IMT 573: Problem Set 2 -Data Manipulation

Kulraj Singh Kohli

Due: Wednesday, October 16 2019

Instructions

Before beginning this assignment, please ensure you have access to R and RStudio.

1. Replace the "Insert Your Name Here" text in the `author:` field with your own full name.
2. Be sure to include well-documented (e.g. commented) code chucks, figures and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text.
3. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the R Markdown file to `YourLastName-YourFirstName-ps1.Rmd`, knit a PDF and submit the PDF file on Canvas.
4. List any collaborators below:

Collaborators:

Setup:

Do whatever setup you do here, such as loading libraries

```
# Load standard libraries
library("tidyverse")
library("nycflights13")
data(flights)
```

Problem 1: Exploring the NYC Flights Data

(a) Importing and Inspecting Data:

1.In the Year 2013 the total number of flights out of NYC are :-

```
data(flights)
# Completing the NA cases
flightsx<- flights[complete.cases(flights),]
summary(flights)
```

```
##       year          month             day          dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907
##  Median :2013   Median : 7.000   Median :16.00   Median :1401
##  Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400
##                                                  NA's   :8255
##  sched_dep_time   dep_delay          arr_time    sched_arr_time
##  Min.   : 106   Min.   : -43.00   Min.   :   1   Min.   :   1
```

```
##  1st Qu.: 906   1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124
##  Median :1359   Median :  -2.00   Median :1535   Median :1556
##  Mean   :1344   Mean   :  12.64   Mean   :1502   Mean   :1536
##  3rd Qu.:1729   3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945
##  Max.   :2359   Max.   :1301.00   Max.   :2400   Max.   :2359
##                 NA's   :8255      NA's   :8713
##    arr_delay          carrier              flight       tailnum
##  Min.   : -86.000   Length:336776      Min.   :   1   Length:336776
##  1st Qu.: -17.000   Class :character   1st Qu.: 553   Class :character
##  Median :  -5.000   Mode  :character   Median :1496   Mode  :character
##  Mean   :   6.895                      Mean   :1972
##  3rd Qu.:  14.000                      3rd Qu.:3465
##  Max.   :1272.000                      Max.   :8500
##  NA's   :9430
##     origin              dest              air_time        distance
##  Length:336776      Length:336776      Min.   : 20.0   Min.   :  17
##  Class :character   Class :character   1st Qu.: 82.0   1st Qu.: 502
##  Mode  :character   Mode  :character   Median :129.0   Median : 872
##                                        Mean   :150.7   Mean   :1040
##                                        3rd Qu.:192.0   3rd Qu.:1389
##                                        Max.   :695.0   Max.   :4983
##                                        NA's   :9430
##       hour            minute        time_hour
##  Min.   : 1.00   Min.   : 0.00   Min.   :2013-01-01 05:00:00
##  1st Qu.: 9.00   1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00
##  Median :13.00   Median :29.00   Median :2013-07-03 10:00:00
##  Mean   :13.18   Mean   :26.23   Mean   :2013-07-03 05:22:54
##  3rd Qu.:17.00   3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00
##  Max.   :23.00   Max.   :59.00   Max.   :2013-12-31 23:00:00
##
```

```
dim(flightsx)
```

```
## [1] 327346     19
```

There are 327346 flights with all the data available about them out of NYC in the year 2013

2.How many NYC airports are included in this data? Which airports are these?

```
# by inspecting the origin column

unique(flightsx$origin)
```

```
## [1] "EWR" "LGA" "JFK"
```

We see that there are three unique airports, EWR,LGA and JFK

3.Into how many airports did the airlines fly from NYC in 2013?

```
#counting the unique elements in dest column

unique(flightsx$dest) %>% length()
```

```
## [1] 104
```

We see that there are 104 destinations to which these flights are flying.

