Lab Exercise - Jan15th - EDA Data Visualization

Luis Correia - Student No. 1006508566

January 15th 2020

Lab Exercises

To be handed in via submission of Rmd file to GitHub by Thursday 16 January, 5pm.

Setup libraries to be used in this exercise

```
library(opendatatoronto)
library(tidyverse)
library(skimr)
library(visdat)
library(janitor)
```

Question 1

Using the `opendatatoronto` package, download the data on mayoral campaign contributions for 2014. (note: the 2014 file you will get from `get_resource`, so just keep the sheet that relates to the Mayor election).

a. Get list of data-bases and download the Election Database

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

```
## # A tibble: 6 x 10
##   title id    topics civic_issues excerpt dataset_category num_resources formats
##   <chr> <chr> <chr>  <chr>        <chr>   <chr>                    <int> <chr>
## 1 Traf... ae4e... Trans... Mobility     This d... Document                     2 XLS,XL...
## 2 Deve... 4443... Finan... Fiscal resp... Develo... Document                     4 XLSX
## 3 Body... c405... City ... <NA>         This d... Table                        2 XML,WE...
## 4 Stre... 1db3... City ... Mobility     Transi... Map                          1 SHP,CS...
## 5 Stre... 74f6... City ... <NA>         Public... Map                          1 SHP,CS...
## 6 Stre... 821f... City ... <NA>         Public... Map                          1 SHP,CS...
## # ... with 2 more variables: refresh_rate <chr>, last_refreshed <date>
```

```
list_package_resources("f6651a40-2f52-46fc-9e04-b760c16edd5c")
```

```
## # A tibble: 2 x 4
##   name                          id                          format last_modified
##   <chr>                         <chr>                       <chr>  <date>
## 1 campaign-contributions-2014-... d99bb1f3-949a-4497-bb96-c9... ZIP    2019-07-23
## 2 campaign-contributions-2014-... 7c05def5-b39d-44cb-a163-0d... XLS    2019-07-23
```

```
mcc_data <- get_resource("d99bb1f3-949a-4497-bb96-c93bbd203130")
```

b. Retain just the contribution data for mayoral campaign...

```
mcc_data <- mcc_data$`2_Mayor_Contributions_2014_election.xls`
mcc_data
```

```
## # A tibble: 10,200 x 13
##    `2014 Municipal... ...2  ...3  ...4  ...5  ...6  ...7  ...8  ...9  ...10 ...11
##    <chr>            <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
##  1 Contributor's N... Cont... Cont... Cont... Cont... Good... Cont... Rela... Pres... Auth... Cand..
##  2 A D'Angelo, Tul... <NA>  M6A ... 300   Mone... <NA>  Indi... <NA>  <NA>  <NA>  Ford...
##  3 A Strazar, Mart... <NA>  M2M ... 300   Mone... <NA>  Indi... <NA>  <NA>  <NA>  Ford...
##  4 A'Court, K Susan <NA>  M4M ... 36    Mone... <NA>  Indi... <NA>  <NA>  <NA>  Chow...
##  5 A'Court, K Susan <NA>  M4M ... 100   Mone... <NA>  Indi... <NA>  <NA>  <NA>  Chow...
##  6 A'Court, K Susan <NA>  M4M ... 100   Mone... <NA>  Indi... <NA>  <NA>  <NA>  Chow...
##  7 Aaron, Robert B  <NA>  M6B ... 250   Mone... <NA>  Indi... <NA>  <NA>  <NA>  Tory...
##  8 Abadi, Babak     <NA>  M5S ... 500   Mone... <NA>  Indi... <NA>  <NA>  <NA>  Tory...
##  9 Abadi, Babak     <NA>  M5S ... 500   Mone... <NA>  Indi... <NA>  <NA>  <NA>  Chow...
## 10 Abadi, David     <NA>  M5S ... 300   Mone... <NA>  Indi... <NA>  <NA>  <NA>  Stin...
## # ... with 10,190 more rows, and 2 more variables: ...12 <chr>, ...13 <chr>
```

Question 2

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

a. Fix 1st row problem which contains the column names using janitor and then cleans-up the column names

