HW2

Chenxi Liu 1010615050

2024-01-19

Lab Exercise 2

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

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

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

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

```
library(lubridate)
library(ggrepel)
```

```
res <- list_package_resources("996cfe8d-fb35-40ce-b569-698d51fc683b") # obtained code from searching da
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")
```

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

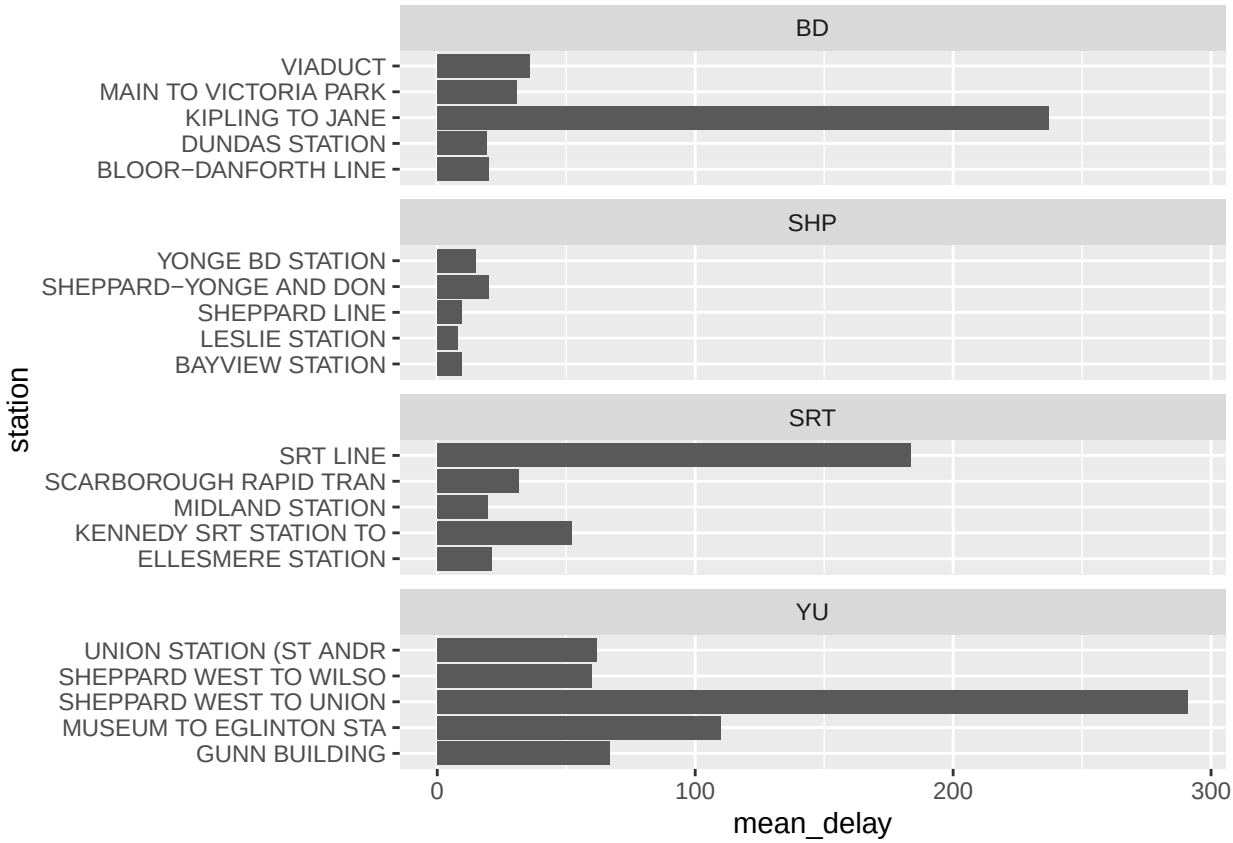
```
delay_data_codebook <- get_resource("ca43ac3d-3940-4315-889b-a9375e7b8aa4")

head(delay_2022)
```

```
## # A tibble: 6 x 10
##   date                time  day      station code  min_delay min_gap bound line
##   <dttm>              <chr> <chr>    <chr>   <chr>     <dbl>   <dbl> <chr> <chr>
## 1 2022-01-01 00:00:00 15:59 Saturday LAWREN~ SRDP          0       0 N     SRT
## 2 2022-01-01 00:00:00 02:23 Saturday SPADIN~ MUIS          0       0 <NA>  BD
## 3 2022-01-01 00:00:00 22:00 Saturday KENNED~ MRO           0       0 <NA>  SRT
## 4 2022-01-01 00:00:00 02:28 Saturday VAUGHA~ MUIS          0       0 <NA>  YU
## 5 2022-01-01 00:00:00 02:34 Saturday EGLINT~ MUATC         0       0 S     YU
## 6 2022-01-01 00:00:00 05:40 Saturday QUEEN ~ MUNCA         0       0 <NA>  YU
## # i 1 more variable: vehicle <dbl>
```

Question1

```
delay_2022 <- delay_2022 |> distinct()

## Removing the observations that have non-standardized lines
delay_2022 <- delay_2022 |> filter(line %in% c("BD", "YU", "SHP", "SRT"))

delay_2022 %>%
  filter(min_delay>0) %>%
  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()
```

