The objective of this project is to perform exploratory data analysis on the flight delay dataset and investigate the relationship between the flight arrival delay time and a certain set of variables that may contribute to it.

Part 1: Exploratory Data Analysis

The dataset contains **5821 observations** and **31 features/columns**. The features are:

YEAR	ORIGIN_AIRPORT	SCHEDULED_TIME	ARRIVAL_TIME	AIRLINE_DELAY
MONTH	DESTINATION_AIRPORT	ELAPSED_TIME	ARRIVAL_DELAY	LATE_AIRCRAFT_DELAY
DAY	SCHEDULED_DEPARTURE	AIR_TIME	DIVERTED	WEATHER_DELAY
DAY_OF_WEEK	DEPARTURE_TIME	DISTANCE	CANCELLED	
AIRLINE	DEPARTURE_DELAY	WHEELS_ON	CANCELLATION_REASON	
FLIGHT_NUMBER	TAXI_OUT	TAXI_IN	AIR_SYSTEM_DELAY	
TAIL_NUMBER	WHEELS_OFF	SCHEDULED_ARRIVAL	SECURITY_DELAY	

There are 14 unique airlines and their count is as follows:

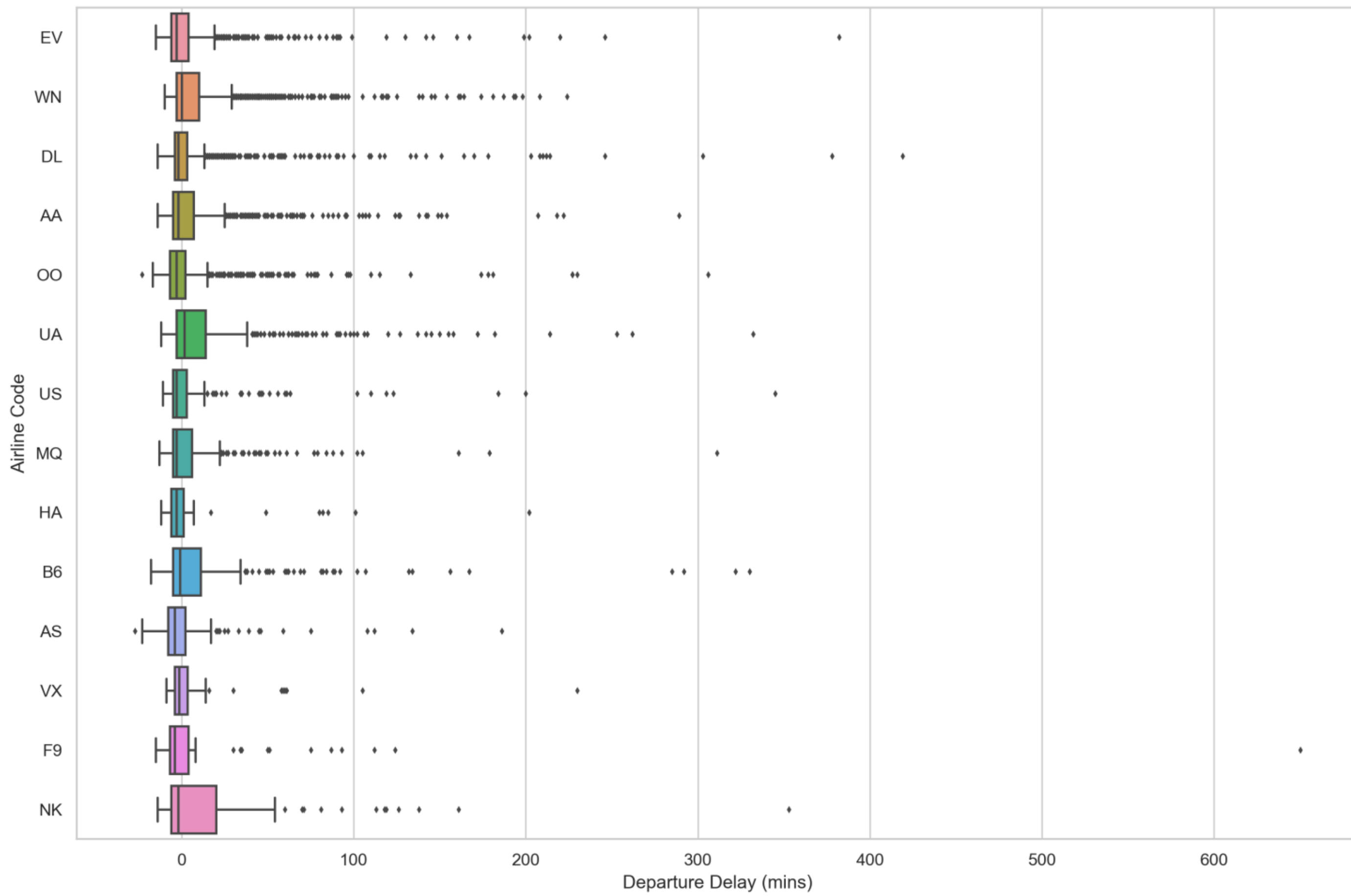
Airline	WN	DL	AA	OO	EV	UA	MQ	B6	US	AS	NK	F9	VX	HA
Count	1285	922	722	593	563	512	288	263	212	145	119	74	66	57

Since we will be analyzing the reasons for arrival delays later, we can clean our data a bit by dropping the observations where either the value of DEPARTURE_DELAY is null (**91** cases) or value of ARRIVAL_DELAY is null (**108** cases). In all cases where ARRIVAL_DELAY is null, DEPARTURE_DELAY is also null (means that the flights were cancelled). However, in the rest of the cases (17), ARRIVAL_DELAY is null because the flights were **diverted** (15 cases) and never reached their destination or 2 cases which are cancelled.

