DataFinalProject

2022-07-05

Question: WHAT ARE THE CAUSES OF DELAYED DEPARTURES AT CHICAGO O'HARE INTERNATIONAL AIRPORT???

Notes: Time after 12 am for departure Day of week starts on Monday representing 1 Length in minutes representing the duration of the flight

Final: Data set for ANALYZATION of Delays on Chicago O'hare Airport Departures -Analyze time of departure -Analyze duration of flight -Analyze airline of operation

Abbreviations: Alaska Airlines(AS) American Airlines(AA) Continental Airlines(CO) Delta Airlines(DL) United Airlines (UA)

Coding: -How to convert time after 12 into actual time for the 'Time' column(Possibly function feature from lecture 9?) -How to convert day of the week from numbers into actual days -Insert summary of basic statistics -Trends in form of graphs and charts

```
#Importing Tidyverse
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

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

```
#Reading the CSV
Airlines <- read.csv("~/Desktop/Data Science Final Project/Airlines.csv", stringsAsFactors=TRUE)
#Filtering/Removing
Chicago = Airlines %>% filter(AirportFrom == "ORD")
#Selecting
Chicago = Chicago%>% select(-c("id","Flight"))
#Renaming
Chicago = Chicago %>%rename(`Depart_@` = "Time") %>%rename(`Duration` = "Length")
#Adding column representing duration in terms of hours
Chicago = Chicago %>% mutate(Duration_Hours = Duration/60)
#Adding column representing departure time in hour format
Chicago = Chicago %>% mutate(`Depart_@Time` = `Depart_@` /60)

#Function to change numbers into strings representing days of the week
Days = function(Number){
```

```
  x = length(Number)
  result = rep(0,x)
  for(i in 1:x){
    if(Number[i] == 1){
      result[i] = "Monday"
    } else if(Number[i] == 2){
      result[i] = "Tuesday"
    } else if(Number[i] == 3){
      result[i] = "Wednesday"
    } else if(Number[i] == 4){
      result[i] = "Thursday"
    } else if(Number[i] == 5){
      result[i] = "Friday"
    } else if(Number[i] == 6){
      result[i] = "Saturday"
    } else if(Number[i] == 7){
      result[i] = "Sunday"
    }
  }
  return(result)
 }
Chicago = Chicago %>% mutate(Day_Week = Days(DayOfWeek))
```

```
#Basic Summary of data
summary(Chicago)
```

```
##     Airline      AirportFrom      AirportTo       DayOfWeek
##  MQ     :6634   ORD    :24822   LGA    :  842   Min.   :1.000
##  UA     :5046   ABE    :    0   ATL    :  595   1st Qu.:2.000
##  AA     :4486   ABI    :    0   DCA    :  582   Median :4.000
##  OO     :3171   ABQ    :    0   EWR    :  580   Mean   :3.921
##  XE     :2060   ABR    :    0   PHL    :  575   3rd Qu.:5.000
##  YV     : 899   ABY    :    0   LAX    :  569   Max.   :7.000
##  (Other):2526   (Other):    0   (Other):21079
##     Depart_@         Duration         Delay        Duration_Hours
##  Min.   : 300.0   Min.   : 38.0   Min.   :0.0000   Min.   :0.6333
##  1st Qu.: 590.0   1st Qu.: 86.0   1st Qu.:0.0000   1st Qu.:1.4333
##  Median : 830.0   Median :115.0   Median :0.0000   Median :1.9167
##  Mean   : 834.7   Mean   :130.1   Mean   :0.4797   Mean   :2.1681
##  3rd Qu.:1079.0   3rd Qu.:155.0   3rd Qu.:1.0000   3rd Qu.:2.5833
##  Max.   :1350.0   Max.   :560.0   Max.   :1.0000   Max.   :9.3333
##
##   Depart_@Time      Day_Week
##  Min.   : 5.000   Length:24822
##  1st Qu.: 9.833   Class :character
##  Median :13.833   Mode  :character
##  Mean   :13.912
##  3rd Qu.:17.983
##  Max.   :22.500
##
```

```
#Delay filtered sets
Delay = Chicago%>% filter(Delay == 1)
Delay = Delay %>% select(c("Airline","Delay"))
```

Airline specific delay/no delay subsets

```
#Numbers of Delay per Airline
table(Delay$Airline)
```

```
##
##   9E   AA   AS   B6   CO   DL   EV   F9   FL   HA   MQ   OH   OO   UA   US   WN
##   43 2632   85   74  323  300  256    0    0    0 2827   76 1720 1853  252    0
##   XE   YV
## 1030  435
```

