Data Visualisation Task - US Presidential Election Donations

RUG @ HSG

Task Description

The US Election campaign of 2020 between Joe Biden and Donald Trump was a hotly debated topic. Have you ever wondered who donated to the campaign of those two and if so, how much was donated but also which genders donated? The fact that women tend to support Democrats more than Republicans is widely known, but is this also reflected in donation behavior to the presidential candidates? And did the fact that Trump's comments and attitude towards women were not always exactly flattering play a role in how many men and women donated in the election? What about Biden's announcement in early August 2020 that he would pick Kamala Harris as a Vice President - Did this boost his donations from women?

The following exercise will allow you to gain an insight into these questions with real data coming directly from the US Government.

The Data

By US law, everyone who cumulatively donates more than 200 USD per election cycle must be reported to a database which is publicly available. We thus have access to all donations over 200 Dollars. Some donations which are less than 200 USD are also reported, but we have to assume that the majority of those donations is not captured in our data set. (Thus, when drawing conclusions, we have to be aware that some donations cannot be analyzed, but the sample we have is extremely large and covers the majority of donations).

The original data set is available at the Federal Election Commission and is over 16 GB large. We have already done some pre-cleaning for you, which leaves us with a file of under 400 MB which contains all relevant information for this exercise.

Download the pre-cleaned donation dataset from here

```
library(tidyverse)
donations <- read_csv("donations.csv")
glimpse(donations)
```

```
## Rows: 8,249,137
## Columns: 6
## $ NAME            <chr> "SUSIN, DANIEL", "HALL GUAY, CATHERINE", "KIELY, M", "~
## $ ZIP_CODE        <chr> "48314", "21409", "24401", "30005", "90230", "63034", ~
## $ OCCUPATION      <chr> "BUS DRIVER", "SPECIAL EDUCATOR", "NOT EMPLOYED", "SAL~
## $ TRANSACTION_DT  <chr> "10252020", "10252020", "10252020", "10252020", "10252~
## $ TRANSACTION_AMT <dbl> 15, 10, 20, 25, 10, 25, 20, 20, 25, 15, 25, 10, 25, 15~
## $ CANDIDATE       <chr> "B", "B", "B", "B", "B", "B", "B", "B", "B", "B", "B",~
```

As you can see, the data file contains the name of the donor, the donation date, the amount donated and other information such as employment - and residential information. However, unfortunately, as it is often the case in real life, we miss some crucial data, namely the gender of the donors. This is not reported by the FEC.

So what do we do?

If you dig around for long enough, you will find that the Social Security Administration in the US publishes a text file every year which contains for all babies born in the United States in a given year the names, their gender as well as the number of occurrences. This will help us in assigning a gender to each name in the donations data set.

Download the folder containing all text files from here

Processing the Data

First, lets read in all those baby names. As the files are separated by year, you need to find a way to combine all the files to one data frame.

Below, you find one possible way to do this.

```
# get the pathname of all files in the folder
filenames <- list.files("names")

#create empty data frame
baby_names <- data.frame()

#loop over all files, read them and paste to names df
for (i in filenames){
  temp <- read_delim(paste("names/", i, sep = ""), delim = ",", col_names = FALSE)
  baby_names <- bind_rows(baby_names, temp)
}

# add column names to the data frame
colnames(baby_names) <- c("NAME", "GENDER", "COUNT")
```

If you skim through the data, you can see that some names are used for both boys and girls. So how do we assign a definitive gender to each name? What we could do is group all observations by gender and first name, and sum up over the occurrences. After some modifications, we can calculate the % of times each name is used for males and females and establish the following rule:

- If occurrences Female / total occurences $\geq 0.75 \rightarrow$ Female
- If occurences Male / total occurences $\geq 0.75 \rightarrow$ Male
- else $\rightarrow$ NA (we cannot determine the gender with high enough accuracy)

Use commands such as `summarize()`, `group_by()` and possibly `pivot_wider()`

Your output after performing the calculations may look something like the following:

```
## # A tibble: 4 x 5
##   NAME       M     F RATIO_M GENDER
##   <chr>  <dbl> <dbl>   <dbl> <chr>
## 1 Aari      33    25  0.569  NA
## 2 Aero     362    88  0.804  M
## 3 Aspen   2086 19386  0.0971 F
## 4 Athena     0 41099  0      F
```

In a second step, we need to join the first names of the *baby_names* data frame onto the first names of the *donations* data set.

Problem:

The Donation Data set contains the first name and last name together. Further, all names are in upper case.

