Kaggle_Flight_Data_Analysis_Notebook

October 24, 2021

1 Mount dataset resources

```
Mounted at /content/drive
```

```
/content/drive/MyDrive/github/eda_examples/Kaggle_Flight_Data_Analysis
```

[3]:
```
'/content/drive/MyDrive/github/eda_examples/Kaggle_Flight_Data_Analysis'
```

2 Kaggle 2015 Flight Delay Data Analysis

[5]:
```
   YEAR  MONTH  DAY  DAY_OF_WEEK AIRLINE  FLIGHT_NUMBER TAIL_NUMBER  \
0  2015      1    1            4      EV           4160      N11150
1  2015      1    1            4      AA           1635      N025AA
2  2015      1    1            4      WN            119      N271LV
3  2015      1    1            4      EV           4936      N738EV
4  2015      1    1            4      DL           2319      N960DL

  ORIGIN_AIRPORT DESTINATION_AIRPORT  SCHEDULED_DEPARTURE  DEPARTURE_TIME  \
0            JAX                 EWR                  540           531.0
1            ATL                 DFW                  625             NaN
2            RSW                 ATL                  800           754.0
3            MSP                 IAD                  900           901.0
4            LGA                 MSP                 1010          1010.0

   DEPARTURE_DELAY  TAXI_OUT  WHEELS_OFF  SCHEDULED_TIME  ELAPSED_TIME  \
0             -9.0       9.0       540.0             137         132.0
1              NaN       NaN         NaN             150           NaN
2             -6.0      11.0       805.0             105         100.0
3              1.0      56.0       957.0             148         159.0
4              0.0      22.0      1032.0             200         195.0

   AIR_TIME  DISTANCE  WHEELS_ON  TAXI_IN  SCHEDULED_ARRIVAL  ARRIVAL_TIME  \
0     109.0       820      729.0     14.0                757         743.0
1       NaN       731        NaN      NaN                755           NaN
2      84.0       515      929.0      5.0                945         934.0
3     100.0       908     1237.0      3.0               1228        1240.0
```

```
4     171.0      1020      1223.0      2.0                 1230         1225.0

   ARRIVAL_DELAY  DIVERTED  CANCELLED CANCELLATION_REASON  AIR_SYSTEM_DELAY  \
0          -14.0         0          0                 NaN               NaN
1            NaN         0          1                   B               NaN
2          -11.0         0          0                 NaN               NaN
3           12.0         0          0                 NaN               NaN
4           -5.0         0          0                 NaN               NaN

   SECURITY_DELAY  AIRLINE_DELAY  LATE_AIRCRAFT_DELAY  WEATHER_DELAY
0             NaN            NaN                  NaN            NaN
1             NaN            NaN                  NaN            NaN
2             NaN            NaN                  NaN            NaN
3             NaN            NaN                  NaN            NaN
4             NaN            NaN                  NaN            NaN
```

2.0.1 Part 1: Exploratory Analysis

1. How many observations are there? How many features are there?

```
 5821 - Total number of observations
   31 - Total number of features

The column names in the flights dataset are:
 ['YEAR' 'MONTH' 'DAY' 'DAY_OF_WEEK' 'AIRLINE' 'FLIGHT_NUMBER'
 'TAIL_NUMBER' 'ORIGIN_AIRPORT' 'DESTINATION_AIRPORT'
 'SCHEDULED_DEPARTURE' 'DEPARTURE_TIME' 'DEPARTURE_DELAY' 'TAXI_OUT'
 'WHEELS_OFF' 'SCHEDULED_TIME' 'ELAPSED_TIME' 'AIR_TIME' 'DISTANCE'
 'WHEELS_ON' 'TAXI_IN' 'SCHEDULED_ARRIVAL' 'ARRIVAL_TIME' 'ARRIVAL_DELAY'
 'DIVERTED' 'CANCELLED' 'CANCELLATION_REASON' 'AIR_SYSTEM_DELAY'
 'SECURITY_DELAY' 'AIRLINE_DELAY' 'LATE_AIRCRAFT_DELAY' 'WEATHER_DELAY']
```

2. How many different airlines are there? What are their counts?

```
   14 - total num of different airlines in the dataset

The count for the airlines in the dataset in the descending order are:
```

[23]:

```
AIRLINE
WN    1285
DL     922
AA     722
OO     593
EV     563
UA     512
MQ     288
B6     263
US     212
```

```
AS     145
NK     119
F9      74
VX      66
HA      57
Name: AIRLINE, dtype: int64
```

3. How many missing values are there in the departure delays? How about arrival delays? Do they match? Why or why not? Remove these observations afterwards.

```
 91 - total number of missing values in departure delays
108 - total number of missing values in arrival delays
```

The number of missing values for departure delays and arrival delays DO NOT match. We have more missing values for arrival delays.

[25]:

	DEPARTURE_TIME	DEPARTURE_DELAY	ARRIVAL_TIME	ARRIVAL_DELAY
1	NaN	NaN	NaN	NaN
10	NaN	NaN	NaN	NaN
47	NaN	NaN	NaN	NaN
115	NaN	NaN	NaN	NaN
116	NaN	NaN	NaN	NaN
172	NaN	NaN	NaN	NaN
174	NaN	NaN	NaN	NaN
190	NaN	NaN	NaN	NaN
350	1221.0	31.0	NaN	NaN
359	NaN	NaN	NaN	NaN
362	NaN	NaN	NaN	NaN
363	NaN	NaN	NaN	NaN
365	NaN	NaN	NaN	NaN
367	NaN	NaN	NaN	NaN
371	NaN	NaN	NaN	NaN
372	NaN	NaN	NaN	NaN
431	NaN	NaN	NaN	NaN
432	NaN	NaN	NaN	NaN
434	NaN	NaN	NaN	NaN
437	NaN	NaN	NaN	NaN
446	NaN	NaN	NaN	NaN
447	NaN	NaN	NaN	NaN
449	NaN	NaN	NaN	NaN
453	NaN	NaN	NaN	NaN
465	NaN	NaN	NaN	NaN
467	NaN	NaN	NaN	NaN
478	NaN	NaN	NaN	NaN
498	NaN	NaN	NaN	NaN
513	NaN	NaN	NaN	NaN
545	NaN	NaN	NaN	NaN
551	NaN	NaN	NaN	NaN
638	NaN	NaN	NaN	NaN

683	NaN	NaN	NaN	NaN
689	NaN	NaN	NaN	NaN
740	NaN	NaN	NaN	NaN
741	NaN	NaN	NaN	NaN
760	NaN	NaN	NaN	NaN
778	NaN	NaN	NaN	NaN
782	NaN	NaN	NaN	NaN
786	NaN	NaN	NaN	NaN
801	NaN	NaN	NaN	NaN
826	NaN	NaN	NaN	NaN
830	NaN	NaN	NaN	NaN
856	1302.0	21.0	2008.0	NaN
860	NaN	NaN	NaN	NaN
861	NaN	NaN	NaN	NaN
869	NaN	NaN	NaN	NaN
899	NaN	NaN	NaN	NaN
934	NaN	NaN	NaN	NaN
1015	NaN	NaN	NaN	NaN
1217	NaN	NaN	NaN	NaN
1372	1936.0	-4.0	143.0	NaN
1521	NaN	NaN	NaN	NaN
1605	NaN	NaN	NaN	NaN
1799	1317.0	2.0	1735.0	NaN
1804	NaN	NaN	NaN	NaN
1901	NaN	NaN	NaN	NaN
2029	NaN	NaN	NaN	NaN
2038	NaN	NaN	NaN	NaN
2055	NaN	NaN	NaN	NaN
2089	1316.0	1.0	2225.0	NaN
2110	NaN	NaN	NaN	NaN
2152	1609.0	22.0	NaN	NaN
2153	NaN	NaN	NaN	NaN
2196	NaN	NaN	NaN	NaN
2263	NaN	NaN	NaN	NaN
2277	NaN	NaN	NaN	NaN
2291	NaN	NaN	NaN	NaN
2368	NaN	NaN	NaN	NaN
2478	1821.0	46.0	2322.0	NaN
2533	NaN	NaN	NaN	NaN
2560	NaN	NaN	NaN	NaN
2577	1911.0	-9.0	2241.0	NaN
2716	NaN	NaN	NaN	NaN
2842	NaN	NaN	NaN	NaN
2899	NaN	NaN	NaN	NaN
2926	1733.0	8.0	2311.0	NaN
3011	NaN	NaN	NaN	NaN
3050	NaN	NaN	NaN	NaN

```
3128            NaN             NaN           NaN            NaN
3194            NaN             NaN           NaN            NaN
3208         1255.0             4.0        1807.0            NaN
3251            NaN             NaN           NaN            NaN
3426            NaN             NaN           NaN            NaN
3445            NaN             NaN           NaN            NaN
3522         1831.0           101.0        2313.0            NaN
3568            NaN             NaN           NaN            NaN
3578            NaN             NaN           NaN            NaN
3661            NaN             NaN           NaN            NaN
3699         1613.0            48.0        2159.0            NaN
3766          643.0            -2.0           NaN            NaN
4171            NaN             NaN           NaN            NaN
4295         1442.0            -3.0        2143.0            NaN
4771         1741.0             1.0         246.0            NaN
4809            NaN             NaN           NaN            NaN
5097            NaN             NaN           NaN            NaN
5183            NaN             NaN           NaN            NaN
5232         1316.0            -2.0        1808.0            NaN
5247            NaN             NaN           NaN            NaN
5250            NaN             NaN           NaN            NaN
5567            NaN             NaN           NaN            NaN
5576            NaN             NaN           NaN            NaN
5587            NaN             NaN           NaN            NaN
5595            NaN             NaN           NaN            NaN
5641         2001.0            29.0         230.0            NaN
5716            NaN             NaN           NaN            NaN
5755            NaN             NaN           NaN            NaN
5764            NaN             NaN           NaN            NaN
```

From the above subset of data we can see that there are flights with departure time but are missing arrival delay values.

