Financial Contributions to Presidential Campaigns in 2016

In this report, we're going to explore the Financial Contributions to president candidates of the 2016 U.S. presidential election. At each individual section of this report, we're going to answer different types of questions, and we'll refine our assumptions as we advance through the analysis. Some of the questions we're going to answer by the end of this report:

- Which candidate received the most number of contributions?
- Which candidate received the highest amount?
- What are the most common occupations from the contributors?
- Which location has the highest amount contributed? To which candidate?
- How the contributions vary over time?
- Does the contribution amount have an impact on the final result of the election?
- The profile of contributors for each candidate (top 3)
- and more.

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    649460 obs. of  18 variables:
##  $ comitee_id               : chr  "C00575795" "C00575795" "C00577130" "C00577130" ...
##  $ candidate_id             : chr  "P00003392" "P00003392" "P60007168" "P60007168" ...
##  $ candidate_name           : chr  "Clinton, Hillary Rodham" "Clinton, Hillary Rodham" "Sanders, Be
##  $ contributor_name         : chr  "JONES TAKATA, LOUISE" "CODY, ERIN" "KEITH, SUSAN H" "LEPAGE, WI
##  $ contributor_city         : chr  "NEW YORK" "BUFFALO" "NEW YORK" "BROOKLYN" ...
##  $ contributor_state        : chr  "NY" "NY" "NY" "NY" ...
##  $ contributor_zip_code     : int  100162783 142221910 100133107 112381202 129011729 110402014 1177
##  $ contributor_employer     : chr  "N/A" "RUPP BAASE PFALZGRAF CUNNINGHAM LLC" "NOT EMPLOYED" "NEW
##  $ contributor_occupation   : chr  "RETIRED" "ATTORNEY" "NOT EMPLOYED" "UNDERGRADUATE ADMINISTRATOR
##  $ contributor_receipt_amount: num  100 67 50 15 100 ...
##  $ contributor_receipt_date : chr  "15-APR-16" "24-APR-16" "06-MAR-16" "04-MAR-16" ...
##  $ receipt_description      : chr  NA NA NA NA ...
##  $ memo_code                : chr  "X" "X" NA NA ...
##  $ memo_text                : chr  "* HILLARY VICTORY FUND" "* HILLARY VICTORY FUND" "* EARMARKED C
##  $ form_type                : chr  "SA18" "SA18" "SA17A" "SA17A" ...
##  $ file_number              : int  1091718 1091718 1077404 1077404 1091718 1077404 1077404 1091718
##  $ transaction_id           : chr  "C4732422" "C4752463" "VPF7BKZ1KR1" "VPF7BKWHRY0" ...
##  $ election_type            : chr  "P2016" "P2016" "P2016" "P2016" ...
##  - attr(*, "problems")=Classes 'tbl_df', 'tbl' and 'data.frame': 649463 obs. of  5 variables:
##   ..$ row     : int  1 2 3 4 5 6 7 8 9 10 ...
##   ..$ col     : chr  NA NA NA NA ...
##   ..$ expected: chr  "18 columns" "18 columns" "18 columns" "18 columns" ...
##   ..$ actual  : chr  "19 columns" "19 columns" "19 columns" "19 columns" ...
##   ..$ file    : chr  "'Data/P00000001-NY.csv'" "'Data/P00000001-NY.csv'" "'Data/P00000001-NY.csv'" "
##  - attr(*, "spec")=List of 2
##   ..$ cols   :List of 18
##   .. ..$ cmte_id          : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ cand_id          : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ cand_nm          : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ contbr_nm        : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ contbr_city      : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ contbr_st        : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
```

```
##   .. ..$ contbr_zip       : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ contbr_employer  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ contbr_occupation: list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ contb_receipt_amt: list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ contb_receipt_dt : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ receipt_desc     : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ memo_cd          : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ memo_text        : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ form_tp          : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ file_num         : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ tran_id          : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ election_tp      : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   ..$ default: list()
##   .. ..- attr(*, "class")= chr  "collector_guess" "collector"
##   ..- attr(*, "class")= chr "col_spec"
```

```
## # A tibble: 6 x 18
##   comitee_id candidate_id candidate_name contributor_name contributor_city
##   <chr>      <chr>        <chr>          <chr>            <chr>
## 1 C00575795  P00003392    Clinton, Hill~ JONES TAKATA, L~ NEW YORK
## 2 C00575795  P00003392    Clinton, Hill~ CODY, ERIN       BUFFALO
## 3 C00577130  P60007168    Sanders, Bern~ KEITH, SUSAN H   NEW YORK
## 4 C00577130  P60007168    Sanders, Bern~ LEPAGE, WILLIAM  BROOKLYN
## 5 C00575795  P00003392    Clinton, Hill~ BIELAT, VEDORA   PLATTSBURGH
## 6 C00577130  P60007168    Sanders, Bern~ LIEBER, MICHAEL  NEW HYDE PARK
## # ... with 13 more variables: contributor_state <chr>,
## #   contributor_zip_code <int>, contributor_employer <chr>,
## #   contributor_occupation <chr>, contributor_receipt_amount <dbl>,
## #   contributor_receipt_date <chr>, receipt_description <chr>,
## #   memo_code <chr>, memo_text <chr>, form_type <chr>, file_number <int>,
## #   transaction_id <chr>, election_type <chr>
```

Data Wrangling & Manipulation

At this section, we're going to perform some data manipulation in order to transform our dataset in a cleaner, more informative data structure.

```
# Converting string dates to date objects to be able to perform date operations
financial_contrib$contributor_receipt_date <- parse_date_time(x = financial_contrib$contributor_receipt_
                orders = c("%d-%b-%y"))
```

```r
# Extracting day from date
financial_contrib$receipt_day <- format(financial_contrib$contributor_receipt_date, "%d")

# Extracting month from date
financial_contrib$receipt_month <- format(financial_contrib$contributor_receipt_date, "%m")

# Extracting year from date
financial_contrib$receipt_year <- format(financial_contrib$contributor_receipt_date, "%Y")

# Converting contribution amount to double
financial_contrib$contributor_receipt_amount <- as.double(financial_contrib$contributor_receipt_amount)

# Converting file number to integer
financial_contrib$file_number <- as.numeric(financial_contrib$file_number)

# Converting occupation to factor
financial_contrib$contributor_occupation <- as.factor(financial_contrib$contributor_occupation)

# Converting city to factor
financial_contrib$contributor_city <- as.factor(financial_contrib$contributor_city)

# Converting election type to factor
financial_contrib$election_type <- as.factor(financial_contrib$election_type)

# Converting employer to factor
financial_contrib$contributor_employer <- as.factor(financial_contrib$contributor_employer)

# Converting zipcode to string
financial_contrib$contributor_zip_code <- as.character(financial_contrib$contributor_zip_code)

# Recover unique cities to geocode
unique_cities_df <- data.frame(city = unique(financial_contrib$contributor_city))
unique_cities_df$city <- as.character(unique_cities_df$city)

# Getting the name of unique locations
unique_cities_df <- na.omit(unique_cities_df)
```

Uncomment this chunk to download geolocation data

```r
# Calling geolocation service to obtain lat and lon
#geocode_output_df <- geocode(unique_cities_df$city,
 #   output = 'latlona',
 #   source = 'dsk',
 #   messaging = FALSE,
 #   sensor = FALSE)

# Changing the 'address' column name to match the original dataframe in order to join the data
#colnames(geocode_output_df)[which(names(geocode_output_df) == "address")] <- "contributor_city"

# Omiting NA values
#geocode_output_df <- na.omit(geocode_output_df)
```

Joining dataframes by `contributur_city` key. Uncomment this chunk to join dataframes with the geolocation data

```
# Joining original dataset with geolocation dataframe by 'contributor_city' key
#financial_contrib <- left_join(financial_contrib, geocode_output_df, by='contributor_city')

# Appending the newly created columns to the original ones
#new_columns <- append(columns_to_select, c('receipt_day', 'receipt_month', 'receipt_year',
#                                           'lon', 'lat'))

# Selecting the columns of interest
#financial_contrib <- select(financial_contrib, new_columns)

# Saving the transformed dataset to avoid iterating over all this process again
#write.csv(financial_contrib, 'Data/financial_contributions.csv')
```

Loading the transformed dataset

Disclaimer: In order to make the process faster, I'll provide the transformed dataset along with my code. If you uncomment the 2 previous chunk above, the result will be the same, but it will take a long time to get the geocode data from Google service. As I already had to go through this process to compile this report, I am providing the full dataset.

Loading the transformed data set and extracting a sample from it, considering the original size of the dataset and performance issues.

