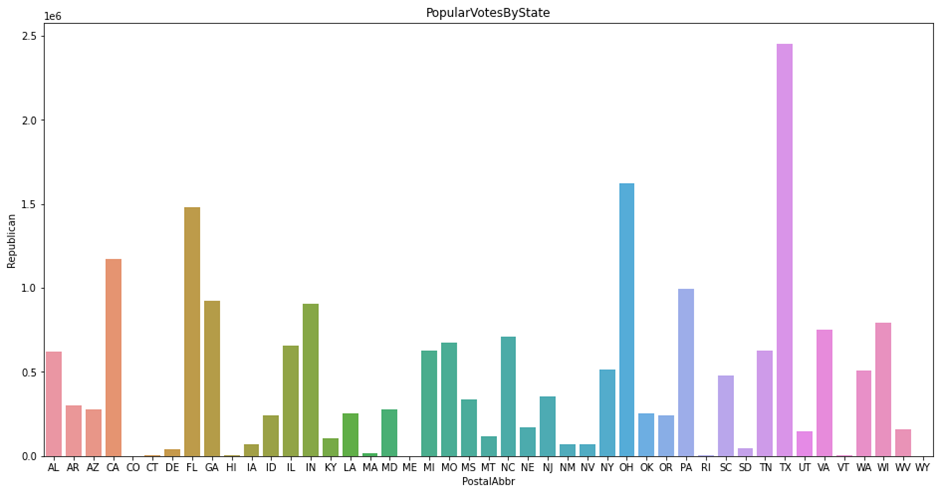
The plot shows distribution of land area in square miles and it ranges from 0 to 10000 square miles. The graph is right skewed with a median closer to 1000 square miles.

* + 1. **Bivariate Analysis**

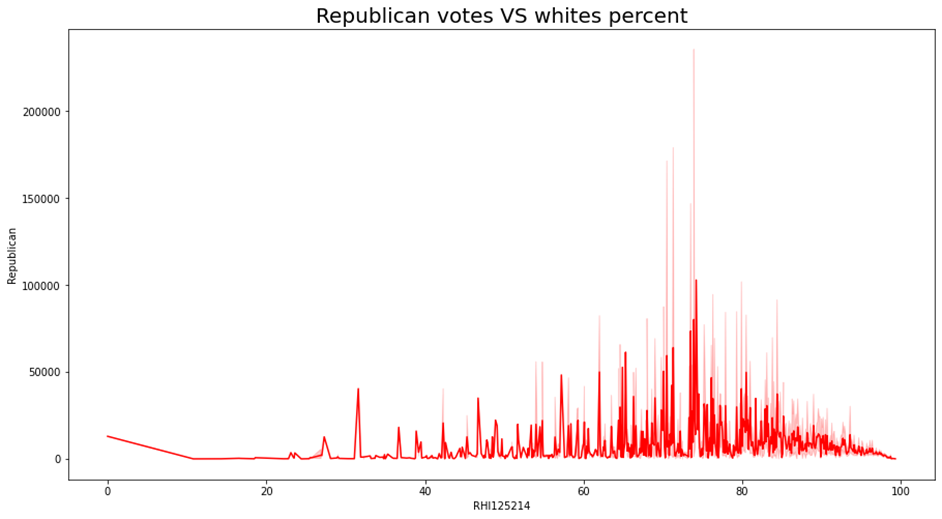
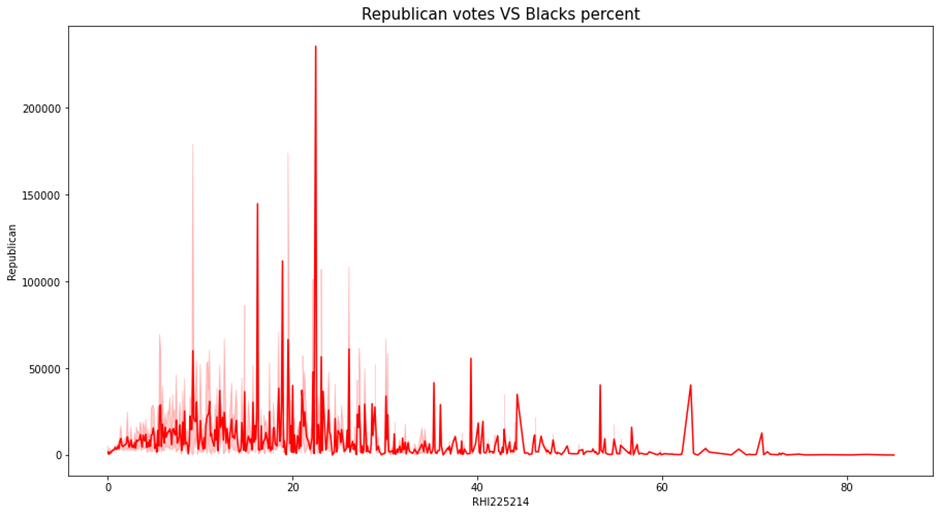
1. In bivariate analysis, state wise Democrat votes and republican votes bar plots were plotted.