[26]:

```
      DEPARTURE_TIME  DEPARTURE_DELAY  ARRIVAL_TIME  ARRIVAL_DELAY  DIVERTED  \
1                NaN              NaN           NaN            NaN         0
10               NaN              NaN           NaN            NaN         0
47               NaN              NaN           NaN            NaN         0
115              NaN              NaN           NaN            NaN         0
116              NaN              NaN           NaN            NaN         0
172              NaN              NaN           NaN            NaN         0
174              NaN              NaN           NaN            NaN         0
190              NaN              NaN           NaN            NaN         0
350           1221.0             31.0           NaN            NaN         1
359              NaN              NaN           NaN            NaN         0
362              NaN              NaN           NaN            NaN         0
363              NaN              NaN           NaN            NaN         0
365              NaN              NaN           NaN            NaN         0
367              NaN              NaN           NaN            NaN         0
```

371	NaN	NaN	NaN	NaN	0
372	NaN	NaN	NaN	NaN	0
431	NaN	NaN	NaN	NaN	0
432	NaN	NaN	NaN	NaN	0
434	NaN	NaN	NaN	NaN	0
437	NaN	NaN	NaN	NaN	0
446	NaN	NaN	NaN	NaN	0
447	NaN	NaN	NaN	NaN	0
449	NaN	NaN	NaN	NaN	0
453	NaN	NaN	NaN	NaN	0
465	NaN	NaN	NaN	NaN	0
467	NaN	NaN	NaN	NaN	0
478	NaN	NaN	NaN	NaN	0
498	NaN	NaN	NaN	NaN	0
513	NaN	NaN	NaN	NaN	0
545	NaN	NaN	NaN	NaN	0
551	NaN	NaN	NaN	NaN	0
638	NaN	NaN	NaN	NaN	0
683	NaN	NaN	NaN	NaN	0
689	NaN	NaN	NaN	NaN	0
740	NaN	NaN	NaN	NaN	0
741	NaN	NaN	NaN	NaN	0
760	NaN	NaN	NaN	NaN	0
778	NaN	NaN	NaN	NaN	0
782	NaN	NaN	NaN	NaN	0
786	NaN	NaN	NaN	NaN	0
801	NaN	NaN	NaN	NaN	0
826	NaN	NaN	NaN	NaN	0
830	NaN	NaN	NaN	NaN	0
856	1302.0	21.0	2008.0	NaN	1
860	NaN	NaN	NaN	NaN	0
861	NaN	NaN	NaN	NaN	0
869	NaN	NaN	NaN	NaN	0
899	NaN	NaN	NaN	NaN	0
934	NaN	NaN	NaN	NaN	0
1015	NaN	NaN	NaN	NaN	0
1217	NaN	NaN	NaN	NaN	0
1372	1936.0	-4.0	143.0	NaN	1
1521	NaN	NaN	NaN	NaN	0
1605	NaN	NaN	NaN	NaN	0
1799	1317.0	2.0	1735.0	NaN	1
1804	NaN	NaN	NaN	NaN	0
1901	NaN	NaN	NaN	NaN	0
2029	NaN	NaN	NaN	NaN	0
2038	NaN	NaN	NaN	NaN	0
2055	NaN	NaN	NaN	NaN	0
2089	1316.0	1.0	2225.0	NaN	1

2110	NaN	NaN	NaN	NaN	0
2152	1609.0	22.0	NaN	NaN	0
2153	NaN	NaN	NaN	NaN	0
2196	NaN	NaN	NaN	NaN	0
2263	NaN	NaN	NaN	NaN	0
2277	NaN	NaN	NaN	NaN	0
2291	NaN	NaN	NaN	NaN	0
2368	NaN	NaN	NaN	NaN	0
2478	1821.0	46.0	2322.0	NaN	1
2533	NaN	NaN	NaN	NaN	0
2560	NaN	NaN	NaN	NaN	0
2577	1911.0	-9.0	2241.0	NaN	1
2716	NaN	NaN	NaN	NaN	0
2842	NaN	NaN	NaN	NaN	0
2899	NaN	NaN	NaN	NaN	0
2926	1733.0	8.0	2311.0	NaN	1
3011	NaN	NaN	NaN	NaN	0
3050	NaN	NaN	NaN	NaN	0
3128	NaN	NaN	NaN	NaN	0
3194	NaN	NaN	NaN	NaN	0
3208	1255.0	4.0	1807.0	NaN	1
3251	NaN	NaN	NaN	NaN	0
3426	NaN	NaN	NaN	NaN	0
3445	NaN	NaN	NaN	NaN	0
3522	1831.0	101.0	2313.0	NaN	1
3568	NaN	NaN	NaN	NaN	0
3578	NaN	NaN	NaN	NaN	0
3661	NaN	NaN	NaN	NaN	0
3699	1613.0	48.0	2159.0	NaN	1
3766	643.0	-2.0	NaN	NaN	0
4171	NaN	NaN	NaN	NaN	0
4295	1442.0	-3.0	2143.0	NaN	1
4771	1741.0	1.0	246.0	NaN	1
4809	NaN	NaN	NaN	NaN	0
5097	NaN	NaN	NaN	NaN	0
5183	NaN	NaN	NaN	NaN	0
5232	1316.0	-2.0	1808.0	NaN	1
5247	NaN	NaN	NaN	NaN	0
5250	NaN	NaN	NaN	NaN	0
5567	NaN	NaN	NaN	NaN	0
5576	NaN	NaN	NaN	NaN	0
5587	NaN	NaN	NaN	NaN	0
5595	NaN	NaN	NaN	NaN	0
5641	2001.0	29.0	230.0	NaN	1
5716	NaN	NaN	NaN	NaN	0
5755	NaN	NaN	NaN	NaN	0
5764	NaN	NaN	NaN	NaN	0

	CANCELLED
1	1
10	1
47	1
115	1
116	1
172	1
174	1
190	1
350	0
359	1
362	1
363	1
365	1
367	1
371	1
372	1
431	1
432	1
434	1
437	1
446	1
447	1
449	1
453	1
465	1
467	1
478	1
498	1
513	1
545	1
551	1
638	1
683	1
689	1
740	1
741	1
760	1
778	1
782	1
786	1
801	1
826	1
830	1
856	0
860	1

861	1
869	1
899	1
934	1
1015	1
1217	1
1372	0
1521	1
1605	1
1799	0
1804	1
1901	1
2029	1
2038	1
2055	1
2089	0
2110	1
2152	1
2153	1
2196	1
2263	1
2277	1
2291	1
2368	1
2478	0
2533	1
2560	1
2577	0
2716	1
2842	1
2899	1
2926	0
3011	1
3050	1
3128	1
3194	1
3208	0
3251	1
3426	1
3445	1
3522	0
3568	1
3578	1
3661	1
3699	0
3766	1
4171	1

```
4295          0
4771          0
4809          1
5097          1
5183          1
5232          0
5247          1
5250          1
5567          1
5576          1
5587          1
5595          1
5641          0
5716          1
5755          1
5764          1
```

From the above subset of data we can conclude that this mismatch in the missing values is due to the flight diversion.

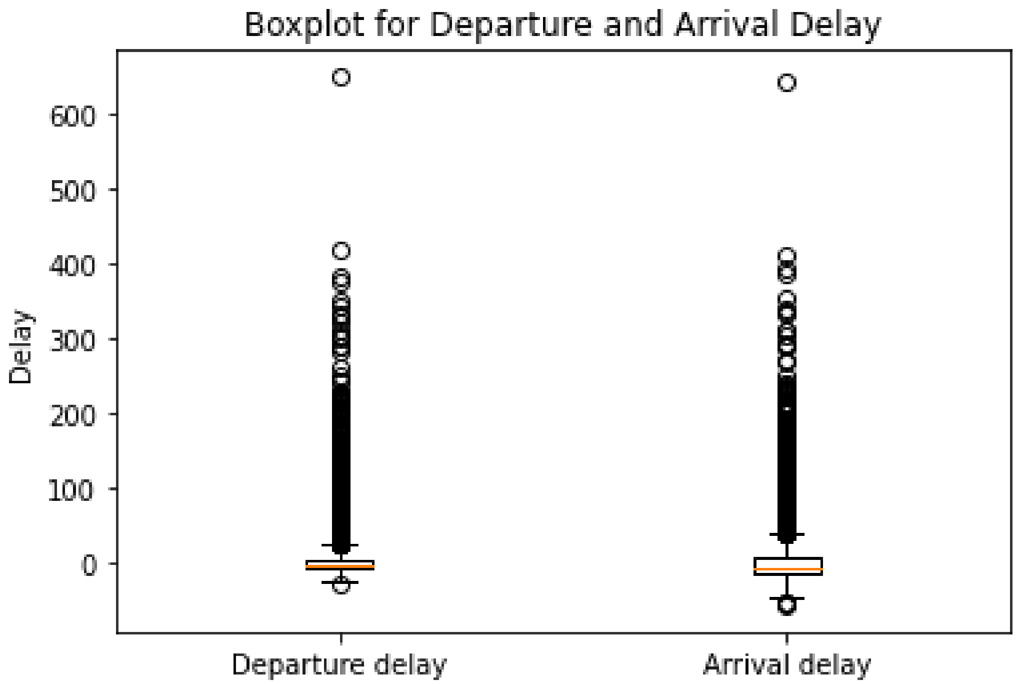
```
0 - total number of missing values in departure delays
0 - total number of missing values in arrival delays
```

4. What is the average and median departure and arrival delay? What do you observe?

```
8.887 - Average departure delay
3.988 - Average arrival delay
-2.000 - Median departure delay
-5.000 - Median arrival delay
```

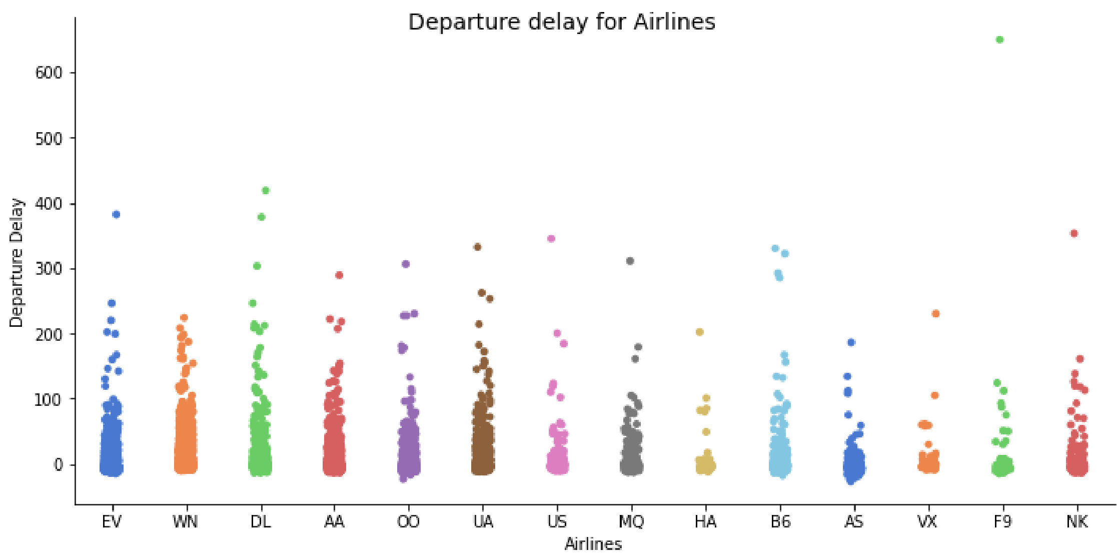
Based on the values above we find that the mean is greater than median for both departure and arrival delay