We look at the five number summary of DEPARTURE_DELAY and ARRIVAL_DELAY and notice that the average and median departure delays are **8.89** mins and **-2** mins respectively, and the average and median arrival delays are **3.99** mins and **-5** mins respectively. The distribution of both DEPARTURE_DELAY and ARRIVAL_DELAY is **positively skewed** (there are some high values pulling the mean up), which makes sense because the majority of flights depart and arrive close to their scheduled times, with only a few flights experiencing significant delays. Also noteworthy is that in the case of DEPARTURE_DELAYS, the mean is even larger than the third quartile (7 min) and the maximum value for DEPARTURE_DELAY is 650 mins and the maximum value for ARRIVAL_DELAY is 644 mins.

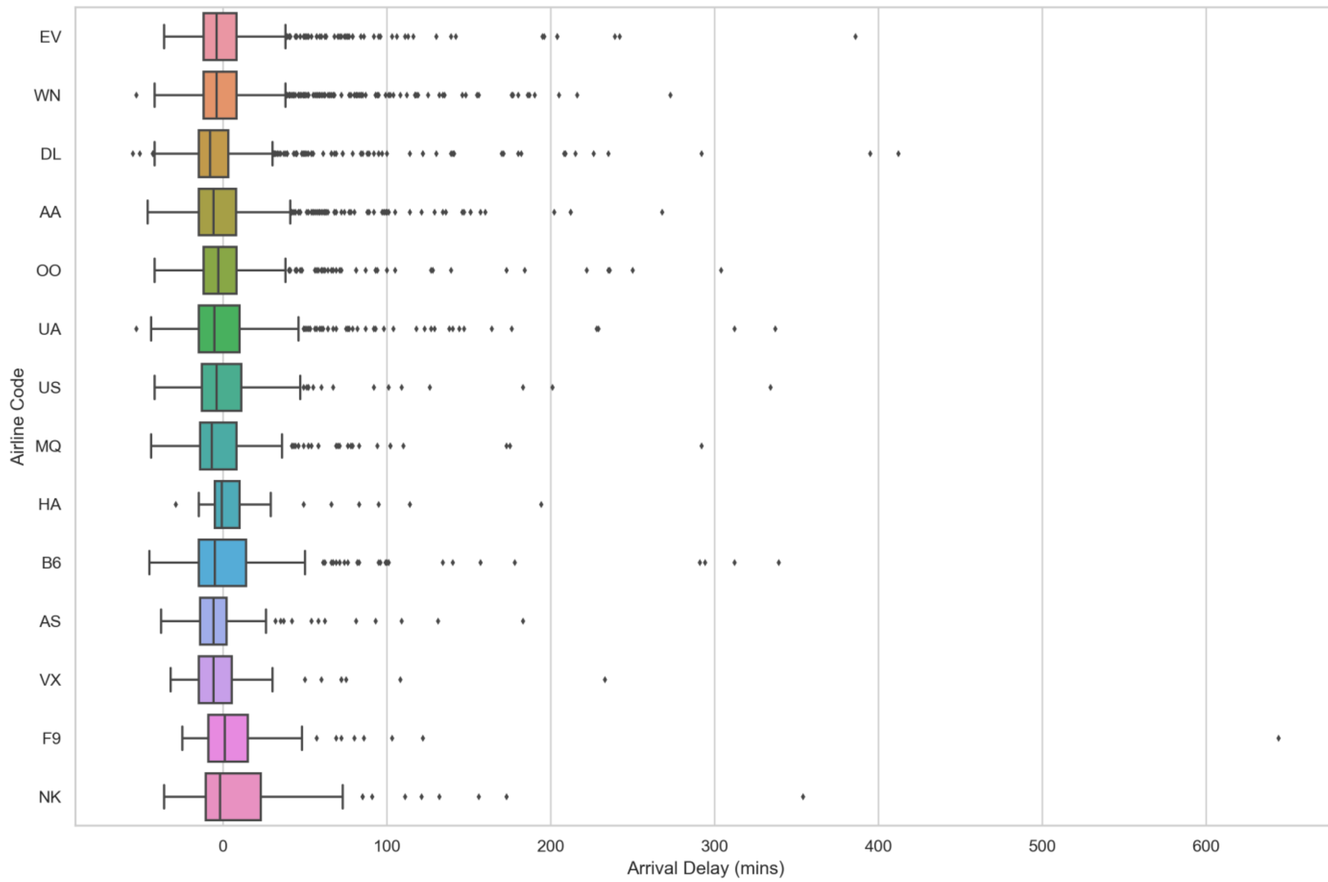
Let us now see the DEPARTURE_DELAY and ARRIVAL_DELAY for each airline via boxplots. We observe that most airlines have similar medians, and most medians are less than 0. Also, there are a lot of extreme values in both types of delay.

