2015 Flights delays and cancellations

Author : Divya Raghunathan

1 Introduction

The aim of this report is to summarize the exploratory data analysis and data modelling performed on the flights dataset in the python notebook. We will draw insights and make suitable assumptions/conclusions on the results obtained in the notebook. The python file will act as a supporting document for this report. The dataset consists of domestic flight details from 2015.

2 Features of the dataset

Flight delay dataset consists of 31 columns. The total number of observations are 5821. There are several null values in the dataset, this will be dealt with on later stage in our analysis. The dataset contains features such as departure delay, arrival delay, distance, airline, airport and more.

3 Exploratory data analysis (EDA)

We perform EDA to investigate patterns, correlation on our dataset. The analysis is limited to the range and variables in the dataset.

3.1 Departure delay and Arrival delay

We shall begin by analyzing the main delays- Departure and Arrival delay

Departure and Arrival delays have missing values of 91 and 108 respectively. The missing values of both correspond to the same observations. The departure and arrival time is also a missing value for the corresponding missing departure and arrival delays in most cases. We can assume that the arrival delays have more missing values than departure delay because- (1) It could be the first time the flight is flying, hence no prior commercial arrival (2) Or simply that entry was missed. It is hard to know for sure without more information. Hence, we shall drop the missing values from departure and arrival delay from our dataset. The current number of observations in our dataset is 5713.

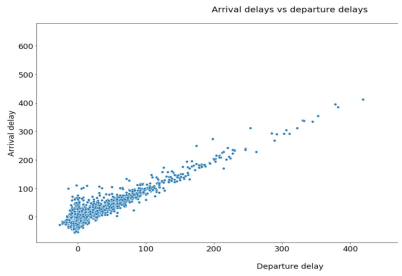

Figure 1: Arrival vs Departure delay

```
Average departure delay : 8.886
Median departure delay : -2.0
Average arrival delay : 3.988
Median arrival delay : -5.0
```

From Figure 1, we can infer that there is a somewhat linear relationship between arrival and departure delay. The flights that depart late will arrive late, in most cases and vice versa.

The average departure delay is more than the average arrival delay. As we do not know the definite reason for this, we can make some assumptions as to why the departure delay is higher than the arrival delay-(1) There are many more factors involved in departure, eg: airline delay, security delay, weather delay, compared to arrival delay which is affected by fewer factors like weather delay. (2) Many times, if the distance is more, the distance makes up for the departure delay. Similarly, the median of departure delay is higher than arrival delay suggesting that departure delay is on the higher side compared to arrival delay.

3.2 Departure and arrival delays based on airlines

There are a total of 14 different airlines. The number of flights for each airline is given in table 1. Table 1 lists the top 10 arrival and departure delay grouped by airlines sorted in descending order of Median. Southwest Airlines Co.(WN) is the most popular with 1269 observations and Hawaiian Airlines Inc.(HA) is the least popular with 57 observations.

Figure 2 & Table 1 details the departure and arrival delay for each airline. We can make several inferences from them - (1) The median for both departure and arrival delay for most of the airlines is negative, which means most of the flights reach or depart early , which is a good thing.

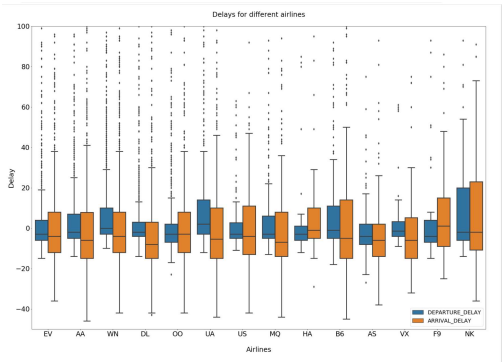

Figure 2: Delays for different airlines

DEPARTURE DELAY SUMMARY BASED ON AIRLINES:

AIRLINE	mean	count	Min	Q1	Median	Q3	Max
UA	13.851779	506.0	-12.0	-3.0	1.5	14.00	332.0
WN	9.894405	1269.0	-10.0	-3.0	0.0	10.00	224.0
B6	13.645914	257.0	-18.0	-5.0	-1.0	11.00	330.0
VX	8.593750	64.0	-9.0	-4.0	-1.5	3.25	230.0
AA	8.349296	710.0	-14.0	-5.0	-2.0	7.00	289.0
DL	7.238562	918.0	-14.0	-4.0	-2.0	3.00	419.0
NK	15.228814	118.0	-14.0	-6.0	-2.0	20.00	353.0
EV	7.461538	546.0	-15.0	-6.0	-3.0	4.00	382.0
HA	7.964912	57.0	-12.0	-6.0	-3.0	1.00	202.0
MQ	7.278810	269.0	-13.0	-5.0	-3.0	6.00	311.0
OO	5.702609	575.0	-23.0	-7.0	-3.0	2.00	306.0
US	7.393204	206.0	-11.0	-5.0	-3.0	2.75	345.0
AS	2.800000	145.0	-27.0	-8.0	-4.0	2.00	186.0
F9	14.835616	73.0	-15.0	-7.0	-4.0	4.00	650.0

ARRIVAL DELAY SUMMARY BASED ON AIRLINE :

AIRLINE	mean	count	Min	Q1	Median	Q3	Max
F9	19.150685	73.0	-25.0	-9.00	1.0	15.00	644.0
HA	10.894737	57.0	-29.0	-5.00	-1.0	10.00	194.0
NK	14.881356	118.0	-36.0	-10.75	-2.0	23.00	354.0
OO	4.269565	575.0	-42.0	-12.00	-3.0	8.00	304.0
EV	5.686813	546.0	-36.0	-12.00	-4.0	8.00	386.0
US	5.067961	206.0	-42.0	-13.00	-4.0	11.00	334.0
WN	3.422380	1269.0	-53.0	-12.00	-4.0	8.00	273.0
B6	9.653696	257.0	-45.0	-15.00	-5.0	14.00	339.0
UA	4.693676	506.0	-53.0	-15.00	-5.5	10.00	337.0
AA	2.807042	710.0	-46.0	-15.00	-6.0	7.75	268.0
AS	-0.282759	145.0	-38.0	-14.00	-6.0	2.00	183.0
VX	4.765625	64.0	-32.0	-15.00	-6.0	5.25	233.0
MQ	3.070632	269.0	-44.0	-14.00	-7.0	8.00	292.0
DL	0.135076	918.0	-55.0	-15.00	-8.0	3.00	412.0

Table 1: Departure delay & Arrival delay summary based on airlines

saw

(2) There are many outliers for both the delays, which mean that the high outlier delays are not that common and it could happen due to some unexpected reason (we don't know for sure). (3) Frontier Airlines Inc.(F9) has high average departure and arrival delay and didn't have many flights (count is 73) in 2015.The maximum departure delay was also by F9, which is 650 minutes (an outlier). This value could have skewed the graph increasing the mean.

3.3 Departure and arrival delays based on airports

Figure 3 and table 2 displays the top 10 airports with the highest average delays. One factor that these airports have in common is that the number of flights departing from the airport is very few. Since the number of flights is very less, one late flight will increase the average delay. Hector International airport(FAR) seems like an unpopular airport with the highest average departure delay of 161 minutes but only one single flight.

DEPARTURE DELAY SUMMARY GROUPEDBY AIRPORT :

ORIGIN_AIRPORT	mean	count	Min	Q1	Median	Q3	Max
FAR	161.000000	1.0	161.0	161.00	161.0	161.00	161.0
12898	119.000000	1.0	119.0	119.00	119.0	119.00	119.0
BMI	101.333333	3.0	-5.0	-3.50	-2.0	154.50	311.0
ERI	92.000000	1.0	92.0	92.00	92.0	92.00	92.0
MYR	88.000000	4.0	-6.0	-3.75	2.5	94.25	353.0
14576	88.000000	1.0	88.0	88.00	88.0	88.00	88.0
14696	88.000000	1.0	88.0	88.00	88.0	88.00	88.0
10157	87.500000	2.0	-3.0	42.25	87.5	132.75	178.0
12992	80.000000	1.0	80.0	80.00	80.0	80.00	80.0
12206	67.500000	2.0	51.0	59.25	67.5	75.75	84.0

