EDA and data visualization

Kishore Basu

21/01/23

Table of contents

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

```
all_data <- list_packages(limit = 500) # find id of table we need
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 Traffic ~ a330~ Trans~ <NA>    Transp~ This d~ Map          12 XSD,SH~ As ava~
2 Polls co~ 7bce~ City ~ <NA>    City C~ Polls ~ Table         5 JSON,X~ Daily
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 Daily Sh~ 21c8~ Commu~ Afford~ Shelte~ Daily ~ Table        12 JSON,C~ Daily
6 BodySafe  c405~ City ~ <NA>    Toront~ This d~ Map           9 SHP,CS~ 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
```

Let’s download the data on TTC subway delays in 2022.

```
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)
```

Let’s also download the delay code and readme, as reference.

```
# note: I obtained these codes from the 'id' column in the `res` object above
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")
```

This dataset has a bunch of interesting variables. You can refer to the readme for descriptions. Our outcome of interest is `min_delay`, which give the delay in mins.

```
head(delay_2022)
```

```
# A tibble: 6 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 LAWRENCE~ 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 EGLINTON~ MUATC       0       0 S     YU
6 2022-01-01 00:00:00 05:40 Saturday QUEEN ST~ MUNCA       0       0 <NA>  YU
# ... with 1 more variable: vehicle <dbl>, and abbreviated variable name
#   1: min_delay
```

```r
delay_2022 <- delay_2022 %>% distinct()
```

```r
## Removing the observations that have non-standardized lines

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

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

```
Joining, by = "code"
```

```r
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 DESCR
  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, by = "code_srt"
```

The largest delay is due to “Signals Other”.

```r
delay_2022 |>
  left_join(delay_codes |> rename(code = `SUB RMENU CODE`, code_desc = `CODE DESCRIPTION..
  arrange(-min_delay) |>
  select(date, time, station, line, min_delay, code, code_desc)
```

```
Joining, by = c("code", "code_desc")
```

```
# A tibble: 17,819 x 7
   date                time  station                line  min_de~1 code  code_~2
   <dttm>              <chr> <chr>                  <chr>    <dbl> <chr> <chr>
 1 2022-08-22 00:00:00 12:20 SRT LINE               SRT        451 PRSO  Signal~
 2 2022-04-28 00:00:00 06:02 JANE STATION           BD         388 PUTR  Rail R~
 3 2022-07-26 00:00:00 07:06 YONGE BD STATION       BD         382 MUPLB Fire/S~
 4 2022-08-15 00:00:00 12:57 DUFFERIN STATION       BD         327 MUPR1 Priori~
 5 2022-01-26 00:00:00 20:15 KENNEDY SRT STATION    SRT        315 MRWEA Weathe~
 6 2022-08-02 00:00:00 21:23 HIGHWAY 407 STATION    YU         312 MUPR1 Priori~
 7 2022-01-17 00:00:00 21:30 SHEPPARD WEST TO UNION YU         291 MUFM  Force ~
 8 2022-01-25 00:00:00 21:03 SCARBOROUGH CTR STATIO SRT        285 PRSL  Loop R~
 9 2022-06-17 00:00:00 12:25 KIPLING STATION        BD         241 SUUT  Unauth~
10 2022-02-09 00:00:00 06:06 DUPONT STATION         YU         240 SUAE  Assaul~
# ... with 17,809 more rows, and abbreviated variable names 1: min_delay,
#   2: code_desc
```

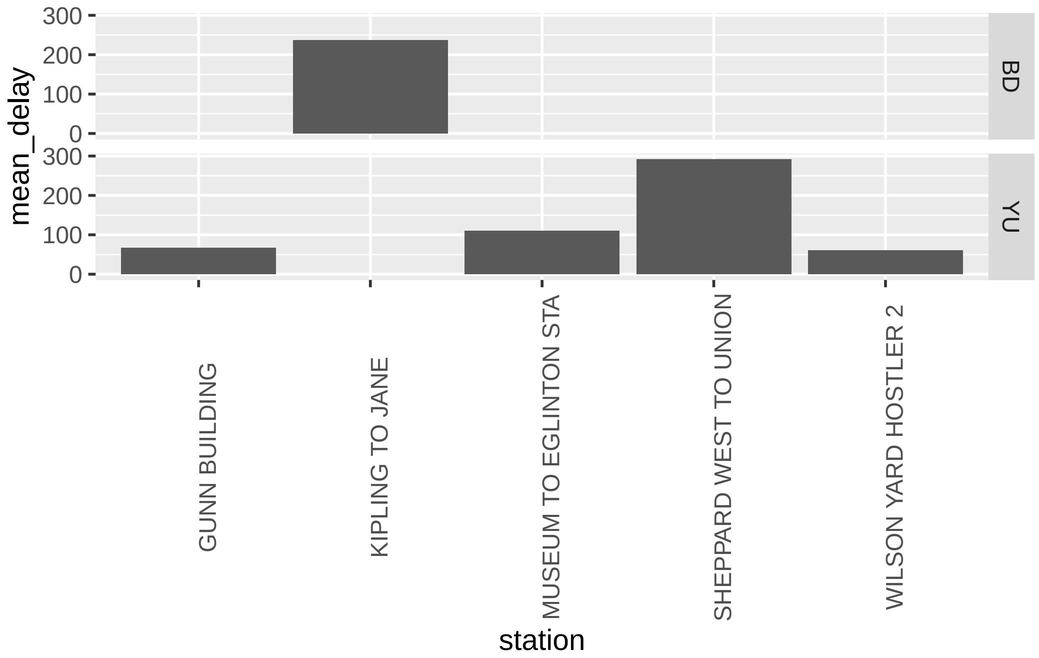
1 Lab Exercises

To be handed in via submission of quarto file (and rendered pdf) to GitHub.

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

