Assignment 4

Krishu Kumar Thapa

2022-10-03

NYC Flights 13 data set

Reading the data from the csv files and storing them in the variables:

```
suppressMessages(library(dplyr))
library(tidyr)

airlines = read.csv('airlines.csv', sep=',', header = TRUE)
weather = read.csv('weather.csv', sep=',', header = TRUE)
airports = read.csv('airports.csv', sep=',', header = TRUE)
flights = read.csv('flights.csv', sep=',', header = TRUE)
planes = read.csv('planes.csv',sep=',', header = TRUE)
```

Question 1a.

Filtering the data set using the left join.

```
flight2 <- flights %>% filter(month == 11 & day == 1 & year == 2013 &
  (hour>=12 & hour<=18) & dest == 'TPA') %>%
  select(tailnum,year,month,day,hour,origin)


weather2<- weather %>%  select(year,month,day,hour,origin,humid)

answer<-left_join(flight2,weather2,by=c('origin','month','day','year','hour'))

answer
```

```
##   tailnum year month day hour origin humid
## 1  N580JB 2013    11   1   14    JFK 63.08
## 2  N337NB 2013    11   1   14    LGA 56.51
## 3  N567UA 2013    11   1   15    EWR 52.80
## 4  N515MQ 2013    11   1   14    JFK 63.08
## 5  N779JB 2013    11   1   15    EWR 52.80
## 6  N561JB 2013    11   1   16    LGA 50.60
## 7  N974DL 2013    11   1   18    JFK 74.75
```

```
count(answer)
```

```
##   n
## 1 7
```

According to the constraints applied above, the number of flighs that happened between the given time frame i.e. 12pm to 6pm is 7.

Question 1b.

Analyzing the two different joins by running the script:

```
anti_join(flights,airports, by = c("origin" = "faa"))
anti_join(flights,airports, by = c("faa" = "origin"))
```

For first command,
anti_join drops all the observations on flights that match with the airports. Here **origin = faa** shows that the join is done by the **origin** column of flights whose corresponding column is **faa** in the airport table. Since all the values of origin in flights have its corresponding value in faa , the final result of this operation is empty.

For second command,
This will give an error because in the portion **faa= origin**, it violates the order of dataframes in the anti_join, i.e, since we have the (x,y) as flights and airports , in the **by** section the left hand side should have column corresponding to flights and right hand should have column corresponding to airports.

The difference between semi_join and anti_join is given by:

semi_join(x,y) : It keeps all the data in x that matches the data in y governed by the **by** section.
anti_join(x,y) : It discards all the data in x that matches the data in y governed by the **by** section.

Question 1c.

The result for the constraints is given by:

```
origin_info <- flights %>% select(origin,dest) %>%inner_join(airports %>%
              select(faa,lat,lon),by = c("origin"="faa"))

origin_info <- origin_info %>% mutate(.,origin_lat = lat , origin_lon = lon)%>%
               select(-lat,-lon)

dest_info <- origin_info %>% inner_join(airports %>% select(faa,lat,lon),
                                        by = c("dest"="faa"))

flight_info <- dest_info %>% mutate(.,dest_lat = lat , dest_lon = lon) %>%
               select(-lat,-lon)

head(flight_info,20)
```

```
##    origin dest origin_lat origin_lon dest_lat   dest_lon
## 1     EWR  IAH   40.69250  -74.16867 29.98443  -95.34144
## 2     LGA  IAH   40.77725  -73.87261 29.98443  -95.34144
## 3     JFK  MIA   40.63975  -73.77893 25.79325  -80.29056
## 4     LGA  ATL   40.77725  -73.87261 33.63672  -84.42807
## 5     EWR  ORD   40.69250  -74.16867 41.97860  -87.90484
## 6     EWR  FLL   40.69250  -74.16867 26.07258  -80.15275
## 7     LGA  IAD   40.77725  -73.87261 38.94453  -77.45581
## 8     JFK  MCO   40.63975  -73.77893 28.42939  -81.30899
## 9     LGA  ORD   40.77725  -73.87261 41.97860  -87.90484
## 10    JFK  PBI   40.63975  -73.77893 26.68316  -80.09559
```

```
## 11    JFK  TPA   40.63975  -73.77893 27.97547  -82.53325
## 12    JFK  LAX   40.63975  -73.77893 33.94254 -118.40807
## 13    EWR  SFO   40.69250  -74.16867 37.61897 -122.37489
## 14    LGA  DFW   40.77725  -73.87261 32.89683  -97.03800
## 15    JFK  BOS   40.63975  -73.77893 42.36435  -71.00518
## 16    EWR  LAS   40.69250  -74.16867 36.08006 -115.15225
## 17    LGA  FLL   40.77725  -73.87261 26.07258  -80.15275
## 18    LGA  ATL   40.77725  -73.87261 33.63672  -84.42807
## 19    EWR  PBI   40.69250  -74.16867 26.68316  -80.09559
## 20    LGA  MSP   40.77725  -73.87261 44.88196  -93.22177
```

```r
#count of flights
count(flight_info)
```

```
##        n
## 1 329174
```

Question 1d.