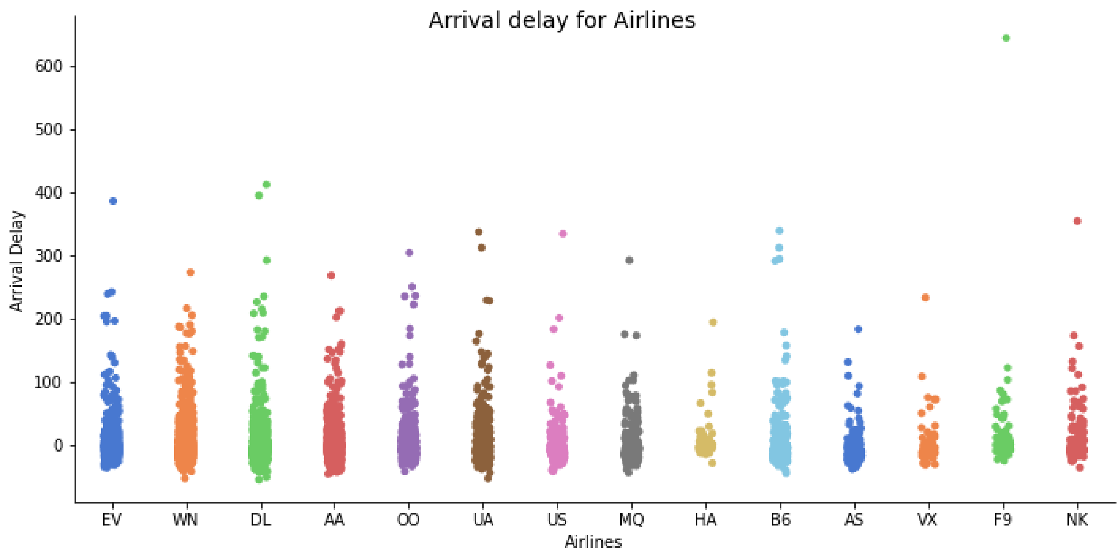
```
Skew DEPARTURE_DELAY: 5.667
Skew ARRIVAL_DELAY:   4.798
```



Observations:
* From the above box plot we can see that there are a lot of outliers and extreme values in the dataset.
* The coefficient of skewness is also significantly higher than zero.
* The distribution is skewed to the right and extremely high values have a significant impact on the mean.

5. Display graphically the departure delays and arrival delays for each airline. What do you notice? Explain






```
Correlation Matrix
```

[38]:

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

Observations:

- We can see that the arrival and departure delay follow the same trend.
- This trend indicate that there might be a strong correlation between the arrival and departure delay.
- From the above correlation matrix we can see that there is no correlation between the distance and delays (0.02 & -0.02).
- However there is a strong positive correlation between the departure and arrival delays (0.93). Hence, delayed flights arrive late.

6. Now calculate the 5 number summary (min, Q1, median, Q3, max) of departure delay for each airline. Arrange it by median delay (descending order). Do the same for arrival delay. Departure delay 5 number summary

[39]:

AIRLINE	Min	Q1	Q3	Max	Median
UA	-12.0	-3.0	14.00	332.0	1.5
WN	-10.0	-3.0	10.00	224.0	0.0
B6	-18.0	-5.0	11.00	330.0	-1.0

VX	-9.0	-4.0	3.25	230.0	-1.5
AA	-14.0	-5.0	7.00	289.0	-2.0
DL	-14.0	-4.0	3.00	419.0	-2.0
NK	-14.0	-6.0	20.00	353.0	-2.0
EV	-15.0	-6.0	4.00	382.0	-3.0
HA	-12.0	-6.0	1.00	202.0	-3.0
MQ	-13.0	-5.0	6.00	311.0	-3.0
OO	-23.0	-7.0	2.00	306.0	-3.0
US	-11.0	-5.0	2.75	345.0	-3.0
AS	-27.0	-8.0	2.00	186.0	-4.0
F9	-15.0	-7.0	4.00	650.0	-4.0

Arrival delay 5 number summary

[40]:

	Min	Q1	Q3	Max	Median
AIRLINE					
F9	-25.0	-9.00	15.00	644.0	1.0
HA	-29.0	-5.00	10.00	194.0	-1.0
NK	-36.0	-10.75	23.00	354.0	-2.0
OO	-42.0	-12.00	8.00	304.0	-3.0
EV	-36.0	-12.00	8.00	386.0	-4.0
US	-42.0	-13.00	11.00	334.0	-4.0
WN	-53.0	-12.00	8.00	273.0	-4.0
B6	-45.0	-15.00	14.00	339.0	-5.0
UA	-53.0	-15.00	10.00	337.0	-5.5
AA	-46.0	-15.00	7.75	268.0	-6.0
AS	-38.0	-14.00	2.00	183.0	-6.0
VX	-32.0	-15.00	5.25	233.0	-6.0
MQ	-44.0	-14.00	8.00	292.0	-7.0
DL	-55.0	-15.00	3.00	412.0	-8.0

7. Which airport has the most averaged departure delay? Give me the top 10 airports. Why do you think the number 1 airport has that much delay?

The airport with the most averaged departure delay is

[41]:

	mean
ORIGIN_AIRPORT	
FAR	161.0

[42]:

	mean
ORIGIN_AIRPORT	
FAR	161.000000
12898	119.000000
BMI	101.333333
ERI	92.000000
MYR	88.000000
14576	88.000000
14696	88.000000

```
10157              87.500000
12992              80.000000
12206              67.500000
```

[43]:
```
      YEAR  MONTH  DAY  DAY_OF_WEEK AIRLINE  FLIGHT_NUMBER TAIL_NUMBER  \
2991  2015      7    6            1      MQ           3195      N658MQ

     ORIGIN_AIRPORT DESTINATION_AIRPORT  SCHEDULED_DEPARTURE  DEPARTURE_TIME  \
2991            FAR                 ORD                 1214          1455.0

      DEPARTURE_DELAY  TAXI_OUT  WHEELS_OFF  SCHEDULED_TIME  ELAPSED_TIME  \
2991            161.0      21.0      1516.0             116         130.0

      AIR_TIME  DISTANCE  WHEELS_ON  TAXI_IN  SCHEDULED_ARRIVAL  ARRIVAL_TIME  \
2991      88.0       557     1644.0     21.0               1410        1705.0

      ARRIVAL_DELAY  DIVERTED  CANCELLED CANCELLATION_REASON  \
2991          175.0         0          0                 NaN

      AIR_SYSTEM_DELAY  SECURITY_DELAY  AIRLINE_DELAY  LATE_AIRCRAFT_DELAY  \
2991             100.0             0.0            0.0                 75.0

      WEATHER_DELAY
2991            0.0
```

Observation:

- Here, we can see that the airport FAR has only one observation in the dataset. Hence, the reason for it being the airport with the maximum average delay.

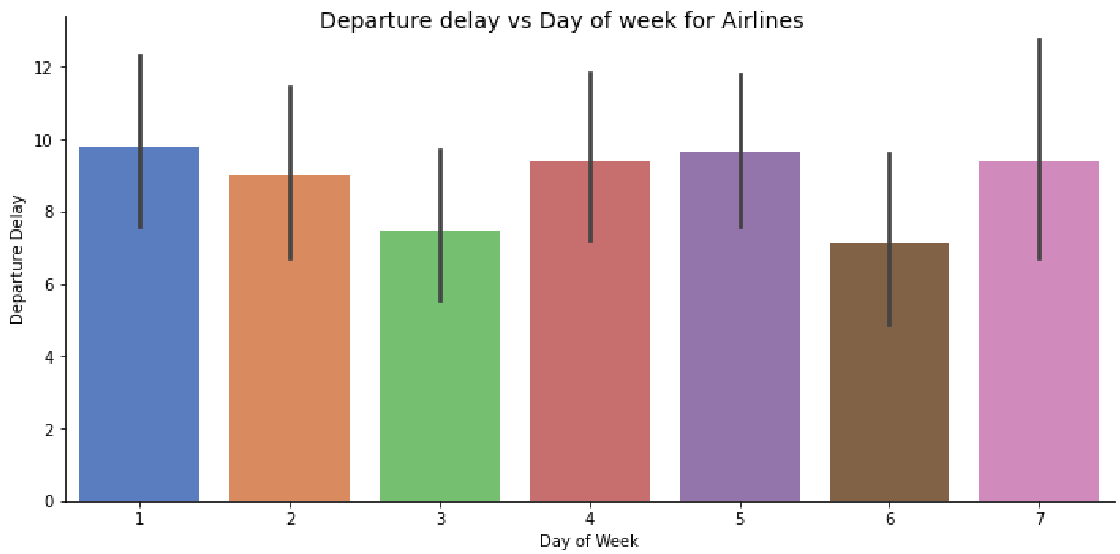
8. Do you expect the departure delay has anything to do with distance of trip? What about arrival delay and distance? Prove your claims.

[44]:
```
                 DISTANCE  DEPARTURE_DELAY  ARRIVAL_DELAY
DISTANCE         1.000000         0.023095      -0.027935
DEPARTURE_DELAY  0.023095         1.000000       0.936069
ARRIVAL_DELAY   -0.027935         0.936069       1.000000
```

Observations:

- The above correlation matrix proves that the distance has nothing to do with the departure and arrival delays.
- There is no correlation between the distance and the departure and arrival delays.