```
delay_2022 %>%
  group_by(station) %>%
  summarize(station, mean_delay = mean(min_delay, na.rm = T), line) %>%
  arrange(-mean_delay) %>%
  head(5)%>%
  ggplot(aes(x = station, y = mean_delay)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90))+
  facet_grid(vars(line))
```

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



2. 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 <- list_packages(limit = 500) # find id of table we need
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 Traffic~ a330~ Trans~ <NA>    Transp~ This d~ Map          12 XSD,SH~ As ava~
 2 Polls c~ 7bce~ City ~ <NA>    City C~ Polls ~ Table         5 JSON,X~ Daily
 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 Daily S~ 21c8~ Commu~ Afford~ Shelte~ Daily ~ Table        12 JSON,C~ Daily
 6 BodySafe c405~ City ~ <NA>    Toront~ This d~ Map           9 SHP,CS~ Daily
 7 Municip~ 57b2~ Busin~ <NA>    Munici~ Some b~ Table         5 JSON,C~ Daily
 8 EarlyON~ earl~ Commu~ Povert~ Childr~ EarlyO~ Map          17 GPKG,S~ Daily
 9 Chemica~ ae8e~ Publi~ <NA>    Toront~ This d~ Table         6 XML,JS~ Daily
10 Committ~ 260e~ City ~ Afford~ City P~ This d~ Table        96 JSON,C~ Weekly
# ... 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
```

```
id <- 'f6651a40-2f52-46fc-9e04-b760c16edd5c'
res <- list_package_resources(id)
res
```

```
# A tibble: 2 x 4
  name                                   id                    format last_mod~1
  <chr>                                  <chr>                 <chr>  <date>
1 campaign-contributions-2014-data       5b230e92-0a22-4a15-9~ ZIP    2019-07-23
2 campaign-contributions-2014-readme-xls aaf736f4-7468-4bda-9~ XLS    2019-07-23
# ... with abbreviated variable name 1: last_modified
```

```
get_resource("5b230e92-0a22-4a15-9572-0b19cc222985")
```