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



Question2

```
delay_2022 <- delay_2022 |>
  left_join(delay_codes |> rename(code = `SUB RMENU CODE`, code_desc = `CODE DESCRIPTION...3`) |> selec
```

```
## Joining with 'by = join_by(code)'
```

```
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)
```

```
## Joining with 'by = join_by(code_srt)'
```

```
# Calculate the most frequent delay reasons
top_delay_reasons <- delay_2022 %>%
  filter((min_delay > 0) & !is.na(code_desc)) %>%
  count(code_desc, sort = TRUE) %>%
  filter(row_number() <= n() / 2) %>%
  pull(code_desc)
```

```r
# Filter delay_2022 based on conditions
filtered_delay_2022 <- delay_2022 %>%
  filter(min_delay > 0, code_desc %in% top_delay_reasons)

library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
#Apply the negative binomial regression model
model_nb <- glm.nb(min_delay ~ line + code_desc, data = filtered_delay_2022)

# Summarize the model
summary(model_nb)
```

```
##
## Call:
## glm.nb(formula = min_delay ~ line + code_desc, data = filtered_delay_2022,
##     init.theta = 3.597029996, link = log)
##
## Coefficients:
##                                                                  Estimate
## (Intercept)                                                      2.740237
## lineSHP                                                          0.113656
## lineSRT                                                          0.343852
## lineYU                                                           0.001405
## code_descAssault / Patron Involved                              -0.336258
## code_descATC Operator Related                                   -1.063996
## code_descATC Project                                            -1.046118
## code_descBody                                                   -1.366932
## code_descBomb Threat                                             1.107634
## code_descBrakes                                                 -0.852931
## code_descConsequential Delay (2nd Delay Same Fault)             -1.161683
## code_descCrew Unable to Maintain Schedule                       -1.085225
## code_descDebris At Track Level - Uncontrollable                 -0.089232
## code_descDisorderly Patron                                      -0.852202
## code_descDivisional Clerk Related                               -1.072479
## code_descDoor Problems - Debris Related                         -1.069239
## code_descDoor Problems - Faulty Equipment                       -1.103473
## code_descDoor Problems - Passenger Related                      -1.136321
## code_descEmergency Alarm Station Activation                     -0.266307
## code_descEquipment - No Trouble Found                           -0.954973
## code_descFire/Smoke Plan B - Source TTC                          0.662259
## code_descGraffiti / Scratchiti                                  -1.177606
## code_descHeld By Polce - Non-TTC Related                        -0.114612
## code_descInjured Employee                                       -1.366759
## code_descInjured or ill Customer (In Station) - Medical Aid Refused -0.959062
## code_descInjured or ill Customer (In Station) - Transported     -0.540033
```

```
## code_descInjured or ill Customer (On Train) - Medical Aid Refused   -0.777368
## code_descInjured or ill Customer (On Train) - Transported           -0.370652
## code_descMisc. Transportation Other - Employee Non-Chargeable       -1.246922
## code_descMiscellaneous Other                                        -0.744012
## code_descMiscellaneous Speed Control                                -1.425190
## code_descNo Operator Immediately Available                          -1.318700
## code_descNo Operator Immediately Available - Not E.S.A. Related     -0.944209
## code_descOperator Not In Position                                   -1.262694
## code_descOperator Overshot Platform                                 -1.288699
## code_descOperator Overspeeding                                      -1.415797
## code_descOperator Violated Signal                                   -1.005271
## code_descOPTO (COMMS) Train Door Monitoring                         -1.213542
## code_descPassenger Assistance Alarm Activated - No Trouble Found    -1.383240
## code_descPassenger Other                                            -0.334282
## code_descPriority One - Train in Contact With Person                 1.817072
## code_descPropulsion System                                          -0.771815
## code_descRC&S Maintenance Error - (Human)                           -0.555628
## code_descRC&S Other                                                 -0.958990
## code_descS/E/C Department Other                                     -0.504376
## code_descSignals - Track Circuit Problems                           -1.344676
## code_descSignals - Train Stops                                      -0.946298
## code_descSignals Axle Counter Block Failure                         -0.760641
## code_descSignals Other                                               0.365263
## code_descSpeed Control Equipment                                    -1.365741
## code_descSubway Car Radio Fault                                     -1.131021
## code_descSubway Radio System Fault                                   0.064332
## code_descSupervisory Error                                          -1.003161
## code_descTimeout                                                    -1.063144
## code_descTR Cab Doors                                               -1.164601
## code_descTrack Switch Failure - Signal Related Problem              -0.033445
## code_descTrain Control - VOBC                                       -1.040275
## code_descTraining Department Related Delays                         -1.044009
## code_descTransit Control Related Problems                           -0.775863
## code_descTransportation Department - Other                          -0.957457
## code_descUnauthorized at Track Level                                -0.167851
## code_descUnsanitary Vehicle                                         -1.150334
## code_descWeather Reports / Related Delays                           -0.269362
## code_descWork Refusal                                               -1.021350
## code_descWork Vehicle                                               -0.110289
## code_descWork Zone Problems - Track                                 -0.207953
## code_descYard/Carhouse Related Problems                             -1.061272
##                                                                     Std. Error
## (Intercept)                                                           0.079414
## lineSHP                                                               0.037595
## lineSRT                                                               0.041215
## lineYU                                                                0.017283
## code_descAssault / Patron Involved                                    0.089742
## code_descATC Operator Related                                         0.124026
## code_descATC Project                                                  0.083216
## code_descBody                                                         0.138634
## code_descBomb Threat                                                  0.157860
## code_descBrakes                                                       0.120824
## code_descConsequential Delay (2nd Delay Same Fault)                   0.120344
## code_descCrew Unable to Maintain Schedule                             0.144396
```

```
## code_descDebris At Track Level - Uncontrollable                          0.140020
## code_descDisorderly Patron                                               0.081534
## code_descDivisional Clerk Related                                        0.120678
## code_descDoor Problems - Debris Related                                  0.098961
## code_descDoor Problems - Faulty Equipment                                0.093214
## code_descDoor Problems - Passenger Related                               0.105737
## code_descEmergency Alarm Station Activation                              0.096630
## code_descEquipment - No Trouble Found                                    0.101524
## code_descFire/Smoke Plan B - Source TTC                                  0.105228
## code_descGraffiti / Scratchiti                                           0.131627
## code_descHeld By Polce - Non-TTC Related                                 0.116140
## code_descInjured Employee                                                0.177479
## code_descInjured or ill Customer (In Station) - Medical Aid Refused      0.172186
## code_descInjured or ill Customer (In Station) - Transported              0.138462
## code_descInjured or ill Customer (On Train) - Medical Aid Refused        0.086639
## code_descInjured or ill Customer (On Train) - Transported                0.087870
## code_descMisc. Transportation Other - Employee Non-Chargeable            0.092013
## code_descMiscellaneous Other                                             0.090822
## code_descMiscellaneous Speed Control                                     0.118256
## code_descNo Operator Immediately Available                               0.086037
## code_descNo Operator Immediately Available - Not E.S.A. Related          0.132961
## code_descOperator Not In Position                                        0.095952
## code_descOperator Overshot Platform                                      0.115299
## code_descOperator Overspeeding                                           0.175128
## code_descOperator Violated Signal                                        0.093172
## code_descOPTO (COMMS) Train Door Monitoring                              0.083229
## code_descPassenger Assistance Alarm Activated - No Trouble Found         0.084970
## code_descPassenger Other                                                 0.084977
## code_descPriority One - Train in Contact With Person                     0.115084
## code_descPropulsion System                                               0.126534
## code_descRC&S Maintenance Error - (Human)                                0.138562
## code_descRC&S Other                                                      0.171086
## code_descS/E/C Department Other                                          0.142896
## code_descSignals - Track Circuit Problems                                0.197519
## code_descSignals - Train Stops                                           0.125798
## code_descSignals Axle Counter Block Failure                              0.153620
## code_descSignals Other                                                   0.141490
## code_descSpeed Control Equipment                                         0.150554
## code_descSubway Car Radio Fault                                          0.177254
## code_descSubway Radio System Fault                                       0.161587
## code_descSupervisory Error                                               0.127595
## code_descTimeout                                                         0.110877
## code_descTR Cab Doors                                                    0.177680
## code_descTrack Switch Failure - Signal Related Problem                   0.143318
## code_descTrain Control - VOBC                                            0.125616
## code_descTraining Department Related Delays                              0.152209
## code_descTransit Control Related Problems                                0.127627
## code_descTransportation Department - Other                               0.087716
## code_descUnauthorized at Track Level                                     0.083729
## code_descUnsanitary Vehicle                                              0.089478
## code_descWeather Reports / Related Delays                                0.121618
## code_descWork Refusal                                                    0.158931
## code_descWork Vehicle                                                    0.163722
## code_descWork Zone Problems - Track                                      0.099812
```

```