4.. How many flights were there from NYC to Seattle (airport code SEA)?

```
# selecting all flights to seattl then counting
flightsx%>% filter(dest == "SEA") %>% nrow()
```

```
## [1] 3885
```

Therefore there are 3885 flights to seattle in 2013 from nyc

5. Were the any flights from NYC to Spokane (GAG)?

```
flightsx%>% filter(dest == "GAG") %>% nrow()
```

```
## [1] 0
```

There were no flights to spokane

6.What about missing destination codes? Are there any destinations that do not look like valid airport codes (three-letter-all-upper case)?

```
# checking unique values in dest
unique(flightsx$dest)
```

```
##   [1] "IAH" "MIA" "BQN" "ATL" "ORD" "FLL" "IAD" "MCO" "PBI" "TPA" "LAX"
##  [12] "SFO" "DFW" "BOS" "LAS" "MSP" "DTW" "RSW" "SJU" "PHX" "BWI" "CLT"
##  [23] "BUF" "DEN" "SNA" "MSY" "SLC" "XNA" "MKE" "SEA" "ROC" "SYR" "SRQ"
##  [34] "RDU" "CMH" "JAX" "CHS" "MEM" "PIT" "SAN" "DCA" "CLE" "STL" "MYR"
##  [45] "JAC" "MDW" "HNL" "BNA" "AUS" "BTV" "PHL" "STT" "EGE" "AVL" "PWM"
##  [56] "IND" "SAV" "CAK" "HOU" "LGB" "DAY" "ALB" "BDL" "MHT" "MSN" "GSO"
##  [67] "CVG" "BUR" "RIC" "GSP" "GRR" "MCI" "ORF" "SAT" "SDF" "PDX" "SJC"
##  [78] "OMA" "CRW" "OAK" "SMF" "TYS" "PVD" "DSM" "PSE" "TUL" "BHM" "OKC"
##  [89] "CAE" "HDN" "BZN" "MTJ" "EYW" "PSP" "ACK" "BGR" "ABQ" "ILM" "MVY"
## [100] "SBN" "LEX" "CHO" "TVC" "ANC"
```

All the destinations seem to be uniform three-letter-all-upper case

(b) Formulating Questions:

1. What is the typical delay of the flights in this data?

```
# filter all dep delays to positive values and find their mean
 flights %>% filter(dep_delay>0) %>% summarise(sum(dep_delay)/n())
```

```
## # A tibble: 1 x 1
##   `sum(dep_delay)/n()`
##                  <dbl>
## 1                 39.4
```

The typical total delay of flights is 39.37 minutes

2.Did you remember to check how good is the delay variable? Are there missings? Are there any implausible or invalid entries? Go and check this.

```
#Checking for invalid/NA entries for dep_delay
flights %>% summarise(count=sum(is.na(flights$dep_delay)))
```

```
## # A tibble: 1 x 1
##   count
##   <int>
## 1  8255
```

We see from the above results that there are 8255 implausible values in the dep_delay variable