The number of carrier/dest unique combination is given by:

```r
combination <- flights %>% group_by(carrier,dest) %>% count()

head(combination,20)
```

```
## # A tibble: 20 x 3
## # Groups:   carrier, dest [20]
##    carrier dest      n
##    <chr>   <chr> <int>
##  1 9E      ATL      59
##  2 9E      AUS       2
##  3 9E      AVL      10
##  4 9E      BGR       1
##  5 9E      BNA     474
##  6 9E      BOS     914
##  7 9E      BTV       2
##  8 9E      BUF     833
##  9 9E      BWI     856
## 10 9E      CAE       3
## 11 9E      CHS     348
## 12 9E      CLE     349
## 13 9E      CLT     291
## 14 9E      CMH      13
## 15 9E      CVG    1559
## 16 9E      DAY     391
## 17 9E      DCA    1074
## 18 9E      DFW     379
## 19 9E      DSM      91
## 20 9E      DTW    1013
```

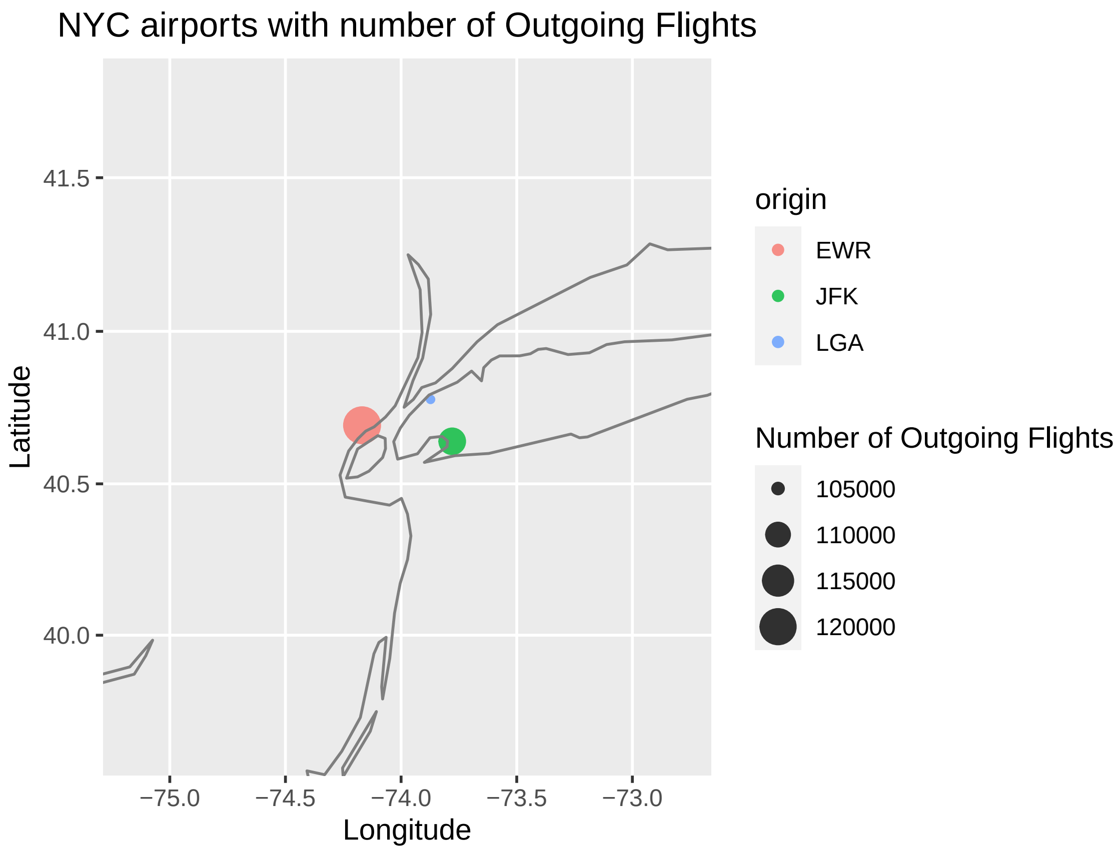
```r
# Number of unique combination of flights

nrow(combination)
```

```
## [1] 314
```

Question 1e.