## code_descYard/Carhouse Related Problems                                0.165308
##                                                                       z value
## (Intercept)                                                            34.506
## lineSHP                                                                 3.023
## lineSRT                                                                 8.343
## lineYU                                                                  0.081
## code_descAssault / Patron Involved                                     -3.747
## code_descATC Operator Related                                          -8.579
## code_descATC Project                                                  -12.571
## code_descBody                                                          -9.860
## code_descBomb Threat                                                    7.017
## code_descBrakes                                                        -7.059
## code_descConsequential Delay (2nd Delay Same Fault)                    -9.653
## code_descCrew Unable to Maintain Schedule                              -7.516
## code_descDebris At Track Level - Uncontrollable                        -0.637
## code_descDisorderly Patron                                            -10.452
## code_descDivisional Clerk Related                                      -8.887
## code_descDoor Problems - Debris Related                               -10.805
## code_descDoor Problems - Faulty Equipment                             -11.838
## code_descDoor Problems - Passenger Related                            -10.747
## code_descEmergency Alarm Station Activation                            -2.756
## code_descEquipment - No Trouble Found                                  -9.406
## code_descFire/Smoke Plan B - Source TTC                                 6.294
## code_descGraffiti / Scratchiti                                         -8.947
## code_descHeld By Polce - Non-TTC Related                               -0.987
## code_descInjured Employee                                              -7.701
## code_descInjured or ill Customer (In Station) - Medical Aid Refused    -5.570
## code_descInjured or ill Customer (In Station) - Transported            -3.900
## code_descInjured or ill Customer (On Train) - Medical Aid Refused      -8.972
## code_descInjured or ill Customer (On Train) - Transported              -4.218
## code_descMisc. Transportation Other - Employee Non-Chargeable         -13.552
## code_descMiscellaneous Other                                           -8.192
## code_descMiscellaneous Speed Control                                  -12.052
## code_descNo Operator Immediately Available                            -15.327
## code_descNo Operator Immediately Available - Not E.S.A. Related        -7.101
## code_descOperator Not In Position                                     -13.160
## code_descOperator Overshot Platform                                   -11.177
## code_descOperator Overspeeding                                         -8.084
## code_descOperator Violated Signal                                     -10.789
## code_descOPTO (COMMS) Train Door Monitoring                           -14.581
## code_descPassenger Assistance Alarm Activated - No Trouble Found      -16.279
## code_descPassenger Other                                               -3.934
## code_descPriority One - Train in Contact With Person                   15.789
## code_descPropulsion System                                             -6.100
## code_descRC&S Maintenance Error - (Human)                              -4.010
## code_descRC&S Other                                                    -5.605
## code_descS/E/C Department Other                                        -3.530
## code_descSignals - Track Circuit Problems                              -6.808
## code_descSignals - Train Stops                                         -7.522
## code_descSignals Axle Counter Block Failure                            -4.951
## code_descSignals Other                                                  2.582
## code_descSpeed Control Equipment                                       -9.071
## code_descSubway Car Radio Fault                                        -6.381
## code_descSubway Radio System Fault                                      0.398
```

```
## code_descSupervisory Error                                          -7.862
## code_descTimeout                                                    -9.588
## code_descTR Cab Doors                                               -6.554
## code_descTrack Switch Failure - Signal Related Problem              -0.233
## code_descTrain Control - VOBC                                       -8.281
## code_descTraining Department Related Delays                         -6.859
## code_descTransit Control Related Problems                           -6.079
## code_descTransportation Department - Other                         -10.915
## code_descUnauthorized at Track Level                                -2.005
## code_descUnsanitary Vehicle                                        -12.856
## code_descWeather Reports / Related Delays                           -2.215
## code_descWork Refusal                                               -6.426
## code_descWork Vehicle                                               -0.674
## code_descWork Zone Problems - Track                                 -2.083
## code_descYard/Carhouse Related Problems                             -6.420
##                                                                  Pr(>|z|)
## (Intercept)                                                       < 2e-16
## lineSHP                                                          0.002502
## lineSRT                                                           < 2e-16
## lineYU                                                           0.935199
## code_descAssault / Patron Involved                               0.000179
## code_descATC Operator Related                                     < 2e-16
## code_descATC Project                                              < 2e-16
## code_descBody                                                     < 2e-16
## code_descBomb Threat                                             2.27e-12
## code_descBrakes                                                  1.67e-12
## code_descConsequential Delay (2nd Delay Same Fault)               < 2e-16
## code_descCrew Unable to Maintain Schedule                        5.67e-14
## code_descDebris At Track Level - Uncontrollable                  0.523942
## code_descDisorderly Patron                                        < 2e-16
## code_descDivisional Clerk Related                                 < 2e-16
## code_descDoor Problems - Debris Related                           < 2e-16
## code_descDoor Problems - Faulty Equipment                         < 2e-16
## code_descDoor Problems - Passenger Related                        < 2e-16
## code_descEmergency Alarm Station Activation                      0.005853
## code_descEquipment - No Trouble Found                             < 2e-16
## code_descFire/Smoke Plan B - Source TTC                          3.10e-10
## code_descGraffiti / Scratchiti                                    < 2e-16
## code_descHeld By Polce - Non-TTC Related                         0.323719
## code_descInjured Employee                                        1.35e-14
## code_descInjured or ill Customer (In Station) - Medical Aid Refused 2.55e-08
## code_descInjured or ill Customer (In Station) - Transported      9.61e-05
## code_descInjured or ill Customer (On Train) - Medical Aid Refused  < 2e-16
## code_descInjured or ill Customer (On Train) - Transported        2.46e-05
## code_descMisc. Transportation Other - Employee Non-Chargeable     < 2e-16
## code_descMiscellaneous Other                                     2.57e-16
## code_descMiscellaneous Speed Control                              < 2e-16
## code_descNo Operator Immediately Available                        < 2e-16
## code_descNo Operator Immediately Available - Not E.S.A. Related  1.24e-12
## code_descOperator Not In Position                                 < 2e-16
## code_descOperator Overshot Platform                               < 2e-16
## code_descOperator Overspeeding                                   6.25e-16
## code_descOperator Violated Signal                                 < 2e-16
## code_descOPTO (COMMS) Train Door Monitoring                       < 2e-16
```

```
## code_descPassenger Assistance Alarm Activated - No Trouble Found     < 2e-16
## code_descPassenger Other                                            8.36e-05
## code_descPriority One - Train in Contact With Person                 < 2e-16
## code_descPropulsion System                                          1.06e-09
## code_descRC&S Maintenance Error - (Human)                           6.07e-05
## code_descRC&S Other                                                 2.08e-08
## code_descS/E/C Department Other                                     0.000416
## code_descSignals - Track Circuit Problems                           9.91e-12
## code_descSignals - Train Stops                                      5.38e-14
## code_descSignals Axle Counter Block Failure                         7.37e-07
## code_descSignals Other                                              0.009836
## code_descSpeed Control Equipment                                     < 2e-16
## code_descSubway Car Radio Fault                                     1.76e-10
## code_descSubway Radio System Fault                                  0.690536
## code_descSupervisory Error                                          3.78e-15
## code_descTimeout                                                     < 2e-16
## code_descTR Cab Doors                                               5.58e-11
## code_descTrack Switch Failure - Signal Related Problem              0.815479
## code_descTrain Control - VOBC                                        < 2e-16
## code_descTraining Department Related Delays                         6.93e-12
## code_descTransit Control Related Problems                           1.21e-09
## code_descTransportation Department - Other                           < 2e-16
## code_descUnauthorized at Track Level                                0.044995
## code_descUnsanitary Vehicle                                          < 2e-16
## code_descWeather Reports / Related Delays                           0.026773
## code_descWork Refusal                                               1.31e-10
## code_descWork Vehicle                                               0.500542
## code_descWork Zone Problems - Track                                 0.037212
## code_descYard/Carhouse Related Problems                             1.36e-10
##
## (Intercept)                                                        ***
## lineSHP                                                            **
## lineSRT                                                            ***
## lineYU
## code_descAssault / Patron Involved                                 ***
## code_descATC Operator Related                                      ***
## code_descATC Project                                               ***
## code_descBody                                                      ***
## code_descBomb Threat                                               ***
## code_descBrakes                                                    ***
## code_descConsequential Delay (2nd Delay Same Fault)                ***
## code_descCrew Unable to Maintain Schedule                          ***
## code_descDebris At Track Level - Uncontrollable
## code_descDisorderly Patron                                         ***
## code_descDivisional Clerk Related                                  ***
## code_descDoor Problems - Debris Related                            ***
## code_descDoor Problems - Faulty Equipment                          ***
## code_descDoor Problems - Passenger Related                         ***
## code_descEmergency Alarm Station Activation                        **
## code_descEquipment - No Trouble Found                              ***
## code_descFire/Smoke Plan B - Source TTC                            ***
## code_descGraffiti / Scratchiti                                     ***
## code_descHeld By Polce - Non-TTC Related
## code_descInjured Employee                                          ***
```

```
## code_descInjured or ill Customer (In Station) - Medical Aid Refused ***
## code_descInjured or ill Customer (In Station) - Transported         ***
## code_descInjured or ill Customer (On Train) - Medical Aid Refused   ***
## code_descInjured or ill Customer (On Train) - Transported           ***
## code_descMisc. Transportation Other - Employee Non-Chargeable       ***
## code_descMiscellaneous Other                                        ***
## code_descMiscellaneous Speed Control                                ***
## code_descNo Operator Immediately Available                          ***
## code_descNo Operator Immediately Available - Not E.S.A. Related     ***
## code_descOperator Not In Position                                   ***
## code_descOperator Overshot Platform                                 ***
## code_descOperator Overspeeding                                      ***
## code_descOperator Violated Signal                                   ***
## code_descOPTO (COMMS) Train Door Monitoring                         ***
## code_descPassenger Assistance Alarm Activated - No Trouble Found    ***
## code_descPassenger Other                                            ***
## code_descPriority One - Train in Contact With Person                ***
## code_descPropulsion System                                          ***
## code_descRC&S Maintenance Error - (Human)                           ***
## code_descRC&S Other                                                 ***
## code_descS/E/C Department Other                                     ***
## code_descSignals - Track Circuit Problems                           ***
## code_descSignals - Train Stops                                      ***
## code_descSignals Axle Counter Block Failure                         ***
## code_descSignals Other                                              **
## code_descSpeed Control Equipment                                    ***
## code_descSubway Car Radio Fault                                     ***
## code_descSubway Radio System Fault
## code_descSupervisory Error                                          ***
## code_descTimeout                                                    ***
## code_descTR Cab Doors                                               ***
## code_descTrack Switch Failure - Signal Related Problem
## code_descTrain Control - VOBC                                       ***
## code_descTraining Department Related Delays                         ***
## code_descTransit Control Related Problems                           ***
## code_descTransportation Department - Other                          ***
## code_descUnauthorized at Track Level                                *
## code_descUnsanitary Vehicle                                         ***
## code_descWeather Reports / Related Delays                           *
## code_descWork Refusal                                               ***
## code_descWork Vehicle
## code_descWork Zone Problems - Track                                 *
## code_descYard/Carhouse Related Problems                             ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(3.597) family taken to be 1)
##
##     Null deviance: 14146.4  on 8540  degrees of freedom
## Residual deviance:  7529.1  on 8474  degrees of freedom
## AIC: 46860
##
## Number of Fisher Scoring iterations: 1
##
```

```
##
##               Theta:  3.5970
##           Std. Err.:  0.0731
##
##  2 x log-likelihood:  -46724.4860
```

