EDA project on flight data

Bhargav Teja Jakku

2/24/2022

Work with a dataset of all domestic outbound flights from Dulles International Airport in 2016.

Airports depend on accurate flight departure and arrival estimates to maintain operations, profitability, customer satisfaction, and compliance with state and federal laws. Flight performance, including departure and arrival delays must be monitored, submitted to the Federal Aviation Agency (FAA) on a regular basis, and minimized to maintain airport operations. **The FAA considered a flight to be delayed if it has an arrival delay of at least 15 minutes**.

The executives at Dulles International Airport have hired you as a Data Science consultant to perform an exploratory data analysis on all domestic flights from 2016 and produce an executive summary of your key insights and recommendations to the executive team.

Before you begin, take a moment to read through the following airline flight terminology to familiarize yourself with the industry: Airline Flight Terms

Dulles Flights Data

The `flights_df` data frame is loaded below and consists of 33,433 flights from IAD (Dulles International) in 2016. The rows in this data frame represent a single flight with all of the associated features that are displayed in the table below.

```
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## Warning: package 'tidyr' was built under R version 4.1.2

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

```
library(ggplot2)
library(RColorBrewer)

flights_df <- readRDS(url('https://gmubusinessanalytics.netlify.app/data/dulles_flights.rds'))
```

Raw Data

```
flights_df
```

```
## # A tibble: 33,433 x 22
##    scheduled_flight_date month_numeric month     day weekday airline   tail_num
##    <date>                        <dbl> <fct>   <dbl> <fct>   <fct>     <fct>
##  1 2016-01-01                        1 January     1 Friday  Southwest N569WN
##  2 2016-01-01                        1 January     1 Friday  Southwest N466WN
##  3 2016-01-01                        1 January     1 Friday  Southwest N458WN
##  4 2016-01-01                        1 January     1 Friday  Southwest N922WN
##  5 2016-01-01                        1 January     1 Friday  Southwest N711HK
##  6 2016-01-01                        1 January     1 Friday  American  N473AA
##  7 2016-01-01                        1 January     1 Friday  American  N3HUAA
##  8 2016-01-01                        1 January     1 Friday  American  N564AA
##  9 2016-01-01                        1 January     1 Friday  American  N4UBAA
## 10 2016-01-01                        1 January     1 Friday  American  N837AW
## # ... with 33,423 more rows, and 15 more variables: flight_num <dbl>,
## #   dest_airport_name <fct>, dest_airport_city <fct>, dest_airport_state <fct>,
## #   dest_airport_region <fct>, sch_dep_time <dbl>, dep_time <dbl>,
## #   dep_delay <dbl>, taxi_out <dbl>, wheels_on <dbl>, taxi_in <dbl>,
## #   arrival_time <dbl>, sch_arrival_time <dbl>, arrival_delay <dbl>,
## #   distance <dbl>
```

