Lab2

Michal Malyska

15/01/2020

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)

set.seed(1337)
library(opendatatoronto)
library(kableExtra)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()     masks stats::filter()
## x dplyr::group_rows() masks kableExtra::group_rows()
## x dplyr::lag()        masks stats::lag()
```

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

```
##
## Attaching package: 'janitor'
```

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

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(ggrepel)
library(tidylog)
```

```
##
## Attaching package: 'tidylog'
```

```
## The following objects are masked from 'package:dplyr':
```

```
##
##     add_count, add_tally, anti_join, count, distinct, distinct_all,
##     distinct_at, distinct_if, filter, filter_all, filter_at, filter_if,
##     full_join, group_by, group_by_all, group_by_at, group_by_if,
##     inner_join, left_join, mutate, mutate_all, mutate_at, mutate_if,
##     rename, rename_all, rename_at, rename_if, right_join, sample_frac,
##     sample_n, select, select_all, select_at, select_if, semi_join,
##     slice, summarise, summarise_all, summarise_at, summarise_if,
##     summarize, summarize_all, summarize_at, summarize_if, tally,
##     top_frac, top_n, transmute, transmute_all, transmute_at,
##     transmute_if, ungroup
```

```
## The following objects are masked from 'package:tidyr':
##
##     drop_na, fill, gather, pivot_longer, pivot_wider, replace_na,
##     spread, uncount
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

Lab Exercises

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

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).
2. Clean up the data format (fixing the parsing issue and standardizing the column names using `janitor`)
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.
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.
5. List the top five candidates in each of these categories:
    - total contributions
    - mean contribution
    - number of contributions
6. Repeat 5 but without contributions from the candidates themselves.
7. How many contributors gave money to more than one candidate?

Data Import (Question 1)

```r
all_data <- opendatatoronto::list_packages(limit = 500)
election_resources <- opendatatoronto::list_package_resources("f6651a40-2f52-46fc-9e04-b760c16edd5c")
elections <- opendatatoronto::get_resource("d99bb1f3-949a-4497-bb96-c93bbd203130")
contributions_orig <- elections$`2_Mayor_Contributions_2014_election.xls`
```

Cleaning names and Col Types (Question 2)

```r
contributions <- contributions_orig %>%
    janitor::row_to_names(1) %>%
    janitor::clean_names()

contributions <- contributions %>% readr::type_convert()
```

Data Wrangling stuff (Question 3)

```r
skimmed <- skimr::skim(contributions)

skimmed %>% select(-numeric.hist) %>% kable()
```

skim_type	skim_variable	n_missing	complete_rate	character.min	character.max	character.em
character	contributors_name	0	1.0000000	4	31	
character	contributors_address	10197	0.0001961	24	26	
character	contributors_postal_code	0	1.0000000	7	7	
character	contribution_type_desc	0	1.0000000	8	14	
character	goods_or_service_desc	10188	0.0010785	11	40	
character	contributor_type_desc	0	1.0000000	10	11	
character	relationship_to_candidate	10166	0.0032356	6	9	
character	president_business_manager	10197	0.0001961	13	16	
character	authorized_representative	10197	0.0001961	13	16	
character	candidate	0	1.0000000	9	18	
character	office	0	1.0000000	5	5	
logical	ward	10199	0.0000000	NA	NA	
numeric	contribution_amount	0	1.0000000	NA	NA	

```r
contributions <- contributions %>% janitor::remove_empty()
```

Ward is all missing so it will get removed by remove_empty. A couple of variables are pretty much all missing contributors_address, goods_or_service_desc, relationship_to_candidate, president_business_manager, authorized_representative. Most of them pertain to companies and there seem to be only 12 corporate contributions. They are the ones with non-missing address. They have all other values non missing except for relationship to candidate. That variable has 12

There are 7545 unique contributors that gave a total of 10199 contributions.

There are a few very large values for contributions, they were all given to the candidates by themselves.

