2016 USA Presidential Campaign Finance in the State of California

Introduction:

This project uses exploratory data techniques to understand the behavior of American contributors to the 2016 Presidential Campaign in the State of California. The purpose of the project is to use different tools in R to acquire, clean, explore and analyze a dataset, and to present three data visualizations that can display insights about the information.

The structure of this document is the following:

1. Objectives
2. Description of the dataset
3. Data wrangling
4. Data exploration
5. Final Plots and Summary
6. Reflections

Objectives:

- To find out how much money the contributors are willing to give to a Presidential Campaign.
- To understand the differences in behavior of contributors by gender and ethnicity.
- To understand how presidential candidates were funded by contributors.

Description of the dataset:

The dataset contains a universe of all individual contributions reported on Form 3P Line 17A, refunds to individuals reported on Form 3P Line 28A and transfers from authorized committees reported on Form 3P Line 18 for the 2016 Presidential Campaign in the State of California.

The data can be downloaded from the following website . Furthermore, the dataset for the State of California contains 19 fields and 1,304,346 entries.

The fields comprised in the dataset are:

- COMMITTEE ID: A 9-character alpha-numeric code assigned to a committee by the Federal Election Commission.
- CANDIDATE ID: A 9-character alpha-numeric code assigned to a candidate by the Federal Election Commission.
- CANDIDATE NAME: Reported candidate name.

- CONTRIBUTOR NAME: Reported name of the contributor.
- CONTRIBUTOR CITY: Reported city of the contributor.
- CONTRIBUTOR STATE: Reported state of the contributor.
- CONTRIBUTOR ZIP CODE: Reported zip code of the contributor.
- CONTRIBUTOR EMPLOYER: Reported employer of the contributor.
- CONTRIBUTOR OCCUPATION: Reported occupation of the contributor.
- CONTRIBUTION RECEIPT AMOUNT: Reported contribution amount.
- CONTRIBUTION RECEIPT DATE: Reported contribution receipt date. The date format is DD-MMM-YYYY.
- RECEIPT DESCRIPTION: Additional information reported by the committee about a specific contribution.
- MEMO CODE: 'X' indicates the reporting committee has provided additional text to describe a specific contribution. See the MEMO TEXT.
- MEMO TEXT: Additional information reported by the committee about a specific contribution.
- FORM TYPE: Indicates what schedule and line number the reporting committee reported a specific transaction.
- FILE NUMBER: A unique number assigned to a report and all its associated transactions.
- TRANSACTION ID: A unique identifier permanently associated with each itemization or transaction appearing in an FEC electronic file.
- ELECTION TYPE / PRIMARY-GENERAL INDICATOR: This code indicates the election for which the contribution was made. EYYYY (election plus election year)

```
## [1] 1304346      19
```

The dataset contains 19 columns and 1,304,346 rows.

Data wrangling:

This section presents a description of how the dataset is cleaned and new variables are added to fulfill the objectives of the project.

Data cleaning:

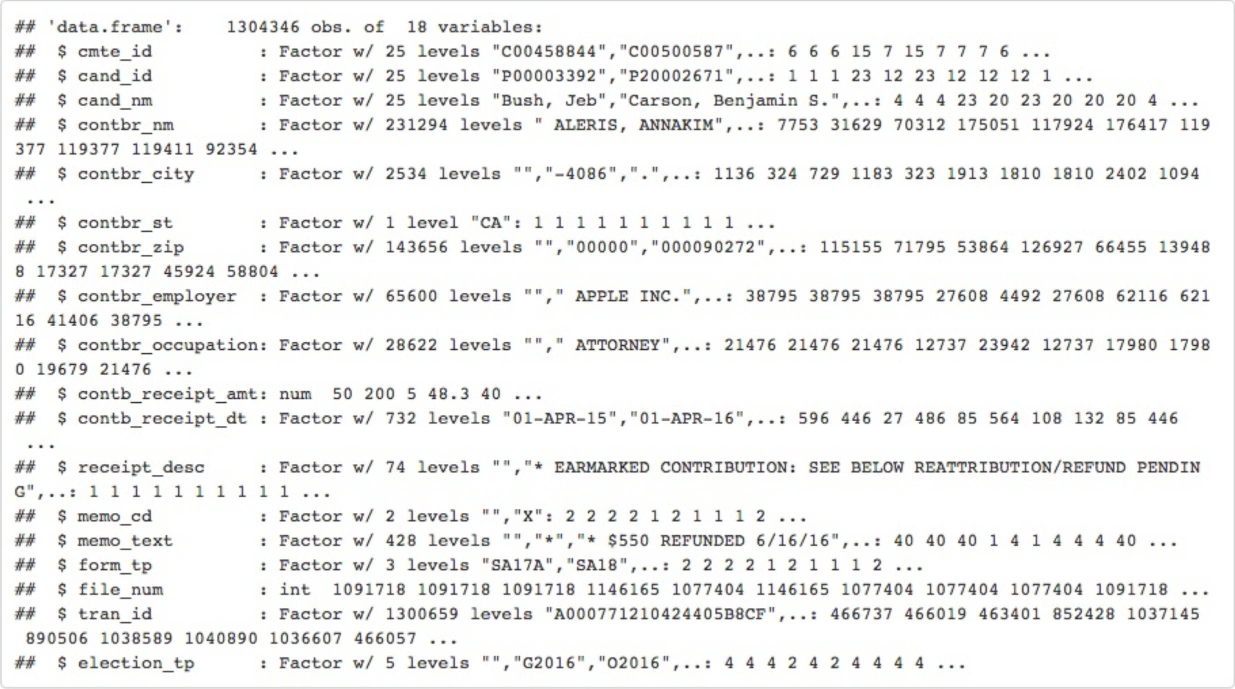
After importing the data that came in a csv format, the first step to clean the data is to rename the columns as needed. The second step is to remove the missing values. The third step is to change the type of the fields as needed. The fourth step is to subset the dataset to the contributions of interest.

a) Renaming columns and dropping empty columns

The first step in the process is to rename the columns of the dataset. Later, the last empty columns will be dropped from the dataset.

Exploring data types:

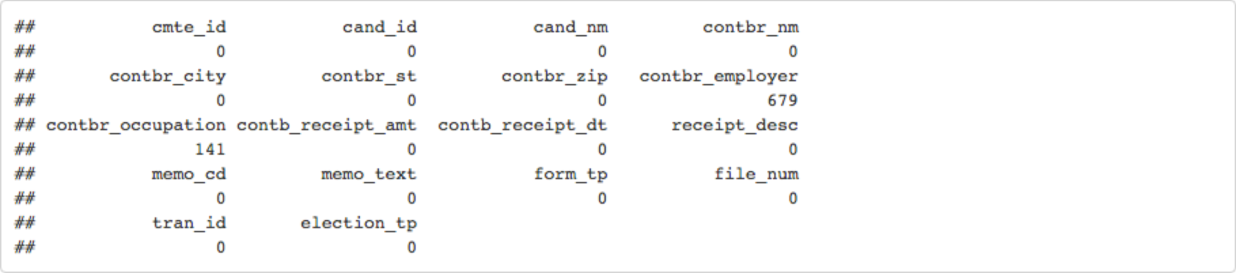
A quick display to the structure of the data shows the variables name and type. With the exception of contb_receipt_amt, R interpreted the fields as factors. This fact should be kept in mind as further type transformations will have to be required.



```
## 'data.frame':    1304346 obs. of  18 variables:
##  $ cmte_id          : Factor w/ 25 levels "C00458844","C00500587",..: 6 6 6 15 7 15 7 7 7 6 ...
##  $ cand_id          : Factor w/ 25 levels "P00003392","P20002671",..: 1 1 1 23 12 23 12 12 12 1 ...
##  $ cand_nm          : Factor w/ 25 levels "Bush, Jeb","Carson, Benjamin S.",..: 4 4 4 23 20 23 20 20 20 4 ...
##  $ contbr_nm        : Factor w/ 231294 levels " ALERIS, ANNAKIM",..: 7753 31629 70312 175051 117924 176417 119
377 119377 119411 92354 ...
##  $ contbr_city      : Factor w/ 2534 levels "","-4086",".",..: 1136 324 729 1183 323 1913 1810 1810 2402 1094
 ...
##  $ contbr_st        : Factor w/ 1 level "CA": 1 1 1 1 1 1 1 1 1 1 ...
##  $ contbr_zip       : Factor w/ 143656 levels "","00000","000090272",..: 115155 71795 53864 126927 66455 13948
8 17327 17327 45924 58804 ...
##  $ contbr_employer  : Factor w/ 65600 levels ""," APPLE INC.",..: 38795 38795 38795 27608 4492 27608 62116 621
16 41406 38795 ...
##  $ contbr_occupation: Factor w/ 28622 levels ""," ATTORNEY",..: 21476 21476 21476 12737 23942 12737 17980 1798
0 19679 21476 ...
##  $ contb_receipt_amt: num  50 200 5 48.3 40 ...
##  $ contb_receipt_dt : Factor w/ 732 levels "01-APR-15","01-APR-16",..: 596 446 27 486 85 564 108 132 85 446
 ...
##  $ receipt_desc     : Factor w/ 74 levels "","* EARMARKED CONTRIBUTION: SEE BELOW REATTRIBUTION/REFUND PENDIN
G",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ memo_cd          : Factor w/ 2 levels "","X": 2 2 2 2 1 2 1 1 1 2 ...
##  $ memo_text        : Factor w/ 428 levels "","*","* $550 REFUNDED 6/16/16",..: 40 40 40 1 4 1 4 4 4 40 ...
##  $ form_tp          : Factor w/ 3 levels "SA17A","SA18",..: 2 2 2 2 1 2 1 1 1 2 ...
##  $ file_num         : int  1091718 1091718 1091718 1146165 1077404 1146165 1077404 1077404 1077404 1091718 ...
##  $ tran_id          : Factor w/ 1300659 levels "A000771210424405B8CF",..: 466737 466019 463401 852428 1037145
 890506 1038589 1040890 1036607 466057 ...
##  $ election_tp      : Factor w/ 5 levels "","G2016","O2016",..: 4 4 4 2 4 2 4 4 4 4 ...
```

b) Searching for missing values (NAs)