Answer2

The results of the Negative Binomial Regression model show that both line and reasons have significant effects on the count of delay minutes.For example, such as "ATC Operator Related" and "ATC Project," are associated with a substantial increase in delay minutes, while delays related to "Assault / Patron Involved" are associated with a decrease in delay minutes. There are also some covariates that do not have significant effects on the delay time, such as "Radio System Fault" and "Track Switch Failure - Signal Related Problem", which also in line with the reality .

Question3

```
res1 <- list_package_resources("e869d365-2c15-4893-ad2a-744ca867be3b") # obtained code from searching d
res1 <- res1 |> mutate(year = str_extract(name, "201.?"))
campaign2014_ids <- res1 |> filter(year==2014 & grepl("Data", name)) |> pull(id)
campaign_2014 <- get_resource(campaign2014_ids)
```

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

```
mayor_data <- campaign_2014[grep("Mayor", names(campaign_2014))][[1]]#select the maylor selection relat
colnames(mayor_data) <- unlist(mayor_data[1, ]) #first row as the column names
mayor_data <- mayor_data[-1, ]

mayor_data <- mayor_data %>%
  clean_names()
head(mayor_data)
```

```
## # A tibble: 6 x 13
##   contributors_name  contributors_address contributors_postal_code
##   <chr>              <chr>                <chr>
## 1 A D'Angelo, Tullio <NA>                 M6A 1P5
## 2 A Strazar, Martin  <NA>                 M2M 3B8
## 3 A'Court, K Susan   <NA>                 M4M 2J8
## 4 A'Court, K Susan   <NA>                 M4M 2J8
## 5 A'Court, K Susan   <NA>                 M4M 2J8
## 6 Aaron, Robert B    <NA>                 M6B 1H7
## # i 10 more variables: contribution_amount <chr>, contribution_type_desc <chr>,
```

```
## #   goods_or_service_desc <chr>, contributor_type_desc <chr>,
## #   relationship_to_candidate <chr>, president_business_manager <chr>,
## #   authorized_representative <chr>, candidate <chr>, office <chr>, ward <chr>
```

Question4

```
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

```
#Check the missing value
mayor_data %>%
  summarize(across(everything(), ~ sum(is.na(.x))))
```

```
## # A tibble: 1 x 13
##   contributors_name contributors_address contributors_postal_code
##               <int>                <int>                    <int>
## 1                 0                10197                        0
## # i 10 more variables: contribution_amount <int>, contribution_type_desc <int>,
## #   goods_or_service_desc <int>, contributor_type_desc <int>,
## #   relationship_to_candidate <int>, president_business_manager <int>,
## #   authorized_representative <int>, candidate <int>, office <int>, ward <int>
```

```r
#Check the duplicates
get_dupes(mayor_data)
```

```
## No variable names specified - using all columns.

## # A tibble: 1,716 x 14
##    contributors_name contributors_address contributors_postal_code
##    <chr>             <chr>                <chr>
##  1 Henery, Marjorie  <NA>                 M9C 4W1
##  2 Henery, Marjorie  <NA>                 M9C 4W1
##  3 Henery, Marjorie  <NA>                 M9C 4W1
##  4 Henery, Marjorie  <NA>                 M9C 4W1
##  5 Henery, Marjorie  <NA>                 M9C 4W1
##  6 Henery, Marjorie  <NA>                 M9C 4W1
##  7 Henery, Marjorie  <NA>                 M9C 4W1
##  8 Amodeo, Merle     <NA>                 M4E 1H2
##  9 Amodeo, Merle     <NA>                 M4E 1H2
## 10 Amodeo, Merle     <NA>                 M4E 1H2
## # i 1,706 more rows
## # i 11 more variables: contribution_amount <chr>, contribution_type_desc <chr>,
## #   goods_or_service_desc <chr>, contributor_type_desc <chr>,
## #   relationship_to_candidate <chr>, president_business_manager <chr>,
## #   authorized_representative <chr>, candidate <chr>, office <chr>, ward <chr>,
## #   dupe_count <int>
```

```r
mayor_data <- mayor_data |> distinct()

#Check the data type of every column
sapply(mayor_data, class)
```

```
##          contributors_name       contributors_address
##                "character"                  "character"
##   contributors_postal_code        contribution_amount
##                "character"                  "character"
##     contribution_type_desc      goods_or_service_desc
##                "character"                  "character"
##      contributor_type_desc  relationship_to_candidate
##                "character"                  "character"
## president_business_manager  authorized_representative
##                "character"                  "character"
##                  candidate                       office
##                "character"                  "character"
##                       ward
##                "character"
```

```r
#Transfer the character column format to numeric
mayor_data$contribution_amount <- as.numeric(mayor_data$contribution_amount)
```

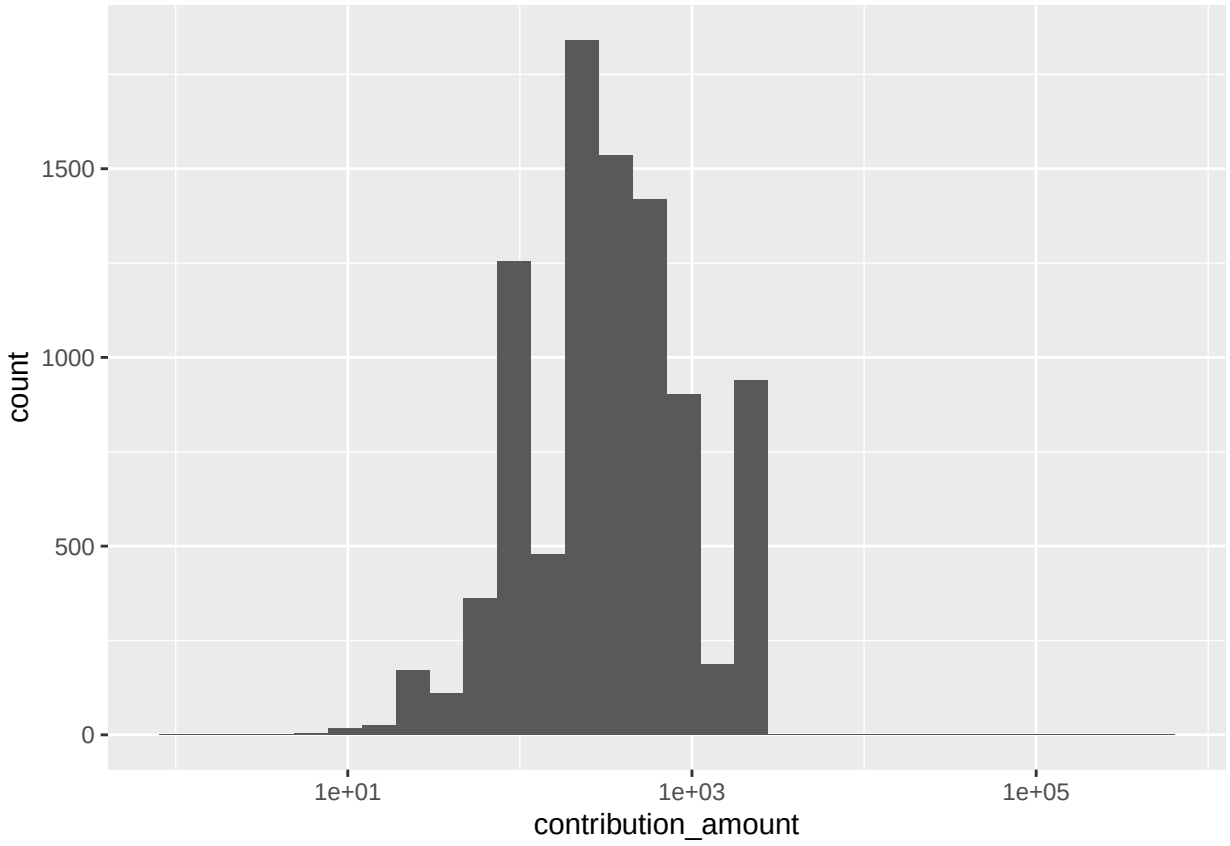
Answer4

There are some empty columns :contributors_address, goods_or_service_desc, relationship_to_candidate, president_business_manager, authorized_representative and ward, so we can exclude these empty columns in our following analysis. On top of these empty columns, there is no missing values for other columns.

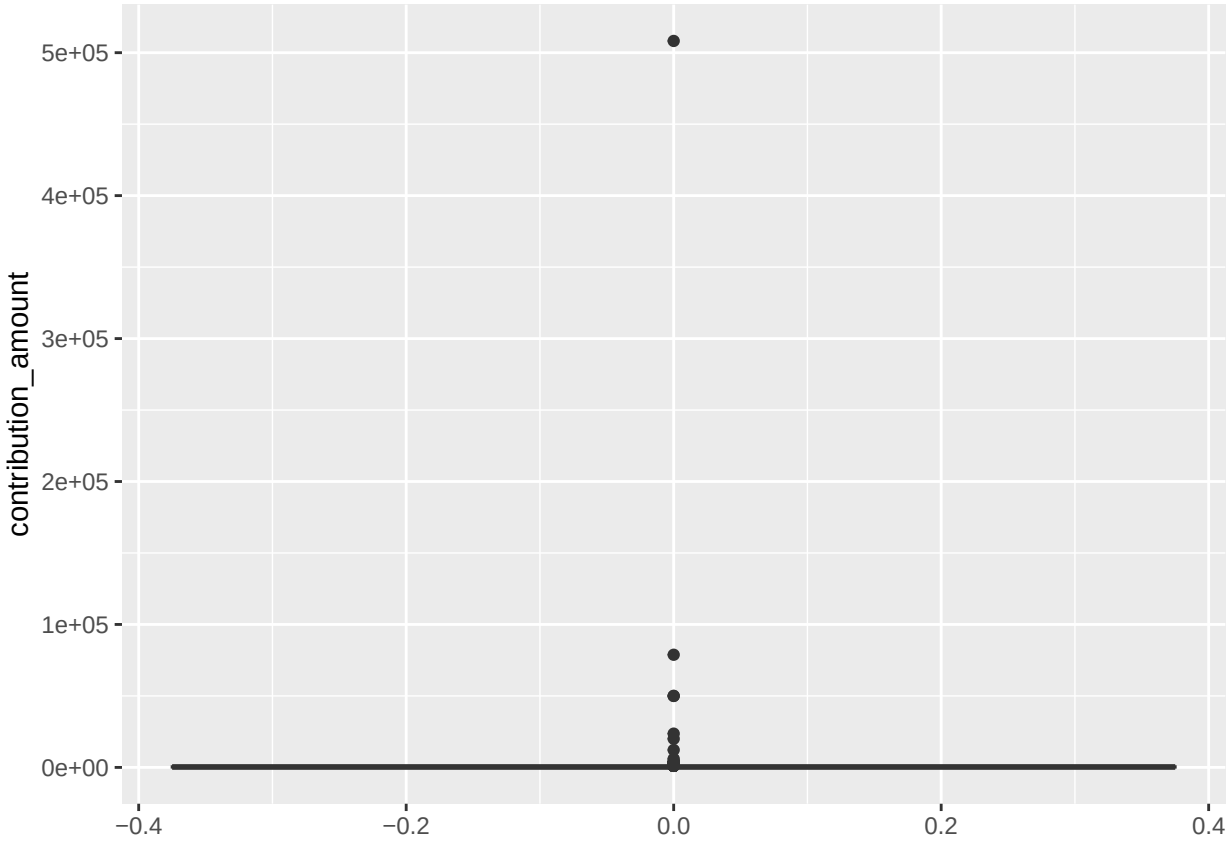
Question5

