Final Project – Milestone 3 – CAP 5610

Team 5

1. Data Description

Our dataset regards arrival times, cancellations and delays for US flights in 2015. This dataset contains information on 5,819,079 flights and includes information on the length of delay (if present, for arrival and departure) as well as the type of delay (weather, security, etc.), the airline of the flight, the destination airport, and the airport of origin. Data on the date and time of the flight as well as the day of the week is also included.

2. Restated Research Question

After the Exploratory data analysis, we decided to trimmed down our research questions as follow:

> Is it possible that we can predict whether a flight will be delayed or canceled before it comes up on departure boards?
>
> Are there certain key predictors that can help you avoid picking a flight that will be delayed or canceled?

3. Exploratory Data Analysis

a. Getting the dataset

First, we import the three separate datasets. Our primary sources for analysis will be the flights data, while airports and airlines serve as data dictionaries of sorts to understand the flight data

```
import numpy as np
import pandas as pd
filename = pd.read_csv(r"C:\Users\Test\Downloads\flights.csv")
filename2 = pd.read_csv(r"C:\Users\Test\Downloads\airlines.csv")
filename3 = pd.read_csv(r"C:\Users\Test\Downloads\airports.csv")
```

Next we look at the columns of the dataframe for flights, which shows that we have 31 columns that can be used as features for analysis

Out[12]:

	YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT	SCHEDULED_DEPARTURE	.
0	2015	1	1	4	AS	98	N407AS	ANC	SEA	5	.
1	2015	1	1	4	AA	2336	N3KUAA	LAX	PBI	10	.
2	2015	1	1	4	US	840	N171US	SFO	CLT	20	.
3	2015	1	1	4	AA	258	N3HYAA	LAX	MIA	20	.
4	2015	1	1	4	AS	135	N527AS	SEA	ANC	25	.
...	...	...	...	...	...	...	...	...	...	...	.
195	2015	1	1	4	UA	1224	N87531	SFO	LAX	600	.
196	2015	1	1	4	UA	1296	N37471	SAT	LAX	600	.
197	2015	1	1	4	UA	1431	N36207	BOS	LAX	600	.
198	2015	1	1	4	UA	1637	N33294	SEA	EWR	600	.
199	2015	1	1	4	UA	1735	N66814	LAX	ORD	600	.

200 rows × 31 columns

Our dates appear to be broken up by year, month, and day so we may need to combine these columns down the line further into our analysis. We also have the airport codes and airline codes that match up with the definitions in our other two data sets.Our most important columns will likely be arrival time and arrival delay as well as canceled. Notably, there are several columns for delays as well that specify the reason for the delays, which could be relevant to making predictions for future flights (i.e. if a flight was canceled each time the past 5 years on April 10th because of rain, our algorithm may predict cancellation of this flight next year because of this factor amongst other factors).

```
In [13]: dfflights.shape

Out[13]: (5819079, 31)
```

Looking at our data shape, we have 5,819,079 flights to work with, which gives us a robust set of data for analysis. We also count the values of the airline column below:

In [18]:
```
dfflights.value_counts('AIRLINE')

```

Out[18]:
```
AIRLINE
WN    1261855
DL     875881
AA     725984
OO     588353
EV     571977
UA     515723
MQ     294632
B6     267048
US     198715
AS     172521
NK     117379
F9      90836
HA      76272
VX      61903
dtype: int64
```

:
```
dfairlines.head(100)
```

:

	IATA_CODE	AIRLINE
0	UA	United Air Lines Inc.
1	AA	American Airlines Inc.
2	US	US Airways Inc.
3	F9	Frontier Airlines Inc.
4	B6	JetBlue Airways
5	OO	Skywest Airlines Inc.
6	AS	Alaska Airlines Inc.
7	NK	Spirit Air Lines
8	WN	Southwest Airlines Co.
9	DL	Delta Air Lines Inc.
10	EV	Atlantic Southeast Airlines
11	HA	Hawaiian Airlines Inc.
12	MQ	American Eagle Airlines Inc.
13	VX	Virgin America

Here we see WN (which in our data for airlines is Southwest) has the most flight data compared to other airlines, followed by DL (Delta Airlines).

```
In [20]:  1 dfflights.value_counts('ORIGIN_AIRPORT')

Out[20]: ORIGIN_AIRPORT
         ATL      346836
         ORD      285884
         DFW      239551
         DEN      196055
         LAX      194673
                   ...
         13964         1
         14025         1
         14222         1
         15497         1
         12265         1
         Length: 930, dtype: int64
```

Doing the same with regard to airports, we can see that most flights originated from ATL. We also see that we have some curious numeric values that don't match our airport data near the bottom, which we may have to drop or find a source for.

```
In [22]:  1 dfflights.value_counts('DESTINATION_AIRPORT')

Out[22]: DESTINATION_AIRPORT
         ATL      346904
         ORD      285906
         DFW      239582
         DEN      196010
         LAX      194696
                   ...
         12265         1
         14025         1
         13459         1
         15497         1
         13964         1
         Length: 930, dtype: int64
```

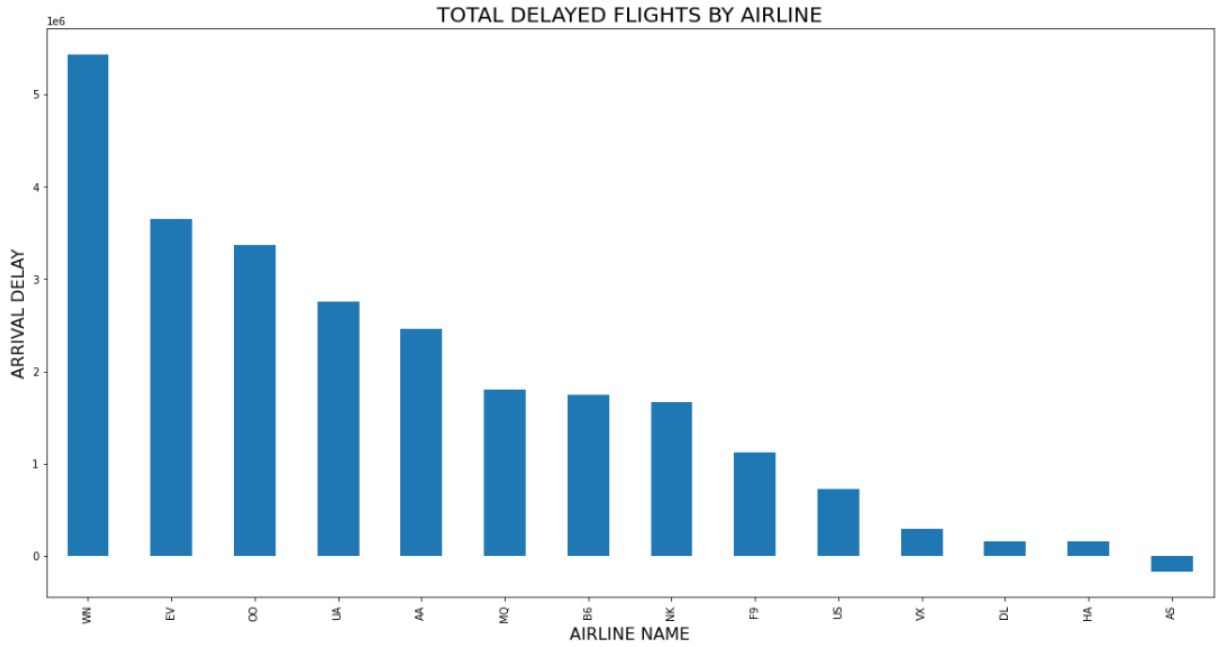
We also see that our destination airports line up exactly with the origin airports, meaning these airports overall are the ones with the most traffic.