An analysis to find out missing values (NAs) in each columns was run indicating as a result that no NAs were found in the dataset, therefore the dataset is ready to be used for exploratory analysis.



```
##           cmte_id           cand_id           cand_nm         contbr_nm
##                 0                 0                 0                 0
##       contbr_city         contbr_st        contbr_zip   contbr_employer
##                 0                 0                 0               679
## contbr_occupation contb_receipt_amt  contb_receipt_dt      receipt_desc
##               141                 0                 0                 0
##           memo_cd         memo_text           form_tp          file_num
##                 0                 0                 0                 0
##           tran_id       election_tp
##                 0                 0
```

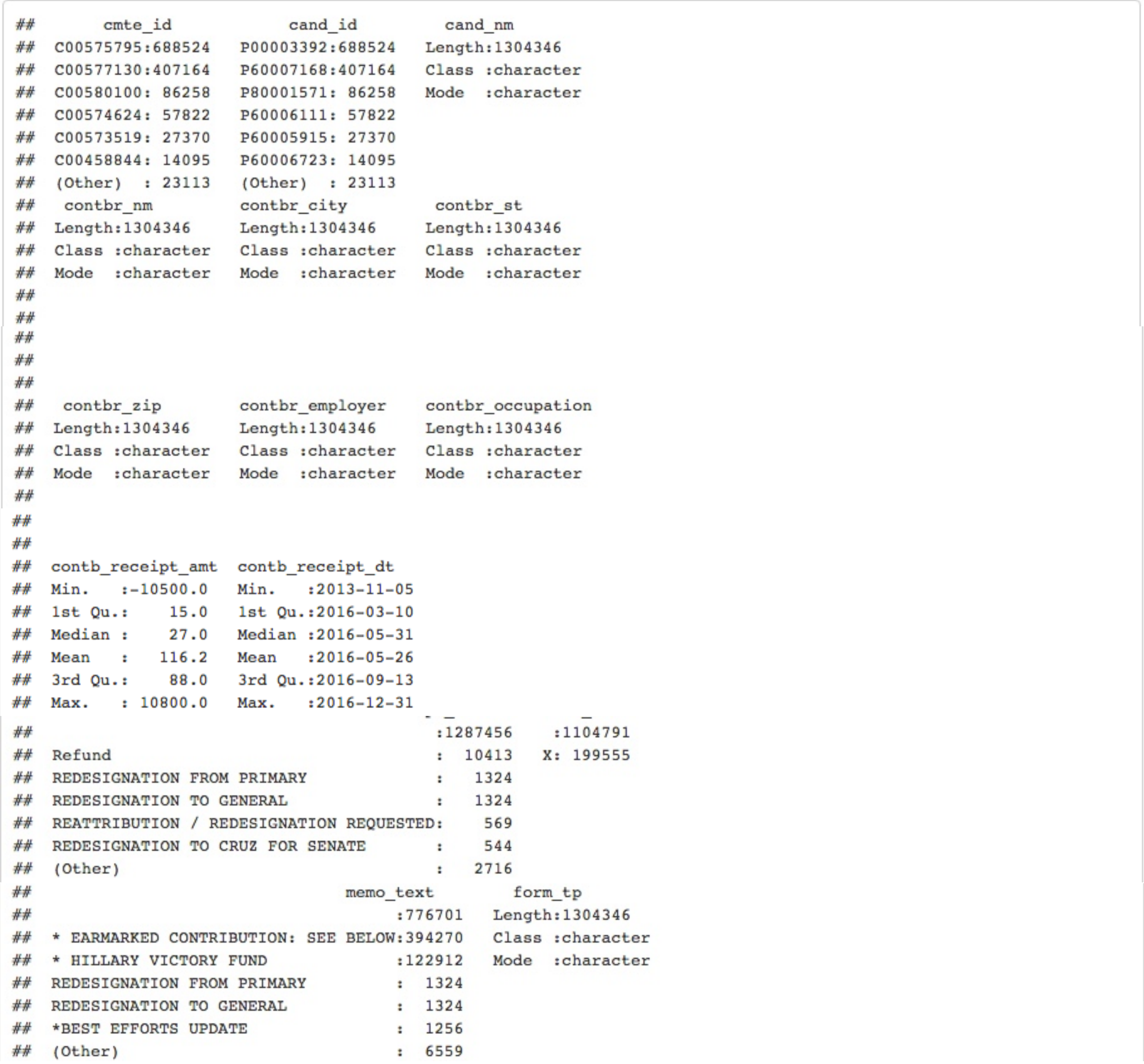
c) Change data types

As previously shown most of the fields were imported as factors, so the next step is to change the type of the fields as required:

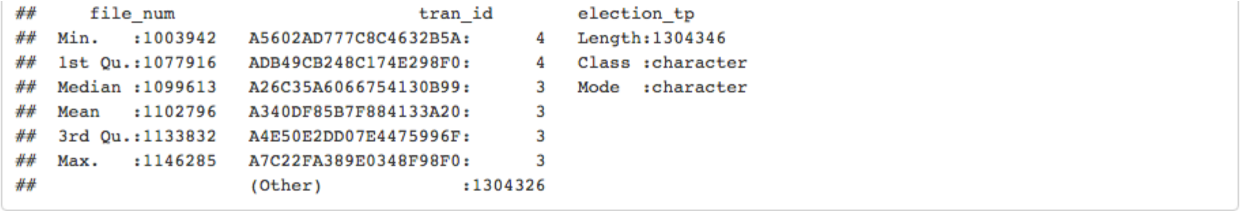
- CONTRIBUTION RECEIPT DATE as date type with the format “dd-mmm-yy”.
- CONTRIBUTION RECEIPT AMOUNT as numeric type.
- OTHER fields as character fields.

In general all types were change to character, except the dates that were formatted as this, and the amounts contributed as numeric type.

A statistical summary at the variables shows that the CONTRIBUTION RECEIPT AMOUNT (contb_receipt_am) takes negative values and very extreme positive values in comparison to the median.



```
##       cmte_id           cand_id           cand_nm
##  C00575795:688524   P00003392:688524   Length:1304346
##  C00577130:407164   P60007168:407164   Class :character
##  C00580100: 86258   P80001571: 86258   Mode  :character
##  C00574624: 57822   P60006111: 57822
##  C00573519: 27370   P60005915: 27370
##  C00458844: 14095   P60006723: 14095
##  (Other)  : 23113   (Other)  : 23113
##   contbr_nm          contbr_city         contbr_st
##  Length:1304346     Length:1304346     Length:1304346
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##
##   contbr_zip         contbr_employer    contbr_occupation
##  Length:1304346     Length:1304346     Length:1304346
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  contb_receipt_amt  contb_receipt_dt
##  Min.   :-10500.0   Min.   :2013-11-05
##  1st Qu.:    15.0   1st Qu.:2016-03-10
##  Median :    27.0   Median :2016-05-31
##  Mean   :   116.2   Mean   :2016-05-26
##  3rd Qu.:    88.0   3rd Qu.:2016-09-13
##  Max.   : 10800.0   Max.   :2016-12-31
##                                         :1287456    :1104791
##  Refund                                 :  10413   X: 199555
##  REDESIGNATION FROM PRIMARY             :   1324
##  REDESIGNATION TO GENERAL               :   1324
##  REATTRIBUTION / REDESIGNATION REQUESTED:    569
##  REDESIGNATION TO CRUZ FOR SENATE       :    544
##  (Other)                                :   2716
##                              memo_text        form_tp
##                                    :776701   Length:1304346
##  * EARMARKED CONTRIBUTION: SEE BELOW:394270   Class :character
##  * HILLARY VICTORY FUND             :122912   Mode  :character
##  REDESIGNATION FROM PRIMARY         :  1324
##  REDESIGNATION TO GENERAL           :  1324
##  *BEST EFFORTS UPDATE               :  1256
##  (Other)                            :  6559
```



```
##     file_num                      tran_id        election_tp
##  Min.    :1003942   A5602AD777C8C4632B5A:      4   Length:1304346
##  1st Qu.:1077916   ADB49CB248C174E298F0:      4   Class :character
##  Median :1099613   A26C35A6066754130B99:      3   Mode  :character
##  Mean    :1102796   A340DF85B7F884133A20:      3
##  3rd Qu.:1133832   A4E50E2DD07E4475996F:      3
##  Max.    :1146285   A7C22FA389E0348F98F0:      3
##                     (Other)             :1304326
```

d) Subsetting dataframe

For this project the variable of interest is the individual contributions for the Presidential Campaign that are capped to 2,700 dollars (See http://www.fec.gov/pages/brochures/citizens.shtml#how_much). In this sense, the "CONTRIBUTION RECEIPT AMOUNT" field is filtered to those values between 0 and 2,700. Also contributions for 2020 elections were removed and also those contributions that were not possible to associate to an election period.

