# Lab 2

Yaqi Shi, 1003813180

2023-01-18

## Lab Exercises

### Q1

1. Using the `delay_2022` data, plot the five stations with the highest mean delays. Facet the graph by `line`

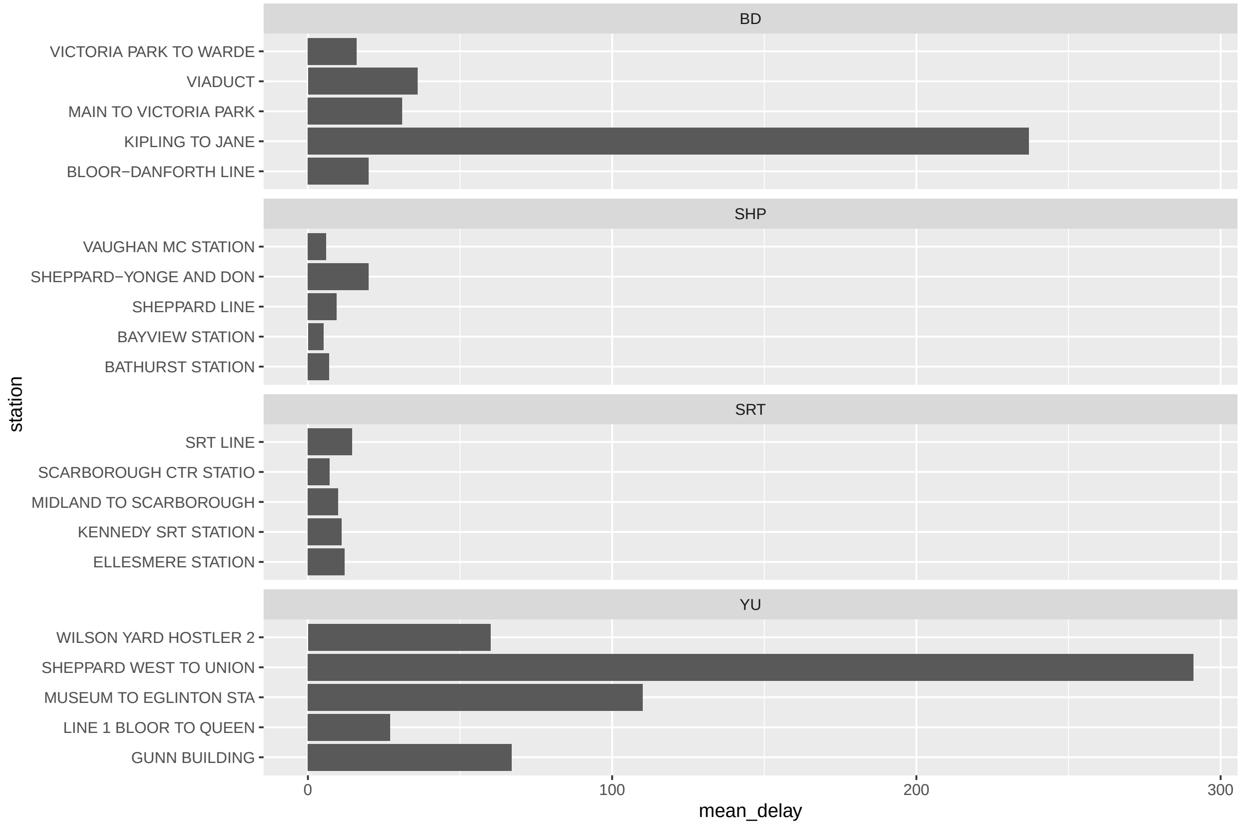
**Answer**

Since I am doing the lab exercise in a separate file, first I need to build the `delay_2022` data.

```
# Based on the lab file to retrieve Data
all_data <- list_packages(limit = 500)
res <- list_package_resources("996cfe8d-fb35-40ce-b569-698d51fc683b")
res <- res %>% mutate(year = str_extract(name, "202.?"))
delay_2022_ids <- res %>% filter(year==2022) %>% select(id) %>% pull()
delay_2022 <- get_resource(delay_2022_ids)
delay_2022 <- clean_names(delay_2022)
delay_2022 <- delay_2022 %>% distinct()
delay_codes <- get_resource("3900e649-f31e-4b79-9f20-4731bbfd94f7")
delay_data_codebook <- get_resource("ca43ac3d-3940-4315-889b-a9375e7b8aa4")

# Since we are doing filtering in the next step, want to save a unfiltered copy
delay_2022_orig <- delay_2022

delay_2022 <- delay_2022 %>% filter(line %in% c("BD", "YU", "SHP", "SRT"))

# There is no filtering after this step, mostly joint new information
delay_2022 <- delay_2022 %>%
  left_join(delay_codes %>% rename(code = `SUB RMENU CODE`, code_desc = `CODE DESCRIPTION...3`) %>% sel

delay_2022 <- delay_2022 %>%
  mutate(code_srt = ifelse(line=="SRT", code, "NA")) %>%
  left_join(delay_codes %>% rename(code_srt = `SRT RMENU CODE`, code_desc_srt = `CODE DESCRIPTION...7`)
  mutate(code = ifelse(code_srt=="NA", code, code_srt),
         code_desc = ifelse(is.na(code_desc_srt), code_desc, code_desc_srt)) %>%
  select(-code_srt, -code_desc_srt)

delay_2022 <- delay_2022 %>%
  mutate(station_clean = ifelse(str_starts(station, "ST"), word(station, 1,2), word(station, 1)))

delay_2022 <- delay_2022 %>%
  mutate(code_red = case_when(
    str_starts(code_desc, "No") ~ word(code_desc, 1, 2),
    str_starts(code_desc, "Operator") ~ word(code_desc, 1,2),
    TRUE ~ word(code_desc,1))
         )
```

Plot the five stations with the highest mean delays. Facet the graph by `line`

```
delay_2022 %>%
  group_by(line, station) %>%
  summarise(mean_delay = mean(min_delay)) %>%
  arrange(-mean_delay) %>%
  slice(1:5) %>%
  ggplot(aes(x = station,
             y = mean_delay)) +
  geom_col() +
  facet_wrap(vars(line),
             scales = "free_y",
