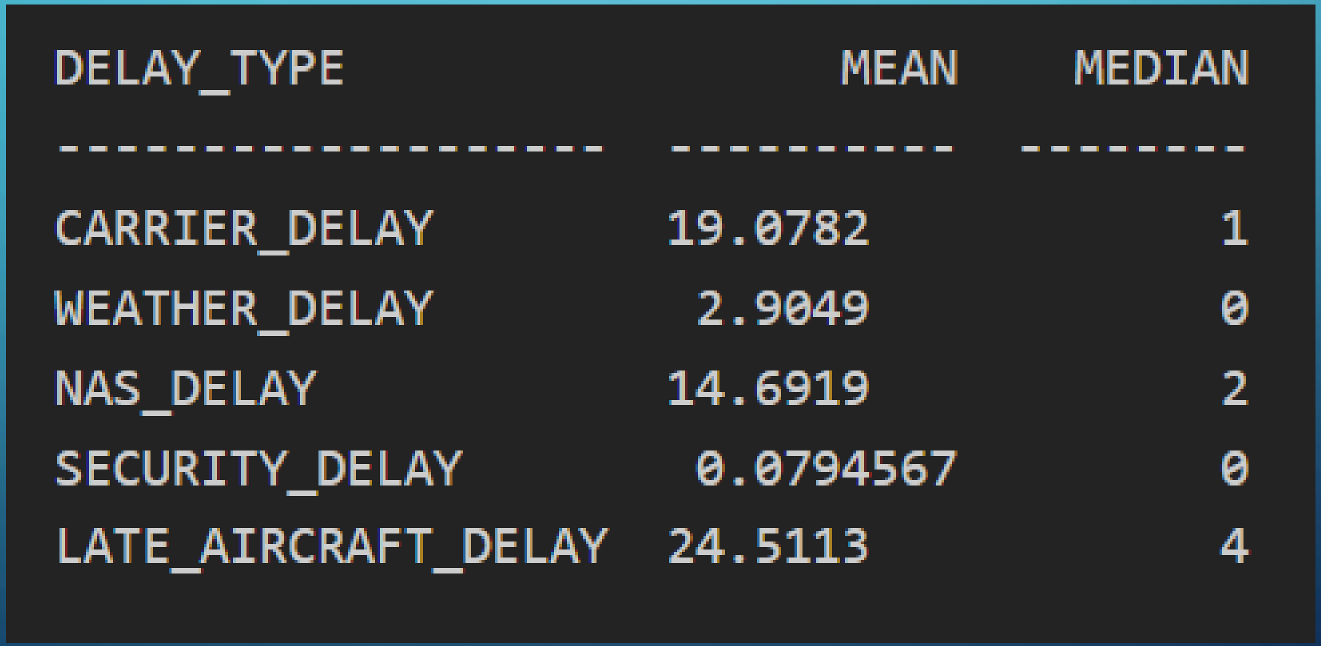
AIRLINE FLIGHT DELAY ANALYSIS

OGNJEN STAMENKOVIĆ 2022.

ALL OF THE RESULTS AND CODE ARE GIVEN AT THE FOLLOWING LINK: HTTPS://GITHUB.COM/OGIMATF/DS-AIRLINE-TRAVEL-ANALYSIS

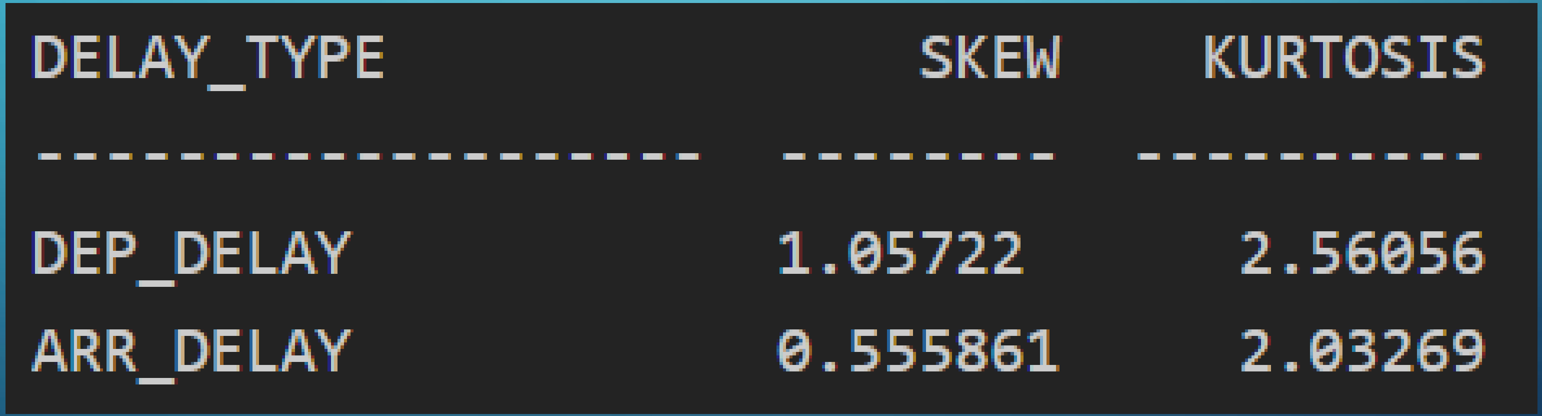
- The mean departure delay time is ~ 9.744 minutes, while the median is -2
- The mean arrival delay time is ~ 4.94 minutes, while the median is -5
- The mean and median delay times in regards to other delay types are given in the following table



```
DELAY_TYPE                 MEAN    MEDIAN
-------------------  ----------  --------
CARRIER_DELAY        19.0782            1
WEATHER_DELAY         2.9049            0
NAS_DELAY            14.6919            2
SECURITY_DELAY        0.0794567         0
LATE_AIRCRAFT_DELAY  24.5113            4
```

