Lab Two Work(STA2201)

SHAOHAN CHANG

2023-01-21

library the package

```
library(opendatatoronto)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stringr)
library(skimr)
library(visdat)
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(lubridate)
```

```
## Loading required package: timechange
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(ggrepel)
```

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

```
## # A tibble: 6 x 11
##   title     id    topics civic~1 publi~2 excerpt datas~3 num_r~4 formats refre~5
##   <chr>     <chr> <chr>  <chr>   <chr>   <chr>   <chr>     <int> <chr>   <chr>
## 1 Polls co~ 7bce~ City ~ <NA>    City C~ Polls ~ Table          5 JSON,C~ Daily
## 2 Traffic ~ a330~ Trans~ <NA>    Transp~ This d~ Map           12 GPKG,S~ As ava~
## 3 Rain Gau~ f293~ Locat~ Climat~ Toront~ This d~ Docume~       11 ZIP,DO~ Monthly
## 4 Developm~ 0aa7~ <NA>   <NA>    City P~ This d~ Table          4 JSON,C~ Monthly
## 5 Web Anal~ 2303~ City ~ <NA>    Inform~ This d~ Docume~        4 XLS,ZIP Weekly
## 6 Daily Sh~ 21c8~ Commu~ Afford~ Shelte~ Daily ~ Table         12 JSON,C~ Daily
## # ... with 1 more variable: last_refreshed <date>, and abbreviated variable
## #   names 1: civic_issues, 2: publisher, 3: dataset_category, 4: num_resources,
## #   5: refresh_rate
```

```
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_2022
```

```
## # A tibble: 18,216 x 10
##    date                time  day      station  code  min_d~1 min_gap bound line
##    <dttm>              <chr> <chr>    <chr>    <chr>   <dbl>   <dbl> <chr> <chr>
##  1 2022-01-01 00:00:00 15:59 Saturday LAWRENC~ SRDP        0       0 N     SRT
##  2 2022-01-01 00:00:00 02:23 Saturday SPADINA~ MUIS        0       0 <NA>  BD
##  3 2022-01-01 00:00:00 22:00 Saturday KENNEDY~ MRO         0       0 <NA>  SRT
##  4 2022-01-01 00:00:00 02:28 Saturday VAUGHAN~ MUIS        0       0 <NA>  YU
##  5 2022-01-01 00:00:00 02:34 Saturday EGLINTO~ MUATC       0       0 S     YU
##  6 2022-01-01 00:00:00 05:40 Saturday QUEEN S~ MUNCA       0       0 <NA>  YU
##  7 2022-01-01 00:00:00 06:56 Saturday DAVISVI~ MUNCA       0       0 <NA>  YU
##  8 2022-01-01 00:00:00 06:58 Saturday ST PATR~ MUNCA       0       0 <NA>  YU
##  9 2022-01-01 00:00:00 07:01 Saturday PAPE ST~ MUNCA       0       0 <NA>  BD
## 10 2022-01-01 00:00:00 07:43 Saturday WILSON ~ TUATC      10       0 S     YU
## # ... with 18,206 more rows, 1 more variable: vehicle <dbl>, and abbreviated
## #   variable name 1: min_delay
```

```
delay_codes <- get_resource("3900e649-f31e-4b79-9f20-4731bbfd94f7")
```

```
## New names:
## * `` -> `...1`
## * `CODE DESCRIPTION` -> `CODE DESCRIPTION...3`
## * `` -> `...4`
## * `` -> `...5`
## * `CODE DESCRIPTION` -> `CODE DESCRIPTION...7`
```

```
delay_data_codebook <- get_resource("ca43ac3d-3940-4315-889b-a9375e7b8aa4")
```

