The team discussed a variety of different topics and subjects to consider for this project, we also looked at and read various topics on data exploration[[1]](#footnote-1). As we looked at the data available to us, we started shaping the discussion around our purpose. This gave us a more defined scope from which to focus. After sorting thru retail data, online sales and crises-specific major events, we opted for donation data focused on presidential election years.

Once we agreed on donation data, we focused on presidential elections and ways that the elections may be affected. This lead us to narrow our focus on six specific states (AZ, FL, MI, NC, PA, and WI) and the counties therein. Historically, these states have been considered battleground and/or swing states, having a significant impact on the presidential election and electoral votes.

We set out to answer the following;

* Do presidential candidates have an effect on how people donate?
* Can we predict who the president will be based on donation data?
* How much money is donated across candidates? States? Counties?
* Are donations affected by employment/unemployment?
* How do rural and urban neighborhoods affect voting?
* Does FL, MI, PA, NC, WI and AZ have a significant impact on elections?
* Do economics effect donations?
* How do major event effects on donations (9/11, real estate market crash, COVID) and/or presidential elections?
* Do demographics, such as age and education have an effect on these data?
* What significance does health information have?

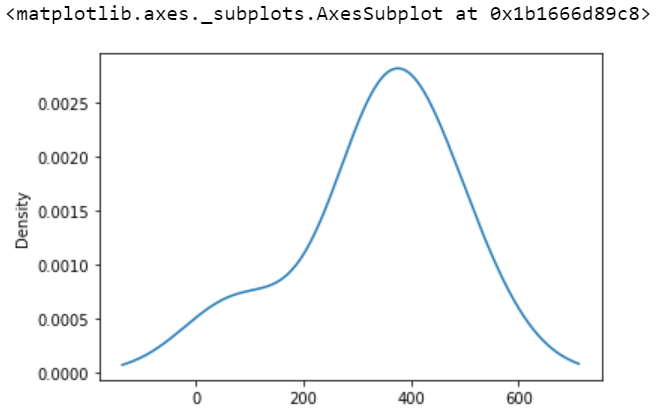
We explored the data and created several visualizations[[2]](#footnote-2), looking at different aspects of the data we sought to explore. We also did some value counts of the datasets to determine the best storage options and how to pull the recall the information to create our models, predictions and perform analysis. We have twelve tables for this analysis. Below is a list and graph with the information.

**Description Table Name Rows Columns**

1. Donors on selected states agg\_county\_donors 6,748 12
2. Voters on selected states agg\_county\_votes 2,020 14
3. Births and deaths birth\_death\_rate 305 6
4. Presidential candidates since 2000 candidates 141,375 17
5. Donations for presidential votes donations 14,954,054 21
6. Committee presidential donations donations 141,375 15
7. Educational levels by state education 3,283 47
8. Health information health\_metrics 3,193 507
9. County postal codes postal\_codes 52,889 5
10. Presidential votes on selected states pres\_votes\_6T 6,464 11
11. Donations for presidential votes six\_state\_donations 14,954,054 21
12. Unemployment rates unemployment 404 114

We created various charts and graphs while exploring the data.