Exploratory Data Analysis

Executives at this company have hired you as a data science consultant to evaluate their flight data and make recommendations on flight operations and strategies for minimizing flight delays.

You must think of **at least 8 relevant questions** that will provide evidence for your recommendations.

The goal of the analysis is discovering which variables drive the differences between flights that are early/on-time vs. flights that are delayed.

The listed questions that would be explored include:

- Are flight delays affected by taxi-out time?
- which weekdays people prefer to fly frequently? How often people face more delayed departures on preferred week days.
- Do certain airlines lead to greater taxi out times (i.e. traffic jams on the runways)?
- Are certain destination or airlines prone to delays?
- Flights of which country delayed frequently? Test the significance for delayed flights in those countries.
- Flights of which country depart on time? Check its significance.
- What is the association between distance and arrival delay?
- Is there any influence of a flight being depart delay on arrive delay too?

Each question must be answered with supporting evidence from summary statistics and plots.

Question 1

Question: Are flight delays affected by taxi-out time?

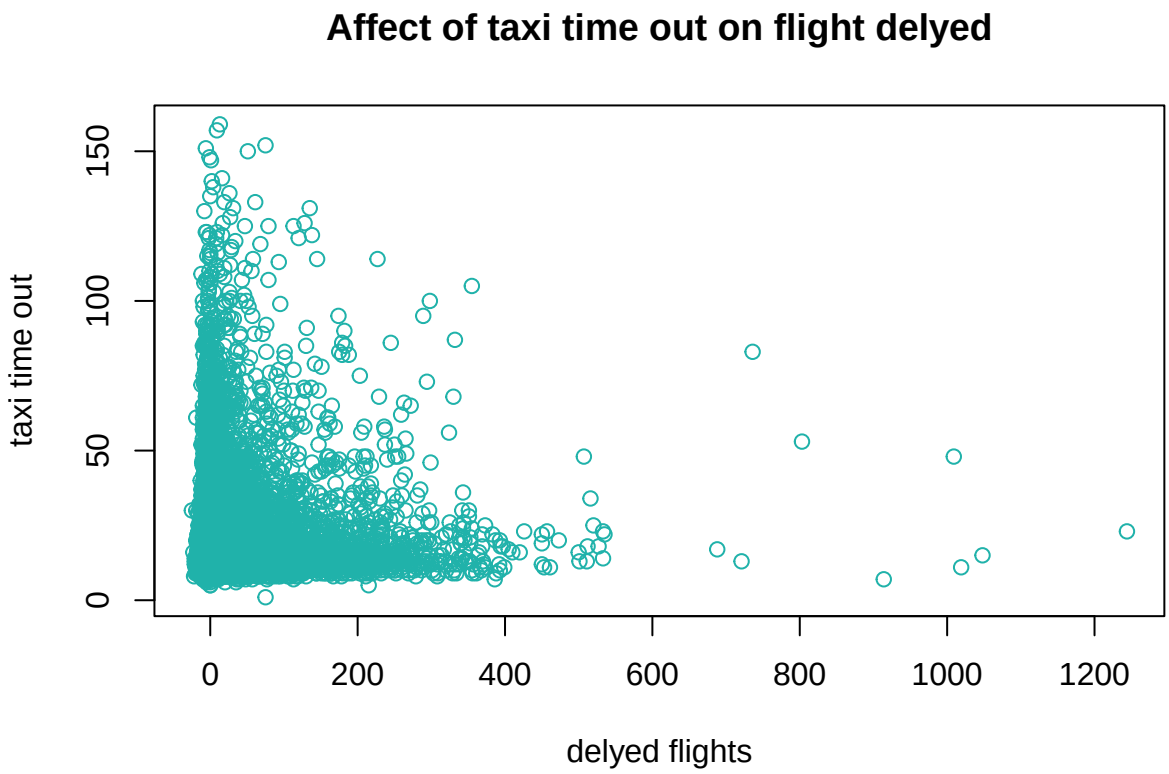
Answer: We are going to test the hypothesis that H0: True correlation between flight delays and taxi time out is equal to 0. H1: True correlation between flight delays and taxi time out is not equal to 0. The hypothesis is tested at 5% level of significance by applying Pearson's product-moment correlation test. We will reject H0 if p-value of respective test is less than level of significance. By observing the summary statistics of correlation test, the p-value of test found to be 2.2e-16 < 0.05 (level of significance). Hence we reject H0 and conclude that the true correlation between flight delays and taxi time out is not equal to 0. The correlation value between them is being observed to be 0.11, which is a sort of weak correlation. So taxi time out affect flight delays a very little.

We can see this little affect with the help of scatter plot as follows. The plot shows a very small association between two selected variables.