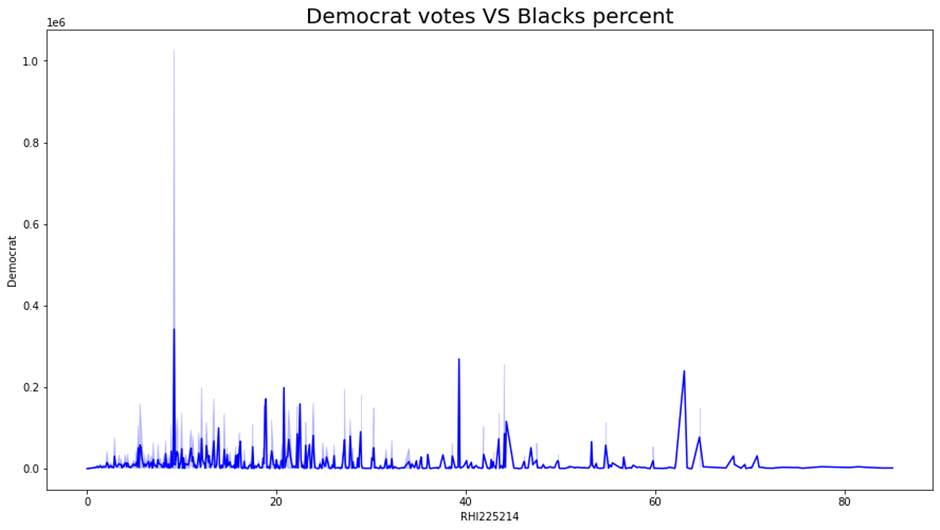
x-axis shows postal abbreviation of states and y-axis shows the party (Democrat or Republican). From the graphs below, it is clear that republicans got more votes from Texas and for democrats more votes marked in California state. There are some states where votes are negligible or nearly zero.

