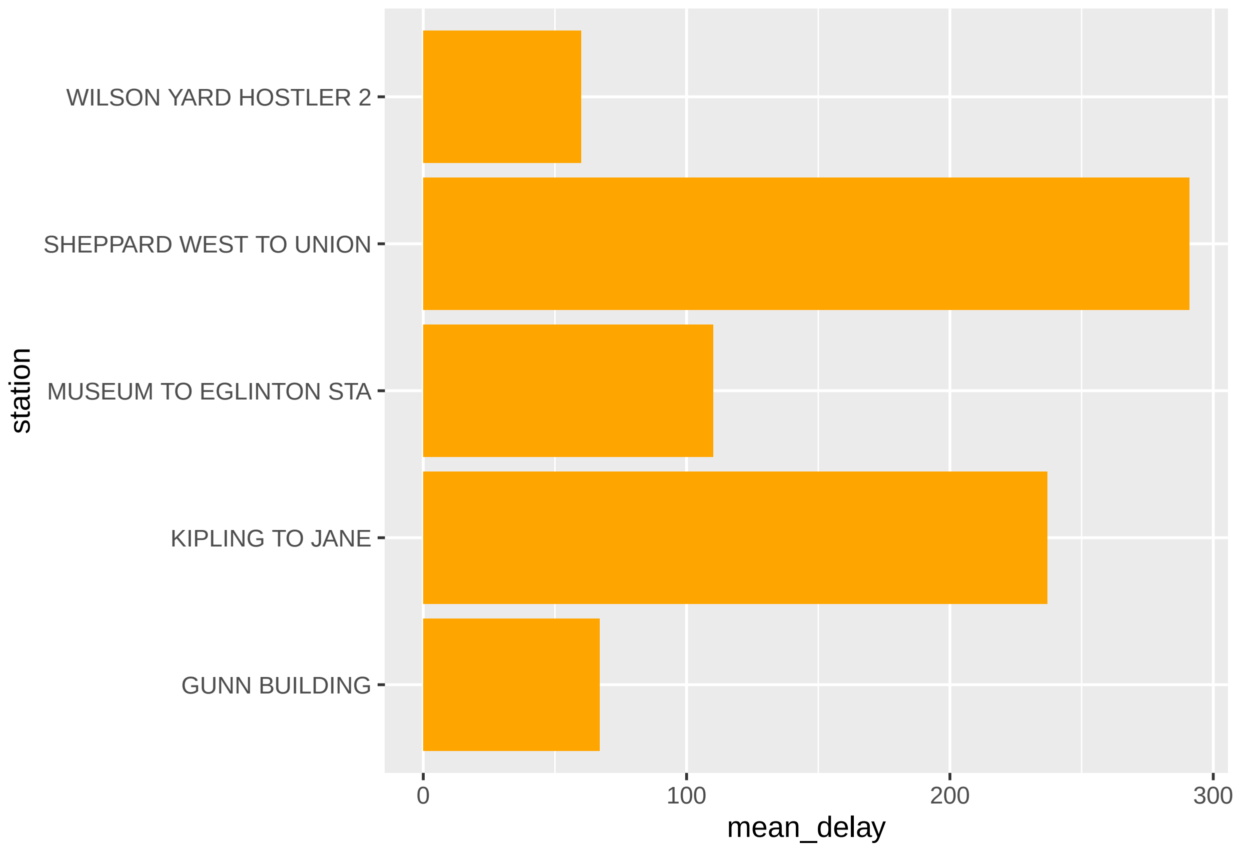
Lab Week 2

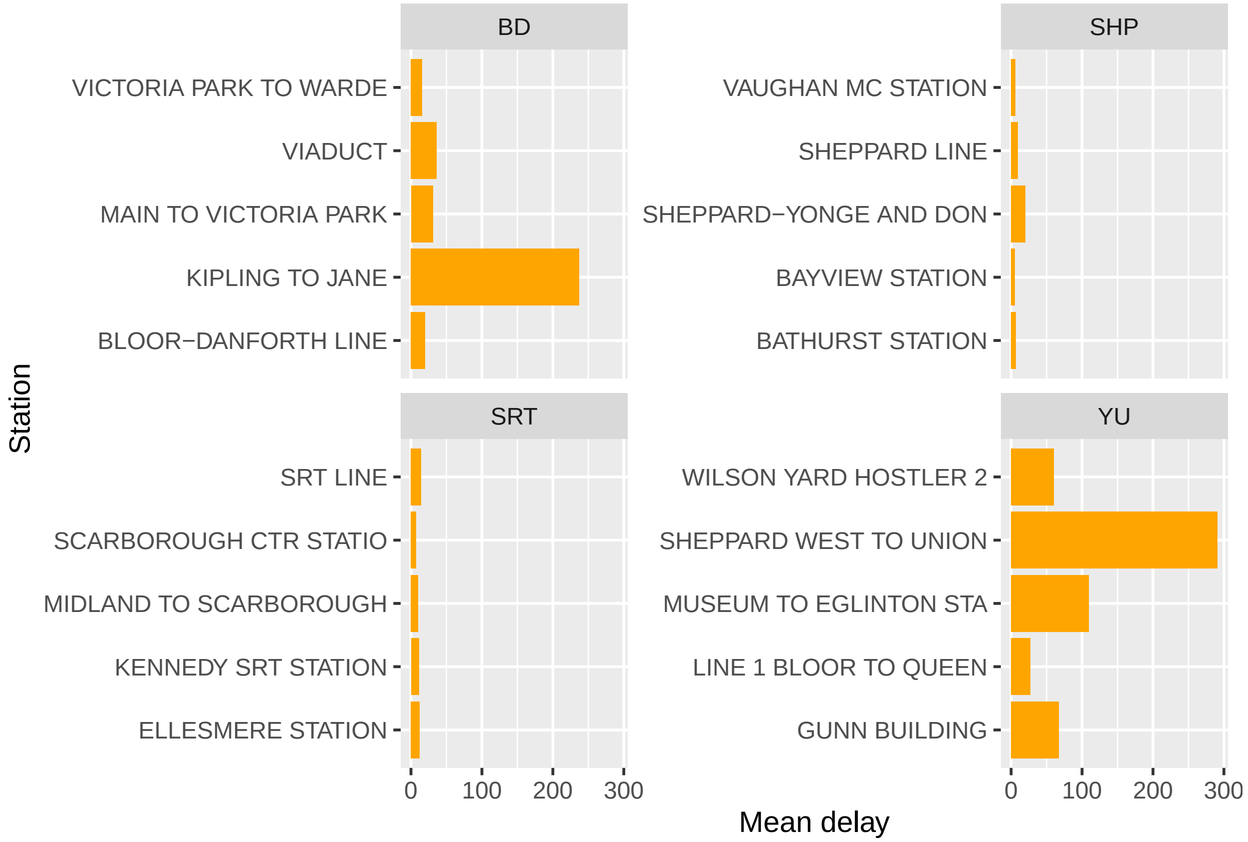
Rosa Fallahpour

- **(1)** The following plots represent the five stations with the highest mean delays and faceting by variable `line` ,respectively:

```
library(opendatatoronto)
library(tidyverse)
library(stringr)
library(skimr)
library(visdat)
library(janitor)
library(lubridate)
library(ggrepel)
res <- list_package_resources("996cfe8d-fb35-40ce-b569-698d51fc683b")
res <- res |> mutate(year = str_extract(name, "2022"))
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 |> filter(line %in% c("BD", "YU", "SHP", "SRT"))

delay_2022 |> group_by(station) |> summarise(mean_delay= mean(min_delay)) |> arrange(desc(mean_delay))
```


WILSON YARD HOSTLER 2
SHEPPARD WEST TO UNION
MUSEUM TO EGLINTON STA
KIPLING TO JANE
GUNN BUILDING
station
0
100
200
300
mean_delay