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

```
$`1_Contribution_Summary_2014_election.xls`
# A tibble: 7 x 3
  `2014 Municipal Election - Summary of Contributions` ...2                 ...3
  <chr>                                                <chr>               <chr>
1 Office                                               # of Contributions~ Tota~
2 Mayor                                                10199               6200~
3 Councillor                                           11035               4532~
4 Toronto District School Board                        1056                6170~
5 Toronto Catholic District School Board               154                 1401~
6 Conseil scolaire Viamonde                            3                   1167
7 Conseil scolaire de district catholique Centre-Sud   5                   900
```

```
$`2_Mayor_Contributions_2014_election.xls`
# 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`

$`3_Counillor_Contributions_2014_election.xls`
# A tibble: 11,036 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 647773 Ont~ 190 ~ M5T ~ 200   Mone~ <NA>  Corp~ <NA>  Miha~ Miha~ Jeff~ Coun~
 3 Abadesso, ~ <NA>  M6H ~ 350   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Bail~ Coun~
 4 Abadesso, ~ <NA>  M6H ~ 350   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Bail~ Coun~
 5 Abadi, Bab~ <NA>  M5S ~ 500   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Wong~ Coun~
 6 Abate, Pao~ <NA>  L4L ~ 375   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Perr~ Coun~
 7 Abbas, Sye~ <NA>  L6S ~ 750   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Baig~ Coun~
 8 Abbott, Da~ <NA>  M6L ~ 300   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Wong~ Coun~
 9 Abbott, Na~ <NA>  L1V ~ 300   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Ains~ Coun~
10 Abboud, Ed~ <NA>  M3K ~ 150   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Augi~ Coun~
# ... with 11,026 more rows, 1 more variable: ...13 <chr>, and abbreviated
#   variable name
#   1: `2014 Municipal Election - List of Contributors to Councillor Candidates`

$`4_TDSB_Trustee_Contributions_2014_election.xls`
# A tibble: 1,057 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 1320215 On~ 8 Ga~ M2M ~ 180   Mone~ <NA>  Corp~ <NA>  Kahn~ Kahn~ Mart~ Toro~
 3 1745573 On~ 630 ~ L4K ~ 200   Mone~ <NA>  Corp~ <NA>  Katz~ Katz~ Mart~ Toro~
 4 2006080N    1238~ M6H ~ 750   Mone~ <NA>  Corp~ <NA>  Mazi~ Mazi~ Wint~ Toro~
 5 2170331 On~ 128 ~ M4J ~ 750   Mone~ <NA>  Corp~ <NA>  Mant~ Mant~ Sara~ Toro~
 6 2214264 On~ 800 ~ L3R ~ 150   Mone~ <NA>  Corp~ <NA>  McGe~ McGe~ Torr~ Toro~
 7 2263053 On~ 885 ~ M1H ~ 500   Mone~ <NA>  Corp~ <NA>  N/A,~ N/A,~ Kand~ Toro~
 8 2418032 On~ 270 ~ L8L ~ 500   Mone~ <NA>  Corp~ <NA>  Zeid~ Zeid~ Torr~ Toro~
