Probability Theory and Statistics

ALY6010, WINTER 2022

Week -6

Module-6- Final Project

Submitted by: Abhinav Jain

NUID: 002938209

Submitted To: Prof. Tom Beur

Date: 04/02/2022

Introduction:

USA flight statistics (2009)

We are investigating US air traffic in this dataset, which will give various visuals and potential indicators to better understand transportation operations through hypothesis testing, correlation, and linear regression. The population's origin and destination parameters, flight details, seats available, people traveling, travel lengths, and origin and destination airports are all included in the dataset.

	Origin_airport	Destination_airport	Origin_city	Destination_city	Passengers	Seats	Flights	Distance	Fly_date	Orig
1	LAX	RDM	Los Angeles, CA	Bend, OR	1962	2354	31	726	12/1/2009	
2	STS	RDM	Santa Rosa, CA	Bend, OR	36	76	1	406	12/1/2009	
3	EUG	RDM	Eugene, OR	Bend, OR	64	150	1	103	12/1/2009	
4	EUG	RDM	Eugene, OR	Bend, OR	726	2280	30	103	12/1/2009	
5	OAK	RDM	Oakland, CA	Bend, OR	68	70	1	454	12/1/2009	
6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	

Explanatory Data Analysis: In this dataset, we have passengers with an average of 2536 and a standard deviation of 4808, seats available on flights with an average of 3363 and a standard deviation of 6131, average flights lasting 33, and an average distance of 692 miles. We are primarily interested in these models.

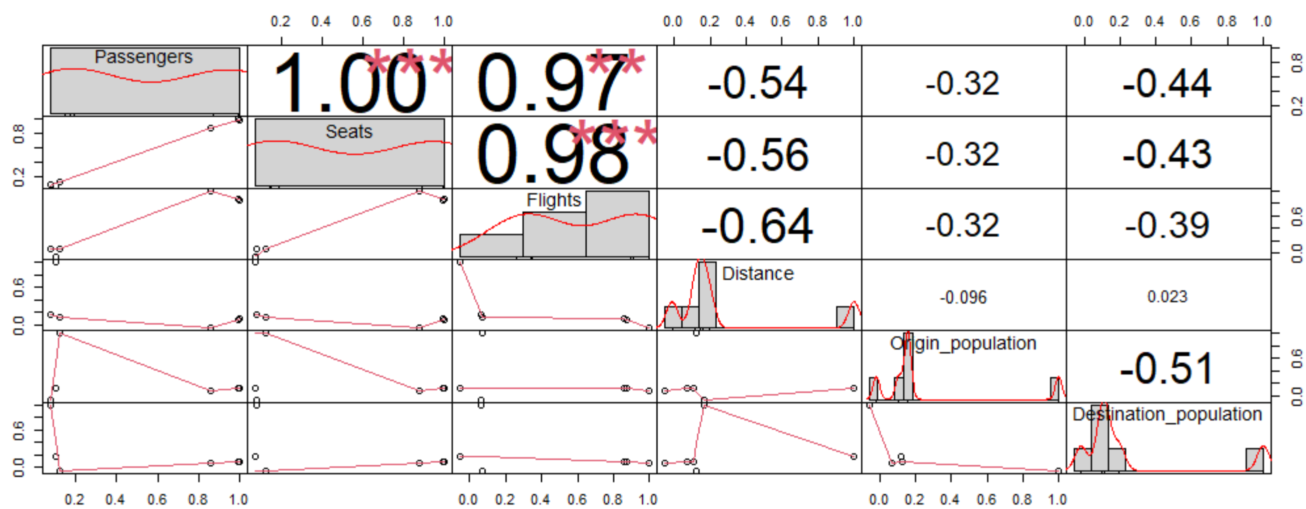
Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Passengers	5211	2536.33	4808.92	0	46	2989.5	67934
Seats	5211	3363.908	6131.183	0	100	4102	89927
Flights	5211	33.894	47.917	0	2	49	491
Distance	5211	692.542	570.302	0	284	920	4962
Fly_date	5211						
... 12/1/2009	5211	100%					
Origin_population	5211	5650104.4	7845570.21	13005	857903	6815696	38139592
Destination_population	5211	9101247.169	7016887.234	13346	1705075	19069796	19161134
Org_airport_lat	5197	37.65	5.903	20.899	33.637	41.709	64.815
Org_airport_long	5197	-90.943	15.628	-157.922	-95.894	-80.291	-68.828
Dest_airport_lat	5198	37.686	5.639	19.721	32.969	41.979	58.355
Dest_airport_long	5198	-87.788	11.528	-155.048	-95.888	-82.533	-68.828

