lab2

Yiyu Wang

```r
library(opendatatoronto)
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.4.0      v purrr   0.3.5
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.1      v stringr 1.5.0
v readr   2.1.3      v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```r
library(stringr)
library(skimr) # EDA
library(visdat) # EDA
library(janitor)
```

```

Attaching package: 'janitor'

The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

```r
library(lubridate)
```

```
Loading required package: timechange

Attaching package: 'lubridate'

The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

```
library(ggrepel)
library(dplyr)
```

```
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)
```

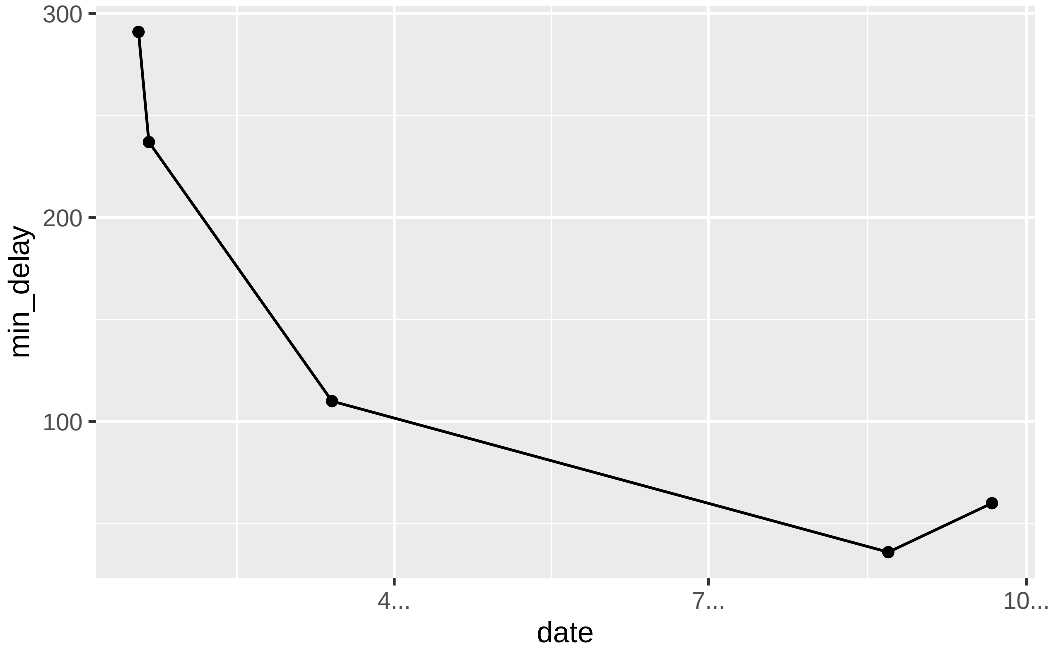
Question 1

```
delay_2022_mean_delay <- delay_2022 |>
  group_by(station)|>
  summarise(mean(min_delay))
delay_2022_mean_delay <- clean_names(delay_2022_mean_delay)
head(arrange(delay_2022_mean_delay,desc(mean_min_delay),n=5))
```

```
# A tibble: 6 x 2
  station                 mean_min_delay
  <chr>                            <dbl>
1 SHEPPARD WEST TO UNION             291
2 KIPLING TO JANE                    237
3 MUSEUM TO EGLINTON STA             110
4 WILSON YARD HOSTLER 2               60
5 VIADUCT                             36
6 MAIN TO VICTORIA PARK               31
```

the five stations with the highest mean delays is SHEPPARD WEST TO UNION, KIPLING TO JANE, MUSEUM TO EGLINTON STA,WILSON YARD HOSTLER 2 and VIADUCT.

```
delay_2022|>
  filter(station=='SHEPPARD WEST TO UNION'|station=='KIPLING TO JANE'|station=='MUSEUM TO
  ggplot(aes(x=date,y=min_delay))+geom_line()+geom_point()
```