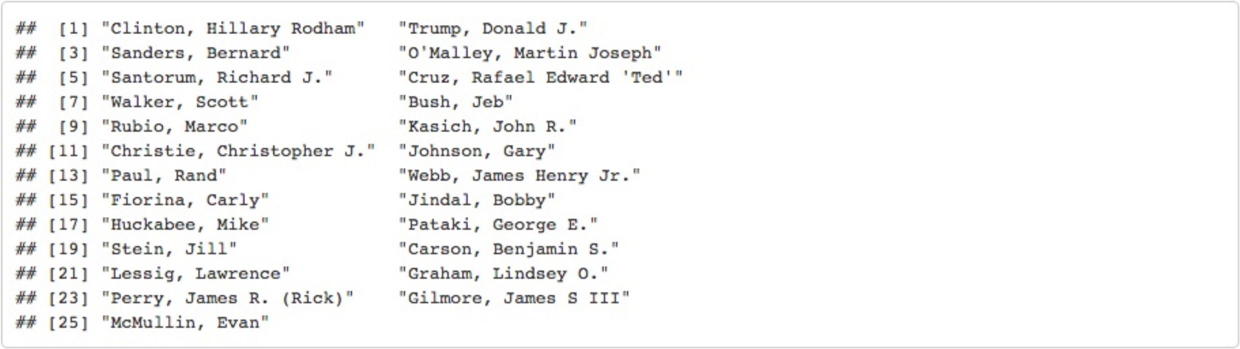
Adding new variables:

In order to achieve the objectives of this project, four additional variables have to be created such as:

1. political party
2. ethnicity
3. gender
4. geographical coordinates

1) Political party

The political party field can be created by classifying the candidate names into Republicans, Democrats, Green, Libertarian and Independent parties.



```
##  [1] "Clinton, Hillary Rodham"    "Trump, Donald J."
##  [3] "Sanders, Bernard"           "O'Malley, Martin Joseph"
##  [5] "Santorum, Richard J."       "Cruz, Rafael Edward 'Ted'"
##  [7] "Walker, Scott"              "Bush, Jeb"
##  [9] "Rubio, Marco"               "Kasich, John R."
## [11] "Christie, Christopher J."   "Johnson, Gary"
## [13] "Paul, Rand"                 "Webb, James Henry Jr."
## [15] "Fiorina, Carly"             "Jindal, Bobby"
## [17] "Huckabee, Mike"             "Pataki, George E."
## [19] "Stein, Jill"                "Carson, Benjamin S."
## [21] "Lessig, Lawrence"           "Graham, Lindsey O."
## [23] "Perry, James R. (Rick)"     "Gilmore, James S III"
## [25] "McMullin, Evan"
```

In this sense the dataset contains 25 unique candidates that were classified as follows.

Republican: * Bush, Jeb * Carson, Benjamin S. * Christie, Christopher J. * Cruz, Rafael Edward 'Ted' * Fiorina, Carly * Gilmore, James S III * Graham, Lindsey O. * Huckabee, Mike * Jindal, Bobby * Kasich, John R. * Pataki, George E. * Paul, Rand * Perry, James R. (Rick) * Rubio, Marco * Santorum, Richard J. * Trump, Donald J. * Walker, Scott

Democrats: * Clinton, Hillary Rodham * Lessig, Lawrence * O'Malley, Martin Joseph * Sanders, Bernard * Webb, James Henry Jr.

Others: * Stein, Jill (Green Party) * McMullin, Evan (Independent) * Johnson, Gary (Libertarian Party)

As expected most of the contributions in the state of California were for the Democrat (84%) and Republican (15%) candidates. The rest of the contributions to the Green and Libertarian parties and to the Independent candidate represent 1%.

```
## [1] "Contributors by political party"
```

```
##
##   Democrats  Republican       Green Libertarian Independent
##     1086377      194690        2813        1752         153
```

```
## [1] "Porportion of contributors by political party"
```

```
##
##    Democrats   Republican        Green  Libertarian  Independent
## 0.8449134187 0.1514172276 0.0021877686 0.0013625917 0.0001189935
```

2) Ethnicity

The ethnicity of each contributor was derived using a Bayesian Prediction of Racial Category using the contributor surname and geolocation (See https://github.com/kosukeimai/wru). The R library used for this purpose is "wru" (See https://cran.r-project.org/web/packages/wru/wru.pdf). The code retrieves information from the US CENSUS to predict the following ethnicities in the US: white, Black, Hispanic and Asian. It is worth to point out that the contributor name was split into name and surname to be able to compute the probabilities.

```
## [1] "Proceeding with surname-only predictions..."
```

The new ethnicity field classified most of the contributors in the dataset as white (81%). In a smaller proportion Hispanic (7%), Asian (4%) and Black (0.4%) contributors were classified. It is worth to point out that those contributors with probabilities less than 0.6 in any ethnicity were classified as "not determined" (6.7%).

```
## [1] "Contributors by ethnicity group"
```

```
##
##          asian          black       hispanic not_determined          white
##          53568           5412          95658          84948        1046199
```

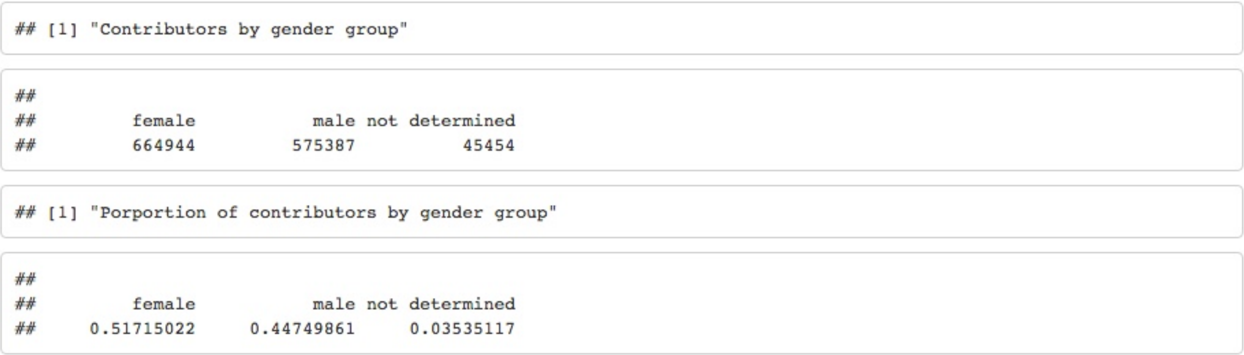
```
## [1] "Porportion of contributors by ethnicity group"
```

```
##
##          asian          black       hispanic not_determined          white
##    0.041661709    0.004209102    0.074396575    0.066067033    0.813665582
```

3) Gender

The gender variable was created by predicting gender from first name with information of the U.S. Social Security Administration. The library used for this purpose is “gender” (See https://github.com/ropensci/gender and https://cran.r-project.org/web/packages/gender/index.html)

The contributors gender was classified as mostly female (52%), male (45%) and “not determined” (4%).



```
## [1] "Contributors by gender group"
```

```
##
##         female          male not determined
##         664944        575387          45454
```

```
## [1] "Porportion of contributors by gender group"
```

```
##
##         female          male not determined
##     0.51715022    0.44749861     0.03535117
```

4) Geographical coordinates

To get the geographical coordinates (latitude and longitud), the zip code of the contributors was merged with the latitudes and coordinates provided by the United States Census Bureau (See https://www.census.gov/geo/maps-data/data/gazetteer.html and https://gist.github.com/erichurst/7882666).

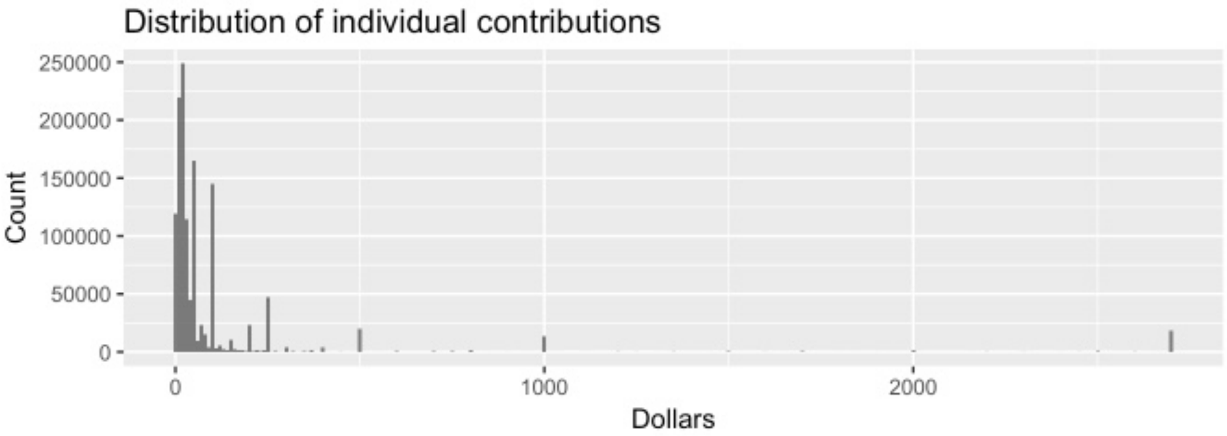
It is worth to point out that the zip code provided in the dataset was cleaned to five numeric characters.

