UNIVERSITY
OF LONDON

ST2195 Programming for Data Science
Coursework Project

Flight Delays in 1997 & 1998

Name: Owen Lee Rui Jie
Student ID: 10219528

Table of Contents

Question 1

In order to analyze flight delays, arrival delay is the more important factor as compared to departure delay, as there are instances where flights could experience departure delays while still arriving at the destination on time.

Best Time of Day to Minimize Flight Delays

The diagrams below show bar plots to display non-delays throughout the time of day in 1997 and 1998.

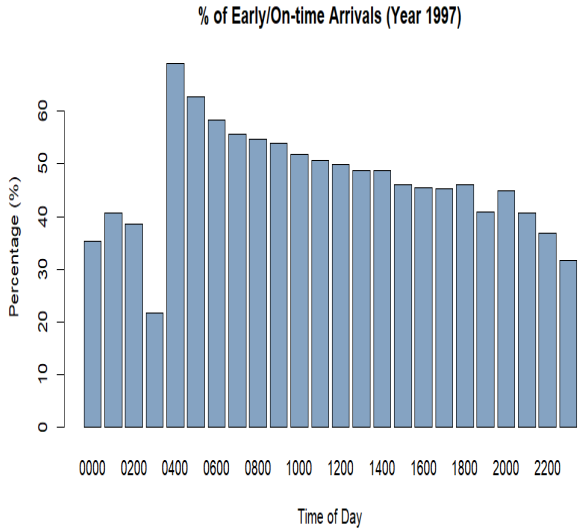

Fig. 1.1.1

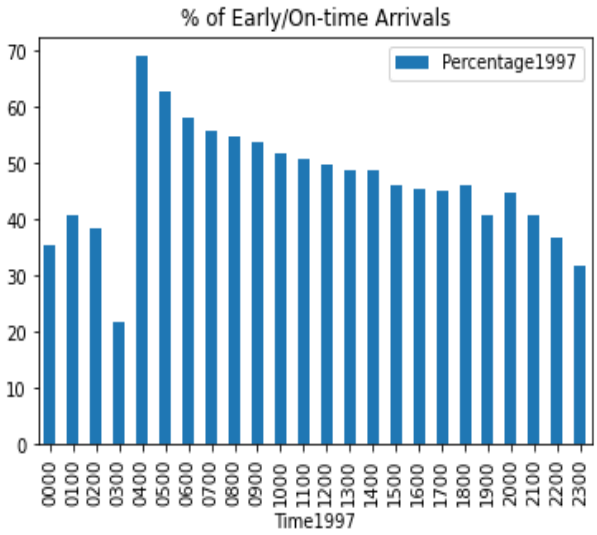

Fig. 1.1.2

Fig. 1.1.1(R) and Fig. 1.1.2(Python) show that the percentage of early/on-time arrivals is the highest between 4am to 5am which is the best time to minimize delays in 1997.