Map that sizes each origin airport by the number of outgoing flights is shown below:

```
library(ggplot2)

origin_count <- flights %>% group_by(origin) %>% count()

final_info <- origin_count %>% left_join(airports, by = c("origin" = "faa")) %>%
    select(origin,lat,lon,n)

ggplot(final_info, aes(x=lon,y=lat, size= n , color = origin)) +
  geom_point(alpha= 0.8) +
  scale_size_continuous(name = "Number of Outgoing Flights") +
  borders("world") +
  coord_map(xlim = c(min(final_info$lon)-1 , max(final_info$lon)+1 ),
            ylim = c(min(final_info$lat)-1, max(final_info$lat)+1))+
  labs(x="Longitude", y="Latitude", title="NYC airports with number of Outgoing Flights")+
  theme(plot.title = element_text(hjust = 0.5))
```



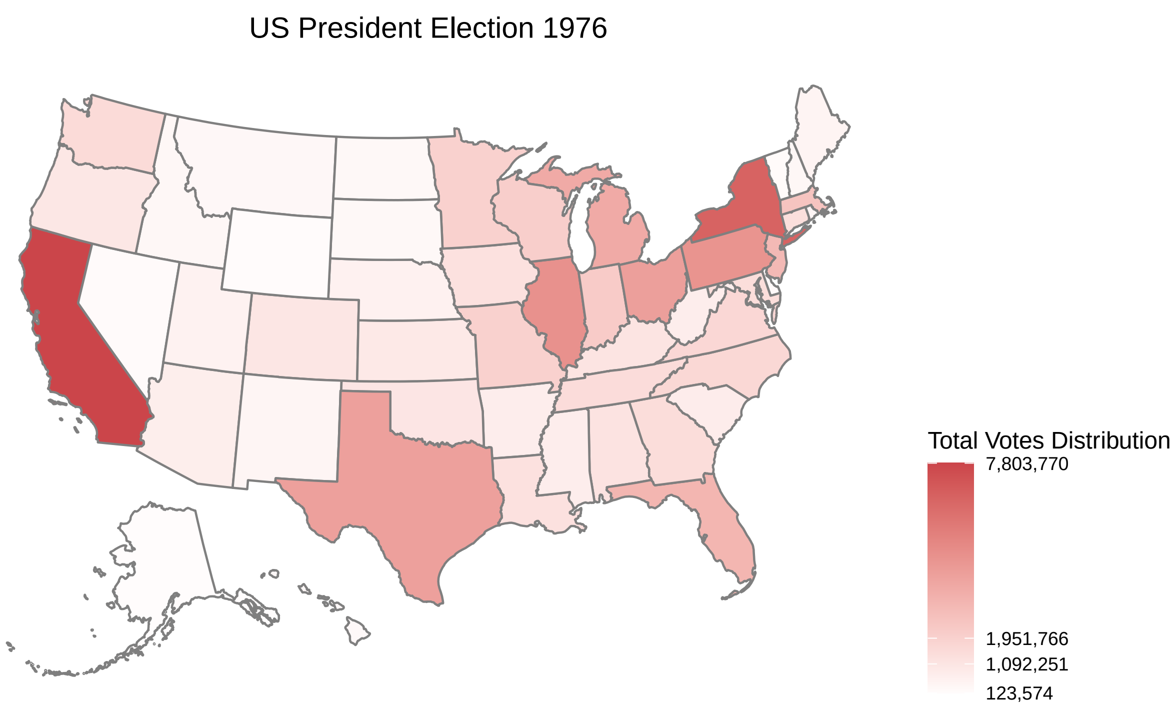
Question 2

Reading the presidential data:

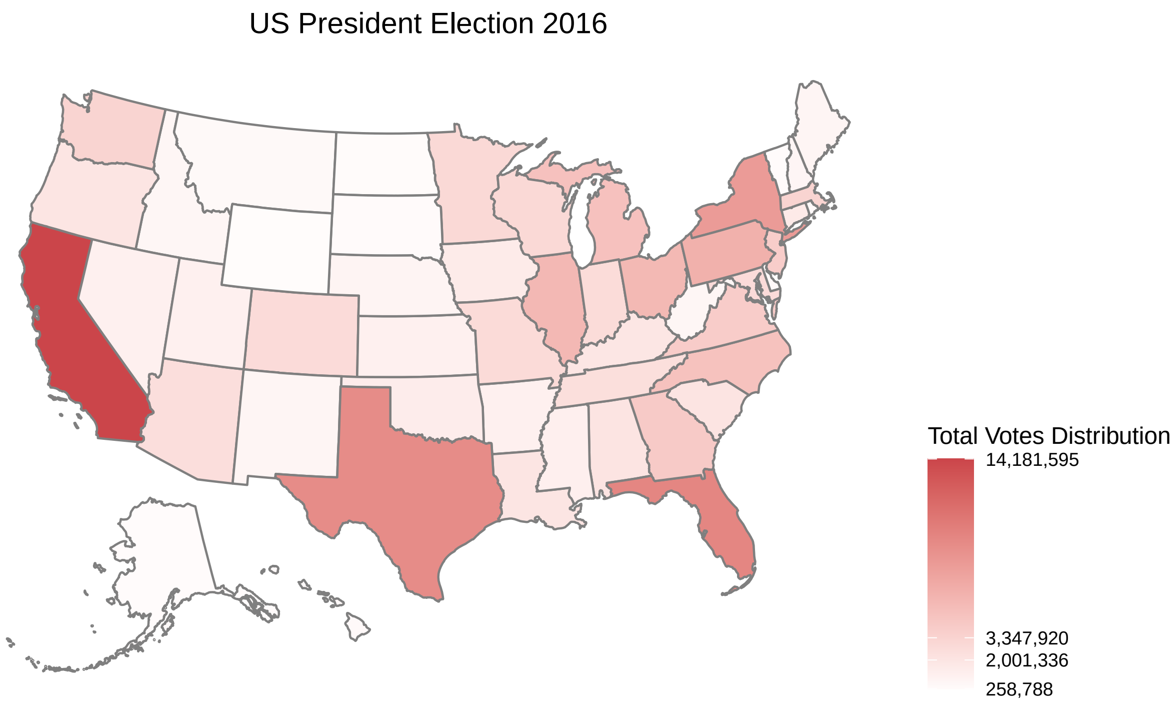
```
us_presidents = read.csv('us-presidents.csv',sep=',', header = TRUE)
```