```
x1 <- flights_df$dep_delay
x2 <- flights_df$taxi_out
cor.test(x1,x2)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x1 and x2
## t = 19.613, df = 33431, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.09604893 0.11724351
## sample estimates:
##       cor
## 0.1066583
```

```
plot(x1, x2,col = "lightseagreen", xlab="delyed flights", ylab = "taxi time out", main = "Affect of taxi
```



Question 2

Question: which weekdays people prefer to fly frequently? How often people face more delayed departures on preferred week days.

Answer: We first see the most of week days when people preferred to fly and then we check the number of delayed flights on those days and see its distribution with the help of box plot. The people preferred to fly mostly on Friday and on these days they often face delayed departures. We can see that the data for delayed departures on Friday have no specific pattern in its distribution but have many outliers.

```
length(which(flights_df$weekday == "Monday"))
```

```
## [1] 4914
```

```
length(which(flights_df$weekday == "Tuesday"))
```

```
## [1] 4917
```

```
length(which(flights_df$weekday == "Wednesday"))
```

```
## [1] 4973
```

```
length(which(flights_df$weekday == "Thursday"))
```

```
## [1] 4993
```

```
length(which(flights_df$weekday == "Friday"))
```

```
## [1] 5015
```

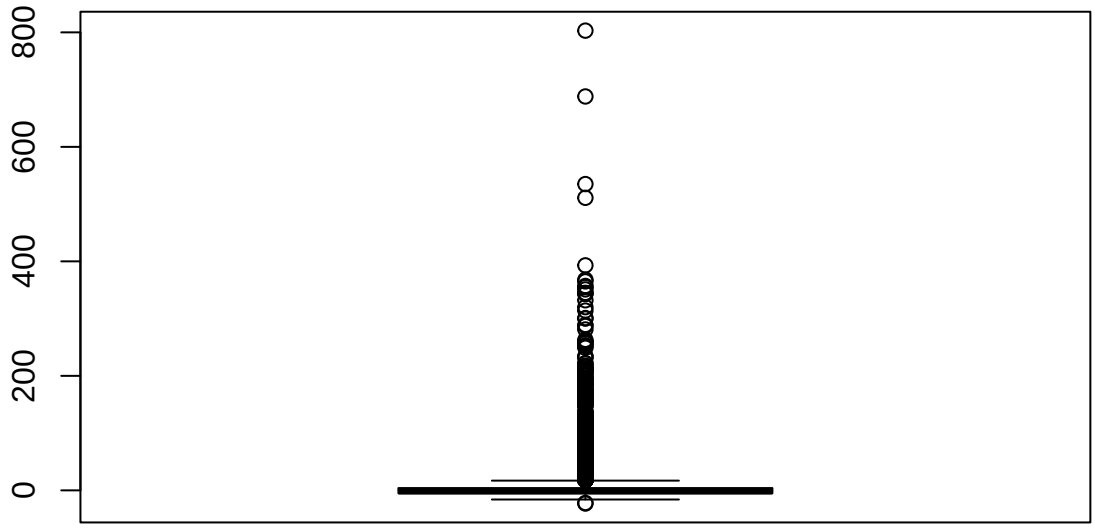
```
length(which(flights_df$weekday == "Saturday"))
```

```
## [1] 3911
```

```
length(which(flights_df$weekday == "Sundayday"))
```

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

```
y1 <- flights_df$weekday[(flights_df$weekday == "Friday")]
y2 <- flights_df$dep_delay[(flights_df$weekday == "Friday")]
boxplot(y2,col = "mediumorchid1", xlab="Delayed flights on Friday")
```



Question 3

Question: Do certain airlines lead to greater taxi out times (i.e. traffic jams on the runways)?

Answer: The highest period of taxi time out due to or traffic jam or any other reason is 159 mints. The airline on this period is observed to be Southwest. Hence, we can say that the Southwest airline leads to be greater taxi out time, which could cause the delayed flights.

```
z1 <- which.max(flights_df$taxi_out)
flights_df$airline[z1]
```

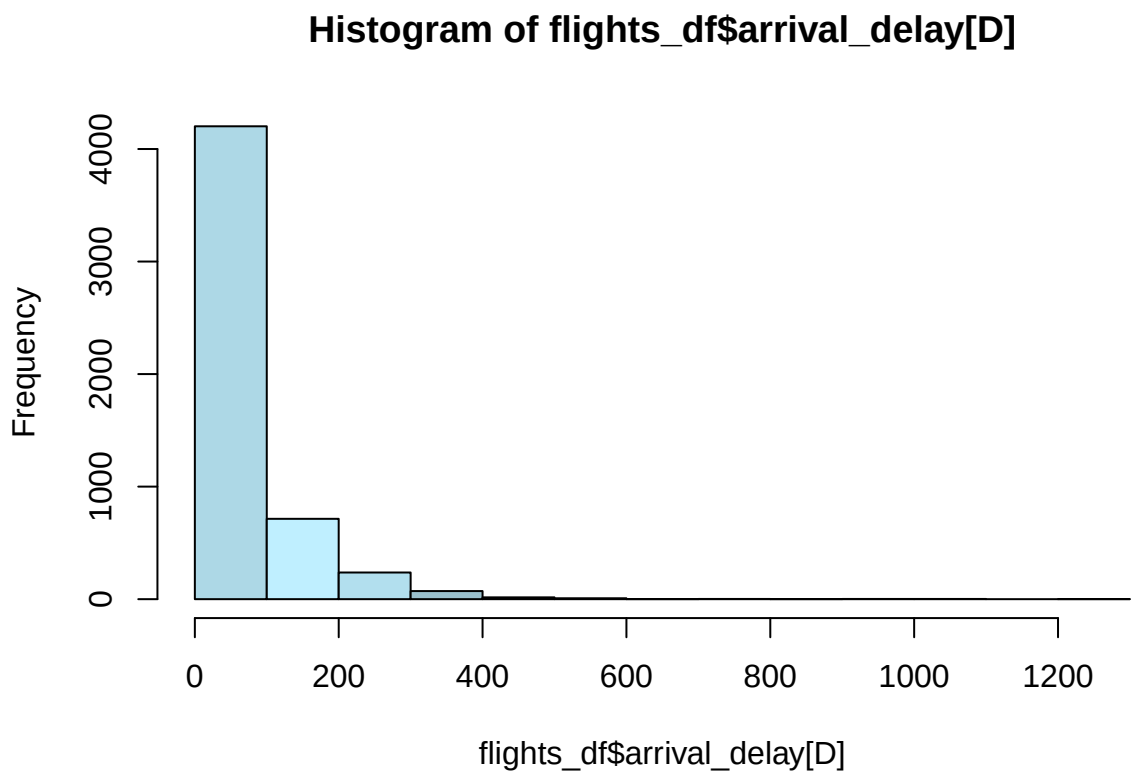
```
## [1] Southwest
## 10 Levels: United American Delta Southwest JetBlue Virgin America ... Alaska
```

Question 4

Question: Are certain destination or airlines prone to delays?

Answer: The airline considered to be delayed by Federal Aviation Agency (FAA) if it has an arrival delay of at least 15 mints. The delayed arrivals are displayed in a following histogram figure, which follows a positively skewed distribution. So the delayed arrival flights observed to be 5258 in number.