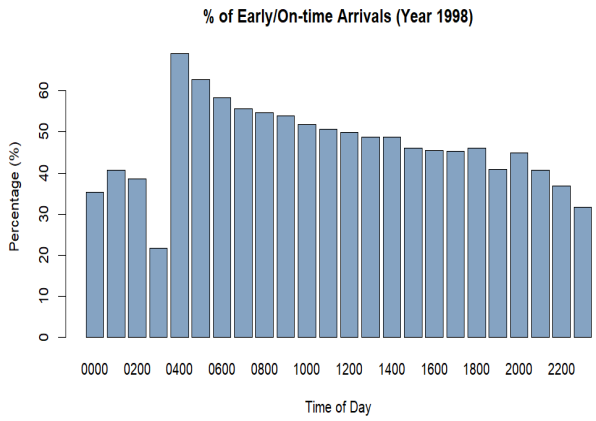

Fig. 1.1.3

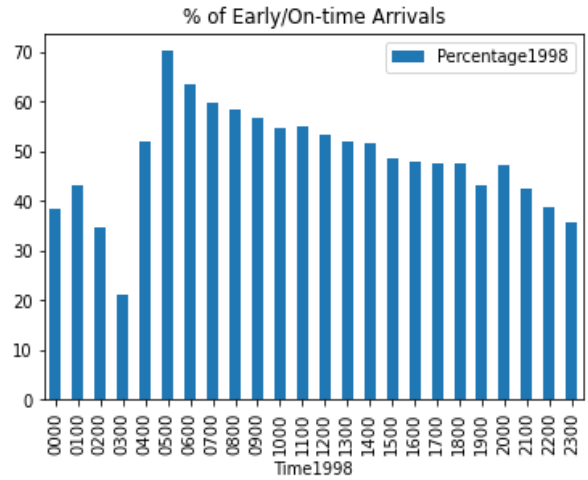

Fig.1.1.4

Fig. 1.1.3(R) and Fig. 1.1.2(Python) show that the percentage of early/on-time arrivals is the highest between 4am to 5am which is the best time to minimize delays in 1998, which is similar to that in 1997.

Best Day of Week to Minimize Delays

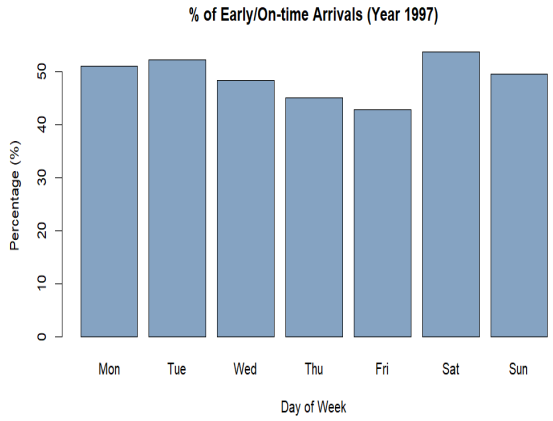

Fig. 1.2.1

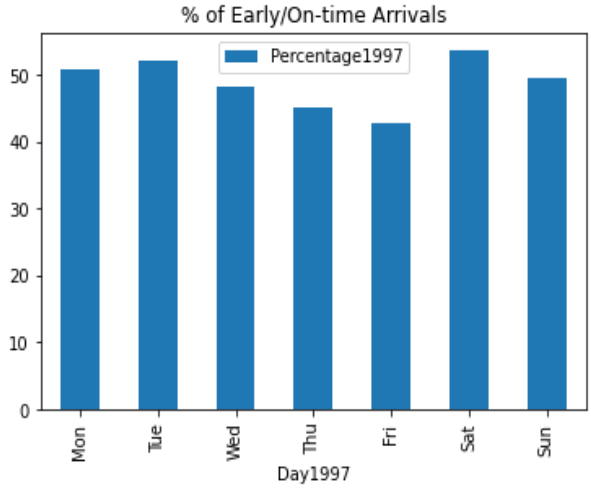

Fig. 1.2.2

Fig. 1.2.1(R) and Fig. 1.2.2(Python) show that the percentage of early/on-time arrivals is the highest on Saturday, which is the best day of week to minimize delays in 1997.

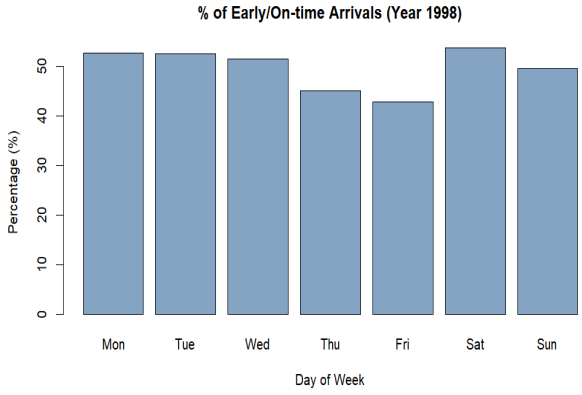

Fig. 1.2.3

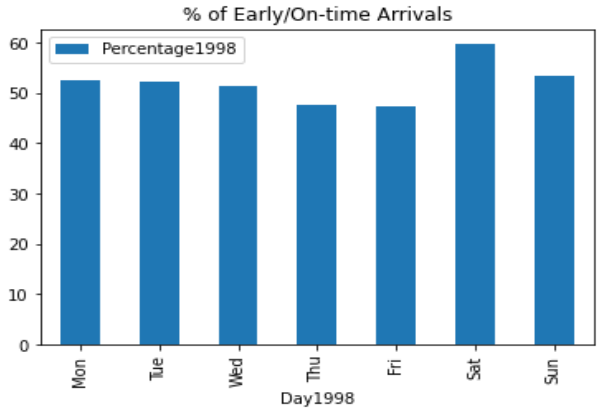

Fig. 1.2.4

Fig. 1.2.3(R) and Fig. 1.2.4(Python) show that the percentage of early/on-time arrivals is the highest on Saturday, which is the best day of week to minimize delays in 1998, similar to that in 1997.