- **(2)** Downloading the data on mayoral campaign contributions for 2014:

```
all_data <- list_packages(limit = 500)
search_res <- all_data %>% filter(str_detect(title, fixed("campaign", ignore_case = T)))
res <- list_package_resources("f6651a40-2f52-46fc-9e04-b760c16edd5c")
may2014 <- get_resource("5b230e92-0a22-4a15-9572-0b19cc222985")$`2_Mayor_Contributions_2014_election.xls
head(may2014)
```

```
## # A tibble: 6 x 13
##   2014 Munic~1 ...2  ...3  ...4  ...5  ...6  ...7  ...8  ...9  ...10 ...11 ...12
##   <chr>        <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 Contributor~ Cont~ Cont~ Cont~ Cont~ Good~ Cont~ Rela~ Pres~ Auth~ Cand~ Offi~
## 2 A D'Angelo,~ <NA>  M6A ~ 300   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Ford~ Mayor
## 3 A Strazar, ~ <NA>  M2M ~ 300   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Ford~ Mayor
## 4 A'Court, K ~ <NA>  M4M ~ 36    Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Chow~ Mayor
## 5 A'Court, K ~ <NA>  M4M ~ 100   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Chow~ Mayor
## 6 A'Court, K ~ <NA>  M4M ~ 100   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Chow~ Mayor
## # ... with 1 more variable: ...13 <chr>, and abbreviated variable name
## #   1: '2014 Municipal Election - List of Contributors to Mayoralty Candidates'
```

- **(3)** Cleaning up the data format:

```
mayoral2014 <- may2014 %>% row_to_names(1) %>% clean_names()
head(mayoral2014)
```

```
## # 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
```

- **(4)** Below displays the table of variables summary. We have large numbers of missing values in some variables such as contributors_address, goods_or_service_desc, relationship_to_candidate, president_business_manager, authorized_representative and ward. Depending on our purpose of data exploration, we can perform an analysis which excludes these variables. Therefore, we should not be worried about them. The contribution_amount variable is in character format which we change it to numeric by creating new variable called "cont_amount".

```
skim(mayoral2014)
```

Table 1: Data summary

Name	mayoral2014
Number of rows	10199
Number of columns	13
Column type frequency:	
character	13
Group variables	None

Variable type: character

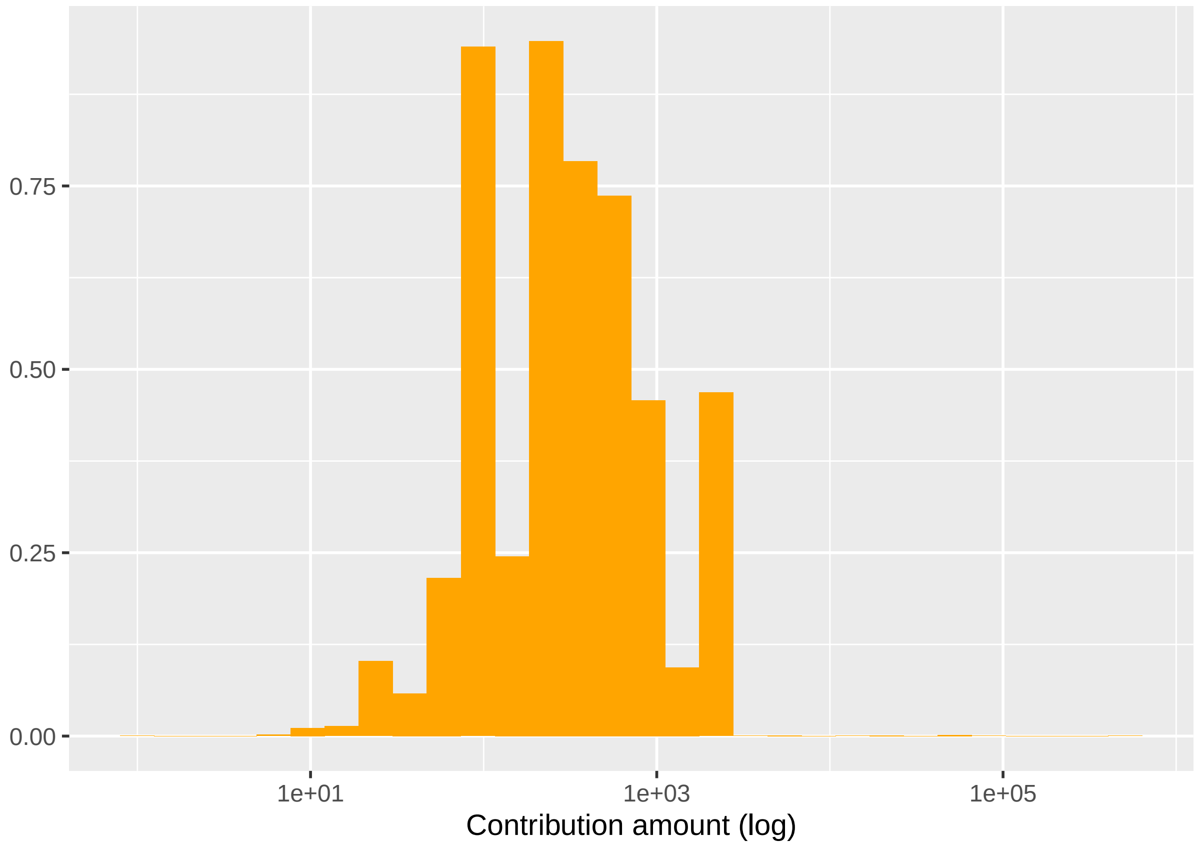
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
contributors_name	0	1	4	31	0	7545	0
contributors_address	10197	0	24	26	0	2	0
contributors_postal_code	0	1	7	7	0	5284	0
contribution_amount	0	1	1	18	0	209	0
contribution_type_desc	0	1	8	14	0	2	0
goods_or_service_desc	10188	0	11	40	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