```
# Loading the transformed dataset to explore on the next sessions
financial_contributions_df <- read.csv('Data/financial_contributions.csv', as.is = TRUE)

# Setting seef for reproducibility
set.seed(1024)

# Making a sample to address performance issues
sample_financial_dataset <- financial_contributions_df[sample(length(financial_contributions_df$contribu

# Taking a look at the top records of the transformed dataset
head(sample_financial_dataset)
```

```
##             X comitee_id candidate_id          candidate_name
## 141641 141641  C00575795    P00003392 Clinton, Hillary Rodham
## 641428 641428  C00575795    P00003392 Clinton, Hillary Rodham
## 226312 226312  C00575795    P00003392 Clinton, Hillary Rodham
## 247474 247474  C00575795    P00003392 Clinton, Hillary Rodham
## 13630   13630  C00575795    P00003392 Clinton, Hillary Rodham
## 486914 486914  C00575795    P00003392 Clinton, Hillary Rodham
##           contributor_name contributor_city contributor_state
## 141641      HERKALO, KEITH             PERU                NY
## 641428        RIGBY, PETER         NEW YORK                NY
## 226312         COHEN, NOEL         NEW YORK                NY
## 247474        LAPMAN, LISA         NEW YORK                NY
## 13630         WIZMUR, DINA         NEW YORK                NY
## 486914 DAVIS-FARAGE, KAREN         NEW YORK                NY
##        contributor_zip_code            contributor_employer
## 141641            129725101                             N/A
```

```
## 641428            100246514 STANHOPE REED RIGBY ADVISORY LLC
## 226312            100166416                              N/A
## 247474            100656465         MONTEFIORE HEALTH SYSTEM
## 13630             100237103  INDEPENDENCE HEALTH CARE SYSTEM
## 486914            100241100            POLE POSITION RACEWAY
##        contributor_occupation contributor_receipt_amount
## 141641               RETIRED                         50
## 641428             CONSULTANT                        100
## 226312               RETIRED                         25
## 247474              PHYSICIAN                         25
## 13630                ATTORNEY                         25
## 486914                 OWNER                        100
##        contributor_receipt_date receipt_description memo_code memo_text
## 141641               2016-05-12                <NA>      <NA>      <NA>
## 641428               2016-11-06                <NA>      <NA>      <NA>
## 226312               2016-05-11                <NA>      <NA>      <NA>
## 247474               2015-10-13                <NA>      <NA>      <NA>
## 13630                2016-08-29                <NA>      <NA>      <NA>
## 486914               2016-11-05                <NA>      <NA>      <NA>
##        form_type file_number transaction_id election_type receipt_day
## 141641     SA17A     1091720       C4962760         P2016          12
## 641428     SA17A     1133832      C15640955         G2016           6
## 226312     SA17A     1091720       C4958526         P2016          11
## 247474     SA17A     1081052       C1356796         P2016          13
## 13630      SA17A     1126762       C9682962         G2016          29
## 486914     SA17A     1133832      C15486739         G2016           5
##        receipt_month receipt_year       lon       lat
## 141641             5         2016 -76.00000 -10.00000
## 641428            11         2016 -74.00597  40.71427
## 226312             5         2016 -74.00597  40.71427
## 247474            10         2015 -74.00597  40.71427
## 13630              8         2016 -74.00597  40.71427
## 486914            11         2016 -74.00597  40.71427
```

Data visualization

The dataset we are going to explore provides data about individual financial contributions to presidentials campaigns of 2016 U.S. presidential election. We are going to start by analysing the structure of some variables of interest such as `contributor_receipt_amount`.

Univariate Plots Section

Summary of the contribution amounts. We notice that there is a negative value (which may indicate refunds made to some reported individuals, as stated on the dataset documentation) and the Max value is clearly an extreme value.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -5400.0    15.0    27.0   139.6   100.0  5400.0
```

Let's explore the datetime data that we extracted from the `contributor_receipt_date` feature. That step was very important because now we can have a sense of how the data is distributed

over time, which it would't be possible by the provided date format. In the next section, we're going to explore the summary statistics for contributions by day and month. It wouldn't make sense for us to analyze year data, since most of the contributions account for the election year itself (2016). The cell bellow confirm this assumption.

```
##
##  2015  2016
##  1848 18152
```

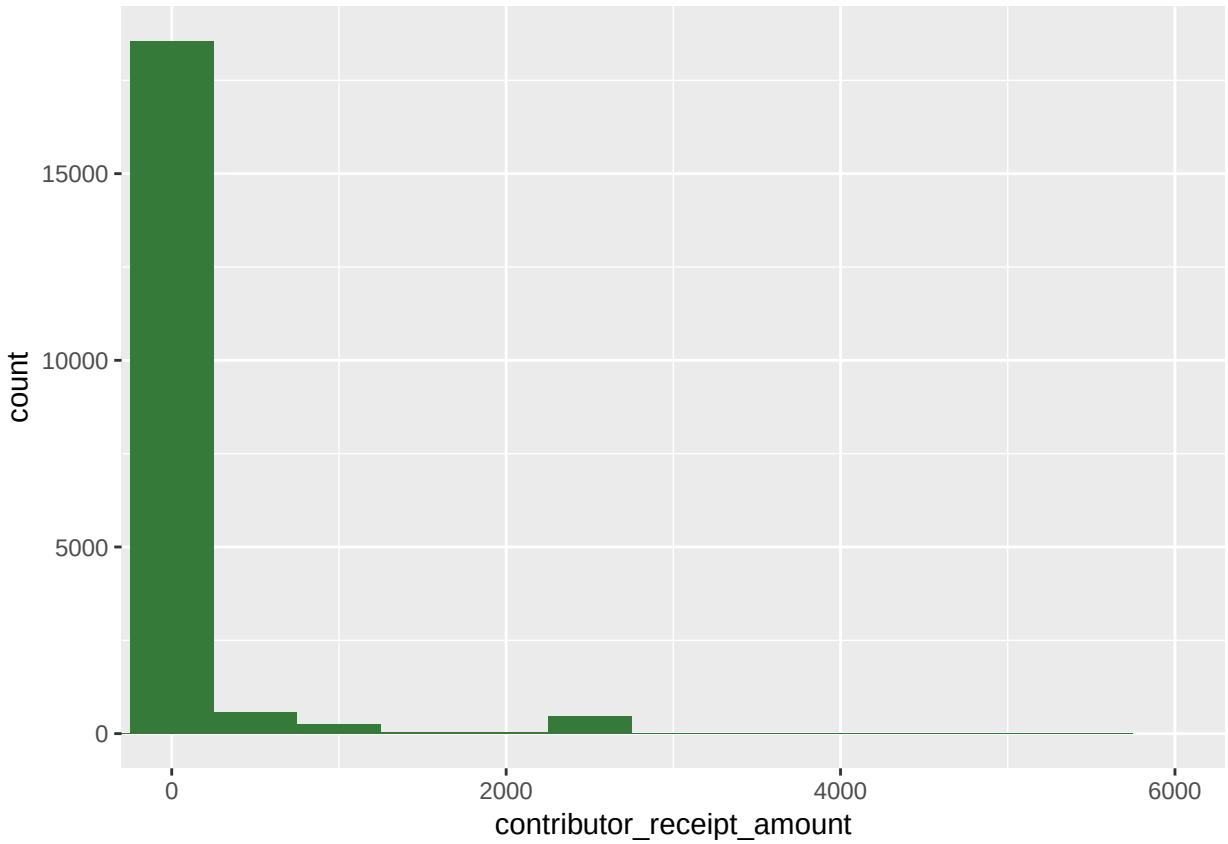
Counting contributions made at each day of the month. Does the number of contribution increase as we get closer to the end of the month? It seems to be the case.