Table 2: Departure delay summary grouped airport

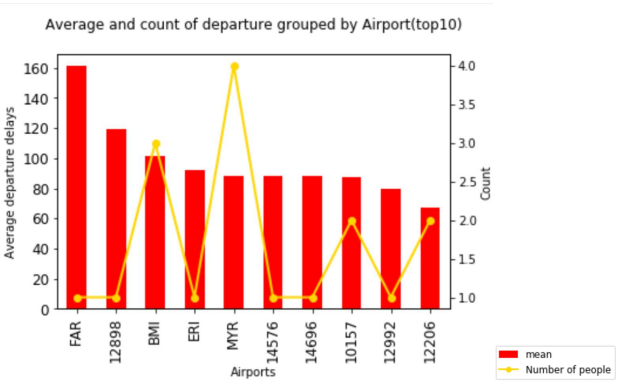

Figure 3: Average and count of departure grouped by airport

3.4 Departure and arrival delays based on distance

Usually, a long distance flight can make up for delays and arrive on time or early. Our analysis on this dataset, doesn't prove this claim.

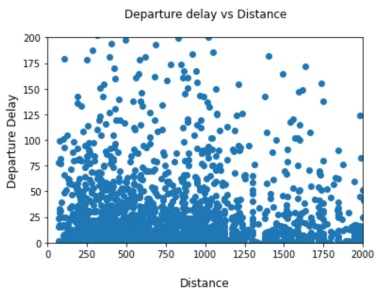

Figure 4: Departure delay vs distance

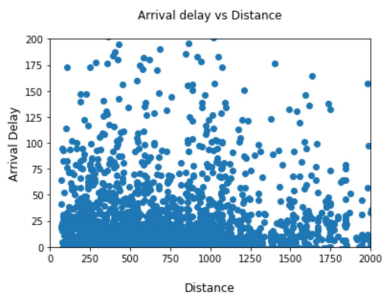

Figure 5: Arrival delay vs distance

	DEPARTURE_DELAY	ARRIVAL_DELAY	DISTANCE
DEPARTURE_DELAY	1.000000	0.936069	0.023095
ARRIVAL_DELAY	0.936069	1.000000	-0.027935
DISTANCE	0.023095	-0.027935	1.000000

Table 3: Correlation matrix - subset

As seen from figures 4 & 5, there is a significant amount of variability associated with arrival delay and distance as well as departure delay and distance. Distance doesn't seem like a good predictor of delay. There seems to very less correlation between departure delay and distance as well as arrival delay and distance as seen in table 3.

	DEPARTURE_DELAY	ARRIVAL_DELAY	DISTANCE
3495	1.0	-18.0	4817
3511	23.0	10.0	3801
2820	19.0	48.0	3417
4522	5.0	-13.0	3365
5762	8.0	-19.0	3329
3969	25.0	26.0	3043
5814	4.0	-13.0	2979
5530	13.0	-6.0	2845
728	1.0	-38.0	2717
2513	4.0	-23.0	2704

Table 4 displays the effect long distance has on arrival delay when departure delay is present i.e. positive values. The table indicates that even though the distance is quite long, the distance doesn't make up for the departure delay. This can be seen by the many arrival delays (positive values). There is so much variability to make inferences.

