IE 6200 Engineering Probability and Statistics
Project Report
Group 04

US International Air Traffic Analysis Using R

Created By: -

Samruddhi Kulkarni

Dev Patel

Shubham Chopade

Contents

List of Figures: -

Abbreviations: -

JFK: - John F Kennedy International Airport
LAX: - Los Angeles International Airport
EWR: - Liberty International Airport
ORD: - O'Hare International Airport
MIA: - Miami International Airport
CDG: - Paris Charles de Gaulle Airport
FRA: - Frankfurt Airport
LHR: - Heathrow Airport
NRT: - Narita International Airport
YYZ: - Toronto Pearson International Airport
AA: - American Airlines
AC: - Air Canada
AF: - Air France
BA: - British Airways
CO: - Continental Airlines
DL: - Delta Airlines
LH: - Lufthansa
NW: - Northwest Airlines
UA: - United Airlines
US: - Puerto Rico International Airlines

1 Introduction

The United States is a highly developed country which accounts for approximately a quarter of global GDP, and it's the world's largest economy, which attracts a lot of population from around the world. Also, USA is one of the largest nations by total area, so the fastest mode of transport is by air. The network of airports is the key in the USA. The dataset that we have identified is obtained from Kaggle website and is sited to be taken from the U.S. International Air Passenger and Freight Statistics Report. As part of the T-100 program, USDOT receives traffic reports of US and international airlines operating to and from US airports.

2 Objective

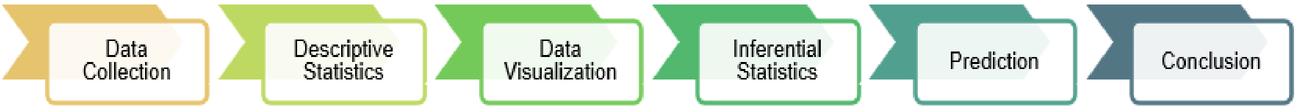
The main objective of the project is to analyse the data and extract useful insights which if used by the authorities will help to get an overall idea of the statistical measures like average number of passenegers, type of distribution of data, probabilities, validation of data through hypothesis testing. The model developed can be used to predict the number of passengers for future.

3 Dataset Description

Sr. No.	Fields	Description
1.	data_dte	Date
2.	Year	Year
3.	Month	Month
4.	usg_apt_id	US Gateway Airport ID - assigned by US DOT to identify an airport
5.	usg_apt	US Gateway Airport Code - usually assigned by IATA, but in absence of IATA designation, may show FAA-assigned code
6.	usg_wac	US Gateway World Area Code - assigned by US DOT to represent a geographic territory
7.	fg_apt_id	Foreign Gateway Airport ID - assigned by US DOT to identify an airport
8.	fg_apt	Foreign Gateway Airport Code - usually assigned by IATA, but in absence of IATA designation, may show FAA-assigned code
9.	fg_wac	Foreign Gateway World Area Code - assigned by US DOT to represent a geographic territory
10.	airline id	Airline ID - assigned by US DOT to identify an air carrier
11.	carrier	IATA-assigned air carrier code. If carrier has no IATA code, ICAO- or FAA-assigned code may be used
12.	carrier group	Carrier Group Code - 1 denotes US domestic air carriers, 0 denotes foreign air carriers
13.	type	The type of the metrics

14.	Scheduled	Metric flown by scheduled service operations
15.	Charter	Metric flown by charter operations
16.	Total	Total Metric flown by scheduled service and charter operations

4 Project Approach

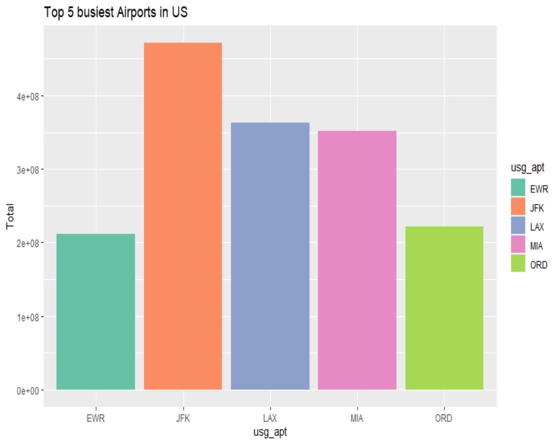


The above image describes the flow of our project.