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
##  598  617  529  668  657  718  694  817  729  603  506  610  505  666  555
##   16   17   18   19   20   21   22   23   24   25   26   27   28   29   30
##  551  481  533  632  592  606  492  555  502  543  748  646  823 1073  959
##   31
##  792
```

We can apply the same counting to each month. For the months, it seems it's a more spread distribution, having a lot of contributions on MARCH, APRIL, SEPTEMBER, and OCTOBER.

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12
##  636 1396 1959 1998 1463 1417 1794 1657 2340 3230 1632  478
```

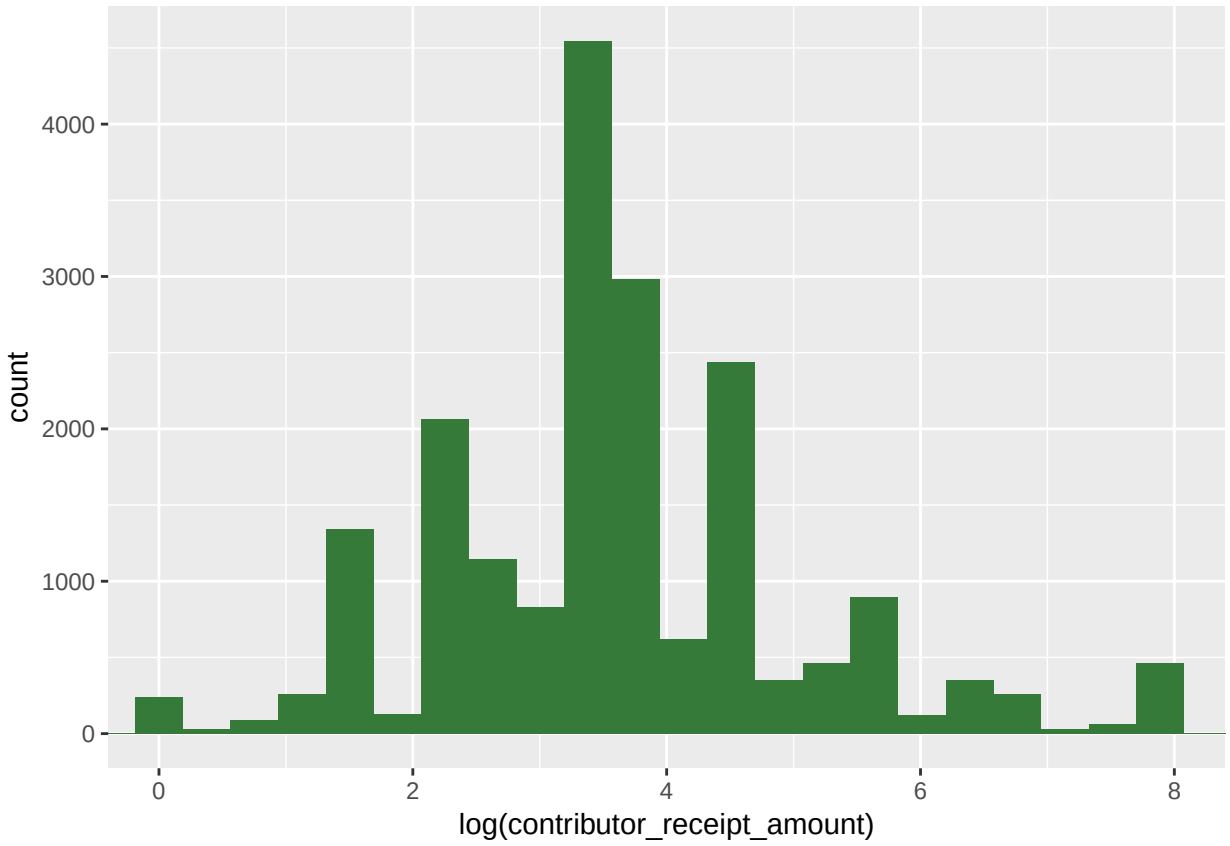
```
## [1] -5400  5400
```



The data for the reported contribution amount is very right skewed. Let's apply a `log10` transformation to get it more like a normal distribution. In fact, the most common amounts contributed are shown below.

```
##
##   25   50  100   10    5   15   27  250 2700   19
## 2976 2217 2080 1739 1299  910  863  785  417  364
```

Now, let's do a log10 transformation to reshaped the data distribution to look like a normal curve.



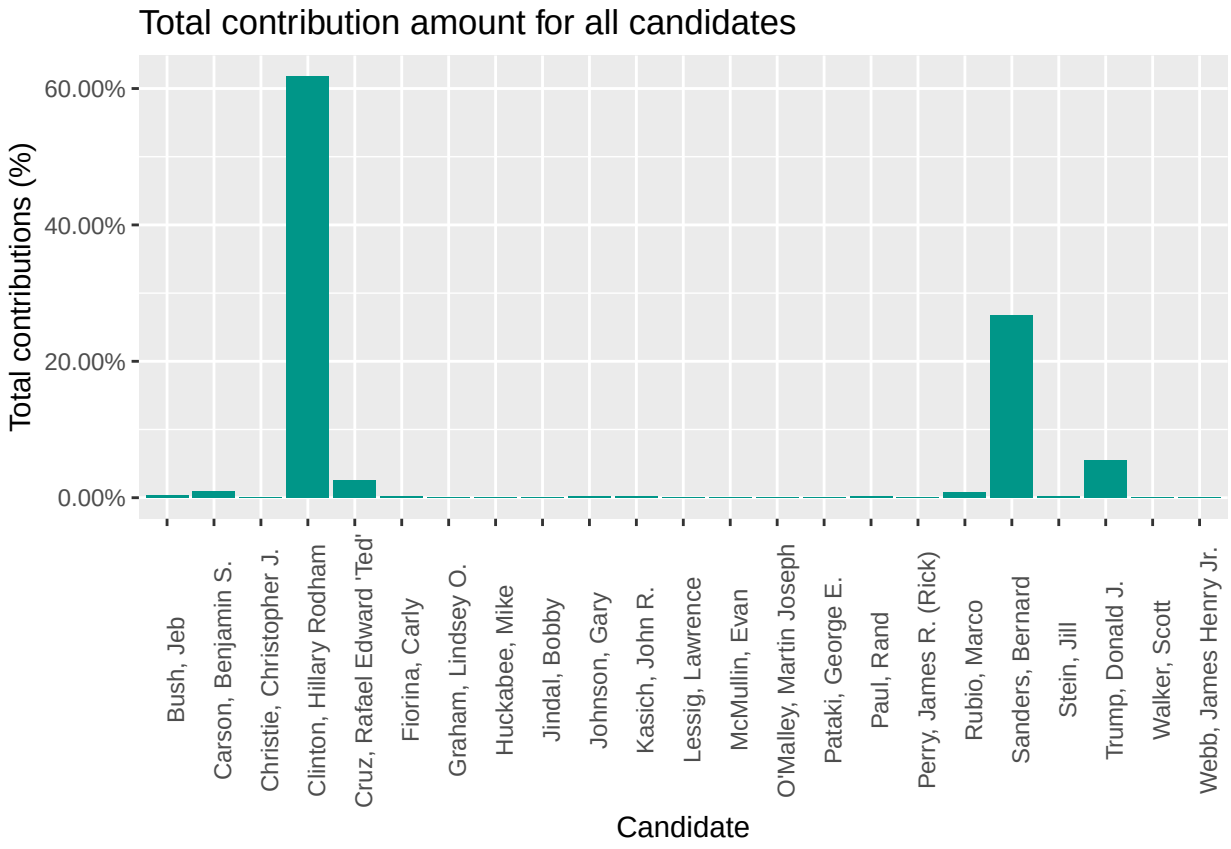
Note that the data for `contributor_receipt_amount` is much more close to a normal distribution than before. Let's continue our analysis by exploring contributions to individual candidates.

First, let's take a look at the `table` command to check the candidates which received the most individual contributions.

```
##
##                 Bush, Jeb      Carson, Benjamin S.
##                         81                      196
##  Christie, Christopher J.   Clinton, Hillary Rodham
##                          9                    12350
## Cruz, Rafael Edward 'Ted'           Fiorina, Carly
##                        506                       31
##        Graham, Lindsey O.           Huckabee, Mike
##                         11                        7
##             Jindal, Bobby            Johnson, Gary
##                          2                       31
##           Kasich, John R.         Lessig, Lawrence
##                         49                        4
##            McMullin, Evan   O'Malley, Martin Joseph
##                          4                       11
##         Pataki, George E.               Paul, Rand
##                          2                       36
##    Perry, James R. (Rick)             Rubio, Marco
##                          1                      154
```

```
##          Sanders, Bernard                Stein, Jill
##                      5355                         35
##          Trump, Donald J.             Walker, Scott
##                      1112                         10
##     Webb, James Henry Jr.
##                         3
```

Plotting the percentage of contributions to each presidential candidate.