DEPARTURE_DELAY:



AIRLINE	count	mean	std	min	25%	50%	75%	max
UA	506.0	13.851779	36.681986	-12.0	-3.0	1.5	14.00	332.0
WN	1269.0	9.894405	27.341367	-10.0	-3.0	0.0	10.00	224.0
B6	257.0	13.645914	46.389097	-18.0	-5.0	-1.0	11.00	330.0
VX	64.0	8.593750	34.744290	-9.0	-4.0	-1.5	3.25	230.0
AA	710.0	8.349296	30.459003	-14.0	-5.0	-2.0	7.00	289.0
DL	918.0	7.238562	34.677903	-14.0	-4.0	-2.0	3.00	419.0
NK	118.0	15.228814	46.313944	-14.0	-6.0	-2.0	20.00	353.0
EV	546.0	7.461538	34.502804	-15.0	-6.0	-3.0	4.00	382.0
HA	57.0	7.964912	35.863812	-12.0	-6.0	-3.0	1.00	202.0
MQ	269.0	7.278810	31.190920	-13.0	-5.0	-3.0	6.00	311.0
OO	575.0	5.702609	30.971892	-23.0	-7.0	-3.0	2.00	306.0
US	206.0	7.393204	36.454659	-11.0	-5.0	-3.0	2.75	345.0
AS	145.0	2.800000	26.927268	-27.0	-8.0	-4.0	2.00	186.0
F9	73.0	14.835616	80.597080	-15.0	-7.0	-4.0	4.00	650.0

ARRIVAL_DELAY



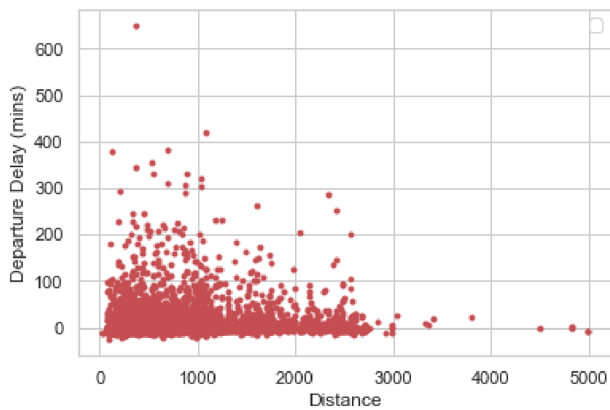
	count	mean	std	min	25%	50%	75%	max
AIRLINE								
UA	506.0	13.851779	36.681986	-12.0	-3.0	1.5	14.00	332.0
WN	1269.0	9.894405	27.341367	-10.0	-3.0	0.0	10.00	224.0
B6	257.0	13.645914	46.389097	-18.0	-5.0	-1.0	11.00	330.0
VX	64.0	8.593750	34.744290	-9.0	-4.0	-1.5	3.25	230.0
AA	710.0	8.349296	30.459003	-14.0	-5.0	-2.0	7.00	289.0
DL	918.0	7.238562	34.677903	-14.0	-4.0	-2.0	3.00	419.0
NK	118.0	15.228814	46.313944	-14.0	-6.0	-2.0	20.00	353.0
EV	546.0	7.461538	34.502804	-15.0	-6.0	-3.0	4.00	382.0
HA	57.0	7.964912	35.863812	-12.0	-6.0	-3.0	1.00	202.0
MQ	269.0	7.278810	31.190920	-13.0	-5.0	-3.0	6.00	311.0
OO	575.0	5.702609	30.971892	-23.0	-7.0	-3.0	2.00	306.0
US	206.0	7.393204	36.454659	-11.0	-5.0	-3.0	2.75	345.0
AS	145.0	2.800000	26.927268	-27.0	-8.0	-4.0	2.00	186.0
F9	73.0	14.835616	80.597080	-15.0	-7.0	-4.0	4.00	650.0

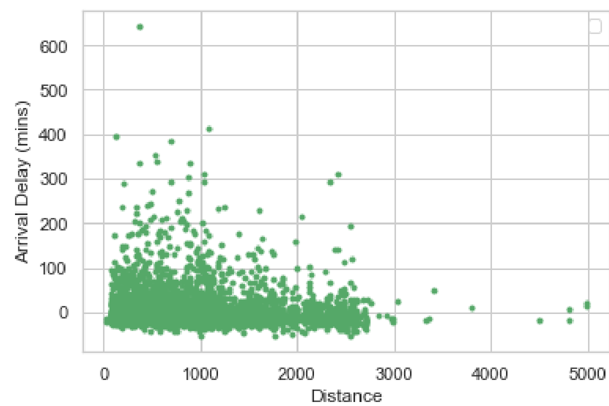
It would also be worthwhile to look at the airports with the highest average DEPARTURE_DELAY

Airport	**FAR**	12898	BMI	ERI	MYR	14576	14696	10157	12992	12206
Avg. DEPARTURE_DELAY (min)	**161**	119	101.33	92	88	88	88	87.5	80	67.5

It is odd that firstly some airports are numbered, and secondly, we don't see any of the major airports here. Airport **'FAR'** (or Hector International Airport) has the **highest average departure delay** of **161 minutes**. Upon analyzing further, we see that in our dataset, there is only 1 observation with the point of origin as FAR. The same is true for the rest of the airports i.e. only a handful of flights (<5) depart from each. That explains the high average!