Correlation Between the Variables

The Correlation Matrix clearly reveals the positive association between the variables of passengers, seats, and flights, whereas the rest of the variables in this dataset are negatively associated. The diagonal value of 1 in the following correlation matrix indicates that all variables in the matrix are correlated, although the distance origin population and destination correlation are not. On this premise, we can observe that if one variable grows, the other increases as well, and if one variable decreases, the other decreases as well, demonstrating a positive association.

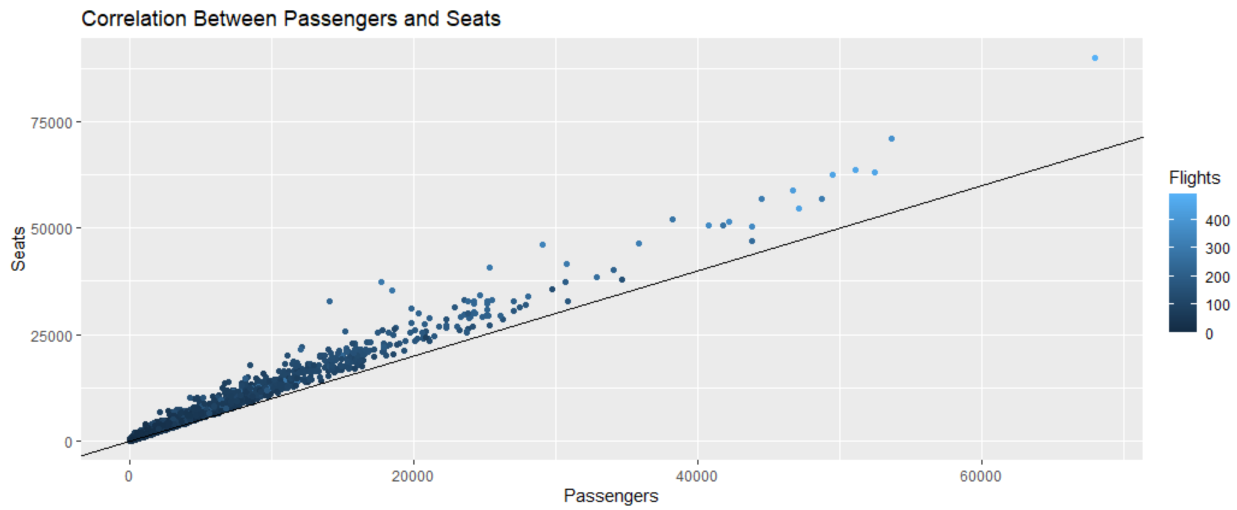
```
> #Correlation
> df$Passengers<- as.numeric(df$Passengers)
> df$Seats<- as.numeric(df$Seats)
> df$Flights<- as.numeric(df$Flights)
> df$Distance<- as.numeric(df$Distance)
> df$Origin_population<- as.numeric(df$Origin_population)
> df$Destination_population<- as.numeric(df$Destination_population)
> df_cor <- df[,c("Passengers","Seats","Flights","Distance","Origin_population","Destination_population")]
> df_corre<- cor(df_cor)
> round(df_corre,2)
                       Passengers Seats Flights Distance Origin_population Destination_population
Passengers                   1.00  0.99    0.86     0.10              0.13                   0.08
Seats                        0.99  1.00    0.88     0.07              0.12                   0.08
Flights                      0.86  0.88    1.00    -0.06              0.07                   0.06
Distance                     0.10  0.07   -0.06     1.00              0.12                   0.16
Origin_population            0.13  0.12    0.07     0.12              1.00                  -0.06
Destination_population       0.08  0.08    0.06     0.16             -0.06                   1.00
```