5 Data Insights

Unique	US Airports	Foreign Airport	Carriers	Airline ID's
Total	946	1464	564	539

In USA there are a total of 946 airports options for the passengers to fly through 564 different air carriers with 539 airline IDs to 1464 foreign destination.



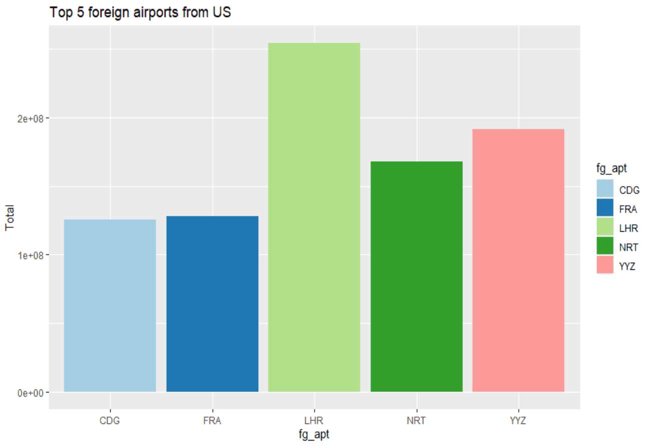


Fig. 1 Fig. 2

From Fig. 1 it can inferred that JFK, LAX, MIA, ORD and EWR are the top USA airport which has got the highest number of passengers over the years.

From Fig. 2 it can be said that the passengers usually fly from USA to LHR, YYZ and NRT.

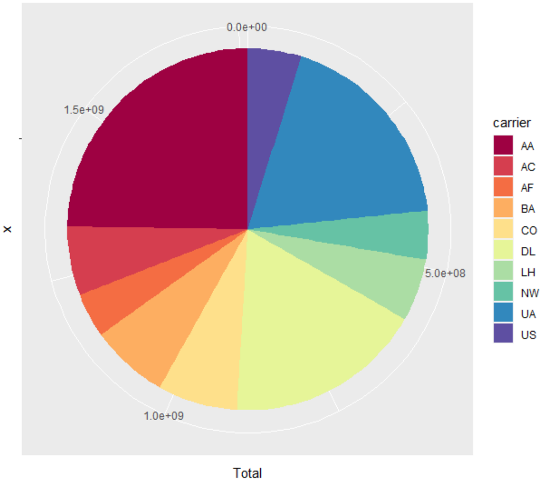


Fig. 3

From Fig. 3 we can see that out of all the airline which fly in USA the top preference of the passengers is AA (American Airlines) followed by UA (United Airlines) and DL (Delta Airlines).

6 Data Visualization

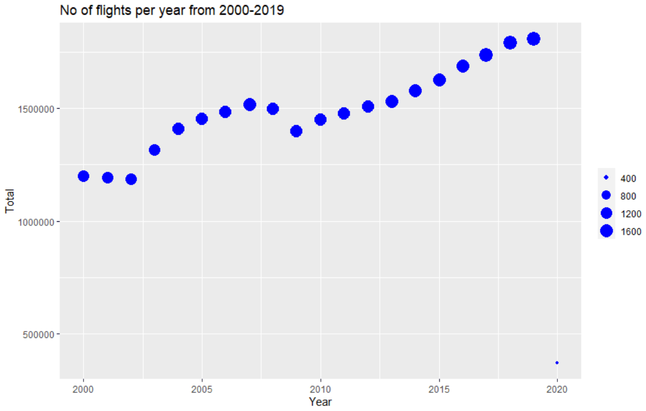


Fig. 4

From Fig.4, it can be observed that with the gradual increase in the number of years there is gradual increase in the number of flights.

