Lab2

Haoluan Chen

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

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`

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

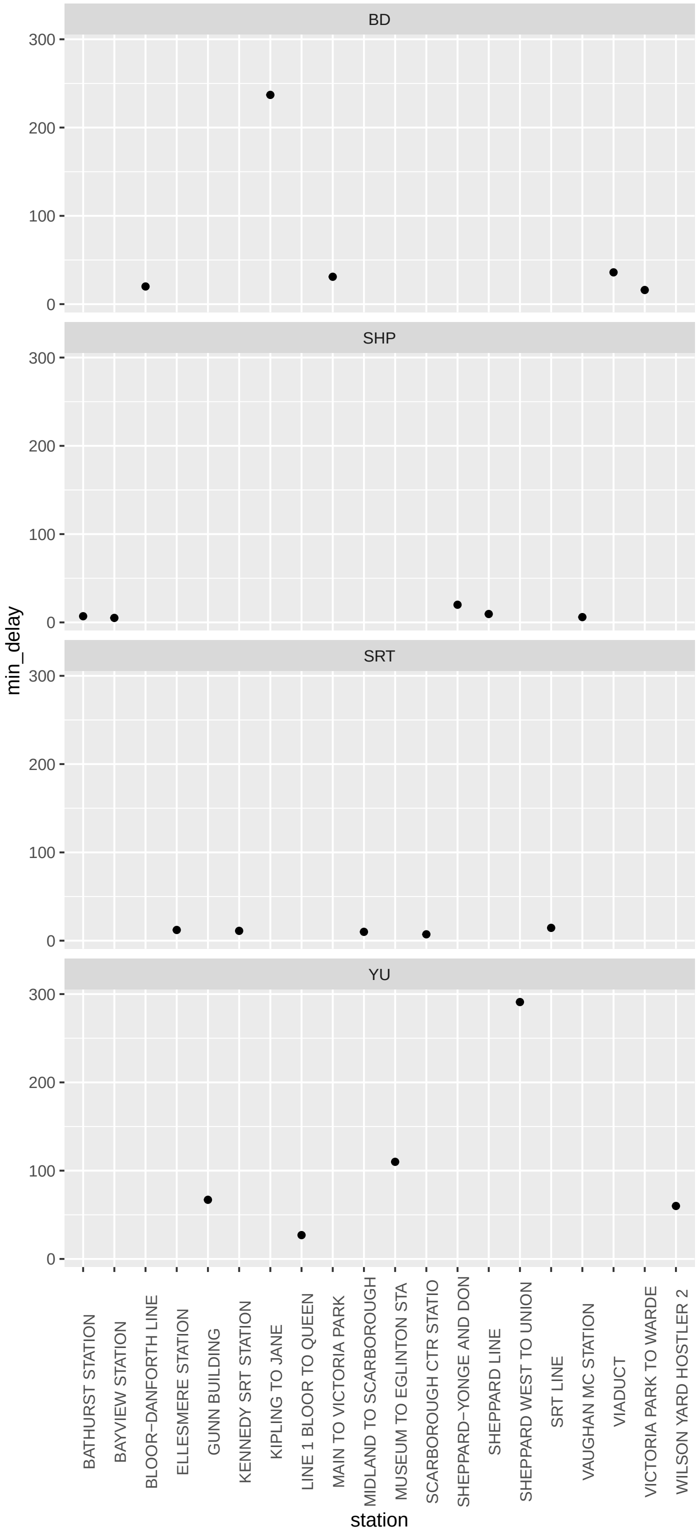
```
# A tibble: 6 x 10
  date                time  day      station   code  min_d~1 min_gap bound line
  <dttm>              <chr> <chr>    <chr>     <chr>   <dbl>   <dbl> <chr> <chr>
1 2022-01-01 00:00:00 15:59 Saturday LAWRENCE~ SRDP        0       0 N     SRT
2 2022-01-01 00:00:00 02:23 Saturday SPADINA ~ 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
```

```
q1 <- delay_2022 %>% group_by(station, line) %>%
  summarise(across(min_delay,mean),  .groups = "keep")
```

```
q1 %>% group_by(line) %>% top_n(n = 5, wt = min_delay)
```

```
# A tibble: 20 x 3
# Groups:   line [4]
   station                line  min_delay
   <chr>                  <chr>     <dbl>
 1 BATHURST STATION       SHP        7
 2 BAYVIEW STATION        SHP        5.09
 3 BLOOR-DANFORTH LINE    BD        20
 4 ELLESMERE STATION      SRT       12.1
 5 GUNN BUILDING          YU        67
 6 KENNEDY SRT STATION    SRT       11.1
 7 KIPLING TO JANE        BD       237
 8 LINE 1 BLOOR TO QUEEN  YU        27
 9 MAIN TO VICTORIA PARK  BD        31
10 MIDLAND TO SCARBOROUGH SRT       10
11 MUSEUM TO EGLINTON STA YU       110
12 SCARBOROUGH CTR STATIO SRT        7.2
13 SHEPPARD-YONGE AND DON SHP       20
14 SHEPPARD LINE          SHP        9.5
15 SHEPPARD WEST TO UNION YU       291
16 SRT LINE               SRT       14.5
17 VAUGHAN MC STATION     SHP        6
18 VIADUCT                BD        36
19 VICTORIA PARK TO WARDE BD        16
20 WILSON YARD HOSTLER 2  YU        60
```

```r
q1 %>% group_by(line) %>% top_n(n = 5, wt = min_delay) %>%
  ggplot(aes(x=station, min_delay)) +
  geom_point() +
  facet_wrap(~line, ncol = 1) +
  theme(axis.text.x = element_text(angle = 90))
```