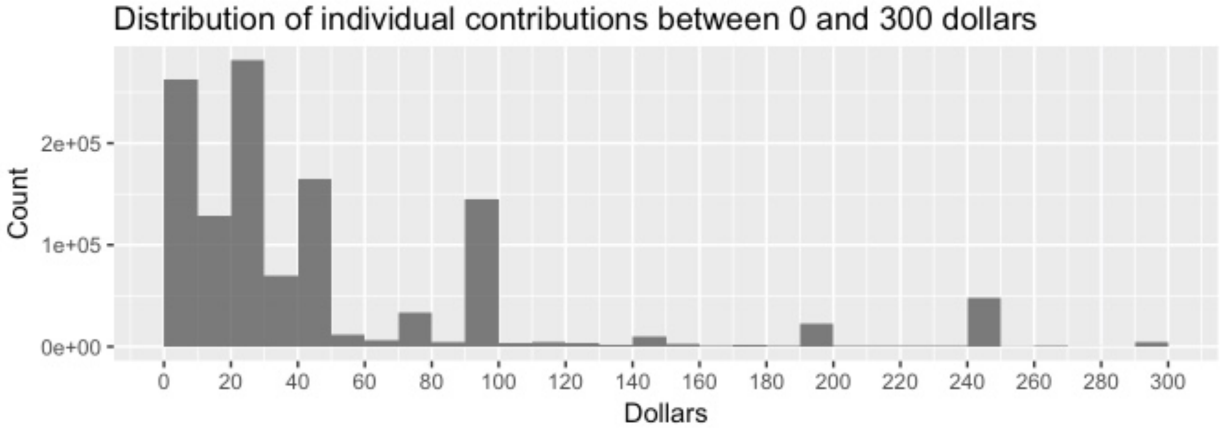
Once the political party, ethnicity, gender and geographical coordinates were created, the dataset is ready to be used for exploratory data analysis.

Exploratory data analysis:

This section explores the dataset variables using univariate and bivariate visualizations to identify patterns in the data that may be useful to produce insights from the data. The results of this analysis will be used to create in the next section final visualizations.

Distribution of individual contributions

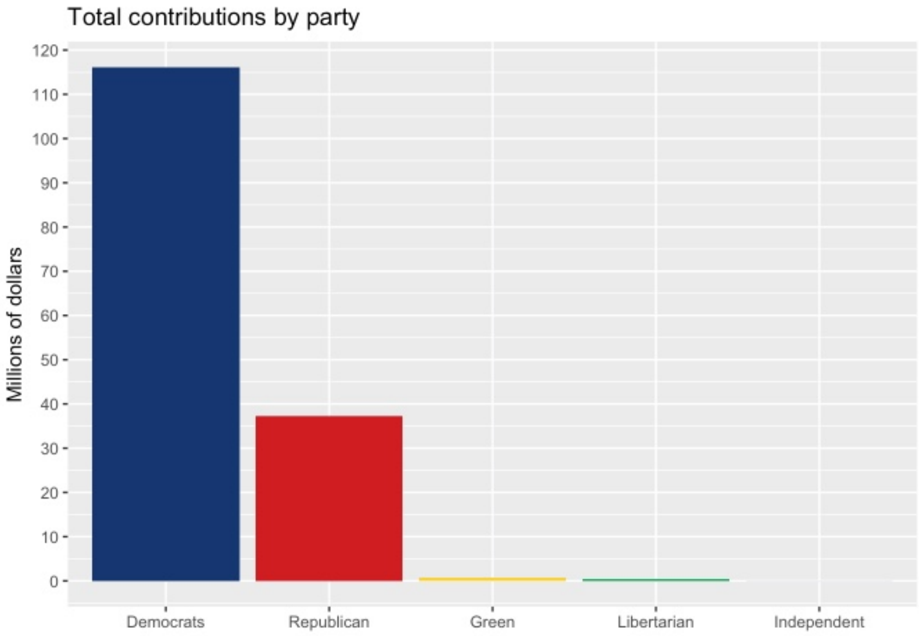




In the first graph it can be seen that the individual contributions to the Presidential Campaign do not show a normal distribution behavior, also that the distribution is skewed to the left.

In the second graph, a zoom was made to the contributions between 0 and 300 dollars, revealing that most of the contribution transactions to the campaign range between 0 and 50 dollars per contributor. Just to give some context to those amounts consider that the minimum salary in the State of California is $10/hour.

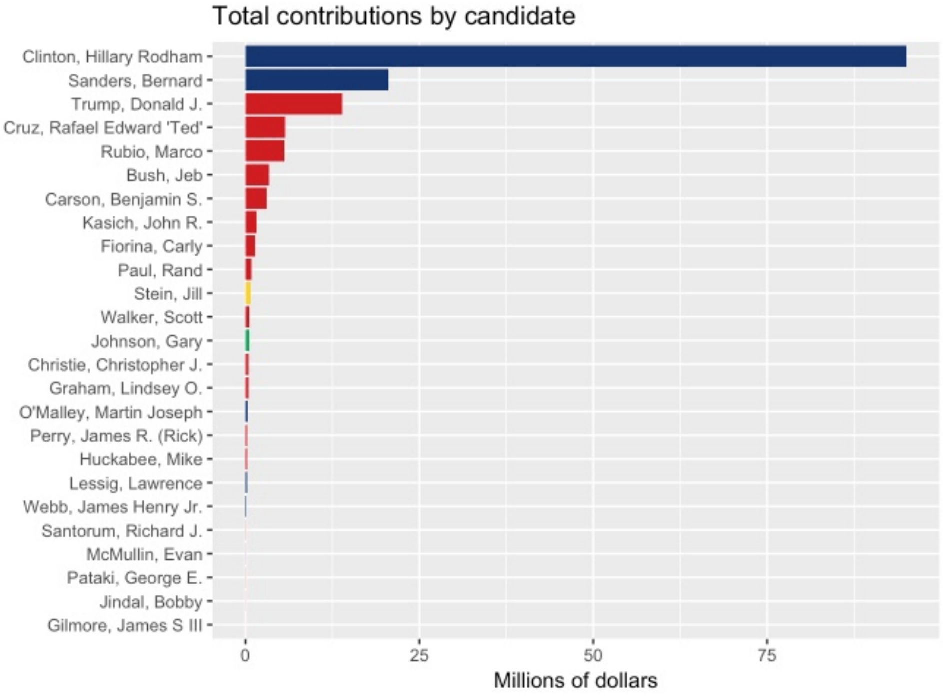
Total contributions by political party



The Democrat Party raised approximately 115 million dollars compared to the Republican Party with approximately 37 million dollars. In other words, Democrats raised 3.10 times more money than Republicans.

The smaller parties raised around 1 million dollars each one.

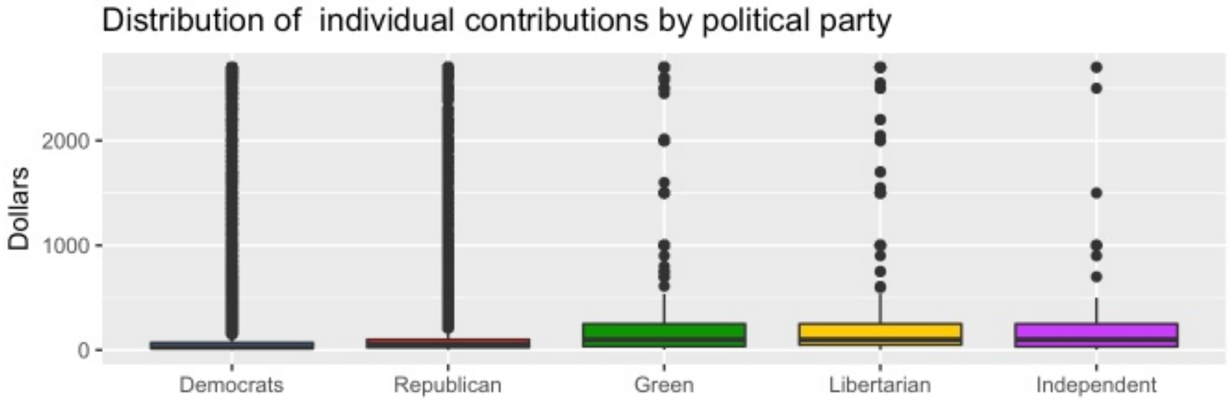
Total contributions by candidate and political party



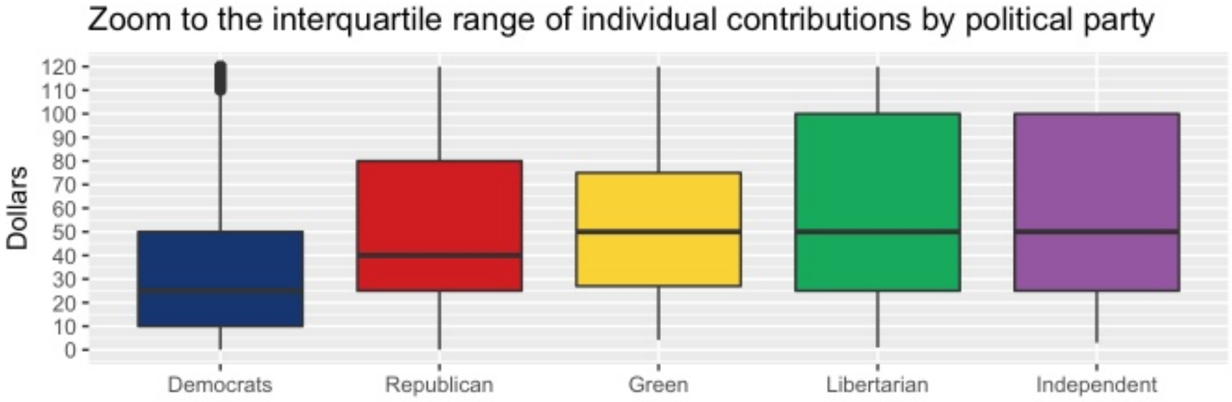
The total individual contributions to the Democrats were to support Hillary Clinton. Also, it is worth to point out that in the State of California Bernard Sanders raised more money than the current President Donald Trump.

The Republican total individual contributions were splitted between Donald Trump, Marco Rubio and Ted Cruz.

