Analysis of Flight Statistics in the UK for June and July 2022

Author: Hamza,Asma Alpha Group

Submission Date: 20-08-2022

Introduction

The data we are going to work on today is retreived from UK CAA Flight Punctuality Statistics. The selected data is for Reporting Month 04 and 05. Report 04 is for June and 05 is for July. It holds all the records which can be used to derive punctuality statistics for the flights and airport. This data has detail about **the flight routes from one airport to another airport**. It involves data of **British and Non British, Airports and Airlines**. The data has **flights departed from the UK and arrived in the UK.** It also tells if a **flight is Chartered or Scheduled.** It contains data about **number of flight on a specific route and the percentage of flights delayed on that route with Average Delay and Cancelled.** The data can be downloaded from here. There was a need for this data to be analysed and visualised. Because United Kingdom is facing a huge **labour shortage** with very high temperatures and the economy is in a **recession**. Which is effecting all sectors of the economy. Moreover June and July are the busiest months of the year for aviation industry because of holidays hence there is a surge in passengers into and out of the country.

Objectives

- We have to clean the data and make it ready for analysis.
- We will observe the distribution of the data.
- We will discuss the Average Delay by Minutes for each Month
- We have to find the Number of Flights Delayed and Cancelled of each airline.
- We have to find the Percentage of Flights Delayed and Cancelled groupby Airlines.
- We have to find the Percentage of Flight Share among Airports in the UK
- We have to find the Average Delay in Minutes by Month at an Airport in the UK
- We have to find the Percentage of Number of Flights Delayed at an Airport in the UK

Data Description

The data is in **csv** format. We have two seperate files for data. One contains data of June and the other contains the data of July.It is pre-seperated. It consists of 26 Features and 11,466 Instances in total.

Feature	Description
reporting_period	**Reporting Period**
number_flights_matched	**Total Number of Flights**
actual_flights_unmatched	**Flights which didn't match the route**
number_flights_cancelled	**Number of Flights Cancelled**

Feature	Description
flights_more_than_15_minutes_early_percent	**Percentage of Flights Delayed or Early Simplified to Duration**
flights_15_minutes_early_to_1_minute_early_percent	**Percentage of Flights Delayed or Early Simplified to Duration**
flights_0_to_15_minutes_late_percent	**Percentage of Flights Delayed or Early Simplified to Duration**
flights_between_16_and_30_minutes_late_percent	**Percentage of Flights Delayed or Early Simplified to Duration**
flights_between_31_and_60_minutes_late_percent	**Percentage of Flights Delayed or Early Simplified to Duration**
flights_between_61_and_120_minutes_late_percent	**Percentage of Flights Delayed or Early Simplified to Duration**
flights_between_121_and_180_minutes_late_percent	**Percentage of Flights Delayed or Early Simplified to Duration**
flights_between_181_and_360_minutes_late_percent	**Percentage of Flights Delayed or Early Simplified to Duration**
flights_more_than_360_minutes_late_percent	**Percentage of Flights Delayed or Early Simplified to Duration**
flights_unmatched_percent	**Percentage of Flights Unmatched**
flights_cancelled_percent	**Percentage of Flights Cancelled**
average_delay_mins	**Average Delay in Minutes**
previous_year_month_flights_matched	**Total of Previous Year Month Flights Matched**
previous_year_month_early_to_15_mins_late_percent	**Percentage of Previous Year Month Flights Delayed or Early Simplified to Duration**
previous_year_month_average_delay	**Average Delay in Minutes of Previous Year Month**

Feature Analysis

Feature Type	Quantity
Categorical	7
Numerical	19

Data Preprocessing

It is a stage in the data analysis process that converts raw data into a format that computers and machine learning software can understand and evaluate.So first **we remove unecessary columns to clear up the data**

and then also rename the columns with long names. This will help us access the data in an easy way.

```
june = june.drop(columns=
['reporting_period','flights_unmatched_percent','previous_year_month_flights_match
ed','previous_year_month_early_to_15_mins_late_percent','previous_year_month_avera
ge_delay','flights_more_than_15_minutes_early_percent',
'flights_15_minutes_early_to_1_minute_early_percent','actual_flights_unmatched'])

july = july.drop(columns=
['reporting_period','flights_unmatched_percent','previous_year_month_flights_match
ed','previous_year_month_early_to_15_mins_late_percent','previous_year_month_avera
ge_delay','flights_more_than_15_minutes_early_percent',
'flights_15_minutes_early_to_1_minute_early_percent','actual_flights_unmatched'])

june.rename(columns=
{'0_to_15_minutes_late_percent':'0_15_mins_late_pct','flights_between_16_and_30_mi
nutes_late_percent':'16_30_mins_late_pct','flights_between_31_and_60_minutes_late_
percent':'31_60_mins_late_pct','flights_between_61_and_120_minutes_late_percent':'
61_120_mins_late_pct','flights_between_121_and_180_minutes_late_percent':'121_180_
mins_late_pct','flights_between_181_and_360_minutes_late_percent':'181_360_mins_la
te_pct','flights_more_than_360_minutes_late_percent':'+_360_mins_late_pct','flight
s_0_to_15_minutes_late_percent':'0_15_mins_late_pct'},inplace=True)

july.rename(columns=
{'0_to_15_minutes_late_percent':'0_15_mins_late_pct','flights_between_16_and_30_mi
nutes_late_percent':'16_30_mins_late_pct','flights_between_31_and_60_minutes_late_
percent':'31_60_mins_late_pct','flights_between_61_and_120_minutes_late_percent':'
61_120_mins_late_pct','flights_between_121_and_180_minutes_late_percent':'121_180_
mins_late_pct','flights_between_181_and_360_minutes_late_percent':'181_360_mins_la
te_pct','flights_more_than_360_minutes_late_percent':'+_360_mins_late_pct','flight
s_0_to_15_minutes_late_percent':'0_15_mins_late_pct'},inplace=True)
```

We have removed the unecessary columns and also renamed the columns with long names. We are left with 18 Columns now.

Statistical Overview

Let's do a statistical analysis on the data. We can do this by using the **describe()** function.

```
june.describe()
july.describe()
```

For June we have 5232 rows and for July we have 6234 rows. Maximum number of flights matched is 344 for a route in June and in July it is 362. 15 Flights were cancelled in June whereas in July there 19 flights cancelled. Average delay in minutes for a route in June was 20.33, it increased in July to 29.61 minutes.

Retreiving Null Values

Next we check the null values in the data. We can do this by using the **isnull()** function with **sum()** to sum the values. Both datasets have the null values in the same column. Which is average_delay_mins. We need to see the rows where there are null values so we can decide that either we have to keep the values and fill accordingly or we have to drop the NaN values. To retreive the data we used the following code:

```
june[june['average_delay_mins'].isnull()].head(5)
```

```
july[july['average_delay_mins'].isnull()].tail(5)
```

I used a head and tail function to get the first 5 rows and last 5 rows of June and July so what ever the situation it is verifiable as we can't see all the null value rows because there are around 10,000 rows. The returned rows had no data in any of their numerical columns. Hence it was understood why they were returning null values in the average_delay_mins column. No data means there were 0's but not NaN values. Otherwise we would have seen the NaN values in other columns as well.