BD
SHP
SRT
YU
300
200
100
0
min_delay
station
BATHURST STATION
BAYVIEW STATION
BLOOR−DANFORTH LINE
ELLESMERE STATION
GUNN BUILDING
KENNEDY SRT STATION
KIPLING TO JANE
LINE 1 BLOOR TO QUEEN
MAIN TO VICTORIA PARK
MIDLAND TO SCARBOROUGH
MUSEUM TO EGLINTON STA
SCARBOROUGH CTR STATIO
SHEPPARD−YONGE AND DON
SHEPPARD LINE
SHEPPARD WEST TO UNION
SRT LINE
VAUGHAN MC STATION
VIADUCT
VICTORIA PARK TO WARDE
WILSON YARD HOSTLER 2

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_
+ 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 cont
```

```r
all_data <- list_packages(limit = 500)
res <- list_package_resources("f6651a40-2f52-46fc-9e04-b760c16edd5c")
campaign_id <- res %>% filter(name == "campaign-contributions-2014-data") %>%
  select(id)
campaign <- get_resource(campaign_id)
campaign <- campaign[["2_Mayor_Contributions_2014_election.xls"]]
campaign
```

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

```r
campaign <- row_to_names(campaign, 1) %>% clean_names()
campaign
```

```
# A tibble: 10,199 x 13
   contributor~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, ~ <NA>    M6A 1P5 300     Moneta~ <NA>    Indivi~ <NA>    <NA>
 2 A Strazar, M~ <NA>    M2M 3B8 300     Moneta~ <NA>    Indivi~ <NA>    <NA>
 3 A'Court, K S~ <NA>    M4M 2J8 36      Moneta~ <NA>    Indivi~ <NA>    <NA>
 4 A'Court, K S~ <NA>    M4M 2J8 100     Moneta~ <NA>    Indivi~ <NA>    <NA>
 5 A'Court, K S~ <NA>    M4M 2J8 100     Moneta~ <NA>    Indivi~ <NA>    <NA>
 6 Aaron, Rober~ <NA>    M6B 1H7 250     Moneta~ <NA>    Indivi~ <NA>    <NA>
 7 Abadi, Babak  <NA>    M5S 2W7 500     Moneta~ <NA>    Indivi~ <NA>    <NA>
 8 Abadi, Babak  <NA>    M5S 2W7 500     Moneta~ <NA>    Indivi~ <NA>    <NA>
 9 Abadi, David  <NA>    M5S 2W7 300     Moneta~ <NA>    Indivi~ <NA>    <NA>
10 Abate, Frank  <NA>    L4H 2K7 150     Moneta~ <NA>    Indivi~ <NA>    <NA>
# ... with 10,189 more rows, 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(campaign)
```

Table 1: Data summary

Name	campaign
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