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ward	10199	0	NA	NA	0	0	0

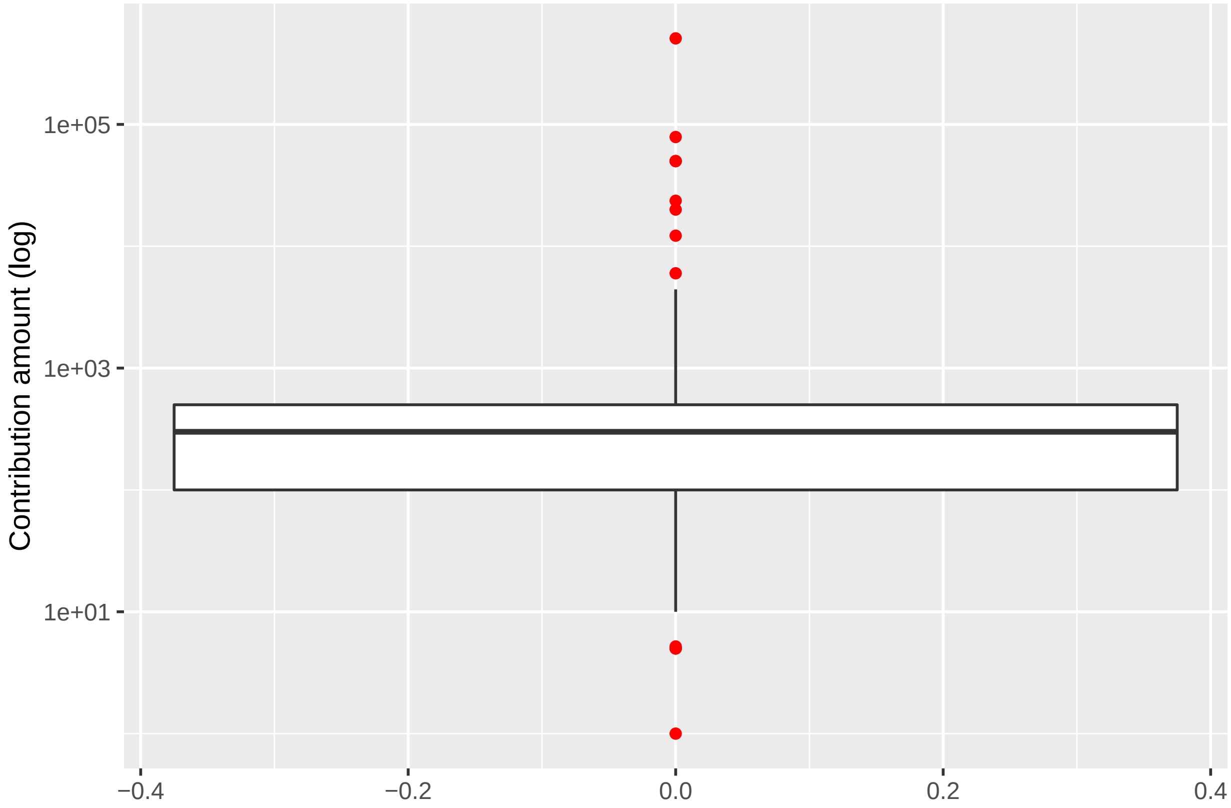
```
mayoral2014 <- mayoral2014 %>% mutate(cont_amount = as.numeric(contribution_amount))
```