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

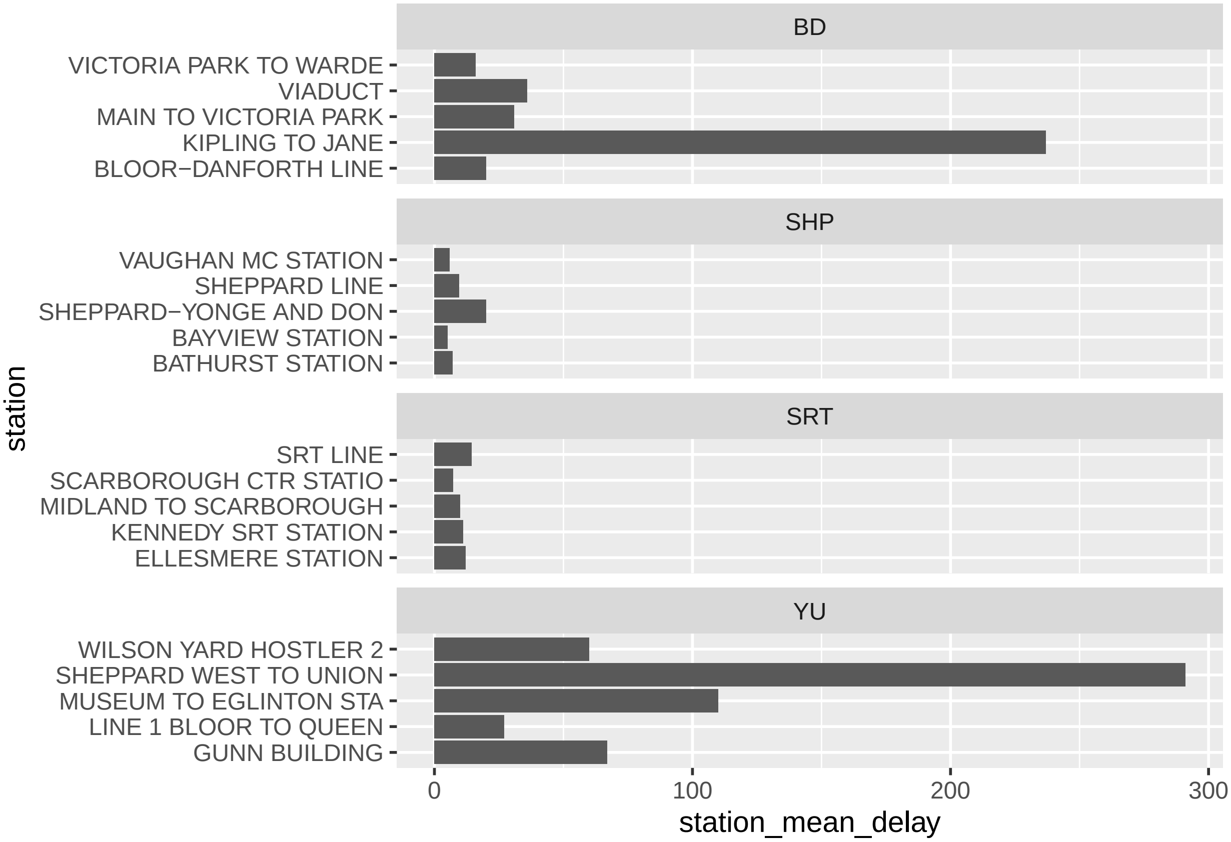
```
## # A tibble: 442 x 11
##    title    id    topics civic~1 publi~2 excerpt datas~3 num_r~4 formats refre~5
##    <chr>    <chr> <chr>  <chr>   <chr>   <chr>   <chr>     <int> <chr>   <chr>
##  1 Polls c~ 7bce~ City ~ <NA>    City C~ Polls ~ Table         5 JSON,C~ Daily
##  2 Traffic~ a330~ Trans~ <NA>    Transp~ This d~ Map          12 GPKG,S~ As ava~
##  3 Rain Ga~ f293~ Locat~ Climat~ Toront~ This d~ Docume~      11 ZIP,DO~ Monthly
##  4 Develop~ 0aa7~ <NA>   <NA>    City P~ This d~ Table         4 JSON,C~ Monthly
##  5 Web Ana~ 2303~ City ~ <NA>    Inform~ This d~ Docume~       4 XLS,ZIP Weekly
##  6 Daily S~ 21c8~ Commu~ Afford~ Shelte~ Daily ~ Table        12 JSON,C~ Daily
##  7 Members~ 7f52~ City ~ <NA>    City C~ Access~ Table        21 JSON,C~ As ava~
##  8 Members~ 9426~ City ~ <NA>    City C~ Access~ Table        21 JSON,T~ As ava~
##  9 City Co~ 3bfa~ City ~ <NA>    City C~ This d~ Table        21 JSON,C~ As ava~
## 10 Registr~ 3538~ City ~ <NA>    City C~ Effect~ Table        21 JSON,C~ As ava~
## # ... with 432 more rows, 1 more variable: last_refreshed <date>, and
## #   abbreviated variable names 1: civic_issues, 2: publisher,
## #   3: dataset_category, 4: num_resources, 5: refresh_rate
```