```
length(Delay$Airline)
```

```
## [1] 11906
```

Delay/No delay tables

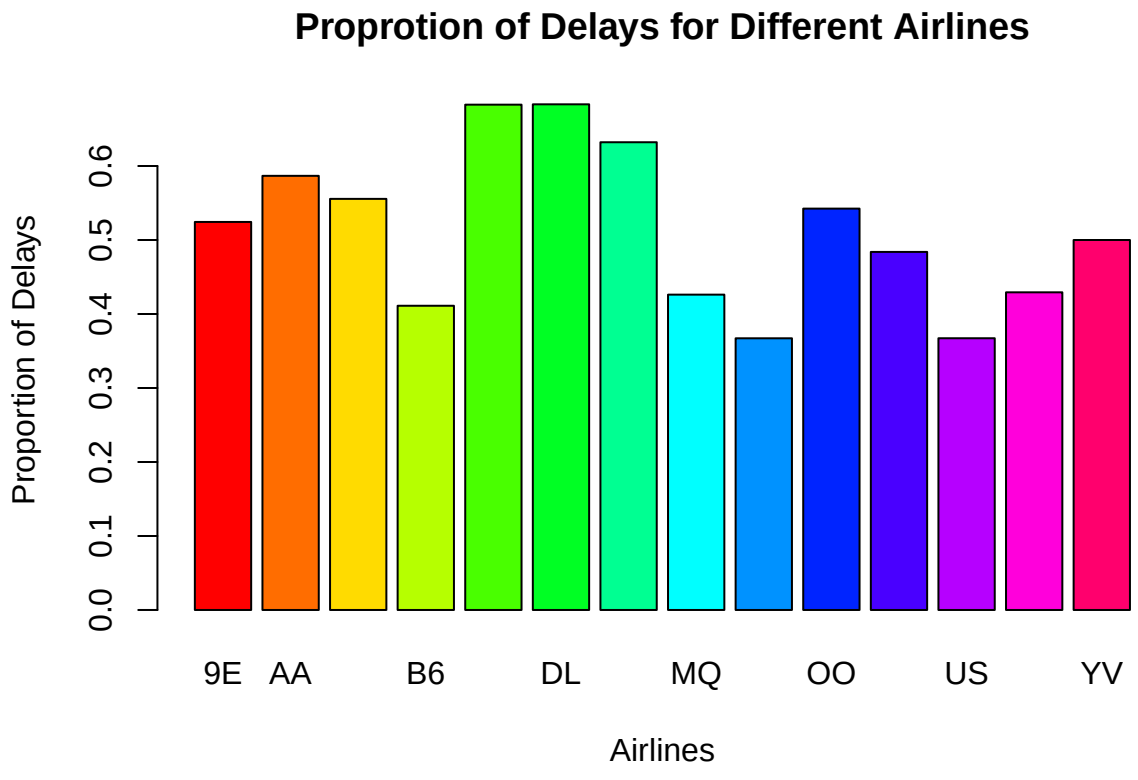
```
#Endeavor Air Subset
`9E_Total` = Chicago%>% filter(Airline == "9E")
E_delAvg <- 43/82
#American Airlines Subset
AA_Total = Chicago%>% filter(Airline == "AA")
AA_delAvg <- 2632/4486
#Alaska Airlines Subset
AS_Total = Chicago%>% filter(Airline == "AS")
AS_delAvg <- 85/153
#JetBlue Subset
`B6_Total` = Chicago%>% filter(Airline == "B6")
B6_delAvg <- 74/180
#Continental Airlines Subset
CO_Total = Chicago%>% filter(Airline == "CO")
CO_delAvg <- 323/473
#Delta Subset
DL_Total = Chicago%>% filter(Airline == "DL")
DL_delAvg <- 300/439
#Eva Air Subset
EV_Total = Chicago%>% filter(Airline == "EV")
EV_delAvg <- 256/405
#Envoy Air Subset
MQ_Total = Chicago%>% filter(Airline == "MQ")
MQ_delAvg <- 2827/6634
#PSA Airlines Subset
OH_Total = Chicago%>% filter(Airline == "OH")
OH_delAvg <- 76/207
#Skywest Airlines Subset
OO_Total = Chicago%>% filter(Airline == "OO")
OO_delAvg <- 1720/3171
#US Airways Subset
```

```
US_Total = Chicago%>% filter(Airline == "US")
US_delAvg <- 252/587
#JSX Airlines Subset
XE_Total = Chicago%>% filter(Airline == "XE")
XE_delAvg <- 1030/2060
#Mesa Airlines Subset
YV_Total = Chicago%>% filter(Airline == "YV")
YV_delAvg <- 435/899
#United Airlines Subset
UA_Total = Chicago%>% filter(Airline == "UA")
UA_delAvg <- 1853/5046

airlines <- c("9E", "AA", "AS","B6","CO","DL","EV","MQ","OH","OO","US","XE","YV","UA")
delays <- c(E_delAvg, AA_delAvg, AS_delAvg, B6_delAvg, CO_delAvg, DL_delAvg,
            EV_delAvg, MQ_delAvg, OH_delAvg, OO_delAvg, UA_delAvg, US_delAvg, XE_delAvg, YV_delAvg)
```

Graphs/Charts