There are missing values, for example, 10197 out of 10199 rows of Contributor's Address, President/ Business Manager, Authorized Representative is missing. Also, we don't have any observations for Ward. Additionally, goods_or_service_desc has 10188 missing value and relationship_to_candidate has 10166 missing value.

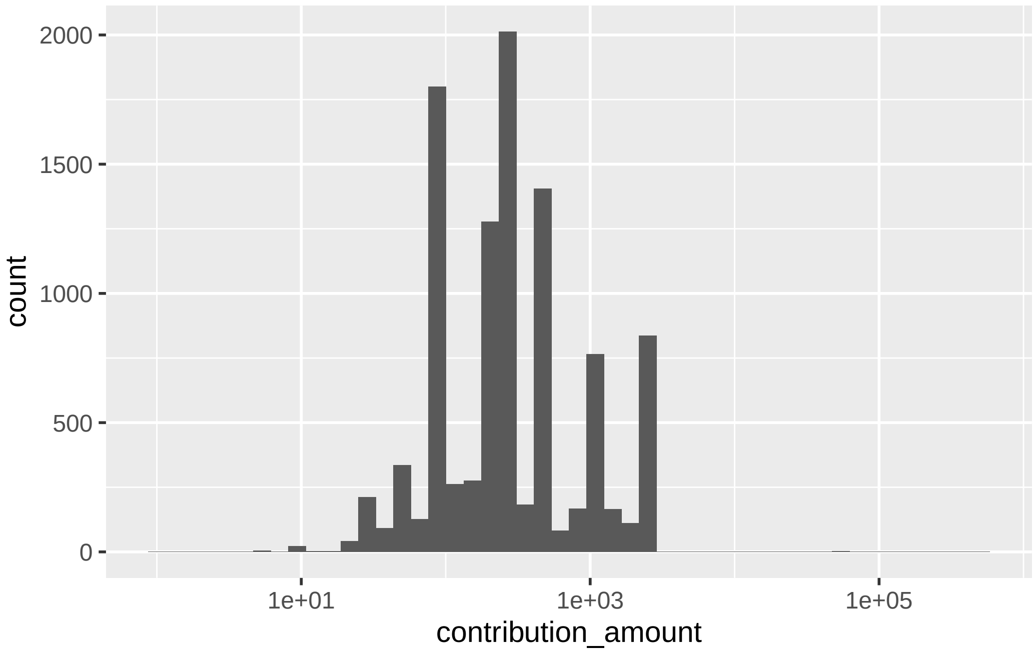
We don't need to worry about it unless we are interested in these variable. In our case, we are interested in the contribution amount, which does not have missing value. However, we also need to pay attention to the missing values that may have meaning to it. For example the missing in relationship_to_candidate may mean that there is no relationship between the contributor and the candidate. The Contribution Amount is character format, but it should be in numeric format.

```
campaign <- campaign %>%
  mutate('contribution_amount' = as.numeric(`contribution_amount`))
campaign
```

```
# A tibble: 10,199 x 13
   contributor~1 contr~2 contr~3 contr~4 contr~5 goods~6 contr~7 relat~8 presi~9
   <chr>         <chr>   <chr>     <dbl> <chr>   <chr>   <chr>   <chr>   <chr>
 1 A D'Angelo, ~ <NA>    M6A 1P5     300 Moneta~ <NA>    Indivi~ <NA>    <NA>
 2 A Strazar, M~ <NA>    M2M 3B8     300 Moneta~ <NA>    Indivi~ <NA>    <NA>
 3 A'Court, K S~ <NA>    M4M 2J8      36 Moneta~ <NA>    Indivi~ <NA>    <NA>
 4 A'Court, K S~ <NA>    M4M 2J8     100 Moneta~ <NA>    Indivi~ <NA>    <NA>
 5 A'Court, K S~ <NA>    M4M 2J8     100 Moneta~ <NA>    Indivi~ <NA>    <NA>
 6 Aaron, Rober~ <NA>    M6B 1H7     250 Moneta~ <NA>    Indivi~ <NA>    <NA>
 7 Abadi, Babak  <NA>    M5S 2W7     500 Moneta~ <NA>    Indivi~ <NA>    <NA>
 8 Abadi, Babak  <NA>    M5S 2W7     500 Moneta~ <NA>    Indivi~ <NA>    <NA>
 9 Abadi, David  <NA>    M5S 2W7     300 Moneta~ <NA>    Indivi~ <NA>    <NA>
10 Abate, Frank  <NA>    L4H 2K7     150 Moneta~ <NA>    Indivi~ <NA>    <NA>
# ... with 10,189 more rows, 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
```

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.

