lab2

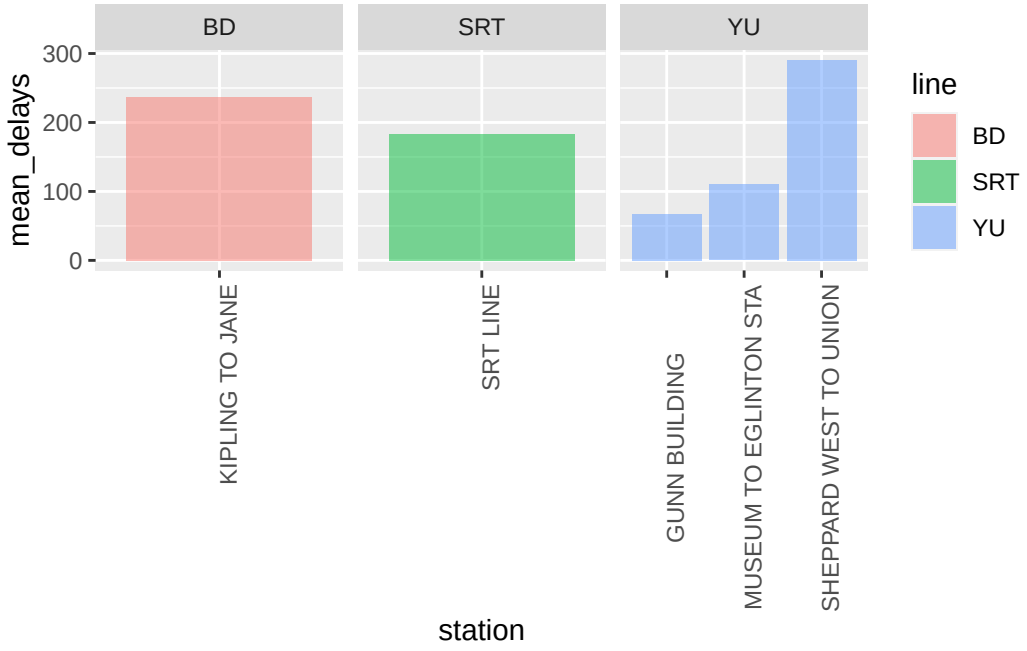
Lab 2

```
library(opendatatoronto)
library(tidyverse)
library(stringr)
library(skimr) # EDA
library(visdat) # EDA
library(janitor)
library(lubridate)
library(ggrepel)
```

1.

```
all_data <- list_packages(limit = 500)
#head(all_data)

res <- list_package_resources("996cfe8d-fb35-40ce-b569-698d51fc683b") # obtained code from
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_2022 %>%
  filter(min_delay>0) %>%
  group_by(station, line) %>%
```

```
summarise(mean_delays = mean(min_delay, na.rm =T)) %>%
distinct() %>% arrange(desc(mean_delays)) %>%
head(5) %>%
ggplot(aes(x = station, y = mean_delays)) +
geom_col(aes(fill = line), alpha = .5) +
facet_wrap(~line, scales = "free_x") +
theme(axis.text.x = element_text(angle = 90))
```



2.

```
all_data <- list_packages(limit = 500)
#head(all_data)

res2 <- list_package_resources("f6651a40-2f52-46fc-9e04-b760c16edd5c") # obtained code fro
campaign_2014 <- get_resource("5b230e92-0a22-4a15-9572-0b19cc222985")[[2]]
```

3.

```
colnames(campaign_2014) <- campaign_2014[1,]
campaign_2014 <- campaign_2014[-1,] %>% clean_names()
head(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
```

4.

There are several variables with a large number of missing values (e.g., contributors address, goods or service, relationship to candidate,...). While some of these variables may not be important (e.g., contributors address), other variables such as relationship to candidate could be necessary to answer research questions such as 'which factors drive contributions?'.

The contribution amount variable which was in character format was changed to numeric.

```
skim(campaign_2014)
```

Table 1: Data summary

Name	campaign_2014
Number of rows	10199
Number of columns	13
Column type frequency:	
character	13

