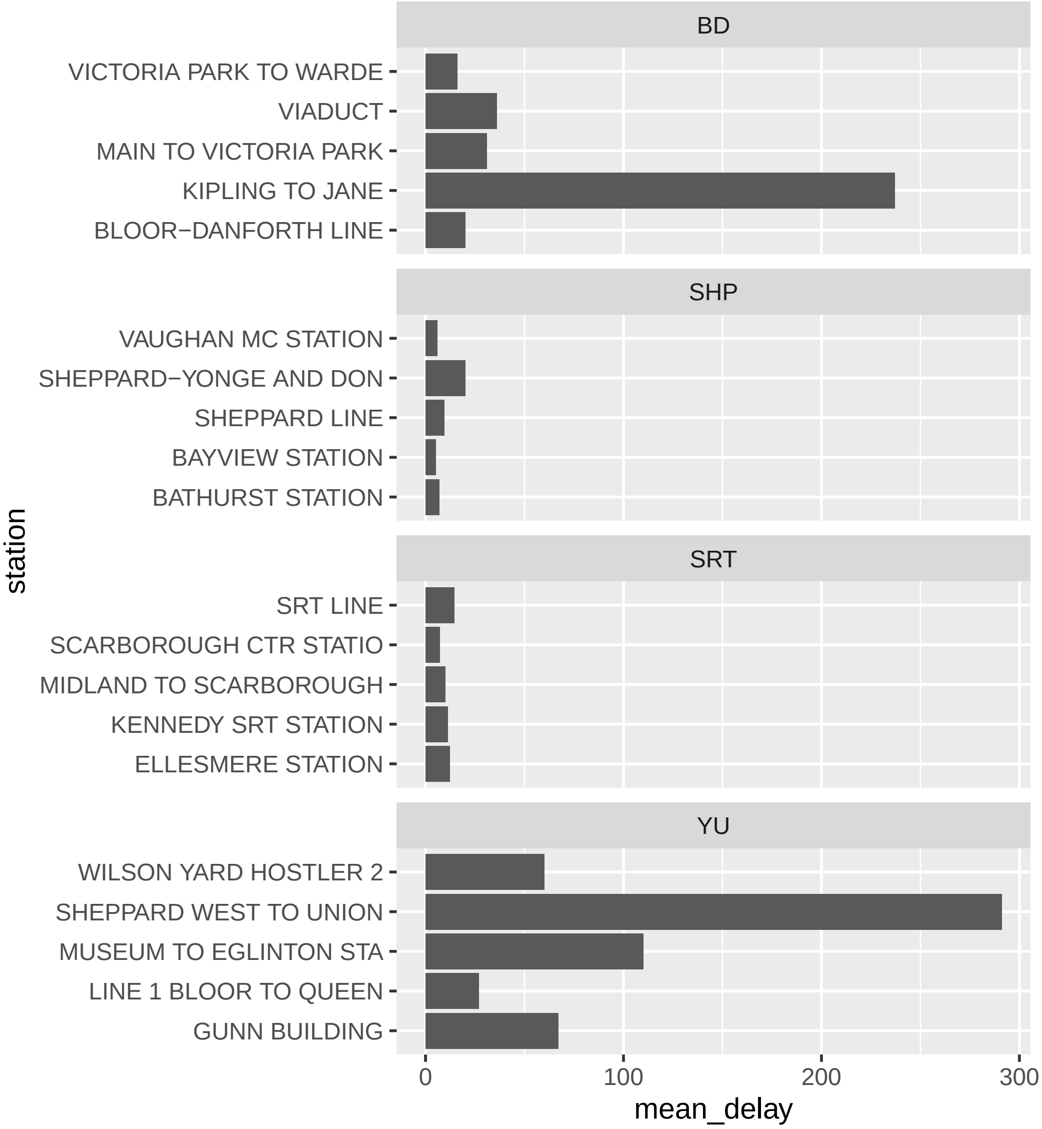
Lab 2

```r
library(opendatatoronto)
library(tidyverse)
library(stringr)
library(skimr) # EDA
library(visdat) # EDA
library(janitor)
library(lubridate)
library(ggrepel)
library(dplyr)
# obtained code from searching data frame above
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)
# make the column names nicer to work with
delay_2022 <- clean_names(delay_2022)
delay_codes <- get_resource("3900e649-f31e-4b79-9f20-4731bbfd94f7")
delay_data_codebook <- get_resource("ca43ac3d-3940-4315-889b-a9375e7b8aa4")
delay_2022=delay_2022 %>% distinct()
delay_2022 <- delay_2022 |> filter(line %in% c("BD", "YU", "SHP", "SRT"))
```

1.

```r
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()
```



2.

```r
all_data <- list_packages(limit = 500)
# obtained code from searching data frame above
res2=list_package_resources("f6651a40-2f52-46fc-9e04-b760c16edd5c")
# obtained this code from the 'id' column in the `res2` object above
mayo=get_resource("5b230e92-0a22-4a15-9572-0b19cc222985")
# just keep the data that relates to the Mayor election
mayo=mayo$"2_Mayor_Contributions_2014_election.xls"
```