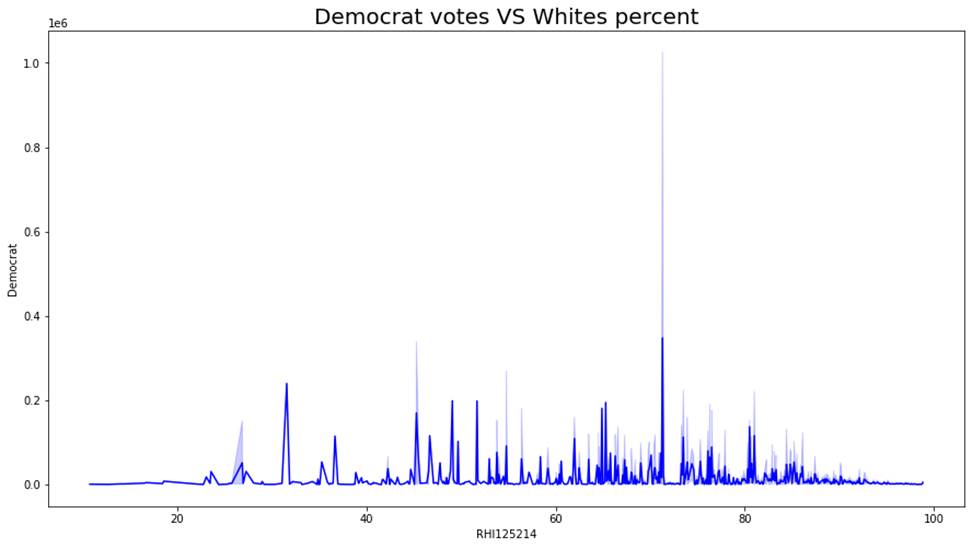




1. Voting distribution among major racial groups whites and blacks were analyzed using line plots. In this analysis, democrat votes were found to be more in counties where whites percent is about 70% and blacks percent about 8%. Republican votes are concentrated on counties where white percent was about 75% and blacks percent was about 23%.