Table 1: Data summary

Group variables	None

Variable type: character

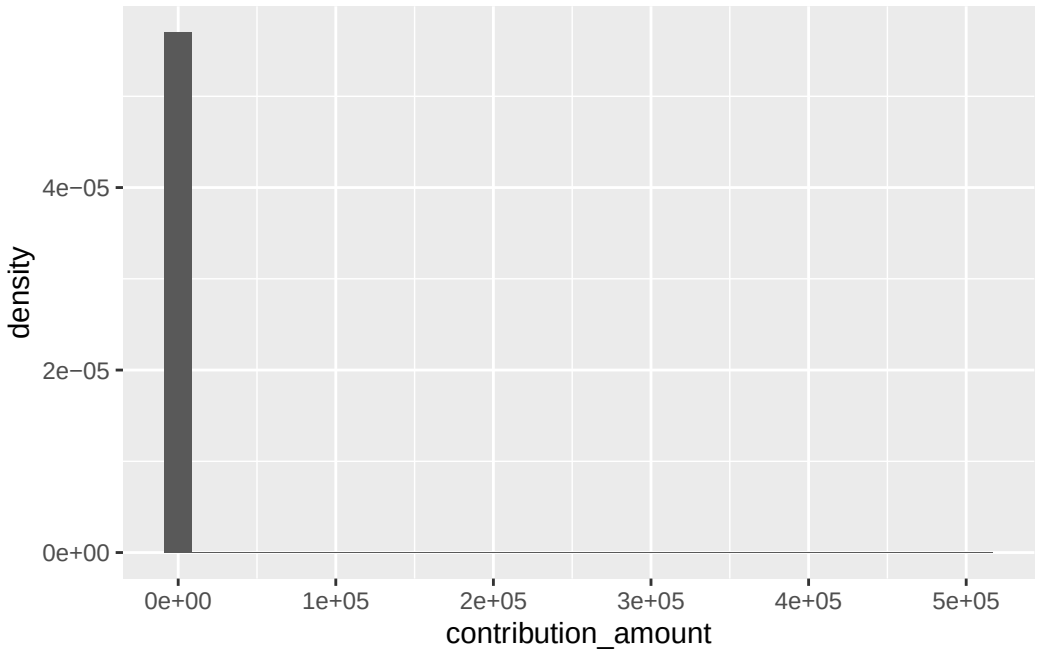
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
contributors_name	0	1	4	31	0	7545	0
contributors_address	10197	0	24	26	0	2	0
contributors_postal_code	0	1	7	7	0	5284	0
contribution_amount	0	1	1	18	0	209	0
contribution_type_desc	0	1	8	14	0	2	0
goods_or_service_desc	10188	0	11	40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

```
campaign_2014$contribution_amount <- as.numeric(campaign_2014$contribution_amount)
```

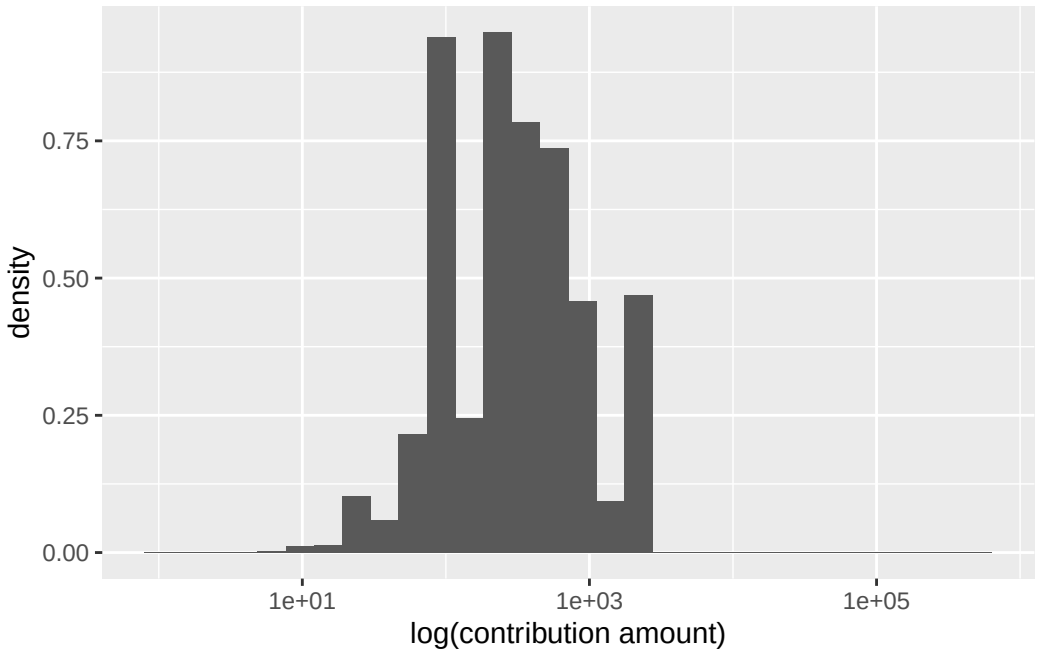
5.

There is one notable outlier with contribution amount 508224.73. However, contributions that exceed 10000 can also be considered as potential outliers. These outliers are mostly monetary contributions from Doug and Rob Ford for their election campaigns.