The plot above shows us that Hillary Clinton, Bernard Sanders , and Donald Trump received the most contributions.

Taking a similar approach, let's see the distribution of the occupation of each individual contributor. First, let's take a look at the table command to get a better sense of this variable.

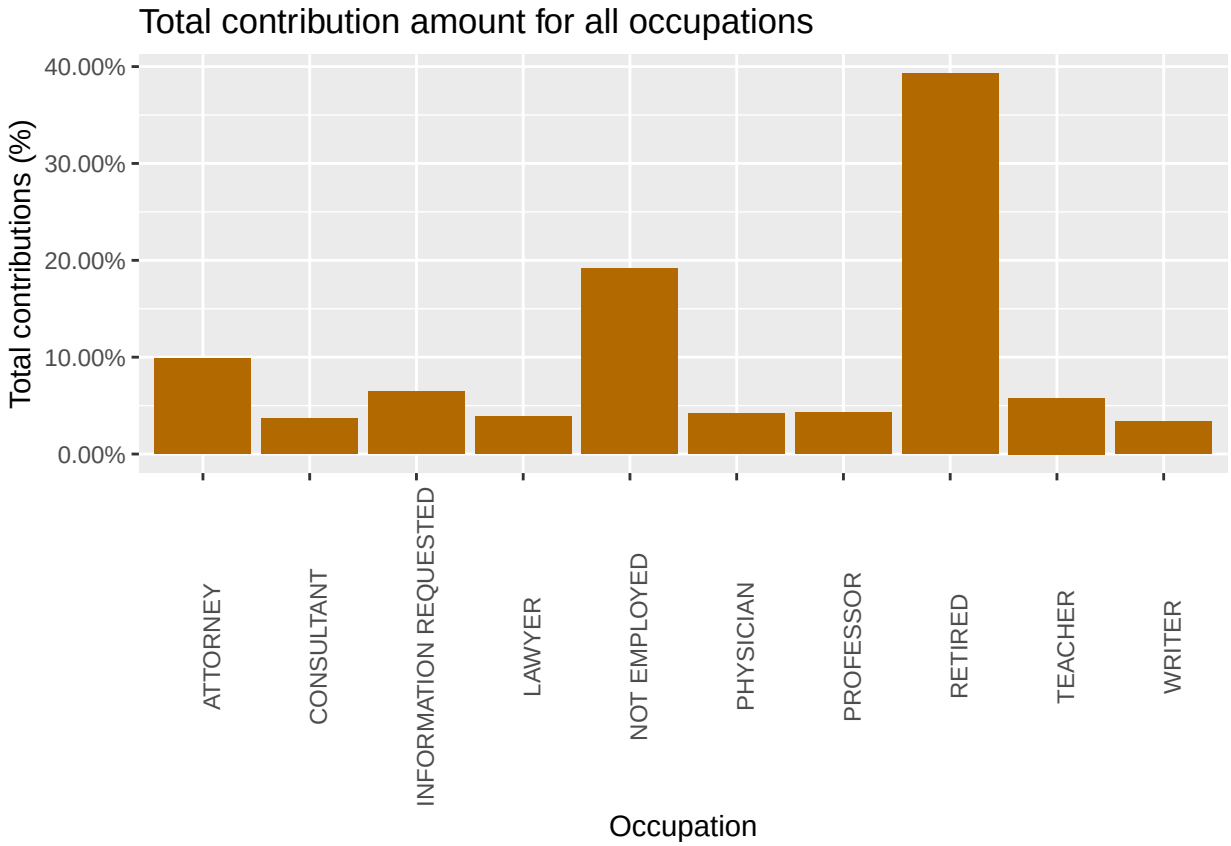
```
## [1] 3656
```

Considering we have over 3,500 unique occupations for financial contributors, let's just have a pick at the most commom positions. In order to do this, lets order the `table` function with `decreasing` parameter as `TRUE` and passing the value 10 to the `head` function to get the top 10 values.

```
##
##               RETIRED          NOT EMPLOYED              ATTORNEY
```

```
##                    3011                    1467                     756
## INFORMATION REQUESTED                 TEACHER               PROFESSOR
##                     497                     443                     332
##               PHYSICIAN                  LAWYER              CONSULTANT
##                     317                     300                     284
##                  WRITER
##                     254
```

Filtering the dataframe to display only the data for the top 10 occupations. This is done in order to make a readable plot. Otherwise, it would be impossible to visualize the data.



This plot indicates that RETIRED peolple are the ones with the most contributions to the presidential

Univariate Analysis

We are done exploring one variable. Not, let's provide a quick summary of what's been observed through this section of the analysis, which is just the starting point of a more elaborated exploration coming on the next sections.

This is an example of a tidy dataset, where each observation is a row and each variable forms a column in the collection of data. The first step we took at the data wrangling and manipulation was to trasnform and clean the original dataset. We converted some features to `factor`, `integer`, and `datetime`. These transformations were performed in order to create a more concise dataset

and prepare it to future data exploration phases. Other than that, we managed to create 5 new features from the original dataset to add more value to our observations. We are going to explain in more details below the transformations executed for some of the most important features we're going to use.

The data distribution for the `amount` variable is highly right skewed. Hence, a `log10` transformation has been applied to transform it to a normal-like distribution.

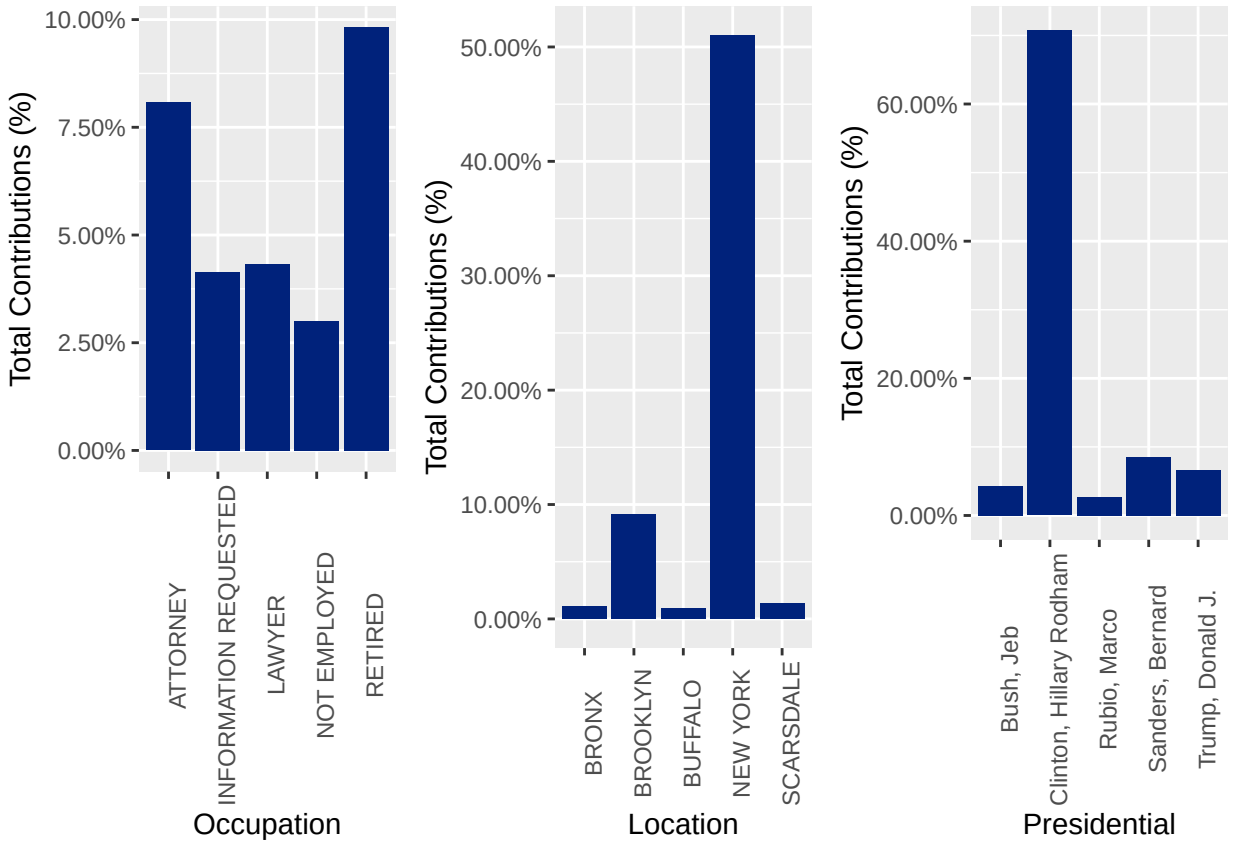
There are several features of interest:

- `candidate_name`: the name of the candidate as a factor variable will be really useful in this analysis in order to group the financial contributions by candidate and see how each candidate compare with each other in terms of amount received.
- `contributor_city`: the location of the person who made the contribution. We are going to use this feature to segment the amount contributed by each candidate by location, to undertstand which presidential performed better in which location.
- `occupation`: to map the most common occupation from the contributors.
- `contributor_receipt_amount`: the amount contributed, perhaps the most important feature to be analyzed, once will provide us the opportunity to aggregate the data by a single number ($), and elaborate our assumptions.
- `contributor_receipt_date`: this feature is very important because it allowed us to create new variables to understand the data movement through time. With this information, we can plot timeseries analysis and understand how the values change over time. It's important to note that this feature didn't came in a usable format. For this reason, we had to apply some transformations on the data manipulation section to make it useful.
- `lat` and `lon`: These two are two features extracted from the original location the dataset provided. With this information, we're now able to perform map operations, like plotting the aggregated values on a map.

At the end of our transformations, we saved the newly formatted dataframe into a new csv dataset. We did this to preserve our operations and avoing doing all the steps once again. Other than that, we created the variables `lat` and `lon` calling the `geocode()` function from the `ggmap` library. This call itself takes a long time to complete (because each location is one request to the Google Service Location API). Therefore, we don't want to perform this step again everytime we run this report.

Bivariate Plots Section

In this section, we're going to summarize financial contributions by candidate, location and occupation. The first step is to group the values by these variables forementioned. To create effective and attractive plots, we're going to limit the aggregated data to the top 5 results.



Looking at the graph above, we identified that the profile of the people who contributed the most is a Retired person who lives in New York and contributed to Hillary Clinton's campaign.

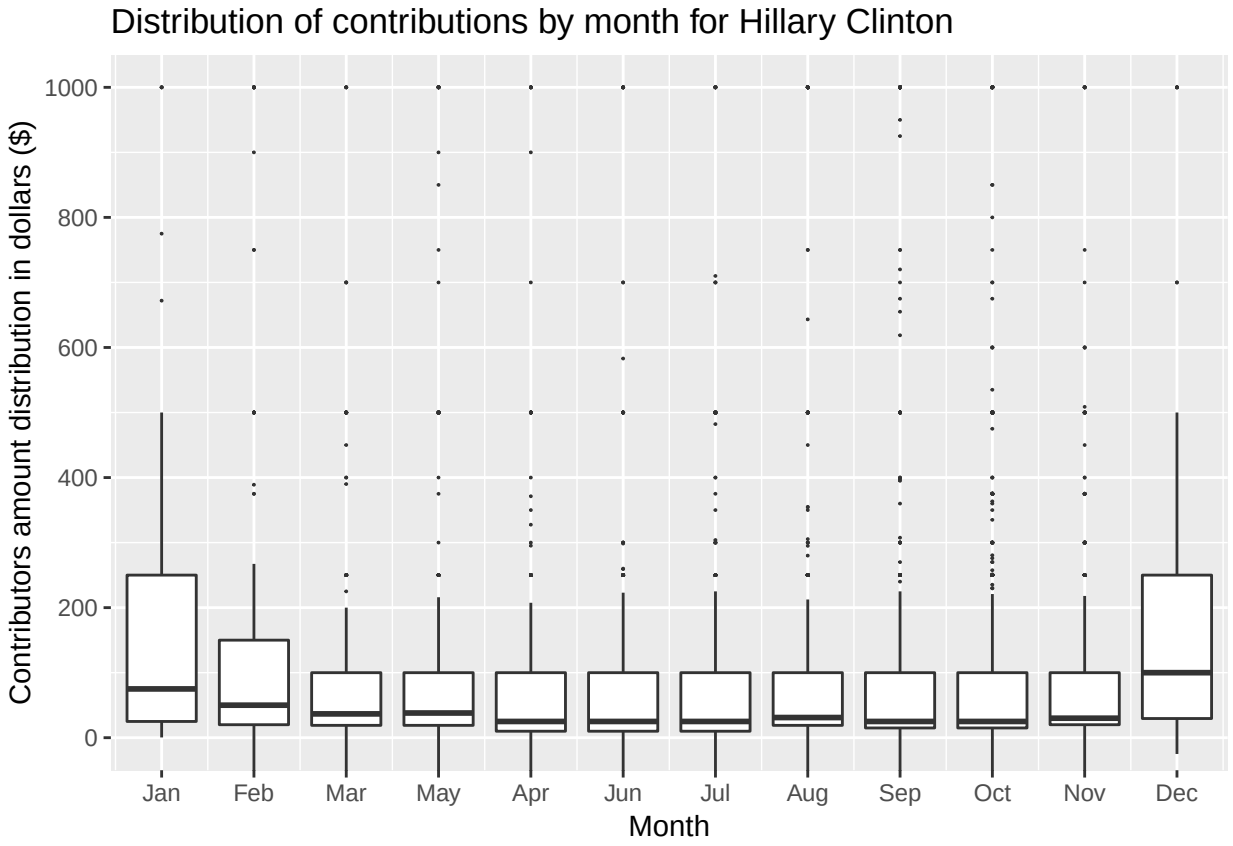
Boxplots

Let's group the data by candidate and filter out the top 3 presidentials with the most contributions amount in order to make easier to read the visualizations. In this next section, we're going to prepare the data to be plotted on the next section using boxplots.

Now that we grouped the data by candidate and filtered the top 3, we're going to create boxplots to understand the distributions for each of the presidentials. Boxplots help us visualize how the data is distributed, the 25%, 50% (median), 75%, Intequartile Range of the data and extreme values (outliers).

Let's visualize the boxplots by month for Hillary Clinton.

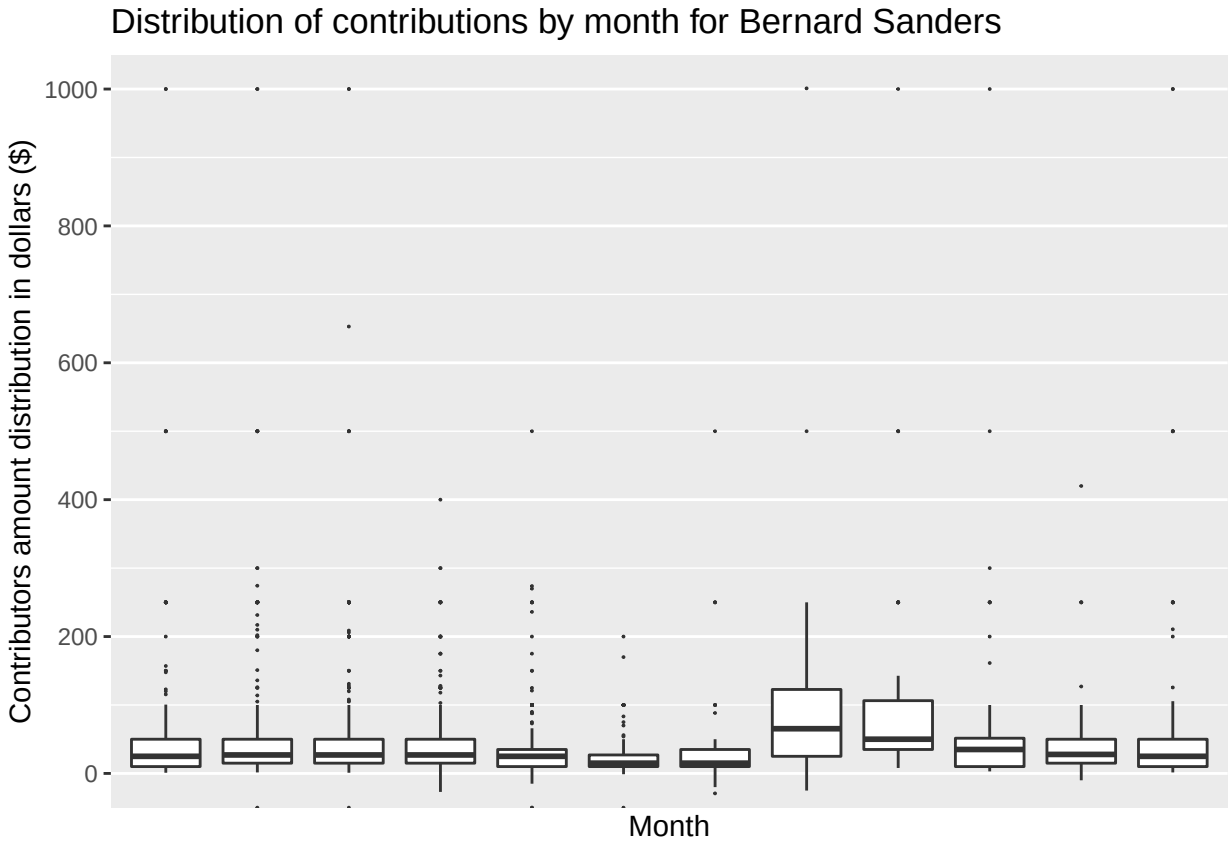
```
## [1] -2700  2700
```



It's interesting to see how there are a lot of outliers for some months for Hillary (which could indicate ups and downs in her campaign), like September and October. The median value is greater for January and December.

