**Analysis of NYC Flights data**

**We have analyzed the Flights dataset based on several reasons which could impact the delay such as airlines, flights, airports ,distance, month, time of day, routes, weather conditions and type of planes**

***Evaluating the delays for the different airlines***

We are merging the Airlines table with the Flights table by the variable carrier to get useful insights on the flight delays. We are calculating the average departure delay and arrival delay per airline.

* We are specifying arrival delay as positive to subset the airlines that arrive on average later than scheduled time. Also, these airlines had a positive departure delay based on which we can conclude that all airlines that depart on average with a delay also arrive with a delay.
* We are following the same procedure as mentioned above, the difference being that we don’t subset the data to include airlines with negative delays. We have not taken the average values here and this will be calculated in tableau. We are ordering by the arrival delay and check the first values which correspond to the airlines with the minimum arrival delay.
* We are checking the departure delays for airlines that arrived on time by subsetting the airlines with arrival delay = 0 and dep delay > 0. In this way we are identifying trends on the corresponding departure delays for these airlines.

Note: Average for some of the variables were not calculated in SQL because they did not give a proper insight in tableau as they had to be converted to attributes.

***Evaluating the delays depending on the destination airports and distances***

* We are evaluating the delays based on destination airports by creating a table with the fields destination (using group by), average of arrival delay and departure delay. Also, we have sorted the values based on average arrival delay to find the destination airports with maximum and minimum arrival delays.
* We are evaluating the relationship between distance and arrival delays by classifying the distance into 8 groups (equally sized). We included another variable ‘med delay’ which is the difference of arrival delay from departure delay. Positive value in the difference signify that the arrival delay is less than that expected by the departure delay (so that the airplane managed to overcome/minimize the initial delay) and vice versa. Based on our analysis, arrival delays increase for Flights with longer distance (compared to the corresponding departure delays) indicating that for longer distances Flights that leave with a certain departure delay tend to arrive even later (i.e. the difference increases in minutes).

***Changes in delays over time (month, day, time of day)***

* We evaluated differences in delays over month. We grouped the average delays (departure & arrival) by the month variable in the Flights table for different origin. The months with the higher delays were July, June, December, April and January, which correspond to the peak season and possibly to the months with higher flight traffic, passenger numbers or more adverse weather conditions (humidity, precipitation). These could be the reasons for the delays. Conversely, the months of autumn (September – November) showed almost no delays (with flights sometimes arriving before scheduled time).
* We then evaluated differences in delays based on days by grouping the average delays by day of month. The results for delays by day of month were not included in the final visualization as we thought that they do not provide great insights.
* We also examined differences in delays across the time of day. For this we used the hour variable (i.e. hour of scheduled departure). We divided the values of the variable into 4 categories: morning (5-11), noon (12-16), evening (17-21) and night (21 –5). The results indicated that evening flights were the ones with higher delays (especially departure delays but also arrival).

***Finding the worst routes (routes with highest delays) and best routes (flights that arrived earlier)***

* For this we combined the variables of origin and destination into a single variable representing the flight routes. By grouping by routes, and further sorting by average arrival delay in descending order, we derived the 5 worst routes with highest delays.
* Similarly, 5 best routes were calculated with the same variables mentioned above, the only difference being sorting was done ascending and the condition where arrival delay values were not blank.

All the 5 worst routes had the same origin airport: **EWR**

***Delays based on airports***

* To get insights on the delays per airport we merged Airports with the Flights table by origin with right join. This was then plotted in Tableau to give more insights on the delay in flights based on the airports. Our inference was that the most delay was in EWR airport.

***Evaluating the relationship with the delays and different weather conditions***

* To evaluate the reasons for the delays we first merged the Weather and the Flights tables by time hour, and origin airport with a left Join. In this way, we could derive the weather metrics for flights with higher delays across all 3 origin airports. The results indicated that flights with longer average arrival delays were linked to higher precipitation. This analysis along with other metrics have been included in our final story.
* For the second table we created a grouping variable for the departure delay based on the length of delays. The following groups were thus created: ontime, earlydeparture, shortdelay ,mediumdelay , longdelay, very long delay. The weather variables included origin, dep\_delay, temp, humid, wind\_speed, precip, visib. The output table was imported into tableau and the delays were analyzed based on Windspeed, Precipitation and Humidity. The results showed that higher the humidity, windspeed and Precipitation the longer the delay.
* The third table was designed to analyze the change in the weather conditions over months. The variables used were month, origin, dep\_delay, arr\_delay, temp, humid, wind\_speed, precip, visib and were derived from the Flights and Weather tables. The values of humidity, precipitation and windspeed were analyzed and graphically plotted in Tableau. The results showed that these values were higher especially during the month of June.

***Evaluating delays based on Planes***

We also created specific tables with summary statistics for the different manufacturers, engine types and year of manufacture. These tables were created by performing an inner join between the Flights and Planes dataset by tail number. A general table was created to give as the main insights on the variables we wanted to focus on. This was not imported in Tableau. Then 3 separate tables were created, each one including and grouping by a different variable of interest. All three tables were ordered by arrival delay descending to identify first features associated with higher arrival delays.

* First a table was created with summary statistics for average delays based on a plane’s manufacturer. We grouped by manufacturer to better examine the differences. The results indicated that planes manufactured by AGUSTA SPA tend to arrive with a higher arrival delay (30.65 mins).
* The same approach was followed to analyze differences in arrival delay as a function of engine type.
* A similar query was performed to examine differences in arrival delay based on the year of a plane’s manufacture. These results however might be correlated with engine types so that during certain years planes were manufactured with specific engine types. This consideration is illustrated in the relevant graph in Tableau.
* A final analysis was made based on seat numbers. A new variable was created that classified planes to different bins (not equally sized as we wanted to focus more on meaningful categories e.g. Small jet etc). The findings indicated longer arrival delays for planes with more seat numbers. This might be due to greater preparation time needed or simply to more passengers.