Interquartile range of contributions by political party



Distribution of individual contributions by political party

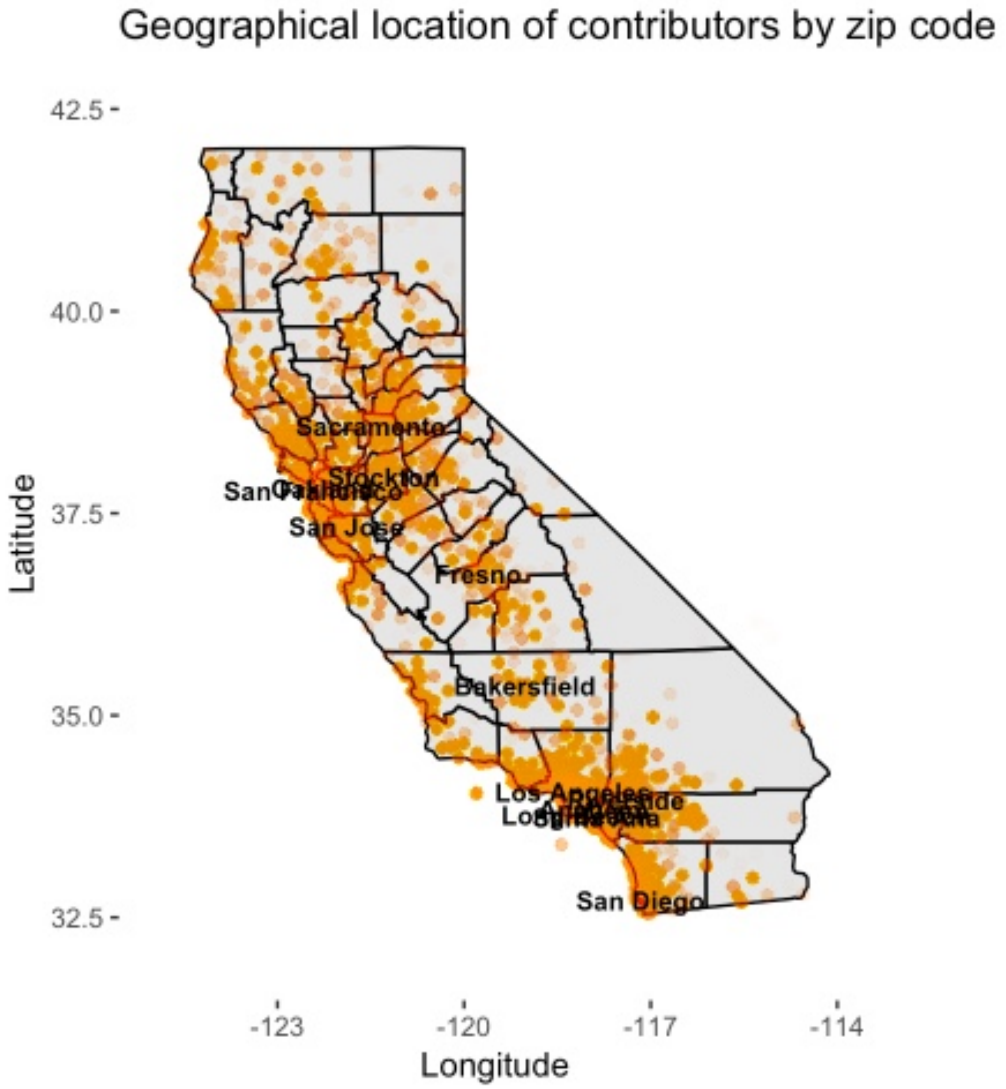


Zoom to the interquartile range of individual contributions by political party

The first set of boxplots show that there are many outliers in the individual contributions by political party. A zoom to the interquartile range where middle 50% group of individual contributions reveals the following patterns.

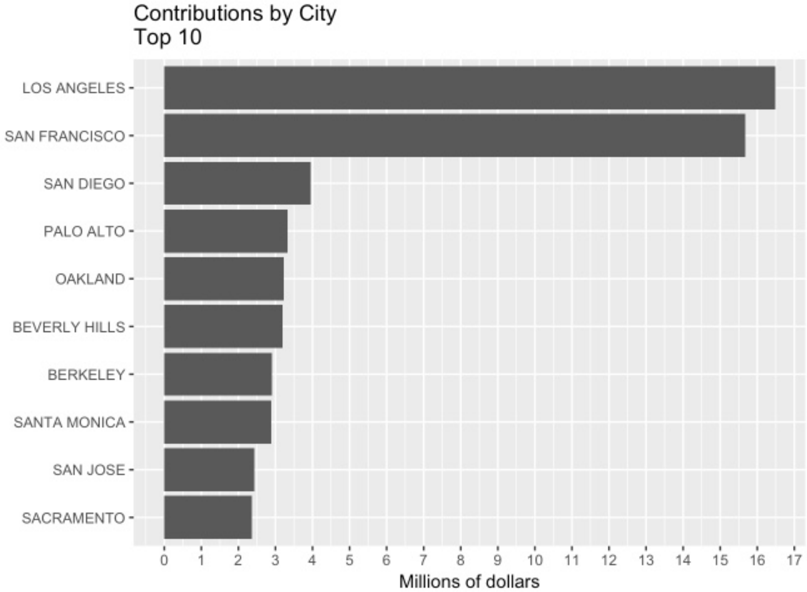
Although Democrats accrued most of the total contributions, their supporters contributed with smaller amounts of money ranging between 10 and 50 dollars. The Republican Party received contributions between 25 and 80 dollars. More interestingly is to see that smaller parties like the Libertarian and the Independent Candidate received half of their contributions in a range of 25 and 100 dollars.

Geographical location of contributors in California



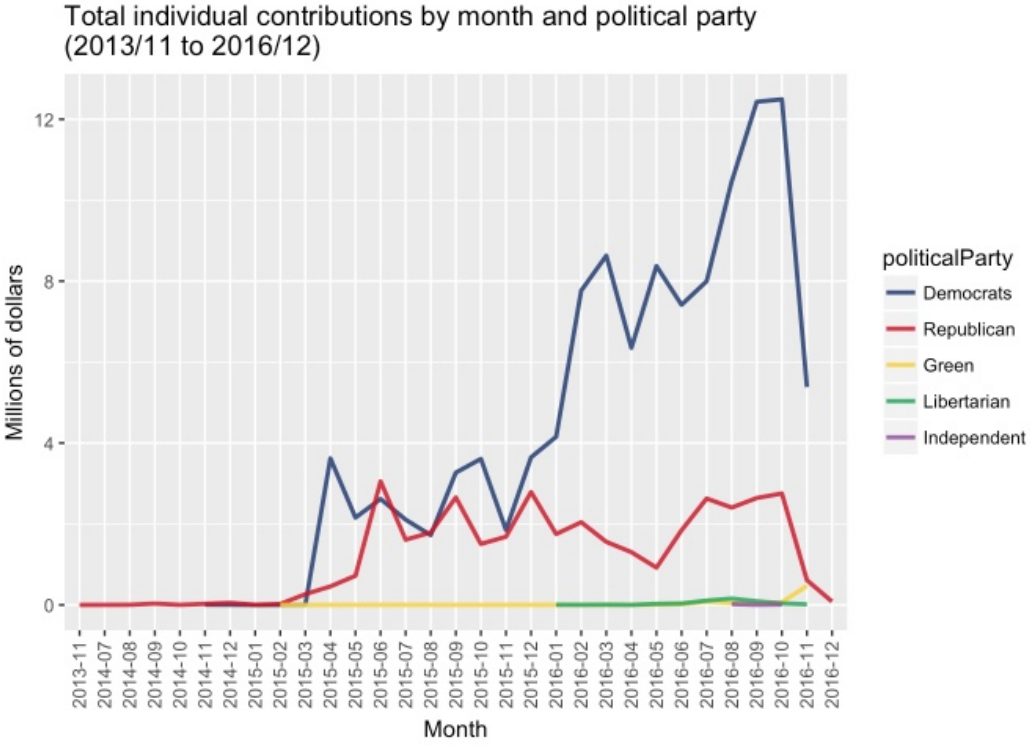
The map shows that in the State of California contributors are clustered around cities with more than 300,000 inhabitants such as: Los Angeles, San Diego, San Jose, San Francisco, Fresno, Sacramento, Long Beach, Oakland, Bakersfield, Anaheim, Santa Ana, Riverside and Stockton.

Top ten cities by total contributions



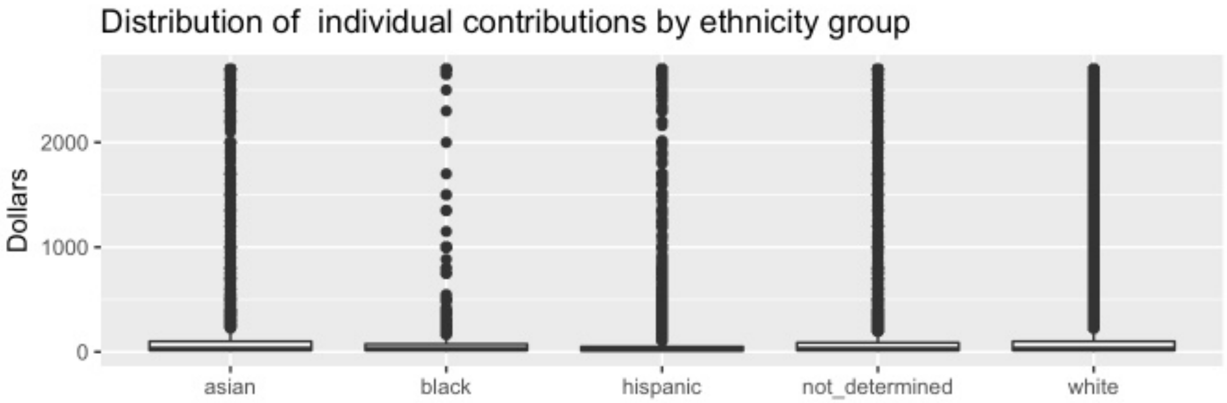
The cities in the State of California with more total individual contributions are Los Angeles and San Francisco. In a smaller proportion are San Diego, Palo Alto, Oakland, Beverly Hills, Berkeley, Santa Monica, San Jose and Sacramento.