             nrow = 4) +
  coord_flip()
```



## Q2

2. Using the `opendatatoronto` package, download the data on mayoral campaign contributions for 2014. Hints:
   - find the ID code you need for the package you need by searching for 'campaign' in the `all_data` tibble above
   - you will then need to `list_package_resources` to get ID for the data file
   - note: the 2014 file you will get from `get_resource` has a bunch of different campaign contributions, so just keep the data that relates to the Mayor election

**Answer**

```
# Find the ID of the dataset
all_data <- list_packages(limit = 500)
res <- list_package_resources("f6651a40-2f52-46fc-9e04-b760c16edd5c")
campaign2014 <- get_resource("5b230e92-0a22-4a15-9572-0b19cc222985")

# Load the Mayor election dataset
mayoral_campaign_2014 <- campaign2014$`2_Mayor_Contributions_2014_election.xls`
```

## Q3

3. Clean up the data format (fixing the parsing issue and standardizing the column names using `janitor`)

**Answer**

```
# Fix the parsing issue
mayoral_campaign_2014 <-  row_to_names(mayoral_campaign_2014,1)

# Standardize the column names
mayoral_campaign_2014 <- clean_names(mayoral_campaign_2014)

# Show the first few lines of the
head(mayoral_campaign_2014)
```

```
## # A tibble: 6 x 13
##   contributors~1 contr~2 contr~3 contr~4 contr~5 goods~6 contr~7 relat~8 presi~9
##   <chr>          <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>
## 1 A D'Angelo, T~ <NA>    M6A 1P5 300     Moneta~ <NA>    Indivi~ <NA>    <NA>
## 2 A Strazar, Ma~ <NA>    M2M 3B8 300     Moneta~ <NA>    Indivi~ <NA>    <NA>
## 3 A'Court, K Su~ <NA>    M4M 2J8 36      Moneta~ <NA>    Indivi~ <NA>    <NA>
## 4 A'Court, K Su~ <NA>    M4M 2J8 100     Moneta~ <NA>    Indivi~ <NA>    <NA>
## 5 A'Court, K Su~ <NA>    M4M 2J8 100     Moneta~ <NA>    Indivi~ <NA>    <NA>
## 6 Aaron, Robert~ <NA>    M6B 1H7 250     Moneta~ <NA>    Indivi~ <NA>    <NA>
## # ... with 4 more variables: authorized_representative <chr>, candidate <chr>,
## #   office <chr>, ward <chr>, and abbreviated variable names
## #   1: contributors_name, 2: contributors_address, 3: contributors_postal_code,
## #   4: contribution_amount, 5: contribution_type_desc,
## #   6: goods_or_service_desc, 7: contributor_type_desc,
## #   8: relationship_to_candidate, 9: president_business_manager
```

