EAST DELTA UNIVERSITY (EDU)
Course Title: Statistics for Business, Course Code**: CBD 501**
Assignment: 02 – Spring 2024
Date of Submission: **11:59PM, 24 May 2024,** Total Marks**: 20**
Submitted by: Shakib-241001661

Highlights of the Dataset

This is an Airline Dataset having 227,496 flight information of an US Airport in the year 2011. There is total 21 columns in the dataset. Below is the brief of the columns.

Column Name	Description
Year	the year of departure
Month	the month of departure
DayofMonth	the day of the month of departure
DayOfWeek	the day of week of departure
DepTime	departure time in local time
ArrTime	arrival time in local time
UniqueCarrier	unique abbreviation for a carrier
FlightNum	flight number
TailNum	airplane tail number
ActualElapsedTime	elapsed time of flight, in minutes
AirTime	flight time, in minutes
ArrDelay	arrival delays in minutes
DepDelay	departure delays in minutes
Origin	origin airport code
Dest	destination airport code
Distance	distance of flight, in miles
TaxiIn	taxi in time in minutes
TaxiOut	taxi out time in minutes
Cancelled	cancelled indicator: 1 = Yes, 0 = No
CancellationCode	A = carrier, B = weather, C = national air system, D = security
Diverted	diverted indicator: 1 = Yes, 0 = No

Requirements

i. Answer the questions given in next page and create a PDF report appropriately explaining the results. Add all the R visualizations in the relevant questions to use in your explanation.
ii. Submit both PDF & R script.

Loading Library and Dataset Preparation

```
#install tidyverse if not available
#install.packages("tidyverse")
#load library
library(tidyverse)
#load csv from github(uploaded for Assignment_02_Group_04)
airline <-
read.csv("https://raw.githubusercontent.com/shakibed/CBD501_Assignment_02_Group_04/main/dataset.csv", stringsAsFactors = FALSE)
#summary of data set
View(airline)
```

Question 1: Consider the months as Winter (January, Feburuary, March) and Summer (June, July, August) seasons. Do planes depart delayed more in winter than in summer? If yes, how significant the delay pattern is? *(Marks: 2)*

Hints: Calculate Mean, Median & SD of both groups. To test significance, use T-test.

Answer:

```
#winter data (January, Feburuary, March)
winter <- subset(airline,subset = (Month==1 | Month==2 | Month ==3))
winter
#mean of winter flights
round(mean(winter$DepDelay, na.rm =T),2)
#median of winter flights
round(median(winter$DepDelay, na.rm =T),2)
#standard deviation of winter flights
round(sd(winter$DepDelay,na.rm=T),2)


#Summer (June,July, August August)
summer <- subset(airline,subset = (Month==6 | Month==7 | Month ==8))
summer
#mean of summer flights
round(mean(summer$DepDelay, na.rm =T),2)
#median of summer flights
round(median(summer$DepDelay, na.rm =T),2)
#standard deviation of summer flights
round(sd(summer$DepDelay,na.rm=T),2)


#To assess whether there is a significant difference in departure delay
between summer and winter flights, we are using a t-test where we compare
whether the mean of departure delays in winter flights is significantly
different from the mean of departure delays in summer flights.


t.test(winter$DepDelay, summer$DepDelay, mu = 0, paired = F, var.equal = F,
conf.level = 0.95)

# Two sample t-test indicates a significant difference in departure delays
between winter and summer, with winter departures experiencing shorter delays
```

```
(mean = 8.97 minutes) compared to summer departures (mean = 10.76 minutes),
t(113811) = -10.845, p < 2.2e-16.
```

Output and Explanation:

Winter

- Mean of winter flights is: 8.97
- Median of winter flights is: 0
- Standard deviation of winter flights is: 26.73

Summer

- Mean of summer flights is: 10.76
- Median of summer flights is: 1
- Standard deviation of summer flights is: 28.9

- Winter departures experience shorter delays (mean = 8.97 minutes) compared to summer departures (mean = 10.76 minutes) which indicates delayed departure of flights more in summer than in winter.
- To assess whether there is a significant difference in departure delay between summer and winter flights, we are using a t-test.
- Two sample t-test indicates a significant difference in departure delays between winter and summer, with winter departures experiencing shorter delays (mean = 8.97 minutes) compared to summer departures (mean = 10.76 minutes), t(113811) = -10.845, p < 2.2e-16

Question 2: Show the actual elapsed time of flights going to Atlanta Airport (ATL) in a histogram. Please add additional reference lines showing the mean and median. *(Marks: 2)*

Hints: First step is to create the histogram and then add mean & median lines using "abline".