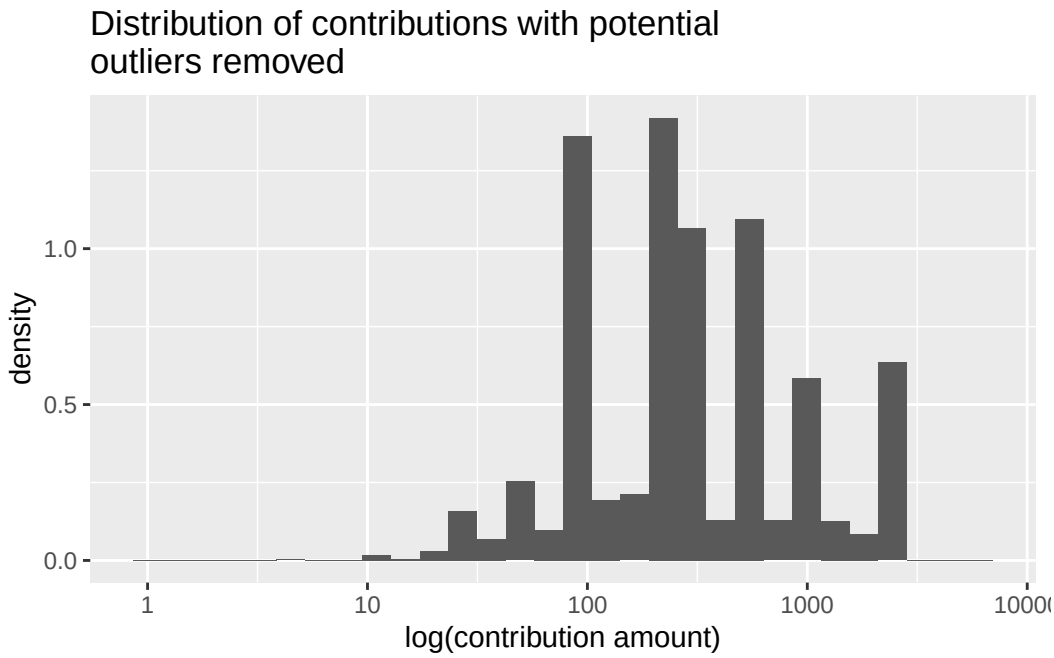
```
campaign_2014 %>%
  ggplot(aes(x = contribution_amount, y = ..density..)) +
  geom_histogram()
```




```
campaign_2014 %>%
  ggplot(aes(x = contribution_amount, y = ..density..)) +
  geom_histogram() + scale_x_log10() +
  labs(x= "log(contribution amount)")
```




```
campaign_2014 %>% filter(contribution_amount <= 10000) %>%
  ggplot(aes(x = contribution_amount, y = ..density..)) +
  geom_histogram() +scale_x_log10() +
  labs(title = "Distribution of contributions with potential \noutliers removed",
       x= "log(contribution amount)")
```



6.

- Highest mean contributions

```
df <- campaign_2014 %>% group_by(contributors_name) %>%
  summarise(ave_cont  = mean(contribution_amount),
            total_cont  = sum(contribution_amount),
            n_cont  = n())

df %>% slice_max(ave_cont, n = 5) %>%
  select(contributors_name)
```

```
# A tibble: 5 x 1
  contributors_name
  <chr>
1 Ford, Doug
2 Ford, Rob
3 Goldkind, Ari
4 Di Paola, Rocco
5 kindred's Muze
```

- Highest total contributions

```
df %>% slice_max(total_cont, n = 5)%>%
  select(contributors_name)
```

```
# A tibble: 5 x 1
  contributors_name
  <chr>
1 Ford, Doug
2 Ford, Rob
3 Goldkind, Ari
4 Thomson, Sarah
5 Pappalardo, Victor
```

- Highest number of contributions

```
df %>% slice_max(n_cont, n = 5)%>%
  select(contributors_name)
```

```
# A tibble: 6 x 1
  contributors_name
  <chr>
1 Italiano, Rob
2 Cranston, Jacqueline
3 Henery, Marjorie
4 Martin, Martha
5 Quin, Derek
6 Stewart, Carol
```

7.

- Highest mean contributions

```
df2 <- campaign_2014 %>% filter(relationship_to_candidate != "Candidate") %>%
  group_by(contributors_name) %>%
  summarise(ave_cont  = mean(contribution_amount),
            total_cont  = sum(contribution_amount),
            n_cont  = n())

df2 %>% slice_max(ave_cont, n = 5) %>%
  select(contributors_name)
```

```
# A tibble: 3 x 1
  contributors_name
  <chr>
1 Hackett, Barbara
2 Yan, Flora
3 Johnson, Suzanne
```

- Highest total contributions

```
df2 %>% slice_max(total_cont, n = 5)%>%
  select(contributors_name)
```

```
# A tibble: 3 x 1
  contributors_name
  <chr>
1 Hackett, Barbara
2 Yan, Flora
3 Johnson, Suzanne
```

- Highest number of contributions

```
df2 %>% slice_max(n_cont, n = 5)%>%
  select(contributors_name)
```

```
# A tibble: 3 x 1
  contributors_name
  <chr>
1 Hackett, Barbara
2 Johnson, Suzanne
3 Yan, Flora
```

8.

```
campaign_2014 %>% select(contributors_name, candidate) %>%
  distinct() %>% group_by(contributors_name) %>%
  summarise(ncand = n()) %>% filter(ncand >1) %>%
  nrow()
```

```
[1] 184
```