Best Time of Year to Minimize Flight Delays

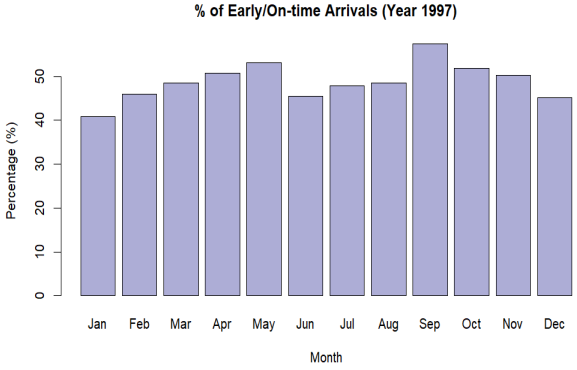

Fig. 1.3.1

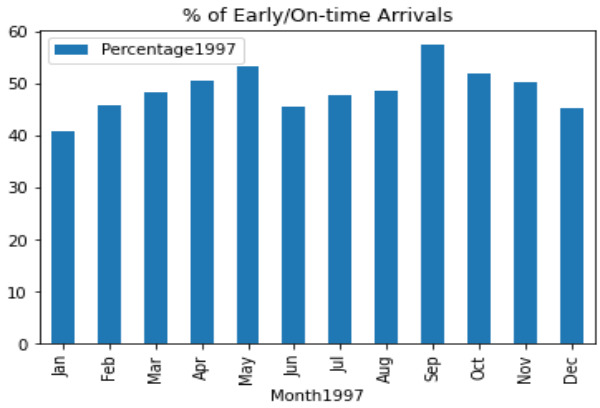

Fig. 1.3.2

Fig. 1.3.1(R) and Fig. 1.3.2(Python) show that the percentage of early/on-time arrivals is the highest in September, which is the best month to minimize delays in 1997.

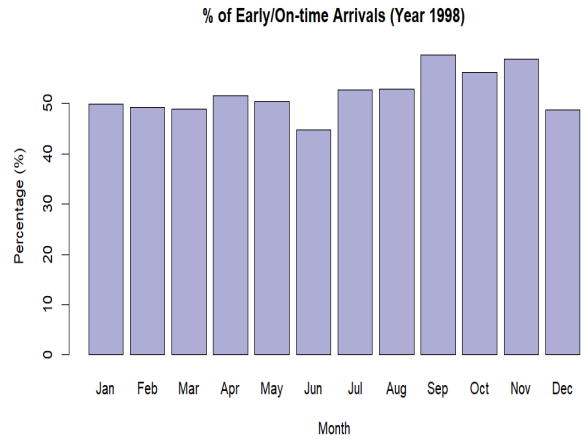

Fig. 1.3.3

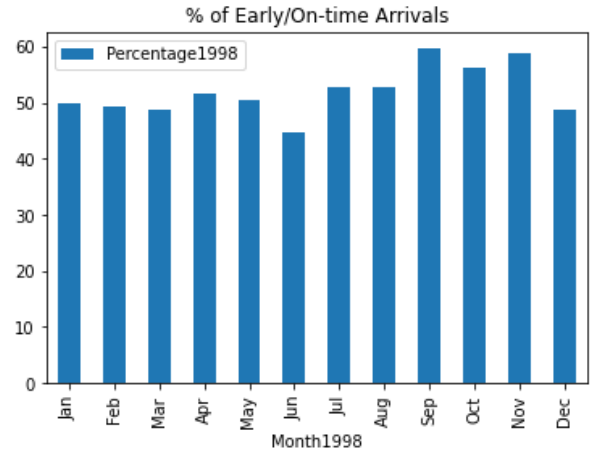

Fig.1.3.4

Fig. 1.3.3(R) and Fig. 1.3.4(Python) show that the percentage of early/on-time arrivals is the highest in September, which is the best month to minimize delays in 1998, which is similar to that in 1997.