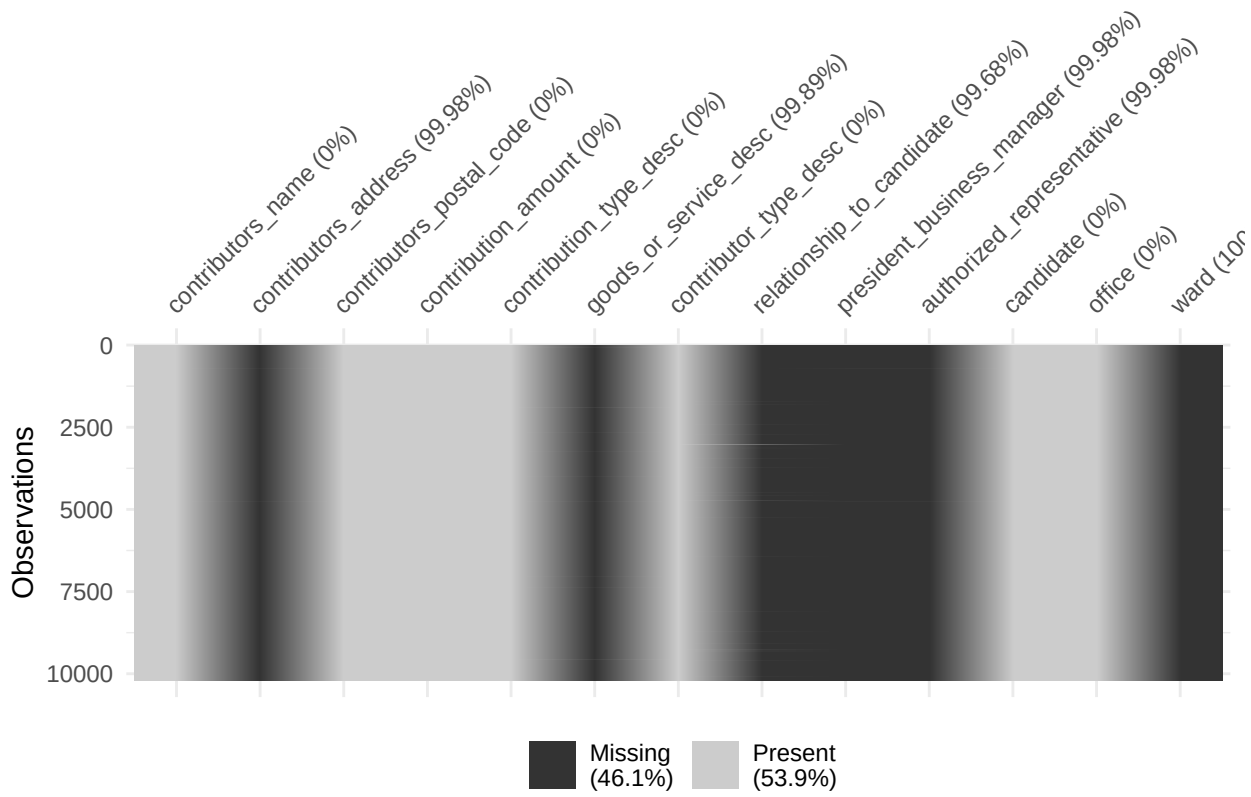
```
mcc_data <- mcc_data %>%
  row_to_names(row_number = 1,remove_row = TRUE)

mcc_data <- clean_names(mcc_data)
```

Question 3

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.

a. Verify if there are NAs present in the database and check if they represent problems