Q1. Using the `delay_2022` data, plot the five stations with the highest mean delays. Facet the graph by `line`.

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

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

```
## `summarise()` has grouped output by 'line'. You can override using the
## `.groups` argument.
```



Q2. 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

```
all_data
```

```
## # A tibble: 442 x 11
##    title    id    topics civic~1 publi~2 excerpt datas~3 num_r~4 formats refre~5
##    <chr>    <chr> <chr>  <chr>   <chr>   <chr>   <chr>     <int> <chr>   <chr>
##  1 Polls c~ 7bce~ City ~ <NA>    City C~ Polls ~ Table         5 JSON,C~ Daily
##  2 Traffic~ a330~ Trans~ <NA>    Transp~ This d~ Map          12 GPKG,S~ As ava~
##  3 Rain Ga~ f293~ Locat~ Climat~ Toront~ This d~ Docume~      11 ZIP,DO~ Monthly
##  4 Develop~ 0aa7~ <NA>   <NA>    City P~ This d~ Table         4 JSON,C~ Monthly
##  5 Web Ana~ 2303~ City ~ <NA>    Inform~ This d~ Docume~       4 XLS,ZIP Weekly
##  6 Daily S~ 21c8~ Commu~ Afford~ Shelte~ Daily ~ Table        12 JSON,C~ Daily
##  7 Members~ 7f52~ City ~ <NA>    City C~ Access~ Table        21 JSON,C~ As ava~
##  8 Members~ 9426~ City ~ <NA>    City C~ Access~ Table        21 JSON,T~ As ava~
##  9 City Co~ 3bfa~ City ~ <NA>    City C~ This d~ Table        21 JSON,C~ As ava~
## 10 Registr~ 3538~ City ~ <NA>    City C~ Effect~ Table        21 JSON,C~ As ava~
## # ... with 432 more rows, 1 more variable: last_refreshed <date>, and
## #   abbreviated variable names 1: civic_issues, 2: publisher,
## #   3: dataset_category, 4: num_resources, 5: refresh_rate
```

```
all_data %>% filter(str_detect(title, "Campaign"))
```

```
## # A tibble: 5 x 11
##   title     id    topics civic~1 publi~2 excerpt datas~3 num_r~4 formats refre~5
##   <chr>     <chr> <chr>  <chr>   <chr>   <chr>   <chr>     <int> <chr>   <chr>
## 1 Civic Is~ 7d0d~ City ~ Afford~ Inform~ "The O~ Table         5 XML,JS~ As ava~
## 2 Election~ 67d2~ Finan~ <NA>    City C~ "This ~ Docume~       2 ZIP,XL~ As ava~
## 3 Election~ f665~ City ~ <NA>    City C~ "This ~ Docume~       2 ZIP,XLS As ava~
## 4 Election~ 28e5~ City ~ <NA>    City C~ "This ~ Docume~       2 ZIP,XLS As ava~
## 5 Election~ 2ee8~ City ~ <NA>    City C~ "This ~ Docume~       2 ZIP,XLS As ava~
## # ... with 1 more variable: last_refreshed <date>, and abbreviated variable
## #   names 1: civic_issues, 2: publisher, 3: dataset_category, 4: num_resources,
## #   5: refresh_rate
```

```
res <- list_package_resources("f6651a40-2f52-46fc-9e04-b760c16edd5c")
res <- res %>% mutate(year = str_extract(name, "2014-data?"))
campaign_2014_id <- res %>% filter(year=='2014-data') %>% select(id) %>% pull()
campaign_2014 <-get_resource(campaign_2014_id)
```

```
## New names:
## New names:
## New names:
## New names:
## New names:
## New names:
## New names:
## * `` -> `...2`
## * `` -> `...3`
```

```
Mayor_data=campaign_2014$`2_Mayor_Contributions_2014_election.xls`

Mayor_data
```

```
## # A tibble: 10,200 x 13
##    2014 Muni~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 Contributo~ 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
##  7 Aaron, Rob~ <NA>  M6B ~ 250   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Tory~ Mayor
##  8 Abadi, Bab~ <NA>  M5S ~ 500   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Tory~ Mayor
##  9 Abadi, Bab~ <NA>  M5S ~ 500   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Chow~ Mayor
## 10 Abadi, Dav~ <NA>  M5S ~ 300   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Stin~ Mayor
## # ... with 10,190 more rows, 1 more variable: ...13 <chr>, and abbreviated
## #   variable name
## #   1: `2014 Municipal Election - List of Contributors to Mayoralty Candidates`
```