Correlation Between the Passengers and the Seats

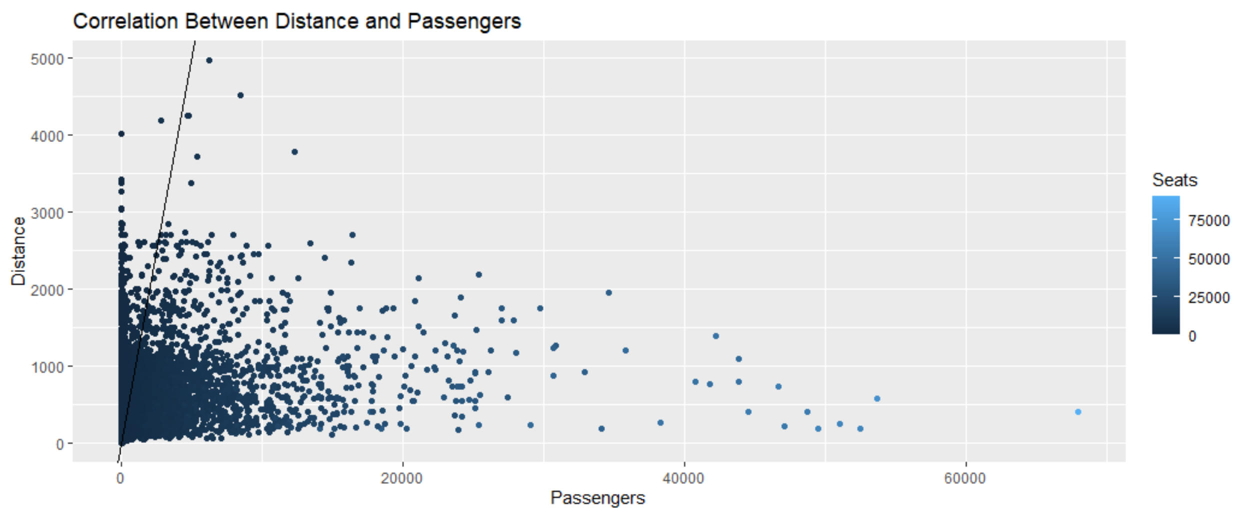
Question1: Is the relationship between the number of passengers booking flight tickets and the availability of the seats?

It's fascinating to see the interaction between the seat and the passenger to the left, which demonstrates that the efficiency of reserving airline tickets is at 1 fill before the scheduled departure time. However, at 20,000 passengers and 25000 seats, practically every aircraft is fully booked.



Question2: What is the connection between the number of people flying and the flight distance?

According to the correlation graph, most passengers travel within 2000 miles of the United States, hence seat availability is lower in contrast to the number of passengers who go long distances via aircraft.



Regression Analysis

The most commonly used statistical tool for comparing two or more variables that are of equivalent importance. Throughout the regression analysis, pay attention to the influence of the dependent variable or one or more independent variables. The coefficient of determination, on the other hand, is a measure for the statistical observation in a regression model that allows us to explain the proportion of variation in the independent variable that can be measured by independent variables. R squared provides information about the regression model's fitness as well as the model's efficiency.