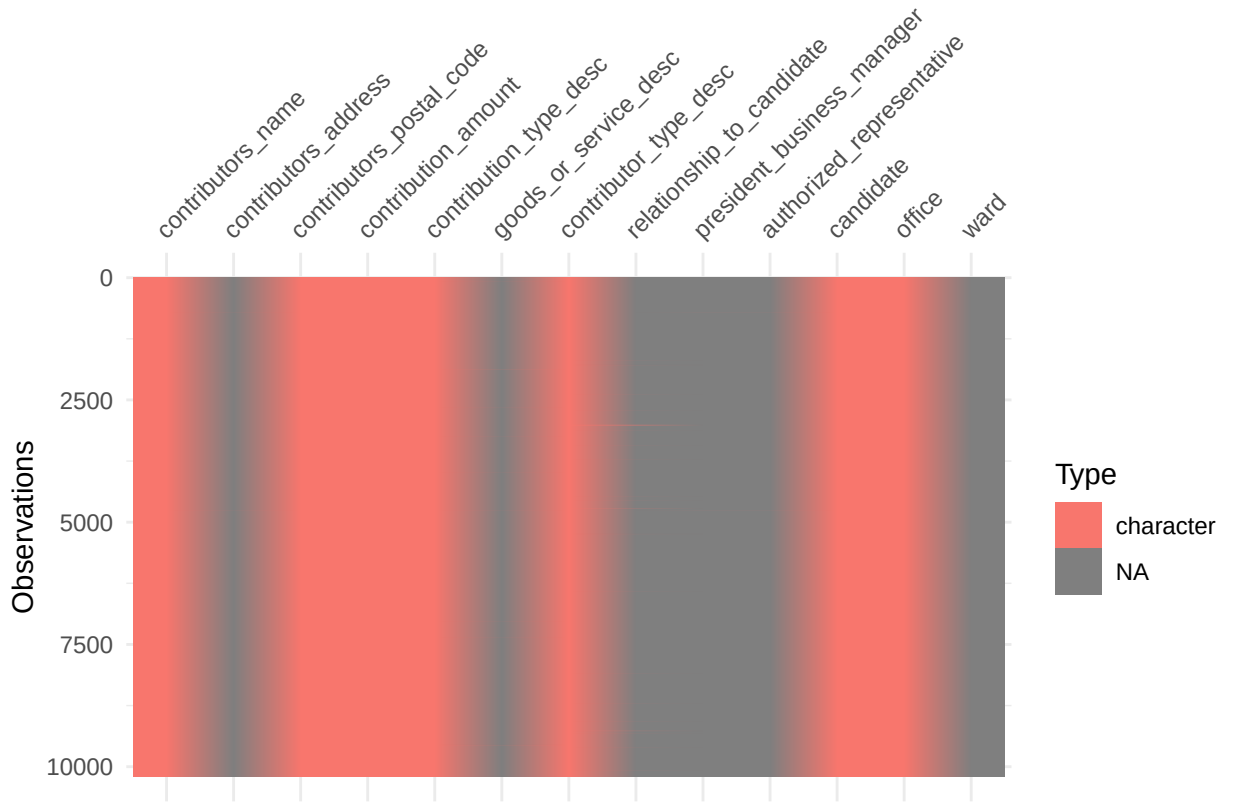
```
vis_miss(mcc_data)
```



Apparently there are no important variables with missing data, assuming that a missing in 'relationship_to_candidate' means the contribution was done by persons really not related w/ those candidates.

b. Skim the data for visualization and identify which columns should be converted

```
vis_dat(mcc_data)
```



c. Transform contributions to numeric and stores into 'amount' column

```
mcc_data <- mcc_data %>%
  mutate(amount = as.numeric(contribution_amount))