```
D <- which(flights_df$arrival_delay >= 15)
hist(flights_df$arrival_delay[D], col = c("lightblue","lightblue1","lightblue2","lightblue3"))
```



```
D_flight <- flights_df$airline[D]
```

Question 5

Question: Flights of which country delayed frequently? Test the significance for delayed flights in those countries.

Answer: We wish to see the name of countries that delayed frequently. The significance of delayed flight (whose arrival time is at lease 15 minutes) is tested bt one sample t test. The summary statistics show as follows. We form the null and alternative hypothesis as H0: The frequent delay flights of respective countries is insignificant. H1: The frequent delay flights of respective countries is significant. Level of significance: 0.05 Test statistics: to be used here is one sample t test Critical region: If the p-value of test is less than significance level then we will reject H0. Calculations: The observed p-value of the test is 2.2e-16. Conclusion: As we can see that the observed p-value of the test is less than significance level i.e., 2.2e-16 $<$ 0.05. Hence we reject H0 and conclude that the frequent delay in flights of respective countries is significant.

```
Countries <- flights_df$dest_airport_city[D]
AV <- flights_df$arrival_delay[D]
t.test(AV)
```

```
##
##  One Sample t-test
##
## data:  AV
## t = 65.456, df = 5257, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  68.83203 73.08238
## sample estimates:
## mean of x
##  70.95721
```

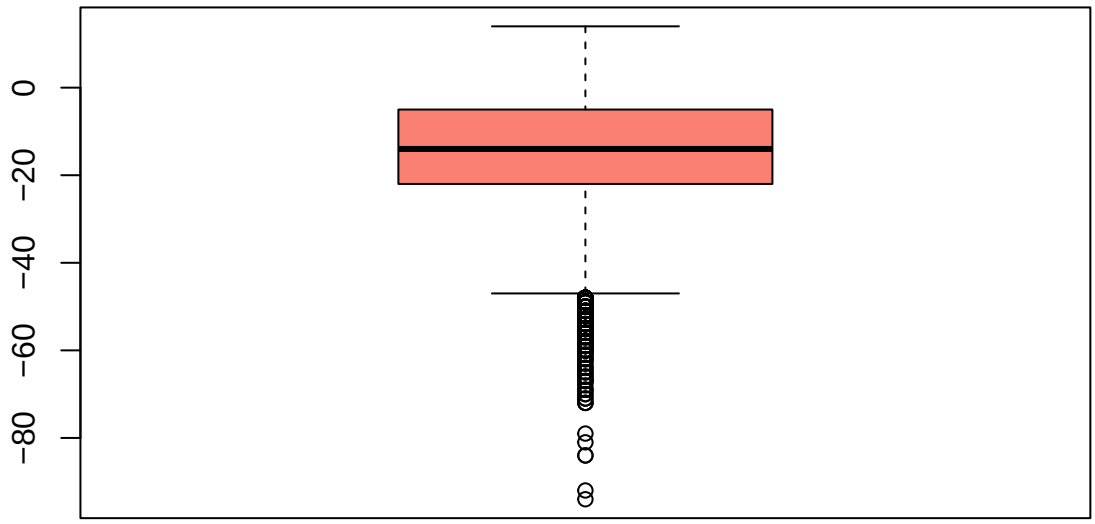
Question 6

Question: Flights of which country depart on time? Check its significance.

Answer: Now we will test the significance for those countries who have on time departure. These are 28175 in number. The box plot of on time departure data for selective countries shows a symmetric pattern in data. However, the data have some outliers too. The significance test has the following interpretations based on summary statistics results.

H0: The on time country departure flights is insignificant. H1: The on time country departure flights is significant. Level of significance: 0.05 Test statistics: to be used here is one sample t test Critical region: If the p-value of test is less than significance level then we will reject H0. Calculations: The observed p-value of the test is 2.2e-16. Conclusion: As we can see that the observed p-value of the test is less than significance level i.e., 2.2e-16 $<$ 0.05. Hence we reject H0 and conclude that the on time departure flights of respective countries is significant.