```
#Bar Plot
barplot(delays~airlines, main= "Proprotion of Delays for Different Airlines",
        xlab="Airlines", ylab="Proportion of Delays", col=rainbow(14))
```



Day of week delay analysis

```
DayofWeeks = Chicago %>% select(c("Day_Week","Delay"))
#Table of Delays per day
table(DayofWeeks$Day_Week)
```

```
##
##    Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
##      3902      3354      2688      3208      4177      3316      4177
```

```
length(DayofWeeks$Day_Week)
```

```
## [1] 24822
```

Number of delays per week using sampling

```
#Data sets representing the delays for each day
Monday = DayofWeeks%>% filter(Day_Week == "Monday")
Tuesday = DayofWeeks%>% filter(Day_Week == "Tuesday")
Wednesday = DayofWeeks%>% filter(Day_Week == "Wednesday")
Thursday = DayofWeeks%>% filter(Day_Week == "Thursday")
Friday = DayofWeeks%>% filter(Day_Week == "Friday")
Saturday = DayofWeeks%>% filter(Day_Week == "Saturday")
Sunday = DayofWeeks%>% filter(Day_Week == "Sunday")
```

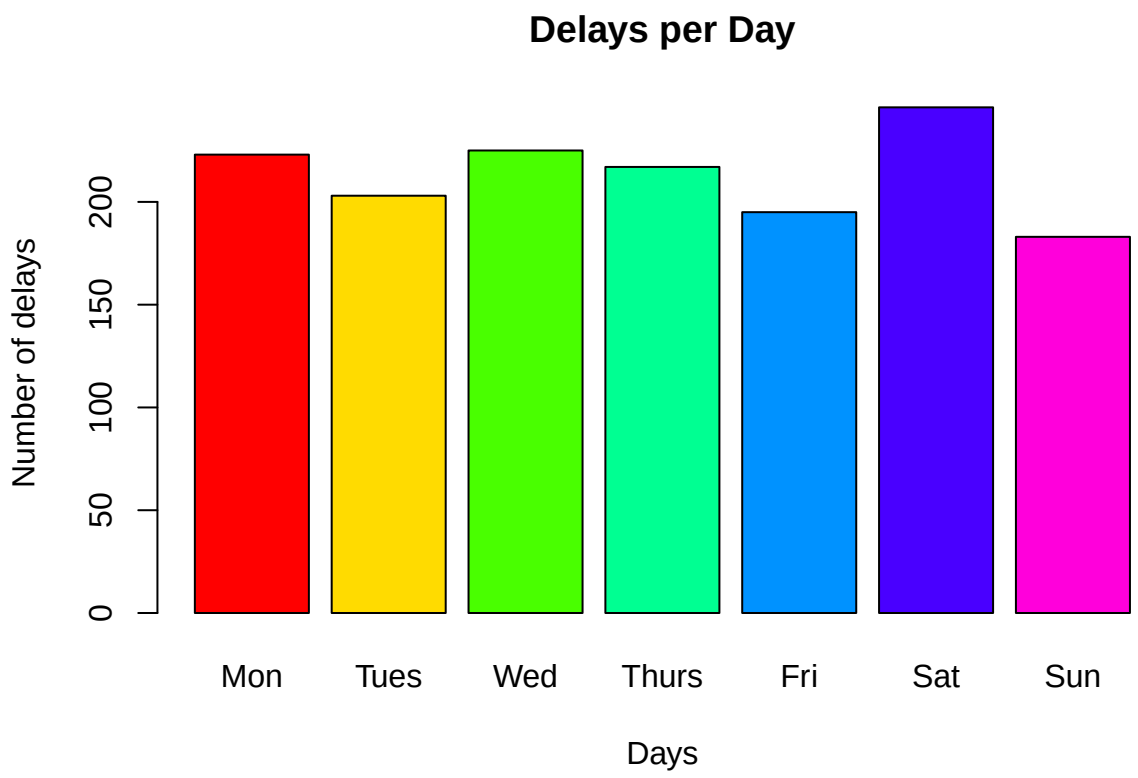
```
#Sampling delays for each day
set.seed(143572)
sampleMon = Monday[sample(nrow(Monday), size = 429, replace = FALSE),]
sampleMonDel = sampleMon %>% filter(Delay == 1)
Mon_del <- 223
sampleTues = Tuesday[sample(nrow(Tuesday), size = 429, replace = FALSE),]
sampleTuesDel = sampleTues %>% filter(Delay == 1)
Tues_del <- 203
sampleWed = Wednesday[sample(nrow(Wednesday), size = 429, replace = FALSE),]
sampleWedDel = sampleWed %>% filter(Delay == 1)
Wed_del <- 225
sampleThurs = Thursday[sample(nrow(Thursday), size = 429, replace = FALSE),]
sampleThursDel = sampleThurs %>% filter(Delay == 1)
Thurs_del <- 217
sampleFri = Friday[sample(nrow(Friday), size = 429, replace = FALSE),]
sampleFriDel = sampleFri %>% filter(Delay == 1)
Fri_del <- 195
sampleSat = Saturday[sample(nrow(Saturday), size = 429, replace = FALSE),]
sampleSatDel = sampleSat %>% filter(Delay == 1)
Sat_del <- 246
sampleSun = Sunday[sample(nrow(Sunday), size = 429, replace = FALSE),]
sampleSunDel = sampleSun %>% filter(Delay == 1)
Sun_del <- 183