9. What about day of week vs departure delay?



[46]:

	DAY_OF_WEEK	DEPARTURE_DELAY
DAY_OF_WEEK	1.000000	-0.004786
DEPARTURE_DELAY	-0.004786	1.000000

Observations:

- From the above graph we can see that the average departure delay for each day of the week is nearly same.
- The correlation matrix also proves that there is no correlation between the departure delay and day of the week.

10. If there is a departure delay (i.e. positive values for departure delay), does distance have anything to do with arrival delay? Explain. (My experience has been that longer distance flights can make up more time.)

[47]:

	YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER	\
3	2015	1	1	4	EV	4936	N738EV	
7	2015	1	1	4	OO	5354	N472CA	
12	2015	1	1	4	US	705	N567UW	
14	2015	1	1	4	UA	1468	N68807	
15	2015	1	1	4	WN	688	N242WN	

	ORIGIN_AIRPORT	DESTINATION_AIRPORT	SCHEDULED_DEPARTURE	DEPARTURE_TIME	\
3	MSP	IAD	900	901.0	
7	ORD	MBS	1317	1349.0	
12	CLT	LAS	1800	1813.0	
14	IAH	SEA	1912	1924.0	
15	MKE	STL	1945	1951.0	

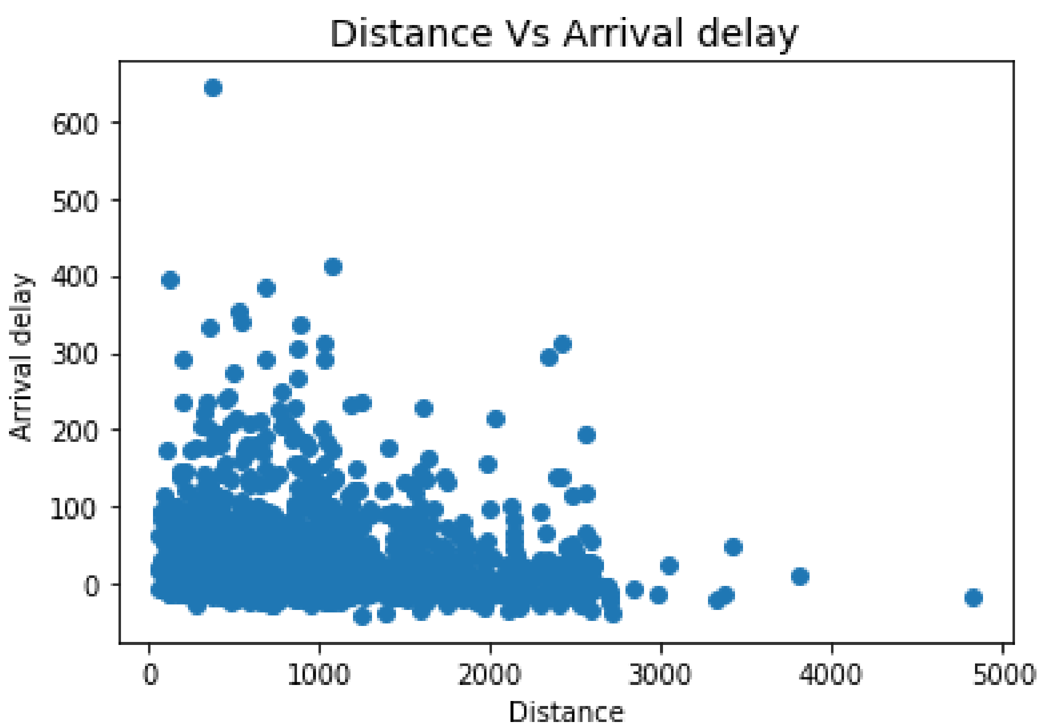
DEPARTURE_DELAY	TAXI_OUT	WHEELS_OFF	SCHEDULED_TIME	ELAPSED_TIME	\

3	1.0	56.0	957.0	148	159.0
7	32.0	27.0	1416.0	66	70.0
12	13.0	15.0	1828.0	295	289.0
14	12.0	9.0	1933.0	289	259.0
15	6.0	9.0	2000.0	75	69.0

	AIR_TIME	DISTANCE	WHEELS_ON	TAXI_IN	SCHEDULED_ARRIVAL	ARRIVAL_TIME	\
3	100.0	908	1237.0	3.0	1228	1240.0	
7	39.0	222	1555.0	4.0	1523	1559.0	
12	266.0	1916	1954.0	8.0	1955	2002.0	
14	245.0	1874	2138.0	5.0	2201	2143.0	
15	55.0	317	2055.0	5.0	2100	2100.0	

	ARRIVAL_DELAY	DIVERTED	CANCELLED	CANCELLATION_REASON	AIR_SYSTEM_DELAY	\
3	12.0	0	0	NaN	NaN	
7	36.0	0	0	NaN	4.0	
12	7.0	0	0	NaN	NaN	
14	-18.0	0	0	NaN	NaN	
15	0.0	0	0	NaN	NaN	

	SECURITY_DELAY	AIRLINE_DELAY	LATE_AIRCRAFT_DELAY	WEATHER_DELAY
3	NaN	NaN	NaN	NaN
7	0.0	11.0	21.0	0.0
12	NaN	NaN	NaN	NaN
14	NaN	NaN	NaN	NaN
15	NaN	NaN	NaN	NaN



[49]:

	DISTANCE	ARRIVAL_DELAY
DISTANCE	1.000000	-0.094924
ARRIVAL_DELAY	-0.094924	1.000000

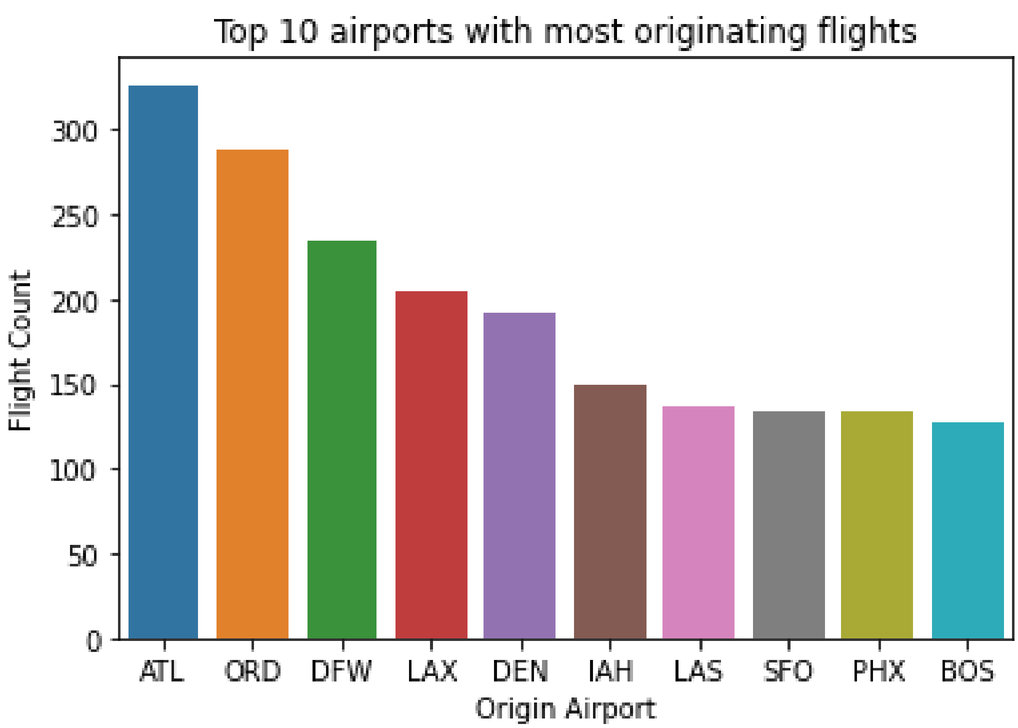
Observations:

- Distance has nothing to do with arrival delay.
- The scatter plot and the correlation matrix suggests the same. All the long distance flight may or may not be able to makeup the lost time.

11. Come up with two interesting questions that you want to answer, then explore it in using this data set. Use any numerical or graphical methods to support your answers. (preferably both). Q1. From which airport does most flights originate?

[50]:

	ORIGIN_AIRPORT	AIRLINE
127	ATL	326
292	ORD	287
180	DFW	235
249	LAX	205
179	DEN	191

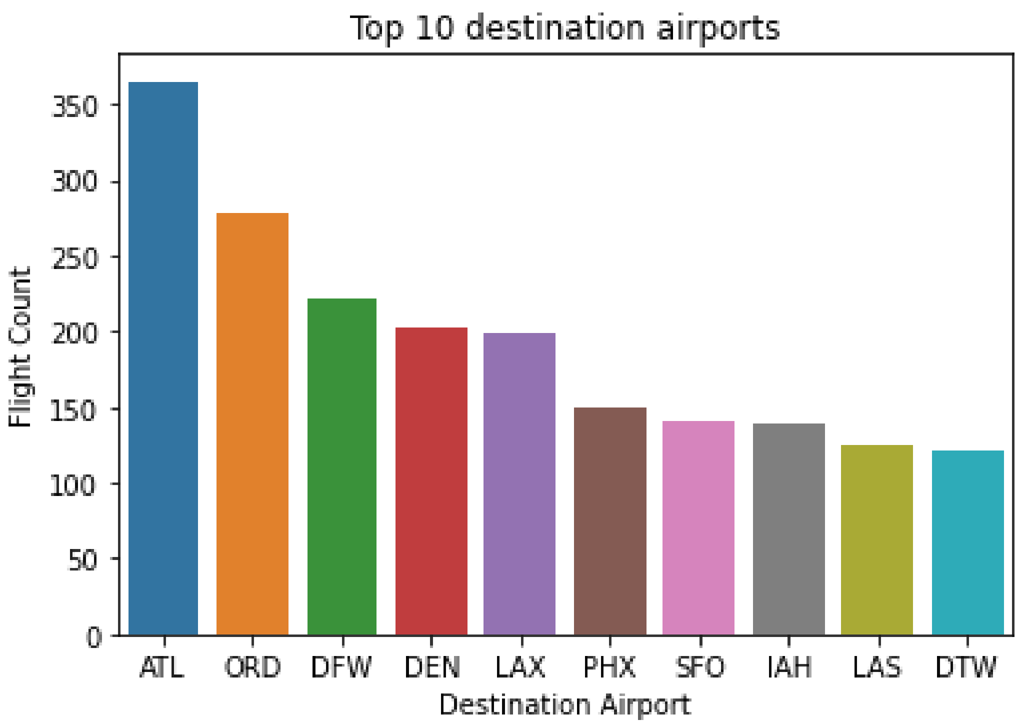


Answer: From the analysis above, we can see that the most flights originate from ATL airport

Q2. Which is the most visited city?