```
on_time <- which(flights_df$arrival_delay <= 14)
boxplot(flights_df$arrival_delay[on_time], col = c("salmon"))
```




```
t.test(on_time)
```

```
##
##  One Sample t-test
##
## data:  on_time
## t = 285.67, df = 28174, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  16424.92 16651.87
## sample estimates:
## mean of x
##   16538.4
```

Question 7

Question: What is the association between distance and arrival delay?

Answer: The association between distance and arrival delay is assessed with the help of correlation test and graph. The hypothesis are defined as follows.

H0: True correlation between arrival delays and distance is equal to 0. H1: True correlation between arrival delays and distance is not equal to 0. Level of significance: 0.05 Test statistics: to be used here is Pearson correlation test. Critical region: If the p-value of test is less than significance level then we will reject H0. Calculations: The observed p-value of the test is 5.174e-13. Conclusion: As we can see that the observed

p-value of the test is less than significance level i.e., 5.174e-13 < 0.05. Hence we reject H0 and conclude that the true correlation between arrival delays and distance is not equal to 0. There is observed to be strong negative association between them with correlation value of -0.0395. The correlation plot also indicate a weak correlation between two variables.

```
a <- flights_df$distance
b <- flights_df$arrival_delay
df <- data.frame(x=a, y=b)
cor.test(a,b)
```

```
##
##  Pearson's product-moment correlation
##
## data:  a and b
## t = -7.2235, df = 33431, p-value = 5.174e-13
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.05017421 -0.02876917
## sample estimates:
##         cor
## -0.03947622
```

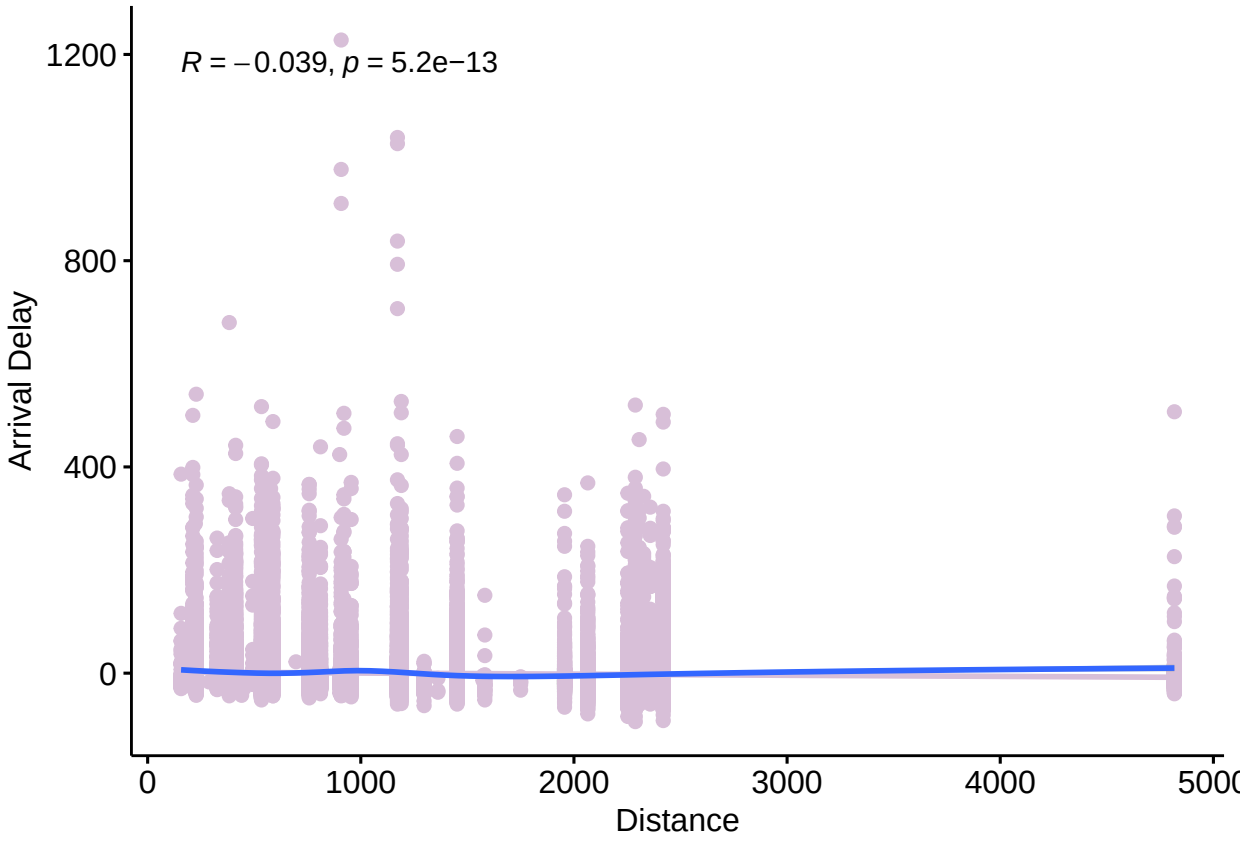
```
library("ggpubr")
```