More Data Cleaning and Manipulation
Now our Data Stands at ((4894, 18), (5820, 18)) from ((5232, 26), (6234, 26)). **We need to narrow our data down to Departures and Scheduled flights** Because UK's situation will only be reflected by departures because the reason why a flight from the UK is known to be labour shortage etc as told earlier, but if a flight is arriving from Germany, and it's delayed it is because of the problems in Germany not in the UK. Hence we have to select Departures only. Second thing is that we only need Scheduled flights not Chartered because Chartered are unscheduled and a lot of times they are private or special flights. Hence this will create inconsistency in the data. So we will select Scheduled flights only.

```
june_updated = june[(june['arrival_departure']=='D')&
(june['scheduled_charter']=='S')]

july_updated = july[(july['arrival_departure']=='D')&
(july['scheduled_charter']=='S')]
```

Next thing Now We need a column of Total percentage of flight dealyed on a route. Because as of now the data is spread into multiple columns hence it is very difficult to read and difficult to retreive it. It also creates confusion despite renaming those columns. Hence we create a new column called **delay_pct** which is the sum of all the columns. The data in it is sum of all other columns with delay percentage.This will get rid of us reading multiple columns of delay percentage

Subsequently we do have the percentage of flights delayed but we don't have the number of the flights delayed. This creates a confusion while interpreting the data. Hence it is necessary that we create another column to get the number of flights delayed.

```
june_updated['no_flight_delayed']= (june_updated['number_flights_matched'] *
june_updated['delay_pct']) /100
```

```
july_updated['no_flight_delayed']= (july_updated['number_flights_matched'] *
july_updated['delay_pct']) /100
```

We need to change the datatype of these columns hence we use the **astype()** function to change the datatype to integer.

Finalized Dataset Info:

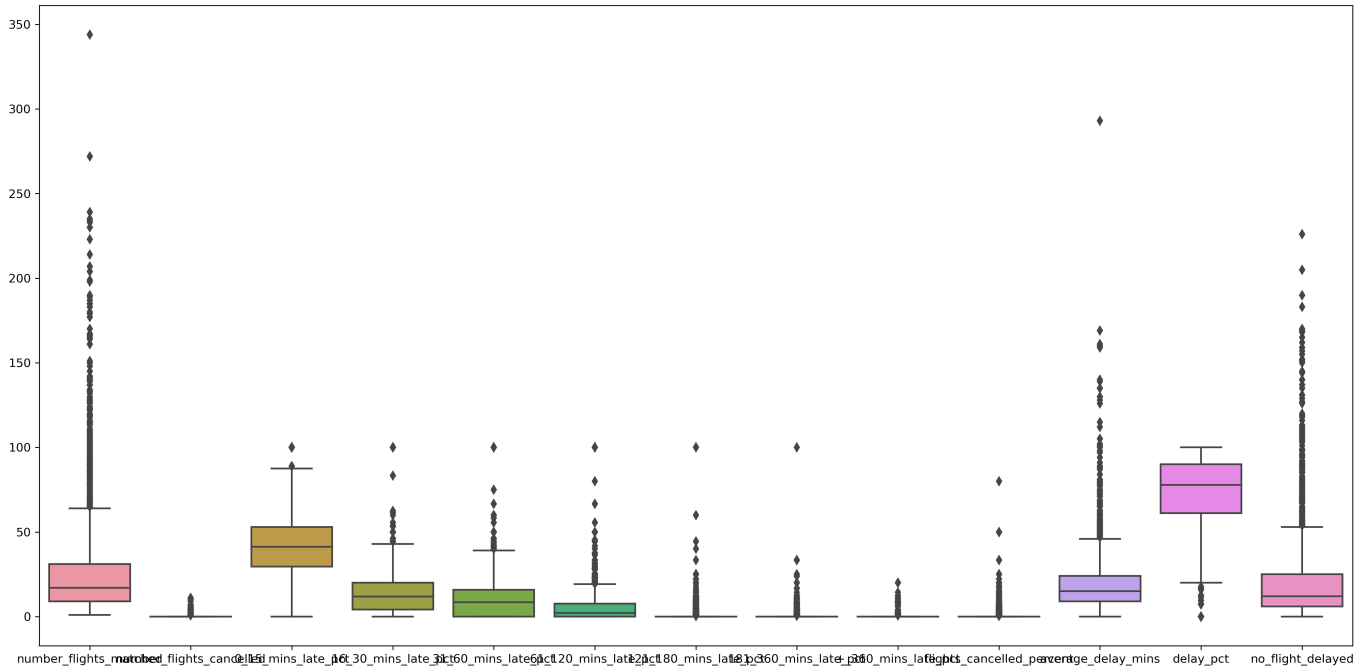
- We have 20 columns in the dataframe
- We have 7 categorical features
- We have 13 Numerical Features
- There are no null values in any of the columns
- 2119 rows for June and 2420 rows for July

As far as **duplicates** are concerned, we do have duplicates in Categorical Column. Which is understood as the same route is being flown on different days. And same airlines fly the same route. But removing the Arrivals and Chartered Flights from the data has made data stable and removed unecessary inconsistencies.