3. Now compute the delay by destinations. Which ones are the worst three destinations in terms of the longest delay?

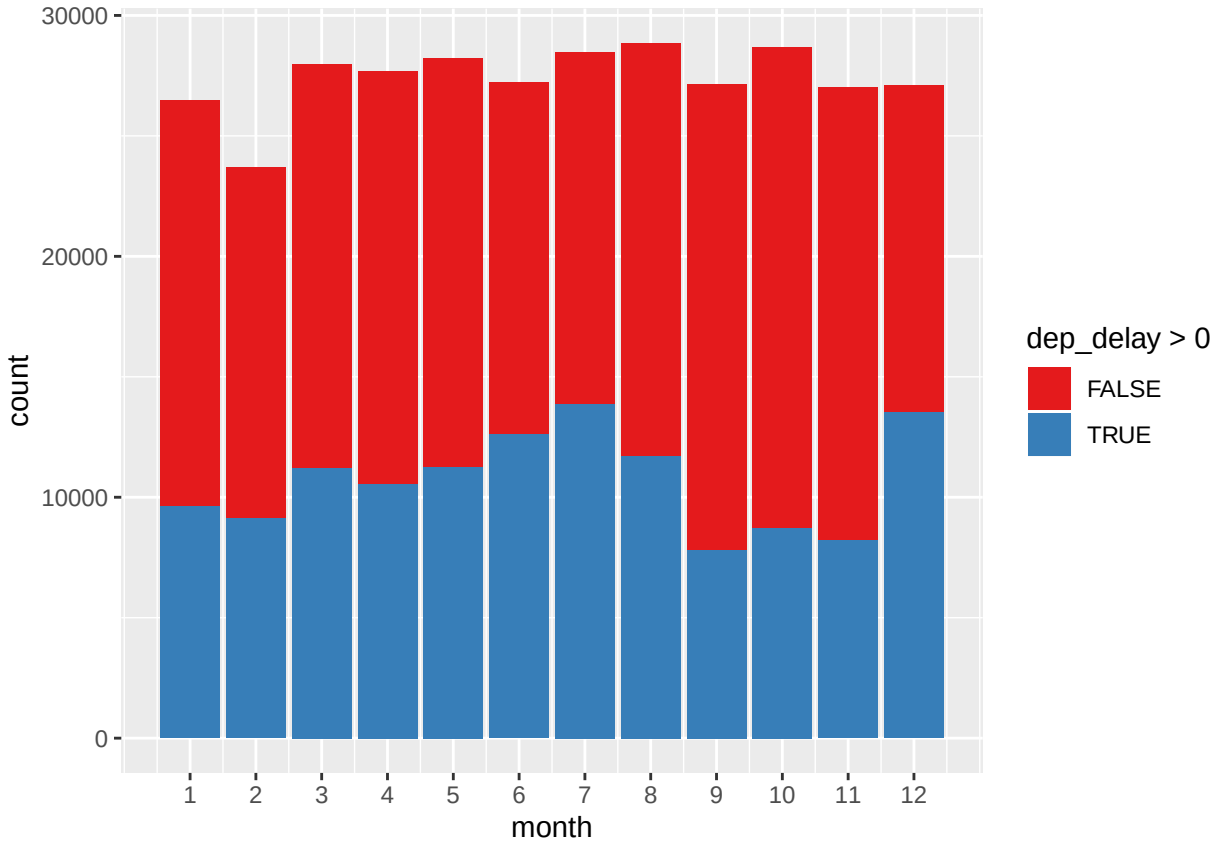
```
# grouping flights data wrt to dest and adding up all delay
flights %>% filter(dep_delay>0) %>% group_by(dest) %>% summarise(max_delay=sum(dep_delay)) %>% arrange(
```

```
## # A tibble: 103 x 2
##    dest  max_delay
##    <chr>     <dbl>
##  1 ORD      275023
##  2 ATL      254414
##  3 SFO      197238
##  4 MCO      195015
##  5 BOS      185833
##  6 LAX      185631
##  7 FLL      182464
##  8 CLT      171878
##  9 DTW      136887
## 10 MIA      135136
## # ... with 93 more rows
```

We see that the top 3 most delayed flights are to the destinations ORD,ATL,SFO.

4.Delays may be partly related to weather. We do not have weather information here but let's analyze how it is related to season. Do it in two (or more) ways: one graphical, and one in a table form.

```
# Plotting a bar graph for delays over various months
flights %>%
  filter(!is.na(dep_delay)) %>%
  mutate(m= factor(month)) %>%
  ggplot()+
  geom_bar(aes(x=month,group=dep_delay>0, fill=dep_delay>0)) +
  scale_x_continuous(breaks=1:12) +
  scale_fill_brewer(palette="Set1")
```




```
#creating a table for total delay minutes per month
flights %>% filter(dep_delay>0) %>% group_by(month)%>% summarise(tot_delay=sum(dep_delay))
```

```
## # A tibble: 12 x 2
##    month tot_delay
##    <int>     <dbl>
##  1     1    341410
##  2     2    322073
##  3     3    444060
##  4     4    465845
##  5     5    443117
##  6     6    630104
##  7     7    678868
##  8     8    436594
##  9     9    278814
## 10    10    275272
## 11    11    236519
## 12    12    504107
```

The above table and bar graph show the delay trends during different months of the year.