 9 443472 Ont~ 10 C~ M4W ~ 750   Mone~ <NA>  Corp~ <NA>  Ruth~ Ruth~ Ward~ Toro~
10 Abbas, Naz~ <NA>  M1V ~ 200   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  de D~ Toro~
# ... with 1,047 more rows, 1 more variable: ...13 <chr>, and abbreviated
#   variable name
#   1: `2014 Municipal Election - List of Contributors to TDSB Trustee Candidates`

$`5_TCDSB_Trustee_Contributions_2014_election.xls`
# A tibble: 155 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 2135784 On~ 35 C~ M6E ~ 200   Mone~ <NA>  Corp~ <NA>  Fatt~ Fatt~ Webs~ Toro~
 3 907037 Ont~ 100 ~ M6L ~ 750   Mone~ <NA>  Corp~ <NA>  Unkn~ Unkn~ Picc~ Toro~
 4 Abrenilla,~ <NA>  M1L ~ 782.~ Mone~ <NA>  Indi~ Cand~ <NA>  <NA>  Abre~ Toro~
 5 Alpuerto, ~ <NA>  L3S ~ 150   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Yang~ Toro~
 6 Alvares, D~ <NA>  M3A ~ 655.~ Mone~ <NA>  Indi~ Cand~ <NA>  <NA>  Alva~ Toro~
 7 Amaida Con~ 19 T~ M9W ~ 750   Mone~ <NA>  Corp~ <NA>  Unkn~ Unkn~ Picc~ Toro~
 8 Amalgamate~ 812 ~ M3K ~ 750   Mone~ <NA>  Trad~ <NA>  Kinn~ Kinn~ Morr~ Toro~
 9 Amalgamate~ 813 ~ M3K ~ 750   Mone~ <NA>  Trad~ <NA>  Kinn~ Mort~ Lacc~ Toro~
10 Amalgated ~ 812 ~ M3K ~ 750   Mone~ <NA>  Trad~ <NA>  Kinn~ Kinn~ Corp~ Toro~
# ... with 145 more rows, 1 more variable: ...13 <chr>, and abbreviated
#   variable name
#   1: `2014 Municipal Election - List of Contributors to TCDSB Trustee Candidates`

$`6_CSV_Trustee_Contributions_2014_election.xls`
# A tibble: 4 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 Baeta, Juli~ <NA>  M1B ~ 361   Mone~ <NA>  Indi~ Cand~ <NA>  <NA>  Baet~ Cons~
3 Baeta, Mrs   <NA>  M1B ~ 189   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>  Baet~ Cons~
4 Boudjenane,~ <NA>  M6P ~ 617   Mone~ <NA>  Indi~ Cand~ <NA>  <NA>  Boud~ Cons~
# ... with 1 more variable: ...13 <chr>, and abbreviated variable name
#   1: `2014 Municipal Election - List of Contributors to CSV Trustee Candidates`

$`7_CSDCCS_Trustee_Contributions_2014_election.xls`
# 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 Bedros, Nat~ <NA>  M9V ~ 40    Mone~ <NA>  Indi~ Cand~ <NA>  <NA>  Bedr~ Cons~
3 Bedros, Nat~ <NA>  M9V ~ 150   Mone~ <NA>  Indi~ Cand~ <NA>  <NA>  Bedr~ Cons~
4 Lutumba-Ntu~ <NA>  L6V ~ 300   Mone~ <NA>  Indi~ Cand~ <NA>  <NA>  Lutu~ Cons~
5 Lutumba-Ntu~ <NA>  L6V ~ 200   Mone~ <NA>  Indi~ Spou~ <NA>  <NA>  Lutu~ Cons~
6 Siani, Robe~ <NA>  L6X ~ 210   Mone~ <NA>  Indi~ Cand~ <NA>  <NA>  Sian~ Cons~
# ... with 1 more variable: ...13 <chr>, and abbreviated variable name
#   1: `2014 Municipal Election - List of Contributors to CSDCCS Trustee Candidates`
```

```
#res <- res |> mutate(year = str_extract(name, "202.?"))
df_id <- res |> select(id) |> pull()

df <- get_resource('5b230e92-0a22-4a15-9572-0b19cc222985')
```