Question 2

```
all_data <- list_packages(limit = 500)
all_data|>
  filter(str_detect(title,pattern = "Campaign"))
```

```
# A tibble: 5 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 Civic Is~ 7d0d~ City ~ Afford~ Inform~ "The O~ Table         5 XML,JS~ As ava~
2 Election~ 67d2~ Finan~ <NA>    City C~ "This ~ Docume~       2 ZIP,XL~ As ava~
3 Election~ f665~ City ~ <NA>    City C~ "This ~ Docume~       2 ZIP,XLS As ava~
4 Election~ 28e5~ City ~ <NA>    City C~ "This ~ Docume~       2 ZIP,XLS As ava~
```

```
5 Election~ 2ee8~ City ~ <NA>    City C~ "This ~ Docume~       2 ZIP,XLS As ava~
# ... 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
```

The ID for 'Elections - Campaign Contributions - 2014 to 2017' is f6651a40-2f52-46fc-9e04-b760c16edd5c

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

```
# 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
```

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

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

```
may_con_2014 <- cap[["2_Mayor_Contributions_2014_election.xls"]]
```

Question 3

```
colnames(may_con_2014)=may_con_2014[1,]
may_con_2014<-may_con_2014[-1,]
may_con_2014<-clean_names(may_con_2014)
may_con_2014$contribution_amount<- as.numeric(may_con_2014$contribution_amount)
```

Question 4

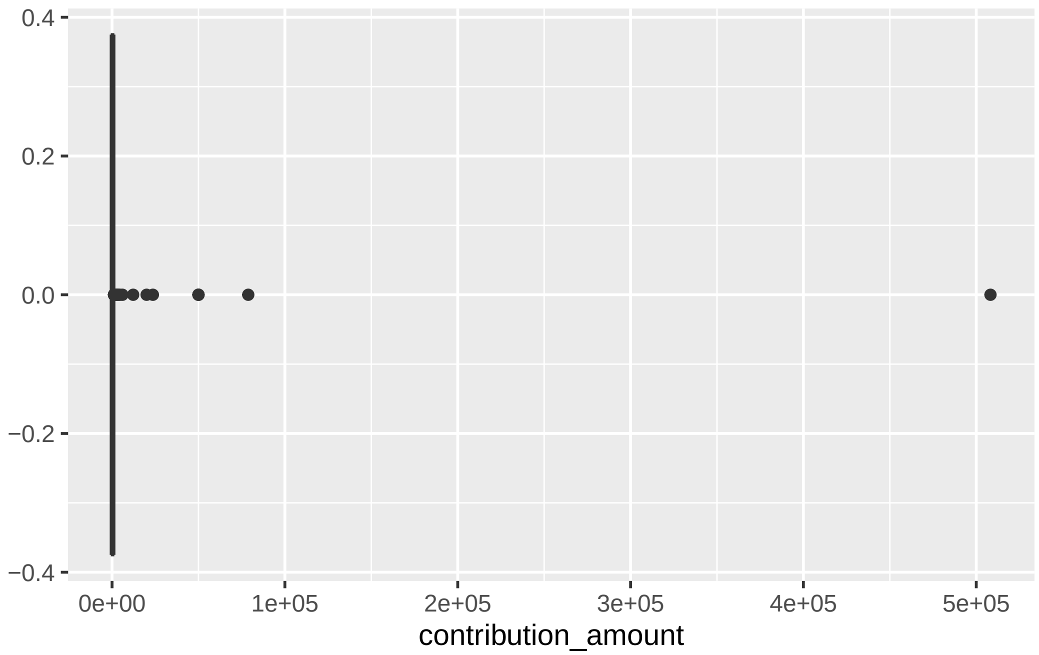
```
may_con_2014 |>
  summarize(across(everything(), ~ sum(is.na(.x))))
```

```
# A tibble: 1 x 13
  contributors~1 contr~2 contr~3 contr~4 contr~5 goods~6 contr~7 relat~8 presi~9
           <int>   <int>   <int>   <int>   <int>   <int>   <int>   <int>   <int>
1              0   10197       0       0       0   10188       0   10166   10197
# ... with 4 more variables: authorized_representative <int>, candidate <int>,
#   office <int>, ward <int>, 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
```

There are a lot of missing variables in the dataset, but they are belong to contributors_address, good/servise description, relationship_to_candidate, president_business_manager, authorized_representative and ward, so we do not need to be worried about them. Every variable in the format it should be.