```
#check the distribution of contribution amount
ggplot(data = mayor_data) +
  geom_histogram(aes(x = contribution_amount)) + scale_x_log10()
```

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




```
#check if there is outlier of thr contribution amount
ggplot(mayor_data, aes(y = contribution_amount)) +
  geom_boxplot()
```



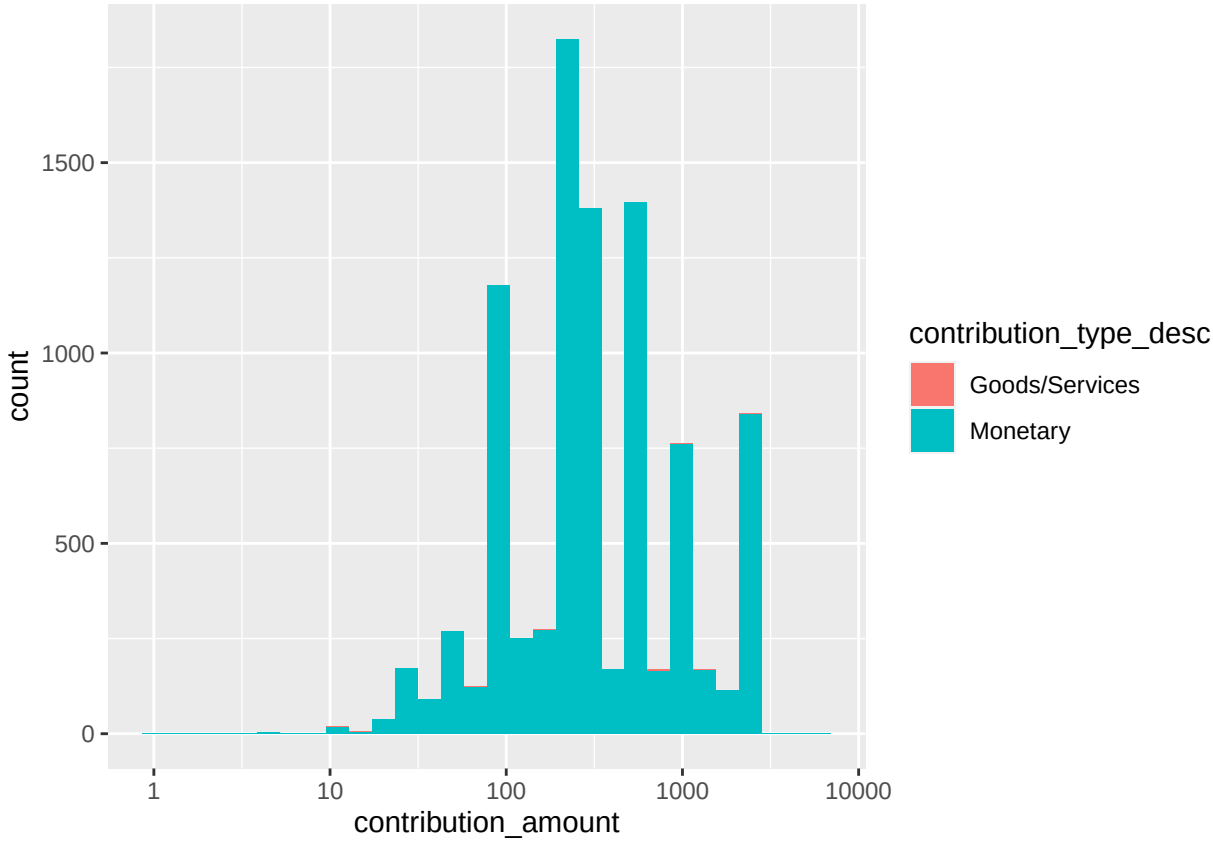

```
#check the outlier
mayor_data %>%
  arrange(-contribution_amount) %>%
  slice(1:10)
```

```
## # A tibble: 10 x 13
##    contributors_name contributors_address       contributors_postal_code
##    <chr>             <chr>                      <chr>
##  1 Ford, Doug        <NA>                       M9A 2C3
##  2 Ford, Rob         <NA>                       M9A 3G9
##  3 Ford, Doug        <NA>                       M9A 2C3
##  4 Ford, Rob         <NA>                       M9A 3G9
##  5 Goldkind, Ari     <NA>                       M5P 1P5
##  6 Ford, Rob         <NA>                       M9A 3G9
##  7 Ford, Rob         <NA>                       M9A 3G9
##  8 Di Paola, Rocco   <NA>                       M3H 2T1
##  9 Thomson, Sarah    <NA>                       M4W 2X6
## 10 kindred's Muze    723 Dovercourt Rd, Toronto M6H 2W7
## # i 10 more variables: contribution_amount <dbl>, contribution_type_desc <chr>,
## #   goods_or_service_desc <chr>, contributor_type_desc <chr>,
## #   relationship_to_candidate <chr>, president_business_manager <chr>,
## #   authorized_representative <chr>, candidate <chr>, office <chr>, ward <chr>
```