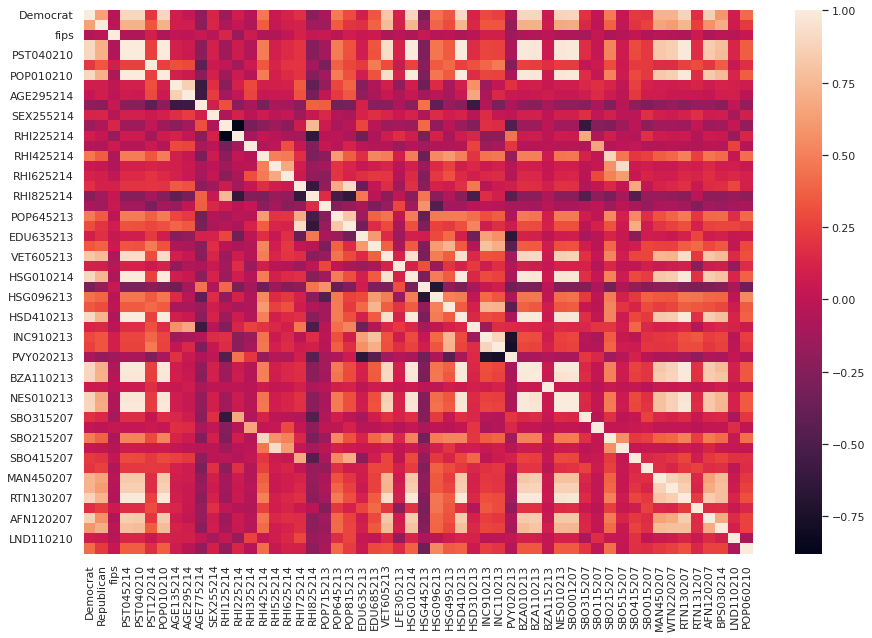






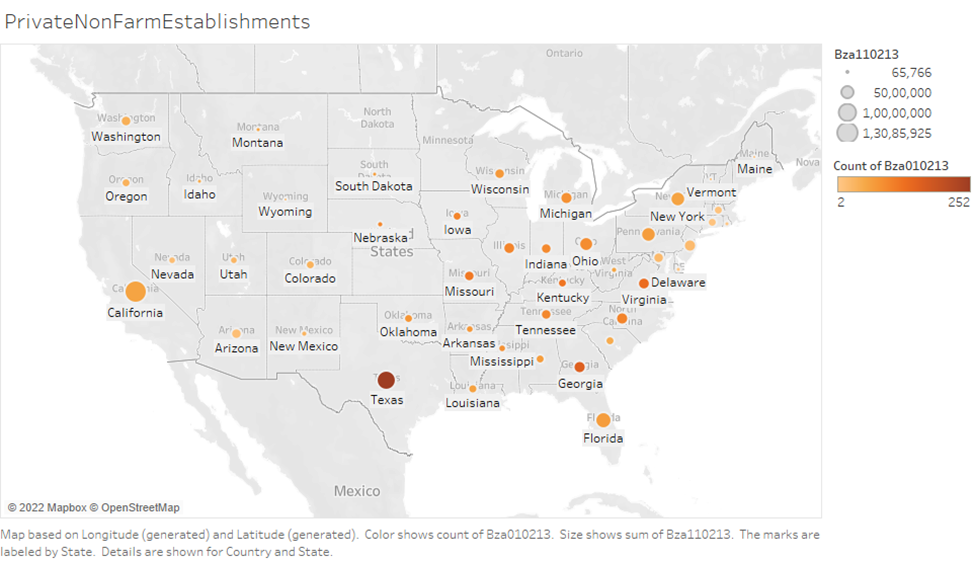
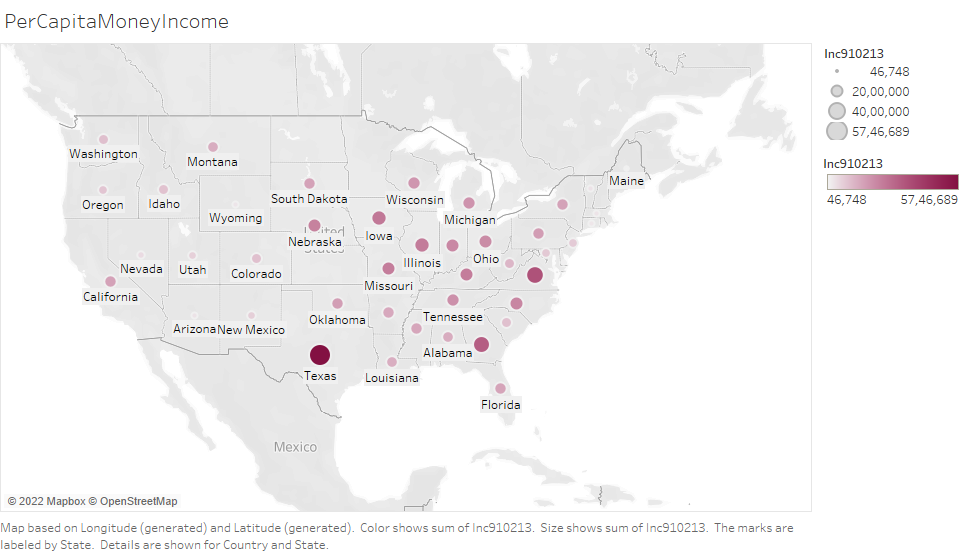
* + 1. **Multivariate Analysis**

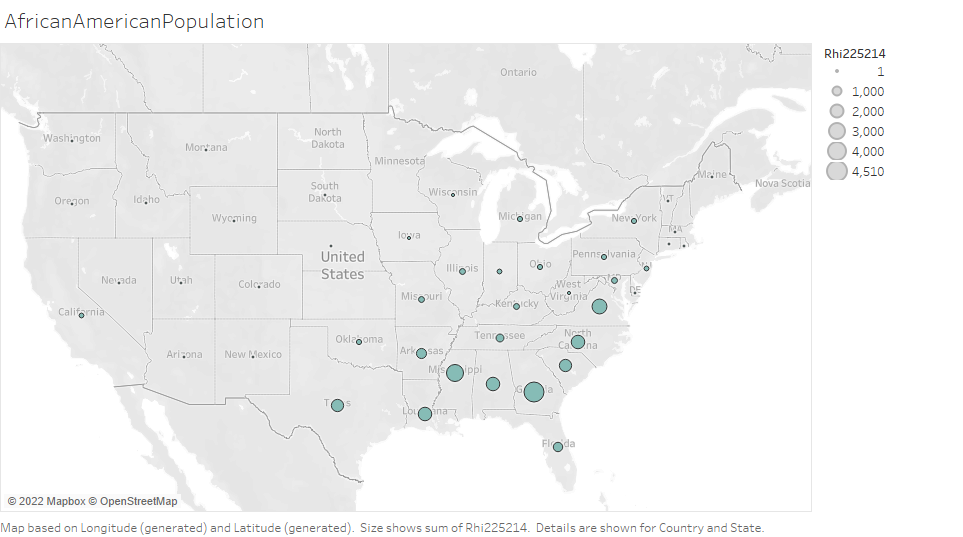
Inorder to find relationships between features of the dataset we have plotted a heatmap where light shades denote strong positive correlation and darker shades denote strong negative correlation. From the heatmap, it is clear that the features are strongly correlated.