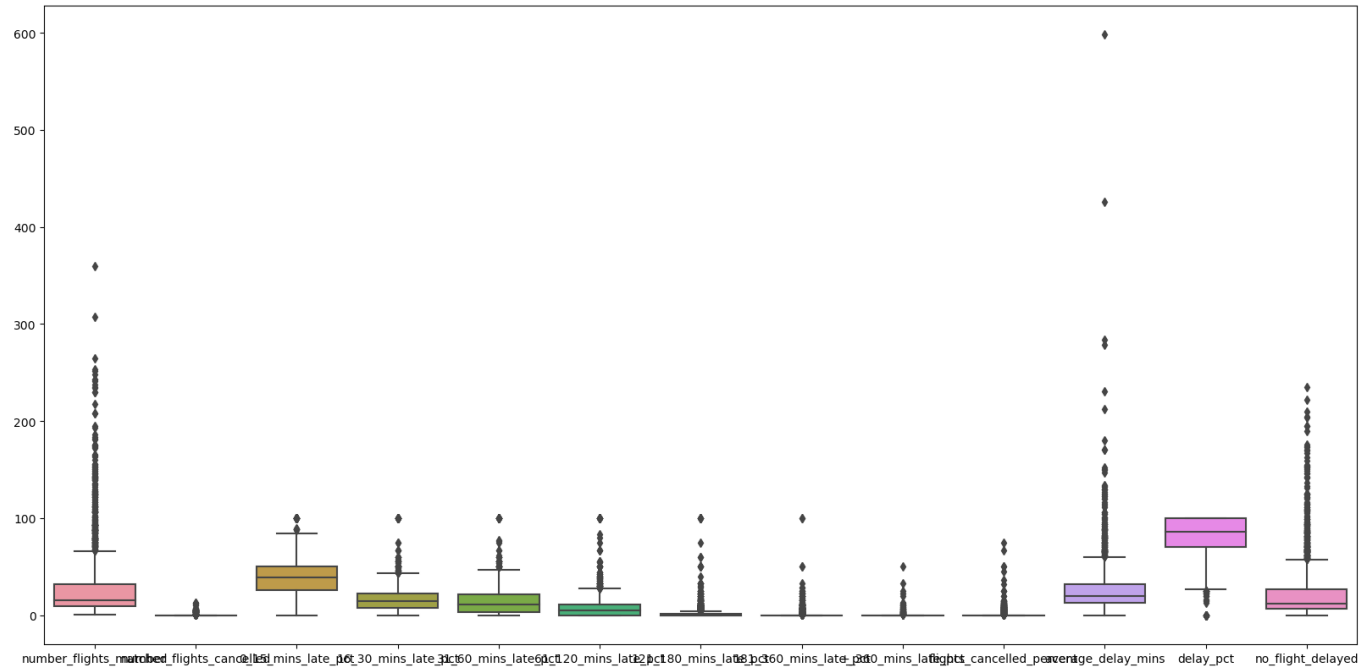
Data Processing Complete

Let's check the outliers in the data

Outlier in June



Outlier in July



Observation :

- There are a lot of outliers in the both datasets
- The most of the outliers are in number of flights
- We have some outliers below lower bound of delay_pct
- There are many outliers in no of flights delayed

The following code is used to calculate the upper bound to retreive the outliers.

```
Q1_june = june_updated['number_flights_matched'].quantile(0.25)
Q3_june = june_updated['number_flights_matched'].quantile(0.75)
IQR = Q3_june - Q1_june
Upper_bound = (Q3_june)+(1.5*IQR)
june_updated[june_updated['number_flights_matched']>Upper_bound].sort_values(by='n
umber_flights_matched',ascending=False).head(20)
```

Observation :

- These are outliers but are realistic and possible values because Aer Lingus Operates on avg 12-15 flights a day between Dublin and Heathrow hence the toll for the month is between 360-400 And especially when there are holidays we can expect such big volumes.

Checking for Outliers in delay_pct

```
Q1_june = june_updated['delay_pct'].quantile(0.25)
Q3_june = june_updated['delay_pct'].quantile(0.75)
IQR = Q3_june - Q1_june
Lower_bound = (Q1_june)-(1.5*IQR)
june_updated[june_updated['delay_pct']
<Lower_bound].sort_values(by='delay_pct',ascending=False).head(20)
```

```
june_updated['delay_pct'].mean()
```

Observation :

- The outliers are below the lower bound because the flights tend to delay as the mean is somewhere close to 75 so the flights which delayed for 15-20 minutes or were on time are considered outlier. But it doesn't mean that they are not realistic

The same steps were performed for the July database as well. And the findings are also the same because we are talking about to months next to each other and the data is of the same country. Hence we can't expect much change in the data.

Should we remove the outliers?

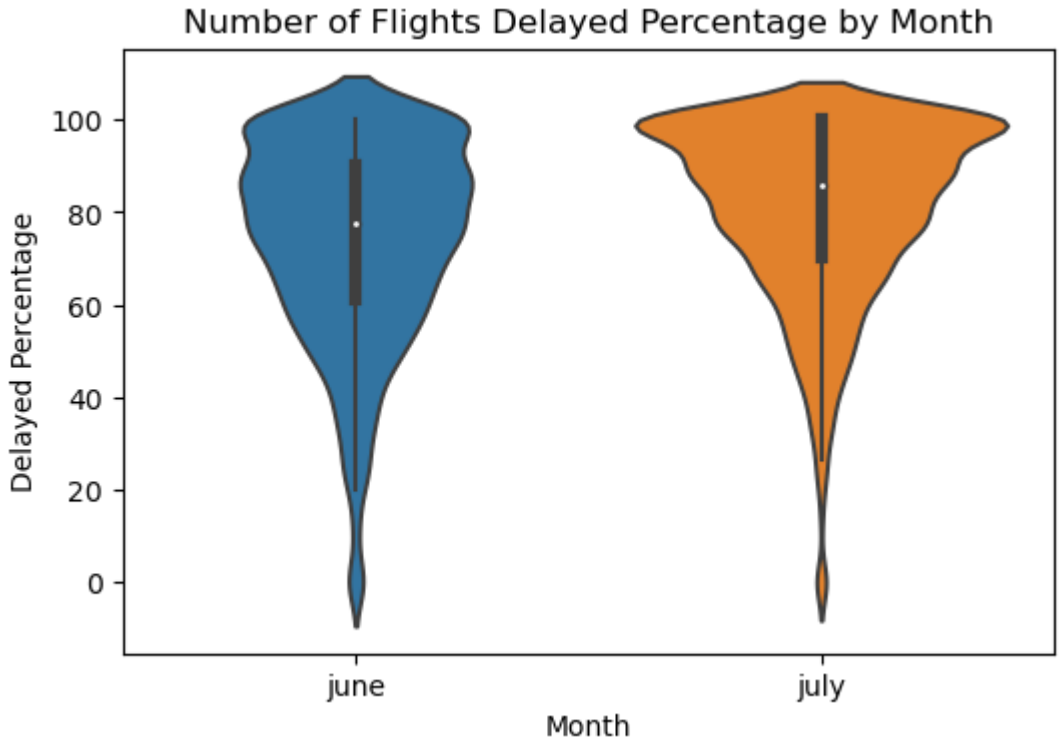
Well this is debateable, but I would say No. Because we need the details as it is on the table. As a Data Scientist our job is to process the data without effecting the useful data. For now we are not going to proceed for Machine Learning hence there is no need to remove the outliers. For instance, it is a possibility that an airline is going through severe financial times hence we can expect huge delays which are way off and not normal. The reason could be failiure to pay the airport costs or other reasons which are not relevant as of now.

Secondly this data is directly published by the Civil Aviation Authority of the United Kingdom. Hence we can expect the data to be stable and reliable. So the data is good to go.

Lastly, we have a huge data so it is normal to get this amount of outliers.