Using dfflights.describe, we see the average delay for each factor:

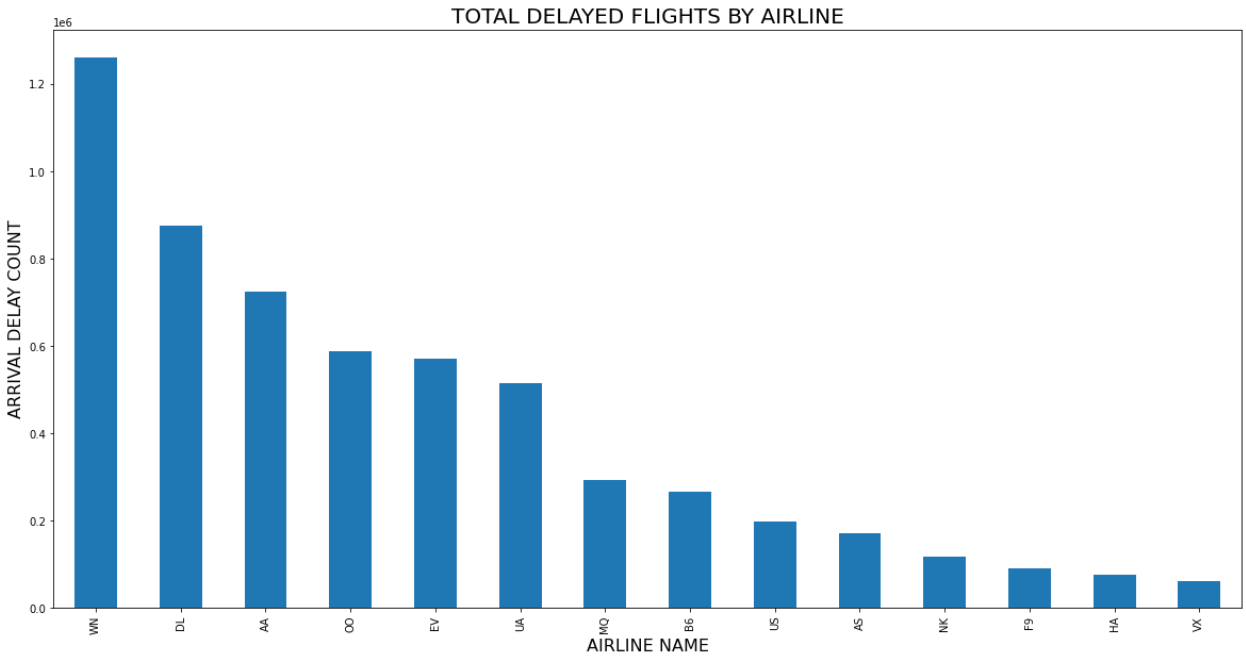
ARRIVAL_DELAY	DIVERTED	CANCELLED	AIR_SYSTEM_DELAY	SECURITY_DELAY	AIRLINE_DELAY	LATE_AIRCRAFT_DELAY	WEATHER_DELAY
5.819079e+06	5.819079e+06	5.819079e+06	5.819079e+06	5.819079e+06	5.819079e+06	5.819079e+06	5.819079e+06
4.327482e+00	2.609863e-03	1.544643e-02	2.463579e+00	1.391715e-02	3.466692e+00	4.289670e+00	5.327704e-01
3.891956e+01	5.102012e-02	1.233201e-01	1.305584e+01	9.167865e-01	2.185500e+01	2.057421e+01	8.807465e+00
-8.700000e+01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
-1.300000e+01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
-5.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
7.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
1.971000e+03	1.000000e+00	1.000000e+00	1.134000e+03	5.730000e+02	1.971000e+03	1.331000e+03	1.211000e+03

Notably, weather delays have the highest value at around 5.328. The average delay overall is around 4.327.

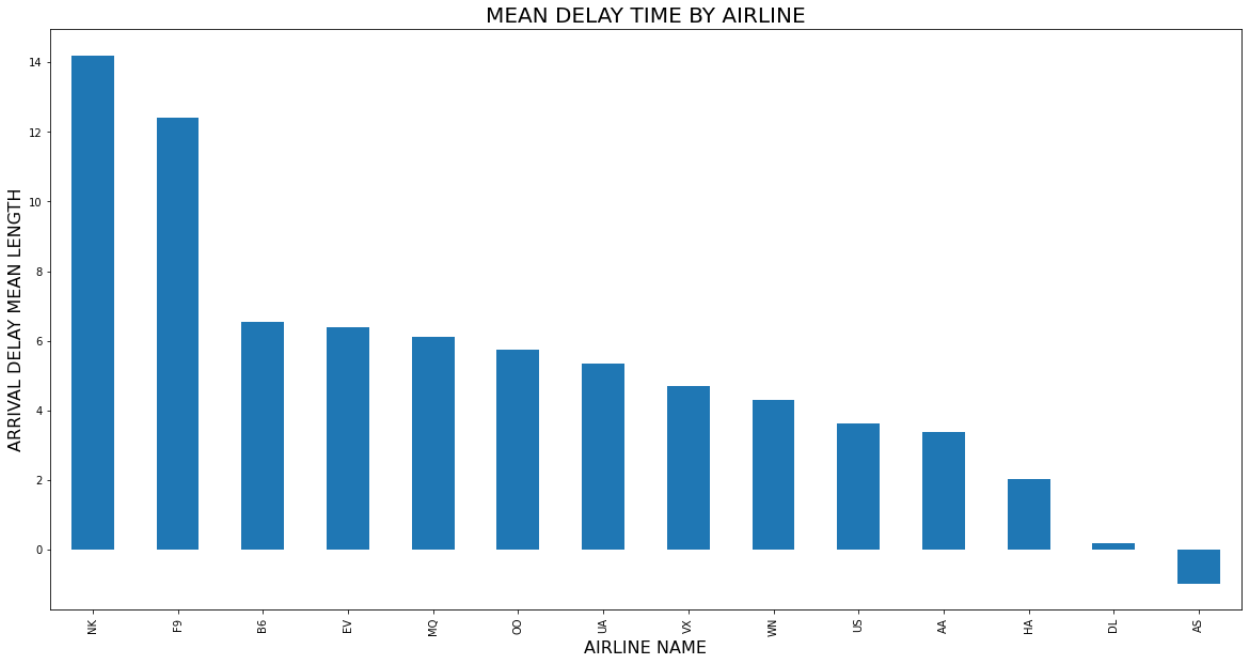
Looking at Arrival Delay andAirlines, we create a plot and can see that Southwest has the longest Arrival Delay by sum