```r
contributions %>% filter(!is.na(relationship_to_candidate)) %>%
    arrange(contribution_amount) %>%
    kableExtra::kable()
```

contributors_name	contributors_address	contributors_postal_code	contribution_amount	contribution_type_des
Khomenko, Klim	NA	M6B 2Z7	200.00	Monetary
Mernagh, Matt	NA	M6E 1E1	200.00	Monetary
Emond, Ryan	NA	M4Y 2J3	220.00	Monetary
Ruel, Jim	NA	M4K 3P3	231.70	Monetary
Khomenko, Klim	NA	M6B 2Z7	319.95	Monetary
Walker, Daniel	NA	M6K 2W9	369.32	Monetary
Ford, Doug	NA	M9A 2C3	500.00	Monetary
French, James	NA	M9R 3T5	500.00	Monetary
Johnson, Suzanne	NA	M5T 1J2	500.00	Monetary
Khomenko, Klim	NA	M6B 2Z7	500.00	Monetary
Tiwari, Ramnarine	NA	M3J 3K3	593.59	Monetary
Lee, Dewitt	NA	M5H 3L9	700.00	Monetary
Ford, Rob	NA	M9A 3G9	853.86	Monetary
Kalevar, Chai	NA	M6E 3C5	900.00	Monetary
Clarke, Kevin	NA	M1E 2R3	1200.00	Monetary
Ford, Rob	NA	M9A 3G9	1270.00	Monetary
Syed, Hïmy	NA	M3C 1C8	2018.00	Monetary
Chow, Olivia	NA	M5T 2B6	2500.00	Monetary
Hackett, Barbara	NA	M5R 2B5	2500.00	Monetary
Sniedzins, Erwin	NA	M4R 1C4	2500.00	Monetary
Thomson, Sarah	NA	M4W 2X6	2500.00	Monetary
Tory, John	NA	M5R 2B5	2500.00	Monetary
Yan, Flora	NA	M4R 1C4	2500.00	Monetary
Thomson, Sarah	NA	M4W 2X6	4425.55	Monetary
Di Paola, Rocco	NA	M3H 2T1	6000.00	Monetary
Ford, Rob	NA	M9A 3G9	12210.00	Monetary
Ford, Rob	NA	M9A 3G9	20000.00	Monetary
Goldkind, Ari	NA	M5P 1P5	23623.63	Monetary
Ford, Doug	NA	M9A 2C3	50000.00	Monetary
Ford, Rob	NA	M9A 3G9	50000.00	Monetary
Ford, Rob	NA	M9A 3G9	50000.00	Monetary
Ford, Rob	NA	M9A 3G9	78804.80	Monetary
Ford, Doug	NA	M9A 2C3	508224.73	Monetary

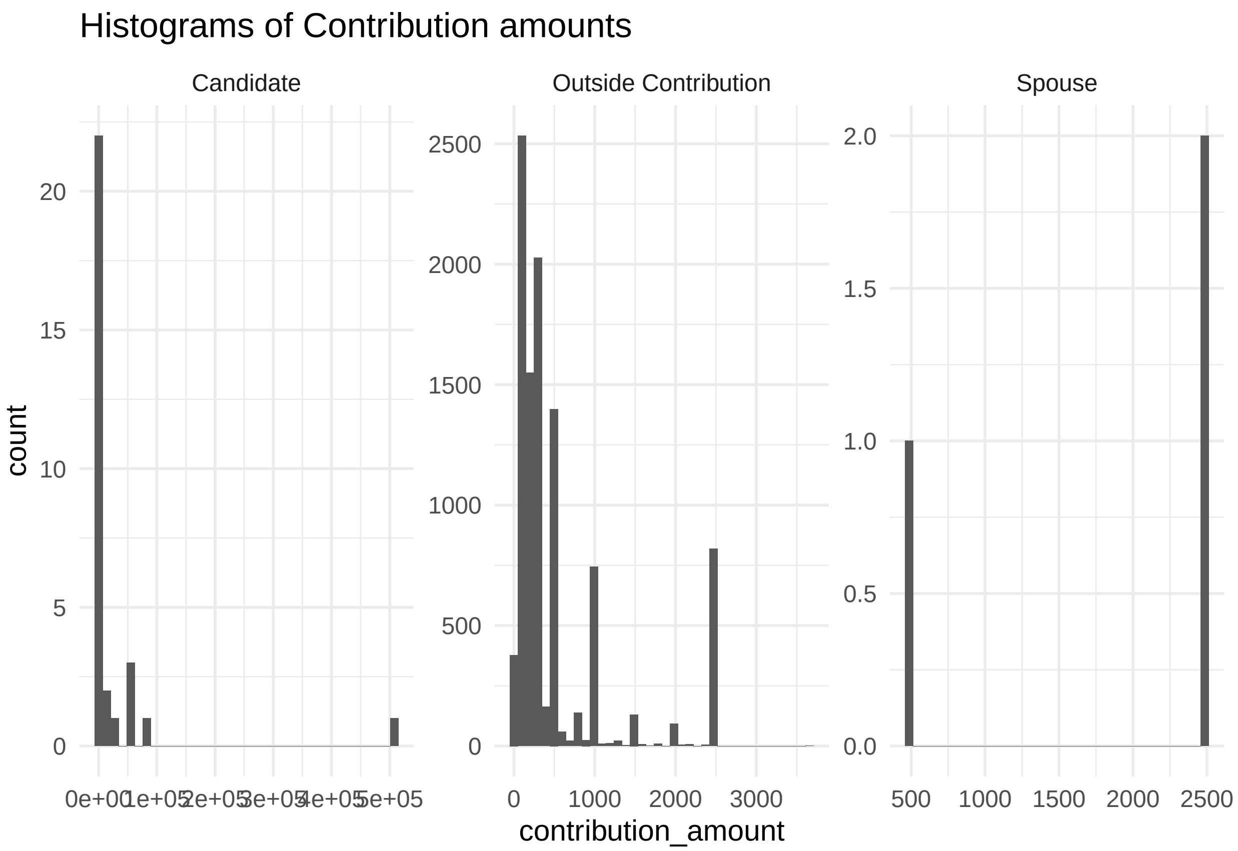
Data Summary (Question 4)

The summary of contribution monetary values is below. It is done separately based on relationship to candidate.