Q3. Clean up the data format (fixing the parsing issue and standardizing the column names using `janitor`).

```
Mayor_data<- Mayor_data %>%
  row_to_names(row_number = 1) %>%
  clean_names()
```

Q4. 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.

```
skim(Mayor_data)
```

Table 1: Data summary

Name	Mayor_data
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

```
Mayor_data %>%
  summarize(across(everything(), ~ sum(is.na(.x))))
```

```
## # A tibble: 1 x 13
##   contributors~1 contr~2 contr~3 contr~4 contr~5 goods~6 contr~7 relat~8 presi~9
##            <int>   <int>   <int>   <int>   <int>   <int>   <int>   <int>   <int>
## 1              0   10197      0       0       0   10188       0   10166   10197
## # ... with 4 more variables: authorized_representative <int>, candidate <int>,
## #   office <int>, ward <int>, 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
```

Explanation(Question 4):

Missing values exist in “contributors_address”, “goods_or_service_desc”, “relationship_to_candidate”, “president_business_manager” and “ward”.

We should not be worried about them.

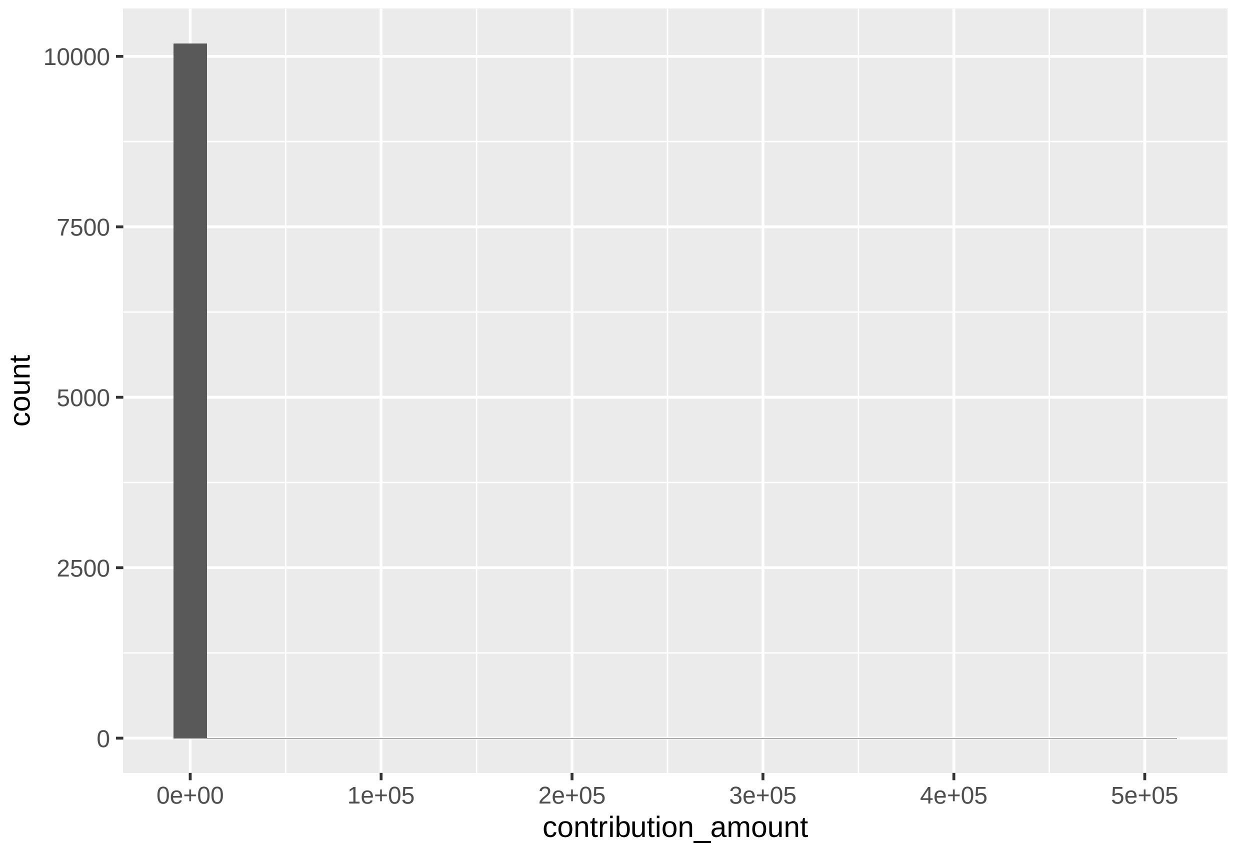
The “contribution_amount” should be numeric, instead of character type.

```
Mayor_data$contribution_amount=as.numeric(Mayor_data$contribution_amount)
```

Q5. 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.

```
ggplot(data = Mayor_data) +
  geom_histogram(aes(x = contribution_amount))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
outlier=which(Mayor_data$contribution_amount>=4000)
outlier
```

```
## [1] 2402 3013 3014 3022 3023 3024 3025 3026 3444 9251
```

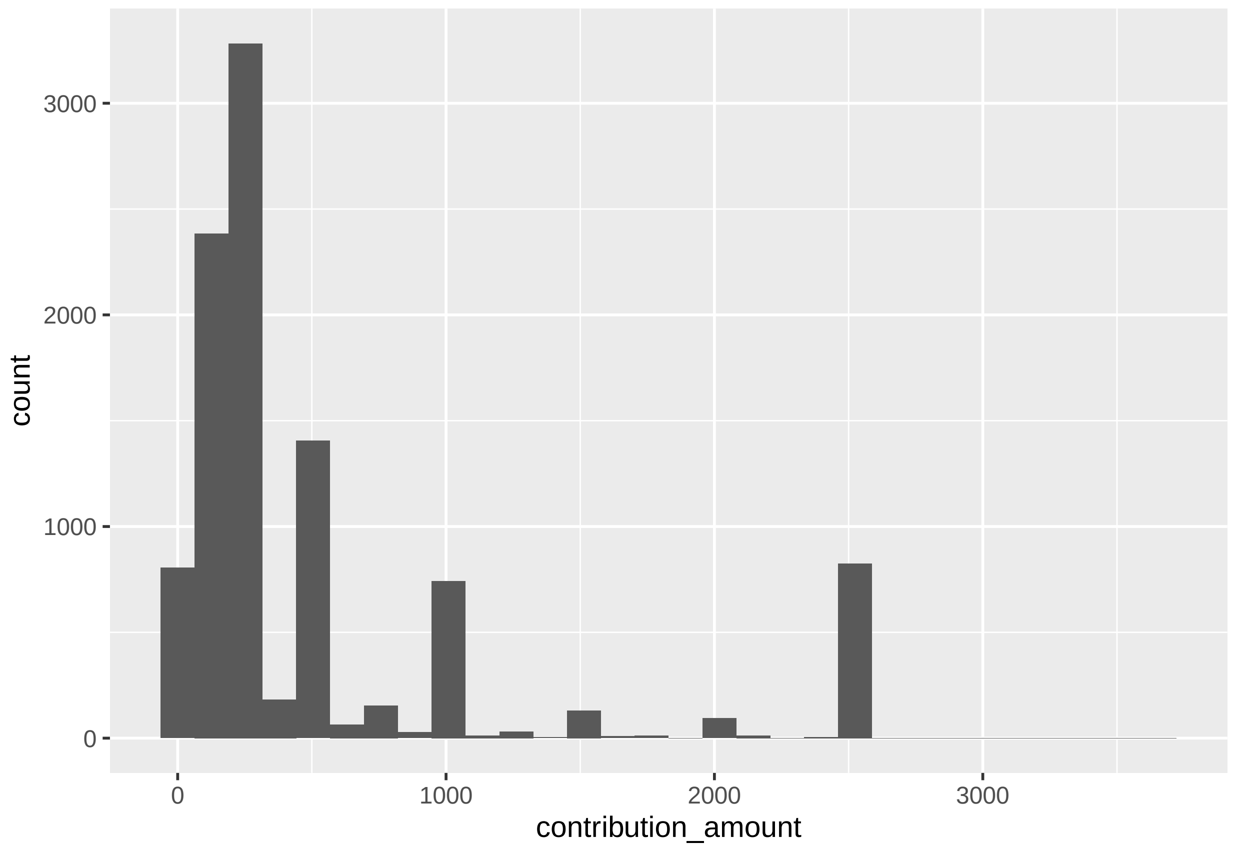
```
Mayor_data$contribution_amount[outlier]
```

```
##  [1]   6000.00 508224.73  50000.00  20000.00  50000.00  50000.00  78804.80
##  [8]  12210.00  23623.63   4425.55
```

