Exploring Contributions and Expenditures in the 2016 Presidential Election

by Amanda Ziegelbauer

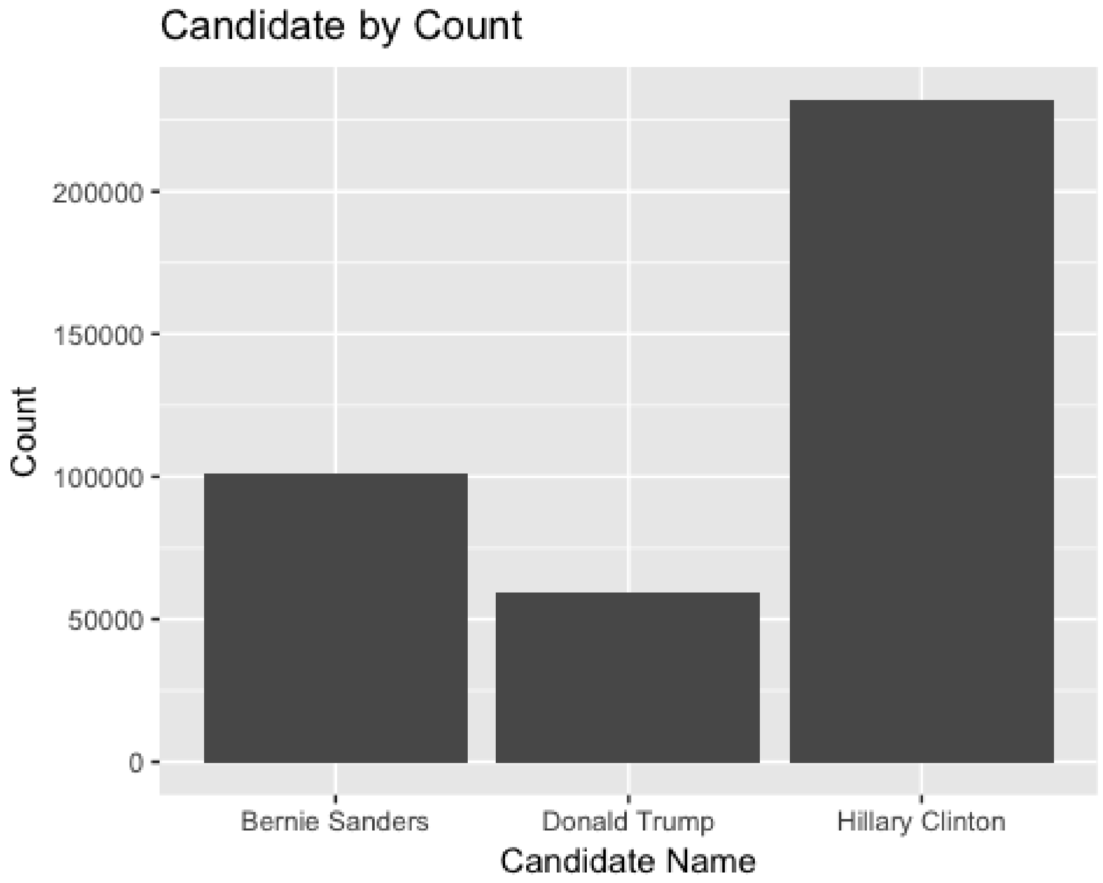
Introduction

This dataset was downloaded from https://classic.fec.gov/disclosurep/pnational.do. There were two files available - one for campaign expenditures, the other for contributions. Each required a decent amount of wrangling. I am using the individual files for this project, as well as a concatenated one that I created using python.

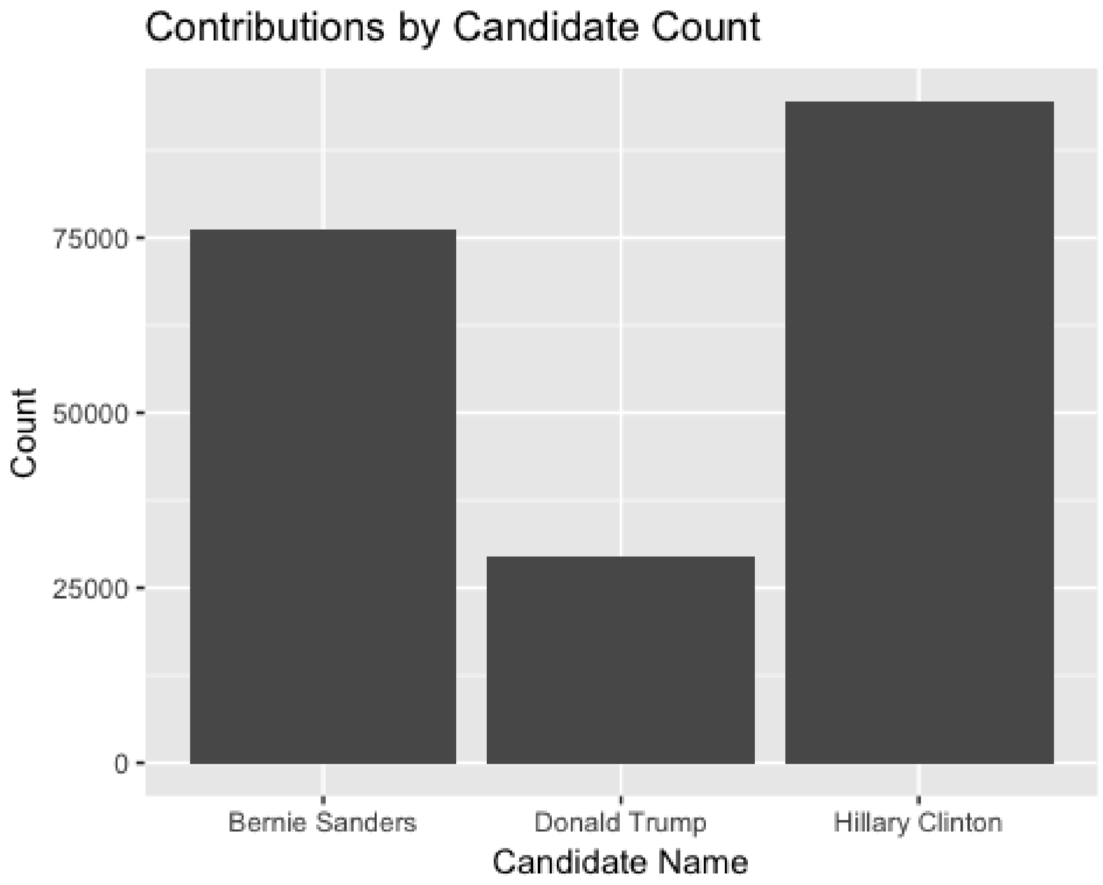
The file contained well over 7 million rows of information about every candidate that ran in the 2016 presidential election. In order to make my analysis more streamlined and meaningful, I decided to look at only the top three candidates in the election - Hillary Clinton, Bernie Sanders, and Donald Trump. I also took a sample of this subset data, since the original file was too large for rStudio to load. The data I used contains 200,000 rows of contribution data and ~190,000 rows of expenditures for these three candidates.

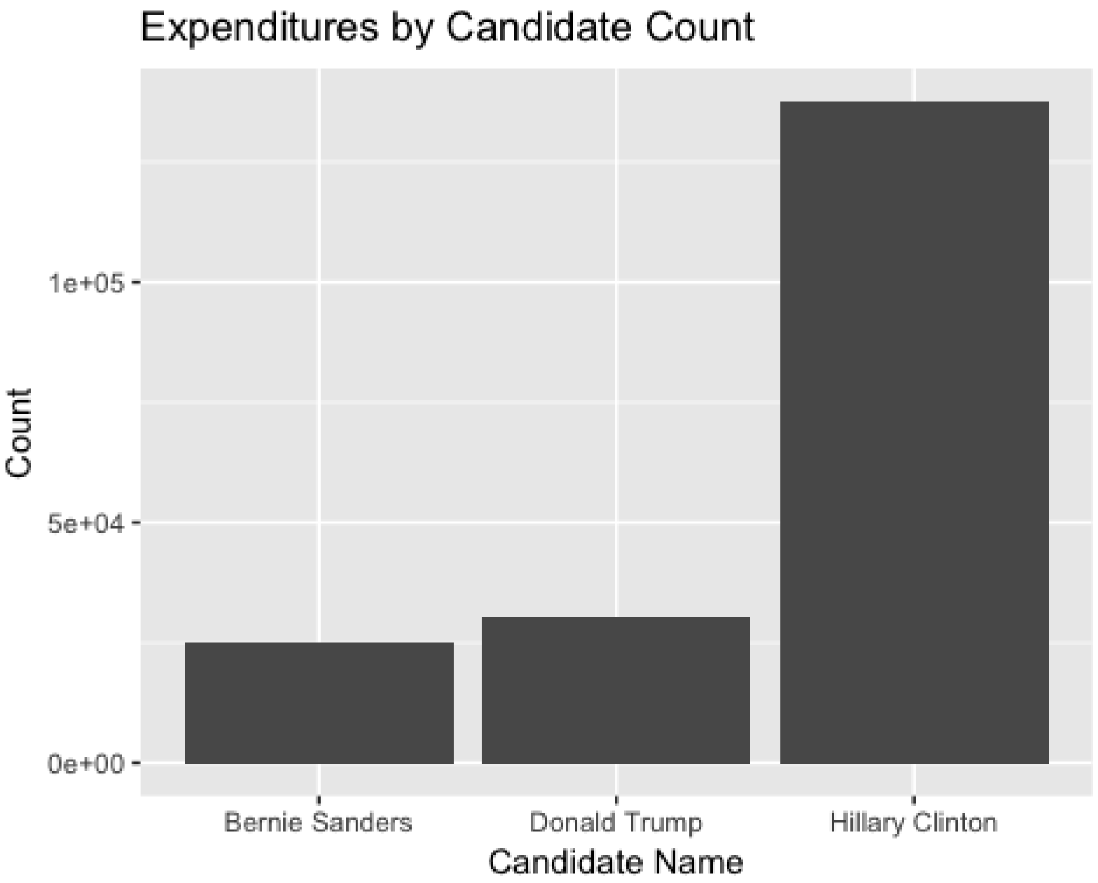
In this project, I will be looking at graphs created in R that delve into the nature of these candidates' campaigns, and will look for insights regarding their campaign transactions.