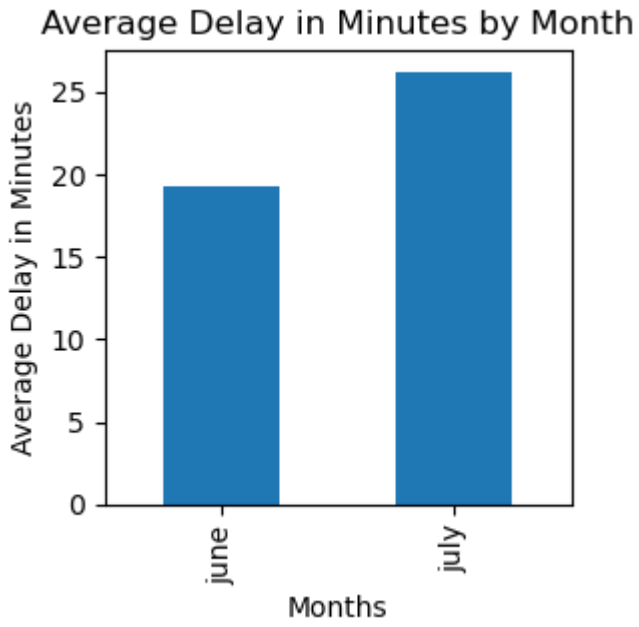
Visualization

Univariate Analysis



Observations :

- A violin plot, which depicts data peaks, is a cross between a box plot and a kernel density plot.
- The Median of July is higher than June which tells us that the many of the values are greater in Jult
- There is more data for July comparitively to June. We can see clearly by the spread of the plot in it's Center that how much it is spreading towards outside.
- For July there are more chances for a flight to get delayed because the Quantiles are greater than June's Quantiles
- Hence we can see that July is the more busy than June

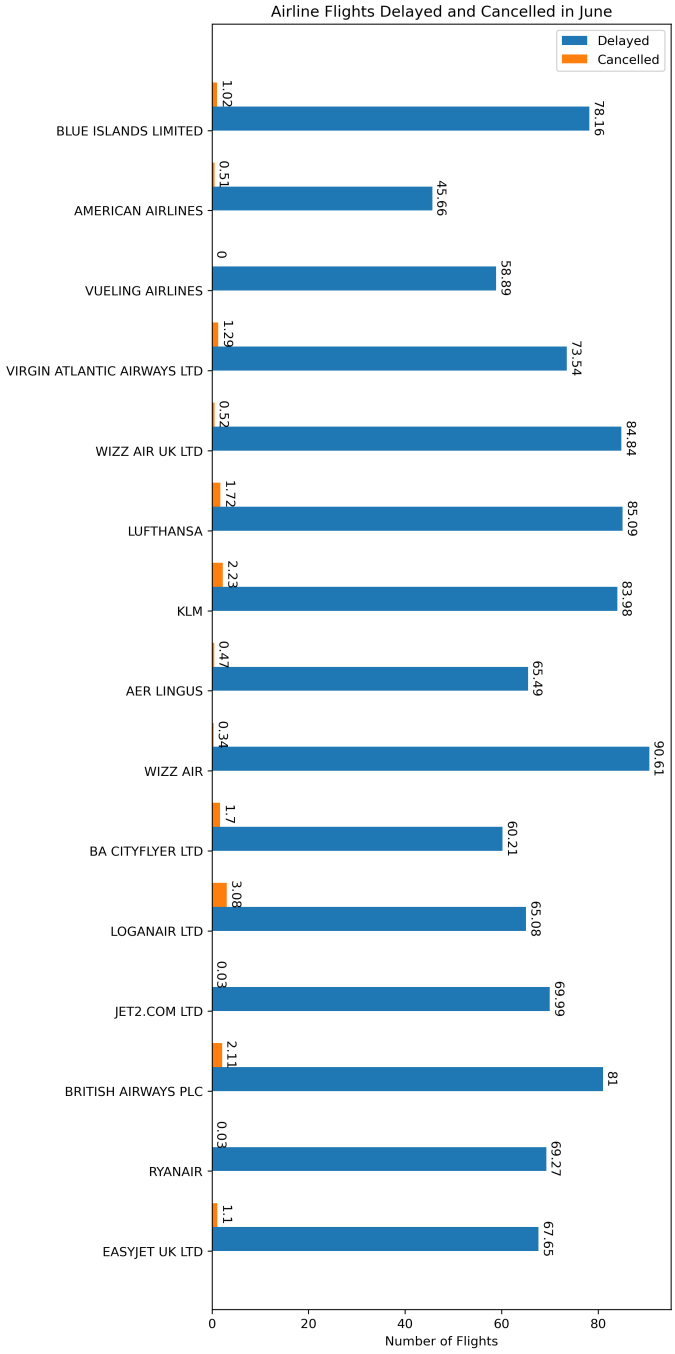


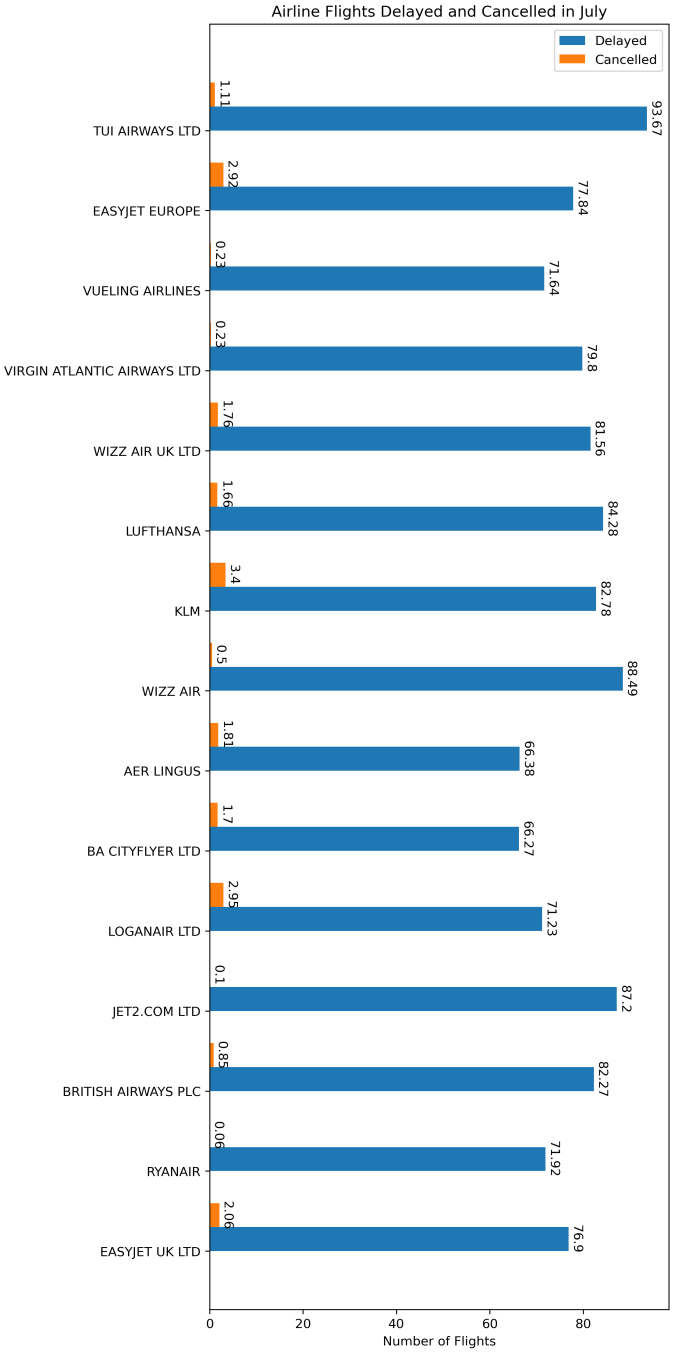
Observations :