Creating two data sets for two years of election:

```r
library(usmap)

election_1 <- us_presidents %>% filter(year == 1976)
election_2 <- us_presidents %>% filter(year == 2016)

# Election Year 1976 mapping

plot_1 <- plot_usmap(regions ="states",data = election_1,
values = "totalvotes", color = "gray50") +
scale_fill_gradient2(low = "white", high = "#CB454A",
name = "Total Votes Distribution",breaks = c(min(election_1$totalvotes),
quantile(election_1$totalvotes, c(0.5,0.75)), max(election_1$totalvotes)),
label = scales::comma)+labs(title = "US President Election 1976") +
theme(legend.position = "right",plot.title = element_text(hjust = 0.5))

plot_1
```




```r
# Election Year 2016 mapping

plot_2 <- plot_usmap(region="states", data = election_2, values = "totalvotes",
      color = "gray50") + scale_fill_gradient2(low = "white", high = "#CB454A"
      , name = "Total Votes Distribution",breaks = c(min(election_2$totalvotes)
    ,quantile(election_2$totalvotes, c(0.5,0.75)), max(election_2$totalvotes)),
  label = scales::comma) + labs(title = "US President Election 2016")+
  theme(legend.position = "right", plot.title = element_text(hjust = 0.5))

plot_2
```



Comparing the election data from 1976 and 2016, we can see that in the period of 50 years the numbers of total votes have almost doubled as all the minimum, maximum and quantiles of total votes seems to have doubled as perceived from the votes distribution scales. Also we can see that, there has been decrease in the density of votes in the north east side and slight increase in votes density in the south east side in this period of 50 years.

Question 3

Creating the wordcloud for the document is shown below. The text used here is **Jane's Statement of Purpose to her medical school**

```
suppressMessages(library(textreadr))
suppressMessages(library(wordcloud))
suppressMessages(library(RColorBrewer))
suppressMessages(library(tm))

# Reading the rich text file
text_content <- read_rtf('WordCloud.rtf',skip = 0, remove.empty = TRUE, trim = TRUE)

# Creating the corpus
text_corpus <- Corpus(VectorSource(text_content))

# Text cleaning

removeSpecialChars <- function(x) gsub("[^a-zA-Z0-9 ]","",x)
text_corpus <- tm_map(text_corpus, removeSpecialChars)

text_corpus <- text_corpus %>%
  tm_map(removeNumbers) %>%
  tm_map(removePunctuation) %>%
  tm_map(stripWhitespace)
```

```r
text_corpus <- tm_map(text_corpus, content_transformer(tolower))
text_corpus <- tm_map(text_corpus, removeWords, stopwords("english"))

# Making a document matrix

document_matrix <- TermDocumentMatrix(text_corpus)
document_matrix <- as.matrix(document_matrix)

words_info <- sort(rowSums(document_matrix),decreasing=TRUE)

word_count_data <- data.frame(word = names(words_info),freq=words_info)

layout(matrix(c(1, 2), nrow=2), heights=c(1,5))
par(mar=rep(0, 4))
plot.new()
text(x=0.5, y=0.5, "Jane's Statement of Purpose to Medical School")

wordcloud(words = word_count_data$word, freq = word_count_data$freq,
          min.freq = 1,max.words=200,
          scale = c(1.5,0.6),
          random.order=FALSE, rot.per=0.35,colors=brewer.pal(8, "Dark2"))
```

Jane's Statement of Purpose to Medical School