Univariate Plots



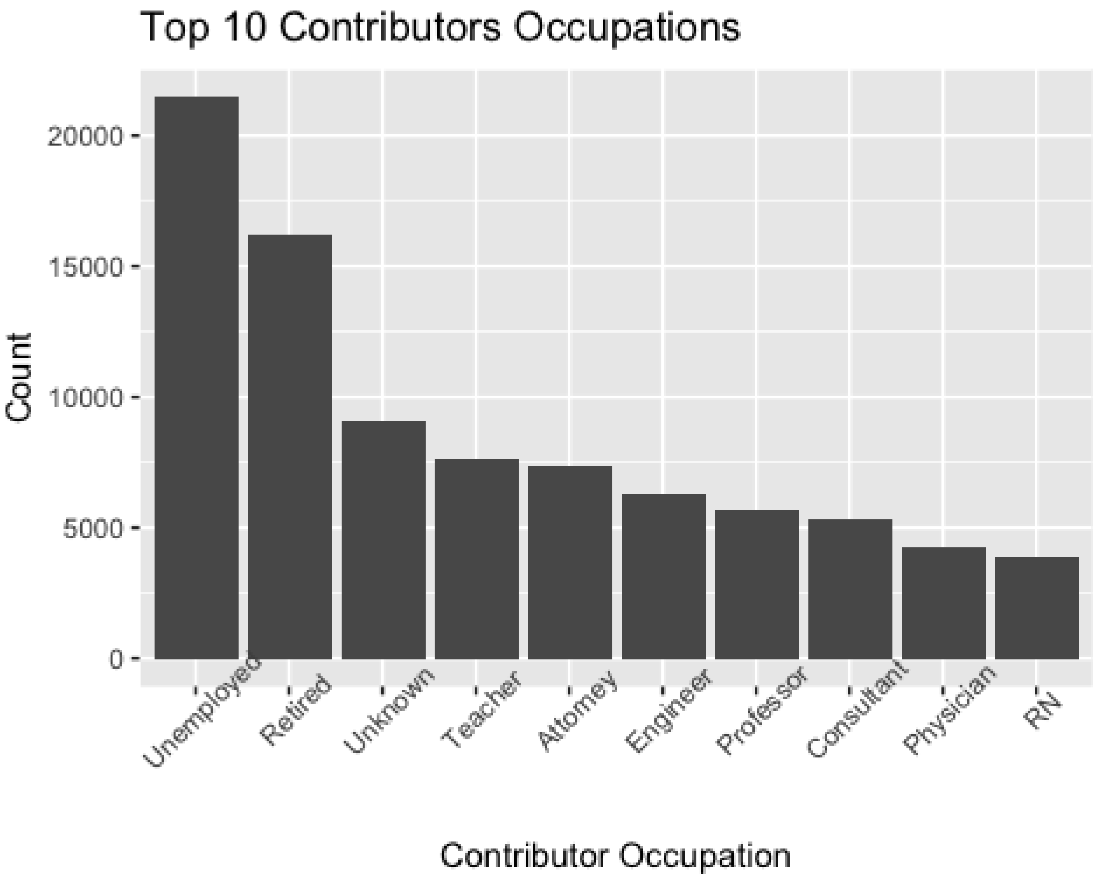
Right away it is easy to see that Hillary Clinton has the most entries for both contributions and expenditures. It would be interesting to see this graph for expenditures and contributions separately, to see if one metric is elevated by the other. I have split the data by transaction type, and remade the plots below.


Contributions by Candidate Count
Count
75000
50000
25000
0
Bernie Sanders
Donald Trump
Hillary Clinton
Candidate Name



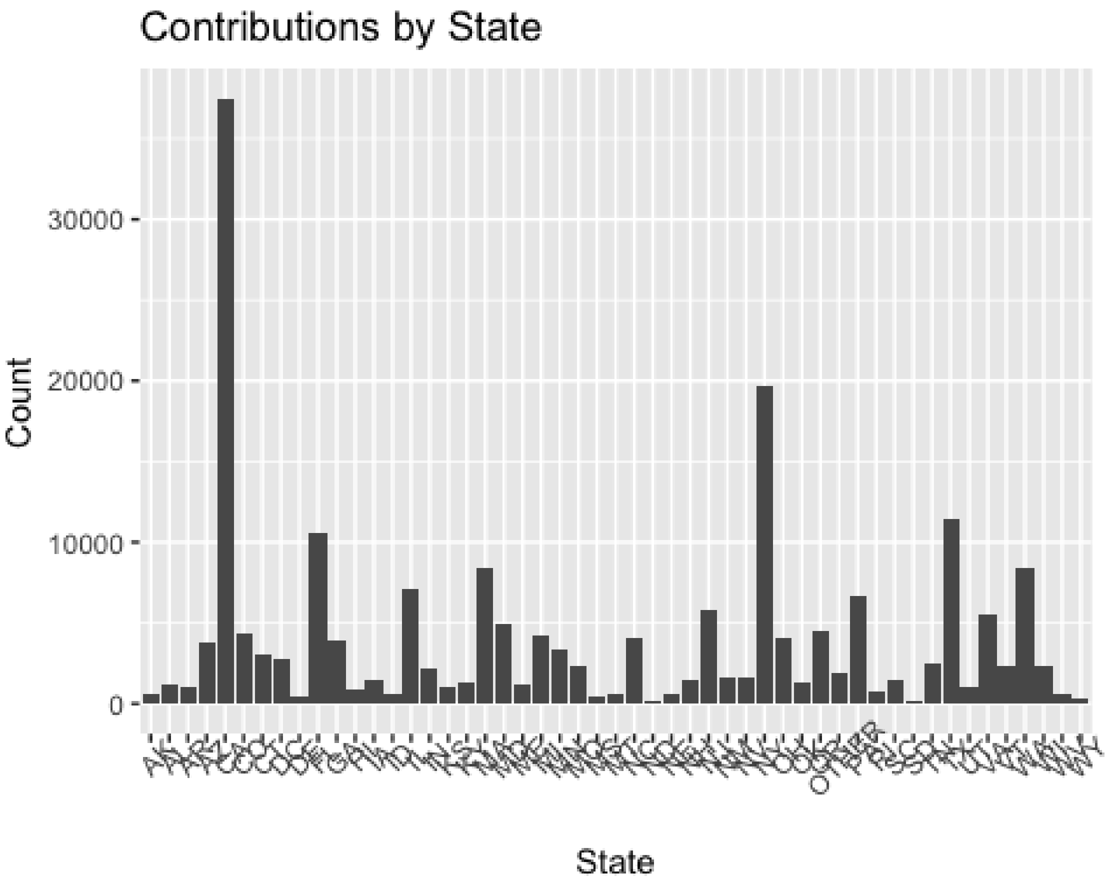
This is very interesting, because it does change the perception of the data. Bernie Sanders doesn't have many expenditures in this data, so on first glance of the amalgamated plot, it looks like he doesn't have many contributions either. However, when splitting the data, it is clear that he received almost as many contributions as Hillary Clinton.

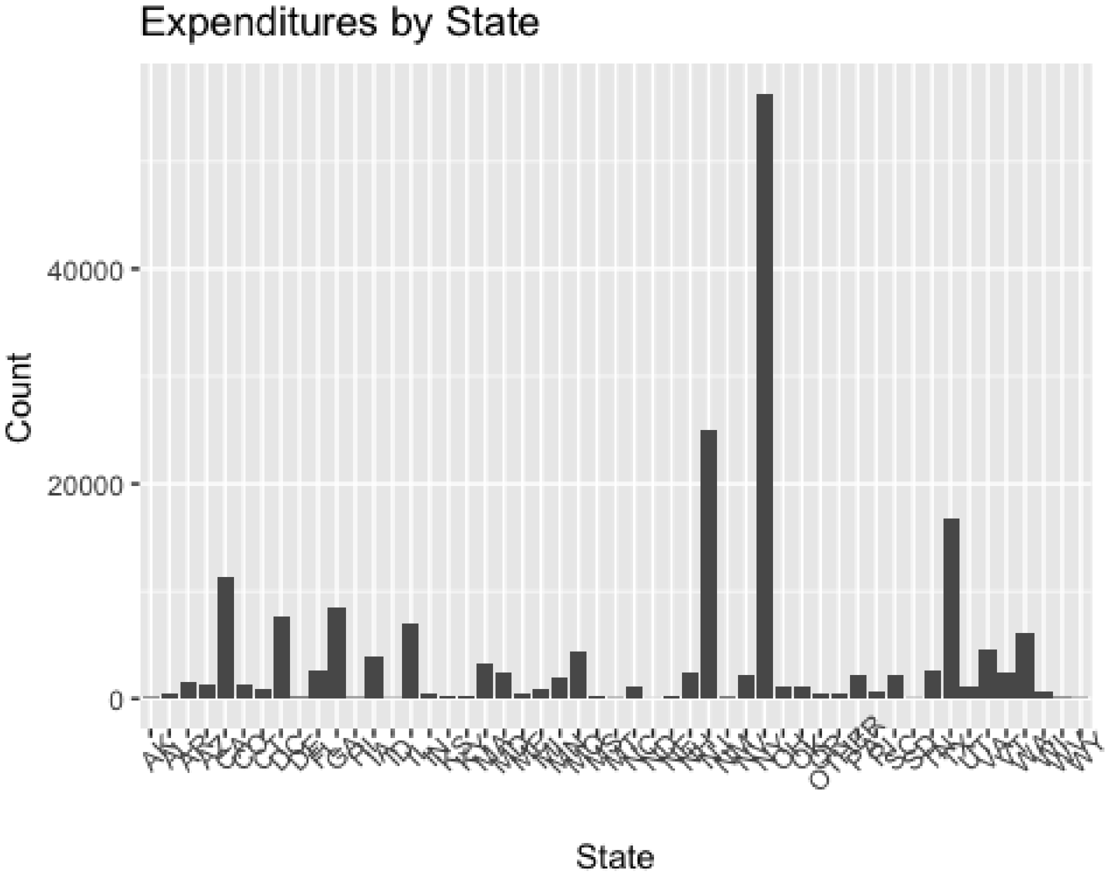
I also interested in looking at the people who donate to campaigns. The next graph shows the top 10 occupations of campaign contributors in this dataset using a subset created through a modified piece of code from Stackoverflow (the link is in the References section). I'd like to be able to include more occupations, but there are well over 100,000 occupations listed, which made the graph completely nonsensical.



This is interesting information, though it might be a more compelling graph if shown with either the candidate they support, or the amount given. This graph will be recreated later in the conclusion with additional information shown.

The next two plots show how many transactions of each type were made in every state. I am curious whether there are any areas that stand out as especially heavy in contributions of expenditures.