Next, we check whether DISTANCE has any correlation with DEPARTURE_DELAY and ARRIVAL_DELAY. Departure and Arrival delays can be caused by a variety of factors such as weather conditions, air system (air traffic control) delay, airline delay (mechanical problems or crew issues). The distance of a trip may play a role in some of these factors. For example, longer flights may be more susceptible due to factors such as weather conditions. However, the **distance of a flight is not necessarily a strong predictor of delay.** There may be some correlation between departure/arrival delay and flight distance but it is not a definitive relationship. The same can be seen using a correlation table between departure delay/arrival delay and distance and also by scatter plots between departure delay/arrival delay and distance. We can see that there is a small positive correlation between departure delay and distance (**.023**) and a negative correlation between arrival delay and distance (**-0.027**) which proves our claim.





	DISTANCE
DISTANCE	1.000000
DEPARTURE_DELAY	0.023095
ARRIVAL_DELAY	-0.027935

Furthermore, we will also investigate if there exists any relationship between DAY_OF_WEEK and DEPARTURE_DELAY. Since DAY_OF_WEEK is a categorical column, a correlation coefficient cannot be calculated. Instead, we will look at the five number summary of DEPARTURE_DELAY for every value of DAY_OF_WEEK. It seems like the 3rd day (Wednesday) and 6th day (Saturday) of the week have a lesser departure delay (lesser average and lesser 3rd quantile value) than the other days of the week.

DAY_OF_WEEK	count	mean	std	min	25%	50%	75%	max
1	835.0	9.786826	35.869736	-17.0	-5.0	-2.0	8.0	382.0
2	801.0	8.995006	33.902581	-18.0	-5.0	-2.0	6.0	330.0
3	816.0	7.488971	30.602255	-16.0	-5.0	-2.0	6.0	345.0
4	858.0	9.390443	34.996486	-18.0	-4.0	-1.0	8.0	419.0
5	906.0	9.661148	32.661177	-16.0	-4.0	-1.0	8.0	311.0
6	699.0	7.125894	32.828087	-27.0	-5.0	-2.0	5.0	353.0
7	798.0	9.385965	38.104675	-23.0	-5.0	-1.0	9.0	650.0

We will also check if there is a correlation between ARRIVAL_DELAY and DISTANCE for flights that have departed late. The Pearson coefficient is **-0.095** i.e. the relationship has become **stronger** for cases with a departure delay. This means that the longer the flight, the lesser would be the arrival delay (as compared to when there was no departure delay).

	ARRIVAL_DELAY
ARRIVAL_DELAY	1.000000
DISTANCE	-0.094924

One factor that contributes to the busyness of an airport is the number of flights it handles. Let us take a look at the 10 busiest origin airports and the average departure delay five number statistics for them, and the 10 busiest destination airports and the average arrival delay five number statistics for them. We can see that in terms of DEPARTURE_DELAY, airports such as ATL, DFW, and PHX have lesser departure delays (mean, 75%) in spite of handling more flights compared to airports such as LAX and SFO. Along the same lines, in terms of ARRIVAL_DELAY, airports such as ATL, DFW, and LAX have lesser arrival delays (in spite of handling more flights) compared to airports such as DEN and SFO.

DESTINATION_AIRPORT	count	mean	std	min	25%	50%	75%	max
ATL	365.0	0.186301	31.124587	-43.0	-15.0	-8.0	4.0	209.0
ORD	278.0	10.442446	49.374008	-44.0	-14.0	-4.0	13.0	337.0
DFW	221.0	1.674208	33.736448	-37.0	-15.0	-7.0	6.0	292.0
DEN	203.0	4.911330	31.504707	-36.0	-12.5	-5.0	10.5	156.0
LAX	199.0	5.864322	41.404346	-53.0	-13.5	-5.0	10.0	294.0
PHX	149.0	2.637584	36.771430	-46.0	-14.0	-4.0	5.0	268.0
SFO	141.0	4.198582	35.645320	-44.0	-13.0	-4.0	10.0	215.0
IAH	139.0	2.942446	37.064528	-35.0	-16.5	-6.0	7.5	204.0
LAS	125.0	3.280000	29.255438	-53.0	-11.0	-4.0	10.0	140.0
DTW	121.0	1.958678	39.753909	-29.0	-14.0	-3.0	7.0	395.0