Table 4: Top 10 long distance flights having departure delays (positive)

3.5 Departure and arrival delays based on day of week

DEPARTURE DELAY SUMMARY GROUPED BY DAY OF WEEK :

DAY_OF_WEEK	mean	count	Min	Q1	Median	Q3	Max
1	9.786826	835.0	-17.0	-5.0	-2.0	8.0	382.0
2	8.995006	801.0	-18.0	-5.0	-2.0	6.0	330.0
3	7.488971	816.0	-16.0	-5.0	-2.0	6.0	345.0
4	9.390443	858.0	-18.0	-4.0	-1.0	8.0	419.0
5	9.661148	906.0	-16.0	-4.0	-1.0	8.0	311.0
6	7.125894	699.0	-27.0	-5.0	-2.0	5.0	353.0
7	9.385965	798.0	-23.0	-5.0	-1.0	9.0	650.0

Table 5 : Departure delay summary grouped by day of week

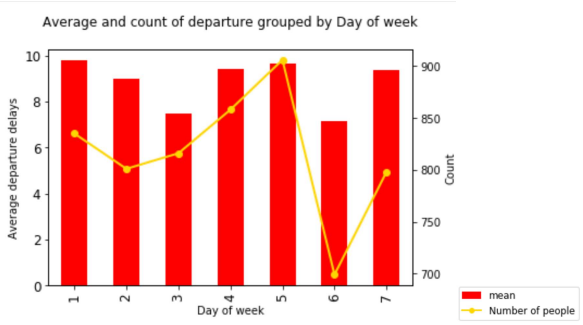

Figure 6: Average and count of departure delay grouped by day of week

According to the table 5 and figure 6 , the best day of the week to travel to experience least departure delay is Saturday-6 with a mean of 7.12 minutes. The highest departure delay is on Monday-1 with a mean of 9.7 minutes.

3.6 Analyzing the different delay types

There are five delays mentioned in the dataset that contribute to the final two delays-departure and arrival. Let's explore these delays some more. The delay types are given in table 6.

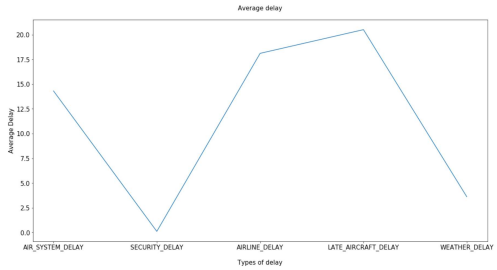

Figure 7: Average delays (graphical)

	mean
AIR_SYSTEM_DELAY	14.319963
SECURITY_DELAY	0.113806
AIRLINE_DELAY	18.119403
LATE_AIRCRAFT_DELAY	20.514925
WEATHER_DELAY	3.616604

Table 6: Average delays (Numerical)

The highest delay is due to delay caused by the aircraft with an average of 20.51 minutes. Whereas, the least delay is caused due to security with an average of 0.11 minutes.

3.7 Month vs Arrival delay

We analyze the arrival delay grouped month wise from table 7 and figure 8. It is interesting to note that, September (9), October (10) and November (11) have a negative average arrival delay. This means that on those months on an average the flights have reached earlier than scheduled. The high average delay is on June (6) with an average delay of 8.57 minutes. We can make assumptions for why this is the case –

(1)Most number of flights (537) in June, so more people, so more delays. (2) A similar reason as to why there is negative delays in months -9,10,11, not many people are travelling, so fewer aircrafts, easier to handle and reduce delays, but surprisingly October has the second highest number of flights, but max delay is the least pulling the mean to the lower side. We would need further information and date to make stronger assumptions.

ARRIVAL DELAY SUMMARY GROUPED BY MONTH :

MONTH	mean	count	Min	Q1	Median	Q3	Max
1	2.663438	413.0	-40.0	-13.0	-5.0	7.0	395.0
2	5.921717	396.0	-42.0	-13.0	-4.0	9.0	292.0
3	7.189555	517.0	-53.0	-13.0	-4.0	11.0	412.0
4	2.601578	507.0	-45.0	-13.0	-6.0	7.0	334.0
5	5.843612	454.0	-36.0	-13.0	-4.0	10.0	226.0
6	8.571695	537.0	-55.0	-12.0	-3.0	12.0	312.0
7	7.605159	504.0	-36.0	-12.0	-3.0	13.0	644.0
8	6.121951	492.0	-44.0	-12.0	-4.0	8.0	354.0
9	-2.253165	474.0	-46.0	-16.0	-8.0	1.0	268.0
10	-1.015595	513.0	-51.0	-15.0	-7.0	3.0	212.0
11	-0.015385	455.0	-42.0	-14.0	-5.0	6.5	337.0
12	3.988914	451.0	-44.0	-15.0	-7.0	10.0	273.0