[52]:

	DESTINATION_AIRPORT	AIRLINE
122	ATL	365
297	ORD	278
177	DFW	221
176	DEN	203
248	LAX	199

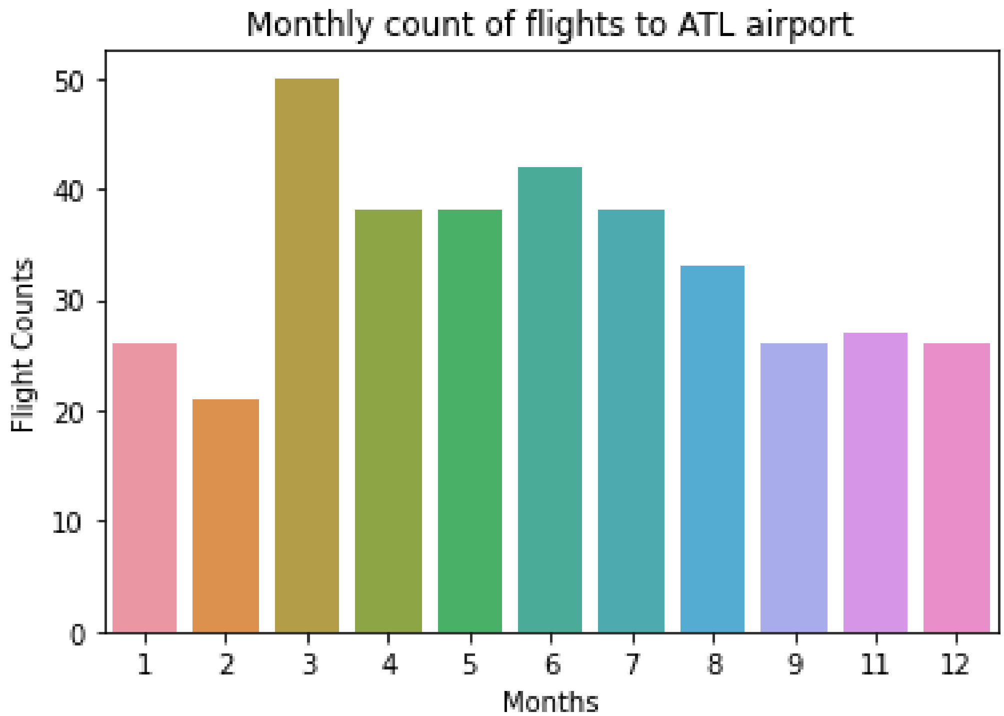


Answer: From the above analysis, we can deduce that ATL is the most visited city. If we combine the results of both Q1 and Q2 we can say ATL is the busiest airport.

Q3. Since ATL is the most visited destination. In which part of the year do people visit it most?

[54]:

	MONTH	AIRLINE
0	1	26
1	2	21
2	3	50
3	4	38
4	5	38
5	6	42
6	7	38
7	8	33
8	9	26
9	11	27
10	12	26



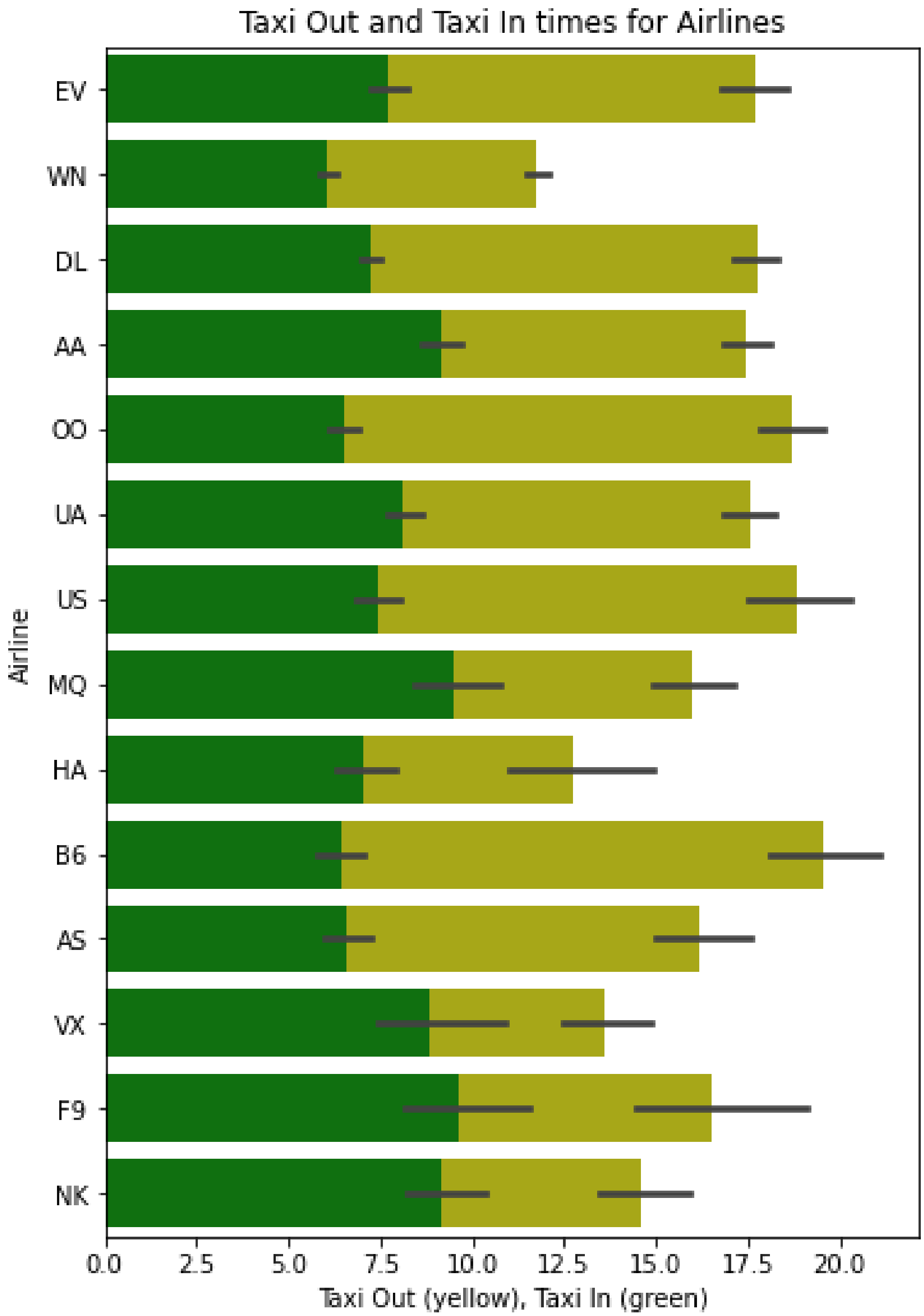
Answer: There are maximum flights in the month of March and in the summer months the count of flights is more. Hence people visit ATL mostly during Spring and Summer.

[56]:

```
   YEAR  MONTH  DAY  DAY_OF_WEEK AIRLINE  FLIGHT_NUMBER TAIL_NUMBER  \
0  2015      1    1            4      EV           4160      N11150
2  2015      1    1            4      WN            119      N271LV
3  2015      1    1            4      EV           4936      N738EV
4  2015      1    1            4      DL           2319      N960DL
5  2015      1    1            4      DL           1806      N594NW

  ORIGIN_AIRPORT DESTINATION_AIRPORT  SCHEDULED_DEPARTURE  DEPARTURE_TIME  \
0            JAX                 EWR                  540           531.0
2            RSW                 ATL                  800           754.0
3            MSP                 IAD                  900           901.0
4            LGA                 MSP                 1010          1010.0
5            LAX                 DTW                 1115          1113.0

   DEPARTURE_DELAY  TAXI_OUT  WHEELS_OFF  SCHEDULED_TIME  ELAPSED_TIME  \
0             -9.0       9.0       540.0             137         132.0
2             -6.0      11.0       805.0             105         100.0
3              1.0      56.0       957.0             148         159.0
4              0.0      22.0      1032.0             200         195.0
5             -2.0      15.0      1128.0             266         248.0

   AIR_TIME  DISTANCE  WHEELS_ON  TAXI_IN  SCHEDULED_ARRIVAL  ARRIVAL_TIME  \
```

```
0     109.0       820      729.0      14.0                   757         743.0
2      84.0       515      929.0       5.0                   945         934.0
3     100.0       908     1237.0       3.0                  1228        1240.0
4     171.0      1020     1223.0       2.0                  1230        1225.0
5     226.0      1979     1814.0       7.0                  1841        1821.0

   ARRIVAL_DELAY  DIVERTED  CANCELLED CANCELLATION_REASON  AIR_SYSTEM_DELAY  \
0          -14.0         0          0                 NaN               NaN
2          -11.0         0          0                 NaN               NaN
3           12.0         0          0                 NaN               NaN
4           -5.0         0          0                 NaN               NaN
5          -20.0         0          0                 NaN               NaN

   SECURITY_DELAY  AIRLINE_DELAY  LATE_AIRCRAFT_DELAY  WEATHER_DELAY
0             NaN            NaN                  NaN            NaN
2             NaN            NaN                  NaN            NaN
3             NaN            NaN                  NaN            NaN
4             NaN            NaN                  NaN            NaN
5             NaN            NaN                  NaN            NaN
```

Q4. Does all airlines have same taxi in and taxi out times?



Answer: Taxi out and Taxi in times for all the airlines is different. However, for all the airlines taxi in times is significantly less than taxi out.