```
## Warning: package 'ggpubr' was built under R version 4.1.2
```

```
ggscatter(df, x = "x", y = "y", color  = "thistle",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Distance", ylab = "Arrival Delay") + geom_smooth(method="gam")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## `geom_smooth()` using formula 'y ~ s(x, bs = "cs")'
```



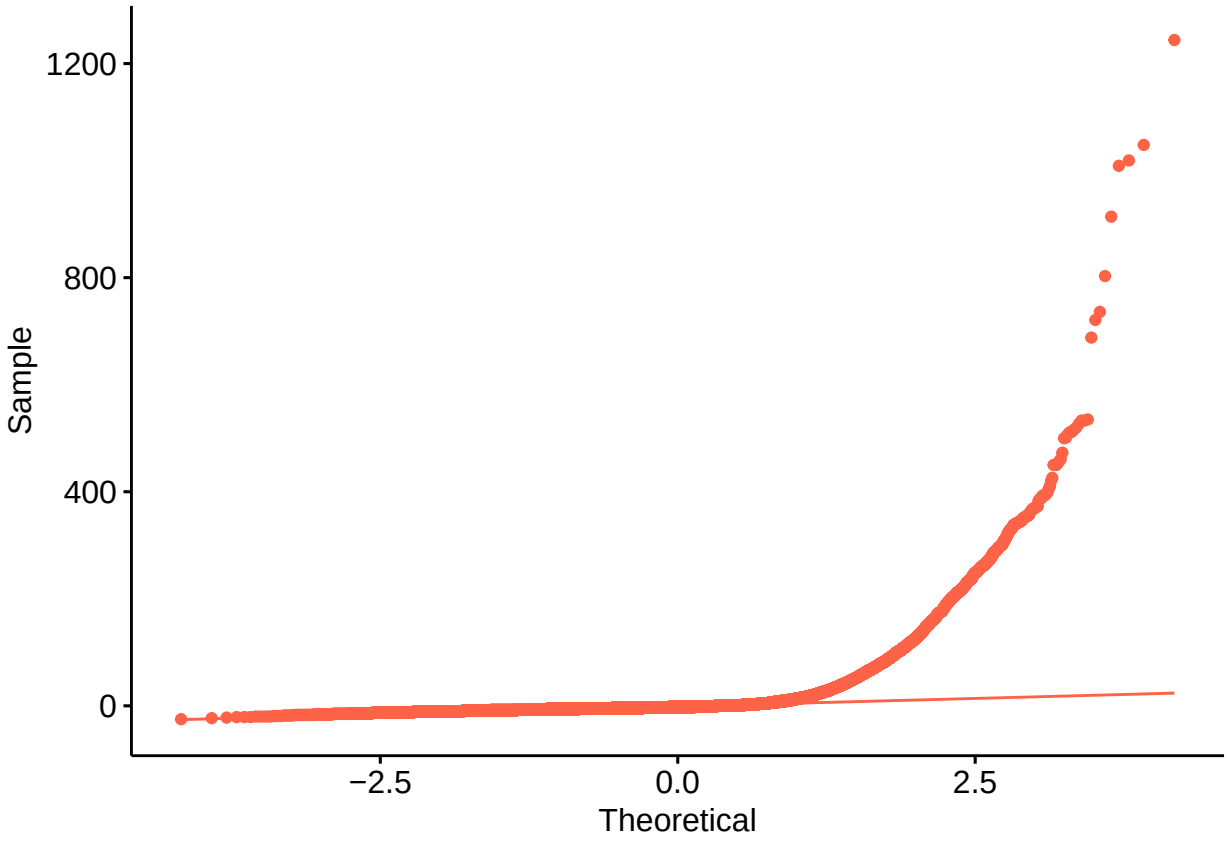
Question 8

Question: Is there any influence of a flight being depart delay on arrive delay too?

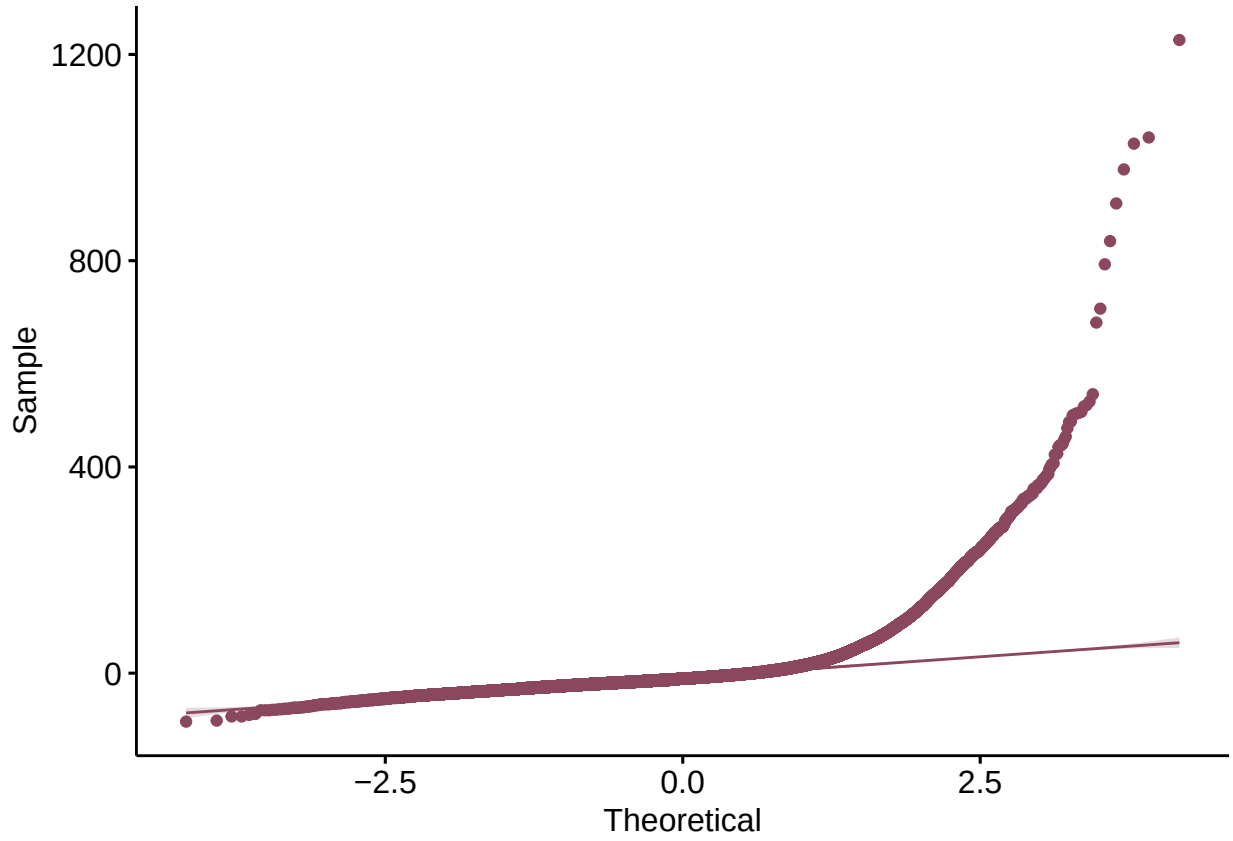
Answer: Now we are going to see the relationship between depart delay on and arrival delay. At first we will check the normality of two variables individually. The data of both variables is not normally distributed. However the significance of two variables is accessed with the help of two sample t test. We form the hypothesis as follows. H0: The difference between two group means is zero. H1: The difference between two group means is not zero. Level of significance: 0.05 Test statistics: to be used here is two sample t test Critical region: If the p-value of respective test is less than level of significance then we will reject H0. Calculation: The observed p-value of test 2.2e-16 Conclusion: Its been observed that 2.2e-16 $<$ 0.05 so we reject H0 and conclude that the difference between two group means is not equal to zero.

However, the correlation plot shows a linear trend with perfect positive correlation among them. So there is a very strong association between a flight who depart delay, will arrive delay with the correlation value of 0.93.