3.

```r
library(janitor)
# fix 1st row of column names
mayo=mayo %>% row_to_names(row_number=1)
# clean up data format
mayo=clean_names(mayo)
```

4.

```r
# Summarize the variables in the dataset
skim(mayo)
```

Table 1: Data summary

Name	mayo
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

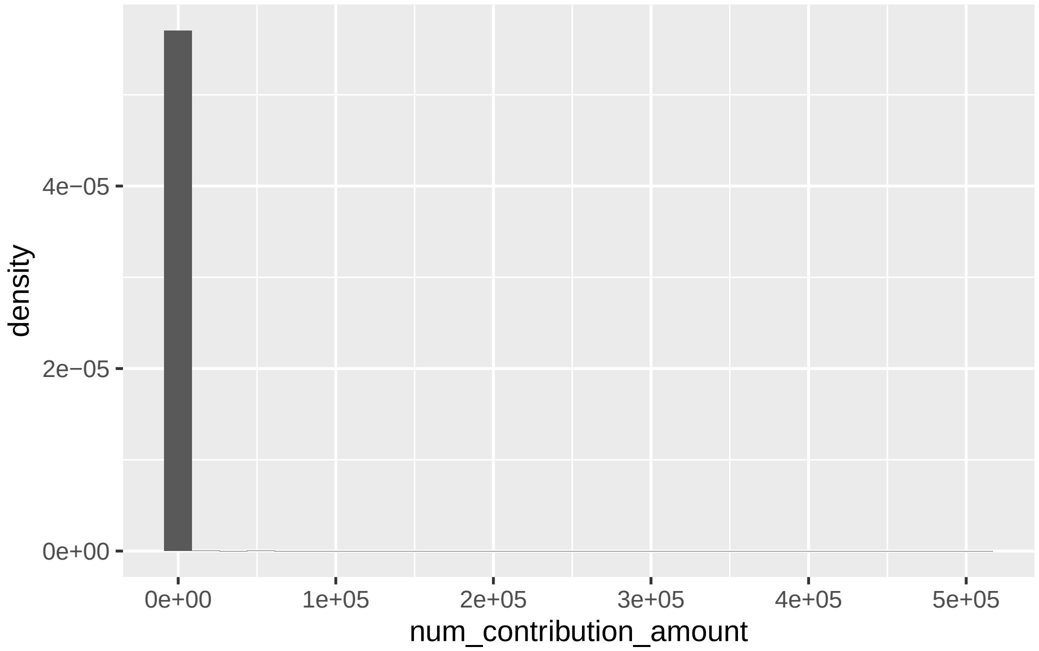
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
contributors_name	0	1	4	31	0	7545	0
contributors_address	10197	0	24	26	0	2	0
contributors_postal_code	0	1	7	7	0	5284	0
contribution_amount	0	1	1	18	0	209	0
contribution_type_desc	0	1	8	14	0	2	0
goods_or_service_desc	10188	0	11	40	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 numeric contribution amount variable
mayo=mayo %>% mutate(num_contribution_amount=as.numeric(contribution_amount))
```

There are missing values for the contributors_address, goods_or_service_desc, relationship_to_candidate, president_business_manager, authorized_representative and ward variables. However, we should not be worried about them since the majority of the values are missing for each of these variables, so for our purposes we can perform our analyses without these variables. The contribution_amount variable is in the character format instead of the numeric format, so we add the variable num_contribution_amount which is the contribution_amount variable in the numeric format.

5.

```
# histogram of contribution amounts
ggplot(data = mayo) +
  geom_histogram(aes(x = num_contribution_amount, y = ..density..),
                 position = 'dodge')
```



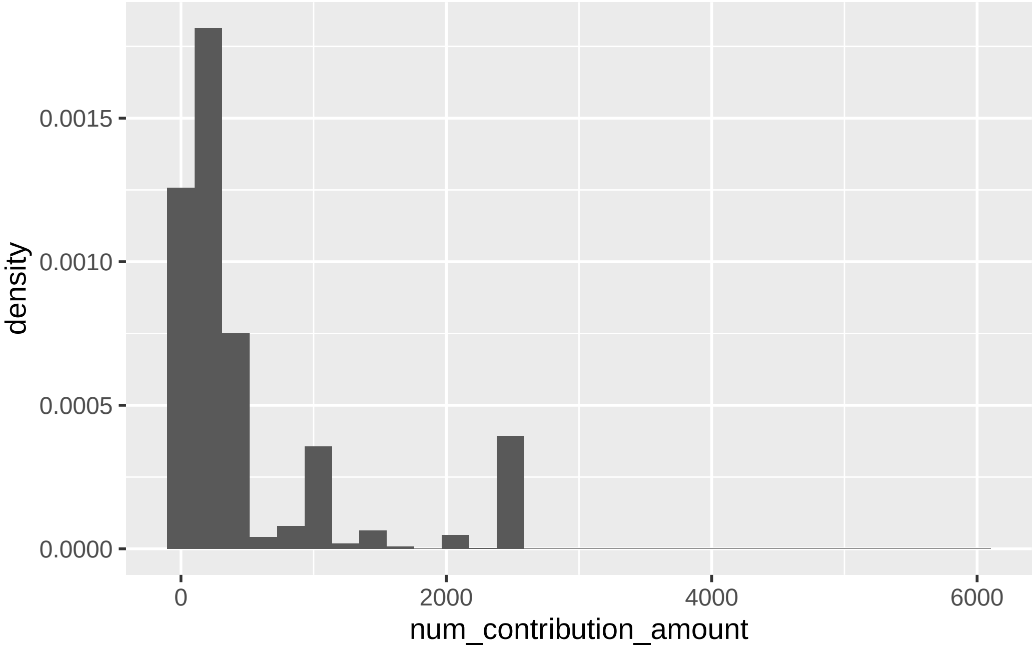
The plot above shows a histogram of the contribution amounts. From the plot, we can see that there are very large contributions that are outliers. We also show the contributors name, contribution type, contributor type, relationship to candidate, candidate and contribution amounts for the top 10 contribution amounts. All these contributions are monetary with an individual contributor type, and the contributions were all made by the candidates themselves.