Contributions by State
Count
30000
20000
10000
0
AK
OTHER
PR
WY
State

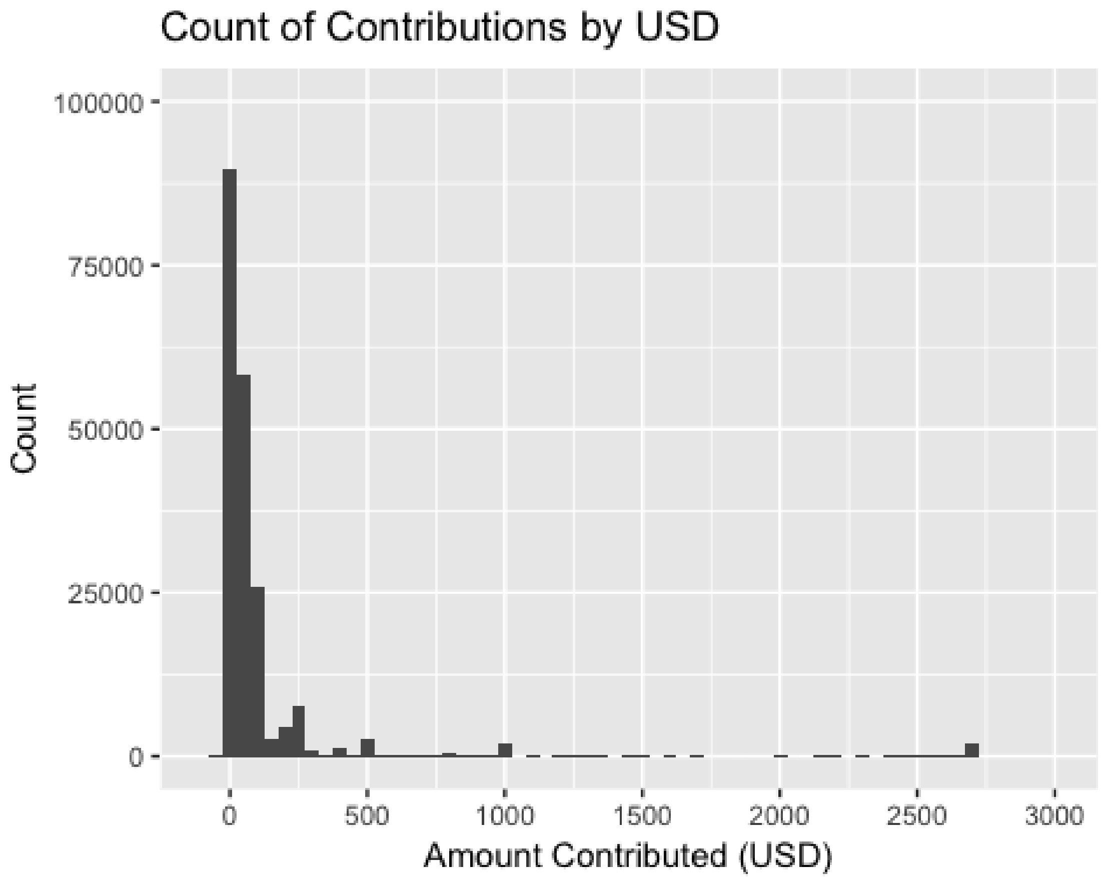


This plot shows that the highest amount of contributions comes from CA, followed by NY and TX.

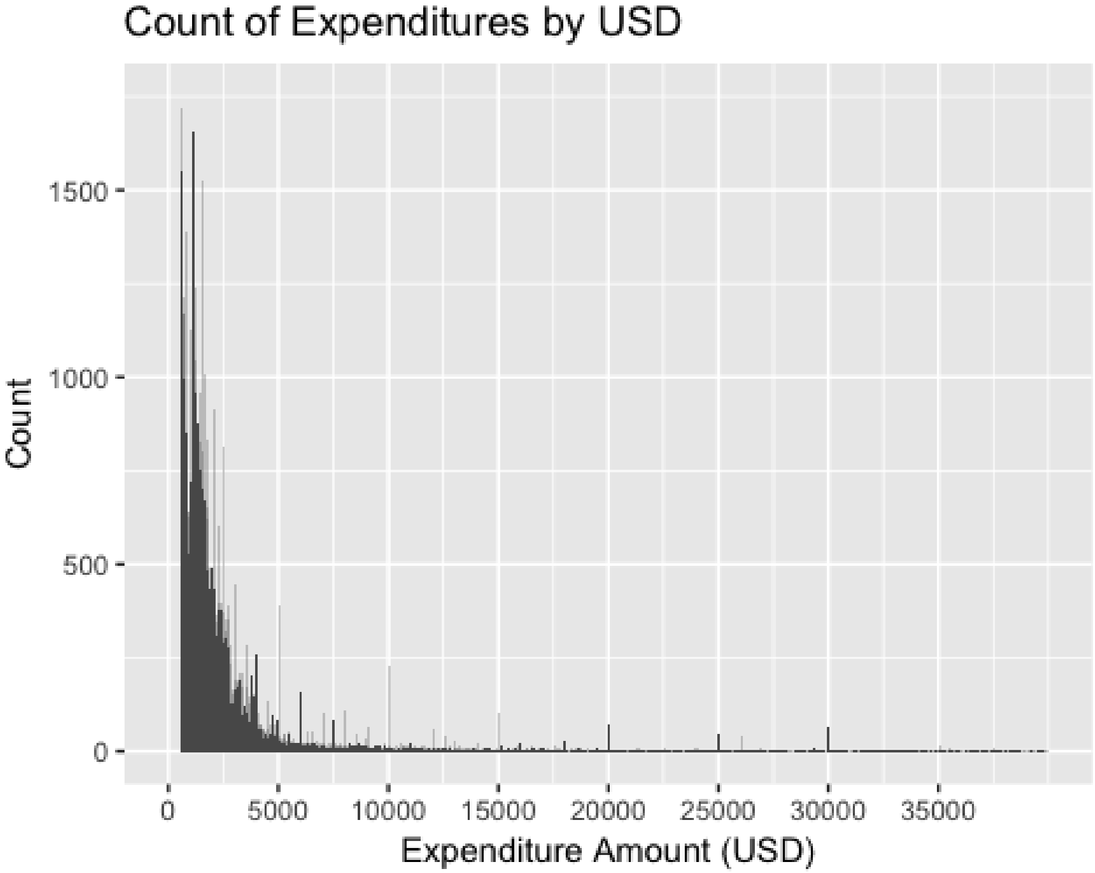
The expenditure plot shows that NY by far has the most expenditures, followed by NJ and TX. I think it is interesting that these two plots do not align. I also wonder if the reason for high expenditures could be due to the fact that all three candidates call NY home. It would be interesting to look at the kinds of expenditures that were made in the tristate area. I would expect that it's mostly travel related. Expenditure types will be explored in the plots below.

Note: The "Other" column contains information about territories and army bases. Because these are combined areas, it is of not too surprising that there are the many transactions here.

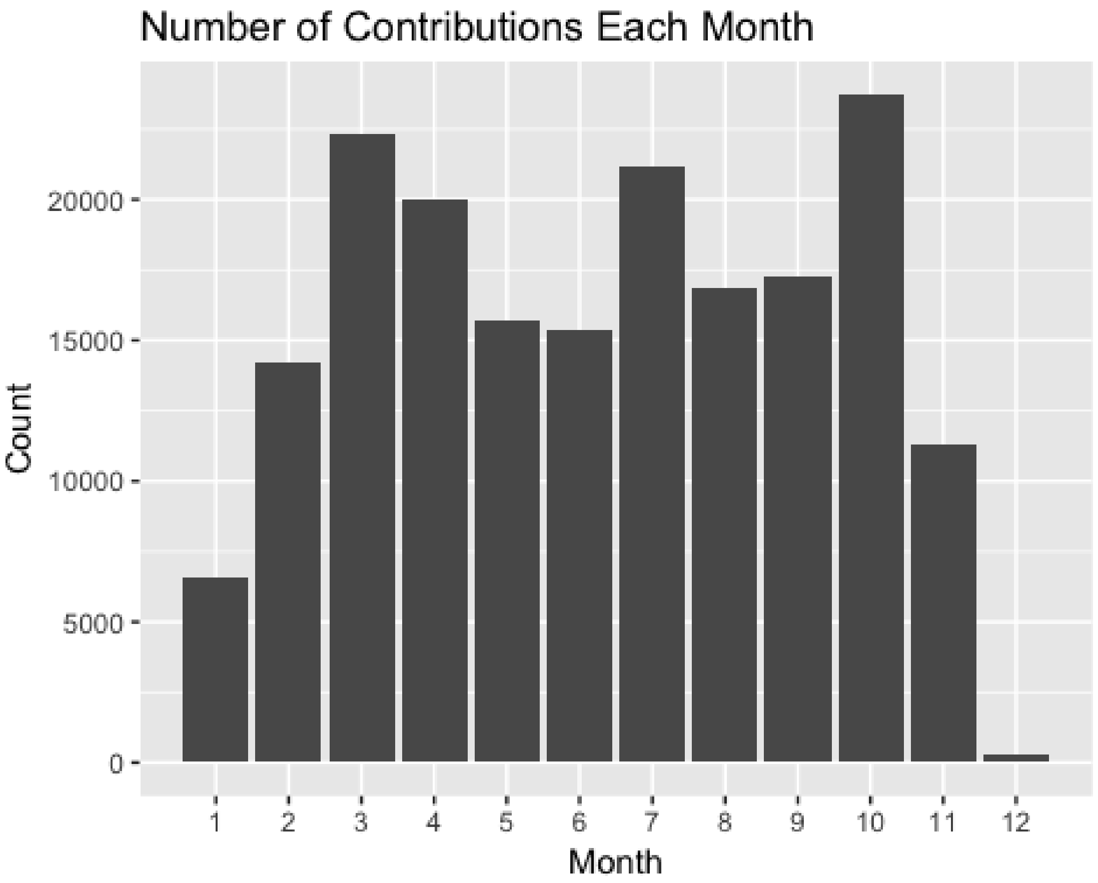
One word of caution about these plots: these states are extremely populated, and it should not be surprising that they have high levels of transactions. CA, NY, and TX have the highest populations in the country (along with FL). Because of this, I'm unsure whether using this data is actually reliable without the added layer of state population (which is not included in this dataset).



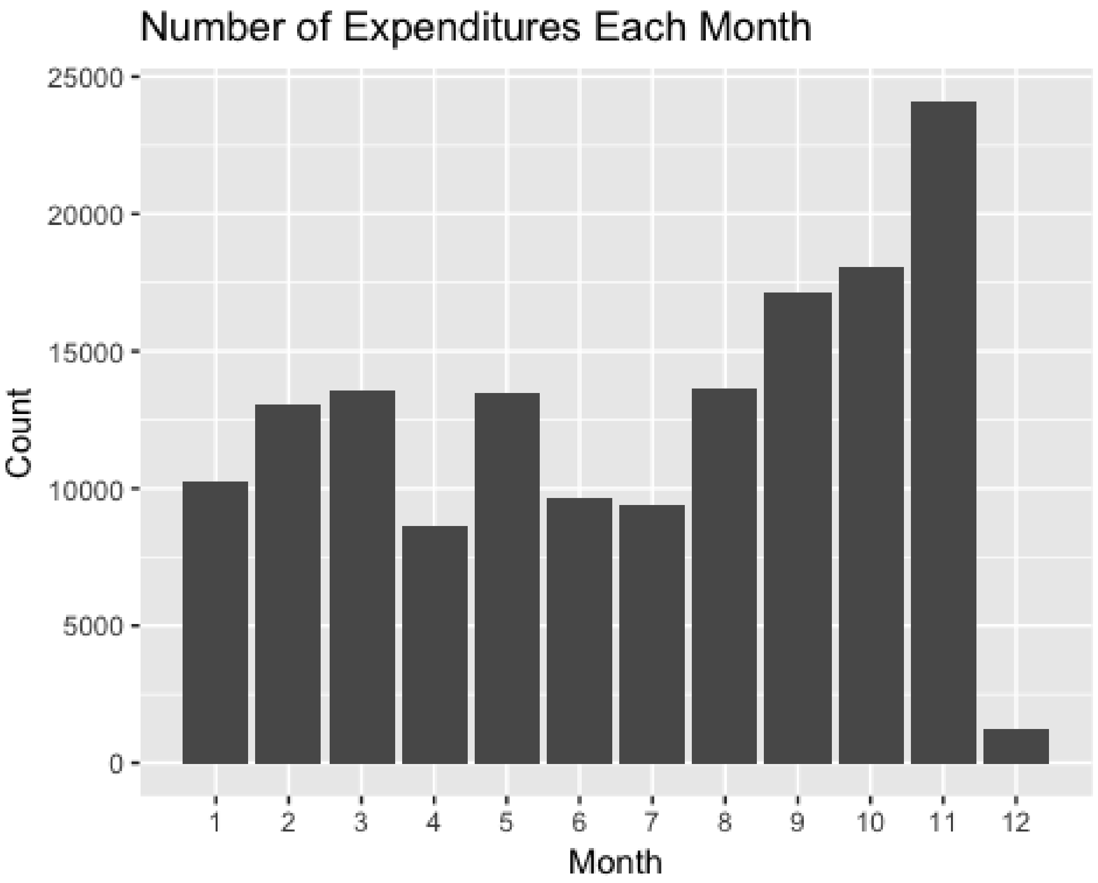
This plot shows the amount of contributions by dollar amount. It appears that most contributions are under 100 dollars, and that there are groups of higher amounts conglomerated around the 500, 1000, and between the 2500 and 3000 dollar marks. Although these could be considered outliers to some degree, I am leaving these data points in the project, so that I can learn about which candidate receives large contributions, etc. If this project were to concentrate on only small donation amounts, it could be preferable to drop these values.



