Week2_lab

Siyi Wei

21/01/2021

Lab Exercises

To be handed in via submission of Rmd file to GitHub.

1. 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
   - yu 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)
all_data[grep("campaign", all_data$title, ignore.case = T), ]
```

```
## # A tibble: 5 x 10
##   title id    topics civic_issues excerpt dataset_category num_resources
##   <chr> <chr> <chr>  <chr>        <chr>   <chr>                     <int>
## 1 Elec~ 28e5~ City ~ <NA>         This d~ Document                      2
## 2 Elec~ 67d2~ Finan~ <NA>         "This ~ Document                      2
## 3 Civi~ 7d0d~ City ~ Affordable ~ "The O~ Document                      2
## 4 Elec~ 2ee8~ City ~ <NA>         This d~ Document                      2
## 5 Elec~ f665~ City ~ <NA>         This d~ Document                      2
## # ... with 3 more variables: formats <chr>, 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-201~ d99bb1f3-949a-4497-bb96~ ZIP    2019-07-23
## 2 campaign-contributions-201~ 7c05def5-b39d-44cb-a163~ XLS    2019-07-23
```

```
data_1418 <- get_resource("d99bb1f3-949a-4497-bb96-c93bbd203130")
```

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

```
## New names:
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * `` -> ...6
## * ... and 7 more problems
## New names:
## * `` -> ...2
## * `` -> ...3
```

```
## * `` -> ...4
## * `` -> ...5
## * `` -> ...6
## * ... and 7 more problems
## New names:
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * `` -> ...6
## * ... and 7 more problems
## New names:
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * `` -> ...6
## * ... and 7 more problems
## New names:
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * `` -> ...6
## * ... and 7 more problems
## New names:
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * `` -> ...6
## * ... and 7 more problems
```

```
data_Mayor <- data_1418[2]
head(data_Mayor)
```

```
## $`2_Mayor_Contributions_2014_election.xls`
## # A tibble: 10,200 x 13
##    `2014 Municipal~ ...2  ...3  ...4  ...5  ...6  ...7  ...8  ...9  ...10
##    <chr>            <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
##  1 Contributor's N~ Cont~ Cont~ Cont~ Cont~ Good~ Cont~ Rela~ Pres~ Auth~
##  2 A D'Angelo, Tul~ <NA>  M6A ~ 300   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>
##  3 A Strazar, Mart~ <NA>  M2M ~ 300   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>
##  4 A'Court, K Susan <NA>  M4M ~ 36    Mone~ <NA>  Indi~ <NA>  <NA>  <NA>
##  5 A'Court, K Susan <NA>  M4M ~ 100   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>
##  6 A'Court, K Susan <NA>  M4M ~ 100   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>
##  7 Aaron, Robert B  <NA>  M6B ~ 250   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>
##  8 Abadi, Babak     <NA>  M5S ~ 500   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>
##  9 Abadi, Babak     <NA>  M5S ~ 500   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>
## 10 Abadi, David     <NA>  M5S ~ 300   Mone~ <NA>  Indi~ <NA>  <NA>  <NA>
## # ... with 10,190 more rows, and 3 more variables: ...11 <chr>,
## #   ...12 <chr>, ...13 <chr>
```

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