## Q4

4. Summarize the variables in the dataset. Are there missing values, and if so, should we be worried about them? Is every variable in the format it should be? If not, create new variable(s) that are in the right format.

### Answer

```
# Summarize the dataset
skim(mayoral_campaign_2014)
```

Table 1: Data summary

| | |
|---|---|
| Name | mayoral_campaign_2014 |
| Number of rows | 10199 |
| Number of columns | 13 |
| Column type frequency: | |
| character | 13 |
| Group variables | None |

**Variable type: character**

| skim_variable | n_missing | complete_rate | min | max | empty | n_unique | whitespace |
|---|---|---|---|---|---|---|---|
| contributors_name | 0 | 1 | 4 | 31 | 0 | 7545 | 0 |
| contributors_address | 10197 | 0 | 24 | 26 | 0 | 2 | 0 |
| contributors_postal_code | 0 | 1 | 7 | 7 | 0 | 5284 | 0 |
| contribution_amount | 0 | 1 | 1 | 18 | 0 | 209 | 0 |
| contribution_type_desc | 0 | 1 | 8 | 14 | 0 | 2 | 0 |
| goods_or_service_desc | 10188 | 0 | 11 | 40 | 0 | 9 | 0 |
| contributor_type_desc | 0 | 1 | 10 | 11 | 0 | 2 | 0 |
| relationship_to_candidate | 10166 | 0 | 6 | 9 | 0 | 2 | 0 |
| president_business_manager | 10197 | 0 | 13 | 16 | 0 | 2 | 0 |
| authorized_representative | 10197 | 0 | 13 | 16 | 0 | 2 | 0 |
| candidate | 0 | 1 | 9 | 18 | 0 | 27 | 0 |
| office | 0 | 1 | 5 | 5 | 0 | 1 | 0 |
| ward | 10199 | 0 | NA | NA | 0 | 0 | 0 |

```
# Create new variable in correct format
mayoral_campaign_2014$contribution_amount_num <- as.numeric(mayoral_campaign_2014$contribution_amount)
```

In this dataset, there are 13 variables with 10199 observations. There are 6 variables that has missing values and we do not need to worry about them.

Among those 6 variables, “contributors_address”, “president_business_manager” and “authorized_representative” has 10197 missing values, which means that only two observation has value in these variables. The variable “ward” is missing for all 10199 observations. Except for those, the variable “goods_or_service_desc” has 10188 missing values and the variable “relationship_to_candidate” has 10166 missing values. For the variable “relationship_to_candidate”, it is reasonable to have missing values as most of the relationship to candidate are unknown. For the variable “goods_or_service_desc”, this is reasonable as only non-monetary contribution needs description and there are 10188 monetary contribution which corresponds to the 10188 missing values. Based on the summary, most of the variables are missing for more than 99% of the observations. We don’t need to worry about or manipulate those variables as the missing proportion is too much.

The other 7 variables: "contributors_name", "contributors_postal_code", "contribution_amount", "contribution_type_desc", "contributor_type_desc", "candidate" and "office" do not have any missing value for all 10199 observations. Thus no missing value issue for these 7 variables.

In terms of format, not every variable in the format it should be. In this dataset, every variable is originally in the character format. Based on the missing value analysis above, we will only discuss the 7 variables that don't have missing value.