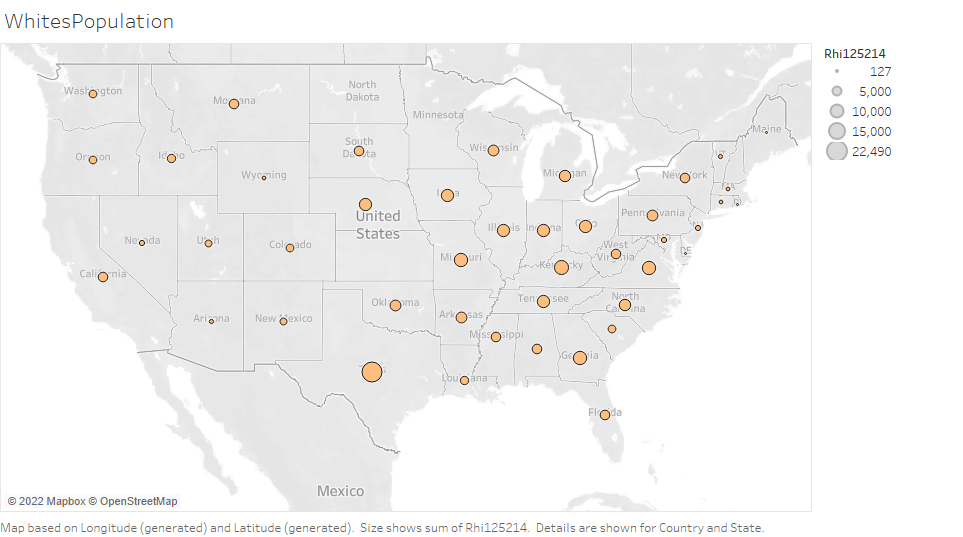
* + 1. **Tableau vizzes**

By comparing graphs of Private nonfarm establishments with per capita income it is clear that Texas has huge per capita income due to more private establishments.



From graphs below, African American population concentrates on eastern side of the country.

From graph below whites population more concentrated on south east parts of country.



* 1. **DATA PREPROCESSING**

Data preprocessing is a prerequisite step in machine learning. The efficiency and accuracy of the ML model can only be achieved through proper preprocessing of Data. It refers to the transformations applied to the data before feeding it to the algorithm which can be interpreted by machine. In simple words it is a technique for converting raw data into a clean dataset. It helps in achieving better results from applied models in ML, and formats data in a proper manner. For example, Random Forest Algorithm does not support null values. Generalization of ML models can be achieved in such a way that more than one ML or Deep learning algorithm are executed in one dataset and best out of them can be chosen.

Major steps involved in preprocessing are Data integration, Data Transformation, Data Reduction, Data cleaning. Data cleaning involves handling missing values, outliers, duplicates, noisy data etc. Feature scaling, attribute selection, encoding etc. are involved in Data Transformation. Data Reduction is used to remove huge amounts and volume of data inorder to make analysis easier. By constructing correlation matrix using heat maps is an easier way to find multicollinearity. Feature Scaling is used when the features involved have different units of measurements. It involves Normalization and standardization.