skim_without_charts(mcc_data)
```

Table 1: Data summary

Name	mcc_data
Number of rows	10199
Number of columns	14
Column type frequency:	
character	13
numeric	1
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

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
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

Variable type: numeric

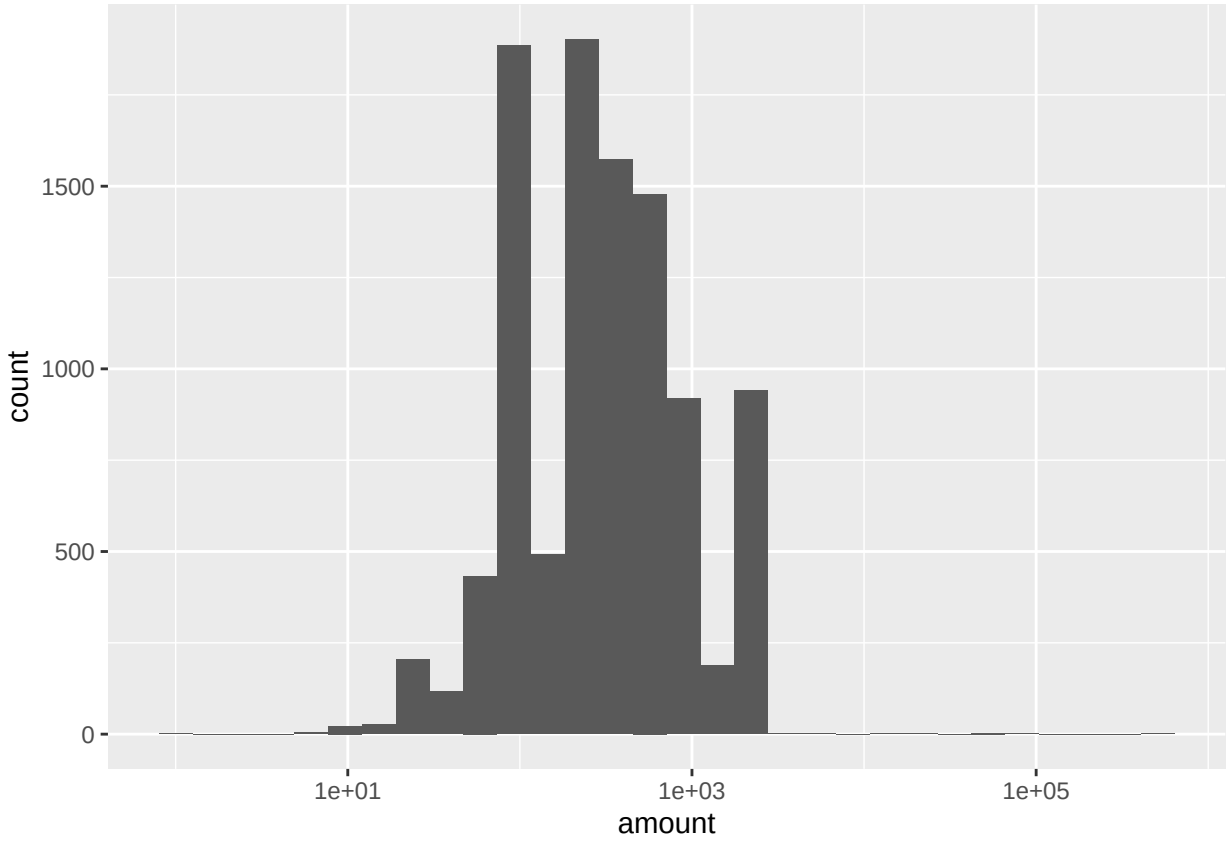
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
amount	0	1	607.95	5211.31	1	100	300	500	508224.7