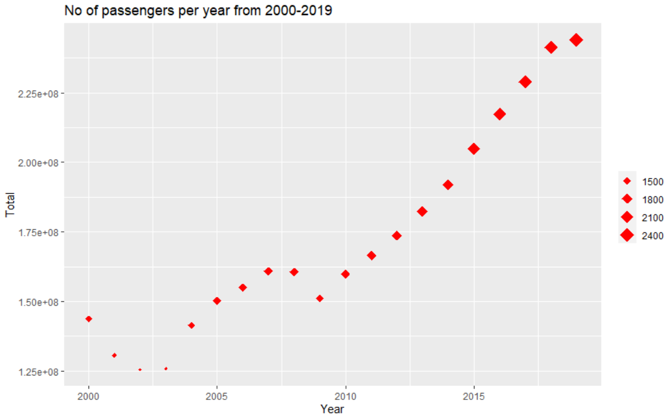


Fig. 5

From the Fig. 5, it can be inferred that with the increase in the year there was an exponential rise in the number of passengers to travel by flights.

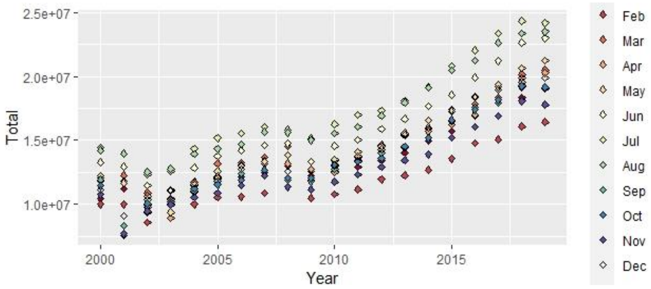


Fig. 6

From Fig. 6, the graph shows number of passengers travelling per month per year in US. It is observed that July and August have peak numbers of passengers travelling which is understandable as there is summer break from June end to August.

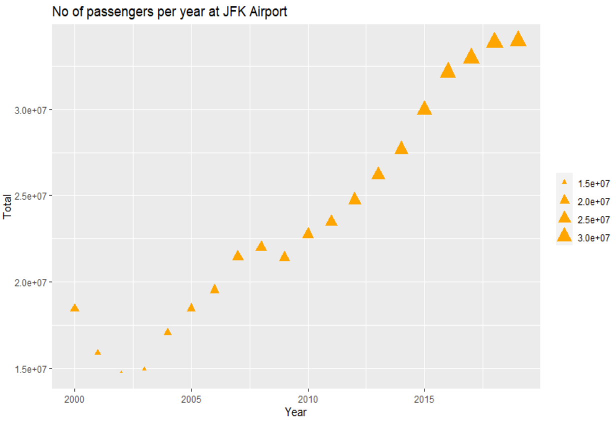


Fig. 7

From Fig. 7, the graph is increasing exponentially. It seems that in year 2001, 2002, 2003 there were less passengers travelling from JFK airport and in year 2018 and 2019 it has recorded the highest number of passengers.

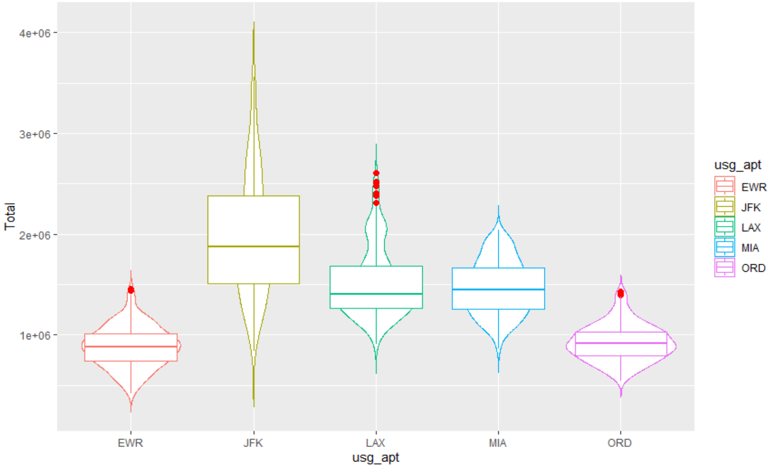


Fig. 8

From Fig. 8, it can be seen that JFK has the highest range as well as highest median as compared to all other airports. JFK passengers range from 0.3e+ 06 to 4.2e+06. Also, we can see there are some outliers in the data. The median of LAX and MIA are nearly same.