Question 2

In order to analyze if older planes suffer from more delays, the plane dataset containing the plane details will be used and merged with the main data in 1997 and 1998. Instead of using arrival delays only, the delays in this scenario will refer to the total delays caused by departure and arrival delays.

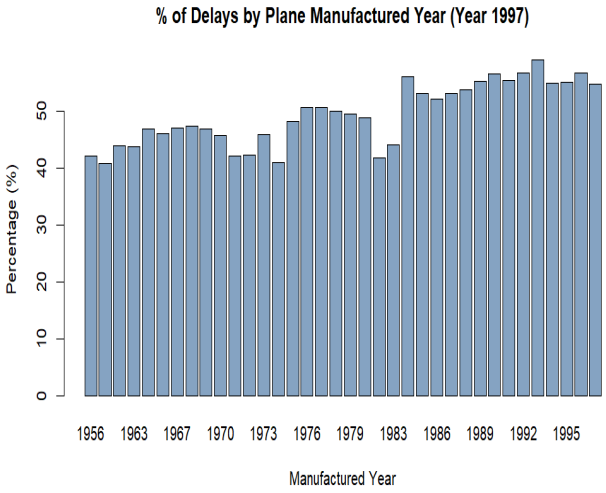

Fig. 2.1

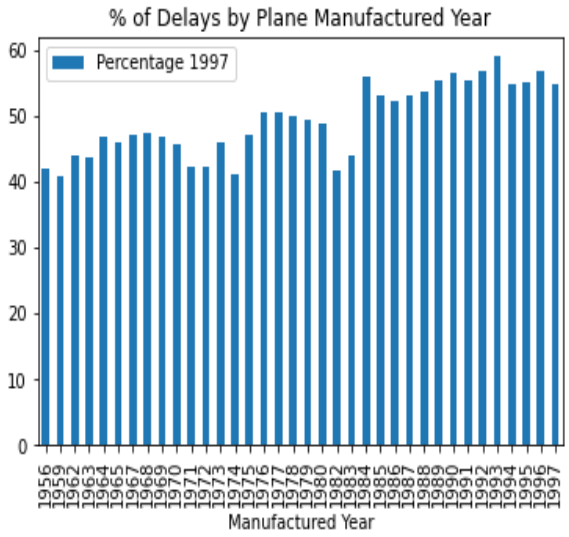

Fig. 2.2

Fig. 2.1(R) and Fig. 2.2(Python) shows the percentage of delays for planes manufactured from 1956 to 1997. In 1997, there were more delays in newer planes compared to older planes with the highest being in planes manufactured in 1993. Therefore, there is no evidence to state that older planes suffer more delays.

