US NATIONAL FLIGHTS DELAY

JANUARY –JUNE 2024

Marta Nores

EDA-THE BRIDGE

2024

1.INTRODUCTION

One of the biggest frustrations these days is flight delays. In the following dataset we have information on US domestic flights between January-June 2024 with variables such as: when the flights depart, the scheduled and actual arrival time, how long they are delayed, reasons for the delay, etc.

The dataset is comprehensive, with 3,461,319 entries and 23 columns, detailing national flights within the US for the period from January to June 2024.

2.DATA

The Raw data has been taken from *Bureau of Transportation Statistics* from US Department of Transportation.

https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGJ&QO_fu146_anzr=b0-gvzr

3.HYPOTESIS

1. Flights

- Between Friday and Sunday, as is the beginning of the weekend, there should supposedly be a greater number of flights.
- The most populated cities in the US are New York, Los Angeles and Chicago in that order. The cities with the greatest number of flights should be those 3.
- June and January should be the months with higher number of flights due to seasonality (winter & christmas season and summer season), while February should be the month with lowest air traffic as there is no festive.

2. Delays

- Delays are more frequent on certain days from Friday to Sunday, when air traffic is expected higher.
- At the end of the day, flights leave with a higher number of minutes of delay, as delays from other flights are accumulated.
- More than 50% of the flights should have less than 30min. Only exceptional cases such weather delays, nas delay, carrier delay can cause a delay higher than 30min.
- Each airline has different operational or assignment processes. Therefore, there will be companies that have a higher volume of delays than others.

4.DATA PROCESSING

The data tha *Bureau of Transportation Statistics* could provide up to 120 variables related to national flights. I decided to choose only 23 that I considered interesting for the EDA (attached table below).

The data was provided on a monthly basis, so it was download 6 times, one for each month, and unified in a data frame.

VARIABLE	DESCRIPTION
MONTH	Month of the Year
DAY_OF_MONTH	The day of the month on which the flight took place, represented by an integer from 1 to 31
DAY_OF_WEEK	The day of the week on which the flight took place.
OP_UNIQUE_CARRIER	Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2). Use this field for analysis across a range of years.
ORIGIN_CITY_NAME	Origin Airport, City Name.
DEST_CITY_NAME	Destination Airport, City Name
CRS_DEP_TIME	The scheduled departure time of the flight, (local time: hhmm)
DEP_TIME	Actual Departure Time (local time: hhmm)
DEP_DELAY	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.
DEP_DELAY_NEW	Difference in minutes between scheduled and actual departure time. Early departures set to 0.
DEP_DEL15	Departure Delay Indicator, 15 Minutes or More (1=Yes)
TAXI_OUT	Taxi Out Time, in Minutes
TAXI_IN	Taxi In Time, in Minutes

CRS_ARR_TIME	The scheduled arrival time of the flight (local time: hhmm)
ARR_TIME	Actual Arrival Time (local time: hhmm)
ARR_DELAY	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
ARR_DELAY_NEW	Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.
ARR_DEL15	Arrival Delay Indicator, 15 Minutes or More (1=Yes)
CARRIER_DELAY	Carrier Delay, in Minutes
WEATHER_DELAY	Weather Delay, in Minutes
NAS_DELAY	National Air System Delay, in Minutes
SECURITY_DELAY	Security Delay, in Minutes
LATE_AIRCRAYFT_DELA	Late Aircraft Delay, in Minutes

4.DATA CLEANING

On the raw data, NANs had no value, therefore they have been replaced with zero

Secondly, I renamed the column names to lowercase, to make it easier to create formulas.

5.DATA ANALYSIS

The EDA started by doing a univariate analysis have a better understandong of the variables and their trends individually.

In the univariate analysis we used 2 types of graph based on the data type:

- For numerical data, we can use a histogram to visualize the data distribution.
- For categorical data, we will use a bar plot to visualize the number of flights of each category.

Secondly, we perform the bivariate analysis, with the “arr_delay” as target column. Several visualizations were made during the analysis.

6.CONCLUSIONS

Hypotesis: **Between Friday and Sunday, as is the beginning of the weekend, there should supposedly be a greater number of flights.**

- **Analysis** Weekend is composed of 3 days - Friday, Saturday and Sunday- It is supposed to have a higher number of flights during those days. However on the analysed data we can see that Tuesday and Friday are the highest days in terms of traffc air, while Saturday is the lowest. This discrepancy could be due to various factors, such as increased prices during the weekend or the preference for traveling on weekdays.
- **Conclusion** The hypothesis that the weekend has the highest number of flights does not hold true according to the analyzed data. We could analyse further into other factors, such as pricing and travel preferences,and see how might influence flight patterns more significantly.