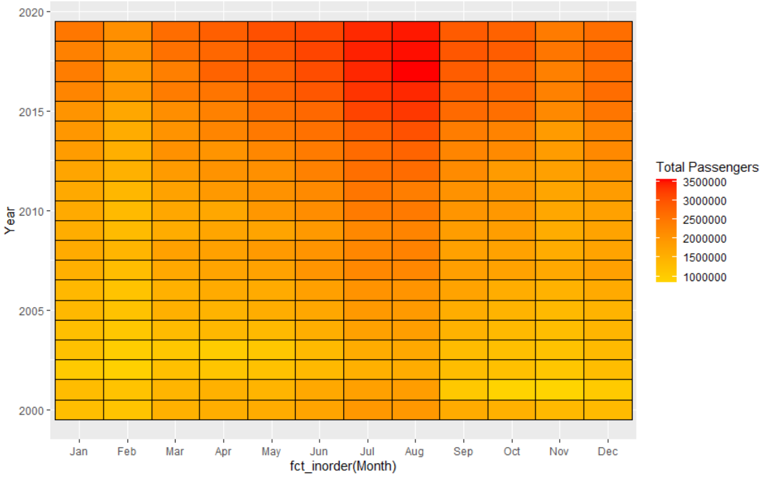


Fig. 9

From Fig. 9, it is noticeable that during the month of July and August over the years USA airport has observed the greatest number of passenger and the least passengers were observed in the month of February and November.

7 Descriptive Statistics

Parameter	Overall data	JFK Airport	LAX Airport	MIA Airport
Mean	172796058	23582251	18143024	17621537
Variance	1.383726e+15	4.070393e+13	1.87669e+13	7.36223e+12
Range (Min -Max)	125464008 - 244063957	14780207 - 33936686	14054403 - 26123058	14436315 - 21297484
Standard Deviation	37198471	6379964	3725143	2713343
Skewness	0.7027084	0.354198	1.225688	0.2376995
Kurtosis	-0.5769966	-1.092538	0.2531447	-1.795048

From the above information, we observe that:

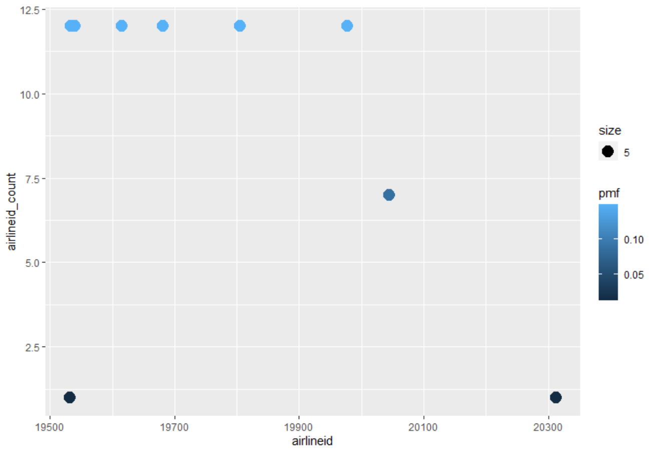
Mean: - We can see that individual means are very less compared to population mean as there are many unique airports in US.

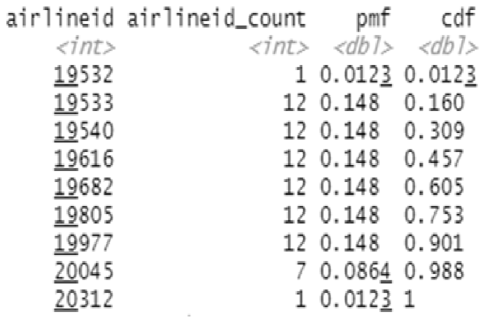
Skewness: - We can say that distribution for combined data of all airports, JFK and MIA airports is moderately skewed (0.5-1) whereas that for LAX is highly skewed (>1).