```
mayo %>% arrange(-num_contribution_amount) %>% select(contributors_name,
                                                      contribution_type_desc,
                                                      contributor_type_desc,
                                                      relationship_to_candidate,
                                                      candidate,
                                                      num_contribution_amount)
```

```
# A tibble: 10,199 x 6
   contributors_name contribution_type_desc contributo~1 relat~2 candi~3 num_c~4
   <chr>             <chr>                  <chr>        <chr>   <chr>     <dbl>
 1 Ford, Doug        Monetary               Individual   Candid~ Ford, ~ 508225.
 2 Ford, Rob         Monetary               Individual   Candid~ Ford, ~  78805.
 3 Ford, Doug        Monetary               Individual   Candid~ Ford, ~  50000
 4 Ford, Rob         Monetary               Individual   Candid~ Ford, ~  50000
 5 Ford, Rob         Monetary               Individual   Candid~ Ford, ~  50000
 6 Goldkind, Ari     Monetary               Individual   Candid~ Goldki~  23624.
 7 Ford, Rob         Monetary                Individual   Candid~ Ford, ~  20000
 8 Ford, Rob         Monetary                Individual   Candid~ Ford, ~  12210
 9 Di Paola, Rocco   Monetary                Individual   Candid~ Di Pao~   6000
10 Thomson, Sarah    Monetary                Individual   Candid~ Thomso~   4426.
# ... with 10,189 more rows, and abbreviated variable names
#   1: contributor_type_desc, 2: relationship_to_candidate, 3: candidate,
#   4: num_contribution_amount
```

Below, we plot the histogram of the contribution amounts with the outliers removed (contributions over $6000).

```
mayo2=mayo %>% filter(num_contribution_amount<=6000)
ggplot(data = mayo2) +
  geom_histogram(aes(x = num_contribution_amount, y = ..density..),
                 position = 'dodge')
```



With the outliers removed, it is easier to see the distribution of the rest of the contributions.

6.

```
# top five candidates in total contributions
mayo %>% group_by(candidate) %>%
  summarize(total_contribution=sum(num_contribution_amount)) %>%
  arrange(-total_contribution) %>% slice(1:5)
```

```
# A tibble: 5 x 2
  candidate     total_contribution
  <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
mayo %>% group_by(candidate) %>%
  summarize(mean_contribution=mean(num_contribution_amount)) %>%
  arrange(-mean_contribution) %>% slice(1:5)
```

```
# A tibble: 5 x 2
  candidate        mean_contribution
  <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
mayo %>% group_by(candidate) %>%
  summarize(num_contribution=n()) %>%
  arrange(-num_contribution) %>% slice(1:5)
```

```
# A tibble: 5 x 2
  candidate       num_contribution
  <chr>                      <int>
```

```
1 Chow, Olivia                5708
2 Tory, John                  2602
3 Ford, Doug                   611
4 Ford, Rob                    538
5 Soknacki, David              314
```

7.

```
# remove contributions from candidates themselves
q7_mayo=mayo %>% filter(relationship_to_candidate =="Spouse"|
                           is.na(relationship_to_candidate))
# top five candidates in total contributions
q7_mayo %>% group_by(candidate) %>%
  summarize(total_contribution=sum(num_contribution_amount)) %>%
  arrange(-total_contribution) %>% slice(1:5)
```

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

```
# top five candidates in mean contribution
q7_mayo %>% group_by(candidate) %>%
  summarize(mean_contribution=mean(num_contribution_amount)) %>%
  arrange(-mean_contribution) %>% slice(1:5)
```

```
# A tibble: 5 x 2
  candidate         mean_contribution
  <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
q7_mayo %>% group_by(candidate) %>%
  summarize(num_contribution=n()) %>%
  arrange(-num_contribution) %>% slice(1:5)
```

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

8.

```
q8_mayo=mayo %>% group_by(contributors_name) %>%
  summarize(count=n_distinct(candidate)) %>% filter(count>1) %>%
  arrange(-count)
nrow(q8_mayo)
```

```
[1] 184
```

184 contributors gave money to more than one candidate.