- The difference between the mean value and the median indicates in departure delay that there are more flights that have a negative delay rather than positive (which means that they depart early), but when a delay occurs it tends to be substantial (greatly above 9 minutes). When it comes to arrival delays time, a similar effect can be noticed. There are more delays that are lower in value, but there are higher positive values that widen the gap between the mean and the median. Both of the mean and median values are lower for arrival delay than the departure delay which could indicate that the pilots tend to try to arrive earlier.
- The value of the mean delay time indicates that the average flight arrives on time (it is considered on time if the arrival delay is less that 15 minutes)

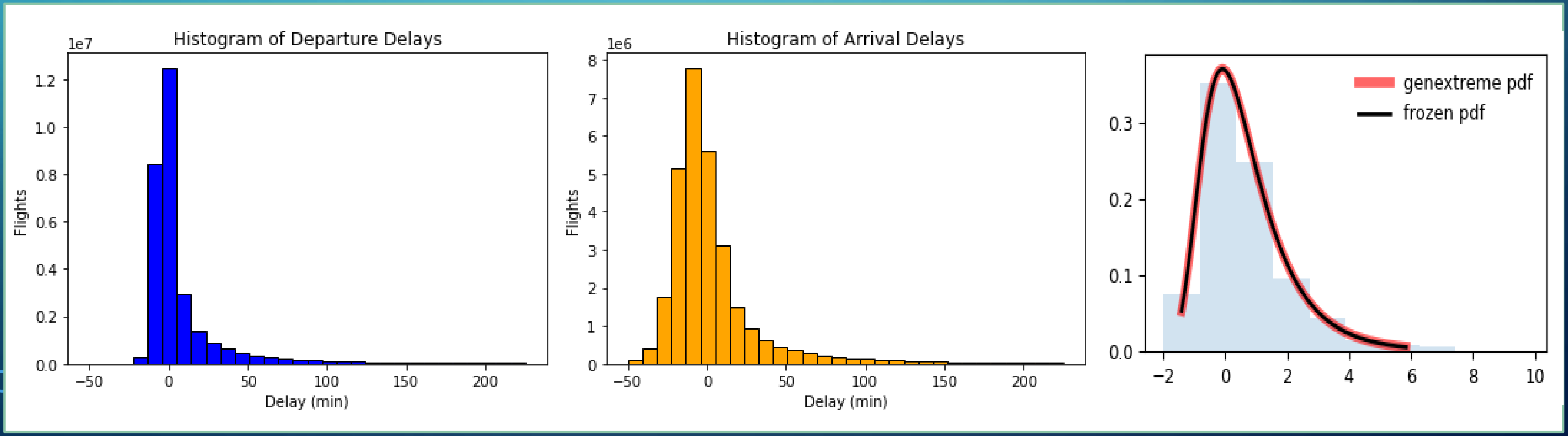
- Looking at the median values of the other types of delays, it can be said that they don't happen often, and most of them are small delays of 0-4 minutes. But judging by the mean values: CARRIER, NAS and LATE_AIRCRAFT delays can contribute a lot to the overall delay of the flights.
- The skew and kurtosis for departure and arrival delay times is given in the following table:



```
DELAY_TYPE                   SKEW    KURTOSIS
-------------------     --------    ----------
DEP_DELAY               1.05722       2.56056
ARR_DELAY               0.555861      2.03269
```

- The skewness of arrival delay and departure delay are both positive, which indicates that the data is skewed right. The tail of the distributions are longer to the right (in the direction of higher values). This means that there are a lot of flights close to the mean, but there are substantially more flights that delay an extreme amount rather than come extremely early.
- Kurtosis was calculated using the Pearson formula (kurtosis for normal distribution is 3). The calculated kurtosis for departure and arrival delay are less than 3 (between 2 and 3), which indicates negative curtosis. This means that the distributions are "light-tailed" and they have most of the values near the mean and fewer values in the tails

- To determine which theoretical distribution would fit the delay a goodness of fit test was used for different distributions. The test returns the probability that our data corresponds to the given distribution type. The highest probability distribution turned out to be the Genextreme Distribution for both the departure delay and arrival delay data. Using the example image for the Genextreme Distribution we can see that it looks coherent to the histograms of departure and arrival delays.



- The calculated average delay times and median delay times per: Year, Quarter of year, Month of year, Week of year and Day of year are given in the Jupyter Notebook that comes with my application where my solution is coded.
- The delays tend to be longer over the summer and winter (months 6-8 and 12 and 1). There are probably more flights at this time of year due to the summer and winter holidays.
- The delays also tent to be longer on Monday, Thursday and Friday. Probably due to the number of people travelling for the weekend.

- The calculated delay averages and by Carrier are given in an .csv file in the given Gitub repository
- The calculations for the average and median delay times with regards to the flight origin and destination are not provided. It took too long to calculate them and I was limited on time.
- I couldn’t provide insightful conclusions based on these calculations.

- For quantifying and identifying the main factors impacting delay I would propose using some kind of regression model
- I would train this regression model to predict the delay times based on other factors
- These models often calculate the Feature Importance of the attributes used in the prediction
- The factors with the highest Importance are most likely to impact delay times.
- I am sorry as I didn't find enough time to implement this.

I AM GREATFUL FOR THE CONSIDERATION
BEST REGARDS, OGNJEN