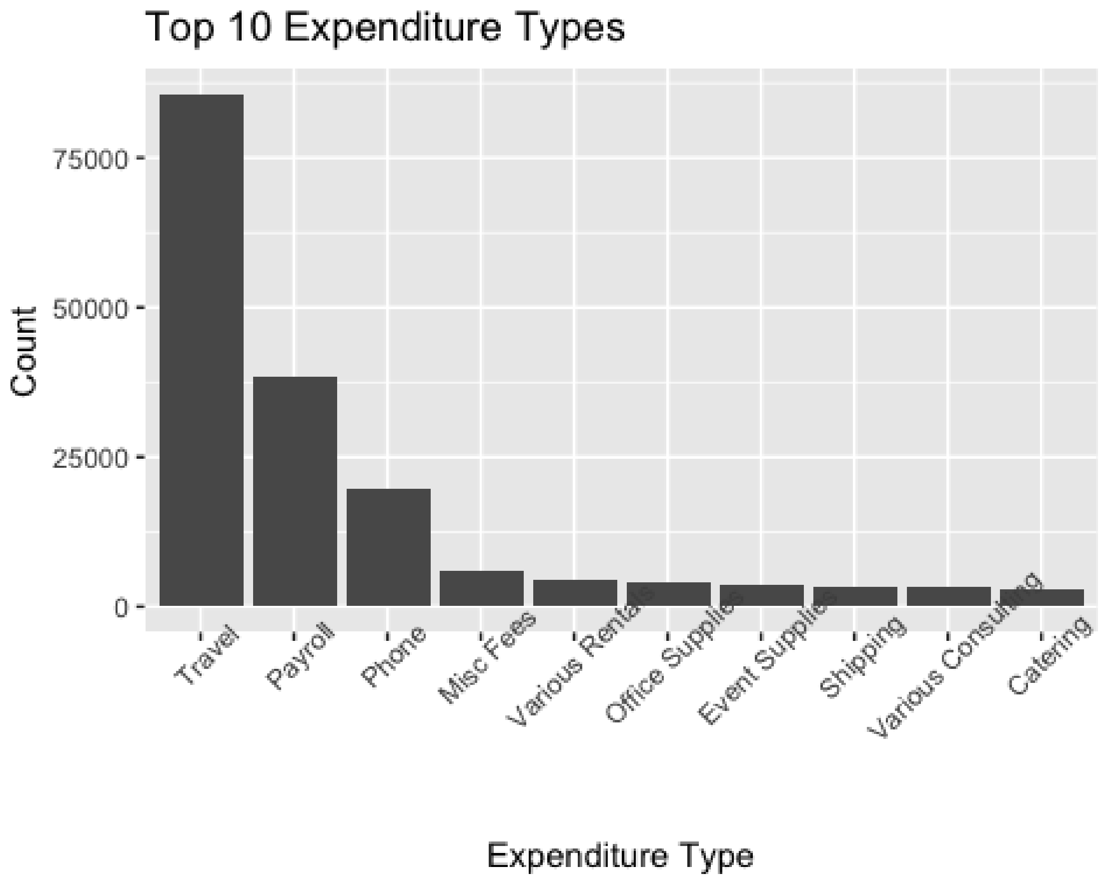
This plot shows the amount of expenditures per dollar amount. Most expenditures tend to be below ~2000 dollars and generally decrease in count as the dollar value increases. The one exception to this is at the rounded dollar marks - 5000, 10000, etc. As with the contributions, I am leaving in these potential outliers because I would like to explore the full picture of expenditures instead of focusing on low expenditure values.



This plot shows that the month with the highest number of contributions is October, followed by March, July, and then April. October contributions would seem to be pre-election contributions. It is interesting to note that the DNC and GOP conventions were both in July, which could account for that surge in contributions.



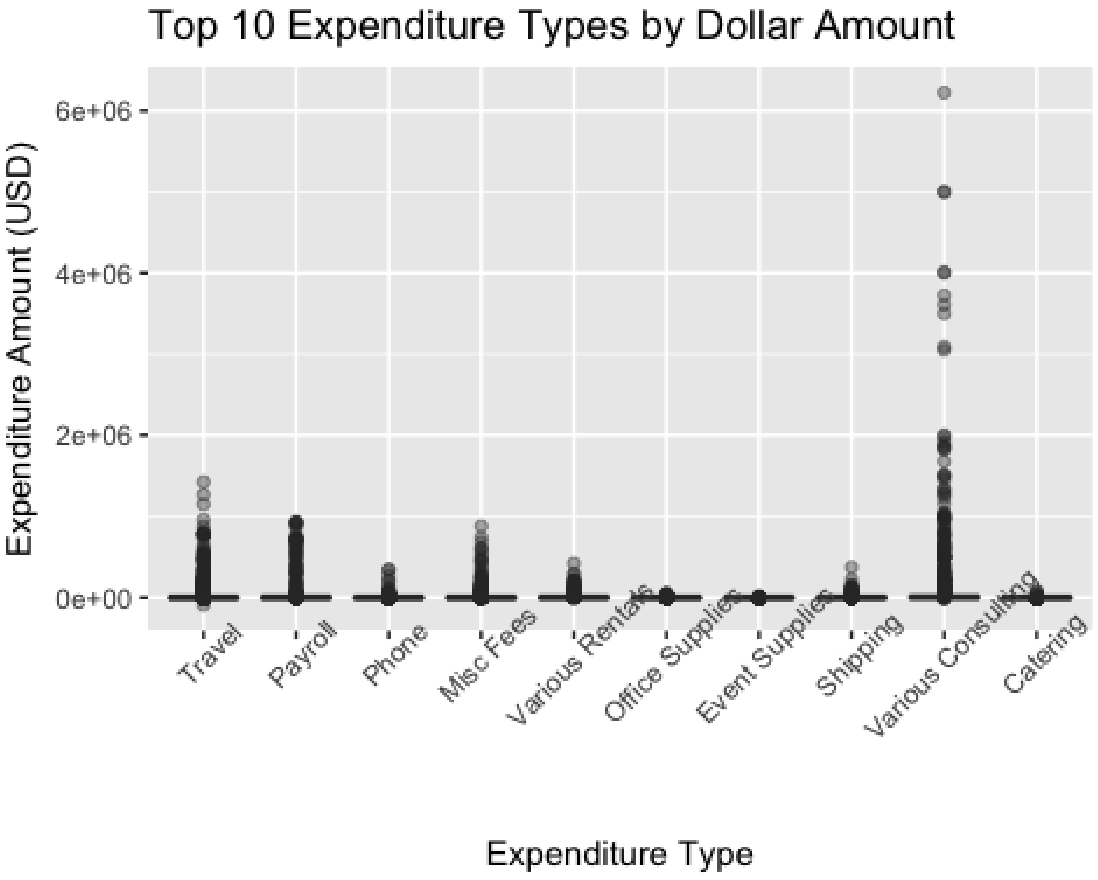
This plot shows that there were generally more expenditures as the 2016 year went on, peaking in November. Since the general election is in November, it makes sense that there would be more expenditures as the candidates reached the end of their campaigns.



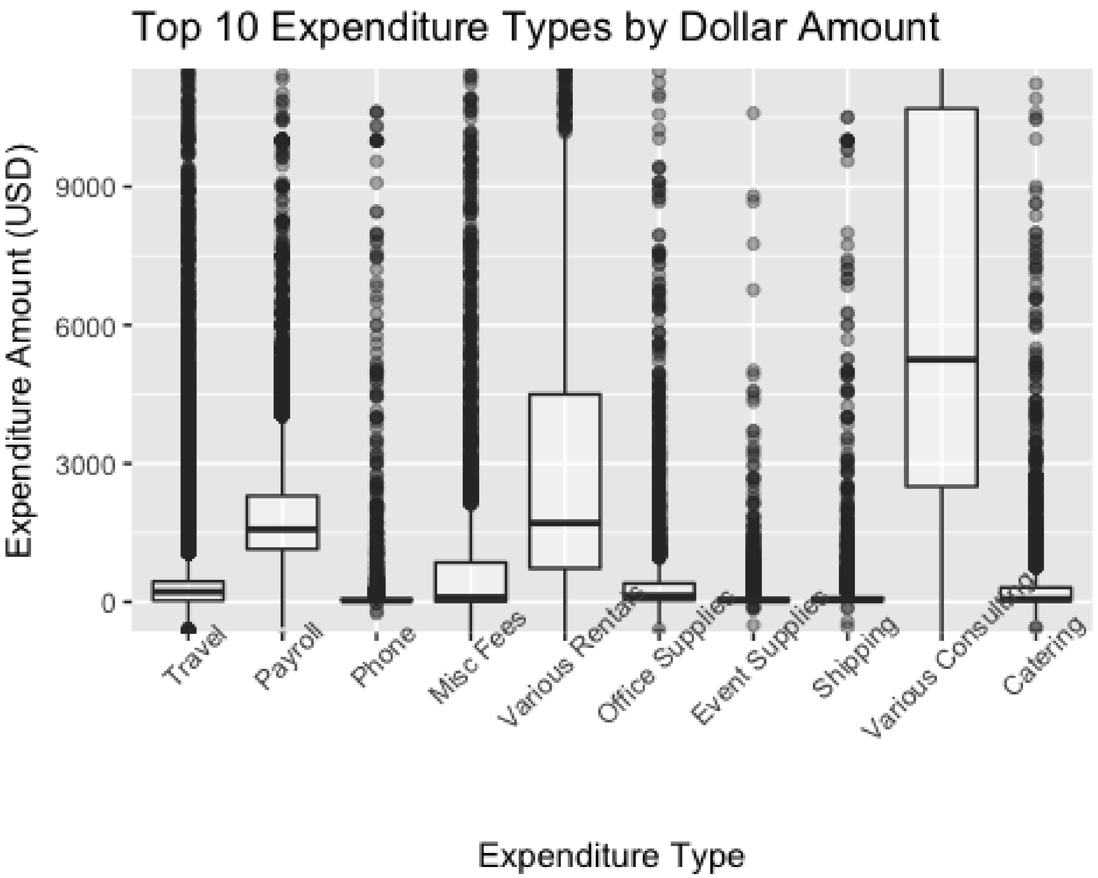
Because there are thousands of types of expenditures in this dataset, the graph with all the data included was impossible to read. To get around this, I've subset the data to so as to only include the top 10 types of transactions in this dataset.