```
campaign %>% ggplot(aes(x = contribution_amount)) +
  geom_histogram(bins = 48) + scale_x_log10()
```

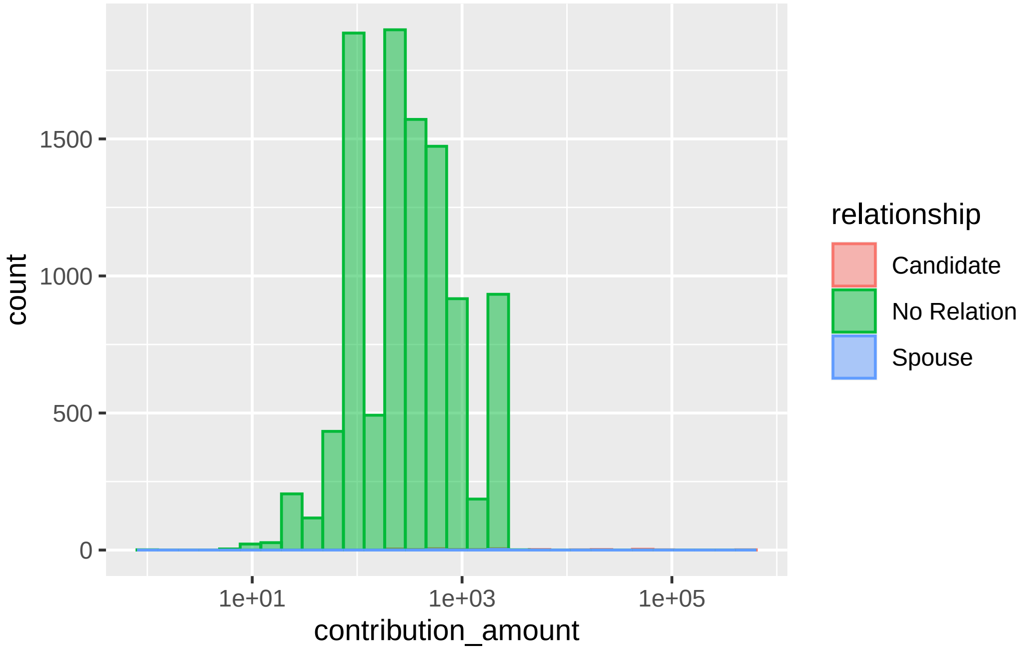