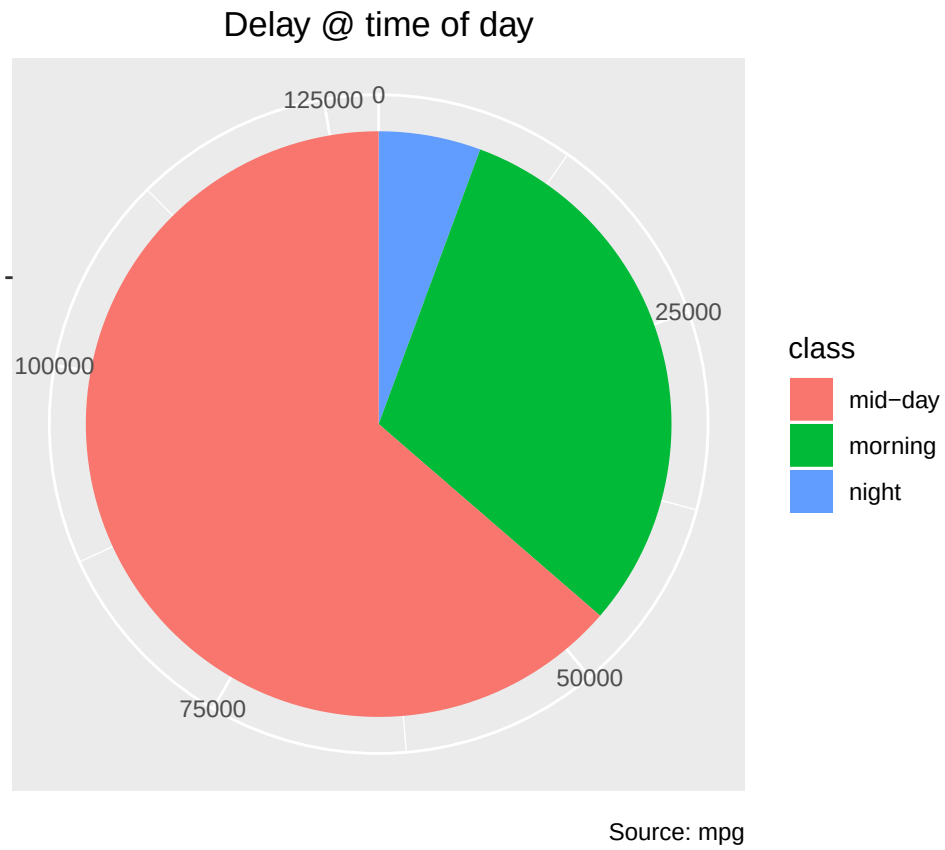
They tend to be very high during the months of june,july and december

5. We'd also like to know how much do delays depend on the time of day. Are there more delays in foggy morning hours? Late night when all the daily delays may accumulate? Create a visualization (graph or table) using a different approach than what you did above.

```r
# partitioning the hours of the day into three categories and creating a boolean for if flights are del
kksk_<-flights %>% filter(dep_delay>0) %>% mutate(timeofday= ifelse(between(hour,4,12),"morning",ifelse

#plotting a pie chart to analyse how many flights are delayed during what tim of the day
pie <- ggplot(kksk_, aes(x = "", fill = factor(timeofday))) +
  geom_bar(width = 1) +
  theme(axis.line = element_blank(),
        plot.title = element_text(hjust=0.5)) +
  labs(fill="class",
       x=NULL,
       y=NULL,
       title="Delay @ time of day",
       caption="Source: mpg")

pie + coord_polar(theta = "y", start=0)
```



We see that the most delays occur during mid-day followed by mornings and then at night fewest flights are delayed

6. Do you see any problems with these questions (and answers)?

FOR EVERY QUESTION WITH DELAY, IT IS NOT SPECIFIED WHICH DELAY ARE WE CONSIDERING!!