- The average delay is higher for July than June

- The reason could be because of the higher number of flights in July

Multivariate Analysis

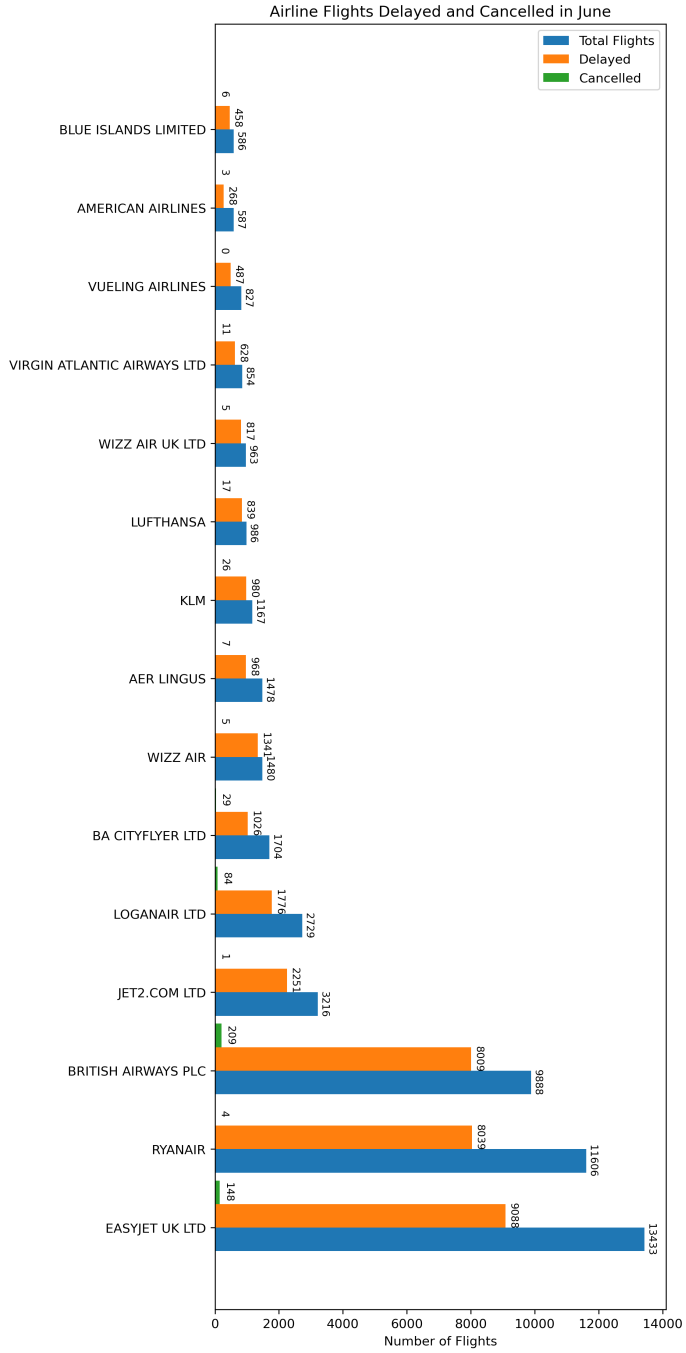


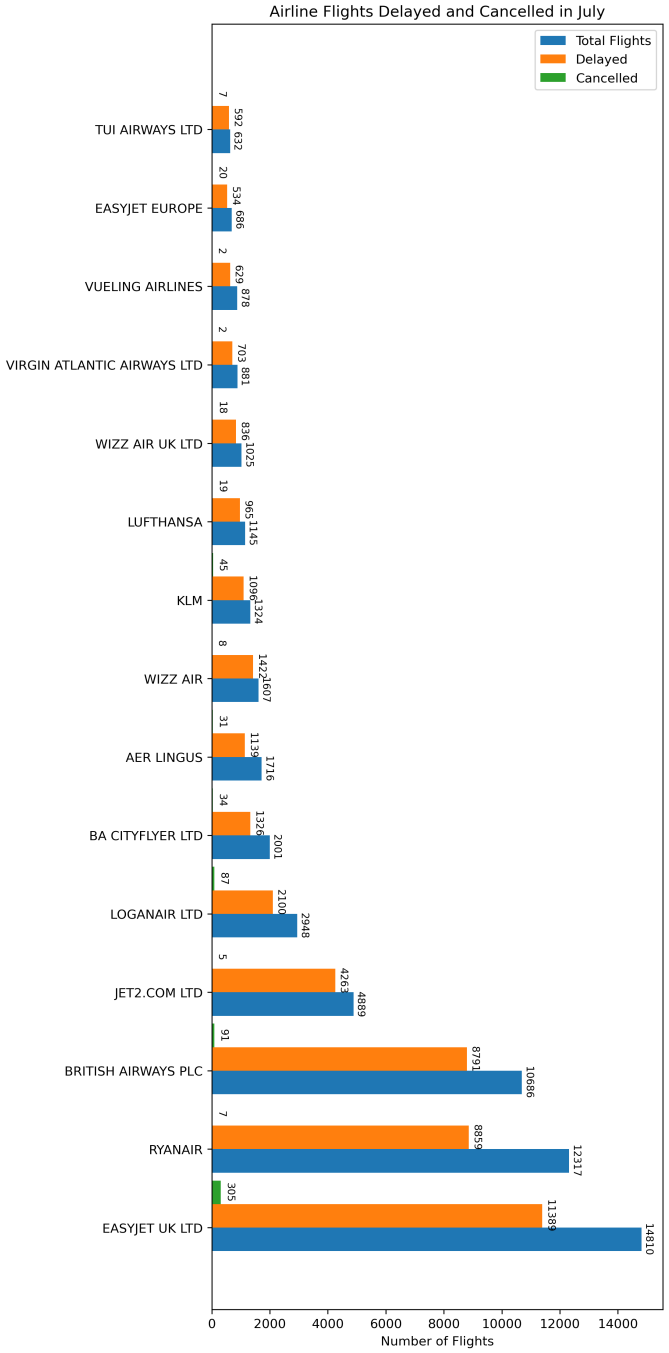


Observation :

- This graph shows us the delay and cancellation percentage of each airlines
- Wizz air had the highest percentage of flights delayed in June
- Surprisingly followed by Lufthansa at 85% of the flights delayed in June
- EasyJet the most popular airline has 67.65% of the flights delayed in June
- RyanAir stood at 69.27%
- British Airways was slightly more than 80 at 81% of the flights delayed in June
- TUI Airways had the highest delay percentage of flights in July at 93% whereas Wizz Air had only 90%
- EasyJet's Flight delay percentage increased by almost 10% in July
- RyanAir's Flight delay percentage increased by about 2%
- Highest Change was observed in Jet2 which increased by 17% in July