Kurtosis: - We can see that the distributions are too flat from the kurtosis value (< -1)

Probability Mass Function: -

PMF of unique airlines going from JFK (top in US) to LHR (top in foreign) in year 2000





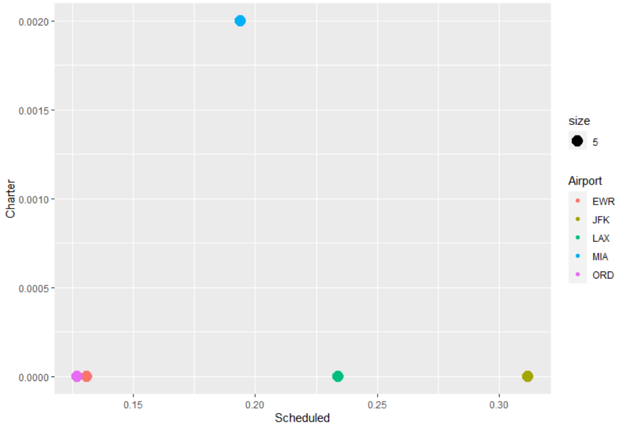
```
airlineid airlineid_count    pmf    cdf
    <int>           <int>  <dbl>  <dbl>
    19532               1 0.0123 0.0123
    19533              12 0.148  0.160
    19540              12 0.148  0.309
    19616              12 0.148  0.457
    19682              12 0.148  0.605
    19805              12 0.148  0.753
    19977              12 0.148  0.901
    20045               7 0.0864 0.988
    20312               1 0.0123 1
```

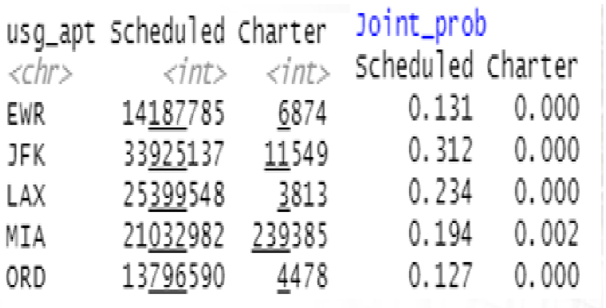
Fig. 10 Fig. 11

Fig. 11 we can find the probability of a given flight flying from JFK to LHR in year 2000. For example, if we take 19616 airline id, we can infer that in year 2000 this airline went from JFK to LHR 12 times and the probability of it is 0.148.

Joint Probability: -

Joint Probability of passengers at top 5 airports taking scheduled flights or chartered flights in the year 2019





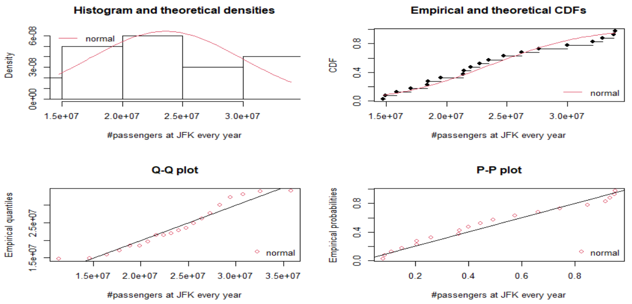
```
                                     Joint_prob
usg_apt Scheduled Charter   Scheduled Charter
<chr>       <int>   <int>
EWR      14187785    6874       0.131   0.000
JFK      33925137   11549       0.312   0.000
LAX      25399548    3813       0.234   0.000
MIA      21032982  239385       0.194   0.002
ORD      13796590    4478       0.127   0.000
```

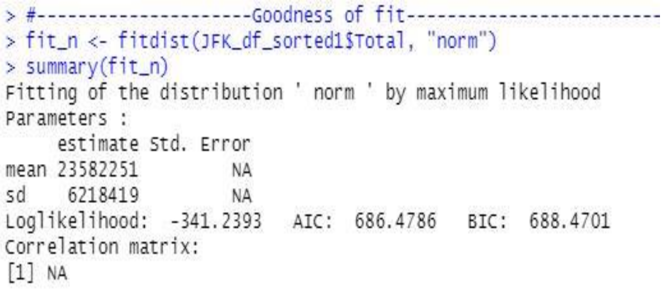
Fig. 12 Fig. 13

From the Fig.13 we can understand what is the probability percentage that a scheduled flight or a charter flight will be boarded by the passenger from top 5 airports in year 2019. For example, if we consider MIA airport, then the probability of the passenger coming to that airport and taking a scheduled flight is 19.4% and a charter flight is 0.2% .

8 Inferential Analysis

For Hypothesis testing, we must see what distribution our data follows, therefore we plot the below graphs to get an idea of the same.
In Fig 14. We have filtered data from Passenger's dataset and considered the top US airport i.e. JFK and tried to understand the distribution of number of passengers per year.