Table 7:Arrival delay summary grouped by month

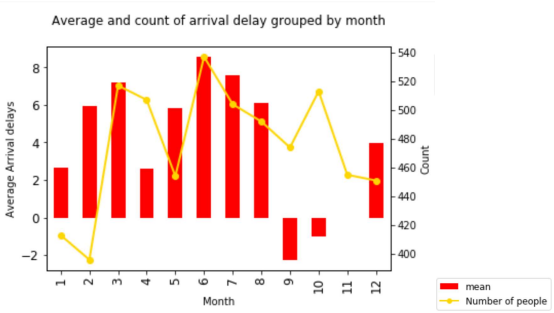

Figure 8: Average and count of arrival delay grouped by month

4 Regression Analysis

4.1 Subpart I

In this section, we will build a model to analyze Arrival delay using Linear regression. We will begin by removing the missing values from weather delay. Our dataset has no missing values now except for Cancellation reason which has 1072 NAN. We have included a total of 9 predictors in our model- 7 numerical & 2 categorical. We shall consider Airline and Day of week as categorical variables and perform one hot encoding on our two categorical variables. We then perform linear regression using statsmodel library and the results of coefficients and p-values is given in table 8.

4.1.1 Interpretations

- With everything else constant, one minute increase in late aircraft delay, increases arrival delay by 0.981362 minutes. With everything else fixed, one minute increase in airline delay, increases arrival delay by 0.982036 minutes. The coefficients suggest that the arrival delay increases with an increase of late aircraft delay, airline delay, air system delay, weather delay and departure delay individually with all others being constant. This makes total sense. As the p values of departure time, airlines, distance and day of week are not significant (high), we will not interpret any of those results, as it might not reveal the truth.
- R-squared: 99.90602987866627%
  The R-squared is really high since there are many variables, but as the p-values suggest, many variables are not that useful.

	Coef	p-values
const	0.574268	0.055692
LATE_AIRCRAFT_DELAY	0.981362	0.000000
AIRLINE_DELAY	0.982036	0.000000
AIR_SYSTEM_DELAY	0.985325	0.000000
WEATHER_DELAY	0.984608	0.000000
DEPARTURE_TIME	-0.000107	0.356129
DEPARTURE_DELAY	0.015849	0.000004
DISTANCE	0.000113	0.269123
Airline_AS	1.890803	0.000015
Airline_B6	0.000871	0.997490
Airline_DL	-0.238513	0.288272
Airline_EV	-0.151854	0.529198
Airline_F9	0.001038	0.998139
Airline_HA	-0.118301	0.829124
Airline_MQ	-0.107958	0.714705
Airline_NK	0.467720	0.160737
Airline_OO	-0.107665	0.654494
Airline_UA	-0.350858	0.132570
Airline_US	-0.169925	0.594466
Airline_VX	-0.139537	0.798189
Airline_WN	-0.173068	0.384351
Dayofweek_2	-0.251671	0.225087
Dayofweek_3	0.089576	0.666804
Dayofweek_4	-0.276833	0.160761
Dayofweek_5	-0.232930	0.240380
Dayofweek_6	-0.290745	0.221801
Dayofweek_7	-0.176871	0.390433

Table 8: Coefficients and p-values of the model (to the left)