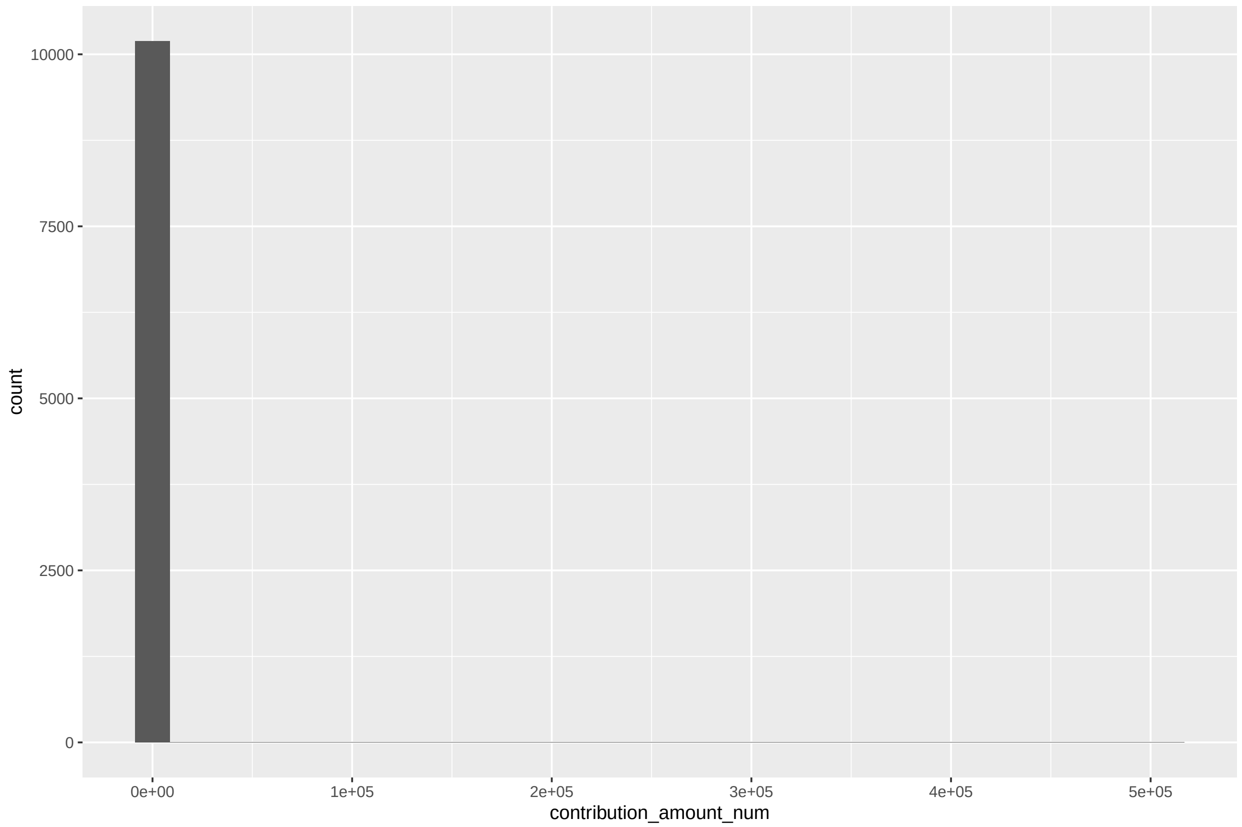
Among the 7 variables, the variable that's in the incorrect formats is "contribution_amount". Since it is an amount variable, we have convert it into numeric and create the new variable "contribution_amount_num".

## Q5

5. Visually explore the distribution of values of the contributions. What contributions are notable outliers? Do they share a similar characteristic(s)? It may be useful to plot the distribution of contributions without these outliers to get a better sense of the majority of the data.