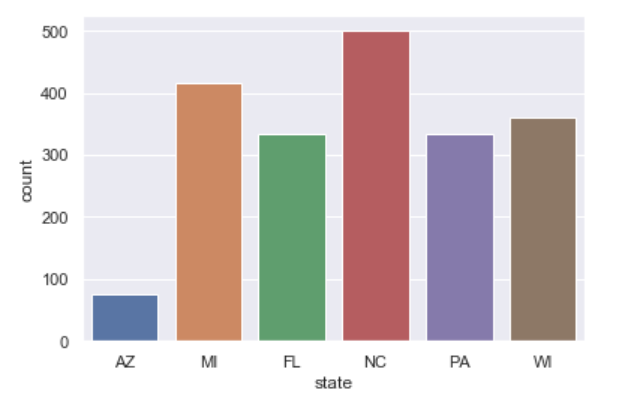
Total Votes graph



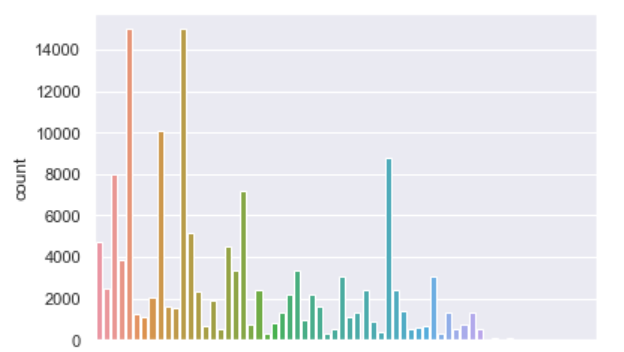
We used several steps out of the 15-step data exploration method[[3]](#footnote-3) to describe our process. We also obtained data exploration information from Medium[[4]](#footnote-4)

**1. Unique value counts: examples of votes by county and candidates by state**

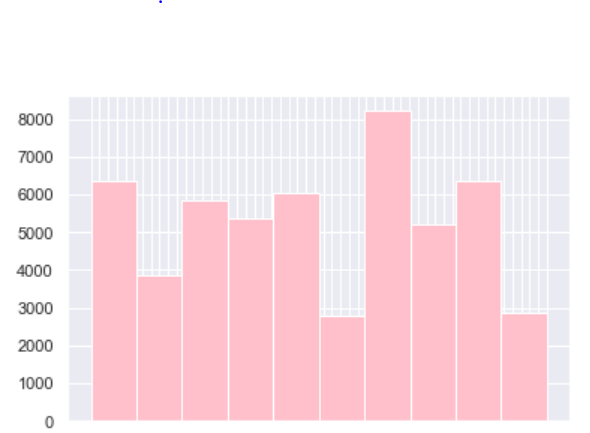
**Aggregated County Votes by State = 2,020**