Not surprisingly, the expenditures with the most entries into the data is for travel. Payroll also makes sense, considering the salaries for numerous campaign employees. I am interested, however, to see this data with the amount of money, instead of the count. This plot is below, in the Bivariate graph section.

Bivariate Plots



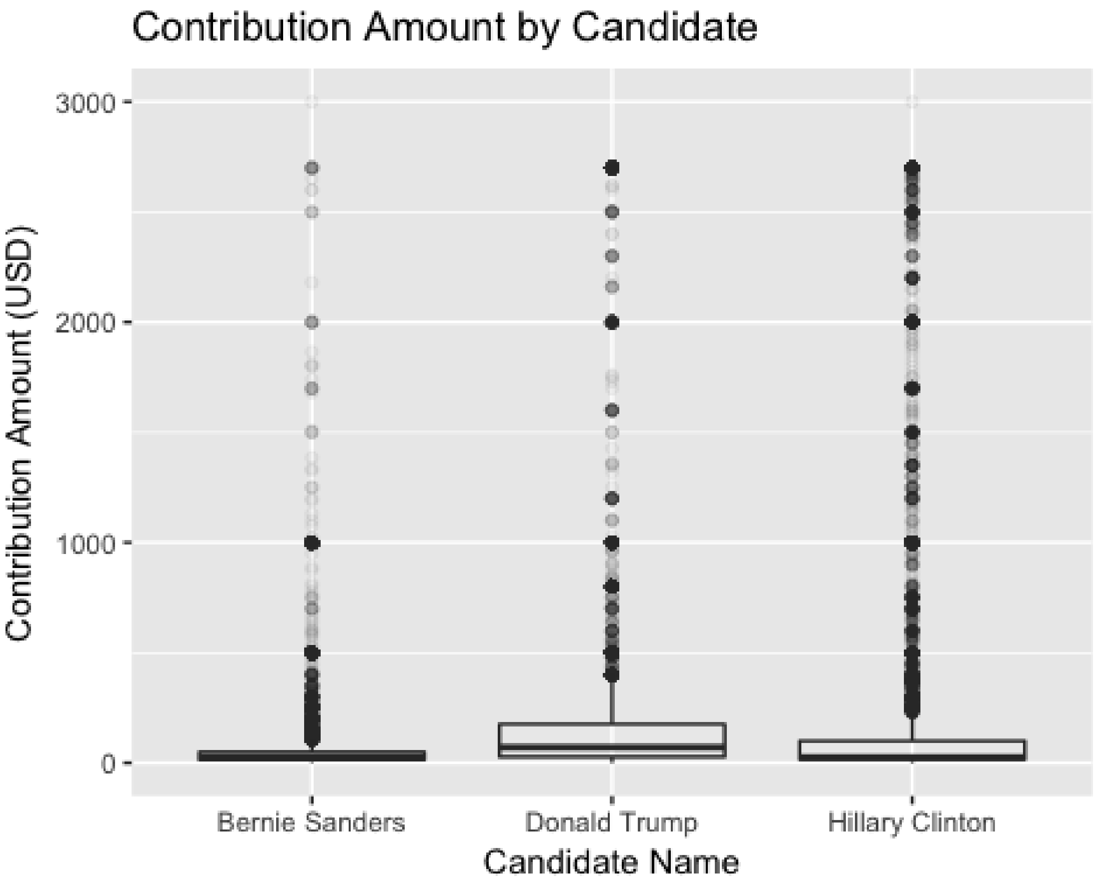
This shows detailed information about the types of expenditures. The notably high expenditures are for Consulting, which makes sense considering that these are the expenditures relating to salaries. I am curious about how this data is delineated across the three candidates. I will revisit this later in the Multivariate graph section. For now, I would like to view only the majority of the spending. The next graph is a zoomed-in version of the above to provide more information of the majority of campaign costs.



It is fascinating that most of the last graph is outliers, and that even with those high costs, the majority of spending for each expenditure type (excepting Consulting) is way under 5,000 dollars.

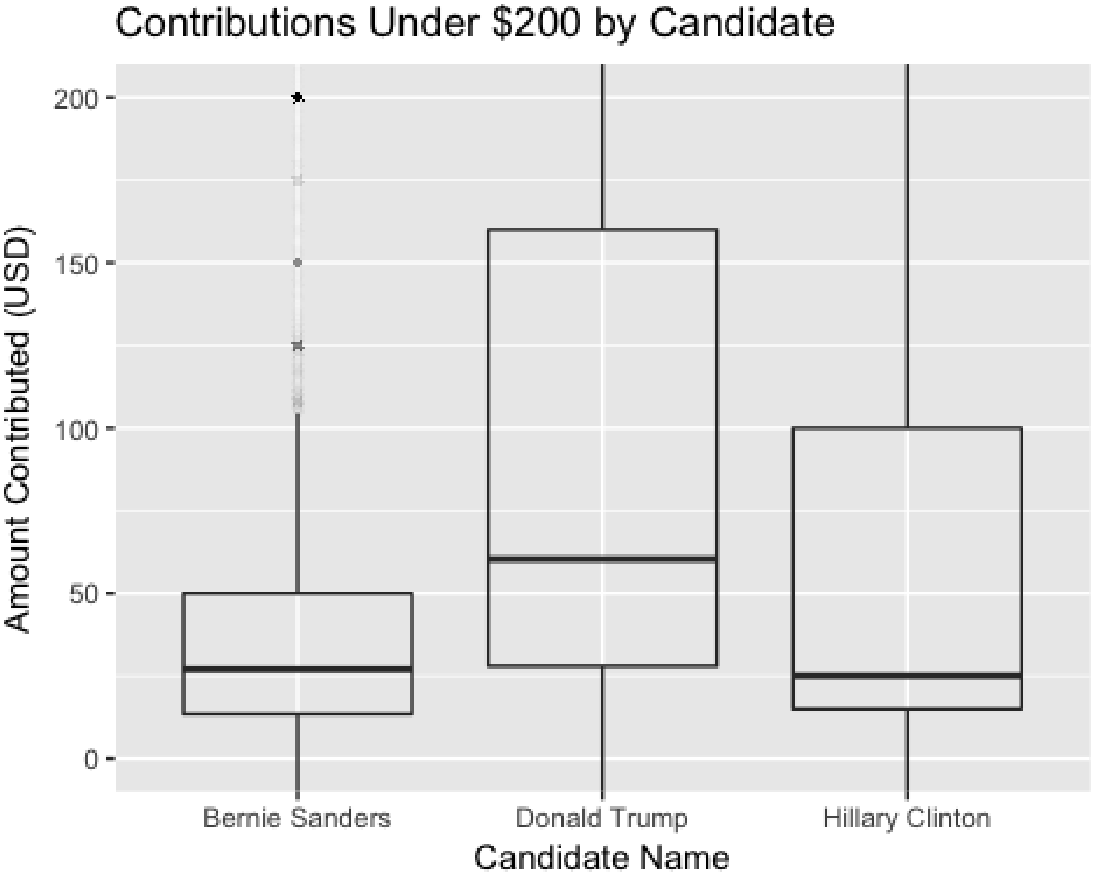
I am also surprised by how low the Payroll costs are. I wonder if these numbers are referring to biweekly paycheck amounts, rather than complete salaries, or if both values are reported, which could be why there is such a large range of costs.

The next group of graphs will look at the dollar amounts of expenditures and contributions by candidate:



The first thing I notice is that there are higher concentrations of donations of a higher value in Donald Trump's and Hillary Clinton's campaign than Bernie Sanders's. I would expect this because Hillary Clinton and Donald Trump became the candidates for their respective parties, and thus, campaigned longer, and also may have garnered more support. This is also consistent with the view of Bernie Sanders's campaign, which was that it was grassroots funded.

There is only so much to see in this graph because it covers such a large range of data. Below is a zoomed-in version. I am curious to look at the small contributions, so I have recreated this graph below with donations equal to and less than 200.

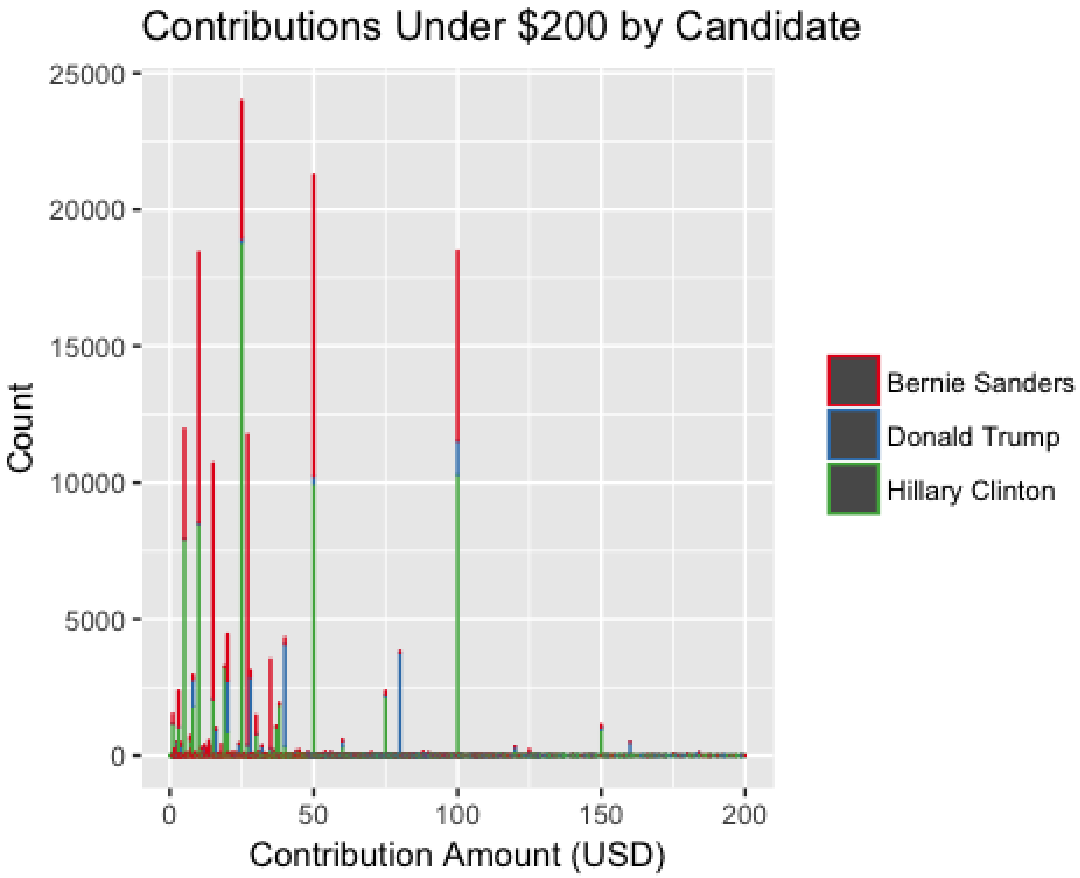


The above graph shows the three candidates' donations less than 200 dollars, with an alpha of 1/500, which means that each point will only be opaque when there are 500 points for that price. It is hard to know exactly who had more small donations overall from this plot. The median for Bernie Sanders and Hillary Clinton is about the same; Hillary Clinton's median looks slightly lower. Bernie Sanders's complete interquartile range is below 50 dollars while Hillary Clinton's caps at 100 dollars. This says to me that they both received many low contribution amounts, but Hillary Clinton also received more high contribution values.