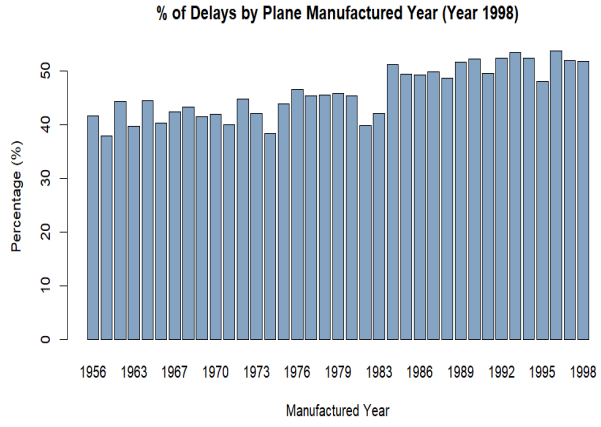

Fig. 2.3

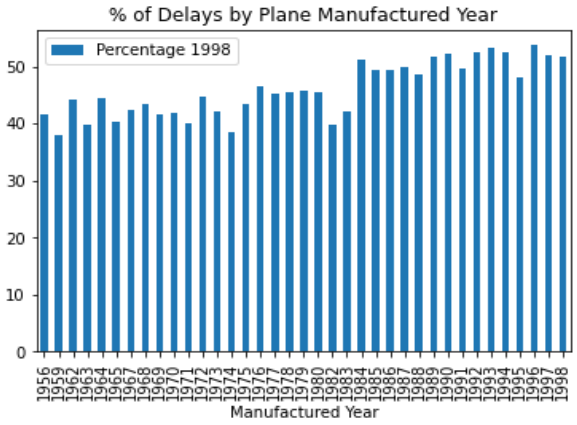

Fig. 2.4

Fig. 2.3(R) and Fig. 2.4(Python) shows the percentage of delays for planes manufactured from 1956 to 1998. In 1998, there were more delays in newer planes compared to older planes with the highest being in planes manufactured in 1996. Therefore, there is also no evidence to state that older planes suffer more delays.

Question 3

In order to analyze the change in number of people flying between locations over time, the data will first be categorized into different seasons - Spring (Mar to May), Summer (Jun to Aug), Fall (Sep to Nov), Winter (Dec to Feb). Secondly, ORD and DFW will be used as an example in 1997 while ORD and ATL will be used as an example in 1998 as these locations have the highest flight frequencies of their year.

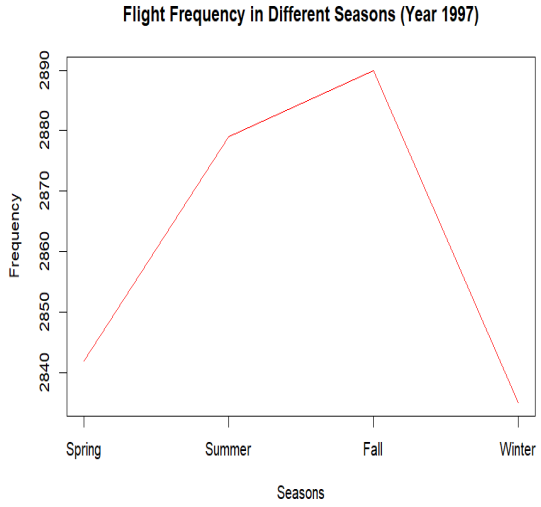

Fig. 3.1

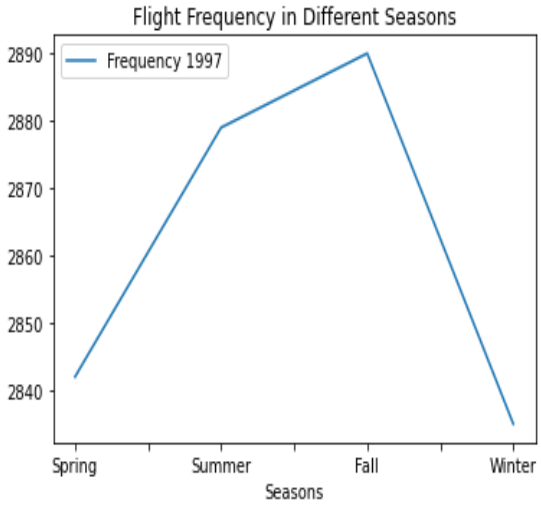

Fig. 3.2

Fig. 3.1(R) and Fig. 3.2(Python) shows the total frequency for flights traveling from ORD to DFW for each season in 1997. The flight frequency is the highest in Fall, followed by Summer, Spring and Winter, which shows that the number of people flying between ORD and DFW is highest in Fall and lowest in Winter.