```
#exclude outlier
mayor_data<- mayor_data %>% filter(contribution_amount <= 10000)
```

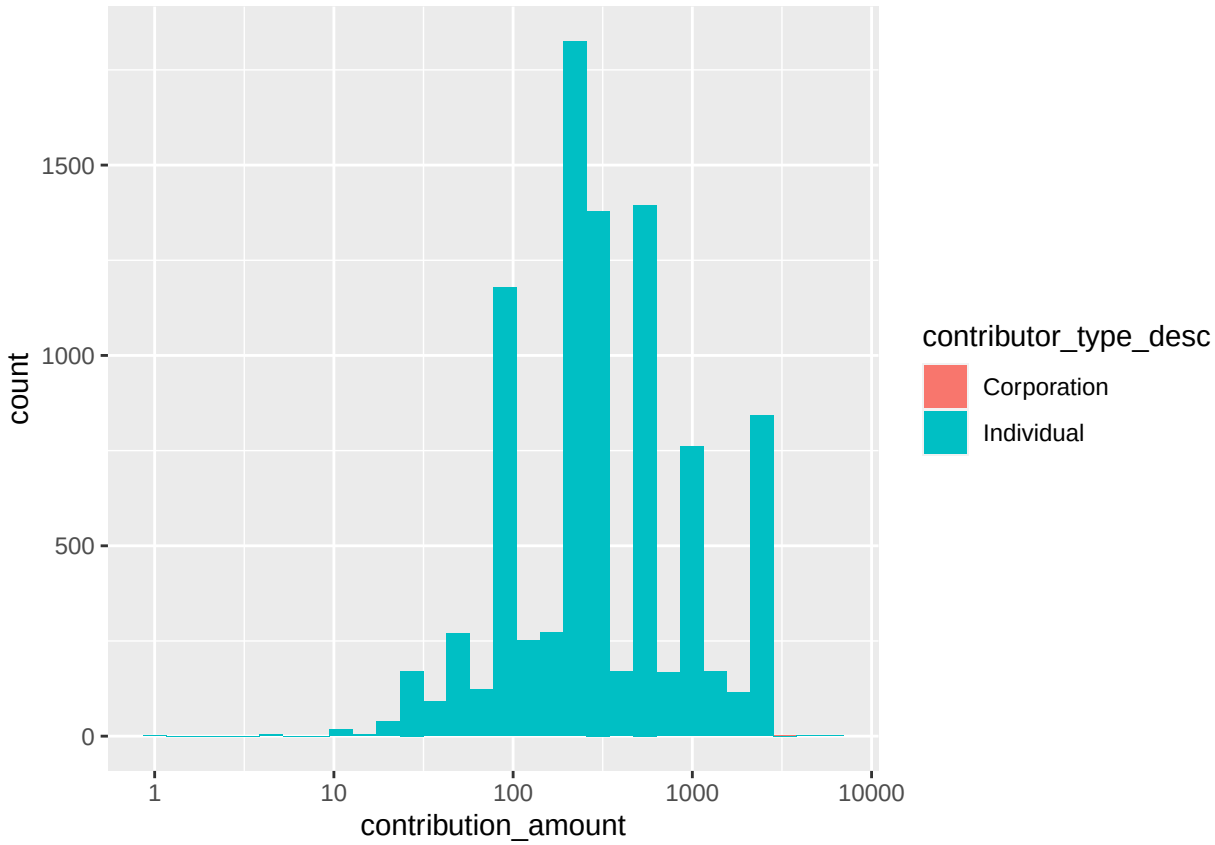
```
#check the distribution of contribution type
ggplot(data = mayor_data) +
  geom_histogram(aes(x = contribution_amount, fill= contribution_type_desc)) + scale_x_log10()
```

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

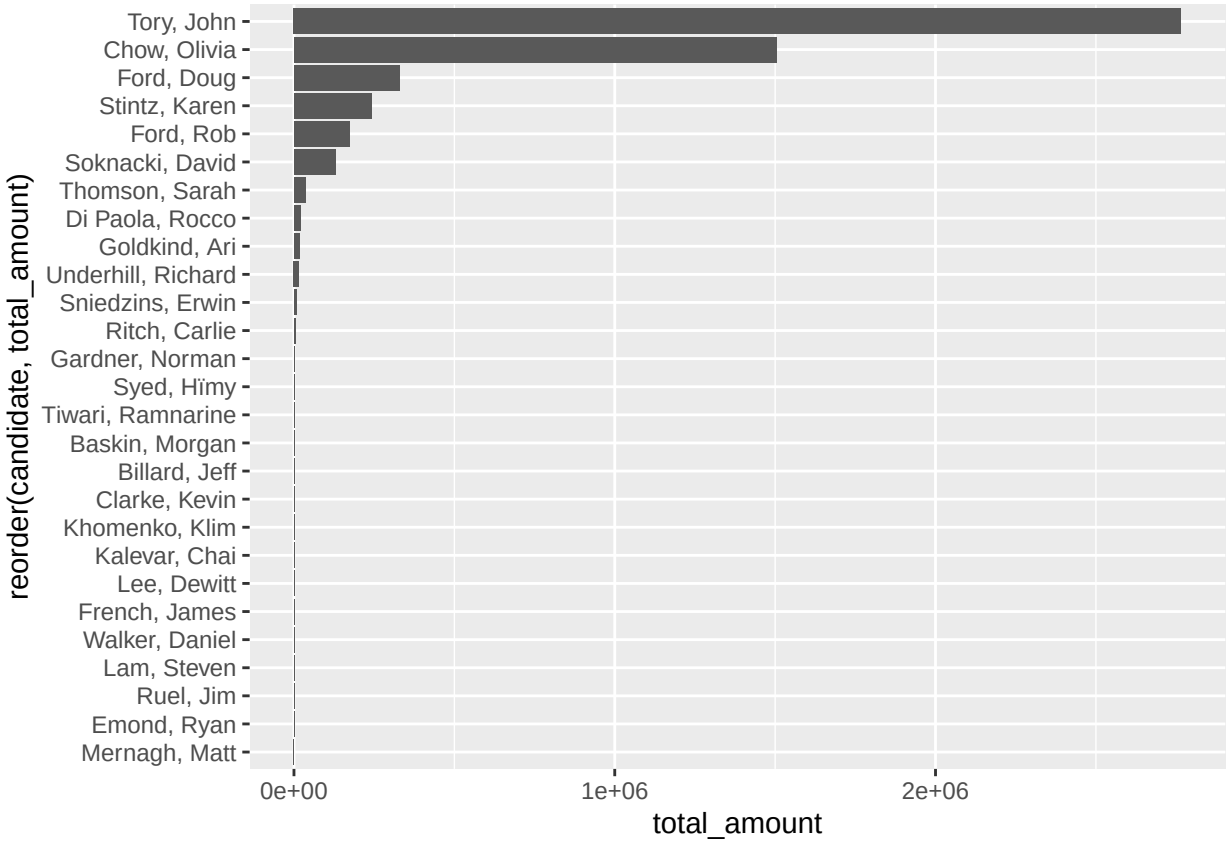



```
#check the distribution of contributor type
ggplot(data = mayor_data) +
  geom_histogram(aes(x = contribution_amount, fill= contributor_type_desc)) + scale_x_log10()
```

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