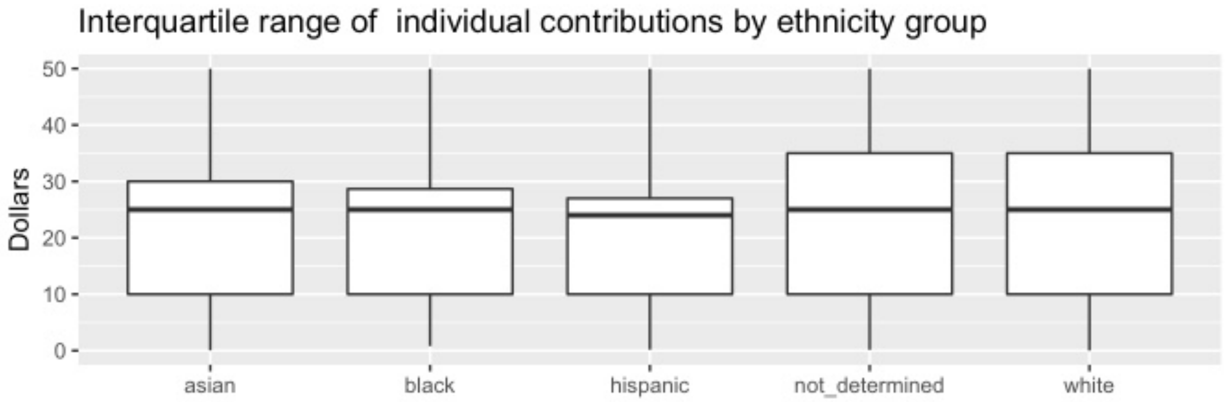
Contributions in time



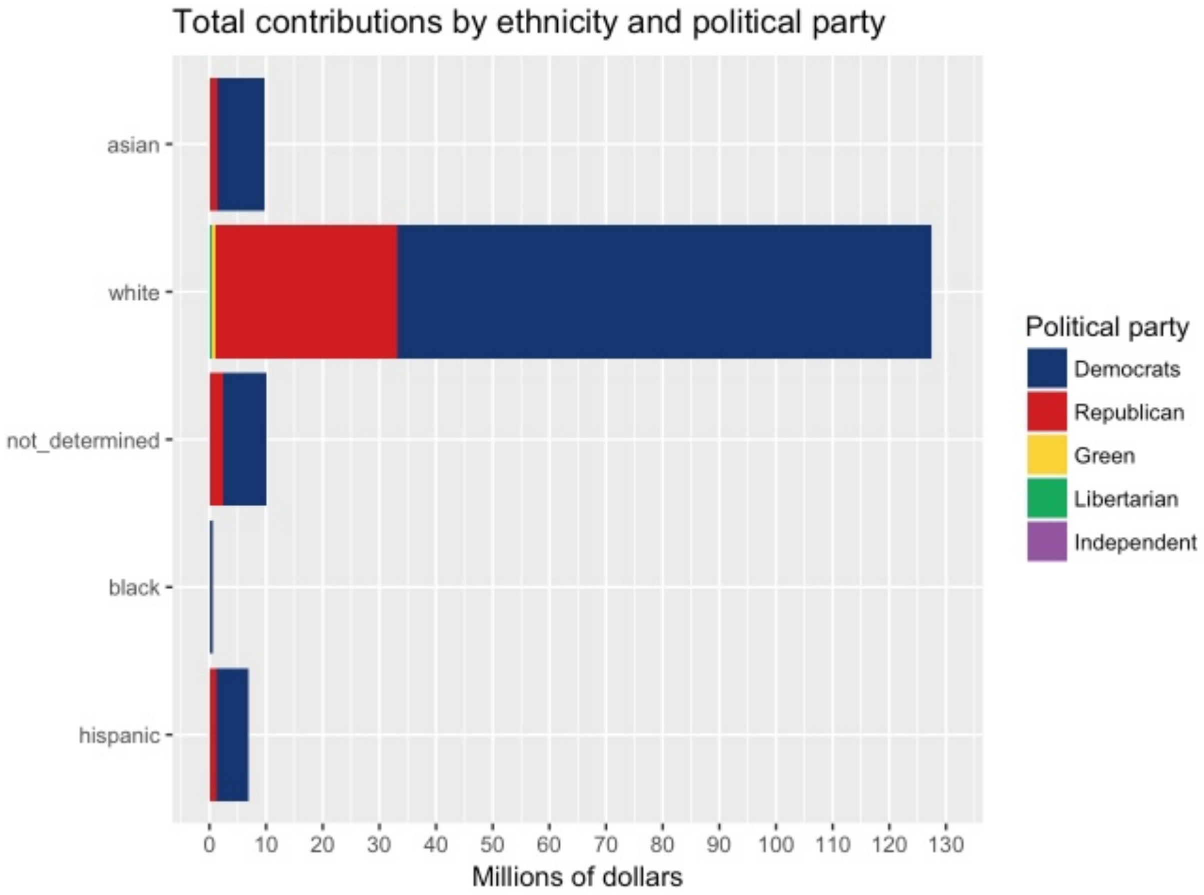
The time series graph shows that individual contributions for the Presidential Campaign started at the end of 2013 and finished on december 2016. In the State of California since April 2015 the Democrat party continuously raised increasing amounts of money.

Contributions by ethnicity


Distribution of individual contributions by ethnicity group


Interquartile range of individual contributions by ethnicity group

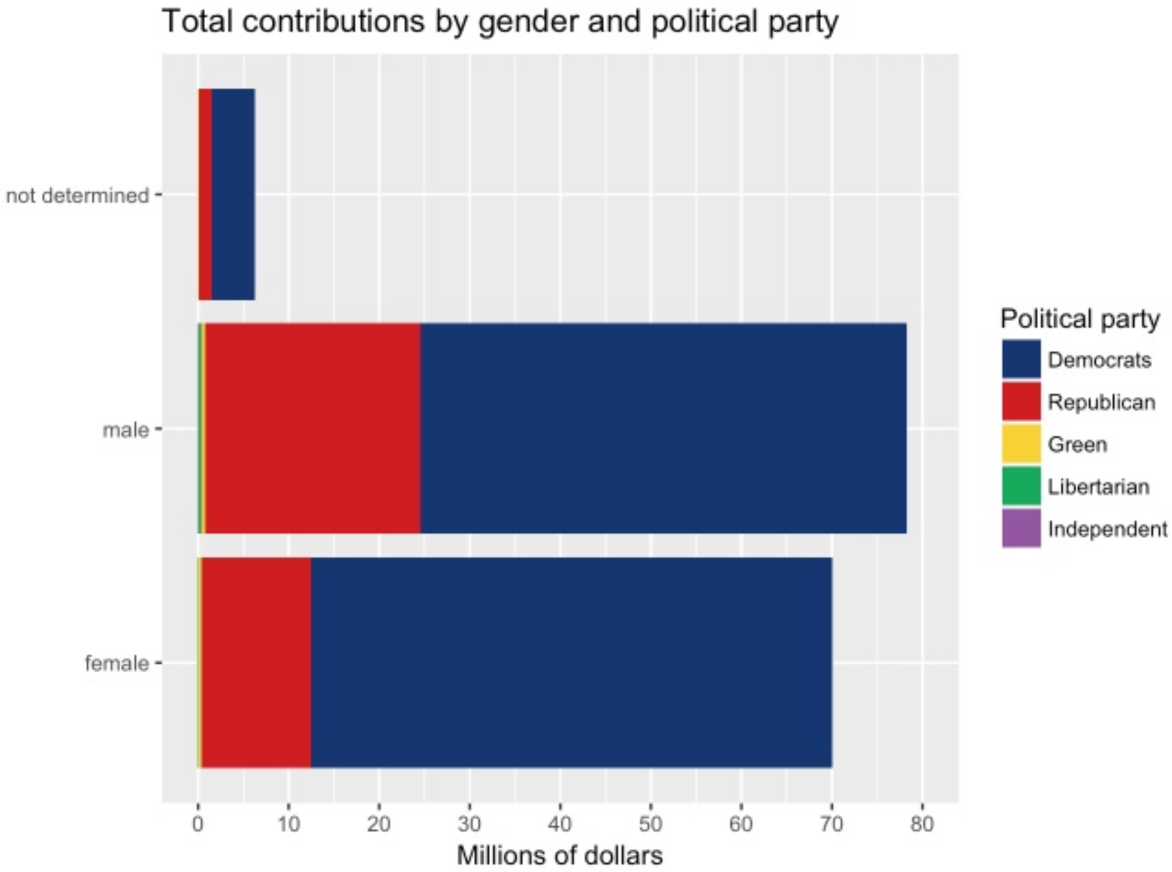
The graphs show that the group with the greatest interquartile range is the white ethnicity group (10-35 dollars). While the others minority groups is more of less equal (10-30 dollars).



The graph shows that the white ethnicity group is the one that contributed with almost 125 million dollars to the presidential campaign. Their support was splitted between Democrats (90 million) and Republicans (34 million).