Now let's visualize the boxplots by month for Bernard Sanders.

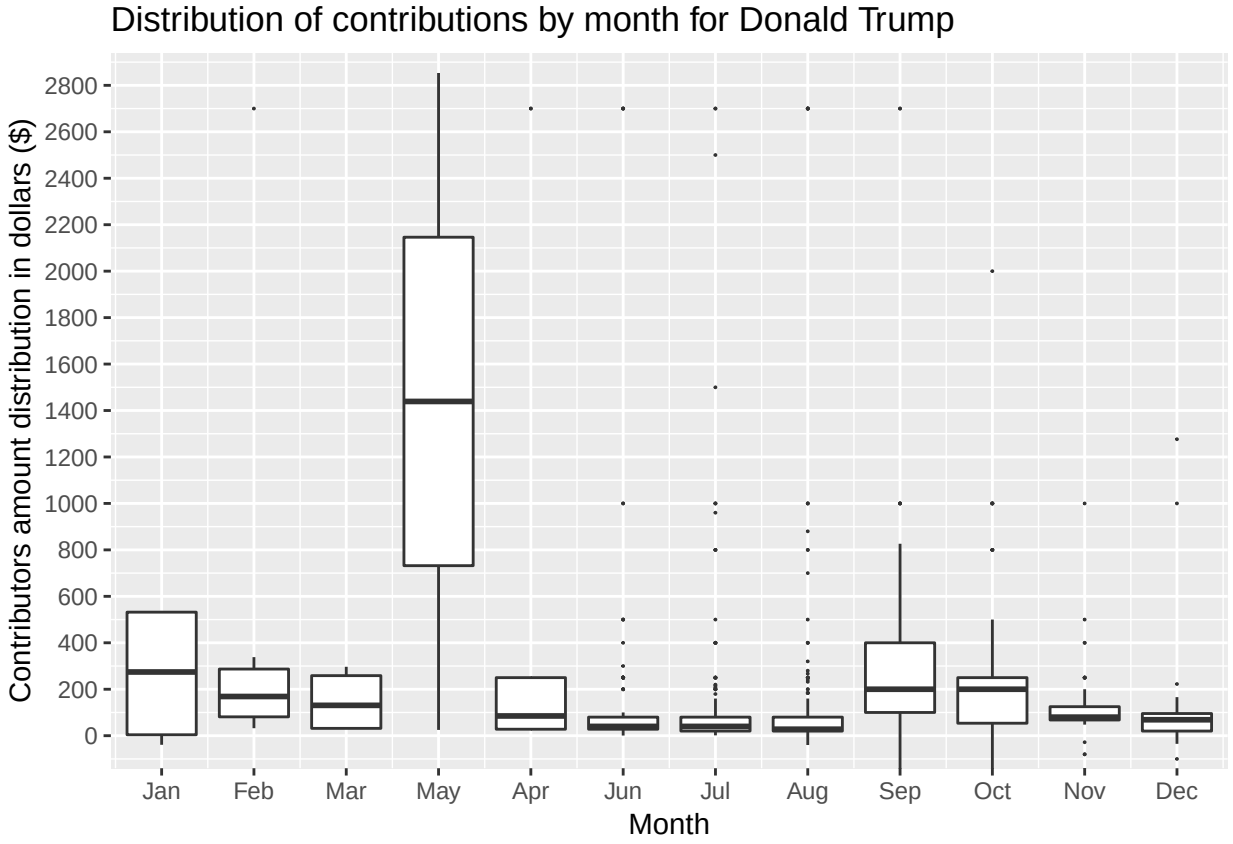
```
## [1] -2000  2700
```



Sanders doesn't have as many outliers as Hillary, and the range of the financial contributions is also low compared to the other 2. Other than that, he received most of his highest contributions on August and September.

Finally, Let's visualize the boxplots by month for Donald Trump.

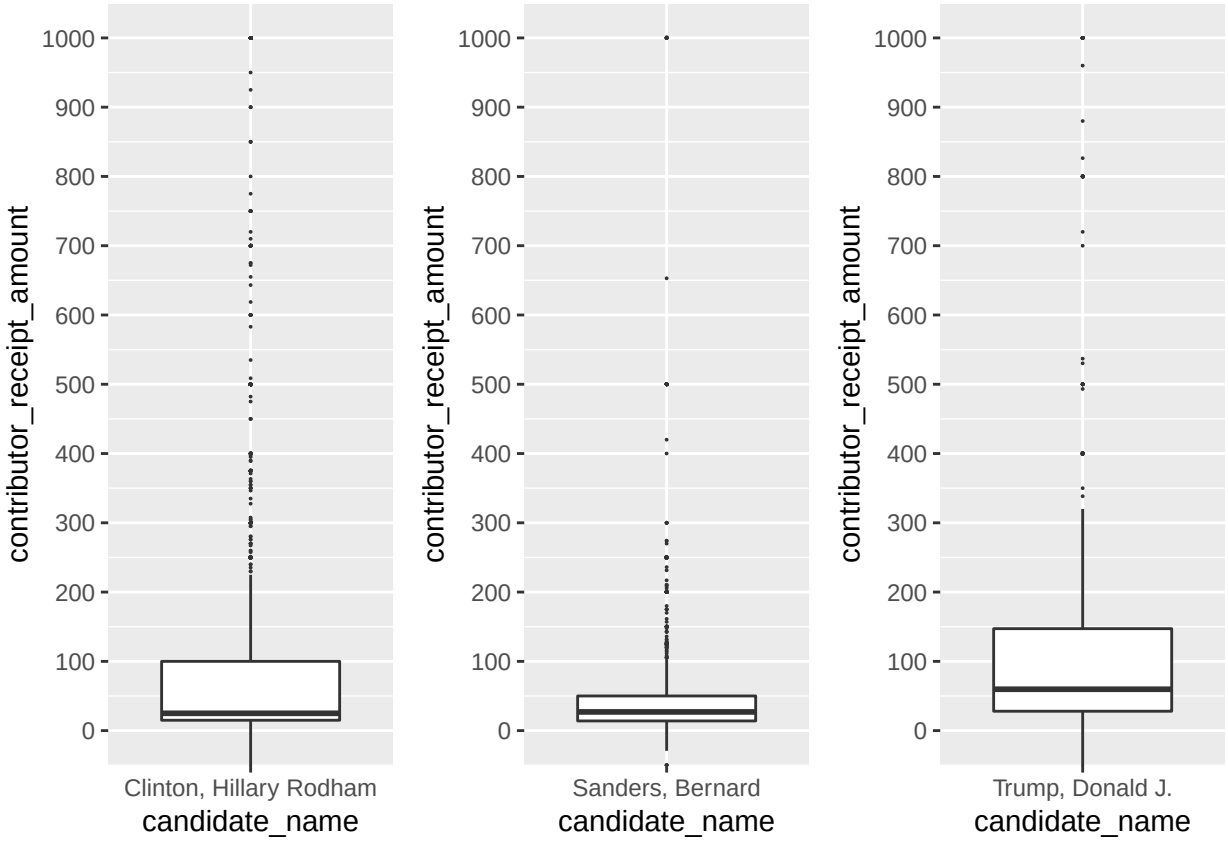
```
## [1] -400.00 2853.18
```



There aren't many outliers for Trump as well. However, the range of his contribution is very high compared to others, and he received a huge amount of donations on May. Let's investigate a little what happened on this month to try to understand a bit more about this data. If we look at the news (https://tinyurl.com/olpm6ur, https://tinyurl.com/y9culvkw, https://tinyurl.com/y9z9gjjy), a couple of things happened:

- Some candidates withdraw their presidency campaigns such as John Kasich and Ted Cruz. Speech by Cruz in his concession: "From the beginning I've said that I would continue on as long as there was a viable path to victory. Tonight I'm sorry to say it appears that path has been foreclosed."
- Nationally televised presidential debates
- Trump crosses delegate threshold

Now that we analyzed the data distribution by month, let's explore each candidate overall distribution to understand a little more of their contributions received.