* + 1. **Data Cleaning**

**Handling missing values:**

Presence of missing values in a data will decrease efficiency of the model. There are only two options to handle missing values either by dropping missing values or by imputing values like mean, median or any other values to it.

In our dataset there are about a total of 592 Missing values in the columns, ‘Democrat’ and ‘Republic’. We have to either drop the rows containing NaN or fill nulls with some value by imputation technique. Since the number of missing values in our data is high, we cannot drop it since it may lead to huge loss of information. So, we imputed zero to fill the missed values.

**Handling Outliers, duplicates and other irrelevant data:**

There are some duplicate columns (‘ua’) and rows (at indexes 151,1267,1266) which we have removed. The ‘fips’ column is removed since it is irrelevant for our study.

* + 1. **Data Transformation:**

Our target column is the ‘winner’ column which we need to create additionally based on the popular political party in each county. It contains two values ‘0’ and ’1’, where ‘0’ represents counties where Democrats dominate and ‘1’ represents counties where Republicans dominate. Inorder to create that column we have to remove counties where both the political parties scores zero votes and counties where both have same votes. Then we found Codington and Owsley counties have equal votes for the two parties, but the winner column stands ‘1’. Inorder to confirm that we compared with the primary\_results.csv file which we used for merging. From that we found the Republican party dominates in those counties because of that the winner column stands ‘1’and so we need not remove those rows.

**Encoding:**

We used the mean encoding for the categorical feature ‘PostalAbbr’ based on target class and it represents the probability of the target variable, conditional on each value of feature. so, by mean encoding we create a feature that is more representative of the target variable and also solves the encoding task. In our data ‘PostalAbbr’ stands for states of the country and we are encoding it based on the ratio of occurrence of the positive class in the target variable.

**Feature scaling:**

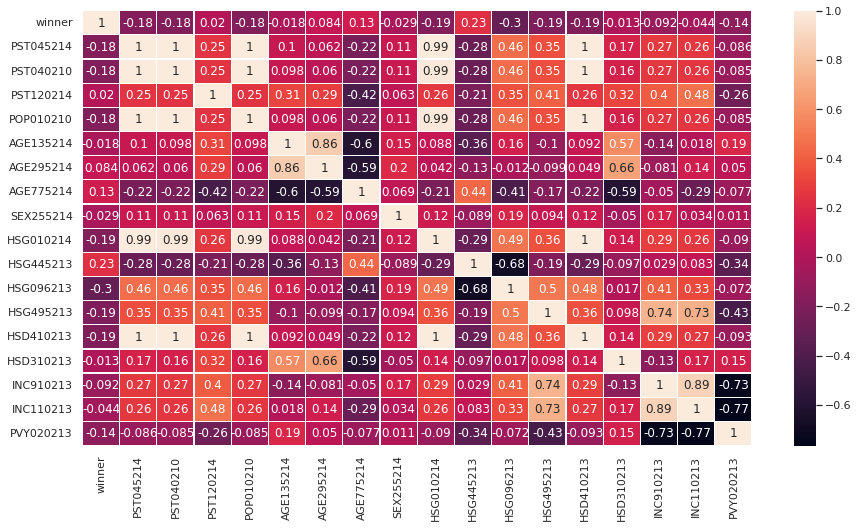
Feature scaling is the process of normalizing the range of features in a dataset. It is a crucial part in the preprocessing stage. It involves mainly normalization and standardization. In normalization values are scaled between 0 and 1, whereas in standardization it does not have a bounding range; instead, the values are centered around the mean(mean=0) with a unit standard deviation.

First of all, the data is split into target and features. The target column is ‘winner’ represented as y and x is the feature data. Our feature column contains different units of measurements, some are in percentage whereas some others are in square miles. The values in these columns contain integer, float type data so we need to scale these into a range of values. We have scaled the Feature data by Standardisation and Min-Max Scaling.