#Combining into single sample
sampleStr  <- c(Mon_del, Tues_del, Wed_del, Thurs_del, Fri_del, Sat_del, Sun_del)
sampleStr
```

```
## [1] 223 203 225 217 195 246 183
```

Plotting days per week delays

```
barplot(sampleStr,main = "Delays per Day", names.arg = c("Mon", "Tues", "Wed", "Thurs", "Fri","Sat", "Su
```

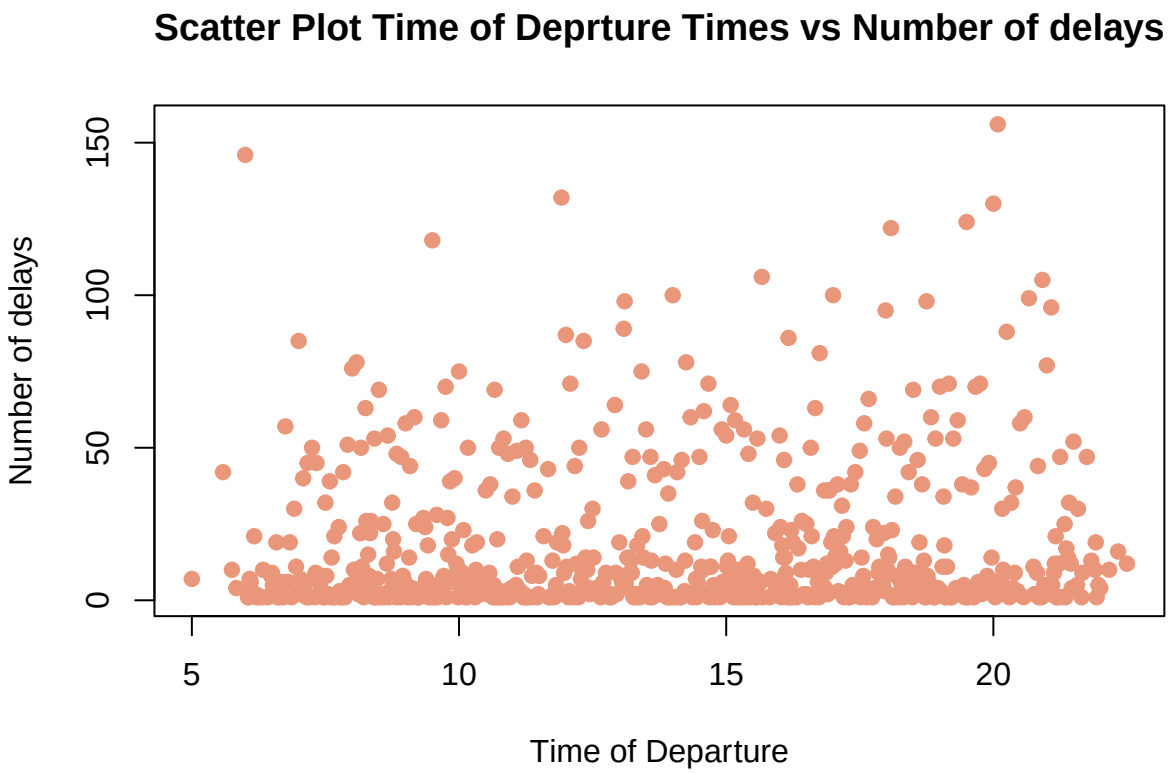


Time of Departure Delay Analysis

```
time = Chicago %>% select("Depart_@Time", "Delay")
time = time %>% filter(Delay == 1)
```

```
tot_Times = data.frame(table(time$`Depart_@Time`))
tot_Times$Var2 = as.numeric(as.character(tot_Times$Var1))

plot(tot_Times$Var2,tot_Times$Freq, xlab = "Time of Departure", ylab = "Number of delays", main = "Scat
```


Scatter Plot Time of Deprture Times vs Number of delays
Number of delays
0
50
100
150
5
10
15
20
Time of Departure