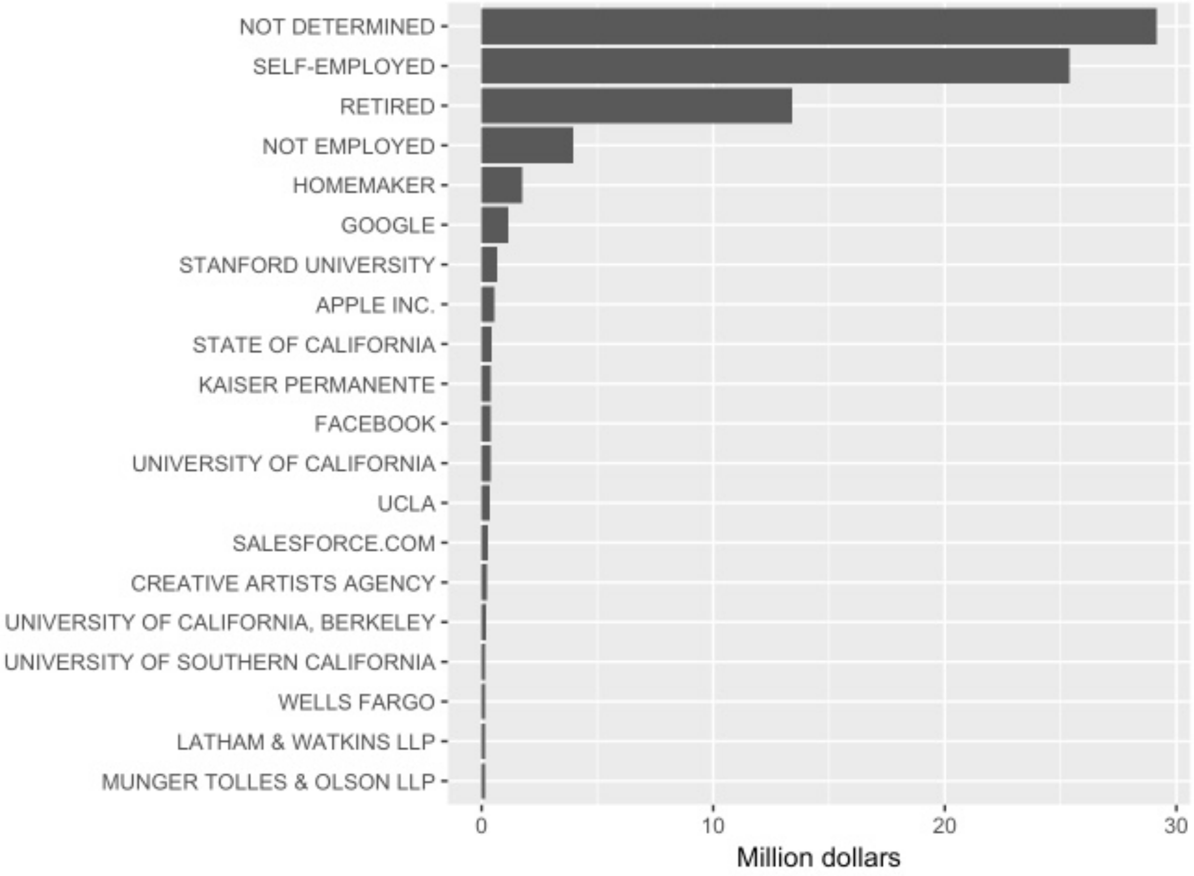
Among the minority groups it is worth to point out that the Asians group contributed with almost 10 million dollars, the Hispanic group with 7 million dollars, and the black group with almost 1 million. It is also worth to point out that in general the support was to the Democrat party.

Gender Plot



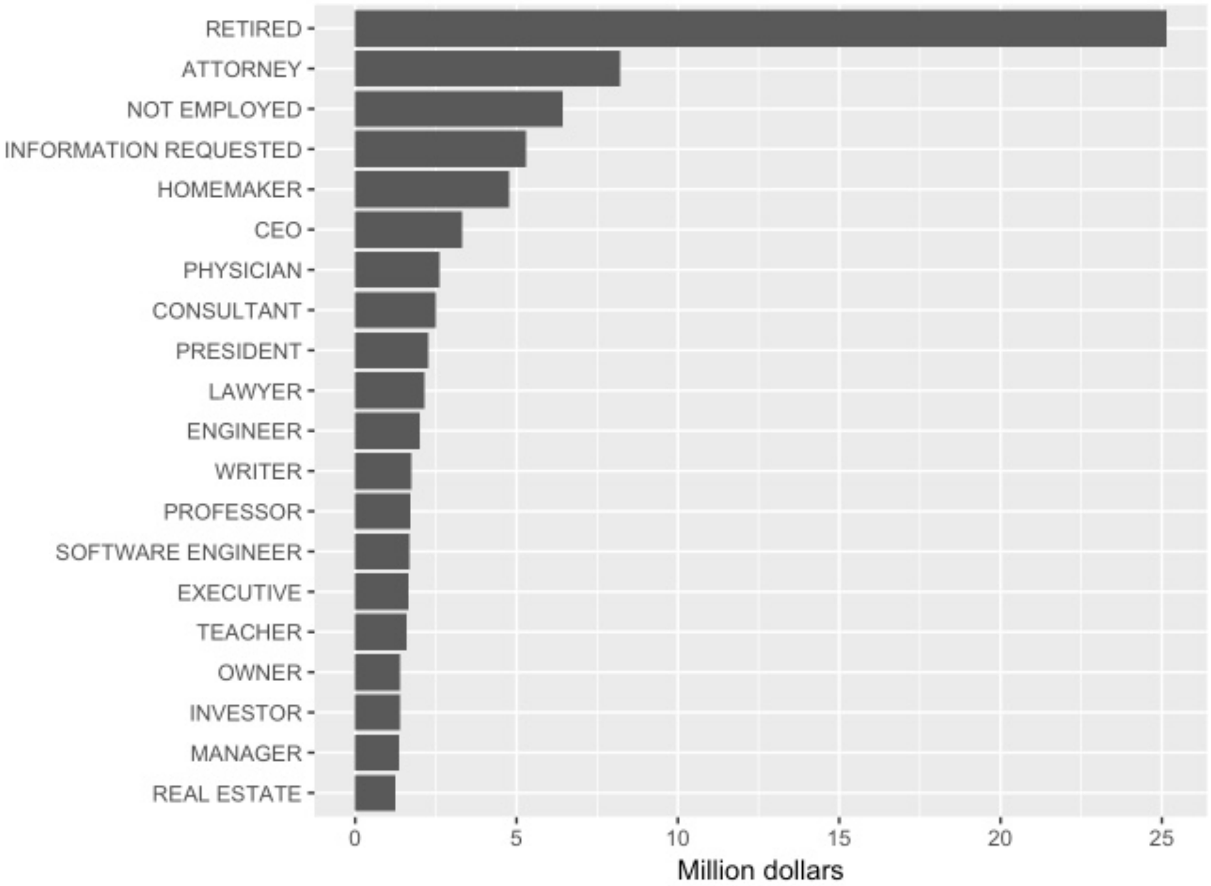
Total individual contributions come mostly from male (78 million dollars) and then female (70 million dollars). There are 6 million dollars that provide from contributors whose gender was not being able to be predicted.

Contributions by employer



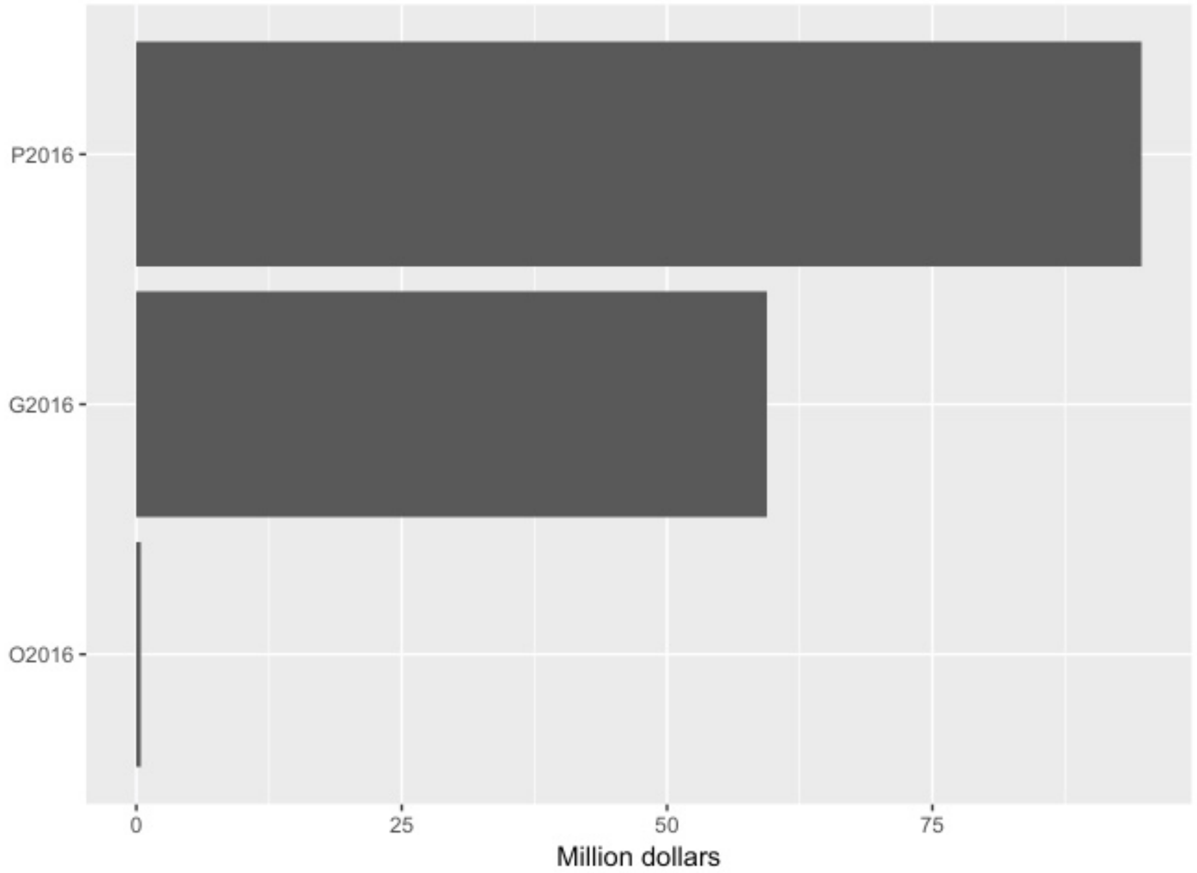
Contributions by occupation

The graph shows that most of the individual contributions came from contributors that did not specified an employer, were self-employed, retired and not employed.