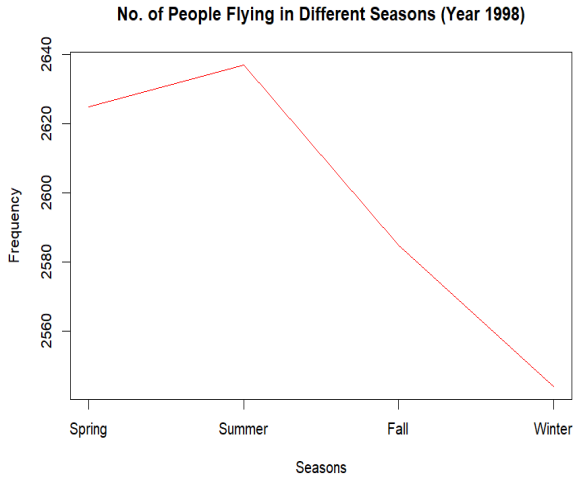

Fig. 3.3

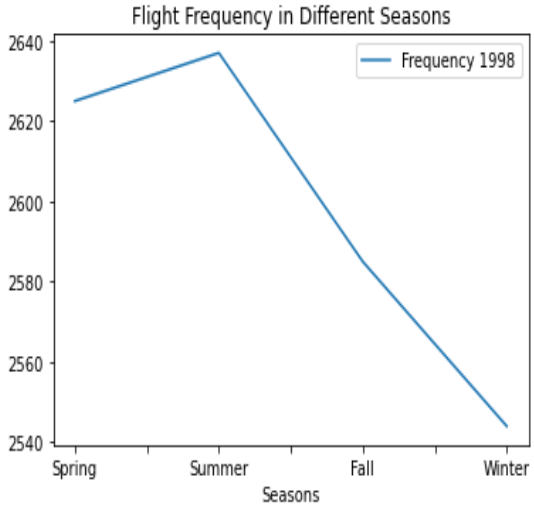

Fig. 3.4

Fig. 3.3(R) and Fig. 3.4(Python) shows the total frequency for flights traveling from ORD to ATL for each season in 1998. The flight frequency is the highest in Summer, followed by Spring, Fall and Winter, which shows that the number of people flying between ORD and ATL is highest in Summer and lowest in Winter.

Question 4

Cascading failures can be detected as illustrated from the examples below.

Month	DayofMonth	DepTime	ArrTime	DepDelay	ArrDelay	CRSDepTime	CRSArrTime	TailNum	Origin	Dest
1	10	1847	2009	72	84	1735	1845	N626	ELP	PHX
1	10	2019	2312	69	67	1910	2205	N626	PHX	SAT
1	10	1446	1532	16	7	1430	1525	N626	DAL	HOU
1	10	1606	1649	71	74	1455	1535	N626	HOU	AUS

Fig. 4.1

Month	ayofMont	DepTime	ArrTime	DepDelay	ArrDelay	RSDepTim	RSArrTim	TailNum	Origin	Dest
1	10	1847	2009	72	84	1735	1845	N626	ELP	PHX
1	10	2019	2312	69	67	1910	2205	N626	PHX	SAT
1	10	1446	1532	16	7	1430	1525	N626	DAL	HOU
1	10	1606	1649	71	74	1455	1535	N626	HOU	AUS

Fig. 4.2

Fig. 4.1(R) and Fig. 4.2(Python) shows a record of flight N626 on 10 January 1997. It is observed that it suffered a departure delay of 72 minutes, which caused an arrival delay of 84 minutes when traveling from ELP to PHX. This further resulted in a cascading failure on its next flight which experienced a departure delay of 69 minutes and arrival delay of 67 minutes when traveling from PHX to SAT.

Month	DayofMonth	DepTime	ArrTime	DepDelay	ArrDelay	CRSDepTime	CRSArrTime	TailNum	Origin	Dest
3	29	1956	2000	56	52	1900	1908	N330AW	CMH	MDW
3	29	2033	2237	53	57	1940	2140	N330AW	MDW	CMH
3	29	1300	1300	95	86	1125	1134	N330AW	CMH	MDW
3	29	1336	1536	81	81	1215	1415	N330AW	MDW	CMH

Fig. 4.3

Month	ayofMont	DepTime	ArrTime	DepDelay	ArrDelay	RSDepTim	RSArrTim	TailNum	Origin	Dest
3	29	1956	2000	56	52	1900	1908	N330AW	CMH	MDW
3	29	2033	2237	53	57	1940	2140	N330AW	MDW	CMH
3	29	1300	1300	95	86	1125	1134	N330AW	CMH	MDW
3	29	1336	1536	81	81	1215	1415	N330AW	MDW	CMH

Fig. 4.4

Fig. 4.3(R) and Fig. 4.4(Python) shows a record of flight N330AW on 29 March 1998. It is observed that it suffered a departure delay of 56 minutes, which caused an arrival delay of 52 minutes when traveling from CMH to MDW. This further resulted in a cascading failure on its return trip to CMH which experienced a departure delay of 53 minutes and arrival delay of 57 minutes.