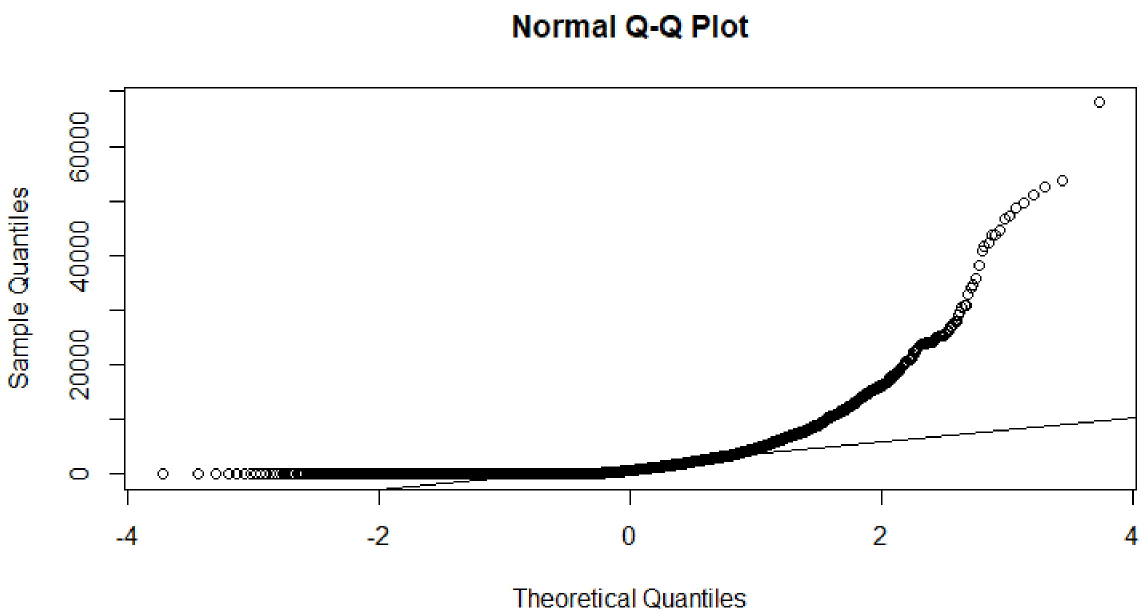
The flight is the dependent variable in the regression model below, although seats, distance, and passengers are the independent variables. After running the model, the data fit the regression model with an r-squared of 79 percent accuracy. In comparison to statistical model analysis, this suggests that the model is better fitted.

```
> fit <- lm(Flights ~ Seats+ Distance+ Passengers, data = df1_r)
> summ(fit)
MODEL INFO:
Observations: 5211
Dependent Variable: Flights
Type: OLS linear regression

MODEL FIT:
F(3,5207) = 6523.06, p = 0.00
R² = 0.79
Adj. R² = 0.79

Standard errors: OLS
-----------------------------------------------
                      Est.    S.E.    t val.      p
----------------- ------- ------ -------- ------
(Intercept)          16.25    0.51     31.60   0.00
Seats                 0.01    0.00     29.49   0.00
Distance             -0.01    0.00    -15.84   0.00
Passengers           -0.01    0.00    -11.13   0.00
-----------------------------------------------
```

Hypothesis testing

Question 3:Find the hypothesis testing to perform the best statistics outcome based on the data? Here, we are collecting dataset from USA flight statistics. IN this case we will do the hypothesis testing wheather is null hypothesis or alternative hypothesis by using **z-test:** as the sample size is more than 30, in this case we decided to use z-test

Normal Q-Q plot: The usual Q-Q plot was used in this work to verify the implementation of hypothesis testing. Whereas y=x shows that the data is connected to the line and there is no departure from the data, it allows us to proceed with the test with the knowledge that the data is normal. We can begin with the hypothesis test.



One Sample z-Test

As we can see the value of p is greater than 0.05, which signifies that we do not have enough sufficient evidence to reject the null hypothesis.

```
> z.test(x=df$Seats, mu= 3363, sigma.x=6131)

        One-sample z-Test

data:  df$Seats
z = 0.010696, p-value = 0.9915
alternative hypothesis: true mean is not equal to 3363
95 percent confidence interval:
 3197.445 3530.372
sample estimates:
mean of x
 3363.908
```

Two sample z-test

The value of p is 1.76, even though the result of the two-sample test z-test is equal to -7.7. We can reject the null hypothesis since we do not have enough evidence, as the p-value is not less than 0.05.

```
> #Two Sample Test
> z.test(x=df$Passengers, sigma.x=4808.92, y=df$Seats, sigma.y=6131)

        Two-sample z-Test

data:  df$Passengers and df$Seats
z = -7.6669, p-value = 1.762e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1039.1388  -616.0173
sample estimates:
mean of x mean of y
 2536.330  3363.908
```

Summary: The interaction between the seat and the customer to the left is intriguing, demonstrating that the efficiency of reserving airline tickets is at 1 fill before the scheduled departure time. However, with a capacity of 20,000 people and 25000 seats, nearly every aircraft is completely booked. According to the correlation graph, most passengers fly within 2000 miles of the United States, hence seat availability is smaller in comparison to the number of passengers who travel vast distances by plane.