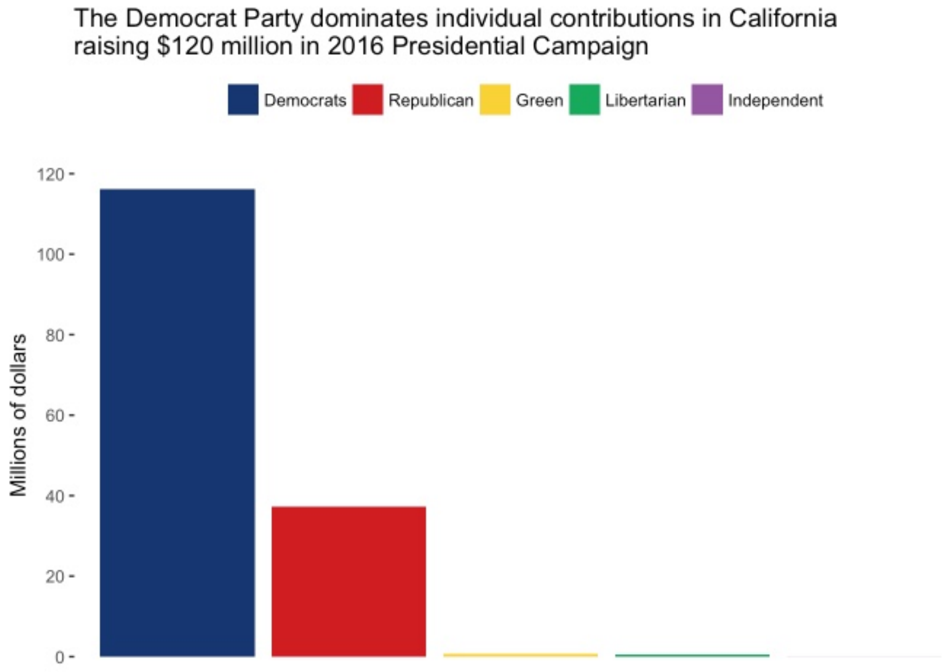
The graph shows that most of the contributions to the Presidential campaign came from retired people, attorneys, not employed, not determined, homemakers and CEOs.

Contributions by Presidential Campaign type



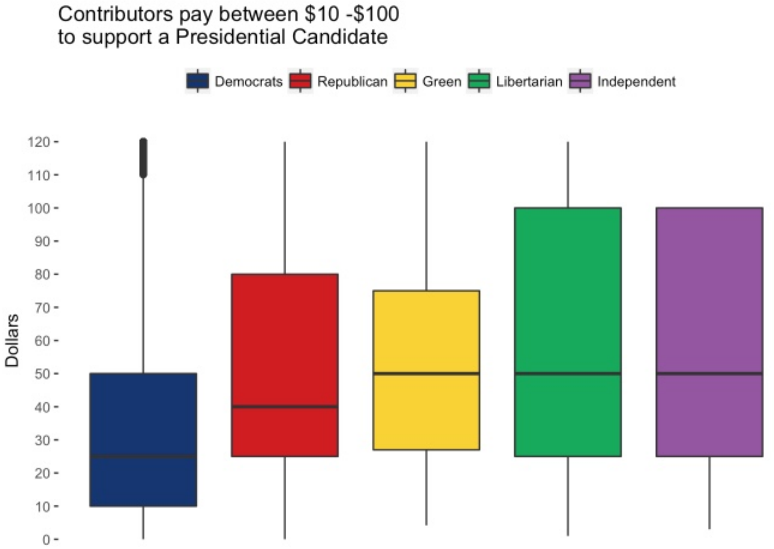
The graph shows that all the contributions considered in this analysis are from 2016 and that most of the contributions took place in the Presidential and the General Campaign.

Final Plots and Summary

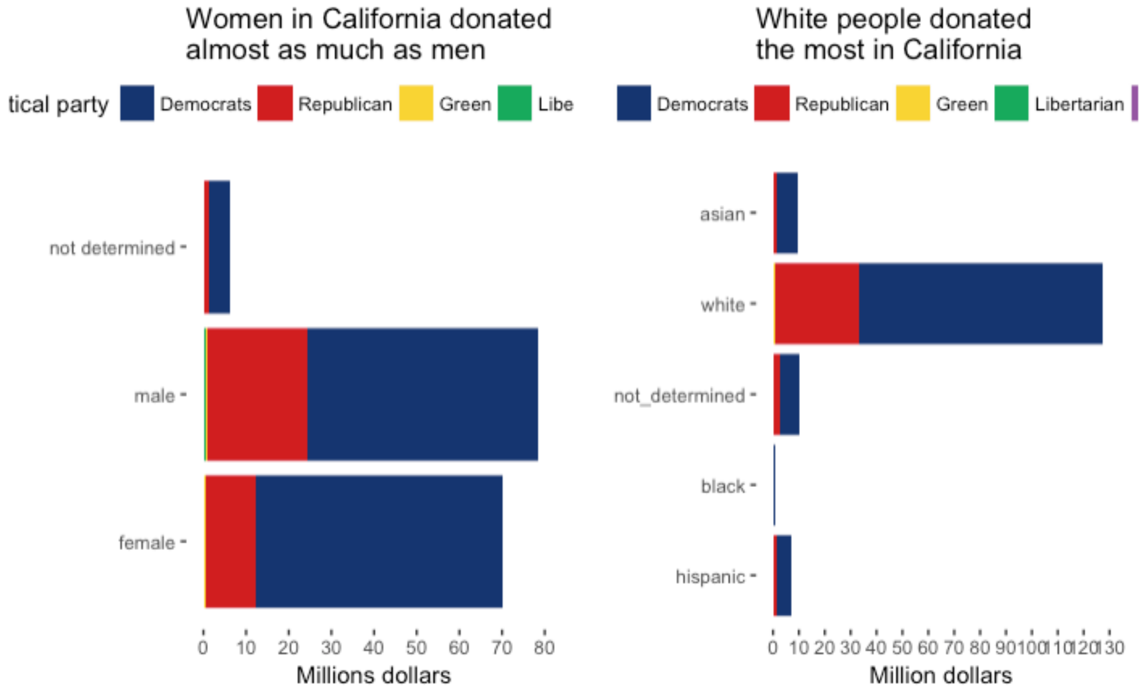


In the State of California during the Presidential Election of 2016, the Democrat party was able to raise almost $120 million in contributions, almost three times more what the Republican Party raised.

Surprisingly, although the Democrats raised the largest amount of contributions, it shows the lowest median contribution. The parties with the highest median contribution are the Green, Libertarian and Independent.



Surprisingly, although the Democrats raised the largest amount of contributions, it shows the lowest median contribution. The parties with the highest median contribution are the Green, Libertarian and Independent.



The previous graph shows contributions to all parties broken by gender and ethnicity. It can be seen that women are an important actors in financing the Presidential Campaign. In 2016 they gave 70 million dls, just 10 million dls less than men.

Furthermore, even that the State of California has one of the largest immigrant population, people of white ethnicity are the largest contributors. From the minorities, it is Asians that gave the most contributions to the Presidential Campaign followed closely by Hispanic. Moreover, there are not differences in voting preference by ethnicity and gender in the State.

Reflections:

The purpose of using Exploratory Data Analysis (EDA) with the 2016 Presidential Campaign contributions was to understand the behavior of contributors. Considering the huge size of the entire dataset for each State of the United States, this analysis was delimited to the State of California.

Data visualization techniques showed that most of the contributions in the State of California significantly supported the Democrat Party in 2016. This result confirms what was already expected. However, interesting results arose when comparing the distribution of contributions by party. It is the Green, Libertarian and Independent parties the ones that receive larger donations. Also, exploring by gender showed there is a small difference of 10 million dollars between men and women total contributions. In this sense women are important contributors to financial campaigns. Furthermore, a relevant discovery is that a State mostly populated by immigrants such as California receives most donations from people of white ethnicity. Bearing the above, the objectives of the EDA were achieved successfully.

The most difficult part of the analysis was the creation of the gender and ethnicity variables. Each of these variables were created using Bayesian methods and accessing government data through APIs. The predictions while not 100% accurate provide a good degree of knowledge about the characteristics of the contributor.

This analysis was successful in addressing the objectives considering socio demographic variables and taking advantage of geographical information as well. Further patterns can be explored using more sophisticated techniques such as machine learning.

This analysis will be very useful if it were conducted in the rest of the data. Also, it would be good to compare this year’s contributions vs previous ones. These with the purpose of understanding if there are changes in patterns over time and election periods. Moreover, the disaggregation of the data is good enough to conduct predictive models for future elections.