Answer:

```
# we are creating histogram with flights going to Atlanta Airport(ATL)
hist(airline$ActualElapsedTime[airline$Dest=="ATL"], breaks=100, col =
"grey",
     border = "black", main = "Histogram of elapsed time for flights to
Atlanta", xlab=" Elaplsed time of Flight(in min)" ,ylab="Number of flights")

#For the mean (red line):
abline(v = mean(airline$ActualElapsedTime[airline$Dest == "ATL"], na.rm = T),
col = "red", lwd = 2)

#For the median (blue line):
abline(v = median(airline$ActualElapsedTime[airline$Dest == "ATL"], na.rm =
T), col = "blue", lwd = 2)
```

Output and Explanation:

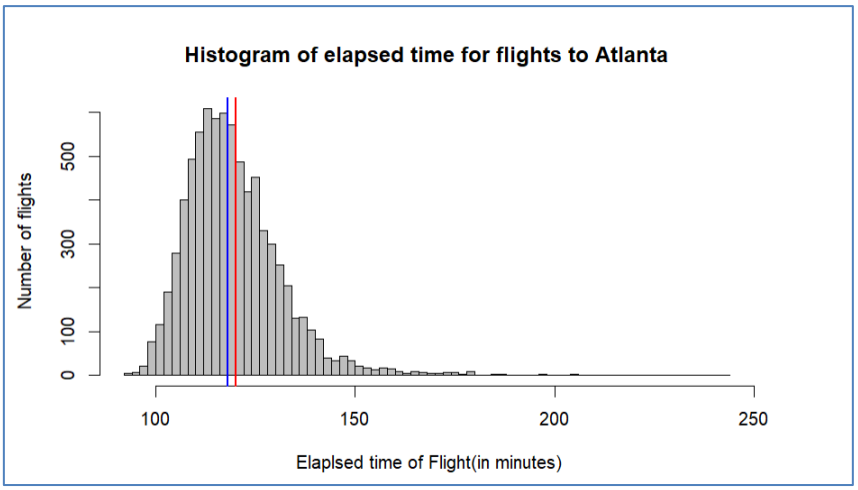


Figure: Histogram of elapsed time for flights to Atlanta

As we observe in the histogram, most of the flights have an elapsed time of about **115 minutes,** and the mean is at **120.05 minutes** and the median at **118 minutes**.

Question 3: Identify relationship between **distance of a flight** and below mentioned variables, if there's any. If there's relation, then please identify the significance of it. *(Marks: 3)*

i. Examine whether average distance of flights differs according to the day of the week.

ii. Examine whether distance of a flight is somehow correlated with the taxi out time. Use appropriate chart to visualize the outcome.

Hints:

For (i), dependent variable 'Distance' is numeric & independent variable 'Day of the week' is categorical variable. You can use boxplot to show the relationship.

For (ii), both variables are numeric. Use correlation to find if there's any relationship and use correlation test to evaluate the significance. Draw a plot to show the correlation.

Answer to 3(i):

```
boxplot(airline$Distance ~ airline$DayOfWeek, data = airline, main =
"Distance variation based on the day of the week.", xlab = "Day of the
Week", ylab = "Distance in Miles", col="grey")

#The boxplot shows that there is no obvious difference in the distance
between the weekdays.
```

Output and Explanation – 3(i):

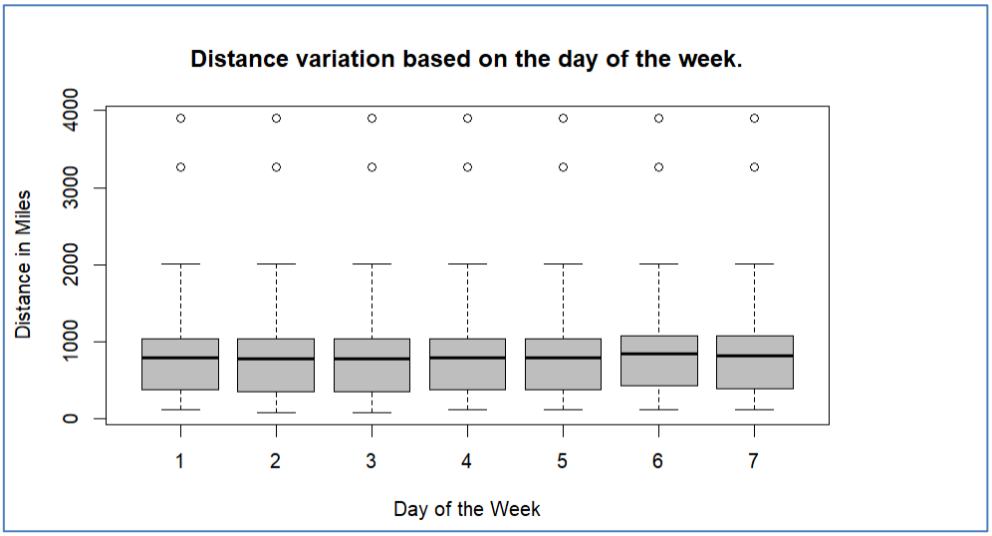


Figure: Distance variation based on the day of the week

The boxplot shows that there is no obvious difference in the distance according to the day of the week.