ORIGIN_AIRPORT	count	mean	std	min	25%	50%	75%	max
ATL	326.0	11.273006	48.008355	-8.0	-3.0	-1.0	5.75	650.0
ORD	287.0	13.059233	36.503185	-14.0	-4.0	0.0	18.50	382.0
DFW	235.0	9.468085	32.086722	-11.0	-4.5	-2.0	9.00	289.0
LAX	205.0	6.985366	20.943546	-10.0	-4.0	0.0	8.00	147.0
DEN	191.0	17.256545	46.052969	-11.0	-4.0	0.0	15.50	332.0
IAH	150.0	8.466667	31.356077	-12.0	-5.0	-2.0	7.00	182.0
LAS	137.0	11.131387	34.313824	-14.0	-4.0	0.0	14.00	306.0
SFO	134.0	8.402985	30.204844	-10.0	-4.0	-1.0	8.50	246.0
PHX	133.0	9.774436	25.609130	-11.0	-4.0	-1.0	10.00	124.0
BOS	127.0	5.874016	27.149788	-11.0	-5.0	-2.0	4.00	161.0

Lastly, we checked **which months** on an average have the highest DEPARTURE_DELAY and ARRIVAL_DELAY, and **June** and **July** were the top 2 months for both (Mean, Median, and 75 quantile values). A possible reason could be that many families take vacations at that time of year and kids have no school.

DEPARTURE_DELAY

MONTH	count	mean	std	min	25%	50%	75%	max
7	504.0	12.644841	42.234210	-18.0	-4.00	0.0	13.00	650.0
1	413.0	7.532688	31.970663	-18.0	-5.00	-1.0	8.00	378.0
2	396.0	9.603535	33.594935	-15.0	-5.00	-1.0	8.00	303.0
3	517.0	12.353965	42.004711	-16.0	-4.00	-1.0	12.00	419.0
5	454.0	10.544053	33.058468	-15.0	-5.00	-1.0	9.75	212.0
6	537.0	11.832402	36.208427	-14.0	-4.00	-1.0	10.00	322.0
8	492.0	10.008130	39.760484	-18.0	-4.25	-1.0	7.00	353.0
4	507.0	7.252465	30.419715	-23.0	-5.00	-2.0	7.00	345.0
10	513.0	4.947368	25.754080	-27.0	-5.00	-2.0	2.00	222.0
11	455.0	5.345055	26.630314	-16.0	-5.00	-2.0	5.00	332.0
12	451.0	9.050998	32.707452	-16.0	-5.00	-2.0	6.00	246.0
9	474.0	4.854430	28.431772	-23.0	-6.00	-3.0	1.00	289.0
	count	mean	std	min	25%	50%	75%	max

ARRIVAL_DELAY

MONTH	count	mean	std	min	25%	50%	75%	max
6	537.0	8.571695	39.842074	-55.0	-12.0	-3.0	12.0	312.0
7	504.0	7.605159	44.483884	-36.0	-12.0	-3.0	13.0	644.0
2	396.0	5.921717	36.834786	-42.0	-13.0	-4.0	9.0	292.0
3	517.0	7.189555	45.092829	-53.0	-13.0	-4.0	11.0	412.0
5	454.0	5.843612	35.916505	-36.0	-13.0	-4.0	10.0	226.0
8	492.0	6.121951	41.932316	-44.0	-12.0	-4.0	8.0	354.0
1	413.0	2.663438	35.620308	-40.0	-13.0	-5.0	7.0	395.0
11	455.0	-0.015385	29.287685	-42.0	-14.0	-5.0	6.5	337.0
4	507.0	2.601578	32.362128	-45.0	-13.0	-6.0	7.0	334.0
10	513.0	-1.015595	28.361088	-51.0	-15.0	-7.0	3.0	212.0
12	451.0	3.988914	35.884783	-44.0	-15.0	-7.0	10.0	273.0
9	474.0	-2.253165	30.731725	-46.0	-16.0	-8.0	1.0	268.0