2.0.2 Part 2: Regression Analysis

Subpart 1

1. Your response is ARRIVAL_DELAY. First, remove all the missing data in the WEATHER_DELAY column. Once you do this, there shouldn't be anymore missing values in the data set(except for the cancellation reason feature). Check that.

```
4641 - Total missing values
```

[59]:
```
YEAR                      0
MONTH                     0
DAY                       0
DAY_OF_WEEK               0
AIRLINE                   0
FLIGHT_NUMBER             0
TAIL_NUMBER               0
ORIGIN_AIRPORT            0
DESTINATION_AIRPORT       0
SCHEDULED_DEPARTURE       0
DEPARTURE_TIME            0
DEPARTURE_DELAY           0
TAXI_OUT                  0
WHEELS_OFF                0
SCHEDULED_TIME            0
ELAPSED_TIME              0
AIR_TIME                  0
DISTANCE                  0
WHEELS_ON                 0
TAXI_IN                   0
SCHEDULED_ARRIVAL         0
ARRIVAL_TIME              0
ARRIVAL_DELAY             0
DIVERTED                  0
CANCELLED                 0
CANCELLATION_REASON    1072
AIR_SYSTEM_DELAY          0
SECURITY_DELAY            0
AIRLINE_DELAY             0
LATE_AIRCRAFT_DELAY       0
WEATHER_DELAY             0
dtype: int64
```

2. Build a regression model using all the observations, and the following predictors: [LATE_AIRCRAFT_DELAY, AIRLINE_DELAY, AIR_SYSTEM_DELAY, WEATHER_DELAY, DAY_OF_WEEK, DEPARTURE_TIME, DEPARTURE_DELAY, DISTANCE, AIRLINE] a total of 9 predictors. Notice the AIRLINE variable is a categorical variable.

[60]:
```
    YEAR  MONTH  DAY  DAY_OF_WEEK AIRLINE  FLIGHT_NUMBER TAIL_NUMBER  \
7   2015      1    1            4      OO           5354      N472CA
9   2015      1    1            4      UA           1062      N73291
19  2015      1    2            5      US           2065      N534UW
21  2015      1    2            5      OO           5211      N943SW
22  2015      1    2            5      HA            335      N477HA

   ORIGIN_AIRPORT DESTINATION_AIRPORT  SCHEDULED_DEPARTURE  DEPARTURE_TIME  \
7             ORD                 MBS                 1317          1349.0
9             DCA                 DEN                 1603          1603.0
19            CLT                 IAH                 1120          1128.0
21            IDA                 DEN                 1338          1428.0
22            OGG                 HNL                 1503          1644.0

    DEPARTURE_DELAY  TAXI_OUT  WHEELS_OFF  SCHEDULED_TIME  ELAPSED_TIME  \
7              32.0      27.0      1416.0              66          70.0
9               0.0      12.0      1615.0             249         272.0
19              8.0      11.0      1139.0             163         176.0
21             50.0      31.0      1459.0              91         122.0
22            101.0      10.0      1654.0              37          50.0

    AIR_TIME  DISTANCE  WHEELS_ON  TAXI_IN  SCHEDULED_ARRIVAL  ARRIVAL_TIME  \
7       39.0       222     1555.0      4.0               1523        1559.0
9      248.0      1476     1823.0     12.0               1812        1835.0
19     154.0       912     1313.0     11.0               1303        1324.0
21      64.0       458     1603.0     27.0               1509        1630.0
22      23.0       100     1717.0     17.0               1540        1734.0

    ARRIVAL_DELAY  DIVERTED  CANCELLED CANCELLATION_REASON  AIR_SYSTEM_DELAY  \
7            36.0         0          0                 NaN               4.0
9            23.0         0          0                 NaN              23.0
19           21.0         0          0                 NaN              13.0
21           81.0         0          0                 NaN              31.0
22          114.0         0          0                 NaN               0.0

    SECURITY_DELAY  AIRLINE_DELAY  LATE_AIRCRAFT_DELAY  WEATHER_DELAY  AS  B6  \
7              0.0           11.0                 21.0            0.0   0   0
9              0.0            0.0                  0.0            0.0   0   0
19             0.0            8.0                  0.0            0.0   0   0
21             0.0            0.0                 50.0            0.0   0   0
22             0.0           25.0                 89.0            0.0   0   0

    DL  EV  F9  HA  MQ  NK  OO  UA  US  VX  WN
7    0   0   0   0   0   0   1   0   0   0   0
9    0   0   0   0   0   0   0   1   0   0   0
19   0   0   0   0   0   0   0   0   1   0   0
21   0   0   0   0   0   0   1   0   0   0   0
```

```
22   0   0   0   1   0   0   0   0   0   0   0
```

[61]:

```
    YEAR  MONTH  DAY  DAY_OF_WEEK AIRLINE  FLIGHT_NUMBER TAIL_NUMBER  \
7   2015      1    1            4      OO           5354      N472CA
9   2015      1    1            4      UA           1062      N73291
19  2015      1    2            5      US           2065      N534UW
21  2015      1    2            5      OO           5211      N943SW
22  2015      1    2            5      HA            335      N477HA

   ORIGIN_AIRPORT DESTINATION_AIRPORT  SCHEDULED_DEPARTURE  DEPARTURE_TIME  \
7             ORD                 MBS                 1317          1349.0
9             DCA                 DEN                 1603          1603.0
19            CLT                 IAH                 1120          1128.0
21            IDA                 DEN                 1338          1428.0
22            OGG                 HNL                 1503          1644.0

    DEPARTURE_DELAY  TAXI_OUT  WHEELS_OFF  SCHEDULED_TIME  ELAPSED_TIME  \
7              32.0      27.0      1416.0              66          70.0
9               0.0      12.0      1615.0             249         272.0
19              8.0      11.0      1139.0             163         176.0
21             50.0      31.0      1459.0              91         122.0
22            101.0      10.0      1654.0              37          50.0

    AIR_TIME  DISTANCE  WHEELS_ON  TAXI_IN  SCHEDULED_ARRIVAL  ARRIVAL_TIME  \
7       39.0       222     1555.0      4.0               1523        1559.0
9      248.0      1476     1823.0     12.0               1812        1835.0
19     154.0       912     1313.0     11.0               1303        1324.0
21      64.0       458     1603.0     27.0               1509        1630.0
22      23.0       100     1717.0     17.0               1540        1734.0

    ARRIVAL_DELAY  DIVERTED  CANCELLED CANCELLATION_REASON  AIR_SYSTEM_DELAY  \
7            36.0         0          0                 NaN               4.0
9            23.0         0          0                 NaN              23.0
19           21.0         0          0                 NaN              13.0
21           81.0         0          0                 NaN              31.0
22          114.0         0          0                 NaN               0.0

    SECURITY_DELAY  AIRLINE_DELAY  LATE_AIRCRAFT_DELAY  WEATHER_DELAY  AS  B6  \
7              0.0           11.0                 21.0            0.0   0   0
9              0.0            0.0                  0.0            0.0   0   0
19             0.0            8.0                  0.0            0.0   0   0
21             0.0            0.0                 50.0            0.0   0   0
22             0.0           25.0                 89.0            0.0   0   0

    DL  EV  F9  HA  MQ  NK  OO  UA  US  VX  WN  DAY_2  DAY_3  DAY_4  DAY_5  \
7    0   0   0   0   0   0   1   0   0   0   0      0      0      1      0
9    0   0   0   0   0   0   0   1   0   0   0      0      0      1      0
```

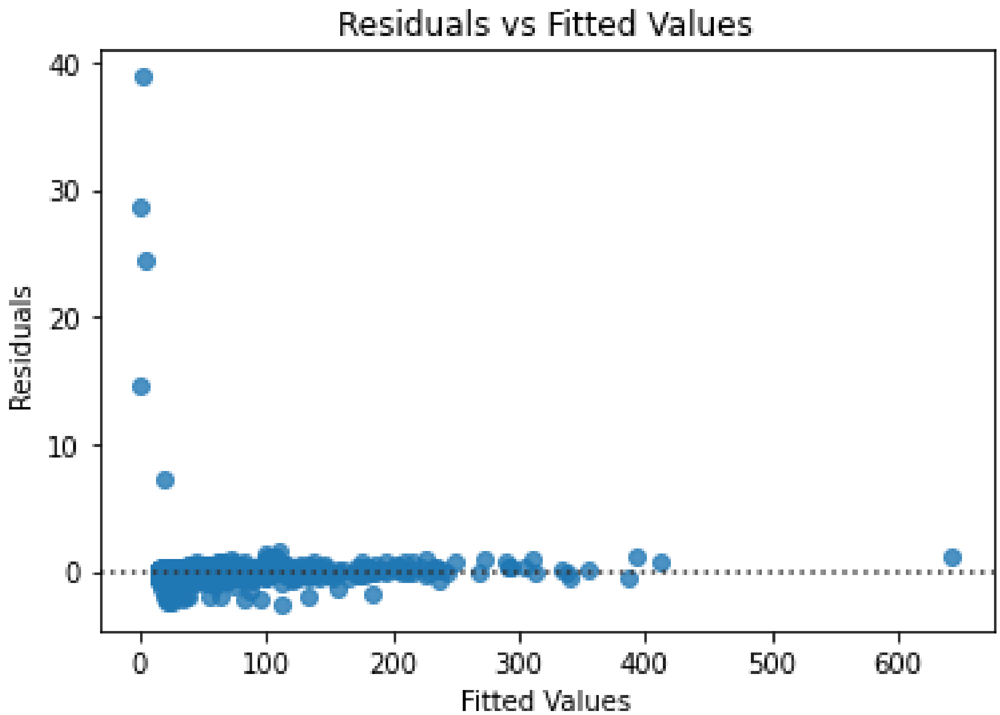
```
19    0    0    0    0    0    0    0    0    1    0    0        0        0        0        1
21    0    0    0    0    0    0    1    0    0    0    0        0        0        0        1
22    0    0    0    1    0    0    0    0    0    0    0        0        0        0        1

    DAY_6  DAY_7
7       0      0
9       0      0
19      0      0
21      0      0
22      0      0
```