Answer to 3(ii):

```
#show a correlation between the two variables (Distance and TaxiOut)
cor(x= airline$Distance, y= airline$TaxiOut, use = "na.or.complete",
method = "pearson" )

#The correlation is 0.1582346. Additionally, we want to know whether this
is significant or not for which we are using the cor.test function.

cor.test(x= airline$Distance, y=airline$TaxiOut , alternative =
"two.sided", method = "pearson")

#According to the correlation test, the correlation is significant
(t(224550) = 75.938, p-value < 2.2e-16), indicating that as the flight
duration increases, the taxi time for passengers to reach the carrier
also increases.
```

Output and Explanation – 3(ii):

According to the correlation test, the correlation is significant (t(224550) = 75.938, p-value < 2.2e-16), indicating that as the flight duration increases, the taxi out time for passengers to reach the carrier also increases.

```
        Pearson's product-moment correlation

data:  airline$Distance and airline$TaxiOut
t = 75.938, df = 224547, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1541994 0.1622646
sample estimates:
      cor
0.1582346
```

```
#Draw a plot to show the correlation:
plot(x = airline$Distance, y=airline$TaxiOut, type =
"p", main = "Relationship between Distance and Taxi
out time", xlab = "Distance of the flight (in
Miles)", ylab = "Taxi time (in minutes)")
abline(lm(airline$TaxiOut ~ airline$Distance ), col =
"blue")
```

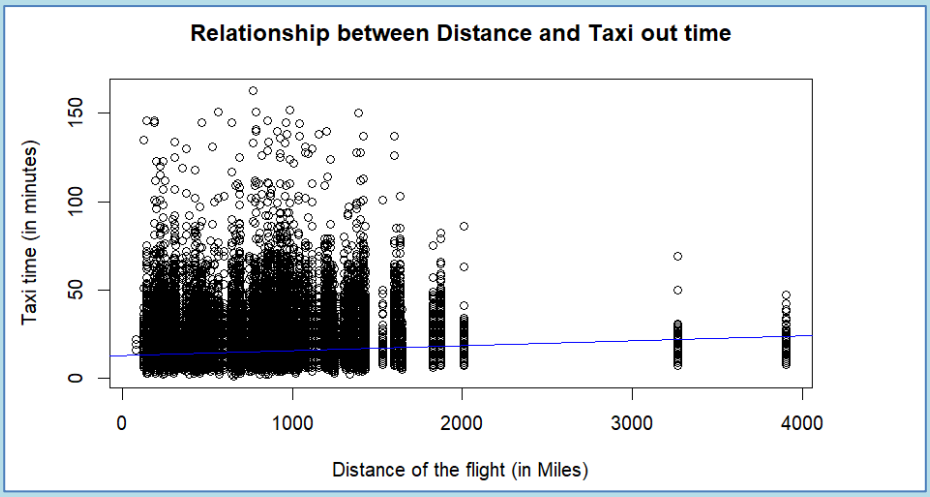


Figure: Relationship between Distance and Taxi Out Time

Question 4: What is the average delay time for flights from Houston to Atlanta in comparison to other destinations? *(Marks: 3)*

Hints:

i. Use aggregate function to assess the mean general delay time for every destination.
ii. Use indexing to get information of Atlanta.
iii. Use summary function to get the maximum and minimum for all the average delay times.