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

```
df <- df['2_Mayor_Contributions_2014_election.xls'][[1]]
head(df)
```

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

```
names(df) <- df[1,]
```

```
Warning: The `value` argument of `names<-` must be a character vector as of tibble
3.0.0.
```

```
df <- df[2:dim(df)[1],1:dim(df)[2]]
```

```
df <- clean_names(df)
```

```
#df <- df %>%
 # select(-contributors_address)
```

```
head(df)
```

```
# 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. 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(df)
```

Table 1: Data summary

Name	df
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

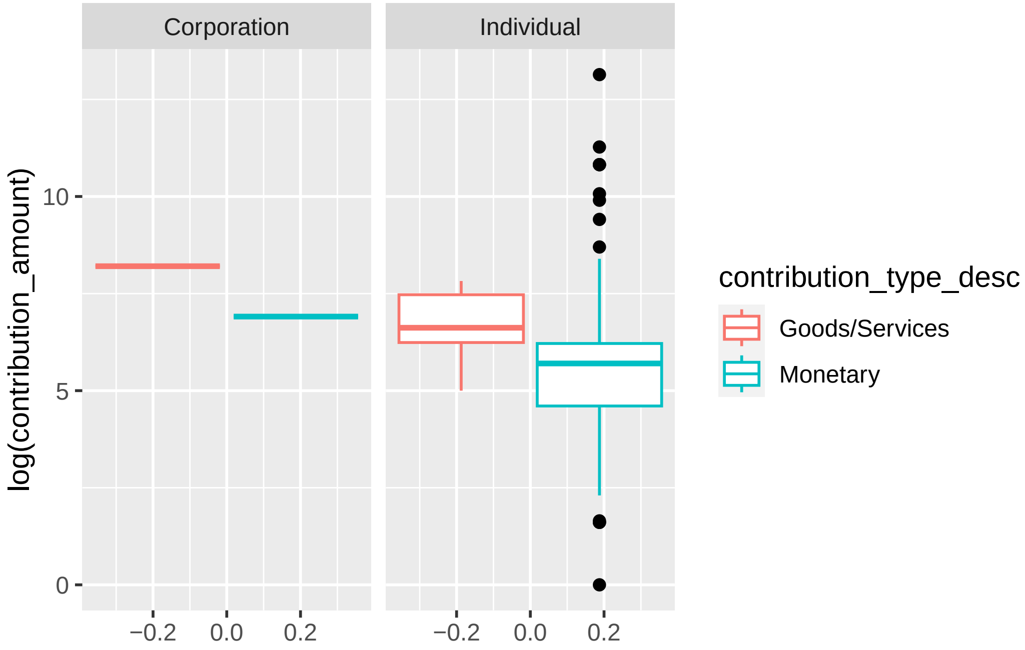
As we can see there are many missing values in the dataset. This is very worrying, as some relationships such as `relationship_to_candidate` might be very influential but we are not able to account for this influence due to a dearth of data. Note that contribution amount should be in floating point precision, so we change that.

```r
df['contribution_amount'] <- as.numeric(df$contribution_amount)
```

5. 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.

First, let's look at outliers on a log-scale.

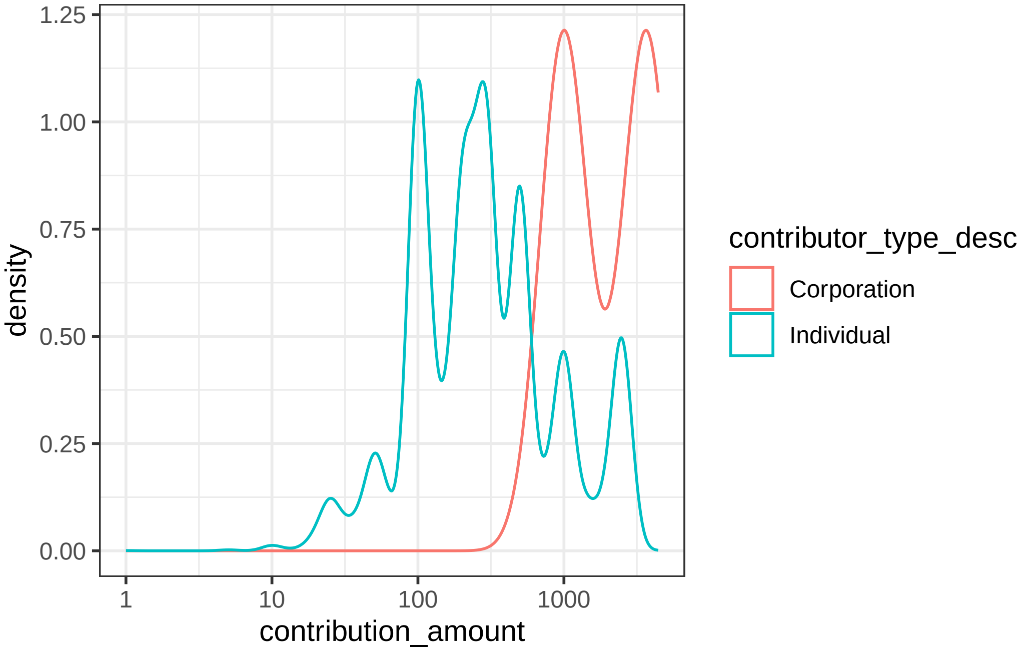
```
df %>%
  ggplot(aes(y = log(contribution_amount), color = contribution_type_desc)) +
  geom_boxplot(outlier.color = 'black', outlier.shape = 16, outlier.size = 2, notch = FALS
  facet_wrap(~contributor_type_desc)
```