Investigating each candidate's distribution, we can confirm what has been observed before. Hillary's distribution has several outlier points. Sanders' has a very low range of values, and Trump's has the greatest median value. Even though Hillary received the highest overall amount, it seems that values contributued to Donald Trump are greater. From this data we could draw some assumptions:

- This might be an indication of engagement from his supporters
- Perhaps better financial condition overall of his audience

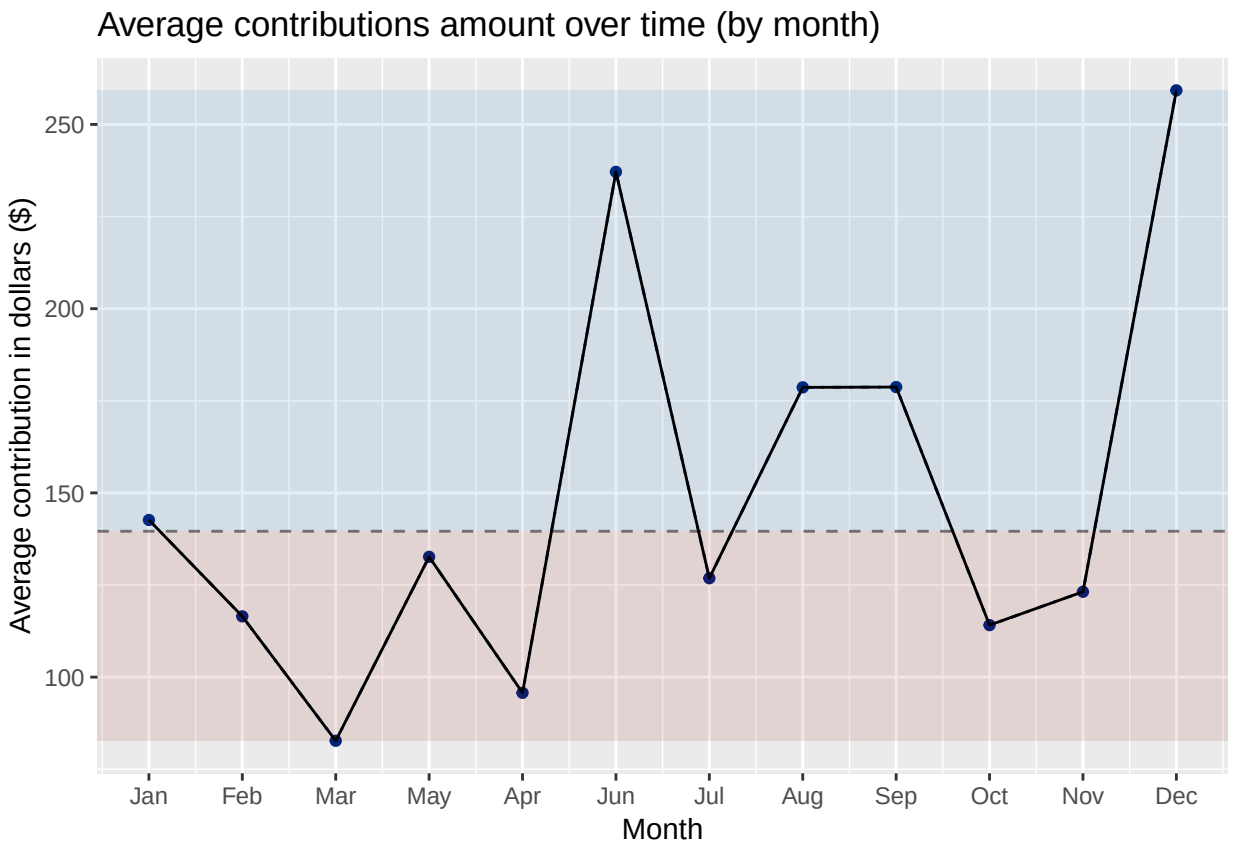
To answer these questions, we would need additional data about the contributors (which we lack on this dataset). Additionally, this analysis is out of the scope of this report.

Time-series Analysis

Now that we aggregated by candidate, occupation and location, let's analyze how the donations change over time. This time series analysis is possible because we pre-processed the dataset to extract the date information from the provided raw, unformatted date string.

```
## # A tibble: 6 x 4
##   receipt_month mean_donation total_donation     n
##           <int>         <dbl>          <dbl> <int>
## 1             1         143.          90728.   636
## 2             2         116.         162625.  1396
## 3             3          82.7        162065.  1959
## 4             4         133.         265042.  1998
```

```
## 5             5          95.7        140063.  1463
## 6             6         237.         336016.  1417
```



In the section above, we did a timeseries analysis, exploring how the contributions have changed over time. Plotting the data as a timeseries is very useful for any analysis. Visualizing the data in this format allows us to identify extreme points, seasonalities and trends in our data. For instance, we can notice how the months March and June have very extreme values, which can indicate seasonal periods. A deeper understanding of these periods can unravel hidden patterns in our data.

Analysis of correlation

In this dataset, we have only one continuous feature of interest (in fact, the most important feature). Therefore, we had to perform a correlation between numerical and categorical features. The test used is the Fligner-Killeen test, a test for homogeneity of variances. For this test, we are going to compare the contribution amount with two other features: `contributor_occupation` and `contributor_city`. The question we're trying to answer here is that if the amount contributed is independent of contributor's location and occupation. The Null hypothesis is that there are independent. Let's check below.