```
ggqqplot(flights_df$dep_delay, color = "tomato")
```




```
ggqqplot(flights_df$arrival_delay, color = "palevioletred4")
```

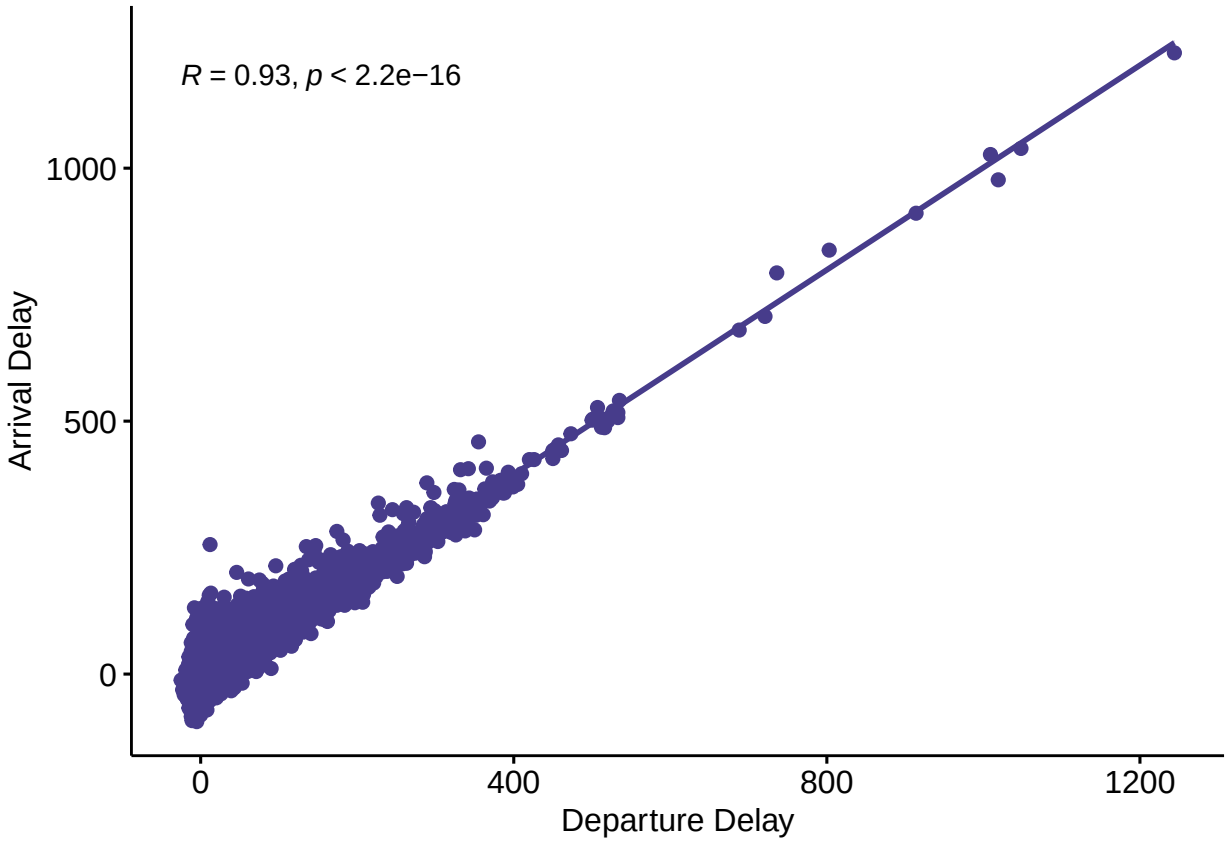



```
x <- flights_df$dep_delay
y <- flights_df$arrival_delay
t.test(x, y)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 28.511, df = 66366, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   8.957839 10.280399
## sample estimates:
##  mean of x  mean of y
##  9.0705291 -0.5485897
```

```
dat <- data.frame(x=x, y=y)
ggscatter(dat, x= "x", y="y", color = "slateblue4",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Departure Delay", ylab = "Arrival Delay")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Summary of Results

Executive Summary

Introduction: In this study we are going to explore the flight data with the following objectives. 1. Study the data with the help of exploratory data analysis techniques. 2. What factors involve in a flight delay?

Results

The exploratory analysis consists of observing the relation, or association between different variables. We visualize them with the help of graphs and plots for clear understanding. Our findings say that there is a little correlation present between flight delays and taxi time out. The majority of people preferred to fly mostly on Friday and on these days they often face delayed departures. We can see that the data for delayed departures on Friday have no specific pattern in its distribution but have many outliers. The highest period of taxi time out due to traffic jams or any other reason is 159 minutes. The airline on this period is observed to be Southwest. Hence, we can say that the Southwest airline leads to be greater taxi-out time, which could cause delayed flights.

On the other hand, if we observe the delayed arrival flights then they are 5258 in number. Among them, the most famous flights are Chicago, Atlanta, Los Angeles, San Francisco, Miami, and New York City. However, about 28175 countries have their flights on time. If we see the association between distance and arrival delay then there is a very small amount of correlation between them. The flight variables of departing delay on arrival delay exhibit a linear trend with a perfect positive correlation among them. So there is a very strong association between a flight that departs delay, will arrival delay with a 0.93 correlation.

Recommendations:

A flight is supposed to be on time if it departs and arrives within 15 minutes of its scheduled time. For on-time flights 1. Passengers should prefer weekdays instead of weekends. 2. Be on time to avoid delays as it's clear from the analysis if a flight departs delayed, then will arrive delayed too. 3. Airlines can reduce reported flight delays simply by inflating the flight's anticipated arrival time. And evidence suggests that airlines have done just that.