- **(5)** The distribution of contribution amount in log scale is as below. We also created the boxplot to better realize the outliers. As we can see, contributions greater than 10000 (in log scale) are outliers. The similar characteristic in these outliers is that they have been contributed by candidates themselves as shown below.

```
#creating histogram for contribution amount in log
mayoral2014 %>% ggplot(aes(x=cont_amount, y=..density..)) +geom_histogram(position="dodge",fill="orange
```



```
#creating boxplot for contribution amount to better look for outliers
mayoral2014 %>% ggplot(aes(y = cont_amount)) +geom_boxplot(outlier.colour = "red")+labs(x="",y="Contribu
```

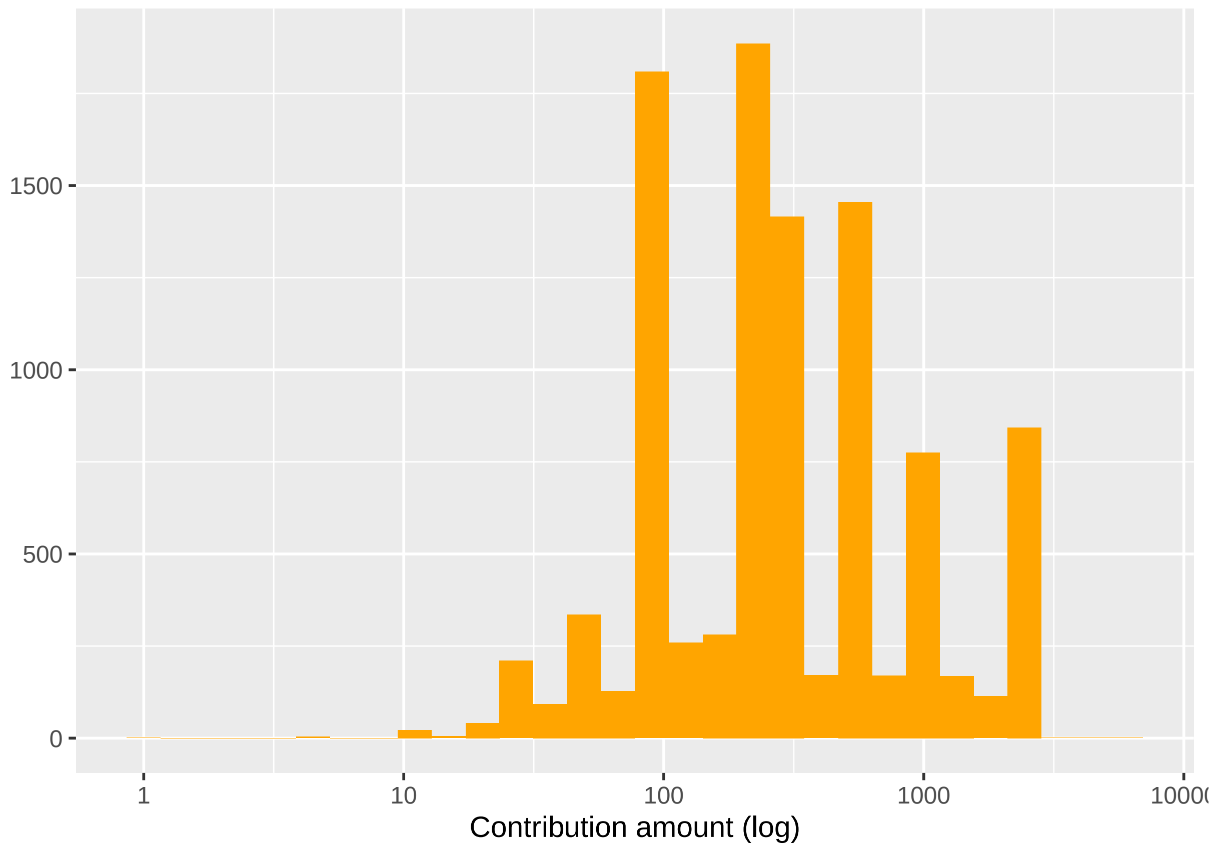


```
mayo1 <- mayoral2014 |> filter(cont_amount > 10000) |> select(cont_amount, relationship_to_candidate)
mayo1
```

```
## # A tibble: 8 x 2
##   cont_amount relationship_to_candidate
##         <dbl> <chr>
## 1     508225. Candidate
## 2      50000  Candidate
## 3      20000  Candidate
## 4      50000  Candidate
## 5      50000  Candidate
## 6      78805. Candidate
## 7      12210  Candidate
## 8      23624. Candidate
```