```
main_data = data.frame(data_Mayor)
colnames(main_data) <- main_data[1,]
main_data <- main_data[-1, ]
main_data <- clean_names(main_data)
```

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.

Answer:

- Yes, there are lots of missing values such as contributors_address, goods or service, relationship, authorized representative and so on.
- I think we should not worry about the NA variables too much, there are various reasons. For example, some of the information are too personal that the contributors might not willing to fill. Such as the contributors address. Then for the relations and authorized representative. NA already represent some useful information like None.
- I have convert the contribution amount to numeric value.

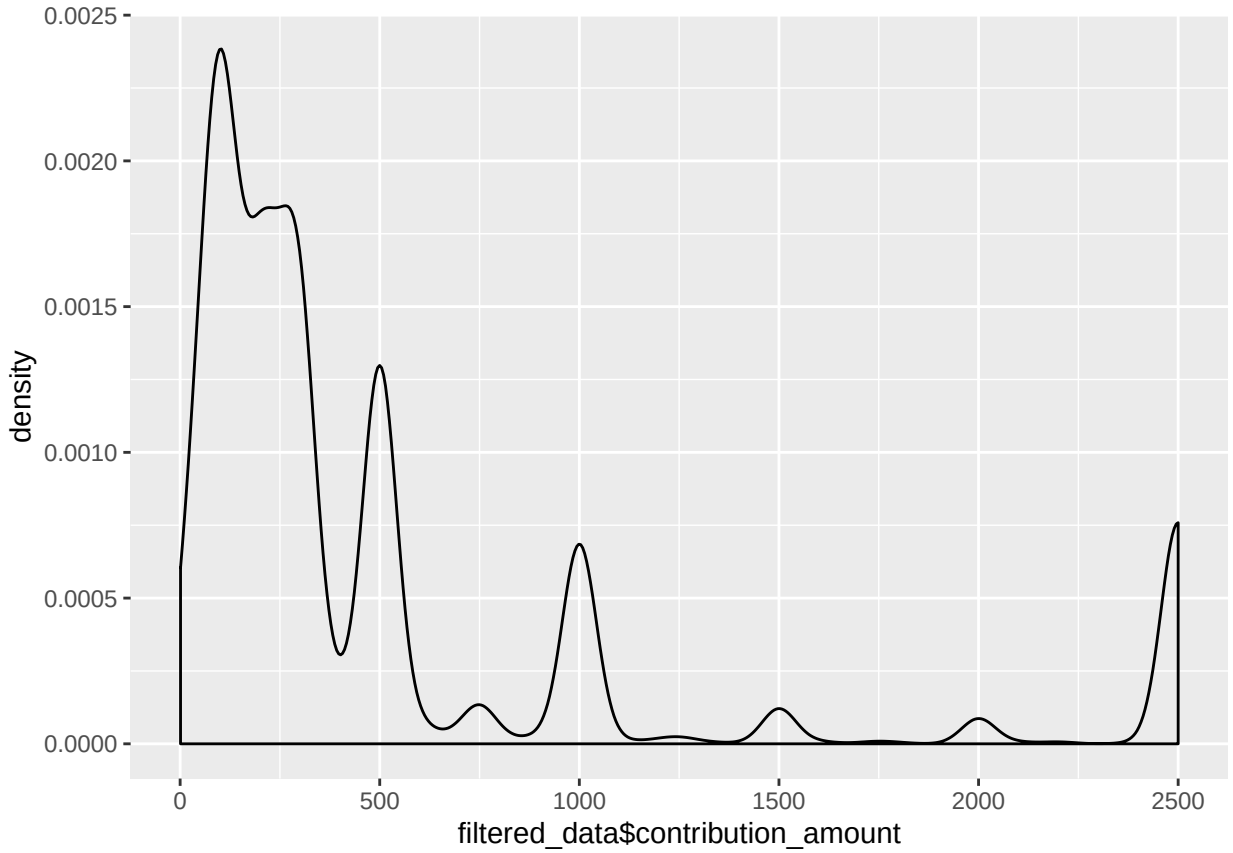
```
main_data$contribution_amount = as.numeric(main_data$contribution_amount)
```

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.

Answer:

- Amount larger than 3000 are most likely to be notable outliers.
- Most of them are contributed by the candidates themselves.

```
filtered_data <- main_data %>% filter(main_data$contribution_amount < 3000)
ggplot(filtered_data, aes(x=filtered_data$contribution_amount)) + geom_density()
```