**Candidates by State = 141,375**

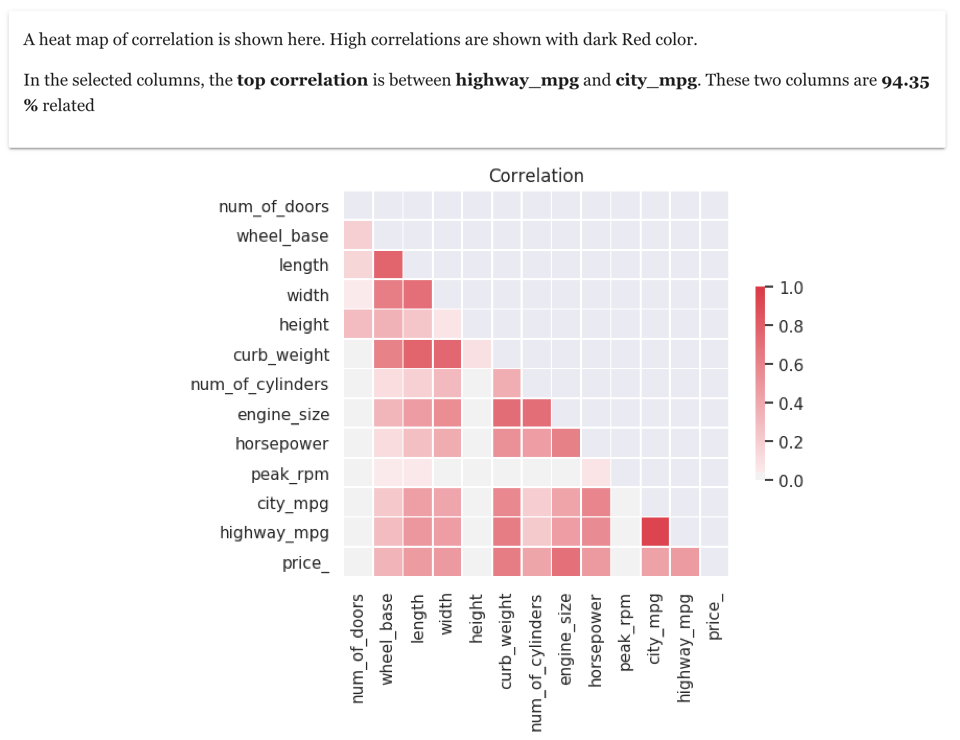


**2. Frequency Count**

**Histogram of Postal Codes by State**

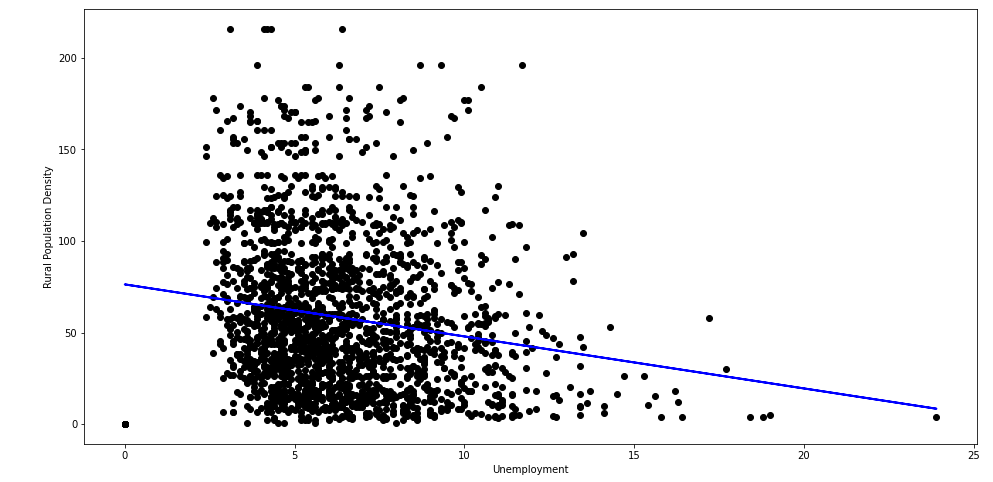
**3. Correlation**

When it comes to analyzing numeric values, some basic information such as minimum, maximum and variance are very useful. Variance gives a good indication how the values are spread.



Correlation Heat Map

We looked at the relationship between different aspects of the data, to determine whether or not the correlation was positive or negative. This example shows a negative correlation.



Correlation: Rural Population w/Unemployment

Data Analysis

Our team used a variety or tools and tables to analyze the data.

* Jupyter Notebook

This has been our main form of data analysis. We read in the tables from our SQL and Cloud connections to write simple to complex coding, focused on understanding and displaying the necessary data. To measure presidential data, we’ve pulled in donation and voter information, along with unemployment, education and demographics. We have discovered that these pieces of information have been useful in helping us transform and mold the data into information that is useful for looking at presidential elections.

* Spreadsheets

Everyone on the team participated in data gathering and located information to assist in our project. This information was located online, using data from the FEC, Kaggle and the Census Bureau. These data were analyzed and saved as CSV files that were uploaded to GitHub, PostgreSQL PGAdmin and Google Cloud for easy access and data sharing.

* Google Cloud

We place large dataset and bucketed information onto this platform. This allows to perform the ETL and join tables.

* PG Admin

We use PGAdmin to perform SQL queries and analyze datasets, as well as review table details.

1. Toward Data Science: 15 Steps to Data Exploration <https://towardsdatascience.com/15-data-exploration-techniques-to-go-from-data-to-insights-93f66e6805df> [↑](#footnote-ref-1)
2. <https://www.sisense.com/glossary/data-exploration/> [↑](#footnote-ref-2)
3. Towards Data Science: “An Extensive Step by Step Guide to Exploratory Data Analysis” <https://towardsdatascience.com/an-extensive-guide-to-exploratory-data-analysis-ddd99a03199e> [↑](#footnote-ref-3)
4. Medium: “Data Exploration” <https://medium.com/@abhinavr8/data-preparation-and-exploration-5e09b92cf00e> [↑](#footnote-ref-4)