Similarly in another trip on the same day, it suffered a departure and arrival delay of 95 and 86 minutes respectively. This resulted in a cascading failure on its next return trip which experienced 81 minutes in both departure and arrival delay.

Question 5

Logistic regression will be used to predict if there is a delay by using data from 1997 and 1998. The data will be split into 70 percent training and 30 percent testing data. The independent variables used to predict the likelihood of delays (dependent variable) are Month, DayOfWeek and FlightNum. In this scenario, the dependent variable will be measured by the sum of departure and arrival delays and categorized into delays and non-delays. The model will then be assessed by sensitivity, specificity and accuracy.

Sensitivity refers to the probability of a true positive test result while specificity refers to the probability of a true negative result, whereas accuracy refers to the model's probability of predicting a correct result.

```
Sensitivity : 0.15082
Specificity : 0.86750
Pos Pred Value : 0.50620
Neg Pred Value : 0.53147
Prevalence : 0.47385
Detection Rate : 0.07147
Detection Prevalence : 0.14118
Balanced Accuracy : 0.50916
```

Fig. 5.1.1

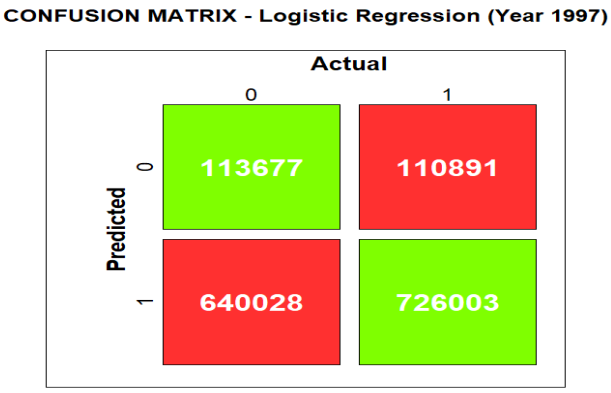

Fig. 5.1.2

Fig 5.1.1(R) shows that the model has a sensitivity and specificity of 0.151 and 0.868 respectively while Fig. 5.1.2(R) shows a confusion matrix of predicted observations with '1' being there is a delay and '0' being there is no delay. The accuracy of the logistic regression model is 50.9% which means the model is average in terms of predicting delays in 1997.

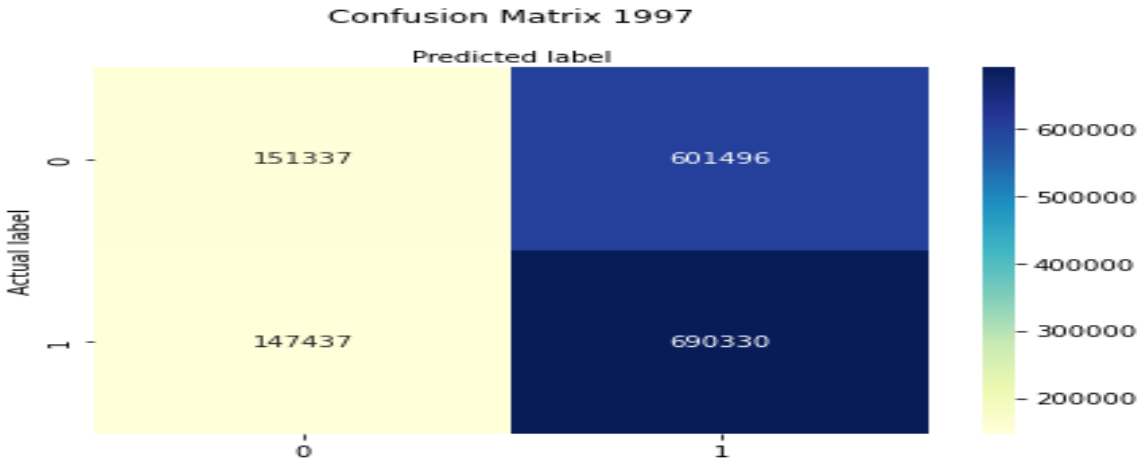

Fig. 5.2

Fig. 5.2(Python) shows a confusion matrix of predicted observations with '1' being there is a delay and '0' being there is no delay. The sensitivity and specificity are 0.201 and 0824 respectively while the accuracy of the logistic regression model is 52.9% which means the model is average in terms of predicting delays in 1997.

```
            Sensitivity : 0.5767
            Specificity : 0.4602
         Pos Pred Value : 0.5211
         Neg Pred Value : 0.5163
             Prevalence : 0.5045
         Detection Rate : 0.2910
   Detection Prevalence : 0.5584
      Balanced Accuracy : 0.5184
```