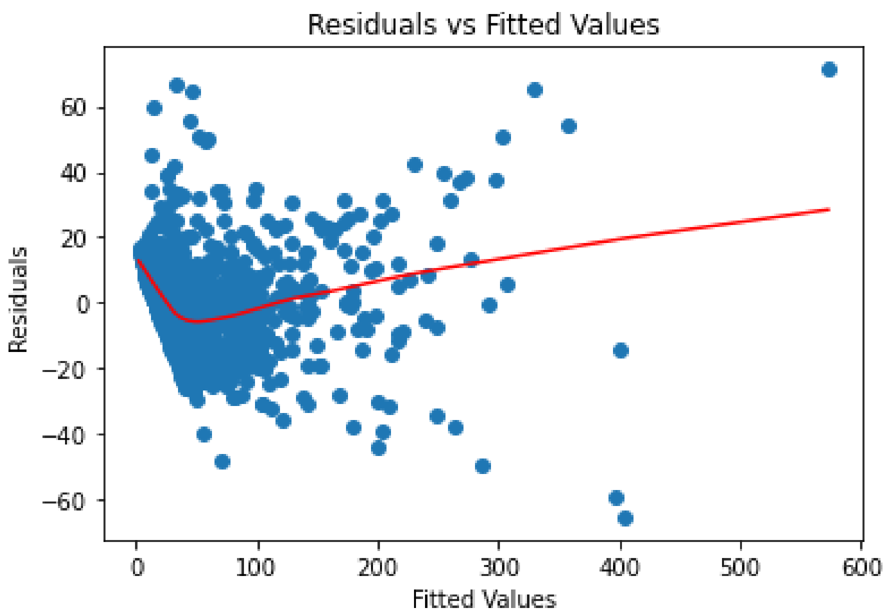
Part 2: Regression Analysis

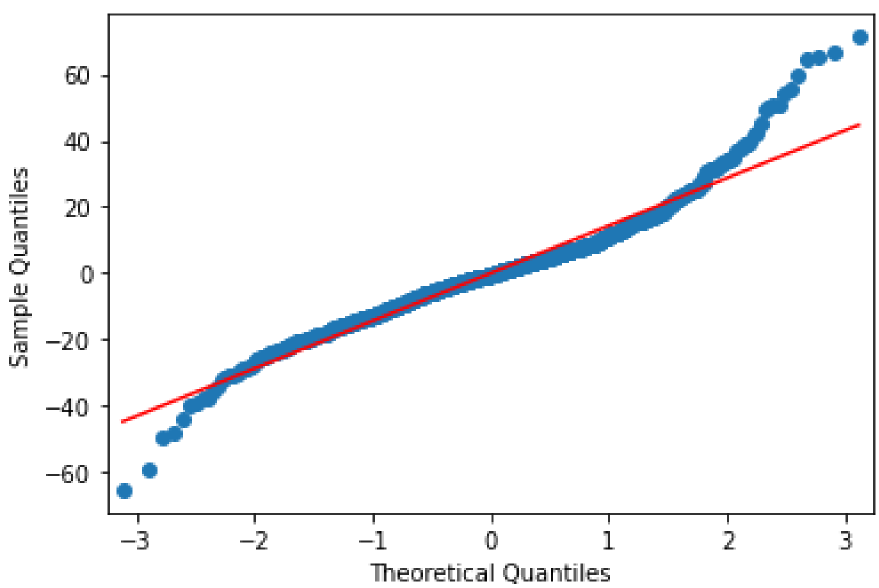
Now, we will use linear regression to model a relationship between ARRIVAL_DELAY and several other predictor variables in the dataset. Before we start about doing that, we **clean** the dataset further by **removing** all **observations** with **null values** in the **WEATHER_DELAY** column (all those observations also have null vales in columns such as AIR_SYSTEM_DELAY, SECURITY_DELAY, AIRLINE_DELAY, and LATE_AIRCRAFT_DELAY. We are now left with 1072 observations.