I did some reading about small donations in this election, and I came across some contradictory reports that Donald Trump actually received more money in donations less than $200 than Bernie Sanders or Hillary Clinton. If this is true, I wonder if that would still be true if the value was under 100 dollars. It looks like Donald Trump did receive a lot of small donations, but that the majority of them were over 50, while Hillary Clinton received most of her under 200 dollar contributions under 100, and Bernie Sander's received most of his small contributions under 50 dollars. Thus, if speaking about actual dollar amounts, this could be true, but if referring to the count of small donations, this might not be true anymore. I think that perhaps that statistic is a little misleading, at least from what I see in my data.

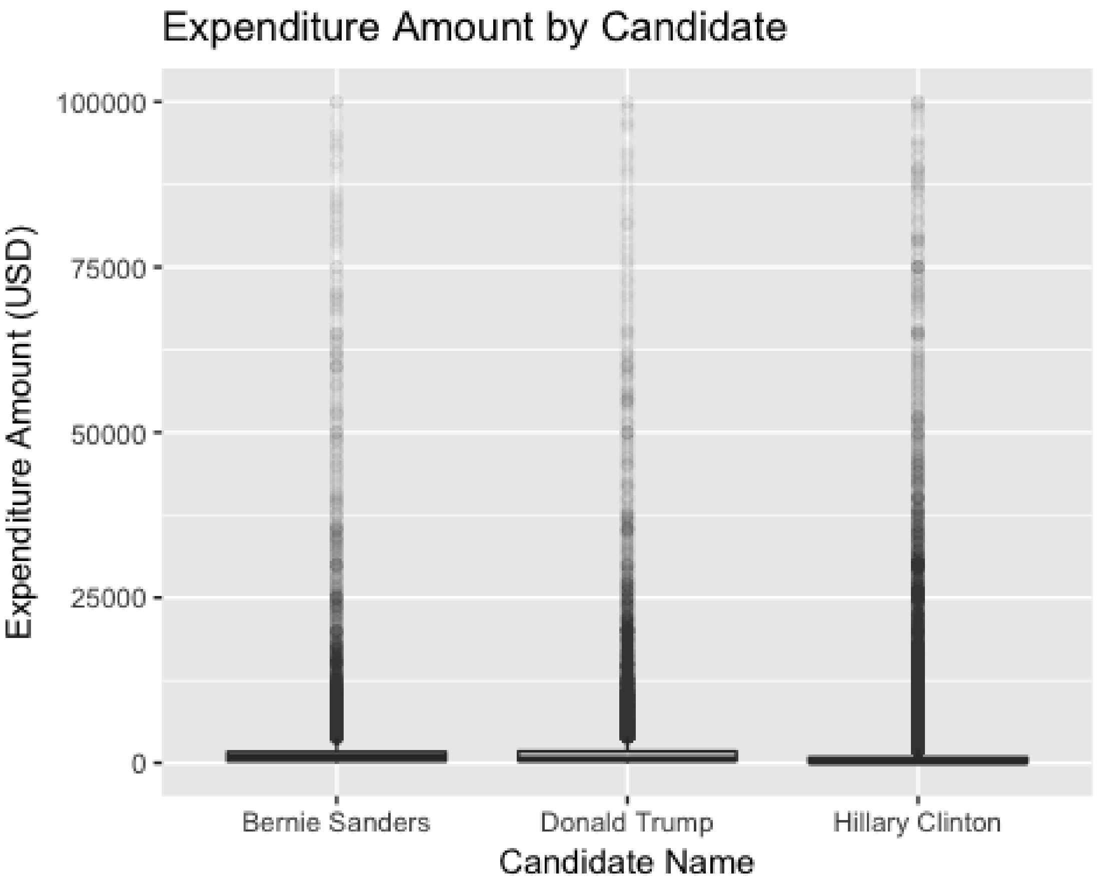
The data in that article is taken from the FEC, which is where the data I am using is derived. Another consideration is that there could be an issue with the integrity of this data, or the one in the article (which can be found in the References section), or with my sample of my data.

I want to see the count of each contribution amount by candidate. I think the following graph is a little easier to read, and will hopefully clarify this contradictory issue.

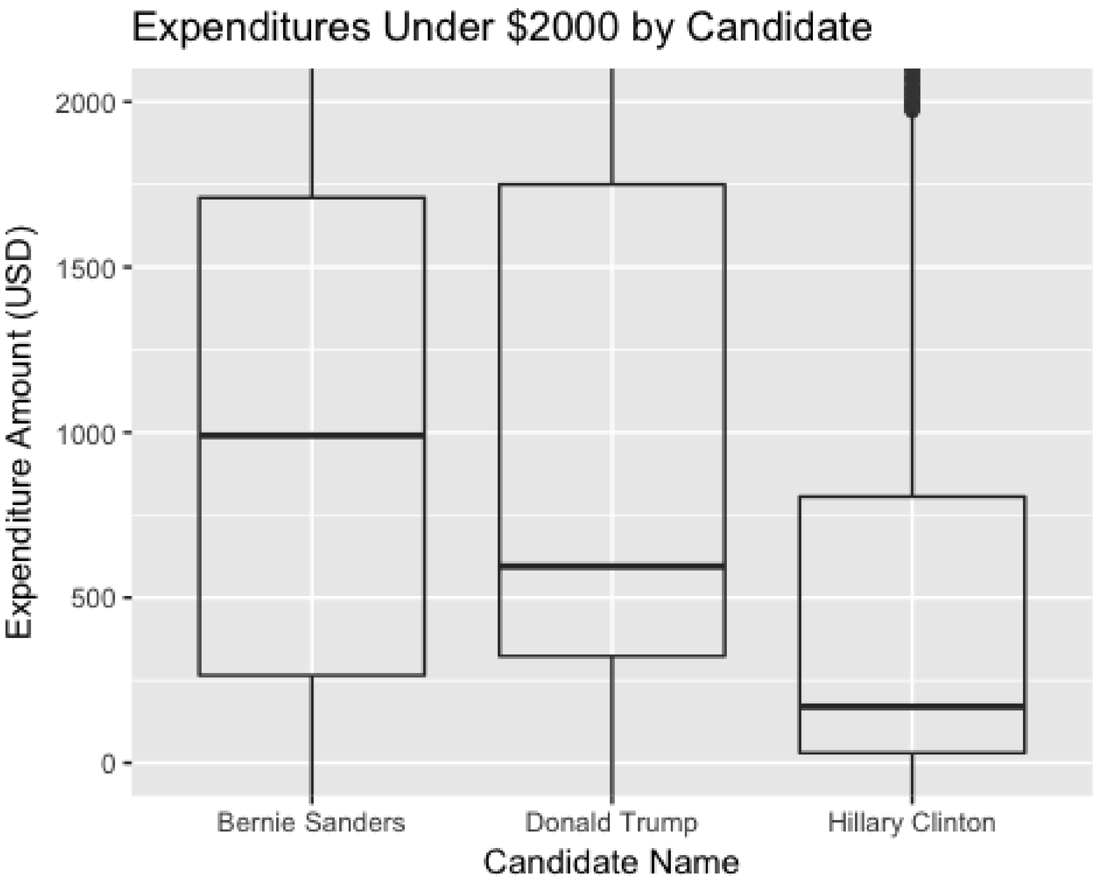


It does not appear that Donald Trump has more money in contributions under $200. I am led to think that there is an issue with the data in the article, the data I am using, or with the sample that I have taken from the FEC data.

The next graph looks at the expenditures for each candidate.

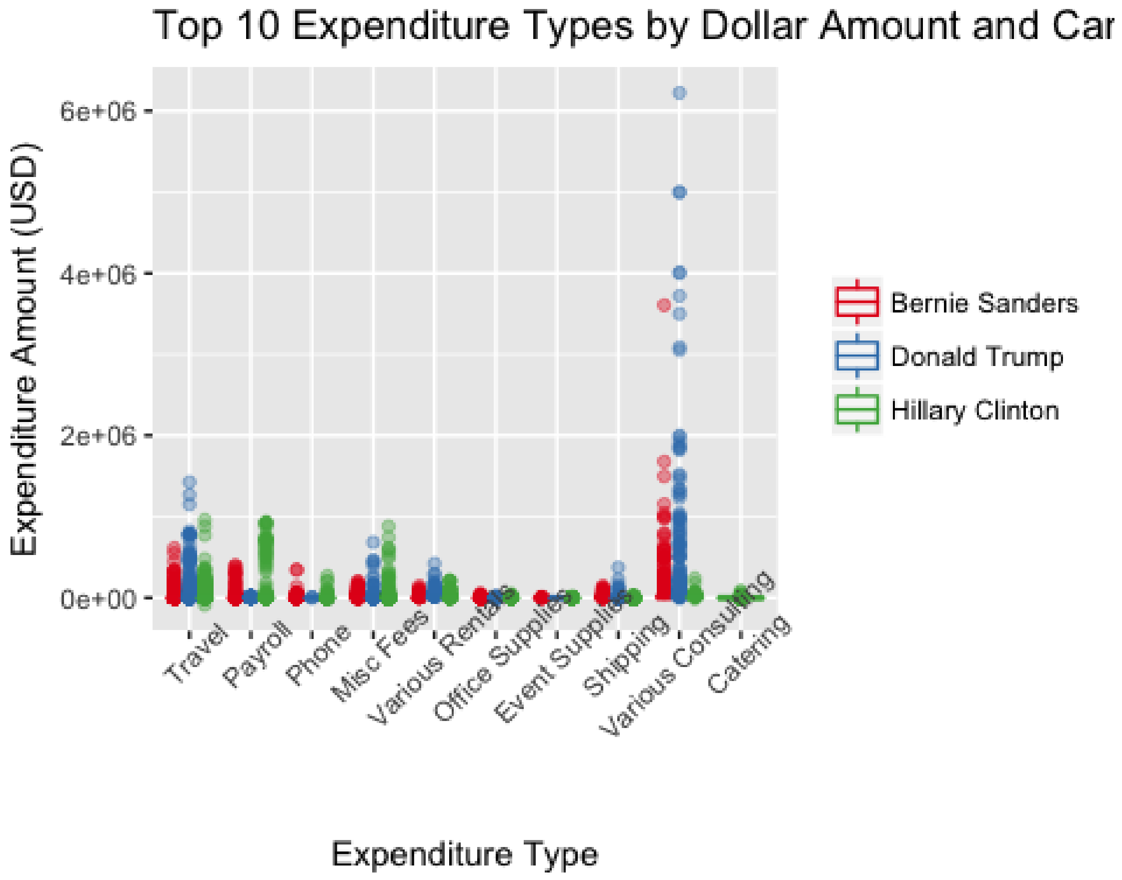


It looks like Hillary Clinton has the most expenditures, but I'd like to recreate this graph a little more zoomed in before drawing any conclusions.