Question 5

```
ggplot(may_con_2014,aes(x=contribution_amount))+geom_boxplot()
```

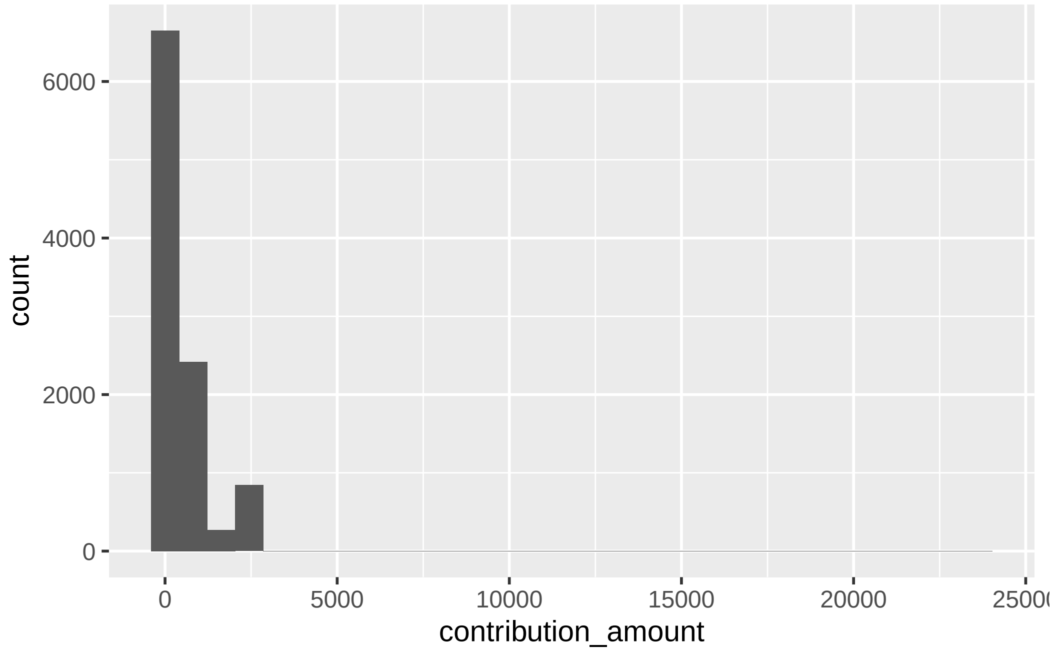


```
may_con_2014|>
  filter(contribution_amount>49999)
```

```
# A tibble: 5 x 13
  contributors~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, Doug     <NA>    M9A 2C3  50000  Moneta~ <NA>    Indivi~ Candid~ <NA>
3 Ford, Rob      <NA>    M9A 3G9  50000  Moneta~ <NA>    Indivi~ Candid~ <NA>
4 Ford, Rob      <NA>    M9A 3G9  50000  Moneta~ <NA>    Indivi~ Candid~ <NA>
5 Ford, Rob      <NA>    M9A 3G9  78805. Moneta~ <NA>    Indivi~ Candid~ <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
```

```
may_con_2014_no <- may_con_2014|>
  filter(contribution_amount<49999)
ggplot(may_con_2014_no, aes(x=contribution_amount)) + geom_histogram()
```

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



There is 5 notable outliers. The outlier shows similar characteristic which is they are all come from Ford family to fund their own campaign. After removing the notable outlier, we can find that the majority of the contribution amount is below 2500.