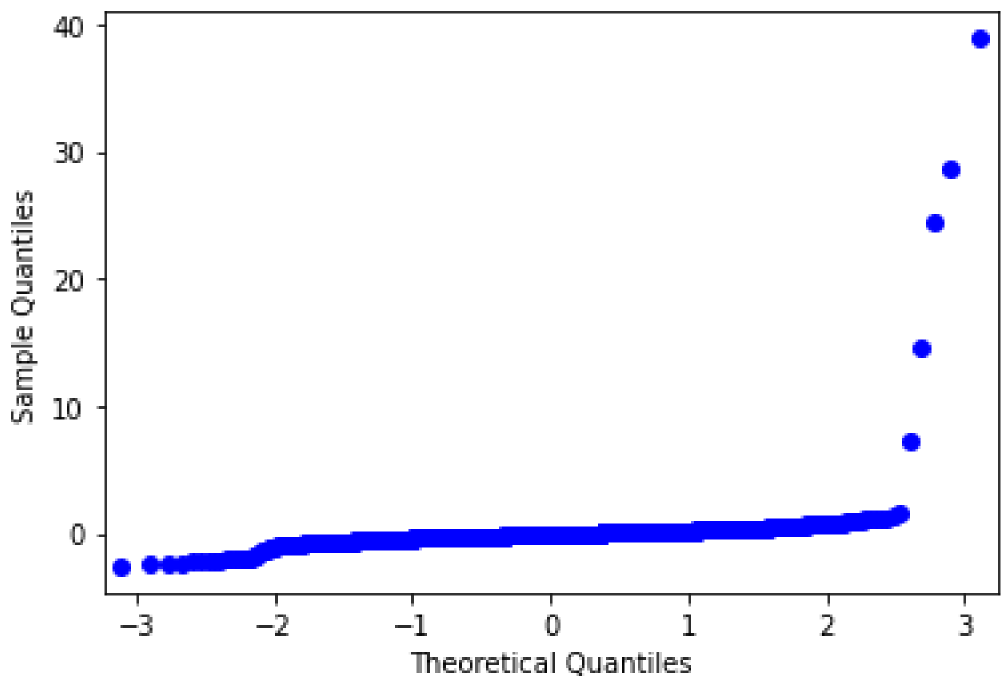
[62]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:          ARRIVAL_DELAY   R-squared:                       0.999
Model:                            OLS   Adj. R-squared:                  0.999
Method:                 Least Squares   F-statistic:                 4.273e+04
Date:                Sun, 24 Oct 2021   Prob (F-statistic):               0.00
Time:                        15:53:33   Log-Likelihood:                -2140.7
No. Observations:                1072   AIC:                             4335.
Df Residuals:                    1045   BIC:                             4470.
Df Model:                          26
Covariance Type:            nonrobust
================================================================================
=======
                          coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
-------
const                   0.5743      0.300      1.916      0.056      -0.014
1.163
LATE_AIRCRAFT_DELAY     0.9814      0.004    251.351      0.000       0.974
0.989
AIRLINE_DELAY           0.9820      0.004    258.523      0.000       0.975
0.989
AIR_SYSTEM_DELAY        0.9853      0.003    311.685      0.000       0.979
0.992
WEATHER_DELAY           0.9846      0.004    239.180      0.000       0.977
0.993
DEPARTURE_TIME         -0.0001      0.000     -0.923      0.356      -0.000
0.000
DEPARTURE_DELAY         0.0158      0.003      4.647      0.000       0.009
0.023
DISTANCE                0.0001      0.000      1.106      0.269   -8.74e-05
0.000
AS                      1.8908      0.434      4.354      0.000       1.039
```

```
2.743
B6                     0.0009      0.277      0.003      0.997      -0.542
0.544
DL                    -0.2385      0.224     -1.062      0.288      -0.679
0.202
EV                    -0.1519      0.241     -0.629      0.529      -0.625
0.322
F9                     0.0010      0.445      0.002      0.998      -0.872
0.874
HA                    -0.1183      0.548     -0.216      0.829      -1.194
0.957
MQ                    -0.1080      0.295     -0.366      0.715      -0.687
0.471
NK                     0.4677      0.333      1.404      0.161      -0.186
1.122
OO                    -0.1077      0.241     -0.448      0.654      -0.580
0.364
UA                    -0.3509      0.233     -1.505      0.133      -0.808
0.107
US                    -0.1699      0.319     -0.533      0.594      -0.796
0.456
VX                    -0.1395      0.546     -0.256      0.798      -1.210
0.931
WN                    -0.1731      0.199     -0.870      0.384      -0.563
0.217
DAY_2                 -0.2517      0.207     -1.214      0.225      -0.659
0.155
DAY_3                  0.0896      0.208      0.431      0.667      -0.319
0.498
DAY_4                 -0.2768      0.197     -1.404      0.161      -0.664
0.110
DAY_5                 -0.2329      0.198     -1.175      0.240      -0.622
0.156
DAY_6                 -0.2907      0.238     -1.222      0.222      -0.757
0.176
DAY_7                 -0.1769      0.206     -0.859      0.390      -0.581
0.227
==============================================================================
Omnibus:                     2216.606   Durbin-Watson:                   2.019
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          4284200.685
Skew:                          16.511   Prob(JB):                         0.00
Kurtosis:                     310.936   Cond. No.                     2.21e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

```
[2] The condition number is large, 2.21e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

3. Perform model diagnostics. What do you observe? Explain.





Observations:

- There are outliers.
- The model does not satisfy the linearity, constant variance and normality.

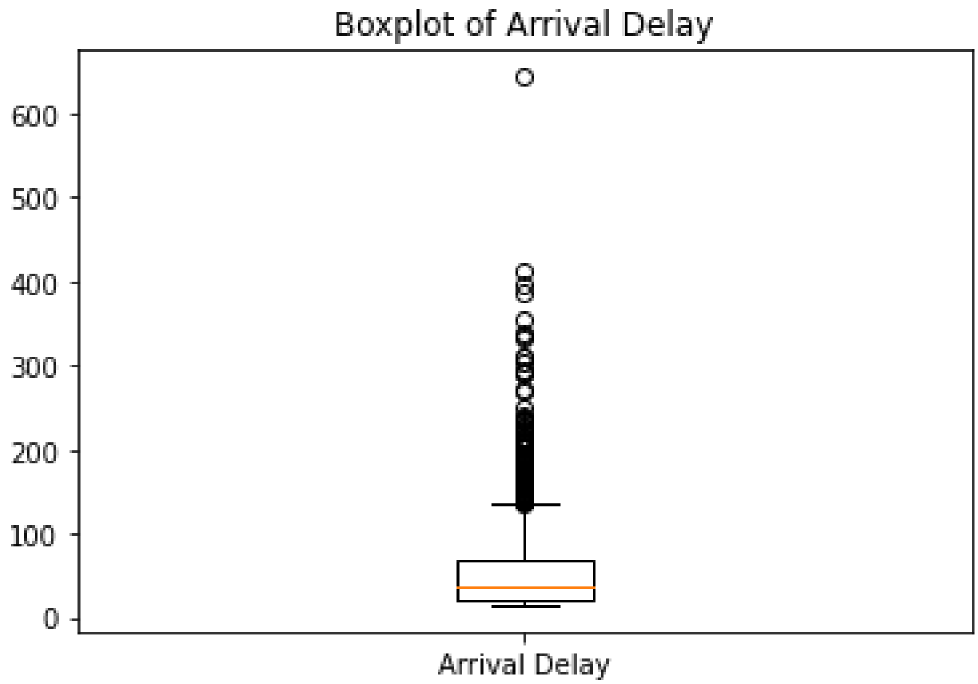
4. Provide interpretations for a few of the coeffcients, and comment on whether they make sense.

Interpretations:

- Every one minute increase in airline delay, results in 0.98 minute increase in arrival (arrival delay).
- There is an impact of 'late aircraft delay', 'air system delay', 'weather delay' and 'departure delay' on aircraft arrivals (arrival delay).
- There is no effect of day of the week on arrivals. This is evident from the high p-values.
- For every one minute increase in departure delay, arrival delay increases by 0.018 minutes.

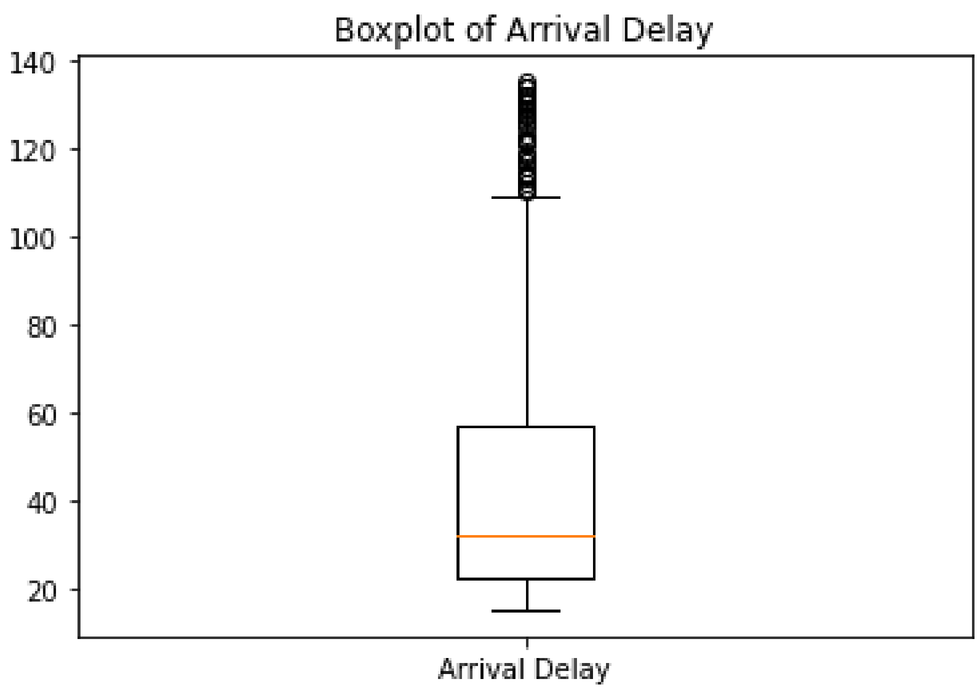
Subpart 2 If you have done the above steps correctly, you will notice a lot of things "doesn't seem right". We will try to fix a couple of these things here.

1. Removing outliers: _first is to remove outliers. Using the boxplot method, remove the outliers in the ARRIVAL_DELAY variable.


Boxplot of Arrival Delay
600
500
400
300
200
100
0
Arrival Delay