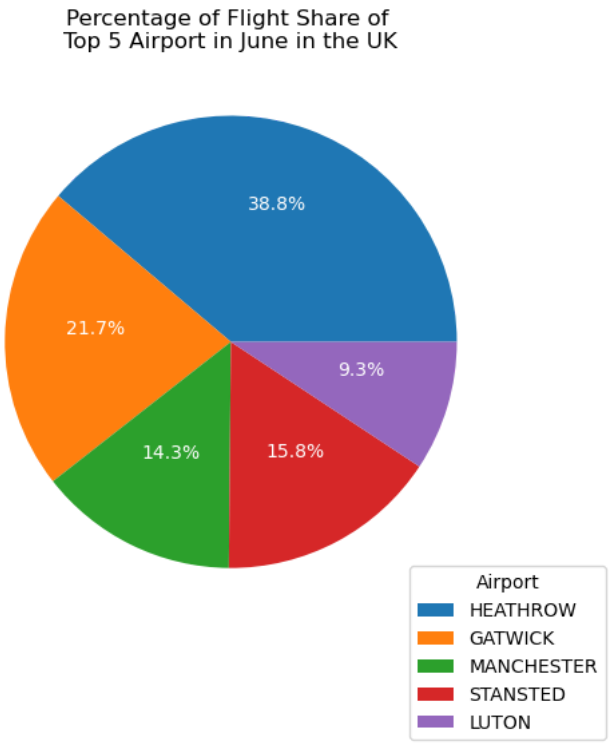
- Most cancellations were observed in Logan Air in June which was replaced by KLM in July

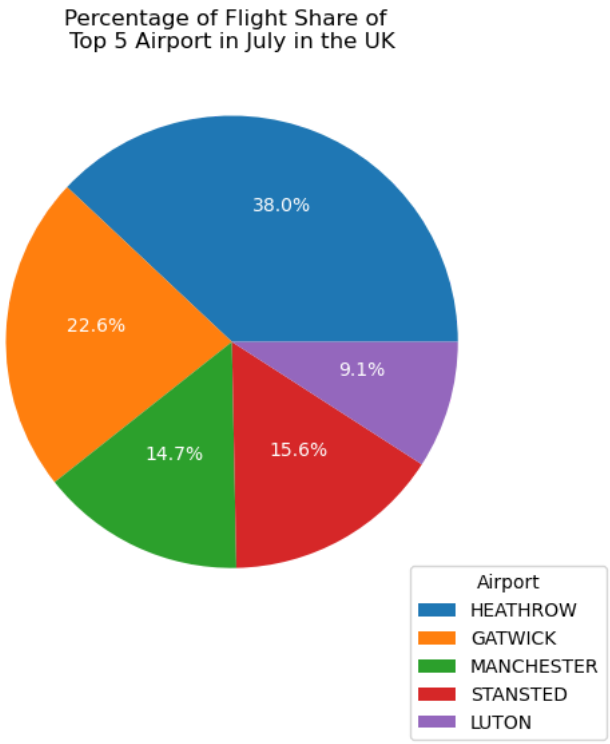




Observation :

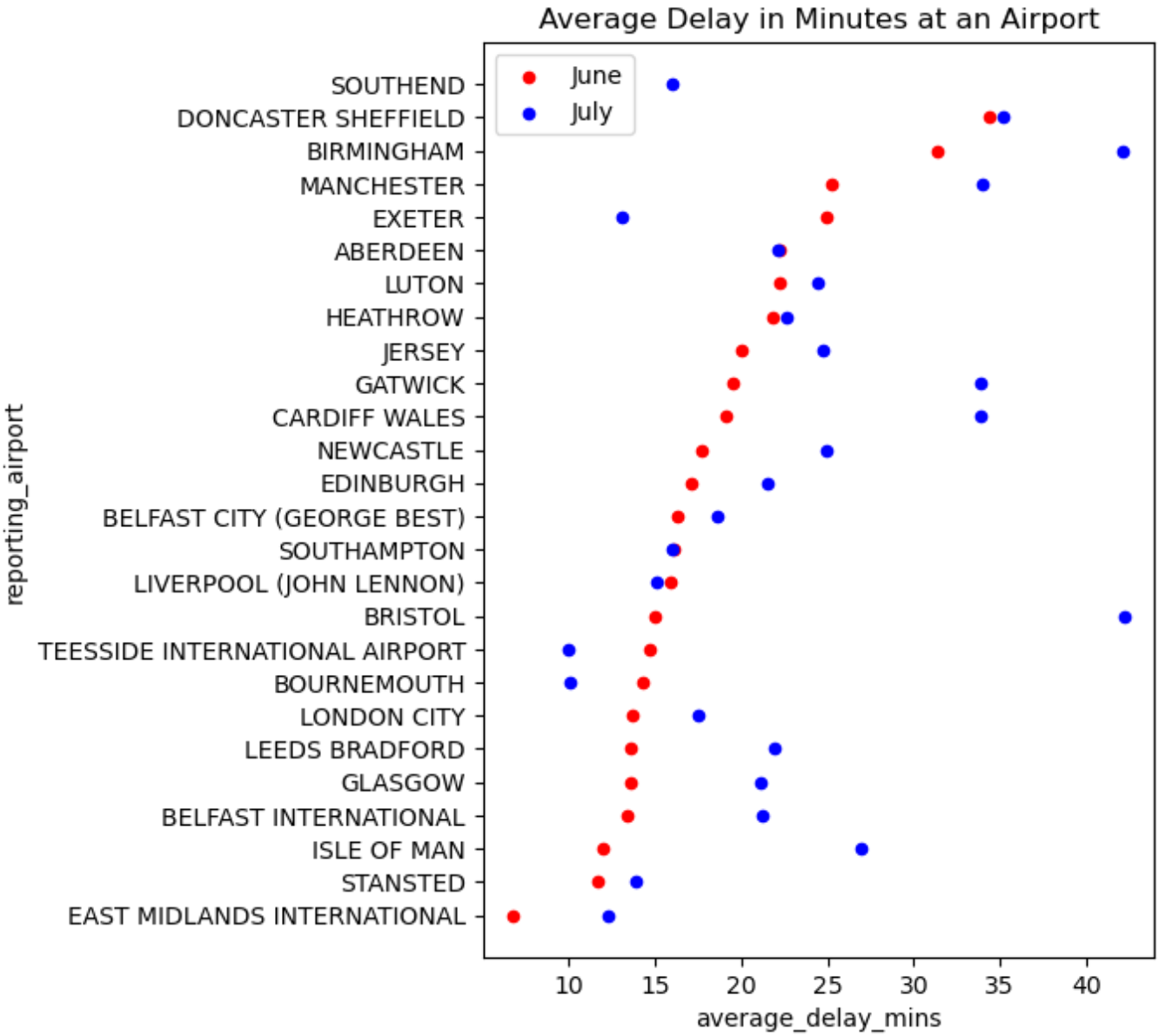
- This graph shows us the number of flights ,delayed flights and cancelled flights of each airlines.
- EasyJet has the most number of flight in the UK followed by RyanAir and British Airways
- Number of Flights which are delayed is higher for EasyJet than RyanAir and British Airways