Solution:

1. transform baby names to upper case or donation names to lower case
2. Split the names in the *donations* data frame into last and first name (at the occurrence of the first comma)

💡 In order to split a string, you can use the `str_split` command.

```
#extract name vector
donor_names <- donations$NAME
#split name vector by comma (separates first and last name)
first_name <- str_split(donor_names, pattern = ",", simplify = TRUE)[,2]
```

⚠ **Attention:** ⚠

As you may suspect, the names are sometimes messed up and you will need to find a solution to handle some common errors, if you want to maximize the matches between your first names in the *donations* data set and the *baby_names* data set

Lets have a look at some names which may cause you a headache:

- HARLAN JR., ANDREW D.
- GUSTAFSON, JOHN & CONNIE
- MEEK, CHARLES RONALD (RON)

After splitting these at the first comma and selecting the second part, we are left with:

- ANDREW D.
- JOHN & CONNIE
- CHARLES RONALD (RON)

Our goal should be to get one unique first name (to maximize the likelihood to match it with a name of the *baby_names* data set)

→ Your task is to think of ways how to do this

💡 if you have multiple names, you could as a rule select the longest string as the "proper name". (eg. First Name = CHARLES RONALD →First Name = CHARLES)

To see how you can use the power of *regular expressions* to achieve your task, check out the regex cheatsheet

Once you have successfully cleaned the names, join the first names vector back onto your *baby_names* dataframe and finally join the dataframe onto the *donations* dataframe (join by first name).

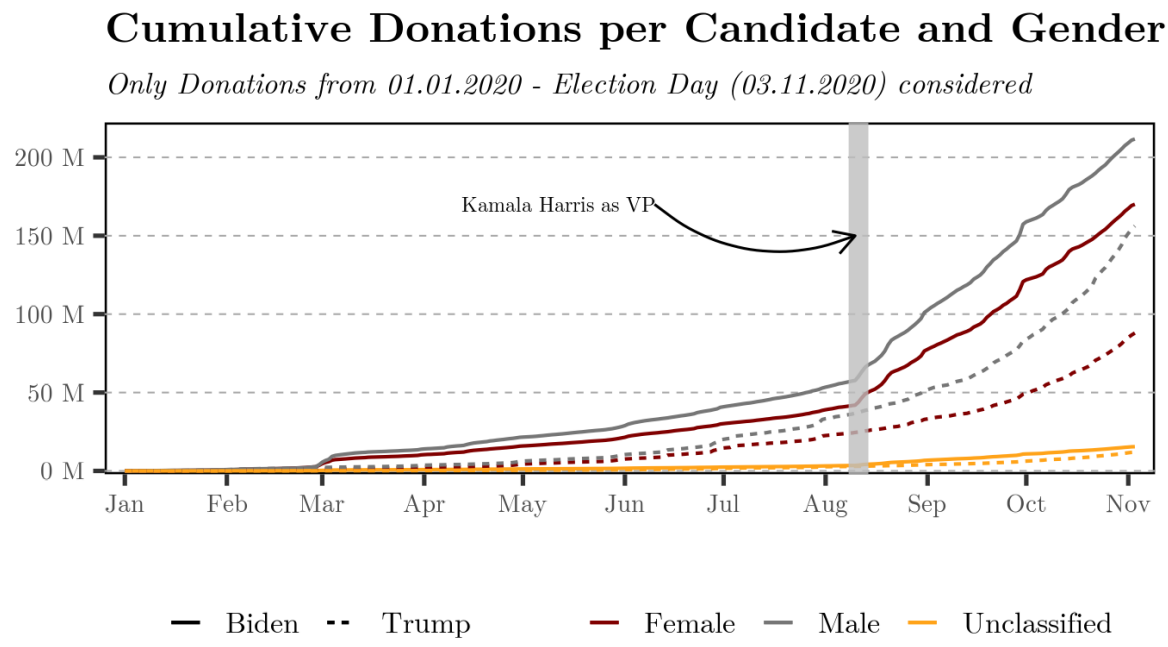
Calculations and Plotting

Now, you are ready for doing some plots and calculations! For example, you could calculate and/or visualize the following:

- How many female and male donors were there for Trump and Biden? (⚠ some people may have donated multiple times. Think of a way to group these people together with the information you have in the data frame)

- Visualize the cumulative donations / monthly donations by Candidate and check, if the appointment of Kamala Harris in August changed the Donation Pattern to Biden
- How much money was donated by each Gender to Biden and Trump? (in absolute & relative terms)
- What employment did the donors have? Is there one group that stands out? Is there a difference between male and female and between those who donated to Trump and Biden?

Sample Plots



Own Depiction | Source: Federal Election Commission

Feel free to ask questions in our Q&A chat and hand in your charts and code. We are looking forward to your submissions!