Answer, Output, and Explanation:

```
airline$TotalDelay <- airline$DepDelay + airline$ArrDelay
mean_delay_dest <- aggregate(TotalDelay ~ Dest, data = airline, FUN =
mean, na.rm = TRUE)
summary(mean_delay_dest)
```

The mean general delay time for every destination is: 15.94

```
     Dest              TotalDelay
Length:116         Min.   :-20.09
Class :character   1st Qu.: 12.55
Mode  :character   Median : 15.59
                   Mean   : 15.94
                   3rd Qu.: 19.73
                   Max.   : 51.26
```

```
# Extract the mean total delay for Atlanta (ATL)
mean_delay_atl <- mean_delay_dest[mean_delay_dest$Dest == "ATL",
"TotalDelay"]
# Use indexing to get information of Atlanta
atl_index <- which(mean_delay_dest$Dest == "ATL")
atl_index
mean_delay_atl <- mean_delay_dest$TotalDelay[atl_index]
mean_delay_atl
```

Using indexing to get information of Atlanta, it has been found that index of the mean total delay for Atlanta (ATL) is 7 and value is: 18.37644

```
[1] 7
[1] 18.37644
```

```
summary(airline$TotalDelay)
```

Maximum average delay time is: 1948 and Minimum average delay times is: -71

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
-71.00  -10.00    0.00   16.51   19.00 1948.00    3622
```

Therefore, the average delay time for flights from Houston to Atlanta is 18.38 minutes, compared to an average delay of 15.94 minutes for other flights.

Question 5: What is the worst destination to go to from Houston? *(Marks: 3)*

Hints:

i. Compare the result of Atlanta with other destinations in terms of mean & median value.
ii. Find the index of the Airport with most average delay time for flights from Houston to that airport.

Answer, Output, and Explanation:

```
airline$TotalDelay <- airline$DepDelay + airline$ArrDelay
mean_delay_dest <- aggregate(TotalDelay ~ Dest, data = airline, FUN =
mean, na.rm = TRUE)

#Comparing the result of Atlanta with other destinations in terms of mean
& median value.
summary(mean_delay_dest) #mean of other destinations is 15.94
# Mean of Atlanta (ATL)
mean_delay_atl <- mean_delay_dest[mean_delay_dest$Dest == "ATL",
"TotalDelay"]
mean_delay_atl #mean of destination Atlanta is 18.37644

#In comparison to other destinations with an average total delay of 15.94
minutes, Atlanta has a significantly higher average total delay of
18.37644 minutes.
```

```
     Dest               TotalDelay
 Length:116         Min.   :-20.09
 Class :character   1st Qu.: 12.55
 Mode  :character   Median : 15.59
                    Mean   : 15.94
                    3rd Qu.: 19.73
                    Max.   : 51.26
[1] 18.37644
```

```r
#Here we are using the function which.max to get the index of the
greatest delaytime.
which.max(mean_delay_dest$`DepDelay + ArrDelay`)
# Index value is [1] 5
#It returns that the index value is 5. So, we use indexing to get the
Destination of the fifth row:
mean_delay_dest$Dest[5]
## [1] "ANC"
```

```
integer(0)
[1] "ANC"
```

Therefore, the most average delay time of 51.26 minutes is for flights from Houston to "ANC" where ANC is Anchorage in Alaska.

Question 6: How many flights are flying from Houston Airport (HOU) and how many from George Bush Intercontinal airport (IAH) per month. Create a loop to calculate the number of flights in each of the 12 months. *(Marks: 4)*

Hints:

i. You can use 'for' loop and repeat it for every month.
ii. Use 'print' function to see a monthly table.

Answer, Output, and Explanation:

```r
# Create a vector with month names
month_names <- c("January", "February", "March", "April", "May", "June",
                 "July", "August", "September", "October", "November",
"December")
for (i in 1:12) {

  monthly.table <- table(airline$Origin[airline$Month == i])
# Print the month name
  cat(month_names[i], "\n")
# Print the table
  print(monthly.table)
  cat("\n")
}
```