Fig. 5.3.1

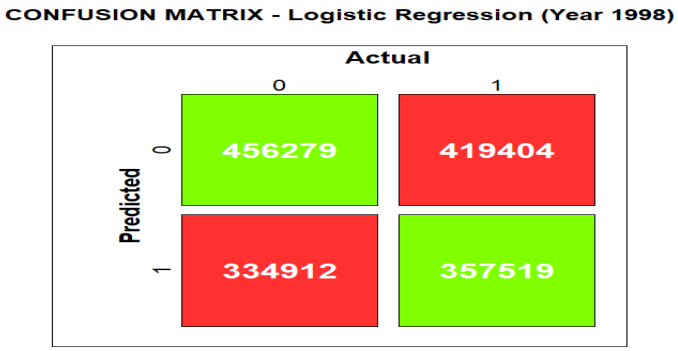

Fig. 5.3.2

Fig 5.3.1(R) shows that the model has a sensitivity and specificity of 0.577 and 0.460 respectively while Fig. 5.3.2(R) shows a confusion matrix of predicted observations with '1' being there is a delay and '0' being there is no delay. The accuracy of the logistic regression model is 51.8% which means the model is average in terms of predicting delays in 1998.

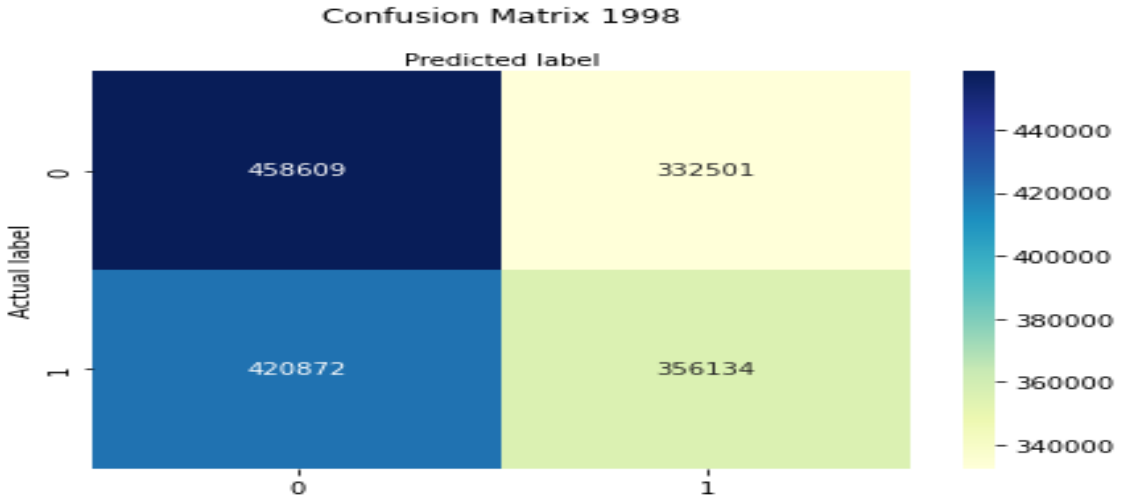

Fig. 5.4

Fig. 5.4(Python) shows a confusion matrix of predicted observations with '1' being there is a delay and '0' being there is no delay. The sensitivity and specificity are 0.580 and 0.458 respectively while the accuracy of the logistic regression model is 52.0% which means the model is average in terms of predicting delays in 1998.

The logistic regression model in 1998 is higher in sensitivity and lower in specificity when predicting delays which means that it is better at predicting no delays but less efficient at predicting delays in 1998 as compared to 1997.

Conclusion

Through this study, it can be observed that flight delays can be statistically analyzed using R and Python. The variables mentioned are significant, and are common factors that help to predict delays. However, given that the current model accuracy is only average, there could be better models available to predict delays other than logistic regression. It could also be due to other factors caused by uncertainties that resulted in a variation in the duration of flight delays.