```
unique(campaign$relationship_to_candidate)
```

```
[1] NA          "Candidate" "Spouse"
```

```
campaign %>% mutate(relationship =
                      if_else(is.na(relationship_to_candidate),
                              "No Relation", relationship_to_candidate)) %>%
  ggplot(aes(x=contribution_amount, fill = relationship, color = relationship)) +
  geom_histogram(position="identity", alpha=0.5) + scale_x_log10()
```

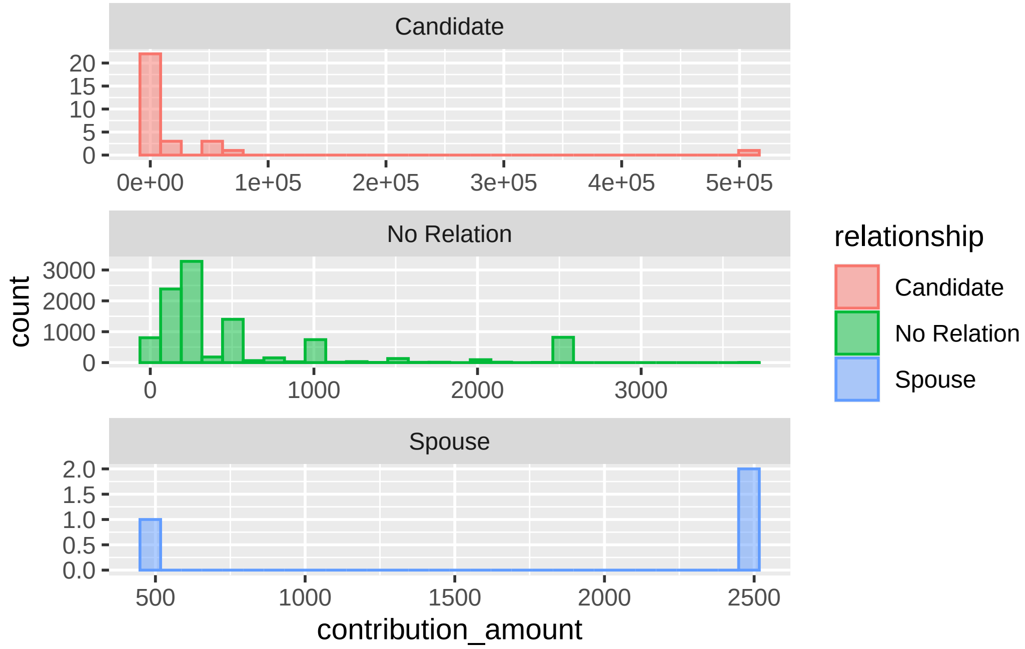
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We can see that the candidates are contributing a large amount of contribution. Lets split the histogram and look into the relationship separately.

```
campaign %>% mutate(relationship =
                      if_else(is.na(relationship_to_candidate),
                              "No Relation", relationship_to_candidate)) %>%
  ggplot(aes(x=contribution_amount, fill = relationship, color = relationship)) +
  geom_histogram(position="identity", alpha=0.5) +
  facet_wrap(~relationship, ncol = 1, scales = "free")
```

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



```r
campaign %>%
  filter(relationship_to_candidate == "Candidate") %>%
  arrange(desc(contribution_amount))
```

```
# A tibble: 30 x 13
   contributor~1 contr~2 contr~3 contr~4 contr~5 goods~6 contr~7 relat~8 presi~9
   <chr>         <chr>   <chr>     <dbl> <chr>   <chr>   <chr>   <chr>   <chr>
 1 Ford, Doug    <NA>    M9A 2C3 508225. Moneta~ <NA>    Indivi~ Candid~ <NA>
 2 Ford, Rob     <NA>    M9A 3G9  78805. Moneta~ <NA>    Indivi~ Candid~ <NA>
 3 Ford, Doug    <NA>    M9A 2C3  50000  Moneta~ <NA>    Indivi~ Candid~ <NA>
 4 Ford, Rob     <NA>    M9A 3G9  50000  Moneta~ <NA>    Indivi~ Candid~ <NA>
 5 Ford, Rob     <NA>    M9A 3G9  50000  Moneta~ <NA>    Indivi~ Candid~ <NA>
 6 Goldkind, Ari <NA>    M5P 1P5  23624. Moneta~ <NA>    Indivi~ Candid~ <NA>
 7 Ford, Rob     <NA>    M9A 3G9  20000  Moneta~ <NA>    Indivi~ Candid~ <NA>
 8 Ford, Rob     <NA>    M9A 3G9  12210  Moneta~ <NA>    Indivi~ Candid~ <NA>
 9 Di Paola, Ro~ <NA>    M3H 2T1   6000  Moneta~ <NA>    Indivi~ Candid~ <NA>
10 Thomson, Sar~ <NA>    M4W 2X6   4426. Moneta~ <NA>    Indivi~ Candid~ <NA>
# ... with 20 more rows, 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
```

The majority of the data contributes range from 0 to about 2500, There are only three contributions from spouse, and they are at the two extreme, one spouse contributed 500 and other two contributed 2500. However, looking at the the candidates, they are contributing a lot of money(outlines) with the highest amount of 508224.73.

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

```
Q6 <- campaign %>%
  group_by(candidate) %>%
  summarize(total_contributions = sum(contribution_amount),
  mean_contributions = mean(contribution_amount),
  number_contributions = n())
```

total contributions