(c) Exploring Data:

1. How many flights were there from NYC airports to Portland in 2013?

```
#selecting flights to portland and then counting
flights%>% filter(dest=="PDX")%>% dim()
```

```
## [1] 1354   19
```

So there are 1354 flights to portland in 2013.

2. How many airlines fly from NYC to Portland?

```
#selecting flights to portland
flights_pdx<-flights%>% filter(dest=="PDX")

unique(flights_pdx$carrier)
```

```
## [1] "DL" "UA" "B6"
```

Three airlines fly from NYC to portland

3.Which are these airlines (and the 2-letter abbreviations)? How many times did each of these go to Portland?

```
#diff types of flights

unique(flights_pdx$carrier)
```

```
## [1] "DL" "UA" "B6"
```

```
flights_pdx %>% group_by(carrier) %>% summarise(tot=n())
```

```
## # A tibble: 3 x 2
##   carrier   tot
##   <chr>   <int>
## 1 B6        325
## 2 DL        458
## 3 UA        571
```

B6 went 325 times DL went 458 times UA went 571 times

4. How many unique airliners fly from NYC to PDX?

```
tail_flights<-flights_pdx %>% filter(is.na(tailnum)==FALSE)
unique(tail_flights$tailnum) %>% length()
```

```
## [1] 491
```

There are 491 unique airliners from NYC to PDX

5. How many different airplanes arrived from each of the three NYC airports to Portland?

```
flights_pdx %>% group_by(origin) %>% summarise(tot=n())
```

```
## # A tibble: 2 x 2
##   origin   tot
##   <chr>  <int>
## 1 EWR      571
## 2 JFK      783
```

571 flights went from EWR and 783 flights went from 571.

6. What percentage of flights to Portland were delayed at departure by more than 15 minutes?

```
#calculate all flights to portland delayed by 15
a <-flights_pdx %>% filter(dep_delay>15) %>% summarise(tot=n())
#calculate all flights to portland
```

```
b <-flights_pdx %>% summarise(tot=n())
#calculating percentage
a/b*100
```

```
##        tot
## 1 26.66174
```

Percentage of flights to portland were delayed at depature were 26.67%

7. And finally answer the question above for each origin airport separately. Is one of the airports noticeably worse than others?

```
flights_pdx %>% filter(dep_delay>15) %>% group_by(origin)%>% summarise(tot=n())
```

```
## # A tibble: 2 x 2
##   origin   tot
##   <chr>  <int>
## 1 EWR      168
## 2 JFK      193
```

```
#making variable a1 the number of flights from EWR
a1<-168
#making variable a2 the number of flights from JFK
a2<-193

# from the previous question we know how many total flights from each airport

#ewr
b1<-571
#jfk
b2<-783
a1/b1*100
```

```
## [1] 29.42207
```

```
a2/b2*100
```

```
## [1] 24.64879
```

There is not significant change in percentages comparing both the airports

(d) Challenge Your Results:

1.4 Think about all this Finally, think about the questions and the analysis. 1. Do you see any issues with data? 2. Ethical concerns? 3. Can these questions be answered? Are these questions meaningful? Your code/explanations here

The way I see it there were not too many errors in this data. Apart from some NAs their seems to be a uniformity along the data. And not too much cleaning was required for the data.

Since all of the variable in data set are public information which can be collected with any privacy or ethical concerns , I coudnt notice any ethical concerns while using data set and performing analysis.

Since I beleive this data is part of a R library which is an open source software it is more or less public information and there are no ethical concerns with using this data.

Yes, I think most of the questions were straightfoward and answerable. Most of them made sense and can be answered meaningfully. Whereas, the need for more data can never be satisfied.