Hypotesis: The most populated cities in the US are New York, Los Angeles and Chicago in that order. The cities with the greatest number of flights should be those 3.

- **Analysis** According to the data, the top 3 cities with the top three cities with the highest number of flights are Chicago, Atlanta, and Dallas. In contrast, the cities with the lowest number of flights are Moab, Gustavus, and Dillingham.
- **Conclusion** The hypothesis that the cities with the highest number of flights correspond to the most populated cities in the US is not supported by the data. Instead, other factors such as airport hub status, regional connectivity, and airline operations may play a more critical role in determining flight traffic.

Hypotesis :**June and January should be the months with higher number of flights due to seasonality (winter & christmas season and summer season), while February should be the month with lowest air traffic as there is no festive.**

- **Analysis** Base on the data , the month with higher number of flights is June, and February is the month with the lowest flights
- **Conclusion** The analysis confirms that June has the highest number of flights, aligning with the expectation of increased air traffic during the summer season. However, while February shows the lowest number of flights, this may not be solely due to a lack of festivities. The shorter duration of February (28 or 29 days) compared to other months likely contributes to its lower flight count. To make a fair comparison, future analyses should standardize the number of days across months, such as comparing flight numbers based on the first 28 days of each month. This approach would provide a more accurate assessment of seasonal trends and ensure that February's shorter length does not skew the results.

Hypotesis: Delays are more frequent on the days with higher air traffic

- **Analysis** The days with the highest number of delays are Friday, Sunday and Thursday, with Tuesday being the day of the week with the fewest delays. This data is correlated to the level of flight volume on those days. We can confirm from the data we have seen that Friday has the highest number of delays as it is the day with the highest number of flights.
- **Conclusion** The analysis confirms the hypothesis that delays are more frequent on days with higher air traffic. Days like Friday, Sunday, and Thursday, which have higher flight volumes, also report the most delays. This suggests a direct correlation between the number of flights and the frequency of delays, indicating that managing high traffic efficiently is crucial to minimizing delays

Hypotesis : At the end of the day, flights leave with a higher number of minutes of delay, as delays from other flights are accumulated.

- **Analysis** According to the data, there is a trend towards longer delays as the day progresses. Arrival delays are shorter for flights leaving early and longer for flights leaving later.
- **Conclusion** The data supports the hypothesis that delays increase as the day progresses. The trend underscores the importance of early departures for reducing delay durations and highlights the compounding nature of delays throughout the day.

Hypotesis: More than 50% of the flights should have less than 30min. Only exceptional cases such weather delays, nas delay, carrier delay can cause a delay higher than 30min

- **Analysis** Data shows that 79% of flights departure have had less than 15 min delay on the departure, while 21% more than 15 in delay. On the arrival delays , we see similar numbers, being 78% for less than 15 min delay and 22% of the flight with more than 15 min delay on the arrival.
- **Conclusion** The hypothesis is validated, as the data shows that a significant majority of flights (79% of departures and 78% of arrivals) experience delays of less than 15 minutes. This indicates that severe delays are not common and usually stem from exceptional factors like weather, NAS, or carrier-related issues. The results demonstrate that the overall punctuality of flights is relatively high, with only a small percentage of flights experiencing significant delays.

Hypotesis : **Each airline has different operational or assignment processes. Therefore, there will be companies that have a higher volume of delays than others.**

- **Analysis** According to the data, airlines with the codes **F9**(Frontier Airlines), **AA** (American Airlines), **B6** (JetBlue Airlines), are the top 3 airlines with higher number of arrival delays. On the other hand, **YX**(Republic Airlines), arrives in advance.

- **Conclusion** The analysis confirms that airlines have varying levels of delays, influenced by their unique operational processes. Frontier Airlines (F9), American Airlines (AA), and JetBlue Airlines (B6) are the top three airlines with the highest number of arrival delays, suggesting potential areas for operational improvements. Conversely, Republic Airlines (YX) often arrives early, indicating more efficient time management. This variation highlights the impact of airline-specific practices on delay performance and suggests that some airlines may benefit from reviewing and optimizing their scheduling and operational strategies.