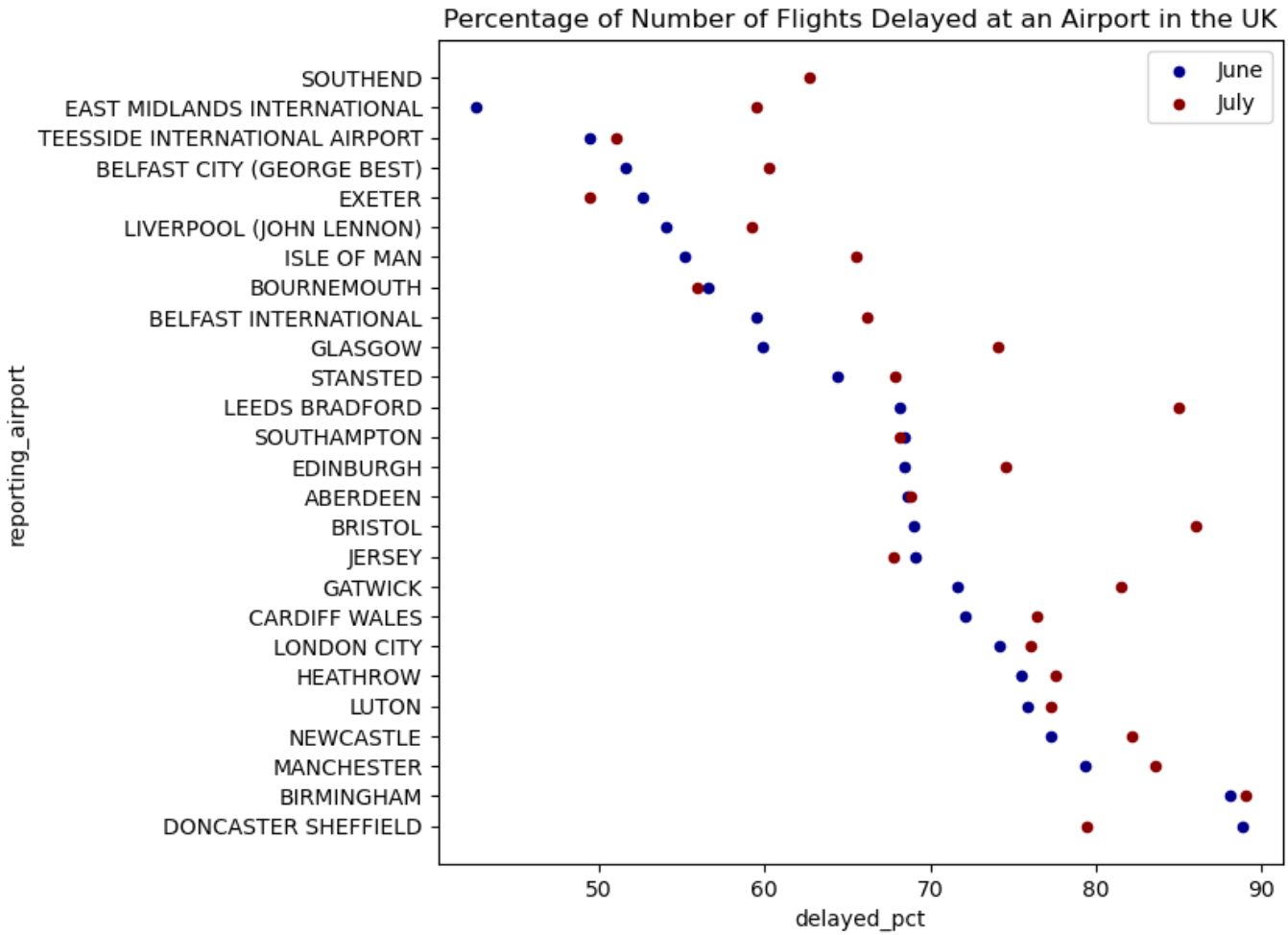
Observation :

- This chart tells us the share of flights from each aiport in UK
- The busiest airport in the UK in June and July was Heathrow
- Almost 38% of the departures in the UK is carried by Heathrow
- There is no notable difference in change between both months
- 4 of the Top 5 busiest airports are in London:



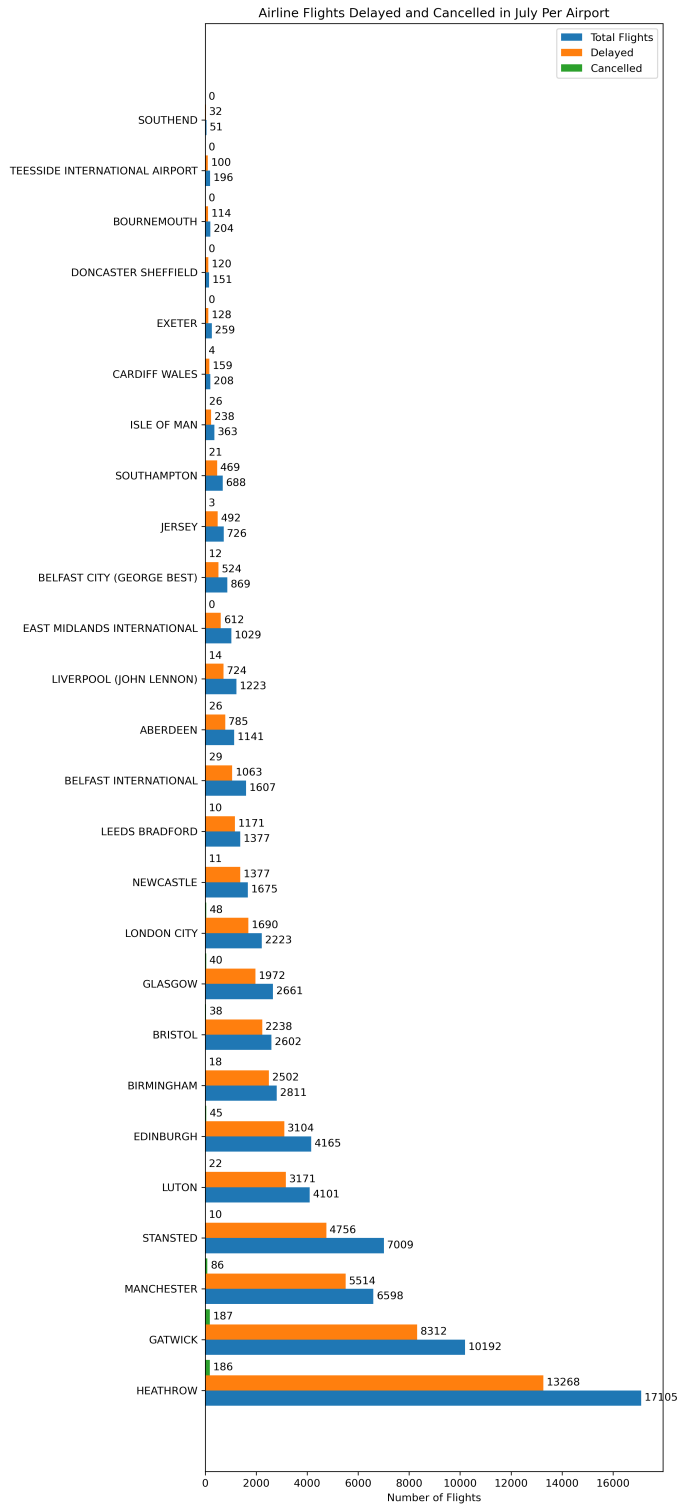
Observation :