```
Q6 %>% arrange(desc(total_contributions)) %>% head(5)
```

```
# A tibble: 5 x 4
  candidate     total_contributions mean_contributions number_contributions
  <chr>                       <dbl>              <dbl>                <int>
1 Tory, John               2767869.              1064.                 2602
2 Chow, Olivia             1638266.               287.                 5708
3 Ford, Doug                889897.              1456.                  611
4 Ford, Rob                 387648.               721.                  538
5 Stintz, Karen             242805                995.                  244
```

mean contribution

```
Q6 %>% arrange(desc(mean_contributions)) %>% head(5)
```

```
# A tibble: 5 x 4
  candidate        total_contributions mean_contributions number_contributions
  <chr>                          <dbl>              <dbl>                <int>
1 Sniedzins, Erwin               8100               2025                     4
2 Syed, Hïmy                     2018               2018                     1
3 Ritch, Carlie                  5660               1887.                    3
4 Ford, Doug                   889897.              1456.                  611
5 Clarke, Kevin                  1200               1200                     1
```

number of contributions

```
Q6 %>% arrange(desc(number_contributions)) %>% head(5)
```

```
# A tibble: 5 x 4
  candidate       total_contributions mean_contributions number_contributions
  <chr>                         <dbl>              <dbl>                <int>
1 Chow, Olivia               1638266.               287.                 5708
2 Tory, John                 2767869.              1064.                 2602
3 Ford, Doug                  889897.              1456.                  611
4 Ford, Rob                   387648.               721.                  538
5 Soknacki, David              132431               422.                  314
```

7. Repeat 6 but without contributions from the candidates themselves.

```
Q7 <- campaign %>%
  filter(relationship_to_candidate == "Spouse" |
           is.na(relationship_to_candidate)) %>%
  group_by(candidate) %>%
  summarize(total_contributions = sum(contribution_amount),
  mean_contributions = mean(contribution_amount),
  number_contributions = n())
```

total contributions

```
Q7 %>% arrange(desc(total_contributions)) %>% head(5)
```

```
# A tibble: 5 x 4
  candidate     total_contributions mean_contributions number_contributions
  <chr>                       <dbl>              <dbl>                <int>
1 Tory, John                2765369.             1063.                 2601
2 Chow, Olivia              1635766.              287.                 5707
3 Ford, Doug                 331173.              545.                  608
4 Stintz, Karen              242805               995.                  244
5 Ford, Rob                  174510.              329.                  531
```

mean contribution

```
Q7 %>% arrange(desc(mean_contributions)) %>% head(5)
```

```
# A tibble: 5 x 4
  candidate         total_contributions mean_contributions number_contributions
  <chr>                           <dbl>              <dbl>                <int>
1 Ritch, Carlie                    5660              1887.                    3
2 Sniedzins, Erwin                 5600              1867.                    3
3 Tory, John                    2765369.             1063.                 2601
4 Gardner, Norman                  3000              1000                     3
5 Tiwari, Ramnarine                1000              1000                     1
```

number of contributions

```
Q7 %>% arrange(desc(number_contributions)) %>% head(5)
```

```
# A tibble: 5 x 4
  candidate       total_contributions mean_contributions number_contributions
  <chr>                         <dbl>              <dbl>                <int>
1 Chow, Olivia                1635766.              287.                 5707
2 Tory, John                  2765369.             1063.                 2601
3 Ford, Doug                   331173.              545.                  608
4 Ford, Rob                    174510.              329.                  531
5 Soknacki, David              132431               422.                  314
```

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

```
campaign %>%
  select(contributors_name, candidate) %>%
  distinct() %>%
  group_by(contributors_name) %>%
  summarize(num_candidates = n()) %>%
  filter(num_candidates > 1)
```

```
# A tibble: 184 x 2
   contributors_name num_candidates
   <chr>                      <int>
 1 Abadi, Babak                   2
 2 Adams, Michael                 2
 3 Anga, John                     2
 4 Argyris, Katerina              2
 5 Atkinson, Tom                  2
 6 Aziz, Peter                    2
 7 Bachir, Salah                  2
 8 Bajwa, Joginder                2
 9 Baker, Norma                   2
10 Banwait, Rav                   2
# ... with 174 more rows
```

184 contributors gave money to more than one candidate.