**Answer**

```r
# Distribution of all contribution amount
ggplot(data = mayoral_campaign_2014) +
  geom_histogram(aes(x = contribution_amount_num))
```



Based on the above plot, we can see that the x axis extends to a very large number with the histogram squeeze in the left end of the graph. This is an indication that there are large outliers in the contribution amount data. Now we take a loot at those large amounts.

```
outlier <- mayoral_campaign_2014 %>%
  arrange(.,-contribution_amount_num) %>%
  select(contributors_name, contribution_amount_num, relationship_to_candidate, contribution_type_desc,
  slice(1:20)

outlier[1:10,]
```

```
## # A tibble: 10 x 5
##    contributors_name contribution_amount_num relationship_to_c~1 contr~2 contr~3
##    <chr>                               <dbl> <chr>               <chr>   <chr>
##  1 Ford, Doug                        508225. Candidate           Moneta~ Indivi~
##  2 Ford, Rob                          78805. Candidate           Moneta~ Indivi~
##  3 Ford, Doug                         50000  Candidate           Moneta~ Indivi~
##  4 Ford, Rob                          50000  Candidate           Moneta~ Indivi~
##  5 Ford, Rob                          50000  Candidate           Moneta~ Indivi~
##  6 Goldkind, Ari                      23624. Candidate           Moneta~ Indivi~
##  7 Ford, Rob                          20000  Candidate           Moneta~ Indivi~
##  8 Ford, Rob                          12210  Candidate           Moneta~ Indivi~
##  9 Di Paola, Rocco                     6000  Candidate           Moneta~ Indivi~
## 10 Thomson, Sarah                      4426. Candidate           Moneta~ Indivi~
## # ... with abbreviated variable names 1: relationship_to_candidate,
## #   2: contribution_type_desc, 3: contributor_type_desc
```

```
outlier[11:20,]
```

```
## # A tibble: 10 x 5
##    contributors_name   contribution_amount_num relationship_to~1 contr~2 contr~3
##    <chr>                                 <dbl> <chr>             <chr>   <chr>
##  1 kindred's Muze                         3660 <NA>              Goods/~ Corpor~
##  2 Achber, Vernon                         2500 <NA>              Moneta~ Indivi~
##  3 Adam, Michael                          2500 <NA>              Moneta~ Indivi~
##  4 Aghaei, Saeid                          2500 <NA>              Moneta~ Indivi~
##  5 Al Zaibak, Mohammad                    2500 <NA>              Moneta~ Indivi~
##  6 Allan, David G. P.                     2500 <NA>              Moneta~ Indivi~
##  7 Allen, Peter A.                        2500 <NA>              Moneta~ Indivi~
##  8 Alper, Laura                           2500 <NA>              Moneta~ Indivi~
##  9 Alter, Robin                           2500 <NA>              Moneta~ Indivi~
## 10 Anderson, Jamie                        2500 <NA>              Moneta~ Indivi~
## # ... with abbreviated variable names 1: relationship_to_candidate,
## #   2: contribution_type_desc, 3: contributor_type_desc
```