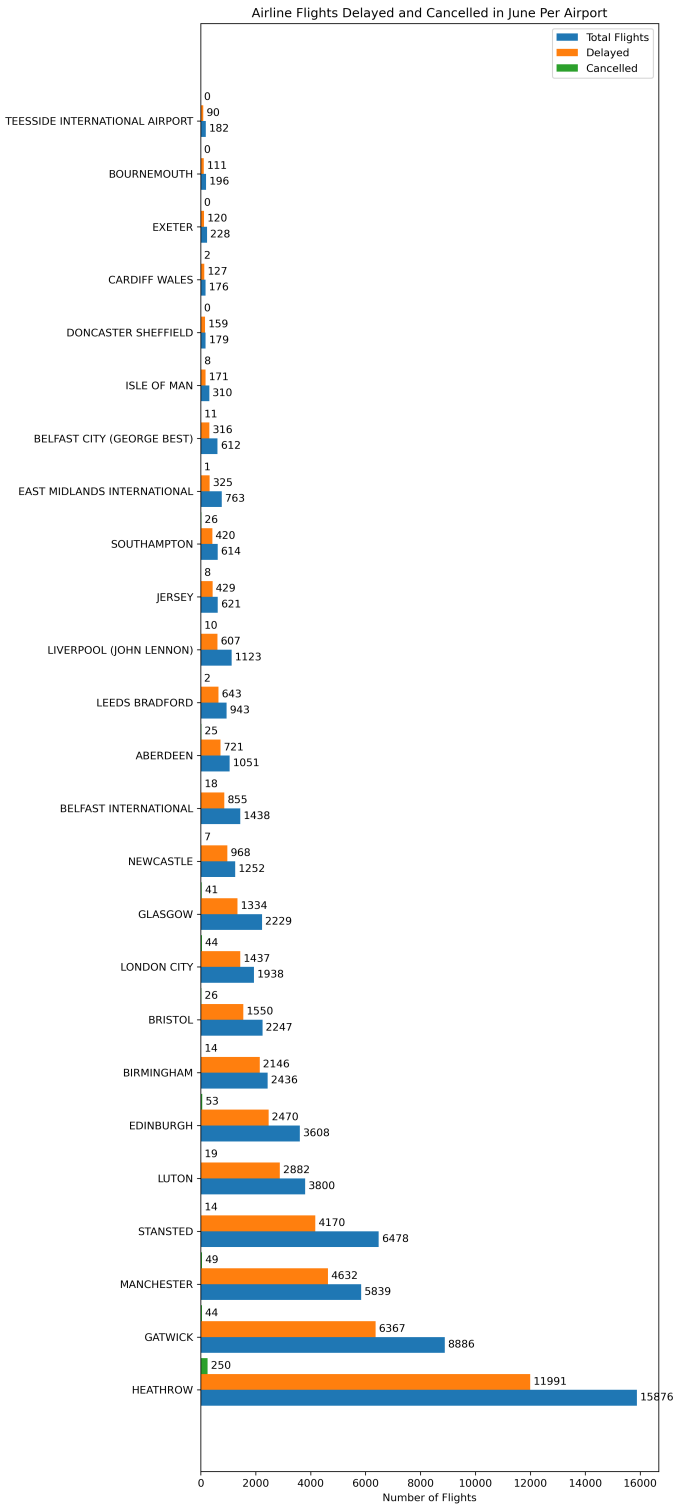
- This graph shows us the average delay in minutes of each airport
- The maximum time delay in Minutes in July was at Bristol which averaged at around 45 minutes followed by Birmingham at around 40 Minutes
- The maximum time delay in Minutes in June was at Doncaster Sheffield which averaged at around 35 minutes
- The maximum time delay had an increase of 10 minutes in July compared to June
- The minimum time delay in Minutes in July was at Teeside Internation Airport which averaged at around 10 minutes
- The minimum time delay in Minutes in June was at East Midlands International which averaged at around 5 minutes
- Maximum Change was seen in Gatwick and Cardiff
- Cardiff's time delay changed from about 17 minutes in June to about 35 minutes in July
- Gatwick's time delay changed from about 20 minutes in June to about 35 minutes in July
- We can see that Exeter,Teeside and Bournemouth had decreased time delay July comparitively to June so it's a positive change



Observation :

- This graph shows us the average delay in percentage of each airport
- The maximum percentage of flights delayed in June was at Birmingham where almost 89% of the flights were delayed
- The maximum percentage of flights delayed in July was at Doncaster Sheffield where almost 88% of the flights were delayed
- The minimum percentage of flights delayed in June was at East Midlands International where only 15% of the flights were delayed
- The minimum percentage of flights delayed in July was at Exeter where about 50% of the flights were delayed
- The notable airport with percent changes are East Midlands, Leeds and Bristol.
- Percentage of flights delayed in East Midlands was at around 15% which increased to 60% in July.
- Percentage of flights delayed in Leeds was at around 70% in June which increased to 85% in July.
- Percentage of flights delayed in Bristol was at around 70% in June which increased to 85% in July.
- Exeter, Doncaster Sheffield and Jersey had decreased percentage of flights delayed in July compared to June.





Observation :

- This graph shows us the number of cancellations and delays of each airport
- This graph shows us number of total flights from an Airport along with the number of flights which are delayed and cancelled in June and July
- Gatwick Airport had 187 cancellations in July while Heathrow only had 186 cancellations in June despite Heathrow having more number of flights
- Heathrow has more cancellation in June than in July
- This also verifies that the Pie Chart stands corrected.

Conclusion

After all the analysis, we have concluded we can say that there are severe delays of flights in the UK and some flights are cancelled as well. The maximum percentage of flights delayed in June was at Birmingham where almost 89% of the flights were delayed The maximum percentage of flights delayed in July was at Doncaster Sheffield where almost 88% of the flights were delayed Wizz air had the highest percentage of flights delayed in June. TUI Airways had the highest delay percentage of flights in July at 93% . Heathrow is the busiest airport in the UK in June and July.

References

- [1]https://www.caa.co.uk/data-and-analysis/uk-aviation-market/flight-punctuality/uk-flight-punctuality-statistics/2022/
- [2] https://www.youtube.com/c/Codanics