After removing the ouliers, we will have the following distribution for the contribution amounts:

```
mayo2 <- mayoral2014 |> filter(cont_amount <10000)
mayo2 |> ggplot(aes(x=cont_amount))+geom_histogram(fill="Orange")+labs(x="Contribution amount (log)",y=
```



- **(6)** Top five candidates in total contributions:

```
mayoral2014  |> group_by(candidate) |> summarise(total=sum(cont_amount)) |> arrange(desc(total)) |> sli
```

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

Top five candidates in mean contribution:

```
mayoral2014  |> group_by(candidate) |> summarise(mean=mean(cont_amount)) |> arrange(desc(mean)) |> slic
```

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

Top five candidates in number of contributions:

```
mayoral2014  |> group_by(candidate) |> summarise(cand_number=n()) |> arrange(-cand_number) |> slice(1:5
```

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

- **(7)** Removing contributions from the candidates themselves we will have the following results:

```
mayo2014_no_cand <- mayoral2014 |> filter(contributors_name!= candidate)
```

Top five candidates in total contributions:

```
mayo2014_no_cand  |> group_by(candidate) |> summarise(total=sum(cont_amount)) |> arrange(desc(total)) |
```

```
## # A tibble: 5 x 2
##   candidate        total
##   <chr>            <dbl>
## 1 Tory, John    2765369.
## 2 Chow, Olivia  1634766.
## 3 Ford, Doug     331173.
## 4 Stintz, Karen  242805
## 5 Ford, Rob      174510.
```

Top five candidates in mean contribution:

```
mayo2014_no_cand  |> group_by(candidate) |> summarise(mean=mean(cont_amount)) |> arrange(desc(mean)) |>
```

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

Top five candidates in number of contributions:

```
mayo2014_no_cand  |> group_by(candidate) |> summarise(cand_number=n()) |> arrange(-cand_number) |> slic
```

```
## # A tibble: 5 x 2
##   candidate       cand_number
##   <chr>                 <int>
## 1 Chow, Olivia           5706
## 2 Tory, John             2601
## 3 Ford, Doug              608
## 4 Ford, Rob               531
## 5 Soknacki, David         314
```

- **(8)** 184 contributors gave money to more than one candidate.

```
mayoral2014 |> group_by(contributors_name) |> distinct(contributors_name, candidate) |> summarise(num_ca
```

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