```
#Show the distribution of contribution amount for different candidates
mayor_data |>
  group_by(candidate) |>
  summarise(total_amount = sum(contribution_amount)) |>
  ggplot(aes(x=reorder(candidate, total_amount), y=total_amount)) +
  geom_col() +
  coord_flip()
```

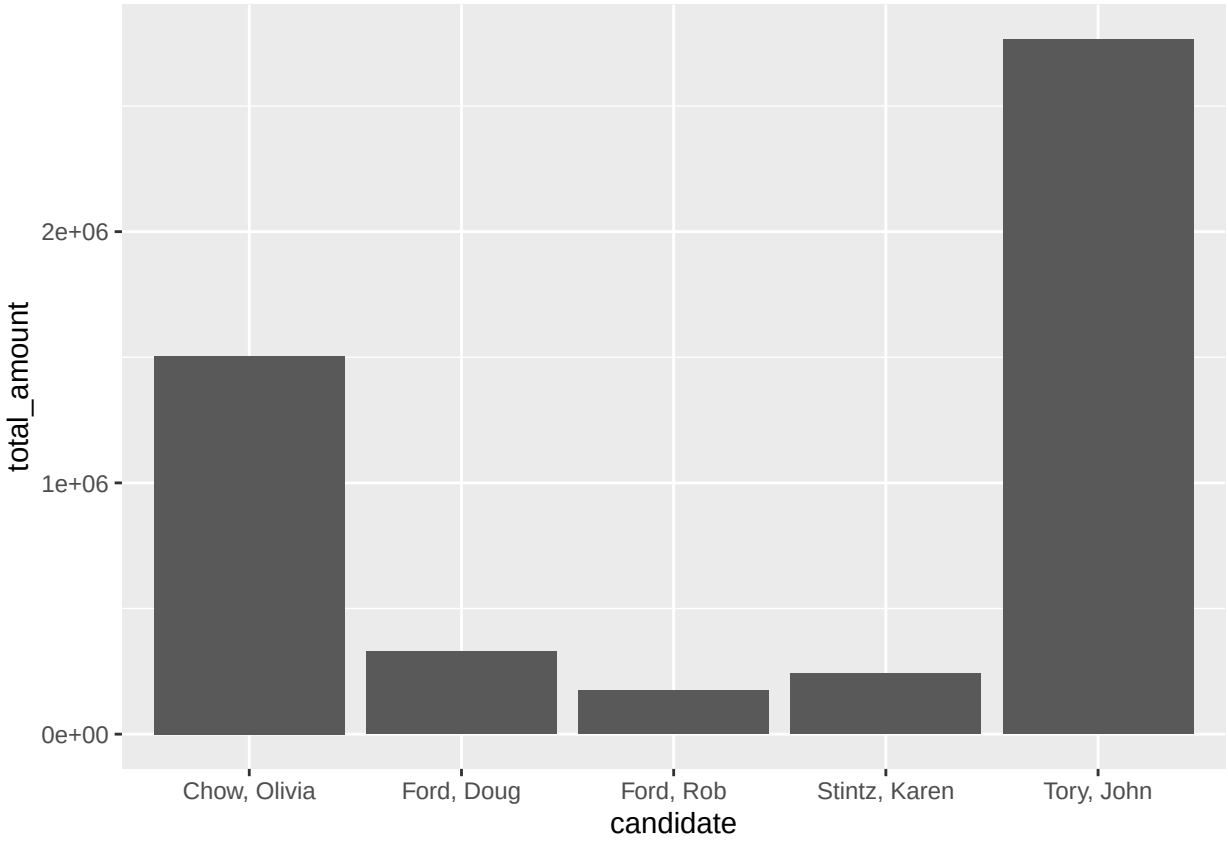


Answer5

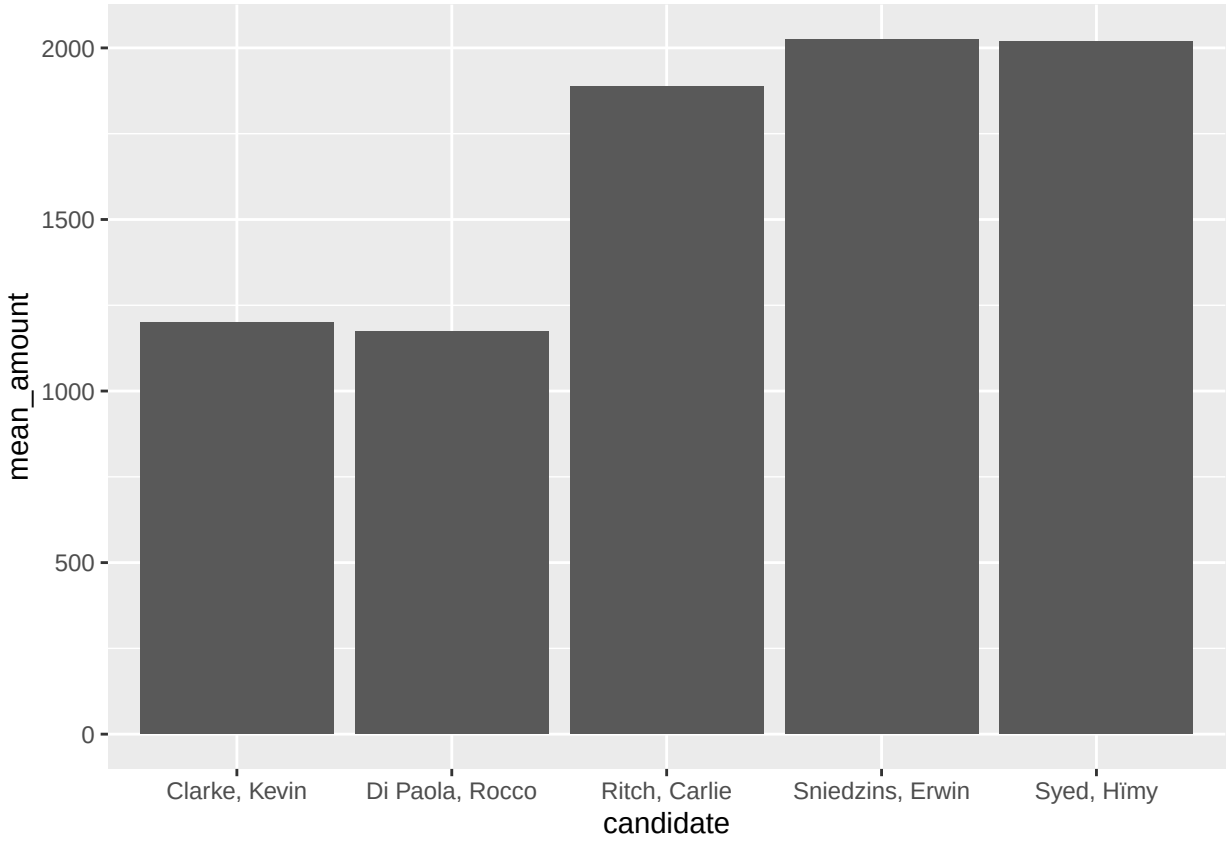
By checking the outliers of the contribution amount, we could see that there are 6 rows of amount which are beyond 10000 dollar, and these contributors are also the candidates themselves, they are : Ford Rob, Ford Doug and Goldkind Ari. And 4 out of 6 outlier contributions are from Ford Rob.

Question6

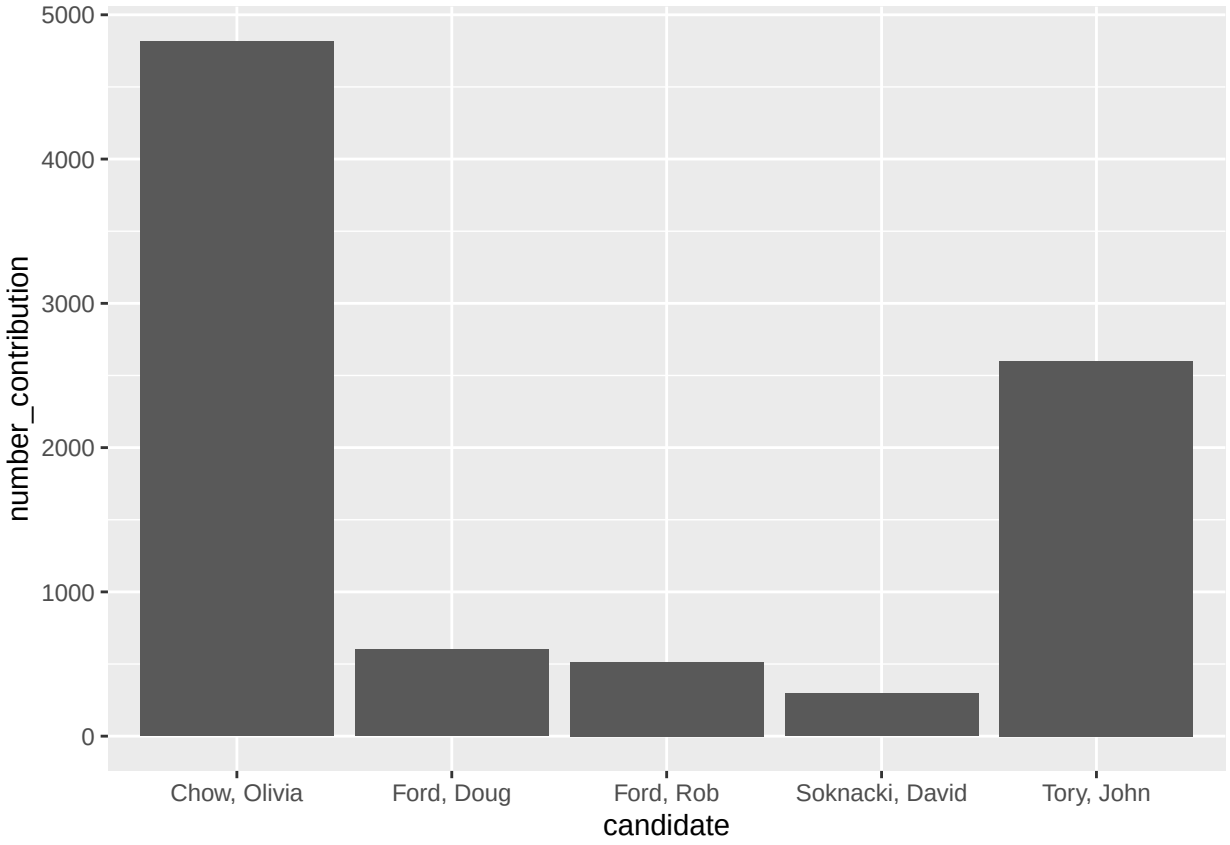
```
candidate_data <- mayor_data |>
  group_by(candidate) |>
  summarise(total_amount = sum(contribution_amount), mean_amount = mean(contribution_amount), number_co

candidate_data %>%
  arrange(-total_amount) %>%
  slice(1:5) %>%
  ggplot(aes(x= candidate, y=total_amount)) +
  geom_col()
```




```
candidate_data %>%
  arrange(-mean_amount) %>%
  slice(1:5) %>%
  ggplot(aes(x= candidate, y=mean_amount)) +
  geom_col()
```