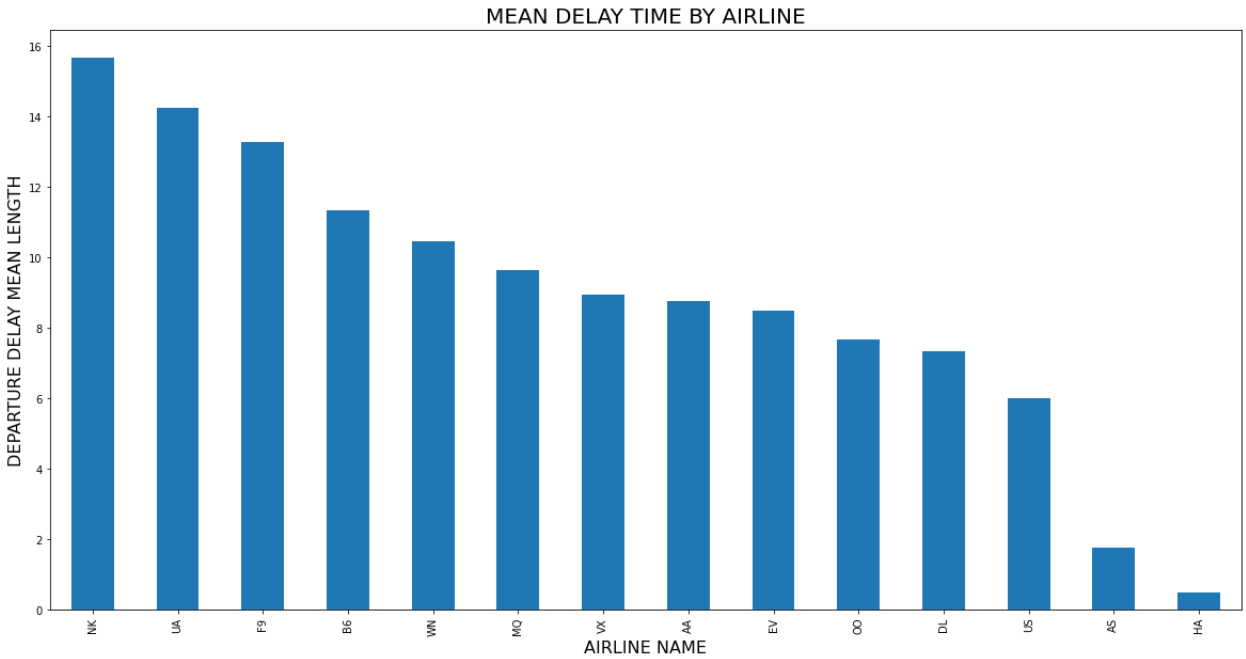
However, when grouping by count (i.e. the number of delays rather than sum of length of delays) we see that while Southwest is still the highest, Delta has moved to second, however, since these airlines are the biggest and thus have the most flights, this does not necessarily mean that these airlines have longer delays. For that we will need to see the count of delays as well as the mean:

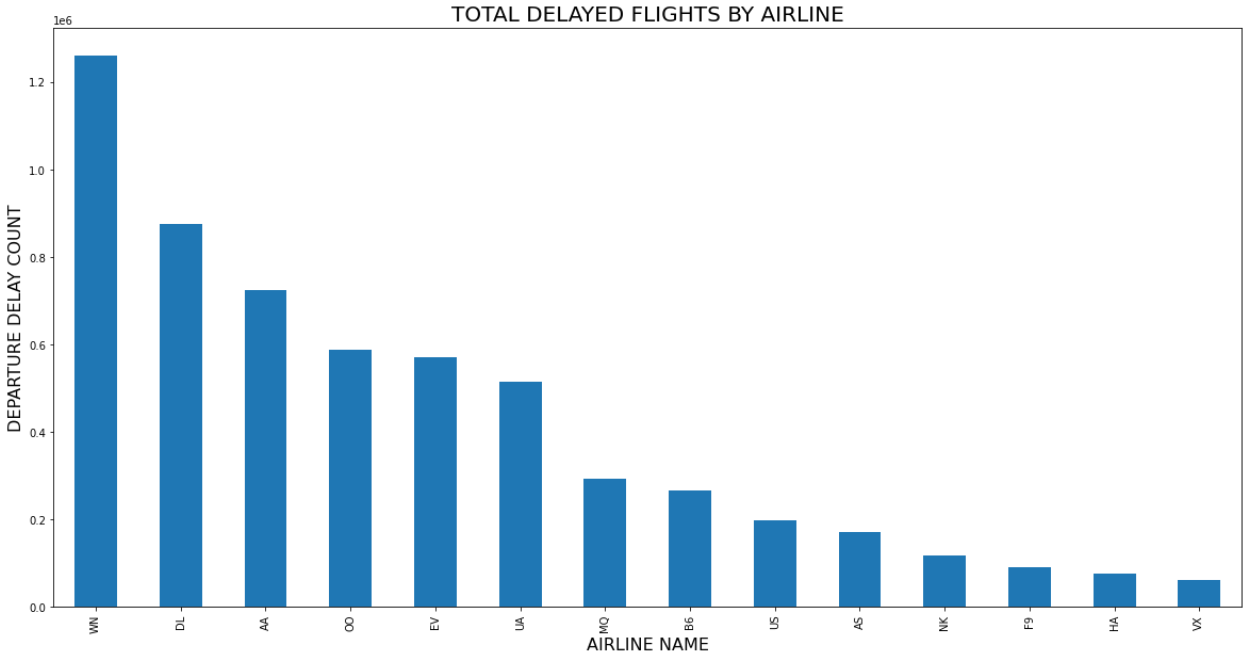


Doing the same with mean Arrival Delay Length, we see that NK airlines has the longest average arrival delay, whence WN is fairly low.