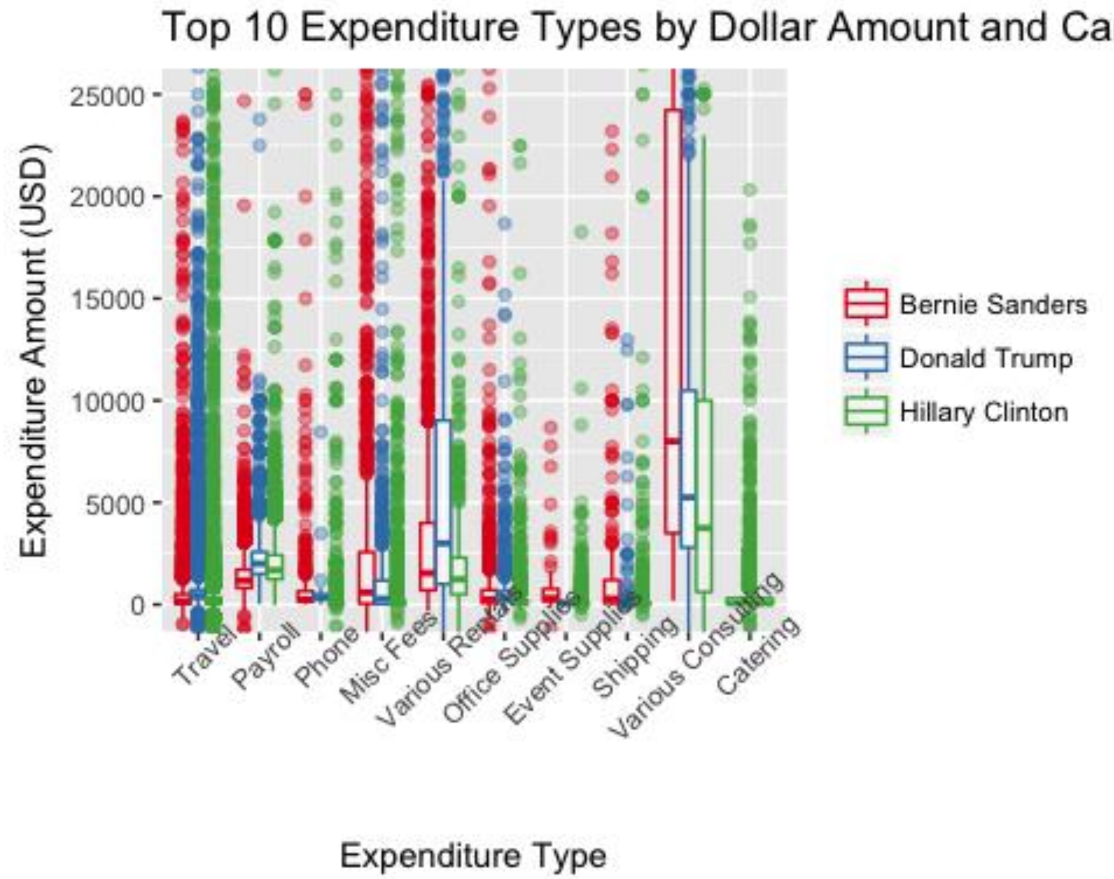


This graph shows campaign spending more clearly. Bernie Sanders has very evenly distributed spending around the 1,000 dollar mark, while Donald Trump and Hillary Clinton have large expenditure values which is pulling up the 75th percentile mark. Even with those high costs, Hillary Clinton and Donald Trump have lower medians than Bernie Sanders, which I think is fascinating.

Multivariate Plots



This plot is recreated from the bivariate section, and shows the kinds of expenditures and the dollar amounts by candidate. This provides a more interesting summary of the previous plot because now there are some striking distributions of candidate spending. The obvious one is that Donald Trump and Bernie Sanders spent much higher amounts of money on consulting. I think this is sort of interesting because both candidates, especially Donald Trump, ran on an “outsider” campaign strategy where he used the fact that he was new to politics as a positive. Because of this, it is not surprising to me that he might have needed to spend higher amounts on consulting than Hillary Clinton, a life-long politician, and who it seems spent smaller amounts on consulting in general. Note: this graph isn’t saying Donald Trump necessarily spent more on consulting than Hillary Clinton, it says that there are higher values of outliers. I am curious to zoom in, so that we can compare the medians by candidate and transaction type.



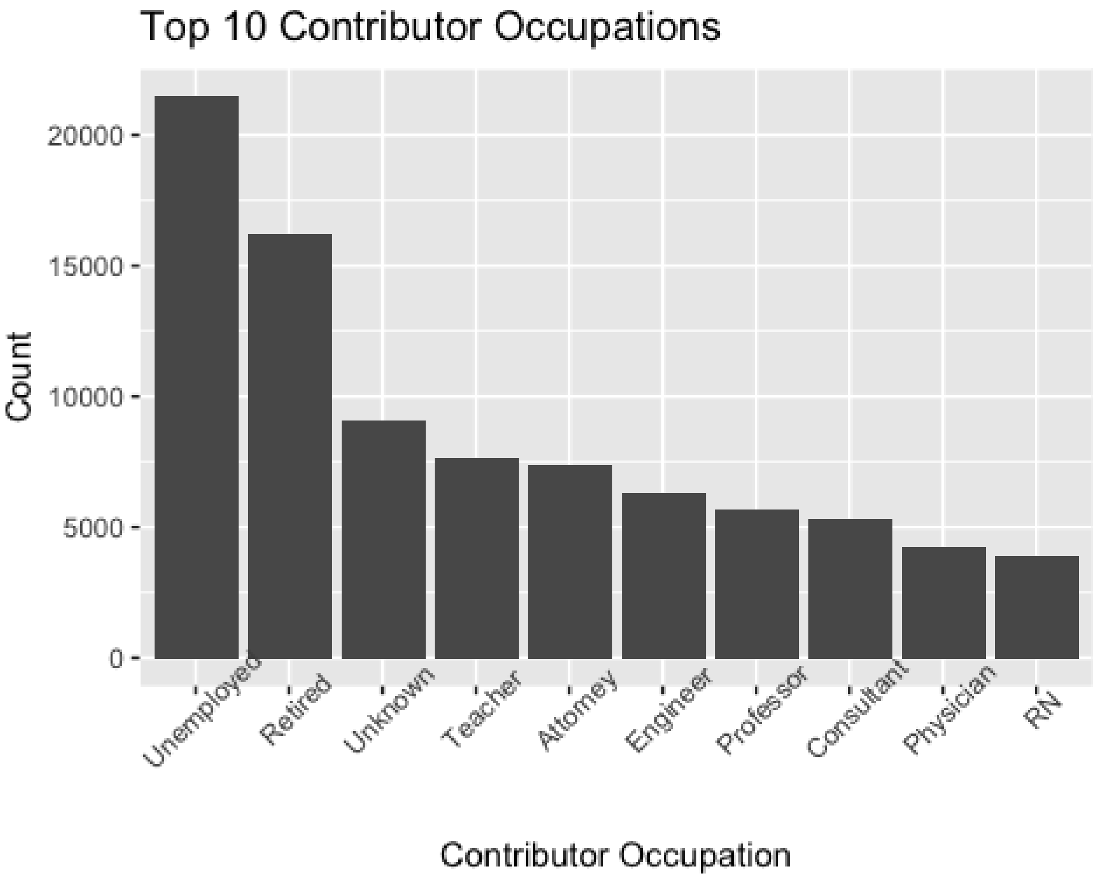
This graph shows more information about the majority of the spending for each candidate. All candidates seem to spend the most on Consulting and Rentals. Although there is a large number of high costs for almost every type (shown in the last graph), it is interesting that the majority of costs for all types (except Consulting and Rentals) are under 2,500 dollars.

The consulting data is clearer now. It appears Bernie Sanders had a higher median for consulting than either of the other candidates, and my guess is that he had less consultants, but paid more for them. Donald Trump has a high concentration of low consulting costs which are pulling the median down. This is not exactly the picture the last graph seemed to show, and I think it is interesting what clarity zooming in on specific data can provide.

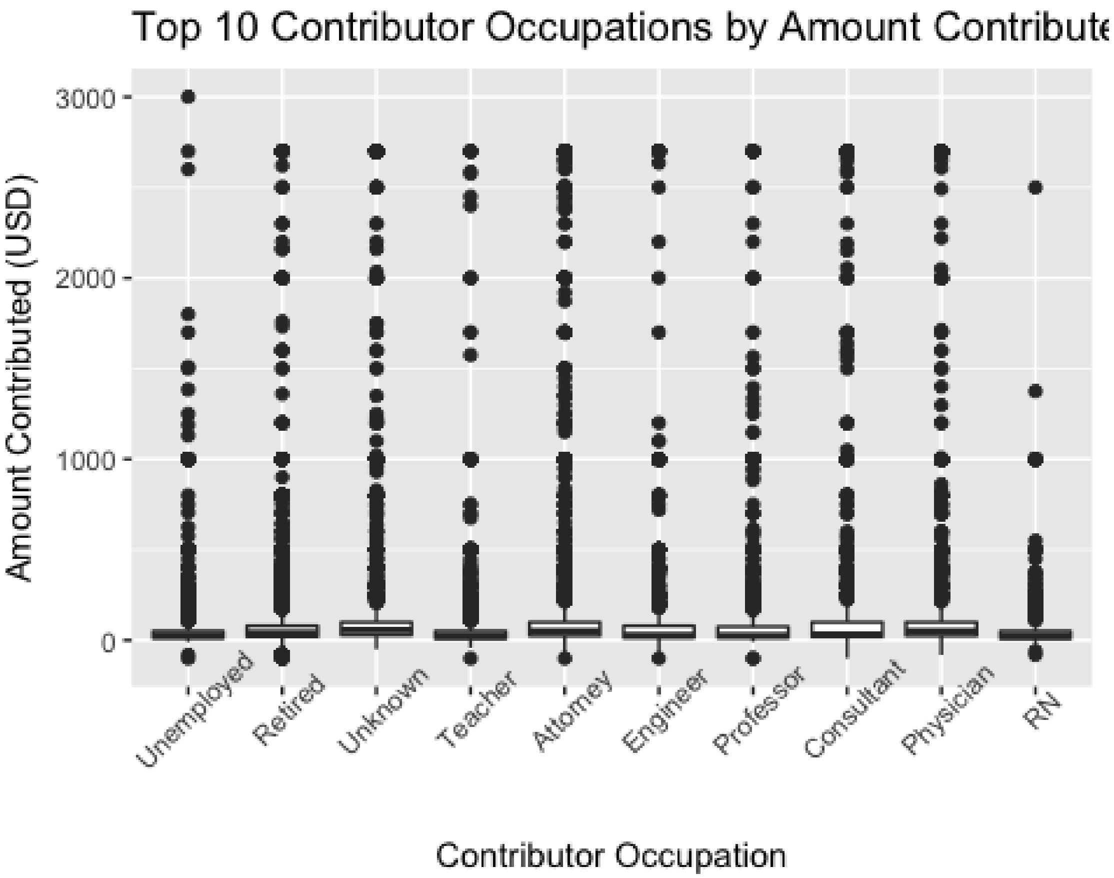
Only Hillary Clinton has value for Catering, which leads me to think that Bernie Sanders and Donald Trump didn't report their catering costs under the name "Catering." That seems more likely to me than the other two candidates spending zero on food for the entire campaign, though I am of course unsure of that. It is striking to me though that there were enough expenditures of this type to push Catering into the top ten expenditure types, even without the other candidates elevating that value.