Question 4

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.

a. Vizualize data

```
ggplot(mcc_data, aes(x = amount)) + geom_histogram() + scale_x_log10()
```

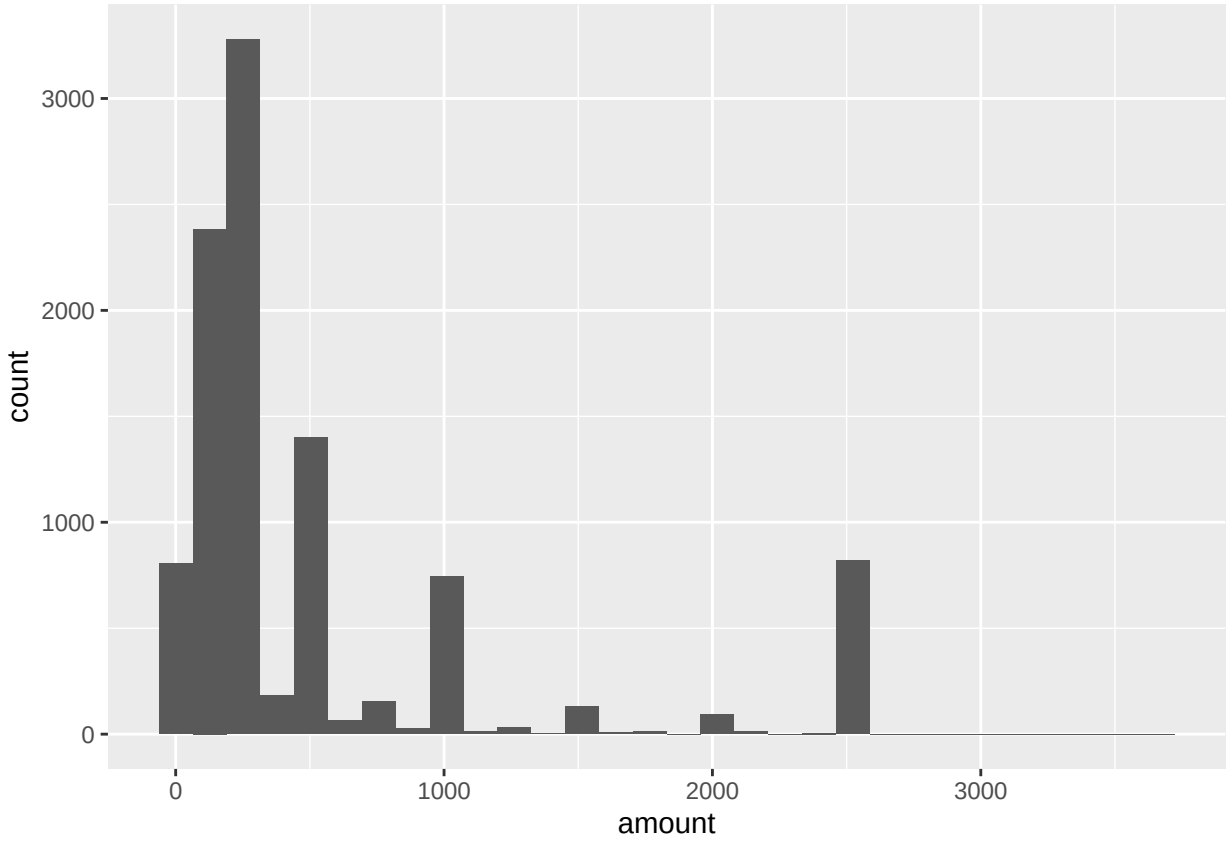


- Shows the contribution amount per contributor in decreasing order

```
mcc_data %>%
  arrange(-amount) %>%
  select(contributors_name, relationship_to_candidate, candidate, amount)
```

```
## # A tibble: 10,199 x 4
##    contributors_name relationship_to_candidate candidate        amount
##    <chr>             <chr>                     <chr>             <dbl>
##  1 Ford, Doug        Candidate                 Ford, Doug      508225.
##  2 Ford, Rob         Candidate                 Ford, Rob        78805.
##  3 Ford, Doug        Candidate                 Ford, Doug       50000
##  4 Ford, Rob         Candidate                 Ford, Rob        50000
##  5 Ford, Rob         Candidate                 Ford, Rob        50000
##  6 Goldkind, Ari     Candidate                 Goldkind, Ari    23624.
##  7 Ford, Rob         Candidate                 Ford, Rob        20000
##  8 Ford, Rob         Candidate                 Ford, Rob        12210
##  9 Di Paola, Rocco   Candidate                 Di Paola, Rocco   6000
## 10 Thomson, Sarah    Candidate                 Thomson, Sarah    4426.
## # ... with 10,189 more rows
```

```
# Apparently the contribution from Doug Ford to himself (~500k) represent an outlier, lets remove them
# people with relationshio with the candidate to have a less biased view of external contributions

mcc_data %>%
  filter(is.na(relationship_to_candidate)) %>% # using contribitions received from people w/ no relatio
  ggplot(aes(x = amount)) + geom_histogram()
```



Question 5

List the top five candidates in each of these categories: + total contributions + mean contribution + number of contributions

```
mcc_data %>%
  group_by(candidate) %>%
  summarise(tot_contrib = sum(amount), mean_contrib = mean(amount), n = n()) %>%
  arrange(-tot_contrib) %>%
  slice(1:5)
```

```
## # A tibble: 5 x 4
##   candidate     tot_contrib mean_contrib     n
##   <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
```

Question 6

Repeat 5 but without contributions from the candidates themselves.

```
mcc_data %>%
  group_by(candidate) %>%
  filter(contributors_name != candidate) %>% # remove contributions from those who contributes to thems
```

```
  summarise(tot_contrib = sum(amount), mean_contrib = mean(amount), n = n()) %>%
  arrange(-tot_contrib) %>%
  slice(1:5)
```

```
## # A tibble: 5 x 4
##   candidate     tot_contrib mean_contrib     n
##   <chr>               <dbl>        <dbl> <int>
## 1 Tory, John       2765369.        1063.  2601
## 2 Chow, Olivia     1634766.         286.  5706
## 3 Ford, Doug        331173.         545.   608
## 4 Stintz, Karen     242805          995.   244
## 5 Ford, Rob         174510.         329.   531
```

Question 7

How many contributors gave money to more than one candidate?

```
ct <- mcc_data %>%
  group_by(contributors_name) %>%  # identifies and group all contributors
  arrange(candidate) %>%  # Arrange the number of candidates they contributes for
  summarise(n = n()) %>%  # Summarizes the number of contributions per contributor
  filter(n > 1) %>%       # Identifies contributors who contributes to more than one candidate
  arrange(-n)
ct
```

```
## # A tibble: 1,883 x 2
##    contributors_name        n
##    <chr>                <int>
##  1 Italiano, Rob           12
##  2 Cranston, Jacqueline    10
##  3 Henery, Marjorie         8
##  4 Martin, Martha           8
##  5 Quin, Derek              8
##  6 Stewart, Carol           8
##  7 Ford, Rob                7
##  8 Lary, Debra              7
##  9 Leeson, John             7
## 10 Amodeo, Merle            6
## # ... with 1,873 more rows
```

```
cat("No. of contributors who gave money to more than one candidate: ", nrow(ct))
```

```
## No. of contributors who gave money to more than one candidate:  1883
```