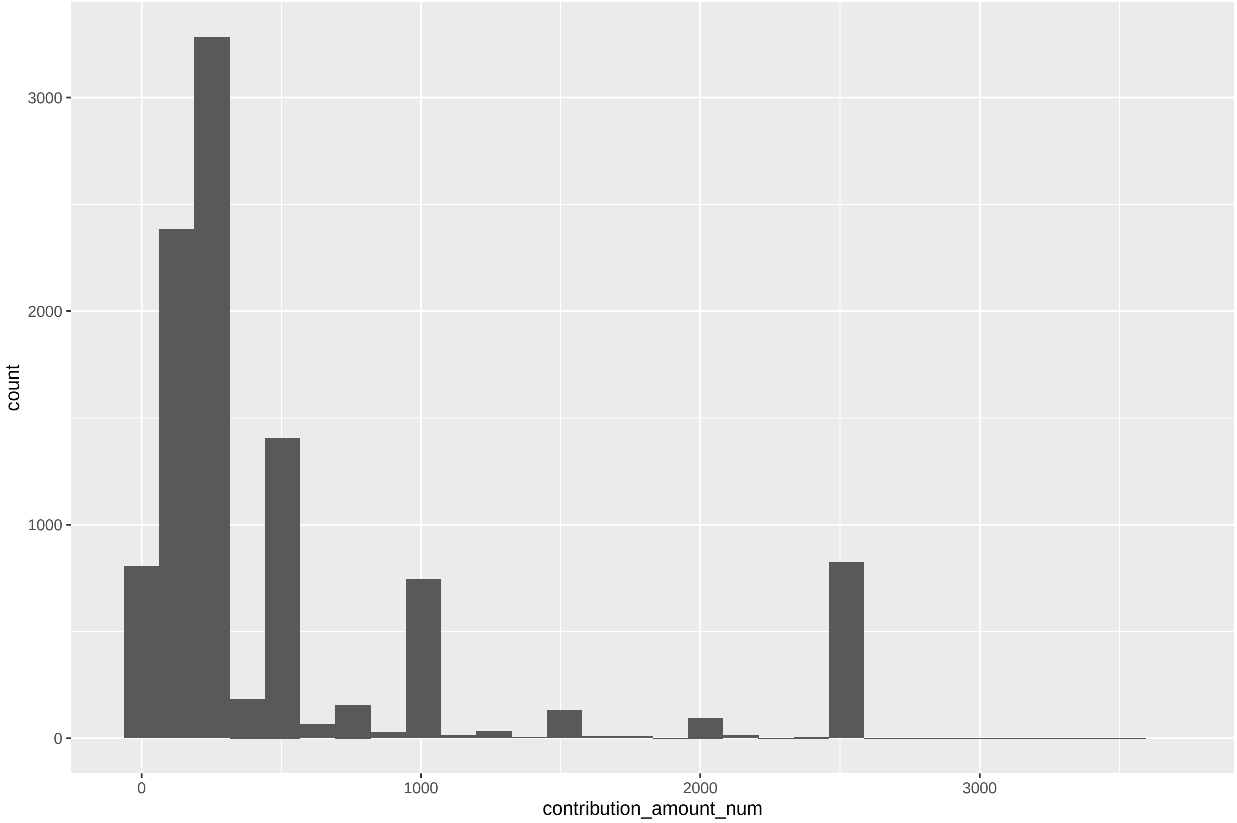
Based on the above table, we can see that there are notable outliers and those outlying contribution are all from the candidates' own contribution.

Here we show the contribution amount in a descending order, the largest contribution is from Ford, Doug, with the amount being 508225. This is a clear outlier as it is about 6 times the second large contribution. Now we look at the top 10 list and we notice that the contribution amount drops dramatically with the tenth largest amount being 4426. We also look at the 11th to 20th largest contribution amount and we find that it drops in a relatively steady way. Thus it is reasonable to conclude that there are outliers in this dataset and based on the scale of the data, it is more clear to see the distribution of the contribution amount without the ten largest contributions as they are so different from the rests.

Now we want to see what is common among those top ten contributions. First thing we noticed is that the contributions are all made by the candidates themselves as indicated by the variable "relationship_to_candidate". Another observation is that 7 out of top 10 contributions are made by "Ford, Doug" and "Ford, Rob" with "Ford, Doug" made the most largest contribution. Last but not least, all those 10 contributions are monetary and from individual.

Here is a distribution of contribution amount removing the ten largest observations in the table above

```
# The tenth largest amount in the data
tenth_largest <- outlier[10,"contribution_amount_num"]$contribution_amount_num

mayoral_campaign_2014 %>%
  filter(contribution_amount_num < tenth_largest) %>%
  ggplot() +
  geom_histogram(aes(x = contribution_amount_num))
```



The plot shows more information regarding the distribution of contribution amount as the outliers are removed. We can see that the majority of the contribution are still below or equal to 1000, with the next peak being 2500. Most people choose to contribute between 100 to 300.