The most frequent statistical method for comparing two or more variables of comparable significance. Pay attention to the effect of the dependent variable or one or more independent variables throughout the regression analysis. The coefficient of determination, on the other hand, is a statistic that allows us to explain the amount of variation in the independent variable that can be assessed by independent variables in a regression model. R squared gives you information about the regression model's fitness as well as its efficiency. In the regression model below, flight is the dependent variable, while seats, distance, and people are the independent variables. The data fit the regression model with an r-squared of 79% after running the model.

In this work, the standard Q-Q plot was employed to validate the hypothesis testing implementation. Whereas y=x indicates that the data is linked to the line and that there is no deviation from the data, it allows us to continue with the test knowing that the data is normal. We'll start with the hypothesis test. As we can see, the value of p is more than 0.05, indicating that we lack sufficient evidence to reject the null hypothesis. Even though the outcome of the two-sample test z-test is -7.7, the value of p is 1.76. We may reject the null hypothesis since there is insufficient evidence, and the p-value is more than 0.05.

Reference:

[1]https://www.datavedas.com/inferential-statistics-in-r/

[2]How to Perform One Sample & Two Sample Z-Tests in R

Zach

https://www.statology.org/z-test-in-r/

[3]R-Squared

https://corporatefinanceinstitute.com/resources/knowledge/other/r-squared/

[4] z.test: Z-test

https://www.rdocumentation.org/packages/BSDA/versions/1.2.1/topics/z.test

[5] **https://cran.r-project.org/web/packages/distributions3/vignettes/one-sample-z-test.html**

[6] Introduction to Hypothesis Testing in R - Learn every concept from Scratch!