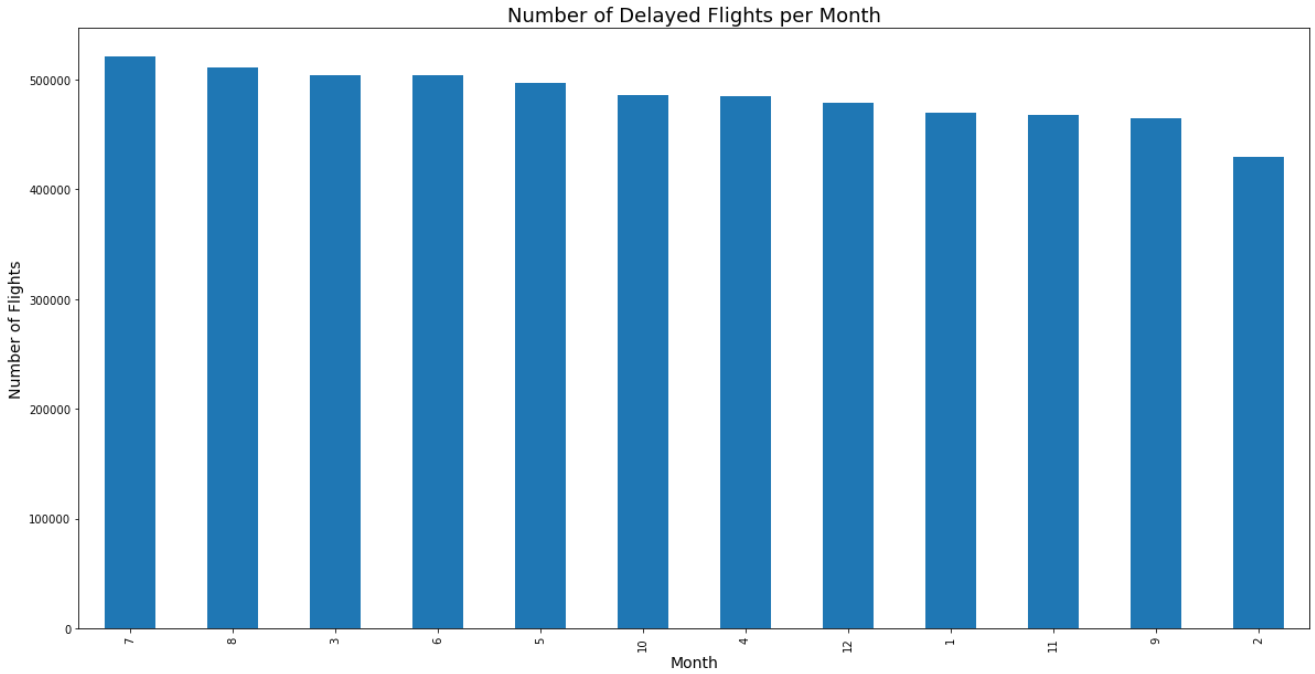


This means that these airlines, while not having the most delays (as airlines with more flights overall will have more delays) have the highest average flight delay. We find similar results when looking at mean departure delay and departure delay count, though on average it looks like mean departure delay is longer than mean arrival delay:

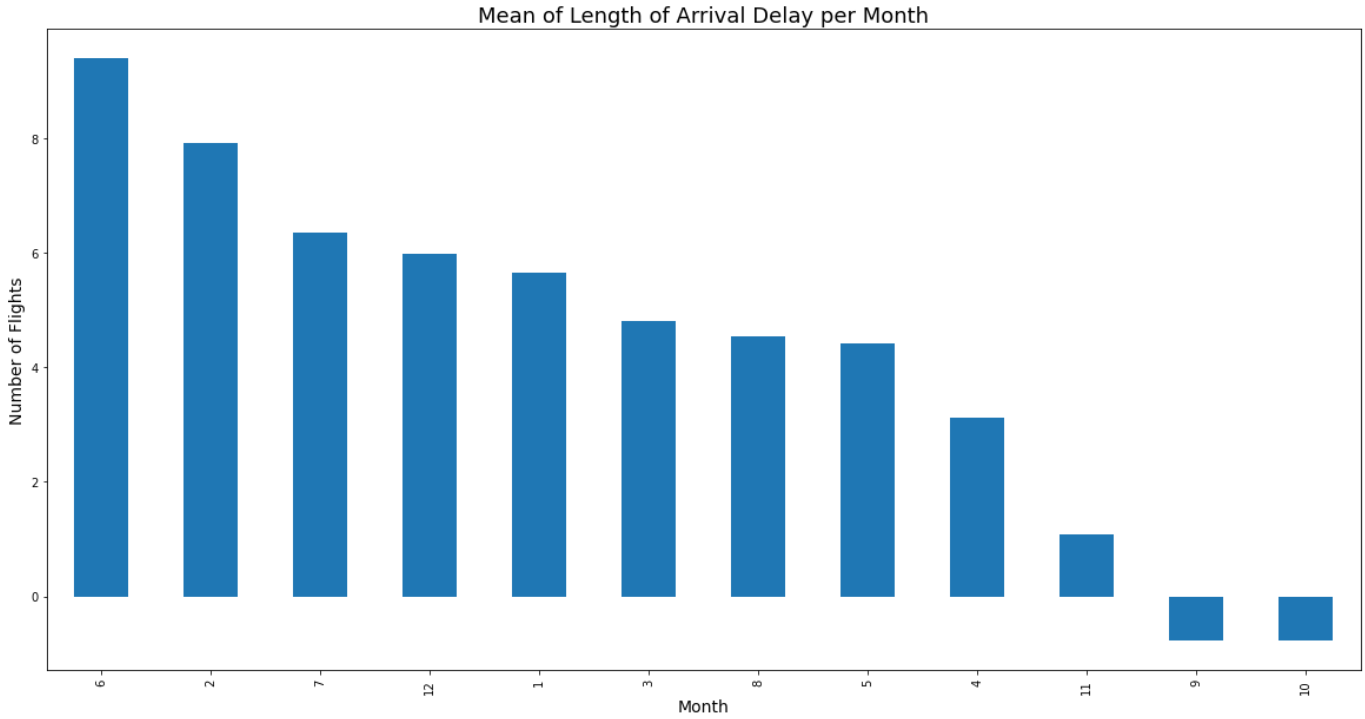




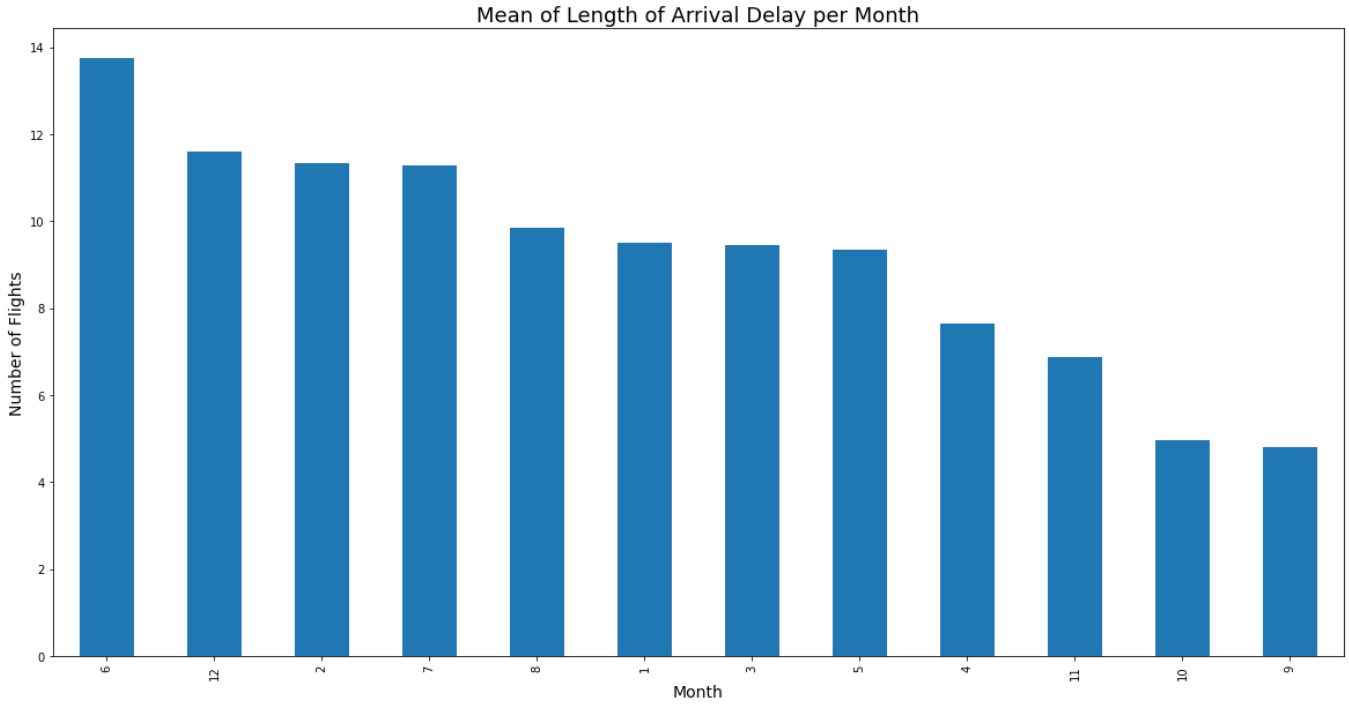
Looking at arrival delayed flights by month we can see that July has the highest count of delayed flights, followed closely by August and March



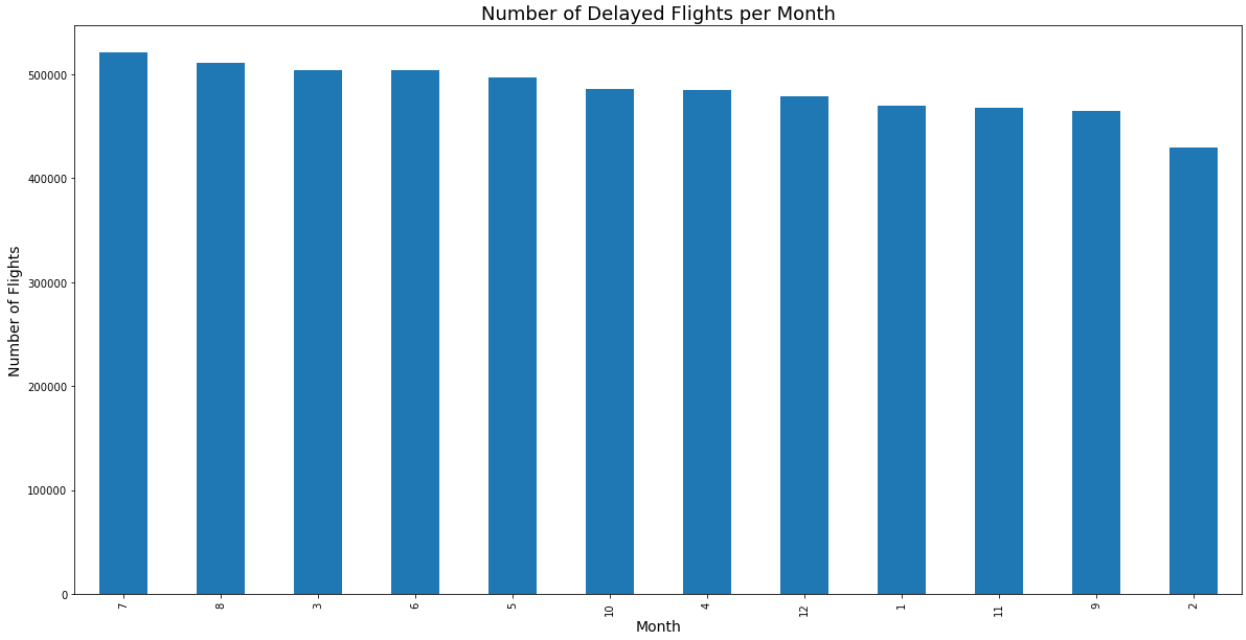
Running with mean Arrival Delay we see however that on average the longest Arrival Delays are in June, followed by February:



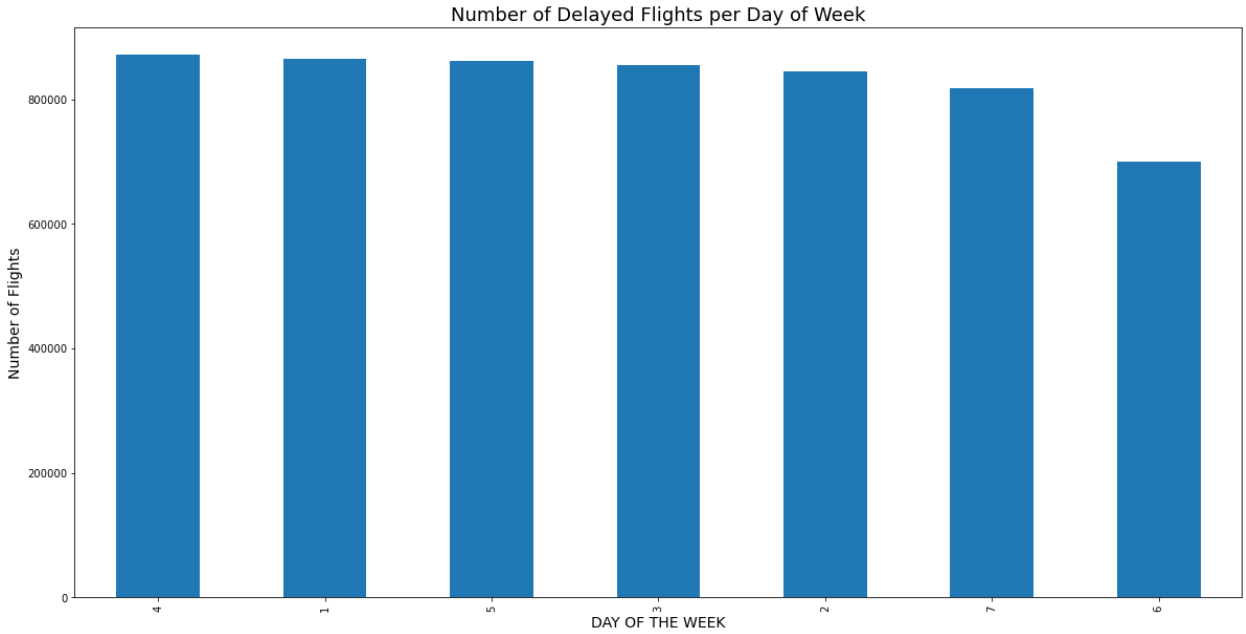
Doing the same with Departure Delays, we see that June also has the longest mean Departure Delays followed by December:



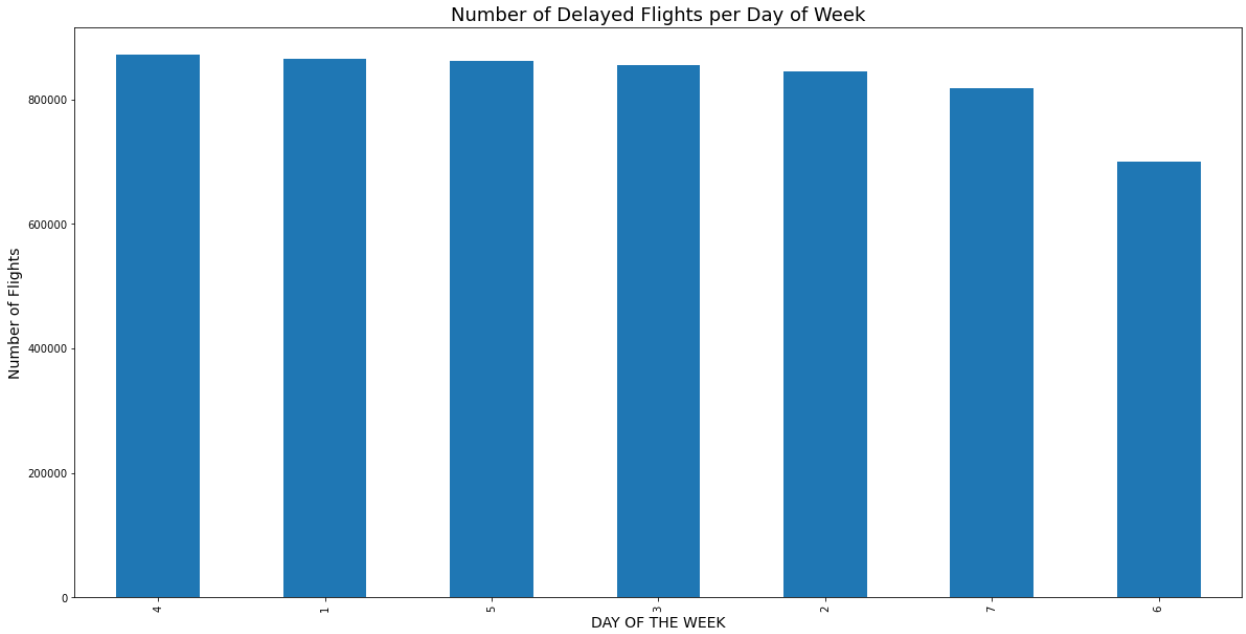
While looking at amount of Departure Delays gives us July and August as with the Arrival Delays



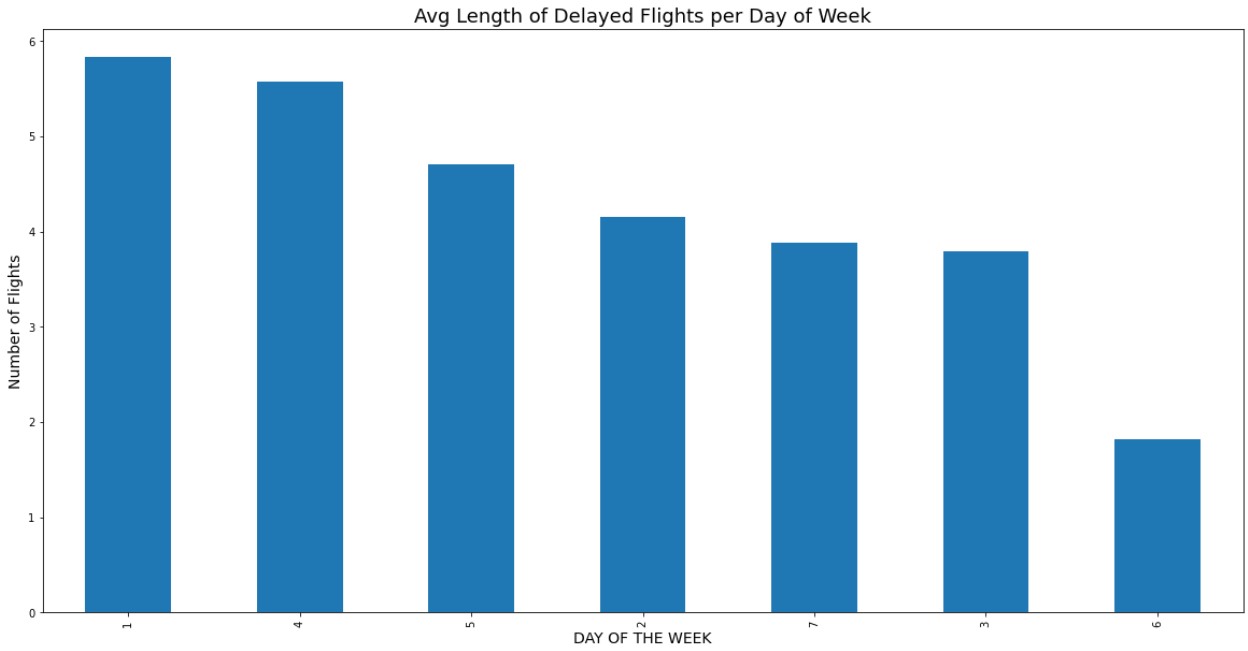
Number of Delays per day of the week is fairly uniform, with Thursdays and Sundays having the most delayed flights (Departure Delays)