```
January

  HOU   IAH
 4270 14640

February

  HOU   IAH
 3884 13244

March

  HOU   IAH
 4544 14926

April

  HOU   IAH
 4420 14173

May

  HOU   IAH
 4533 14639

June

  HOU   IAH
 4499 15101

July

  HOU   IAH
 4519 16029

August

  HOU   IAH
 4505 15671

September

  HOU   IAH
 4186 13879

October

  HOU   IAH
 4405 14291

November

  HOU   IAH
 4212 13809

December

  HOU   IAH
 4322 14795
```

Figure: Flights from HOU to IAH per month (Jan-Dec)

Above the number of flights per month departing from Houston Airport (HOU) and George Bush Intercontinental Airport (IAH) is shown.

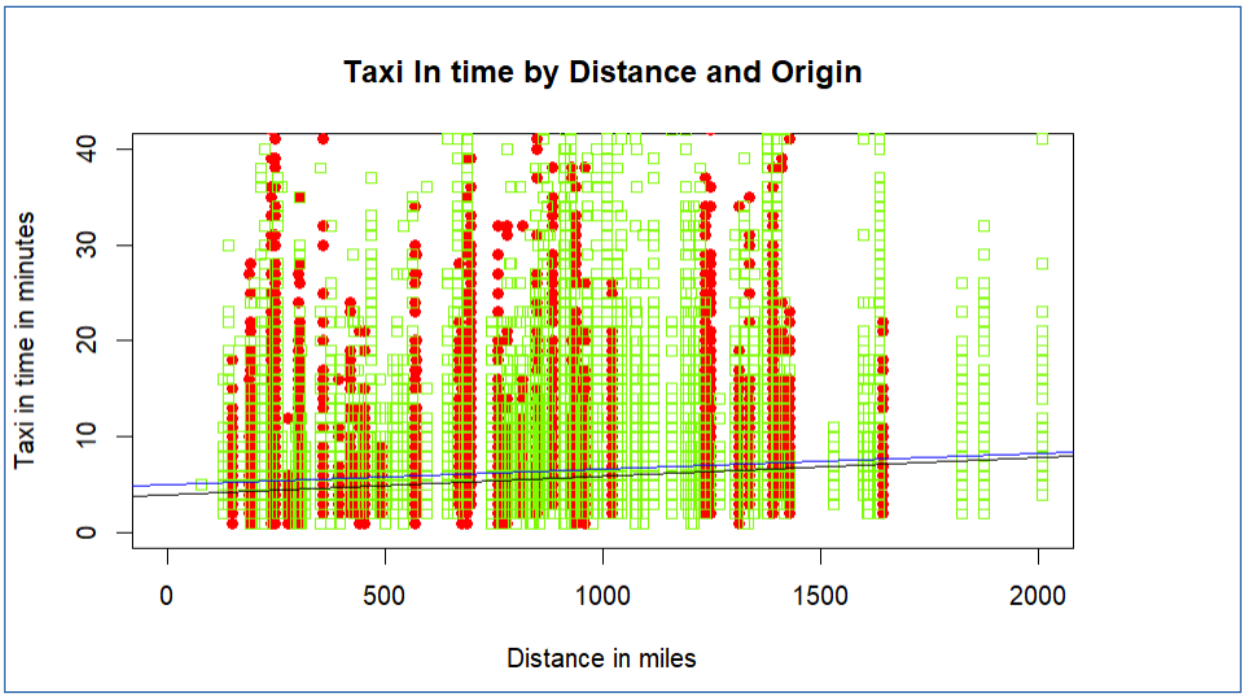
Question 7: Create scatterplot with two groups in different colours: Compare flights starting at HOU (red dots) and flights starting at IAH Airport (green squares) in relation to distance of a flight and Taxi in time. *(Marks: 3)*

Hints:

i. Create two subsets of data – one for HOU & the other for IAH.
ii. Now create a blank plot.
iii. Then add red dots (pch = 16) for HOU and green squares (pch = 22) for IAH. You can use 'points' command to add the points.

Answer, Output, and Explanation:

```
flights_origin_hou <- subset(airline, Origin == "HOU")
flights_origin_iah <- subset(airline, Origin == "IAH")

# Now we are creating a blank plot
plot(x = 1,xlab = "Distance in miles",ylab = "Taxi in time in minutes",
type = "n",
main = "Taxi In time by Distance and Origin",
xlim = c(0, 2000),
ylim = c(0, 40))

# Now adding red dots for flights from Houston Airport:
points(x = (flights_origin_hou$Distance),
y = (flights_origin_hou$TaxiIn),
pch = 16,
col = "red")

# adding lawngreen squares for flights from George Bush International
Airport:
points(x = flights_origin_iah$Distance,
y = flights_origin_iah$TaxiIn,
pch = 22,
col = "lawngreen")

#Finally, we are adding two regression lines, a black one for HOU and a
blue one for IAH:

abline(lm(flights_origin_hou$TaxiIn ~ flights_origin_hou$Distance), col =
"black")
abline(lm(flights_origin_iah$TaxiIn ~ flights_origin_iah$Distance), col =
"blue")
```



Here,

- HOU (Red Dots): Flights from HOU are mostly of short distance (less than 500 miles). These flights have varied taxi in times, spread widely on the y-axis. There are fewer long-distance flights from HOU.
- IAH (Green Squares): Flights from IAH cover a wider range of distances, including many long-distance flights (up to 2000 miles). Taxi in times for these flights are more spread out but tend to cluster in the mid-range.
- **Black Line**: This line shows the relationship between distance and taxi in time for HOU flights. It indicates a slight increase in taxi in time as the distance increases.
- **Blue Line:** This line shows the relationship between distance and taxi in time for IAH flights. It indicates that taxi in time increases with distance, but less so compared to HOU flights.