```
candidate_data %>%
  arrange(-number_contribution) %>%
  slice(1:5) %>%
  ggplot(aes(x= candidate, y=number_contribution)) +
  geom_col()
```

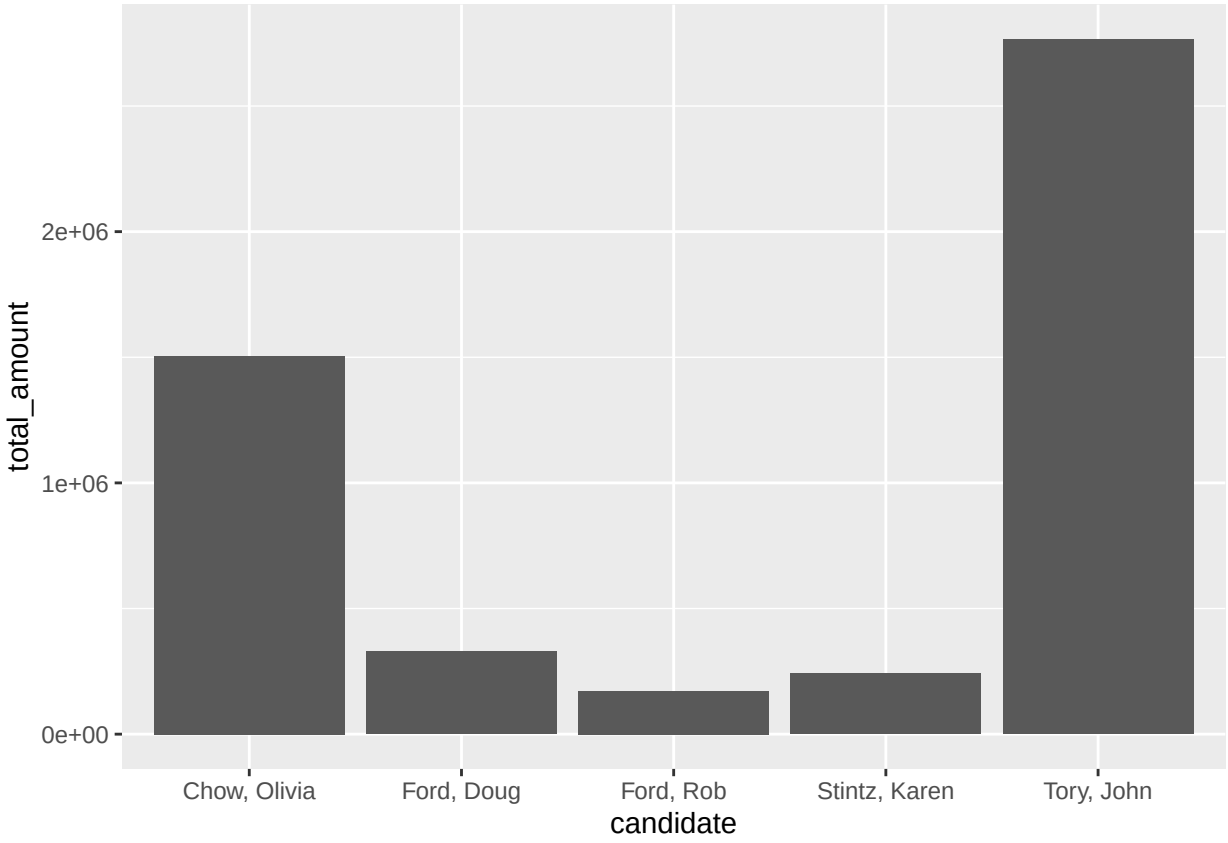


Question7

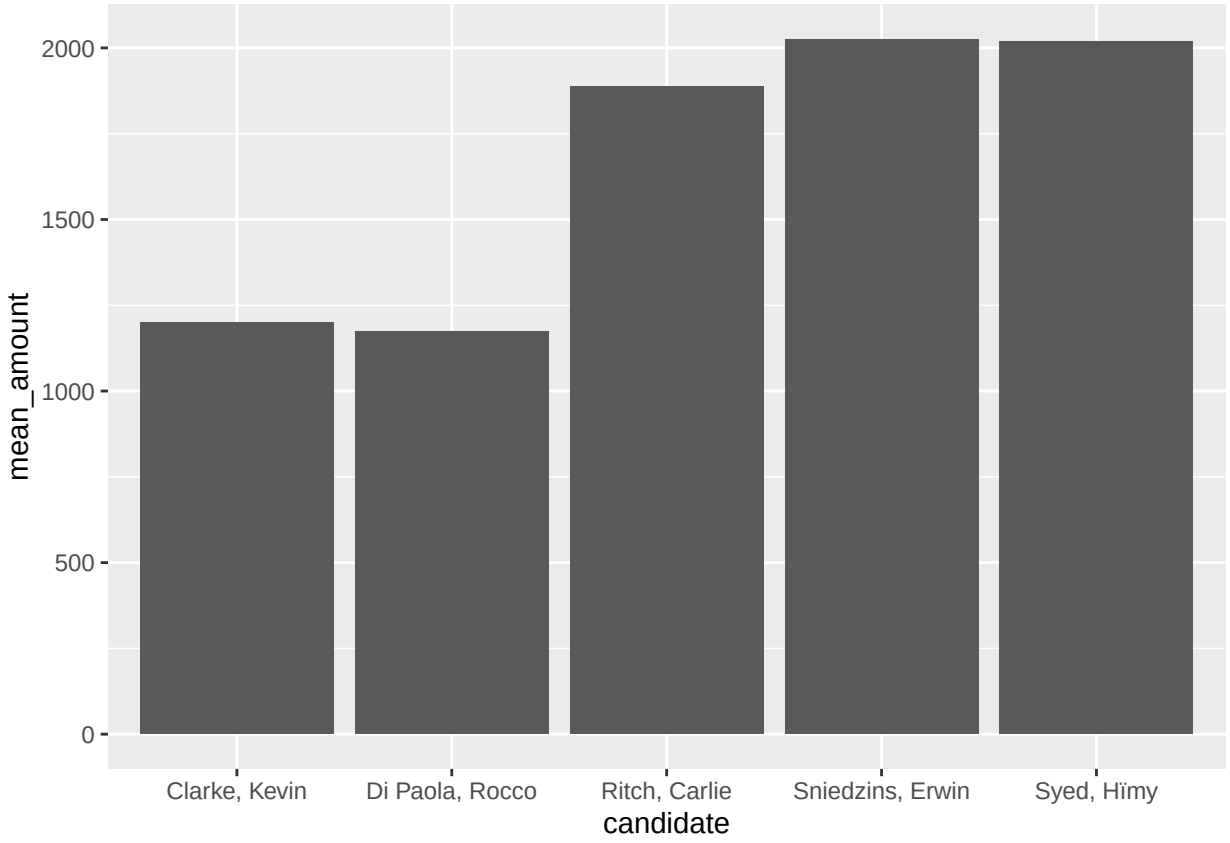
```
candidate_data_without_self <- mayor_data %>%
  filter(!grepl(candidate, contributors_name)) %>%
  group_by(candidate) %>%
  summarise(
    total_amount = sum(contribution_amount),
    mean_amount = mean(contribution_amount),
    number_contribution = n()
  )
```

```
## Warning: There was 1 warning in `filter()`.
## i In argument: `!grepl(candidate, contributors_name)`.
## Caused by warning in `grepl()`:
## ! argument 'pattern' has length > 1 and only the first element will be used
```

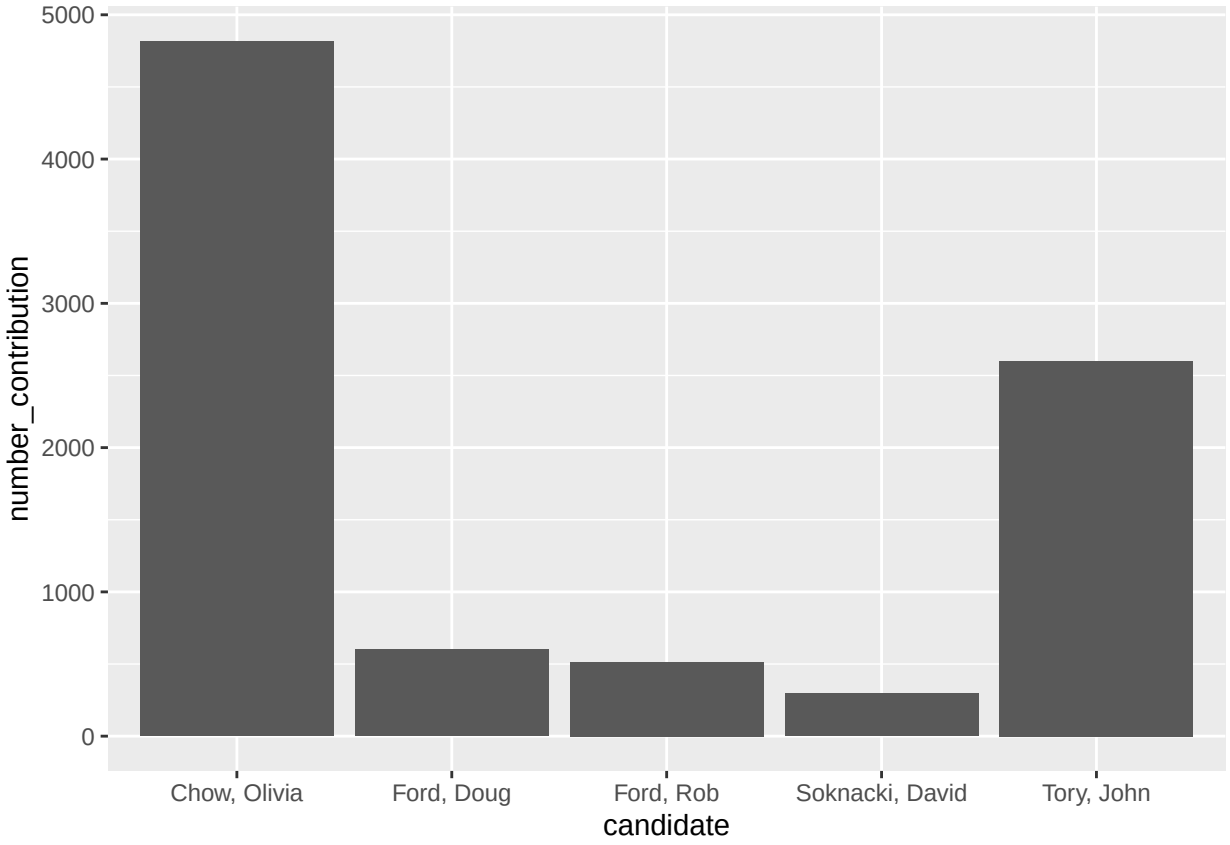
```
candidate_data_without_self %>%
  arrange(-total_amount) %>%
  slice(1:5) %>%
  ggplot(aes(x= candidate, y=total_amount)) +
  geom_col()
```




```
candidate_data_without_self %>%
  arrange(-mean_amount) %>%
  slice(1:5) %>%
  ggplot(aes(x= candidate, y=mean_amount)) +
  geom_col()
```




```
candidate_data_without_self %>%
  arrange(-number_contribution) %>%
  slice(1:5) %>%
  ggplot(aes(x= candidate, y=number_contribution)) +
  geom_col()
```



Question8

```
# Count the number of contributors who gave money to more than one candidate
contributors_multiple_candidates <- mayor_data %>%
  group_by(contributors_name) %>%
  summarise(num_candidates = n_distinct(candidate)) %>%
  filter(num_candidates > 1) %>%
  count()

contributors_multiple_candidates
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   184
```

Answer8

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