```
main_data %>% filter(main_data$contribution_amount > 3000)
```

```
##    contributors_name       contributors_address contributors_postal_code
## 1    Di Paola, Rocco                       <NA>                  M3H 2T1
## 2         Ford, Doug                       <NA>                  M9A 2C3
## 3         Ford, Doug                       <NA>                  M9A 2C3
## 4          Ford, Rob                       <NA>                  M9A 3G9
## 5          Ford, Rob                       <NA>                  M9A 3G9
## 6          Ford, Rob                       <NA>                  M9A 3G9
## 7          Ford, Rob                       <NA>                  M9A 3G9
## 8          Ford, Rob                       <NA>                  M9A 3G9
## 9      Goldkind, Ari                       <NA>                  M5P 1P5
## 10    kindred's Muze 723 Dovercourt Rd, Toronto                  M6H 2W7
## 11    Thomson, Sarah                       <NA>                  M4W 2X6
##    contribution_amount contribution_type_desc goods_or_service_desc
## 1              6000.00               Monetary                  <NA>
## 2            508224.73               Monetary                  <NA>
## 3             50000.00               Monetary                  <NA>
## 4             20000.00               Monetary                  <NA>
## 5             50000.00               Monetary                  <NA>
## 6             50000.00               Monetary                  <NA>
## 7             78804.80               Monetary                  <NA>
## 8             12210.00               Monetary                  <NA>
## 9             23623.63               Monetary                  <NA>
## 10             3660.00         Goods/Services           photography
## 11             4425.55               Monetary                  <NA>
##    contributor_type_desc relationship_to_candidate
```

```
## 1             Individual                Candidate
## 2             Individual                Candidate
## 3             Individual                Candidate
## 4             Individual                Candidate
## 5             Individual                Candidate
## 6             Individual                Candidate
## 7             Individual                Candidate
## 8             Individual                Candidate
## 9             Individual                Candidate
## 10           Corporation                     <NA>
## 11            Individual                Candidate
##    president_business_manager authorized_representative       candidate
## 1                        <NA>                      <NA> Di Paola, Rocco
## 2                        <NA>                      <NA>      Ford, Doug
## 3                        <NA>                      <NA>      Ford, Doug
## 4                        <NA>                      <NA>       Ford, Rob
## 5                        <NA>                      <NA>       Ford, Rob
## 6                        <NA>                      <NA>       Ford, Rob
## 7                        <NA>                      <NA>       Ford, Rob
## 8                        <NA>                      <NA>       Ford, Rob
## 9                        <NA>                      <NA>   Goldkind, Ari
## 10           Pharell, Colleen          Pharell, Colleen   Ritch, Carlie
## 11                       <NA>                      <NA>  Thomson, Sarah
##    office ward
## 1   Mayor <NA>
## 2   Mayor <NA>
## 3   Mayor <NA>
## 4   Mayor <NA>
## 5   Mayor <NA>
## 6   Mayor <NA>
## 7   Mayor <NA>
## 8   Mayor <NA>
## 9   Mayor <NA>
## 10  Mayor <NA>
## 11  Mayor <NA>
```

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

```
main_data %>% group_by(candidate) %>% summarize(Total=sum(contribution_amount, na.rm = T)) %>% arrange(
```

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

```
main_data %>% group_by(candidate) %>% summarize(Mean=mean(contribution_amount, na.rm = T)) %>% arrange(
```

```
## # A tibble: 5 x 2
##   candidate         Mean
```

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

```
main_data %>% group_by(candidate) %>% summarize(Count=n()) %>% arrange(desc(Count)) %>% slice(1:5)
```

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

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

```
main_data %>% filter(contributors_name != candidate) %>% group_by(candidate) %>%
  summarize(Total=sum(contribution_amount, na.rm = T)) %>% arrange(desc(Total)) %>% slice(1:5)
```

```
## # A tibble: 5 x 2
##   candidate        Total
##   <chr>            <dbl>
## 1 Tory, John    2765369.
## 2 Chow, Olivia  1634766.
## 3 Ford, Doug     331173.
## 4 Stintz, Karen  242805
## 5 Ford, Rob      174510.
```

```
main_data %>% filter(contributors_name != candidate) %>% group_by(candidate) %>%
  summarize(Mean=mean(contribution_amount, na.rm = T)) %>% arrange(desc(Mean)) %>% slice(1:5)
```

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

```
main_data %>% filter(contributors_name != candidate) %>% group_by(candidate) %>%
  summarize(Count=n()) %>% arrange(desc(Count)) %>% slice(1:5)
```

```
## # A tibble: 5 x 2
##   candidate       Count
##   <chr>           <int>
## 1 Chow, Olivia     5706
## 2 Tory, John       2601
## 3 Ford, Doug        608
## 4 Ford, Rob         531
## 5 Soknacki, David   314
```

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

Answer

- 184

```
main_data %>%
  group_by(contributors_name) %>%
  summarise(uni = length(unique(candidate))) %>%
  filter(uni > 1) %>% dim()
```

```
## [1] 184   2
```