We then build a model **using stats models library**, with the following 8 independent variables: LATE_AIRCRAFT_DELAY, AIR_SYSTEM_DELAY, WEATHER_DELAY, DAY_OF_WEEK, DEPARTURE_TIME, DEPARTURE_DELAY, DISTANCE, AIRLINE. It is important to note here that since **DAY_OF_WEEK** and **AIRLINE** are **categorical variables**, we create **dummy variables** for them (The **reference level** for DAY_OF_WEEK is **Friday** and the reference level for AIRLINE is **Airline AA**). The summary of the model is below:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          ARRIVAL_DELAY   R-squared:                       0.939
Model:                            OLS   Adj. R-squared:                  0.938
Method:                 Least Squares   F-statistic:                     643.6
Date:                Sun, 05 Mar 2023   Prob (F-statistic):               0.00
Time:                        02:06:32   Log-Likelihood:                -4377.8
No. Observations:                1072   AIC:                             8808.
Df Residuals:                    1046   BIC:                             8937.
Df Model:                          25
Covariance Type:            nonrobust
=================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const                  14.0775      2.325      6.055      0.000       9.516      18.639
LATE_AIRCRAFT_DELAY     0.0959      0.015      6.351      0.000       0.066       0.126
AIR_SYSTEM_DELAY        0.3532      0.016     21.881      0.000       0.321       0.385
WEATHER_DELAY           0.1923      0.022      8.687      0.000       0.149       0.236
DEPARTURE_TIME         -0.0050      0.001     -5.407      0.000      -0.007      -0.003
DEPARTURE_DELAY         0.8416      0.010     87.442      0.000       0.823       0.861
DISTANCE                0.0006      0.001      0.714      0.476      -0.001       0.002
DAY_OF_WEEK_Monday     -1.1038      1.597     -0.691      0.490      -4.237       2.029
DAY_OF_WEEK_Saturday   -0.9512      1.874     -0.507      0.612      -4.629       2.727
DAY_OF_WEEK_Sunday     -0.1068      1.605     -0.067      0.947      -3.257       3.043
DAY_OF_WEEK_Thursday    0.9111      1.533      0.594      0.552      -2.097       3.919
DAY_OF_WEEK_Tuesday    -1.8314      1.625     -1.127      0.260      -5.021       1.358
DAY_OF_WEEK_Wednesday  -0.7913      1.626     -0.487      0.627      -3.982       2.399
AIRLINE_AS             -2.0719      3.496     -0.593      0.554      -8.933       4.789
AIRLINE_B6              0.7711      2.229      0.346      0.729      -3.602       5.144
AIRLINE_DL             -2.6744      1.807     -1.480      0.139      -6.220       0.871
AIRLINE_EV             -0.4124      1.943     -0.212      0.832      -4.226       3.401
AIRLINE_F9              5.5515      3.579      1.551      0.121      -1.471      12.574
AIRLINE_HA              7.0726      4.409      1.604      0.109      -1.578      15.724
AIRLINE_MQ             -1.3426      2.378     -0.565      0.572      -6.009       3.324
AIRLINE_NK             -1.1258      2.684     -0.419      0.675      -6.392       4.141
AIRLINE_OO              1.1141      1.937      0.575      0.565      -2.687       4.915
AIRLINE_UA             -6.9494      1.866     -3.723      0.000     -10.612      -3.287
AIRLINE_US             -0.0638      2.570     -0.025      0.980      -5.108       4.980
AIRLINE_VX              6.9887      4.389      1.592      0.112      -1.624      15.602
AIRLINE_WN             -4.0197      1.598     -2.516      0.012      -7.154      -0.885
==============================================================================
Omnibus:                      139.151   Durbin-Watson:                   2.022
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              530.760
Skew:                           0.575   Prob(JB):                    5.58e-116
Kurtosis:                       6.249   Cond. No.                     2.20e+04
==============================================================================
```

To test for linear model's assumptions, we plot a scatter plot of the residuals against the fitted values. It is observed that the **variance of the residuals increases** as the fitted values increase and there is **heteroscedasticity**. To check for linearity, we fit a non-parametric curve (or a lowess line) to the scatterplot and see that the **lowess line is not linear and has a curve**, which means linearity assumption is also violated. Examining the **QQ plot**, we can see that the plot looks somewhat linear. The center follows a straight line but both the ends deviate quite a lot (**heavy tails**). The data is not precisely normally distributed, but it’s not too far off. The R-squared is 93.89%, and at 5% significance level the following predictors are significant: LATE_AIRCRAFT_DELAY, AIR_SYSTEM_DELAY, WEATHER_DELAY, DEPARTURE_TIME, DEPARTURE_DELAY, AIRLINE_UA, AIRLINE_WN.
Overall, the **linear model is not fitting the data well**.





We can **interpret** a few of the significant coefficients as follows:
Fixing everything else, for every 1 minute delay due to air systems, the arrival delay of a flight increases by 0.3532 minutes.

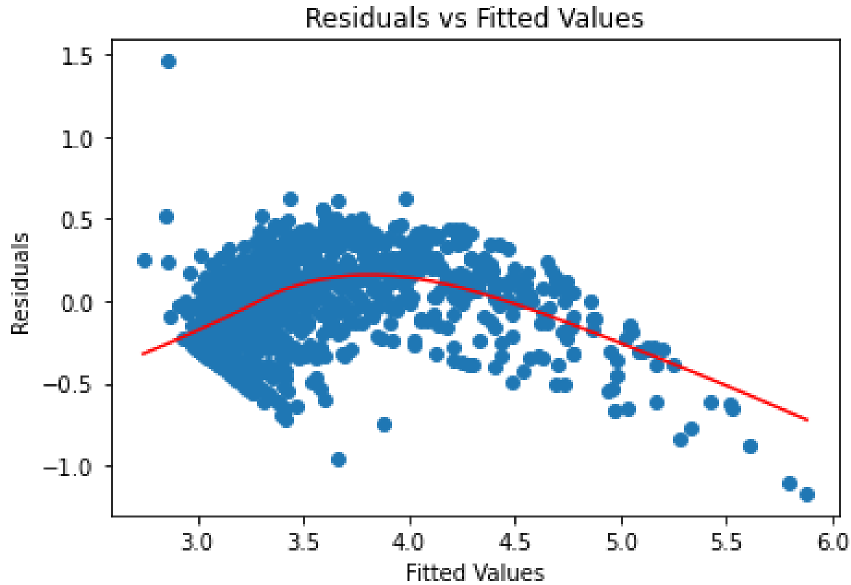
Fixing everything else, for every 1 minute delay in the departure of a flight, its arrival delay increases by 0.8416 minutes. This **makes sense because** if a flight departs late, it can only catch up on time mid-air to a certain extent and will ultimately arrive late.
Fixing everything else, Airline UA on average has a 6.95 mins lesser arrival delay than Airline AA (our reference airline)
Fixing everything else, Airline WN on average has a 4.02 min lesser arrival delay than Airline AA (our reference airline)

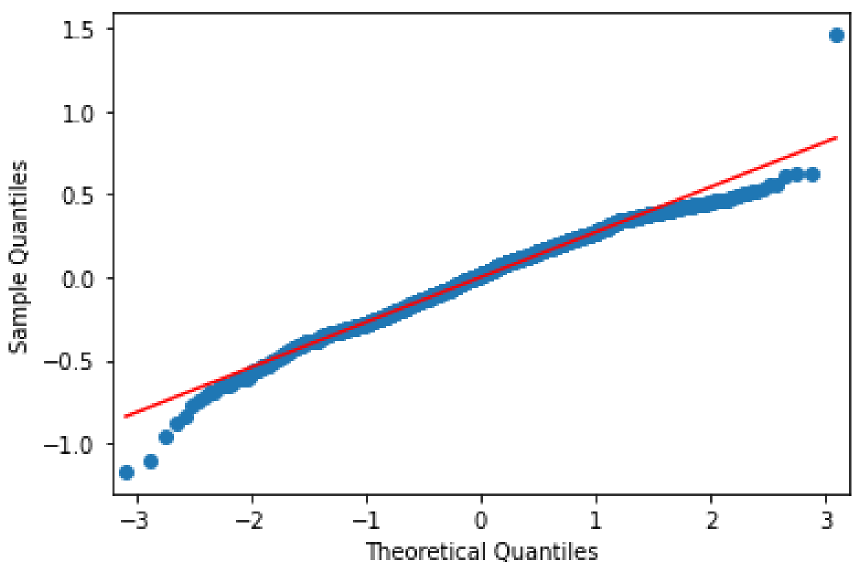
Model Improvement
To improve our model, let us drop the outliers from ARRIVAL_DATA, make the response variable as log(ARRIVAL_DATA), and remove the insignificant predictors from the above model.

The summary of the refined model is below:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          ARRIVAL_DELAY   R-squared:                       0.789
Model:                            OLS   Adj. R-squared:                  0.788
Method:                 Least Squares   F-statistic:                     523.3
Date:                Sun, 05 Mar 2023   Prob (F-statistic):               0.00
Time:                        00:16:21   Log-Likelihood:                -115.36
No. Observations:                 986   AIC:                             246.7
Df Residuals:                     978   BIC:                             285.9
Df Model:                           7
Covariance Type:            nonrobust
=========================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                   2.9142      0.031     93.812      0.000       2.853       2.975
LATE_AIRCRAFT_DELAY     0.0041      0.000      9.060      0.000       0.003       0.005
AIR_SYSTEM_DELAY        0.0131      0.000     27.278      0.000       0.012       0.014
WEATHER_DELAY           0.0054      0.001      5.710      0.000       0.004       0.007
DEPARTURE_TIME      -5.004e-05   1.87e-05     -2.676      0.008   -8.67e-05   -1.33e-05
DEPARTURE_DELAY         0.0135      0.000     39.505      0.000       0.013       0.014
AIRLINE_UA             -0.0941      0.030     -3.162      0.002      -0.152      -0.036
AIRLINE_WN             -0.0343      0.022     -1.558      0.119      -0.077       0.009
==============================================================================
Omnibus:                       29.604   Durbin-Watson:                   2.041
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               46.039
Skew:                          -0.263   Prob(JB):                     1.01e-10
Kurtosis:                       3.919   Cond. No.                     6.04e+03
==============================================================================
```