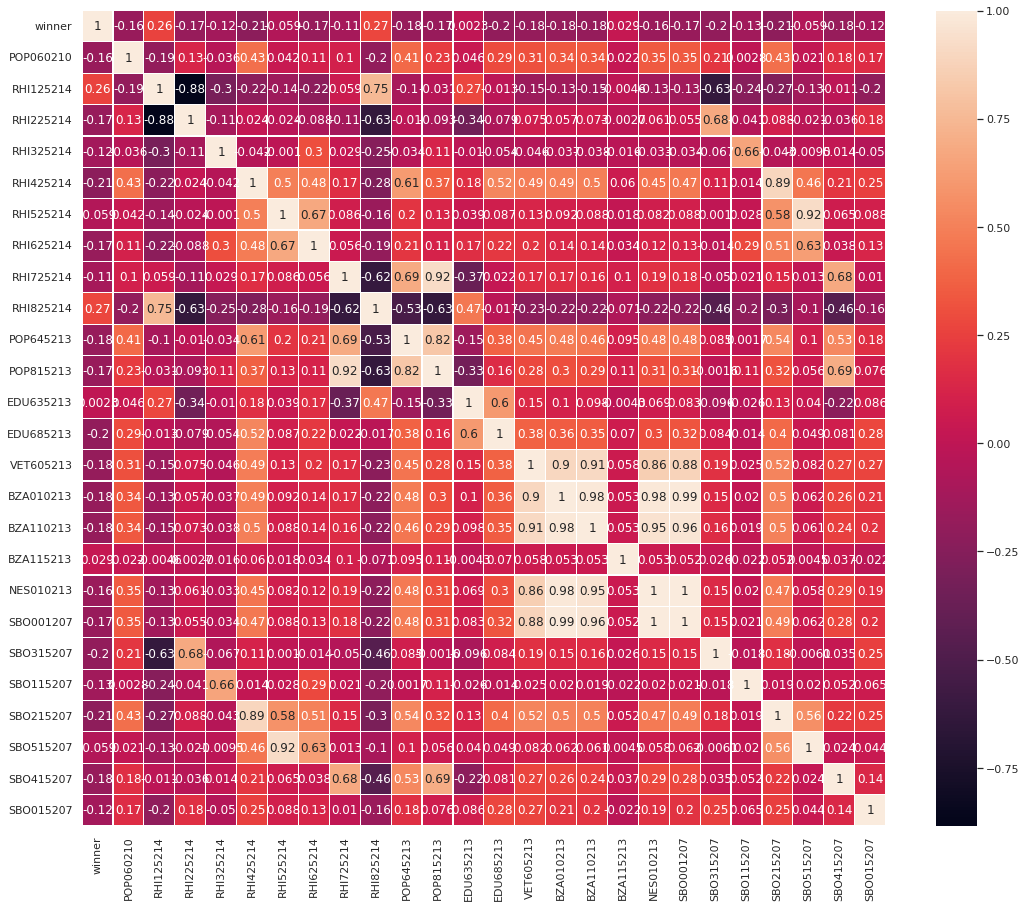
* + 1. **Feature selection and reduction:**

Analyzed correlation between the variables in the dataset using heatmaps. We create separate heatmaps for the features for better understanding. Some columns are highly correlated and some are negatively correlated.

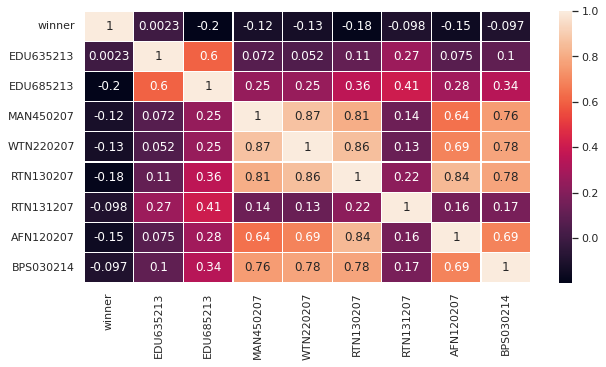
* 1. **correlation of population, Age, Number of households, Per Capita Income/Poverty and House ownership data.**