```r
numerics <- contributions %>% select_if(is.numeric)

numerics_summary <- numerics %>%
    summarize_all(.funs = funs(n_unique = sum(!is.na(.)),
                               mean = mean(.),
                               median = median(.),
                               sd = sd(.),
                               min = min(.),
                               max = max(.)))

numerics_summary %>% kableExtra::kable()
```

n_unique	mean	median	sd	min	max
10199	607.9521	300	5211.311	1	508224.7

```
numerics_hist <- contributions %>%
    mutate(relation_contributed = if_else(is.na(relationship_to_candidate),
                                          "Outside Contribution",
                                          relationship_to_candidate)) %>%
    ggplot() +
    aes(x = contribution_amount) +
    geom_histogram(bins = 38) +
    theme_minimal() +
    labs(title = "Histograms of Contribution amounts") +
    facet_wrap(.~relation_contributed, scales = "free", shrink = TRUE)

numerics_hist
```



Clearly the contributions from the general public tend to be between 0 and 3000, with clear spikes at round numbers like 1000, 2000, 2500. The contributions from people with a relation to the candidate are extremely varied, the spousal ones tend to be very close to the general public, while the candidates like to give themselves a lot of money.

Question 5

```
contributions_summary_q5 <- contributions %>%
    group_by(candidate) %>%
    summarize(total_contributions = sum(contribution_amount),
              mean_contributions = mean(contribution_amount),
              number_contributions = n())
```

Top total contributions:

```
top_5_total <- contributions_summary_q5 %>%
    arrange(desc(total_contributions)) %>%
    top_n(n = 5, wt = total_contributions)

top_5_total %>% kable()
```

candidate	total_contributions	mean_contributions	number_contributions
Tory, John	2767868.7	1063.7466	2602
Chow, Olivia	1638265.9	287.0122	5708
Ford, Doug	889897.3	1456.4604	611
Ford, Rob	387648.2	720.5356	538
Stintz, Karen	242805.0	995.1025	244

Top average contributions:

```
top_5_mean <- contributions_summary_q5 %>%
    arrange(desc(mean_contributions)) %>%
    top_n(n = 5, wt = mean_contributions)

top_5_mean %>% kable()
```

candidate	total_contributions	mean_contributions	number_contributions
Sniedzins, Erwin	8100.0	2025.000	4
Syed, Hïmy	2018.0	2018.000	1
Ritch, Carlie	5660.0	1886.667	3
Ford, Doug	889897.3	1456.460	611
Clarke, Kevin	1200.0	1200.000	1

Top number of contributions:

```
top_5_number <- contributions_summary_q5 %>%
    arrange(desc(number_contributions)) %>%
    top_n(n = 5, wt = number_contributions)

top_5_number %>% kable()
```

candidate	total_contributions	mean_contributions	number_contributions
Chow, Olivia	1638265.9	287.0122	5708
Tory, John	2767868.7	1063.7466	2602
Ford, Doug	889897.3	1456.4604	611
Ford, Rob	387648.2	720.5356	538
Soknacki, David	132431.0	421.7548	314

Question 6

```
contributions_summary_q6 <- contributions %>%
    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())
```

Top total contributions:

```r
top_5_total <- contributions_summary_q6 %>%
    arrange(desc(total_contributions)) %>%
    top_n(n = 5, wt = total_contributions)

top_5_total %>% kable()
```

candidate	total_contributions	mean_contributions	number_contributions
Tory, John	2765368.7	1063.1944	2601
Chow, Olivia	1635765.9	286.6245	5707
Ford, Doug	331172.6	544.6917	608
Stintz, Karen	242805.0	995.1025	244
Ford, Rob	174509.5	328.6431	531

Top average contributions:

```r
top_5_mean <- contributions_summary_q6 %>%
    arrange(desc(mean_contributions)) %>%
    top_n(n = 5, wt = mean_contributions)

top_5_mean %>% kable()
```

candidate	total_contributions	mean_contributions	number_contributions
Ritch, Carlie	5660	1886.667	3
Sniedzins, Erwin	5600	1866.667	3
Tory, John	2765369	1063.194	2601
Gardner, Norman	3000	1000.000	3
Tiwari, Ramnarine	1000	1000.000	1

Top number of contributions:

```r
top_5_number <- contributions_summary_q6 %>%
    arrange(desc(number_contributions)) %>%
    top_n(n = 5, wt = number_contributions)

top_5_number %>% kable()
```

candidate	total_contributions	mean_contributions	number_contributions
Chow, Olivia	1635765.9	286.6245	5707
Tory, John	2765368.7	1063.1944	2601
Ford, Doug	331172.6	544.6917	608
Ford, Rob	174509.5	328.6431	531
Soknacki, David	132431.0	421.7548	314

Question 7

```r
contributions_q7 <- contributions %>%
    select(contributors_name, candidate) %>%
    distinct() %>%
    group_by(contributors_name) %>%
    summarize(num_candidates = n()) %>%
    filter(num_candidates > 1)
```

There were 184 people who gave money to more than one candidate.