```
45.25
```

[68]: 986


Boxplot of Arrival Delay
140
120
100
80
60
40
20
Arrival Delay

2. Refit the linear regression model, but now with log(ARRIVAL_DELAY) as your response. Also, remove the nonsignificant predictors from the previous model (with p-values larger than 0.05) and the AIRLINE variable. (Remember that when removing nonsignificant predictors one can only eliminate one variable per step.)

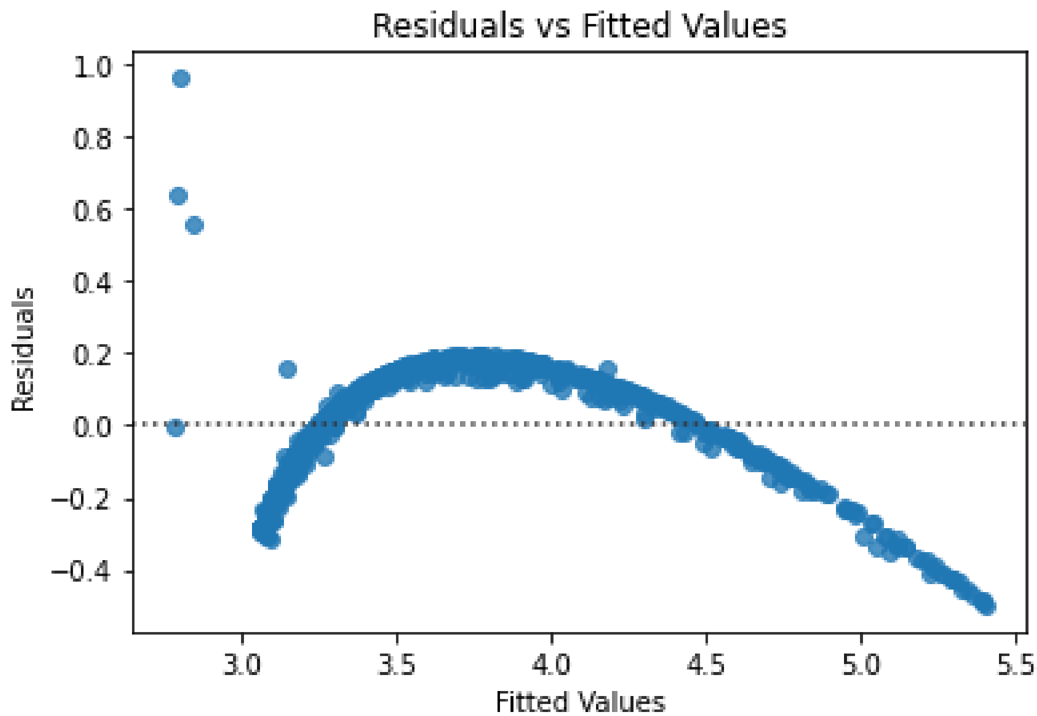
[70]:
```
    YEAR  MONTH  DAY  DAY_OF_WEEK AIRLINE  FLIGHT_NUMBER TAIL_NUMBER  \
7   2015      1    1            4      OO           5354      N472CA
9   2015      1    1            4      UA           1062      N73291
19  2015      1    2            5      US           2065      N534UW
21  2015      1    2            5      OO           5211      N943SW
22  2015      1    2            5      HA            335      N477HA

   ORIGIN_AIRPORT DESTINATION_AIRPORT  SCHEDULED_DEPARTURE  DEPARTURE_TIME  \
7             ORD                 MBS                 1317          1349.0
9             DCA                 DEN                 1603          1603.0
19            CLT                 IAH                 1120          1128.0
21            IDA                 DEN                 1338          1428.0
22            OGG                 HNL                 1503          1644.0

    DEPARTURE_DELAY  TAXI_OUT  WHEELS_OFF  SCHEDULED_TIME  ELAPSED_TIME  \
7              32.0      27.0      1416.0              66          70.0
9               0.0      12.0      1615.0             249         272.0
19              8.0      11.0      1139.0             163         176.0
21             50.0      31.0      1459.0              91         122.0
22            101.0      10.0      1654.0              37          50.0

    AIR_TIME  DISTANCE  WHEELS_ON  TAXI_IN  SCHEDULED_ARRIVAL  ARRIVAL_TIME  \
7       39.0       222     1555.0      4.0               1523        1559.0
9      248.0      1476     1823.0     12.0               1812        1835.0
19     154.0       912     1313.0     11.0               1303        1324.0
21      64.0       458     1603.0     27.0               1509        1630.0
22      23.0       100     1717.0     17.0               1540        1734.0

    ARRIVAL_DELAY  DIVERTED  CANCELLED CANCELLATION_REASON  AIR_SYSTEM_DELAY  \
7            36.0         0          0                 NaN               4.0
9            23.0         0          0                 NaN              23.0
19           21.0         0          0                 NaN              13.0
21           81.0         0          0                 NaN              31.0
22          114.0         0          0                 NaN               0.0

    SECURITY_DELAY  AIRLINE_DELAY  LATE_AIRCRAFT_DELAY  WEATHER_DELAY  AS  B6  \
7              0.0           11.0                 21.0            0.0   0   0
9              0.0            0.0                  0.0            0.0   0   0
19             0.0            8.0                  0.0            0.0   0   0
21             0.0            0.0                 50.0            0.0   0   0
```

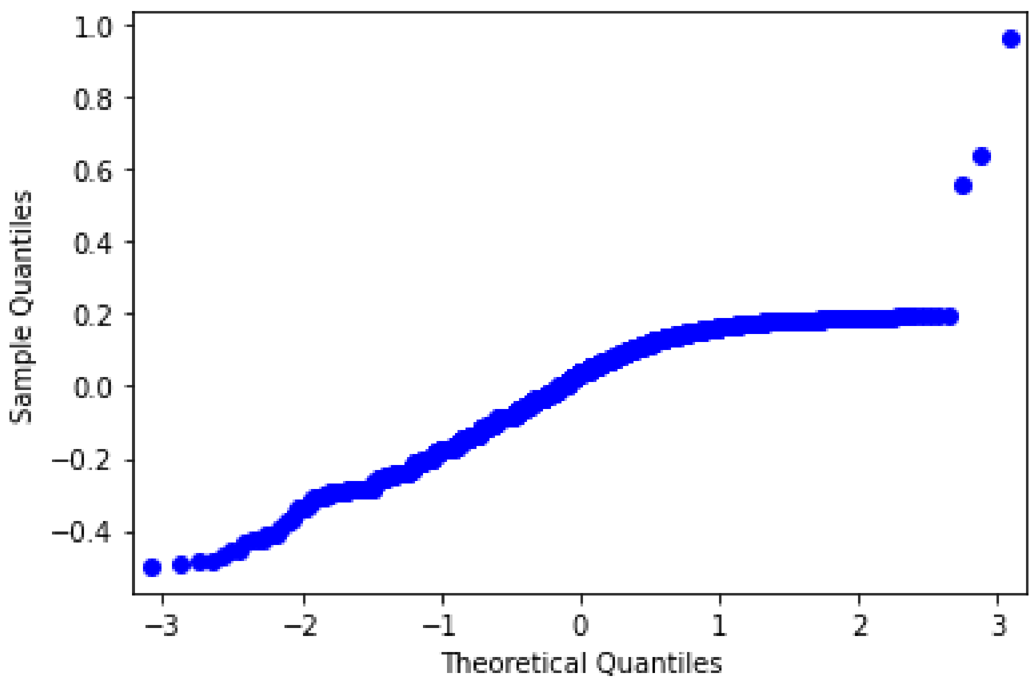
```
22             0.0           25.0                 89.0              0.0   0   0

    DL  EV  F9  HA  MQ  NK  OO  UA  US  VX  WN  DAY_2  DAY_3  DAY_4  DAY_5  \
7    0   0   0   0   0   0   1   0   0   0   0      0      0      1      0
9    0   0   0   0   0   0   0   1   0   0   0      0      0      1      0
19   0   0   0   0   0   0   0   0   1   0   0      0      0      0      1
21   0   0   0   0   0   0   1   0   0   0   0      0      0      0      1
22   0   0   0   1   0   0   0   0   0   0   0      0      0      0      1

    DAY_6  DAY_7  LOG_ARRIVAL_DELAY
7       0      0           3.610918
9       0      0           3.178054
19      0      0           3.091042
21      0      0           4.406719
22      0      0           4.744932
```

[71]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:      LOG_ARRIVAL_DELAY   R-squared:                       0.920
Model:                            OLS   Adj. R-squared:                  0.920
Method:                 Least Squares   F-statistic:                     2269.
Date:                Sun, 24 Oct 2021   Prob (F-statistic):               0.00
Time:                        15:54:58   Log-Likelihood:                 391.96
No. Observations:                 986   AIC:                            -771.9
Df Residuals:                     980   BIC:                            -742.6
Df Model:                           5
Covariance Type:            nonrobust
================================================================================
=======
                          coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
-------
const                   2.7613      0.010    287.307      0.000       2.742
2.780
LATE_AIRCRAFT_DELAY     0.0186      0.000     41.198      0.000       0.018
0.020
AIRLINE_DELAY           0.0188      0.000     40.344      0.000       0.018
0.020
AIR_SYSTEM_DELAY        0.0198      0.000     60.354      0.000       0.019
0.020
WEATHER_DELAY           0.0190      0.001     28.730      0.000       0.018
0.020
DEPARTURE_DELAY         0.0008      0.000      2.283      0.023       0.000
0.002
```

```
==============================================================================
Omnibus:                       37.990   Durbin-Watson:                   1.914
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               56.372
Skew:                          -0.344   Prob(JB):                     5.74e-13
Kurtosis:                       3.948   Cond. No.                         108.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

3. Perform model diagnostics. Did anything improve?





Observations:

- The model needs improvement.
- The model does not satisfy the constraints of linearity, constant variance and normality.

4. Provide interpretations to a few of the coeffcients. Do you think they make sense?

Interpretations:

- Weather delay has an impact on arrival delays. For every one minute increase in weather delay there is an increase of 0.0190 minutes in arrival delay.
- For every one minute increase in air system delay, there is an increase of 0.0198 minutes in arrival delay.

5. Obviously there's still a lot that needs to be done. Provide a few suggestions on how we can further improve the model fit (you don't need to implement them).

Suggestions:

- We can add interaction among the independent variables in the model.
- Using Tukey;s ladder transformation, we may increase or decrease the power of independent variables and use them in the model.