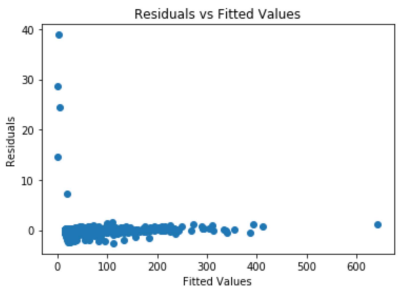

Figure 10: Residual plot

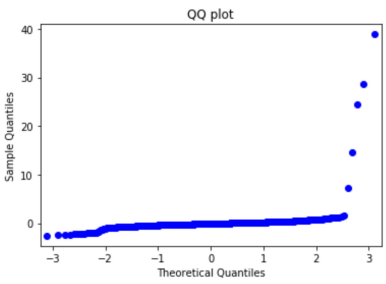

Figure 11: QQ plot

4.1.2 Model diagnostics

- As per figure 10,The residuals then fall within a horizontal band centered around 0, but it displays slight tendencies to be positive. Hence, Linearity is not met.
- Residual plot exhibits slight funnel shape, constant variance is not satisfied.
- As per figure 11,The points fall nowhere close to a straight line The distribution of the error terms departs substantially from a normal distribution. Normality is not satisfied.

4.2 Subpart II

ARRIVAL_DELAY has several outliers, so we can start off our second model by removing those outliers. We are left with 986 observations after removing the outliers. On an attempt to improve our model, we shall perform a log transformation on our response variable. We will refit the above model by removing the nonsignificant predictors from the previous model (p-values>0.5) and airline. Departure time, airline, distance and day of week fit this criteria, so we shall remove those variables from model.

	Coef	p-values
const	2.707571	0.000000e+00
LATE_AIRCRAFT_DELAY	0.019111	1.781589e-209
AIRLINE_DELAY	0.019317	4.990089e-204
AIR_SYSTEM_DELAY	0.020272	0.000000e+00
WEATHER_DELAY	0.019478	1.745813e-127
DEPARTURE_DELAY	0.000854	2.570335e-02

Figure 11: Coefficients and p-values of the model

As per figure 11,All our predictors now are significant as the p-values <0.5.With everything else constant, one minute increase in late aircraft delay, increases arrival delay by 0.0.019 minutes. Similarly, With everything else fixed, one minute increase in airline delay, increases arrival delay by 0.0.019 minutes.

The coefficients suggest that the arrival delay increases with an increase of late aircraft delay, airline delay, air system delay, weather delay and departure delay individually with all others being constant. This makes total sense.The R-squared is has decreased from our previous model to R-squared: 91.68525991830104% since the total of number of variables is less.

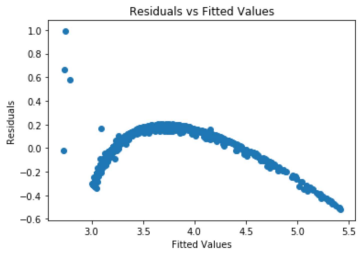

Figure 12: Residual plot

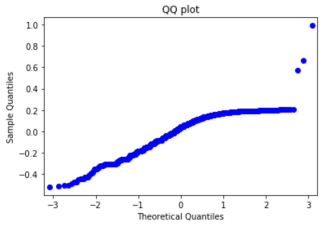

Figure 13: QQ plot

4.2.1 Model diagnostics

- As per figure 12,The residuals depart from zero in a systematic fashion, they are negative for smaller values, positive for medium values and negative again for larger values. This tendency indicates the lack of linearity and lack of constant variance of the regression function.
- As per figure 13,The points fall reasonably close to a straight line initially but the points quickly departs from normality. The distribution of the error terms departs substantially from a normal distribution since the points are not in a straight line. Hence, normality is not satisfied
- No improvement from the previous model in section 4.1. The normality improved slightly, but linearity and constant variance has gotten worse.

4.3 Suggestions on how to improve model fit

1) Tukey's transformation - Performing multiple transformation, till linearity is achieved (at least max level possible)
2) Suitable variable selection technique
3) Adding interaction terms to the model
4) Adding polynomial terms to the model