We test again for linear model's assumptions. The Residuals vs Fitted Values plot **still looks heteroskedastic**. The **lowess line is not linear and rather is quadratic-like**, means the relationship between the response and predictors **is still not linear**. What improved is that we now have less outliers. Examining the **QQ plot**, we can see that the plot looks somewhat linear. The data is not precisely normally distributed and there is some negative skewness. The R-squared is 78.9%, and at 5% significance level the following predictors are significant: LATE_AIRCRAFT_DELAY, AIR_SYSTEM_DELAY, WEATHER_DELAY, DEPARTURE_TIME, DEPARTURE_DELAY, AIRLINE_UA.





We can **interpret** a few of the significant coefficients as follows:
Fixing everything else, for every 1 minute delay due to air systems, the average arrival delay of a flight will be multiplied by exp(0.0131) = 1.0132 times
Fixing everything else, for every 1 minute delay in the departure of a flight, its average arrival delay will be multiplied by exp(0.0135) = 1.0136 times
Fixing everything else, Airline UA average arrival delay will be multiplied by exp(-0.0941) = 0.91 times to that of Airline AA (our reference airline)

<u>Suggestions to improve the model:</u>
Firstly, since the relationship is still non-linear, perhaps some higher degree terms can be added to the model. Secondly, interaction terms can also be added (interactions between various kinds of delays). Thirdly, the summary output suggests a strong multicollinearity between the independent variables. A correlation matrix between them indicated that the Pearson correlation coefficient between LATE_AIRCRAFT_DELAY and DEPARTURE_DELAY was 0.6 (which is between moderate and strong). Hence, to tackle this issue, we can consider removing one of them from the model or address

the multicollinearity in other ways. Lastly, the departure time variable does not make much sense because the model treats it like an integer whereas it's actually a time value. Ideally, the predicted arrival delay (keeping everything else the same) when time is 0000 and when it is 2359 should be close, but it will not be close in this model. Perhaps it can be removed as well.