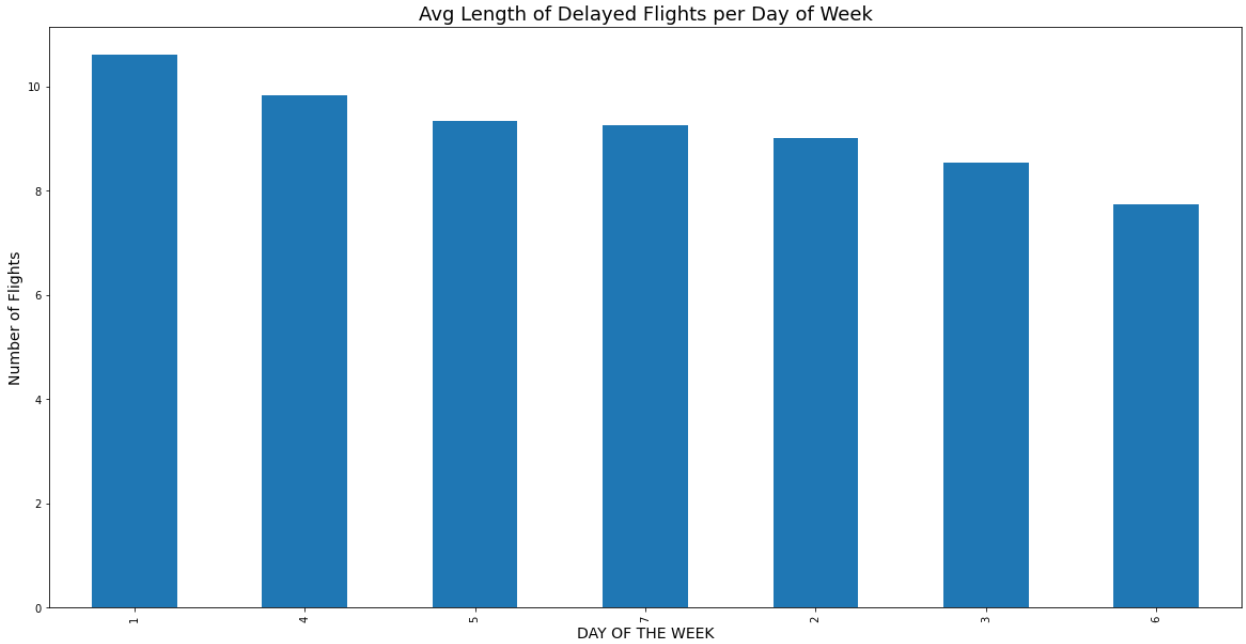
As well as similar results with Arrival Delays



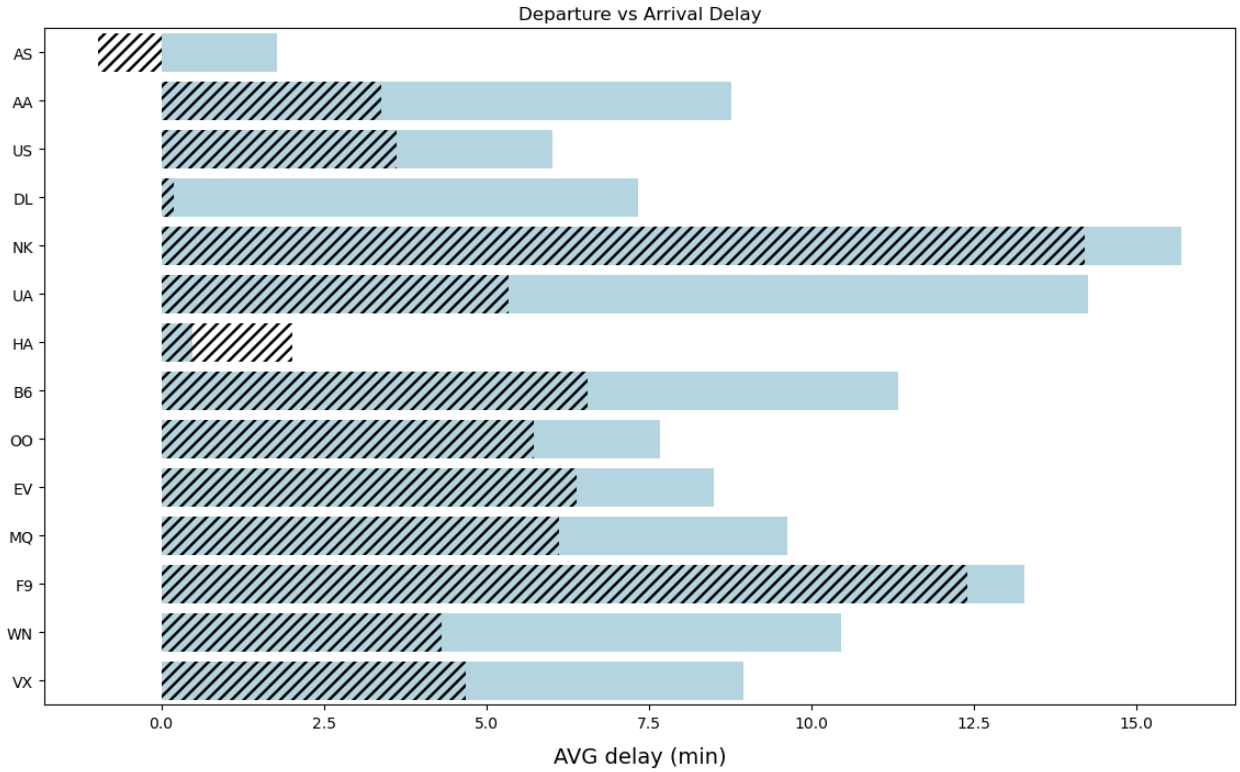
While with average length of delay,, Sunday is highest with thursday close behind (Arrival Delay)



Departure delay length is fairly similar:

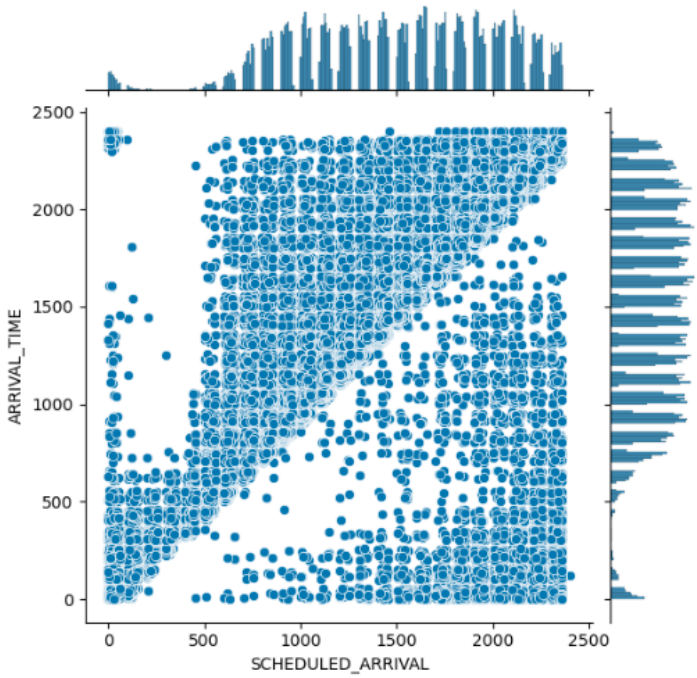


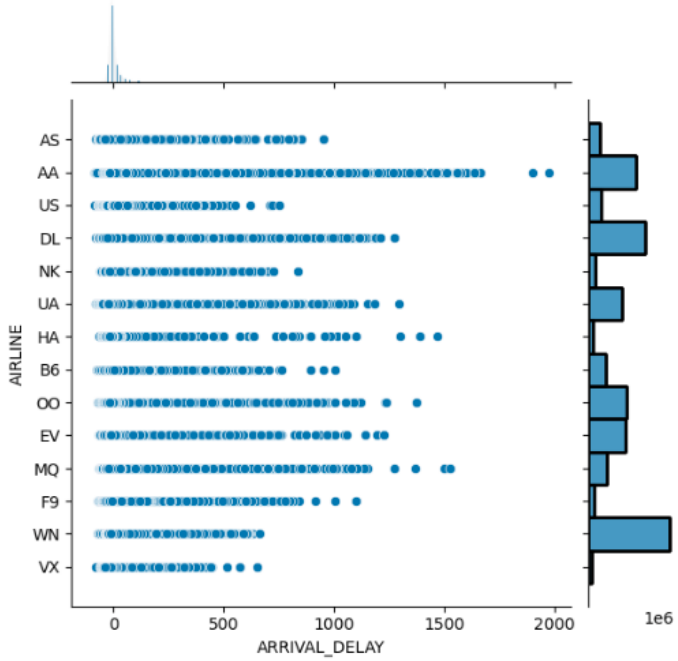
We also plot average arrival and departure delays by airline together:



Through this we can confirm that average departure delays are longer than average arrival delays, and that NK and UA (Spirit and United Airlines respectively) have the highest average delay time for Departures while NK and F9 (Frontier Airlines) have the longest average Arrival delay.

We also created joint plots to see the correlation between ARRIVAL_TIME & SCHEDULED_ARRIVAL as well as AIRLINES & ARRIVAL DELAY





b. Data Pre-Processing

First, we decided to filter out unnecessary columns in the dataset. Below is the shape the dataset after filtered out:

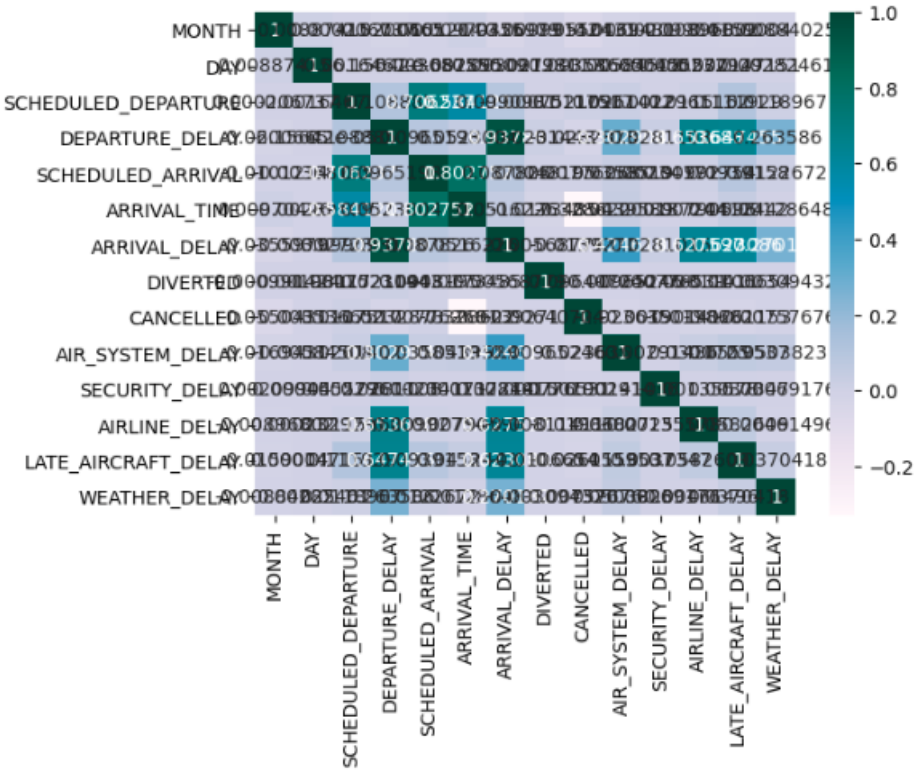
```
#Current shape of dataset

dfflights.shape
```

```
(5819079, 16)
```

Next, we will identify the NaN value within the dataset.

```
MONTH                      0
DAY                        0
ORIGIN_AIRPORT             0
DESTINATION_AIRPORT        0
SCHEDULED_DEPARTURE        0
DEPARTURE_DELAY        86153
SCHEDULED_ARRIVAL          0
ARRIVAL_TIME           92513
ARRIVAL_DELAY         105071
DIVERTED                   0
CANCELLED                  0
AIR_SYSTEM_DELAY     4755640
SECURITY_DELAY       4755640
AIRLINE_DELAY        4755640
LATE_AIRCRAFT_DELAY  4755640
WEATHER_DELAY        4755640
dtype: int64
```

We also can see that we have NaN values for Delay when no delay was present. Since this means there was no delay in these categories, we can safely set Null to 0.



Since there is no dependent variable which serves as the results whether flight is delayed or not, we will create a new feature called **'result'** which takes values 0 and 1. In which, 0 means flight is not delayed and 1 means flight is delayed. Below is the value count in 'result' feature:

```
dfflights['result'].value_counts()
```

```
0    3732183
1    2086896
Name: result, dtype: int64
```

c. Building Model

For model building, we select only these columns 'MONTH', 'DAY', 'SCHEDULED_DEPARTURE', 'DEPARTURE_DELAY', 'SCHEDULED_ARRIVAL', 'DIVERTED', 'CANCELLED', 'AIR_SYSTEM_DELAY','SECURITY_DELAY', 'AIRLINE_DELAY', 'LATE_AIRCRAFT_DELAY','WEATHER_DELAY', 'result'

```
#Splitting dataset into Training and Testing with 70:30 ratio
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import roc_auc_score

data = dfflights.values
```

```
X, y = data[:,:-1], data[:,-1]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=42)
```

```
#Feature Scalling

scaled_features = StandardScaler().fit_transform(X_train, X_test)
```

```
#Model: Decision Tree Classifier

clf = DecisionTreeClassifier()
clf = clf.fit(X_train,y_train)
pred = clf.predict_proba(X_test)
```

```
#AUC score of the Model
auc_score = roc_auc_score(y_test, pred[:,1])
auc_score
```

```
0.7639666446049135
```

At this stage, our baseline model achieved 76.39% accuracy.