Question 6

total contributions

```
may_con_2014_total <- may_con_2014|>
  group_by(contributors_name)|>
  summarise(sum(contribution_amount))
may_con_2014_total<-clean_names(may_con_2014_total)
head(arrange(may_con_2014_total,desc(sum_contribution_amount)),n=5)
```

```
# A tibble: 5 x 2
  contributors_name  sum_contribution_amount
  <chr>                                <dbl>
```

```
1 Ford, Doug                          561225.
2 Ford, Rob                           213139.
3 Goldkind, Ari                        23624.
4 Thomson, Sarah                        6926.
5 Pappalardo, Victor                    6300 
```

mean contribution

```
may_con_2014_mean <- may_con_2014|>
  group_by(contributors_name)|>
  summarise(mean(contribution_amount))
may_con_2014_mean<-clean_names(may_con_2014_mean)
head(arrange(may_con_2014_mean,desc(mean_contribution_amount)),n=5)
```

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

number of contributions

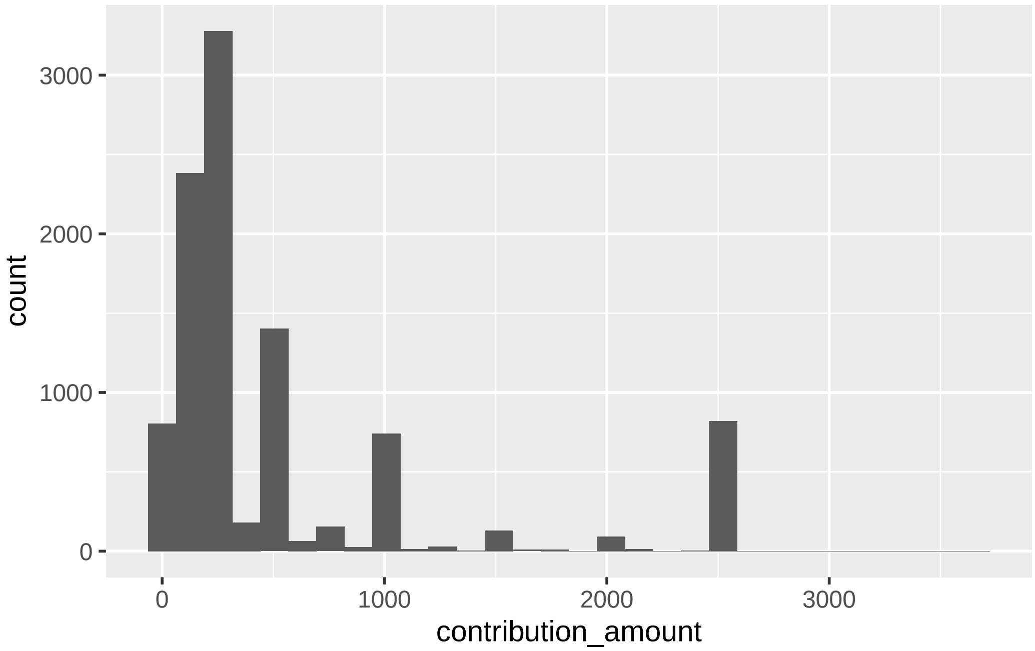
```
may_con_2014_count <- may_con_2014|>
  group_by(contributors_name)|>
  count(contributors_name)
head(arrange(may_con_2014_count,desc(n)),n=5)
```

```
# A tibble: 5 x 2
# Groups:   contributors_name [5]
  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
```

Question 7

```
may_con_2014|>
  filter(!contributors_name==candidate)|>
  ggplot(aes(x=contribution_amount)) +geom_histogram()
```

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



Question 8

```
contributor_name <-may_con_2014|>
  group_by(contributors_name)|>
  count(candidate)|>
  get_dupes(contributors_name)
length(unique(contributor_name$contributors_name))
```

```
[1] 184
```