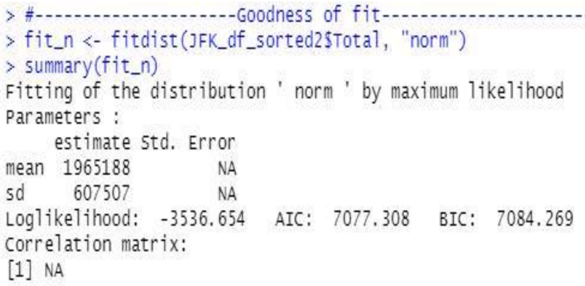




```
> #--------------------Goodness of fit------------------------
> fit_n <- fitdist(JFK_df_sorted1$Total, "norm")
> summary(fit_n)
Fitting of the distribution ' norm ' by maximum likelihood
Parameters :
     estimate Std. Error
mean 23582251         NA
sd    6218419         NA
Loglikelihood:  -341.2393   AIC:  686.4786   BIC:  688.4701
Correlation matrix:
[1] NA
```

Fig. 14

In Fig 15. We have filtered data from Passenger's dataset and considered the top US airport i.e. JFK and tried to understand the distribution of number of passengers per year per month.



```
> #--------------------Goodness of fit------------------------
> fit_n <- fitdist(JFK_df_sorted2$Total, "norm")
> summary(fit_n)
Fitting of the distribution ' norm ' by maximum likelihood
Parameters :
     estimate Std. Error
mean  1965188         NA
sd     607507         NA
Loglikelihood:  -3536.654   AIC:  7077.308   BIC:  7084.269
Correlation matrix:
[1] NA
```

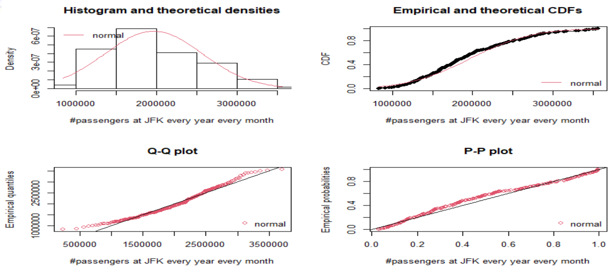


Fig. 15

From the above plots we see that our distribution is normal with some outliers at higher and lower end

8.1 Test of Hypothesis

Mean

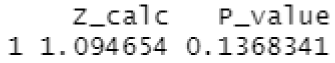
Data considered: JFK data with number of passengers for every month of year is taken.
Assumption: Confidence interval – 95%
Sample method: Stratified samples.
Sample size: 60
Population mean 1965188

Population Variance 370608948985

Sample mean 1999028.

Variance obtained was 393861383240

H0: The mean number of passengers per month of a year < 1000000
H1: The mean number of passengers per month of a year > 1000000
This is a right tailed test

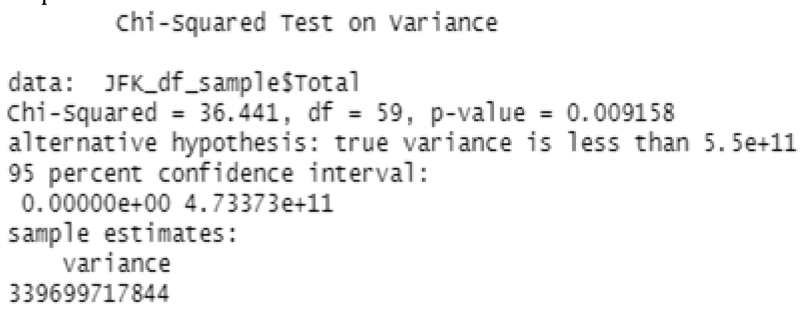


```
    Z_calc   P_value
1 1.094654 0.1368341
```

Conclusion: - As p<<0.05, we reject the null hypothesis and conclude that the sample mean is >1000000

Variance

Data considered: Same as for Mean.
H0: The variance of number of passengers per month of a year > 55x10^10
H1: The variance of number of passengers per month of a year < 55x10^10
This is a left tailed test.
Output:



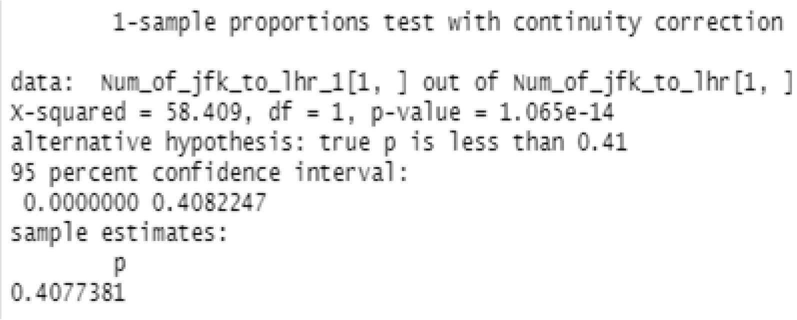
```
        Chi-Squared Test on Variance

data:  JFK_df_sample$Total
Chi-Squared = 36.441, df = 59, p-value = 0.009158
alternative hypothesis: true variance is less than 5.5e+11
95 percent confidence interval:
 0.00000e+00 4.73373e+11
sample estimates:
    variance
339699717844
```

Conclusion: As p<0.05, we reject the null hypothesis and conclude that sample variance is < 55x10^10.

Proportion

Data considered: Dataset filtered for passengers going by BA airlines from JFK to LHR in 2008.
H0: The proportion of number of passengers going by BA airlines from JFK to LHR in 2008 > 0.41.
H1: The proportion of number of passengers going by BA airlines from JFK to LHR in 2008 < 0.41.
This is a left tailed test.
Output:



```
        1-sample proportions test with continuity correction

data:  Num_of_jfk_to_lhr_1[1, ] out of Num_of_jfk_to_lhr[1, ]
X-squared = 58.409, df = 1, p-value = 1.065e-14
alternative hypothesis: true p is less than 0.41
95 percent confidence interval:
 0.0000000 0.4082247
sample estimates:
        p
0.4077381
```

Conclusion: - As p<0.05, we reject the null hypothesis and conclude that sample proportion is < 0.41.

9 Linear Regression

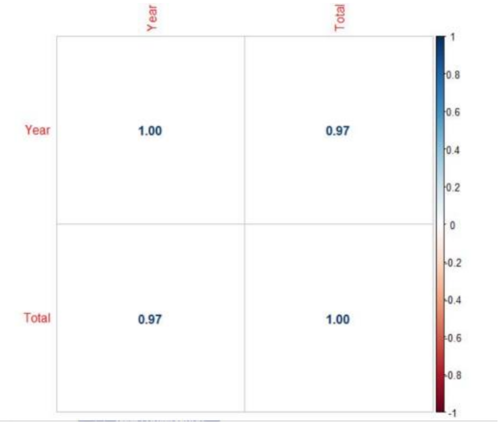
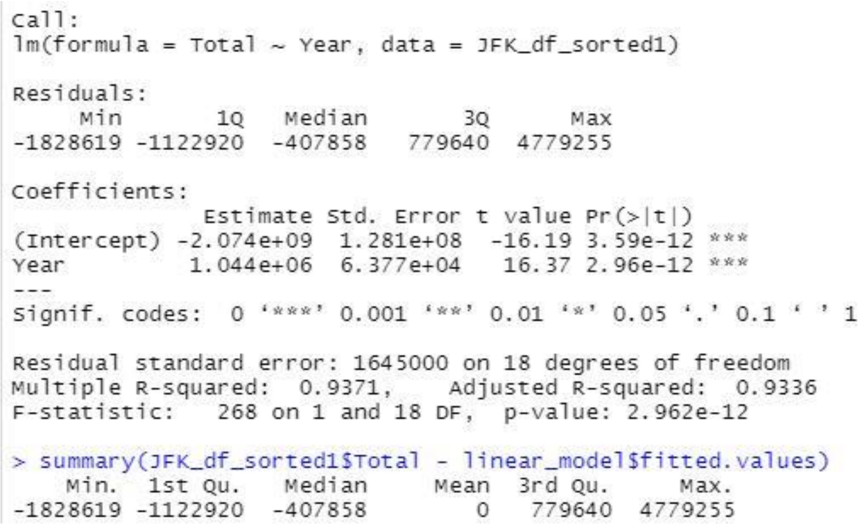


Fig. 16

Fig.16 explains the relation between the strength of the linear relationship between Year and total number of passengers variables.
As the correlation coefficient is high =0.97, the two variables are highly correlated and can be used for regression analysis.