https://data-flair.training/blogs/hypothesis-testing-in-r/

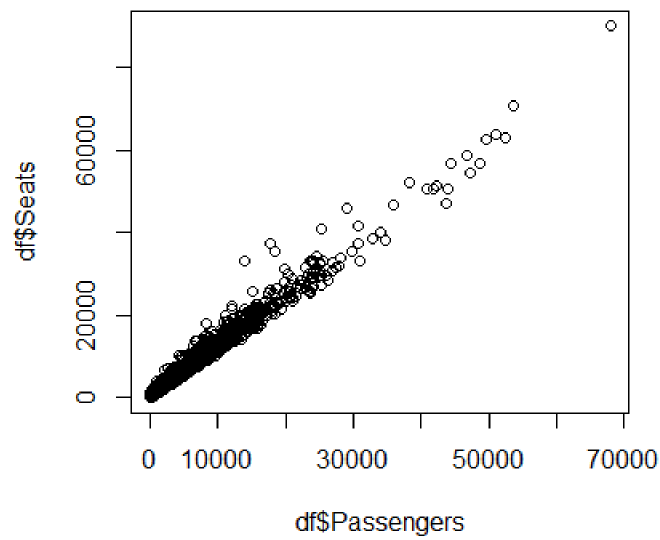
Appendix:

```
install.packages("BSDA")
install.packages("dplyr")

library(ggplot2)
library(dplyr)
library(vtable)
library("Hmisc")
library(fs)
library(corrplot)
library(PerformanceAnalytics)
library(jtools)
library(car)
library(BSDA)

df<- read.csv("Airports2.1.csv")
df


unique(df$Fly_date)
summary(df)
st(df)
df1<- plot(df$Passengers,df$Seats,main = "Scatterplot between Passangers and Seats")

boxplot(df$Passengers,df$Seats, main ="Box Plot of Passangers and Seats")


#make the qqplot

qqnorm(df$Seats)
qqline(df$Seats)
sd(df$Seats)


#One-Sample z-Test Passengers
z.test(x=df$Passengers, mu= 2536.33, sigma.x=4808)

#As we can see the value of p is greater than 0.05,
#which signify that we do not have the enough of sufficient evidence
#to reject the null hypothesis

z.test(x=df$Seats, mu= 3363, sigma.x=6131)

#Two Sample Test
z.test(x=df$Passengers, sigma.x=4808.92, y=df$Origin_population, sigma.y=7845570.21)

#As the value of two sample test z-test is equal to -7.7,
#however the value of p is 1.76. As we can see the p-value is
#not less then 0.05, we can reject the null hypotheisis
#because we do not have enough evidence.
```

```
#Correlation
df$Passengers<- as.numeric(df$Passengers)
df$Seats<- as.numeric(df$Seats)
df$Flights<- as.numeric(df$Flights)
df$Distance<- as.numeric(df$Distance)
df$Origin_population<- as.numeric(df$Origin_population)
df$Destination_population<- as.numeric(df$Destination_population)

df_cor <- df[,c("Passengers","Seats","Flights","Distance","Origin_population","Destination_population")]


df_corre<- cor(df_cor)
round(df_corre,2)
ggplot(df_cor, aes(Passengers,Seats, color = Flights))+
  geom_point()+
  geom_abline()+
  labs(title = "Correlation Between Passengers and Seats")
chart.Correlation(df_corre, histogram= TRUE, pch=19)
#
df_corre<- cor(df_cor)
round(df_corre,2)
ggplot(df_cor, aes(Passengers,Distance, color=Seats))+
  geom_point()+
  geom_abline()+
  labs(title = "Correlation Between Distance and Passengers")


chart.Correlation(df_corre, histogram= TRUE, pch=19)

#Regression Testing

df1_r <- as.data.frame(df_cor)
fit <- lm(Flights ~ Seats+ Distance+ Passengers, data = df1_r)
summ(fit)
```

```
> summary(df)
 Origin_airport     Destination_airport Origin_city        Destination_city     Passengers        Seats           Flights           Distance         Fly_date        
 Length:5211        Length:5211         Length:5211        Length:5211        Min.   :    0   Min.   :    0   Min.   :  0.00   Min.   :   0.0   Length:5211       
 Class :character   Class :character    Class :character   Class :character   1st Qu.:   46   1st Qu.:  100   1st Qu.:  2.00   1st Qu.: 284.0   Class :character  
 Mode  :character   Mode  :character    Mode  :character   Mode  :character   Median :  702   Median : 1000   Median : 16.00   Median : 554.0   Mode  :character  
                                                                              Mean   : 2536   Mean   : 3364   Mean   : 33.89   Mean   : 692.5                     
                                                                              3rd Qu.: 2990   3rd Qu.: 4102   3rd Qu.: 49.00   3rd Qu.: 920.0                     
                                                                              Max.   :67934   Max.   :89927   Max.   :491.00   Max.   :4962.0                     
                                                                                                                                                                  
 Origin_population  Destination_population Org_airport_lat Org_airport_long  Dest_airport_lat Dest_airport_long
 Min.   :   13005   Min.   :   13346       Min.   :20.90   Min.   :-157.92   Min.   :19.72    Min.   :-155.05  
 1st Qu.:  857903   1st Qu.: 1705075       1st Qu.:33.64   1st Qu.: -95.89   1st Qu.:32.97    1st Qu.: -95.89  
 Median : 2082421   Median : 8806874       Median :38.75   Median : -86.75   Median :40.69    Median : -84.43  
 Mean   : 5650104   Mean   : 9101247       Mean   :37.65   Mean   : -90.94   Mean   :37.69    Mean   : -87.79  
 3rd Qu.: 6815696   3rd Qu.:19069796       3rd Qu.:41.71   3rd Qu.: -80.29   3rd Qu.:41.98    3rd Qu.: -82.53  
 Max.   :38139592   Max.   :19161134       Max.   :64.82   Max.   : -68.83   Max.   :58.35    Max.   : -68.83  
                                           NA's   :14      NA's   :14        NA's   :13       NA's   :13       
```

Scatterplot between Passangers and Seats



Box Plot of Passangers and Seats