Final Plots and Summary

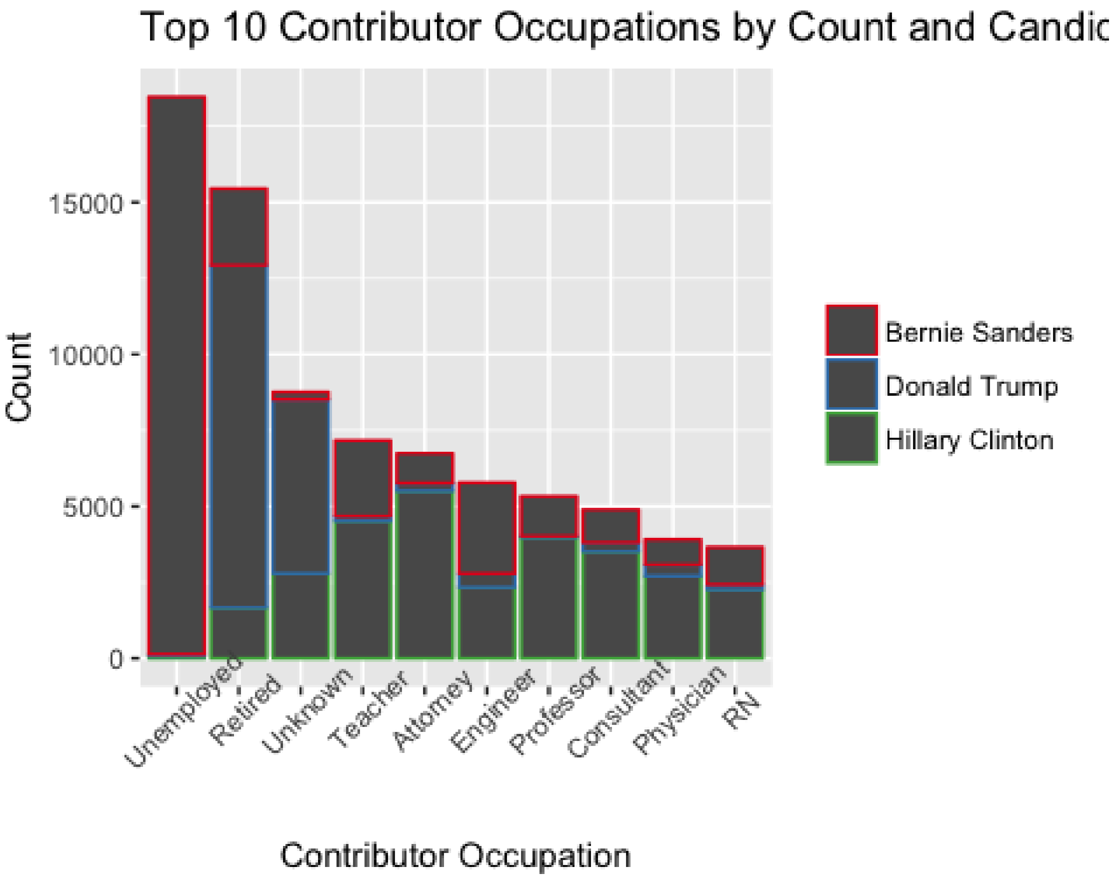
Let's revisit the contributor occupations, and hopefully expound upon the information that has already been plotted. Below is the Univariate plot of the top 10 contributor occupations.



I think it would be useful to look at this information with another variable. Below, is the same plot with the amount contributed added to the y axis.



It’s interesting to see this data plotted, however, even with the alpha value, it is difficult to distinguish the points in areas with a lot of overlap. Most contributions seem to be under 1000 dollars with some larger values spread throughout the occupations, but particularly in the Attorney category. I am curious whether people of any particular occupation tend to contribute to a certain campaign. This plot is created below with count of contributor instead amount contributed.



There is a striking distribution of support in this graph. The most obvious one is that of the Unemployed column, which is mostly made up of Bernie Sanders's contributions. The Retired and Unknown column is very heavily weighted toward Donald Trump, and the rest are made up mostly of Hillary Clinton with some Bernie Sanders support. Considering the platform of Bernie Sanders's, which often seemed to speak to a young/college-aged demographic, and those who were in the need of financial stability, it is interesting that he is supported by those who are not employed. I also wonder if some people marked the "not employed" demarcation, instead of "student," which would make sense with this analysis. I think it would be nice to overlay this with an age variable, but this information was not included in the file.

In my exploration, I also noticed a decent amount of contributions that are less than zero. I'm unsure whether these points are mistakes, or whether they refer to a return of funds, etc. There are also a number of negative expenditures, and I wonder if these could also be refunds of expenditures made.

Conclusion

This analysis has afforded a few points of interest, as well as bringing up a few concerns.

Points of Interest:

1. This analysis demonstrates the individual campaigns' divergent distributions of contribution and expenditure amounts. I was interested in the high levels of small donations in Bernie Sanders's and Hilary Clinton's campaigns, as well as the seemingly contradictory reporting about FEC small donation data in regards to Donald Trump.

2. It suggests that the likelihood of a person's choice to contribute to a certain campaign and their occupation is, at times, extremely correlated. It was surprising to see such a striking distribution of support by occupation!

3. It shows insights about campaign transactions around the country, and that some states received more expenditures and had populations that tended to give more to specific campaigns.

4. It explored the types of expenditures made by each campaign, and that certain candiates had more costs of certain types than others. It was interesting for example to see the amounts Bernie Sanders and Donald Trump spent on consulting, and the issue of Hillary Clinton's campaign and the catering column.

Concerns:

1. This analysis found the contradictory information about Donald Trump's small donation data. Could this simply be due to an issue with the sample used in this project?

2. It found some confusing data points, such as the expenditures and contributions less than zero.

3. The state data might be misleading when considering states that have large populations. In order for this analysis to be more meaningful, the state data would need to be shown in conjunction with population.

4. This project would be more complete if it included information about more candidates, but it was difficult to accomplish this without compromising the integrity of the sample, as the sample size was capped at 200,000 (rStudio coudn't load more than that on my computer).

5. I also have a few concerns about the data quality in general, especially because I am uncertain if the data in the file is complete. I also wonder if sampling from the data could have thrown off the ratios of data included for each candidate. If I were to try this analysis again, I would like to figure out how to efficiently use a larger subset of data, and preferably the entire dataset, as it would insure the most reliable results.

References

Udacity Material

https://www.statmethods.net/graphs/bar.html https://stackoverflow.com/questions/22305023/how-to-get-a-barplot-with-several-variables-side-by-side-grouped-by-a-factor

http://www.sthda.com/english/wiki/ggplot2-axis-ticks-a-guide-to-customize-tick-marks-and-labels

http://moderndata.plot.ly/create-colorful-graphs-in-r-with-rcolorbrewer-and-plotly/

http://www.sthda.com/english/wiki/colors-in-r

http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization

http://www.sthda.com/english/wiki/ggplot2-title-main-axis-and-legend-titles

https://constitutioncenter.org/blog/what-are-the-really-swing-states-in-the-2016-election/

https://www.reddit.com/r/learnpython/comments/5avhzb/how_to_pass_through_multiple_conditions_in_a/?st=jk0oy9zg&sh=87d86c7e

https://stackoverflow.com/questions/17216358/eliminating-nas-from-a-ggplot

https://stackoverflow.com/questions/27422229/how-to-subset-long-dataframe-based-on-top-n-frequent-occurrences-of-variable

https://www.politifact.com/truth-o-meter/statements/2017/nov/13/kayleigh-mcenany/trump-raised-more-dollars-small-donations/

https://en.wikipedia.org/wiki/Bernie_Sanders_presidential_campaign,_2016#Conclusion

https://en.wikipedia.org/wiki/2016_Democratic_National_Convention

https://en.wikipedia.org/wiki/2016_Republican_National_Convention

http://kbroman.org/knitr_knutshell/pages/Rmarkdown.html

https://rmarkdown.rstudio.com/lesson-9.html

https://www.neonscience.org/rmd-use-knitr

https://stackoverflow.com/questions/11714951/remove-extra-legends-in-ggplot2

http://www.cookbook-r.com/Graphs/Legends_(ggplot2)/

https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population

https://www.datacamp.com/community/tutorials/make-histogram-ggplot2

http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization

https://stackoverflow.com/questions/28742870/use-of-scale-x-discrete-in-r-ggplot2 https://community.rstudio.com/t/discrete-x-axis-ticks-in-ggplot2/4102

https://rpubs.com/kaz_yos/ggplot2-axis

https://github.com/tidyverse/ggplot2/issues/1465

https://stats.stackexchange.com/questions/206/what-is-the-difference-between-discrete-data-and-continuous-data

http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization

http://flowingdata.com/2008/02/15/how-to-read-and-use-a-box-and-whisker-plot/

References for Wrangling

Udacity Material

https://stackoverflow.com/questions/20297317/python-dataframe-pandas-drop-column-using-int

https://stackoverflow.com/questions/21902080/python-pandas-not-reading-first-column-from-csv-file

https://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_csv.html

https://chrisalbon.com/python/data_wrangling/pandas_saving_dataframe_as_csv/

https://stackoverflow.com/questions/27060098/replacing-few-values-in-a-pandas-dataframe-column-with-another-value

https://stackoverflow.com/questions/11346283/renaming-columns-in-pandas

https://stackoverflow.com/questions/44068913/python-pandas-seed-for-random-generator

https://stackoverflow.com/questions/38085547/random-sample-of-a-subset-of-a-dataframe-in-pandas

https://stackoverflow.com/questions/12555323/adding-new-column-to-existing-dataframe-in-python-pandas

https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.groupby.html

https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.sort_values.html

https://stackoverflow.com/questions/34913546/remove-low-counts-from-pandas-data-frame-column-on-condition

https://stackoverflow.com/questions/35678083/pandas-delete-rows-of-a-dataframe-if-total-count-of-a-particular-column-occurs

https://erikrood.com/Python_References/rows_cols_python.html

https://datascience.stackexchange.com/questions/12645/how-to-count-the-number-of-missing-values-in-each-row-in-pandas-dataframe

https://stackoverflow.com/questions/13413590/how-to-drop-rows-of-pandas-dataframe-whose-value-in-certain-columns-is-nan