There are a lot! Notice that all of these appear to be donated by individuals rather than corporations. This could be because corporations are limited by how much they can legally donate (so they might have large contributions but not outlying large contributions). In addition, they are all monetary donations rather than goods and services.

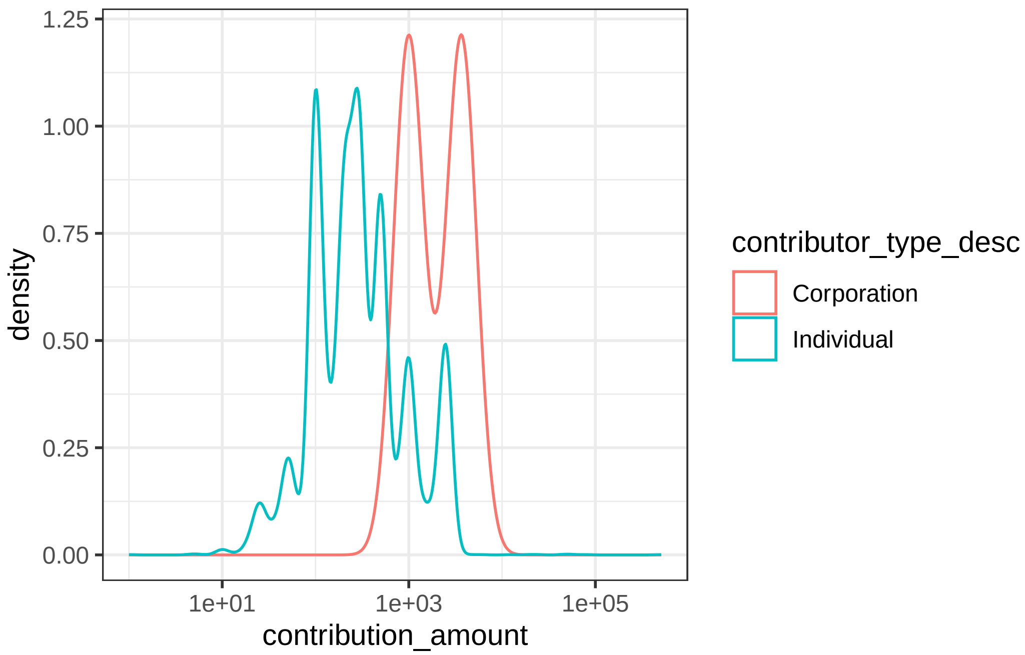
Let's plot the contribution amount without these outliers. We see that corporations tend to contribute more on average!

```
df %>%
  filter(between(contribution_amount, mean(contribution_amount, na.rm=TRUE) - (1.0 * sd(co
  ggplot() +
  geom_density(aes(x = contribution_amount, color = contributor_type_desc)) +
  scale_x_log10() +
  theme_bw()
```