## Q6

6. List the top five candidates in each of these categories:
    - total contributions
    - mean contribution
    - number of contributions

### Answer

```
# Top 5 Total Contributions
mayoral_campaign_2014 %>%
  group_by(candidate) %>%
  summarise(Total_Contributon = sum(contribution_amount_num)) %>%
  arrange(., -Total_Contributon) %>%
  slice(1:5)
```

```
## # A tibble: 5 x 2
##   candidate     Total_Contributon
##   <chr>                     <dbl>
## 1 Tory, John             2767869.
## 2 Chow, Olivia           1638266.
## 3 Ford, Doug              889897.
## 4 Ford, Rob               387648.
## 5 Stintz, Karen           242805
```

```
# Top 5 Mean Contributions
mayoral_campaign_2014 %>%
  group_by(candidate) %>%
  summarise(Mean_Contributon = mean(contribution_amount_num)) %>%
  arrange(., -Mean_Contributon) %>%
  slice(1:5)
```

```
## # A tibble: 5 x 2
##   candidate        Mean_Contributon
##   <chr>                       <dbl>
## 1 Sniedzins, Erwin            2025
## 2 Syed, Hïmy                  2018
## 3 Ritch, Carlie               1887.
## 4 Ford, Doug                  1456.
## 5 Clarke, Kevin               1200
```

```
# Top 5 Number of Contributions
mayoral_campaign_2014 %>%
  group_by(candidate) %>%
  summarise(Number_of_Contributon = n()) %>%
  arrange(., -Number_of_Contributon) %>%
  slice(1:5)
```

```
## # A tibble: 5 x 2
##   candidate       Number_of_Contributon
##   <chr>                           <int>
## 1 Chow, Olivia                     5708
## 2 Tory, John                       2602
## 3 Ford, Doug                        611
## 4 Ford, Rob                         538
## 5 Soknacki, David                   314
```

## Q7

7. Repeat 6 but without contributions from the candidates themselves.

**Answer**

```
# Remove the contribution from the candidate
mayoral_campaign_2014_remove <- filter(mayoral_campaign_2014, relationship_to_candidate %in% c("Spouse"
# Top 5 Total Contributions
mayoral_campaign_2014_remove %>%
  group_by(candidate) %>%
  summarise(Total_Contributon = sum(contribution_amount_num)) %>%
  arrange(., -Total_Contributon) %>%
  slice(1:5)
```

```
## # A tibble: 5 x 2
##   candidate     Total_Contributon
##   <chr>                     <dbl>
## 1 Tory, John             2765369.
## 2 Chow, Olivia           1635766.
## 3 Ford, Doug              331173.
## 4 Stintz, Karen           242805
## 5 Ford, Rob               174510.
```

```
# Top 5 Mean Contributions
mayoral_campaign_2014_remove %>%
  group_by(candidate) %>%
  summarise(Mean_Contributon = mean(contribution_amount_num)) %>%
  arrange(., -Mean_Contributon) %>%
  slice(1:5)
```

```
## # A tibble: 5 x 2
##   candidate         Mean_Contributon
##   <chr>                        <dbl>
## 1 Ritch, Carlie                1887.
## 2 Sniedzins, Erwin             1867.
## 3 Tory, John                   1063.
## 4 Gardner, Norman              1000
## 5 Tiwari, Ramnarine            1000
```

```
# Top 5 Number of Contributions
mayoral_campaign_2014_remove %>%
  group_by(candidate) %>%
  summarise(Number_of_Contributon = n()) %>%
  arrange(., -Number_of_Contributon) %>%
  slice(1:5)
```

```
## # A tibble: 5 x 2
##   candidate       Number_of_Contributon
##   <chr>                           <int>
## 1 Chow, Olivia                     5707
## 2 Tory, John                       2601
## 3 Ford, Doug                        608
## 4 Ford, Rob                         531
## 5 Soknacki, David                   314
```

We only keep the contribution record from people who doesn't have any relationship with the candidate and the spouse of the candidate (who is not the candidate himself or herself).

## Q8

8. How many contributors gave money to more than one candidate?

**Answer**

```
contributor_candidate <- mayoral_campaign_2014 %>%
  group_by(contributors_name, candidate) %>%
  summarise(Number_of_Contributon = n())

contributor <- contributor_candidate %>%
  group_by(contributors_name) %>%
  summarise(Number_of_Candidate = n()) %>%
  filter(Number_of_Candidate > 1)

nrow(contributor)
```

```
## [1] 184
```

There are 184 contributors that gave money to more than one candidate.