Assumption: Data is from normal distribution. Regression is run on Total number of passengers from JFK airport every year as dependent variable and Years as independent variable.



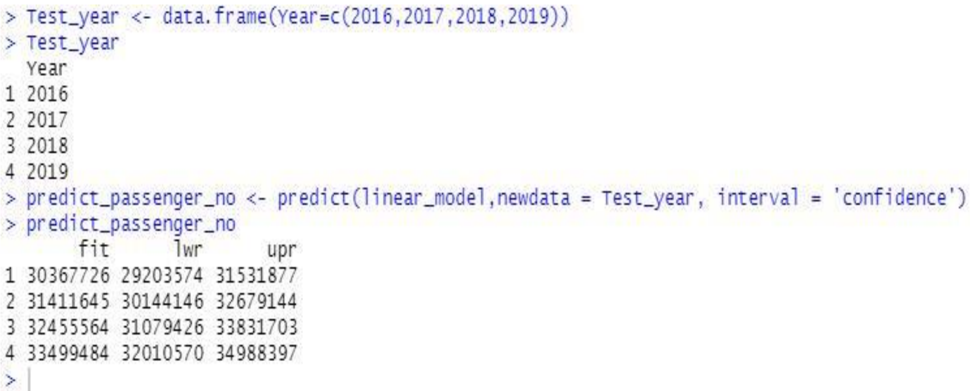
```
Call:
lm(formula = Total ~ Year, data = JFK_df_sorted1)

Residuals:
     Min       1Q   Median       3Q      Max 
-1828619 -1122920  -407858   779640  4779255 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.074e+09  1.281e+08  -16.19 3.59e-12 ***
Year         1.044e+06  6.377e+04   16.37 2.96e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1645000 on 18 degrees of freedom
Multiple R-squared:  0.9371,	Adjusted R-squared:  0.9336 
F-statistic:   268 on 1 and 18 DF,  p-value: 2.962e-12

> summary(JFK_df_sorted1$Total - linear_model$fitted.values)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
-1828619 -1122920  -407858        0   779640  4779255 
```



```
> Test_year <- data.frame(Year=c(2016,2017,2018,2019))
> Test_year
  Year
1 2016
2 2017
3 2018
4 2019
> predict_passenger_no <- predict(linear_model,newdata = Test_year, interval = 'confidence')
> predict_passenger_no
       fit      lwr      upr
1 30367726 29203574 31531877
2 31411645 30144146 32679144
3 32455564 31079426 33831703
4 33499484 32010570 34988397
> |
```

Looking at the output we can say that the distribution is not symmetrical but rightly skewed.
From Coefficients output we can conclude following points:

1. Equation of the model: Total number of passengers for x year = 1.044e+06(x) - 2.074e+09.
2. From model we get predicted value as 31411645 versus actual value as 32936207 for 2017 year which is very close.
3. t-value is 16.37 which means that our Year coefficient is 16.37 standard errors away from 0 which is far, and we can say that the year coefficient is away from value 0 which is true naturally as years cannot be 0.
4. As p-values in our model are extremely small we can say that there is strong evidence that there is strong relationship between Year and Number of passengers.
5. The multiple asterisks indicate that Year is more significant to the model.
6. For our model, we can say that on average, the actual values of number of passengers per year at JFK airport would be 1645000 (1M) away from predicted values. As our max actual value is 33M, having all our predicted values off by 1M proves that model is a good fit for data.
7. Here, Year explains ~93.71% of the variation within Number of passengers, our dependent variable. Thus, we can conclude that our model fits the data very well

10 Conclusion

- From analysing this dataset, we were able to understand the power of R language and how we can increase the efficiency of airports and the most used airlines used by passengers.
- We also implemented our knowledge we gained in lab exercises and tried to analyse different substantial statistics around the most visited JFK Airport.
- We also predicted number of passengers which will visit JFK airport in the future.

11 References

https://www.kaggle.com/parulpandey/us-international-air-traffic-data?select=International_Report_Passengers.csv

https://www.kaggle.com/parulpandey/us-international-air-traffic-`data?select=International_Report_Departures.csv