Explanation(Question 5): The 2402 th, 3013 th, 3014th , 3022th , 3023th , 3024th , 3025th , 3026th, 3444th , 9251th contributions seem to be notable outliers.They share a similar characteristic. that is, their contribution_amount are more than 4000.The following graph plots the distribution of contributions without these outliers, from which we could get a better sense of the majority of the data.

```
ggplot(data = Mayor_data[-outlier,]) +
  geom_histogram(aes(x = contribution_amount))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Q6. List the top five candidates in each of these categories:

```
+ total contributions
+ mean contribution
+ number of contributions
```

```
data1=Mayor_data %>% group_by(candidate)%>%
  summarise(total_con=sum(contribution_amount))%>%arrange(-total_con)

head(data1,5)
```

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

```
data2=Mayor_data %>% group_by(candidate)%>%
  summarise(mean_con=mean(contribution_amount))%>%arrange(-mean_con)

head(data2,5)
```

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

```
data3=Mayor_data %>% group_by(candidate)%>%
  summarise(num_con=length(contribution_amount))%>%arrange(-num_con)

head(data3,5)
```

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

Q7. Repeat 5 but without contributions from the candidates themselves.

```
data = Mayor_data[-which(Mayor_data$relationship_to_candidate=="Candidate"),]

data4 <- data %>% group_by(candidate)%>%
  summarise(total_con=sum(contribution_amount))%>%arrange(-total_con)

head(data4,5)
```

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

```
data5 <- data  %>% group_by(candidate)%>%
  summarise(mean_con=mean(contribution_amount))%>%arrange(-mean_con)
head(data5,5)
```

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

```
data6 <- data %>% group_by(candidate)%>%
  summarise(num_con=length(contribution_amount))%>%arrange(-num_con)
```

```
head(data6,5)
```

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

Explanation(Question 7): Without the contributions from the candidates themselves, there are not notable outliers.

Q8. How many contributors gave money to more than one candidate?

```
Mayor_data %>%
  group_by(contributors_name) %>%
  summarize(n_candidates = n_distinct(candidate)) %>%
  filter(n_candidates > 1) %>%
  nrow()
```

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

Explanation(Question 8): There are 184 contributors who gave money to more than one candidate.