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  log10(contributor_receipt_amount) by as.factor(contributor_occupation)
## Fligner-Killeen:med chi-squared = 5587.5, df = 3648, p-value <
## 2.2e-16
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  log10(contributor_receipt_amount) by as.factor(contributor_city)
## Fligner-Killeen:med chi-squared = 1979.4, df = 1144, p-value <
## 2.2e-16
```

As we can see, for a p-value of 0.05, we can reject the Null Hypothesis for both cases. We then can state that the financial contribution is not independent of the contributor's location and occupation.

Bivariate Analysis

This section let us observe interesting patterns and answer a couple of questions. We were able to draw a picture of the most common contributior profile by anylising the amounts donated by all candidates. This analysis of the contributor (donator) profile was possible by crossing data about the main features we selected before: candidate, location, and occupation.

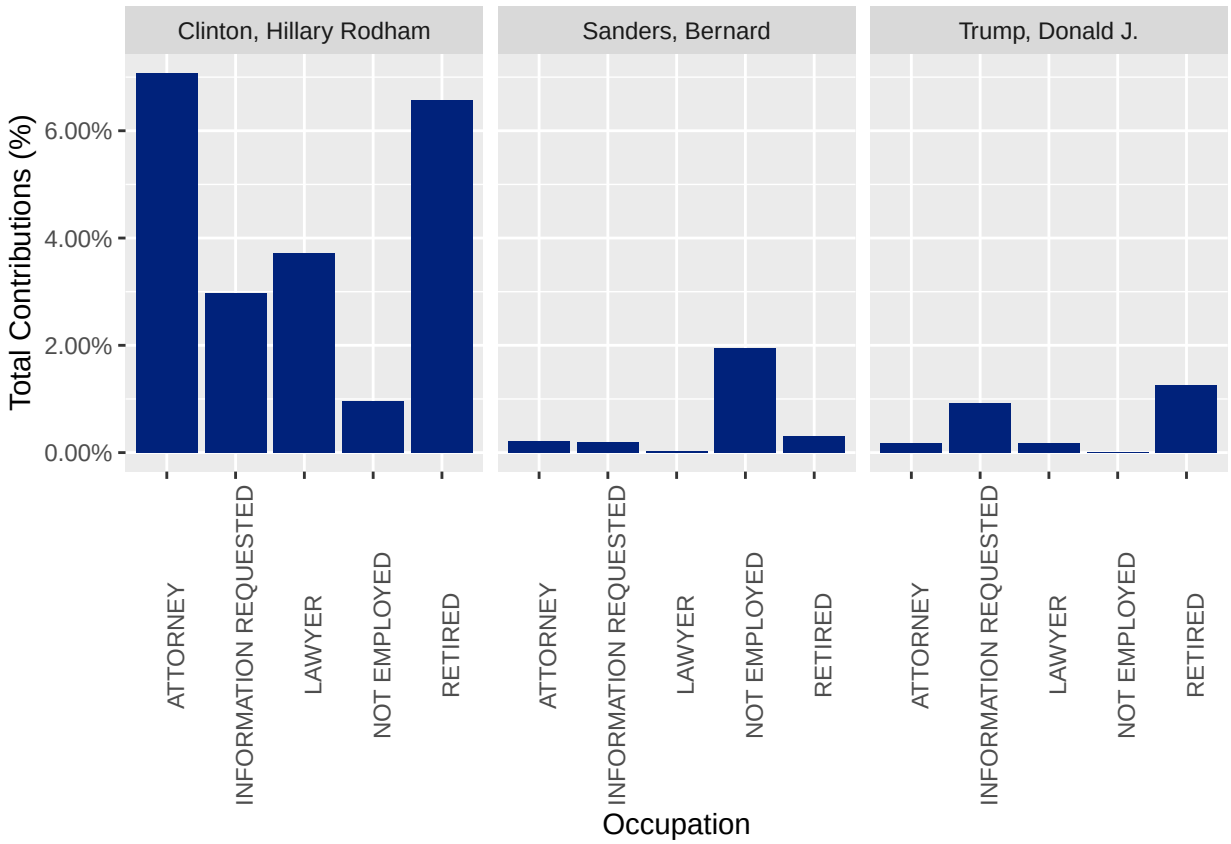
Additionally, the features we created from the raw unformatted date provided were valuable in our exploration. We include date and time in our data, the analysis become much more robust because we have then the opportunity to observe the trends and seasonality of the data, such information impossible to derive without time perspective.

Finally, location and occupation (as we supposed) had the strongest relationship with the amount contributed. As we can assume, the amount contributed to a candidate will greatly depend on where the people live and what they do for living.

In the next section, we're going to add one more dimension to the analysis and I believe we will get an even better sense of the data of our analysis.

Multivariate Plots Section

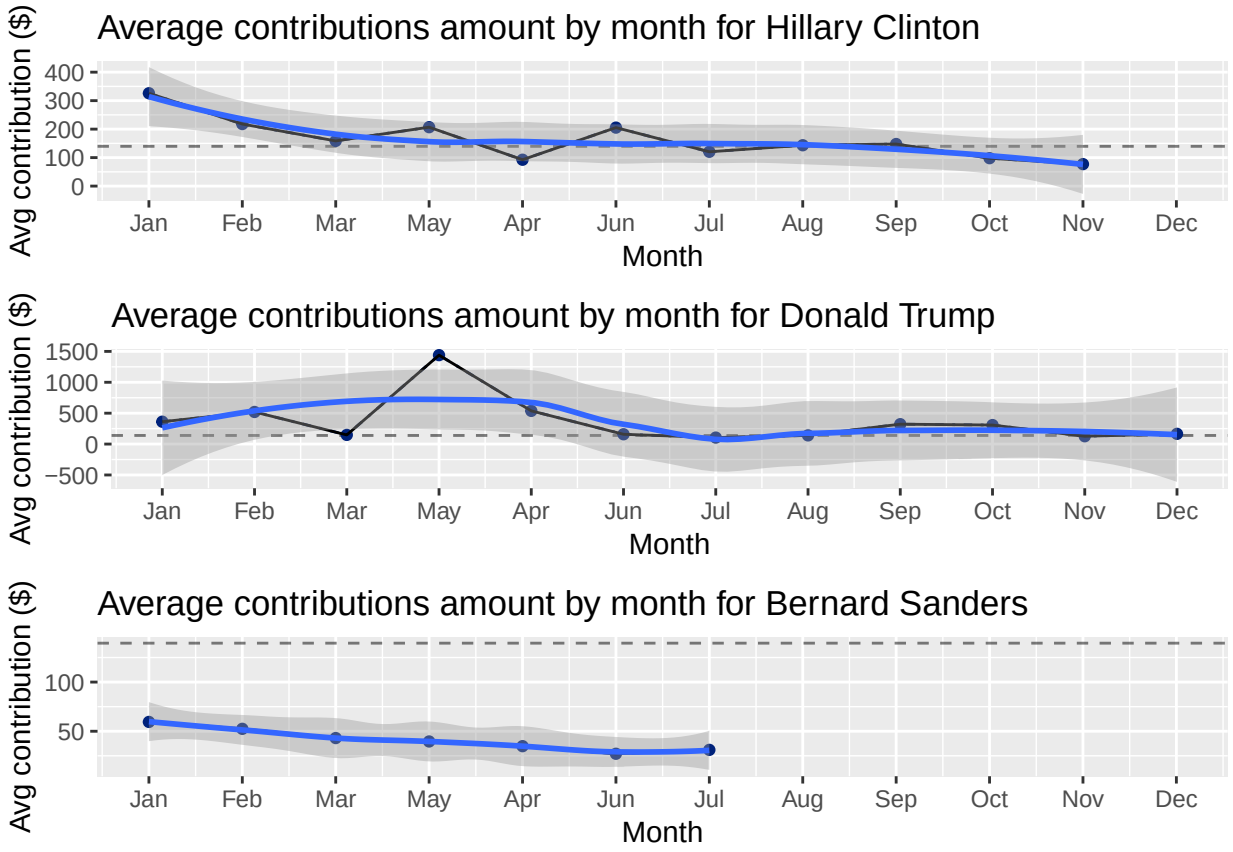
To gain a deeper understanding of the contributor profile (as we did on a previous section), let's run a similar analaysis but now faceting the graph by each candidate. This way, we can have a better understanding of each candidate supporter's profile.



As we can see on the chart above, the people who contributed more for her campaign are Attorney. For Sanders, Not Employed (interesting). Finally, for Trump, Retired people.

Breaking Time-series Analysis by candidate

Now the we have our date information, we can detail a bit more our time-series analysis by adding another dimension: `candidate_name`.



The graph above show the distribution of the contributions over time by each candidate. Also, we plotted a linear model to identify the trend of each candidate's distribution. Hillary had a high increase on June (if we take a look at the news, this importnat event happened on this month: "9 June – Obama endorses Clinton"), but after September, her contributions start to decline, following a negative trend. For Trump, on the other hand, the trend is rather stable. Sanders received contributions until July, when he suspended his campaign and endorsed Hillary Clinton (https://tinyurl.com/olpm6ur). Finally, we can spot some seasonal points on the plot for Donald Trump in March (as observed before).

Multivariate Analysis

Adding another dimension to the plots allowed us to detail the interaction between the variables. By adding the occupation and candidate together, we were able to draw the contributor profile for each candidate.

Additionally, by adding the candidate feature to the time-series analysis helped us understand how the contributions vary over time. Also, researching the news, we were able to find some important events that happened and which could help us justify some changes on the graph.

Finally, for the time-series analysis, we created a linear model to spot the trend of the data for each candidate. Furthermore, we were surprised how much information just a simple date feature can add to our analysis. This information helps us understand what happened in the

past, how each candidate is doing in his campaign, the trend of the data and some extreme points and seasonality.

Final words and Summary

This was a challenging project. I had some trouble manipulating the original dataset to transform it to a cleaner and more informative data structure. For this reason, I wanted to add extra features that would empower our analysis. Features about location and date were added and used extensively. Particularly, the date information was really useful to perform time-series analysis and understand the distribution of the data over time.

Throughout the analysis, we were able to gain some insights about the financial contribution for the 2016 U.S. presidential election. We were able to find the top candidates, location, and occupation which contributed the most to the funding of presidential campaigns. Also, we found very interesting that some important events that happened during the campaign of each candidate had some influence on the amount contributed. Finally, we were able to spot the trends, seasonality and outliers on the data and to draw the profile of people who donated for each presidency candidate.

Our work could be improved by adding more features to our dataframe (for example, more information about the contributors (like address, education, etc)) and performing a more elaborated work on time-series analysis, like forecasting, ARIMA, etc. We had some attempts to work with the Facebook library `prophet`, but issues on our environment limited our resources.