* we shall retain only PST045214(Population by 2014 estimate) since strong positive correlation between PST040210 and PST045214 with 1 as correlation coefficient.
* HSD310213(Persons per household 2009-13) and AGE775214 (percent of persons 65 years and above) have negative correlation of -0.59. In this case, we shall retain the latter AGE775214, alone.
* HSG445213(home ownership rate) and HSG096213 (percentage of housing units in multi-unit structures) have -0.68. Among these two, we can retain one column, ie, HSG096213.
* strong positive correlation between POP010210 (Population, 2010) with PST045214 (Population, 2014 estimate) and PST040210 (Population, 2010 (April 1) estimates base) so dropping POP010210.
* There is an inverse correlation of -0.59 between HSD310213(Persons per household 2009-13) and AGE775214 (percent of persons 65 years and above).
* There is -0.68 correlation between HSG445213(home ownership rate) and HSG096213 (percentage of housing units in multi-unit structures).
* We see a correlation of 0.89 between INC910213 (per capita monetary income in last 12 months) and INC110213(median household income).
* a -0.73 correlation between INC910213 and PVY020213(percent of persons below poverty level) ; so dropping PVY020213 and INC110213.
  1. **Correlation of Racial ethnicity, Education, Employement, Firms data**



* RHI125214(white alone, percent) and RHI225214(Black or African American alone, percent) has strong negative correlation of -0.88 as expected.
* RHI425214(Asians percent) and SBO215207(Asian owned firms) have strong positive correlation of 0.89 so we will drop the latter.
* RHI525214(Hawaiian and other Pacific islanders percent) and SBO515207(Hawaiian owned firms) with strong positive correlation coefficient 0.92; we will drop the latter.
* BZA110213(Private non-farm employment) andBZA010213(Private non-farm establishments) has very high positive correlation coefficient of 0.98 so we shall drop BZA010213.
* RHI725214 (Hispanics and Latinos percent of population) and POP815213 (language other than English) also have high positive correlation (0.92).
* POP815213 (language other than English) and POP645213 (percentage of persons who are foreign born) with correlation 0.82, so POP815213 (language other than English) is dropped.
* NES010213(non employer establishments) has a high correlation to the many others and extreme correlation of 1 with SBO001207(Total number of firms); we will drop NES010213
  1. **Correlation between Education, Shipments, Sales, Building permits data**



The

* MAN450207 (manufacturers shipments) and WTN220207 (merchant wholesaler sales) have highly positive correlation of 0.87 so we will drop MAN450207.
* RTN130207(retail sales) and WTN220207 (merchant wholesaler sales) also have correlation coefficient 0.86 so we are dropping RTN130207.
* BPS030214 (Building permits) and WTN220207 (merchant wholesaler sales) has correlation coefficient 0.69 and we will drop BPS030214.
* AFN120207(Accommodation and Food Service Sales) and WTN220207 (merchant wholesaler sales) shows 0.69 correlation so, dropping AFN120207.

Finally, the columns ‘HSG010214’, ‘PST040210’, ‘POP010210’, ‘HSD310213’, ‘HSG445213’, ‘PVY020213’, ‘INC110213’, ‘POP815213’, ‘SBO215207’, ‘SBO515207’, ‘VET605213’, ‘BZA010213’, ‘NES010213’, ‘MAN450207’, ‘RTN130207’, ‘BPS030214’, ‘AFN120207’ have been identified for dropping.

In Dimensionality reduction, Principal component analysis (PCA) is performed on the data for 96 % variance and after that our features are reduced to 25 columns. Now we have 2767 rows and 25 columns.

The data is splitted into test and train data and we can go for ML modeling.

**3. Literature Survey**

**7. Result**

**8. Conclusion**

**References**