For context, here is without outlier removal.

```
df %>%
  ggplot() +
  geom_density(aes(x = contribution_amount, color = contributor_type_desc)) +
  scale_x_log10() +
  theme_bw()
```



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

    - total contributions
    - mean contribution
    - number of contributions

```
df %>%
  group_by(contributors_name) %>%
  summarize(total_contr = sum(contribution_amount), mean_contr = mean(contribution_amount)
  arrange(-total_contr) %>%
  head(5)
```

```
# A tibble: 5 x 4
  contributors_name  total_contr mean_contr num_contr
  <chr>                    <dbl>      <dbl>     <int>
1 Ford, Doug             561225.    140306.         4
2 Ford, Rob              213139.     30448.         7
3 Goldkind, Ari           23624.     23624.         1
4 Thomson, Sarah           6926.      3463.         2
5 Pappalardo, Victor       6300       2100          3
```

```
df %>%
  group_by(contributors_name) %>%
  summarize(total_contr = sum(contribution_amount), mean_contr = mean(contribution_amount)
  arrange(-mean_contr) %>%
  head(5)
```

```
# A tibble: 5 x 4
  contributors_name total_contr mean_contr num_contr
  <chr>                   <dbl>      <dbl>     <int>
1 Ford, Doug            561225.    140306.         4
2 Ford, Rob             213139.     30448.         7
3 Goldkind, Ari          23624.     23624.         1
4 Di Paola, Rocco         6000       6000          1
5 kindred's Muze            3660       3660          1
```

```
df %>%
  group_by(contributors_name) %>%
  summarize(total_contr = sum(contribution_amount), mean_contr = mean(contribution_amount)
  arrange(-num_contr) %>%
  head(5)
```

```
# A tibble: 5 x 4
  contributors_name    total_contr mean_contr num_contr
  <chr>                      <dbl>      <dbl>     <int>
1 Italiano, Rob                751       62.6        12
2 Cranston, Jacqueline        2718      272.         10
3 Henery, Marjorie             900      112.          8
4 Martin, Martha               900      112.          8
5 Quin, Derek                 1350      169.          8
```

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

```
df %>%
  group_by(contributors_name) %>%
  summarize(candidate, total_contr = sum(contribution_amount), mean_contr = mean(contribut
  filter(candidate != contributors_name) %>%
  arrange(-total_contr) %>%
  distinct(contributors_name) %>%
  head(5)
```

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

```
# A tibble: 5 x 1
# Groups:   contributors_name [5]
  contributors_name
  <chr>
1 Ford, Doug
2 Pappalardo, Victor
3 Block, Sheila
4 Gazzola, Vern
5 Bachir, Salah
```

```r
df %>%
  group_by(contributors_name) %>%
  summarize(candidate, total_contr = sum(contribution_amount), mean_contr = mean(contribut
  filter(candidate != contributors_name) %>%
  arrange(-mean_contr) %>%
  distinct(contributors_name) %>%
  head(5)
```

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

```
# A tibble: 5 x 1
# Groups:   contributors_name [5]
  contributors_name
  <chr>
1 Ford, Doug
2 kindred's Muze
3 Achber, Vernon
4 Adam, Michael
5 Aghaei, Saeid
```

```r
df %>%
  group_by(contributors_name) %>%
  summarize(candidate, total_contr = sum(contribution_amount), mean_contr = mean(contribut
  filter(candidate != contributors_name) %>%
  arrange(-num_contr) %>%
  distinct(contributors_name) %>%
    head(5)
```

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

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

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

```r
df %>%
  group_by(contributors_name) %>%
  summarize(num_donation = length(unique(candidate))) %>%
  filter(num_donation > 1) %>%
  dim()
```

```
[1] 184   2
```

So 184 contributors gave money to more than one candidate.