SQL & Tableau - NYC Flights Data

NYC Flights Data

Technical report

Access the Data Visualization through Tableau Public by clicking here

Team members: Fajar Tri ANGGORO, Mohammad Hadi ALIPOUR MOTLAGH & Sofie GHYSELS

Data cleaning

"Your insights and analysis are only as good as the data you are using" (Tableau, 2021). This quote implies the importance of the very first step we did in this group assignment, namely data cleaning.

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. Especially when combining multiple data sources, as is the case in this assignment, the dataset will be too cluttered and unstructured to use for data visualizations and analysis if there would be no data cleaning. There is no "one right way" to describe the exact steps in the data cleaning process because the processes will vary from dataset to dataset.
Source: ANONYMOUS, 'Guide to Data Cleaning: Definition, Benefits, Components, And How to Clean Your Data', internet, Tableau, 2021, (https://www.tableau.com/learn/articles/what-is-data-cleaning).

We did the data cleaning part in SQLite. Below are the steps that we performed in our data cleaning process:

1. We first had a look at the structure of the different datasets with the following query:

```
select *
from airlines;

select *
from airports;

select *
from flights;

select *
from planes;

select *
from weather;
```

When we browsed through the data, we already could make the following interesting remarks:

- All the dates are from one year, namely 2013
- All the flights are displayed as from 3 airports (EWR or Newark Liberty International Airport, JFK or John F. Kennedy International Airport, LGA or LaGuardia Airport) to multiple other airports.

2. In the second step, we removed irrelevant observations, we renamed certain column names from the initial datasets, and finally, we combined multiple datasets by using left join:

```
SELECT      A. Name AS Airline_Company,
            CASE WHEN dep_delay>0 THEN (SELECT dep_delay
                                        FROM flights
                                        WHERE dep_delay>0)
            ELSE 0 END dep_positive,
            CASE WHEN arr_delay>0 THEN (SELECT arr_delay
                                        FROM flights
                                        WHERE dep_delay>0)
            ELSE 0 END arr_positive,
            CASE WHEN dep_delay<0 THEN (SELECT dep_delay
                                        FROM flights
                                        WHERE dep_delay<0)
            ELSE 0 END dep_delay,
            CASE WHEN arr_delay<0 THEN (SELECT arr_delay
                                        FROM flights
                                        WHERE dep_delay<0)
            ELSE 0 END arr_delay,

            flight AS 'Flight Number',
            origin,
            dest,
            air_time,
            time_hour AS 'flight schedule',
FROM flights;
```

- We used CASE WHEN to remove irrelevant observations. For instance, in the column 'departure delay' there are positive (amount of delay) and negative values (when a flight took off early). We only want to calculate the actual delays in our visualizations, so we use a case when here to filter the initial data.
- Here we separate arrival delay into two columns based on the positive and negative numbers and did the same for departure delays.

3. In our final step, we filter the weather data.

```
SELECT  origin, temp,dewp,humid,wind_dir,
        wind_speed,wind_gust,precip,pressure,visib,time_hour
FROM weather
WHERE time_hour  IN (SELECT time_hour
                     FROM flights);
```

- We filter the weather data on origin and time_hour to have fewer rows and make our analyses easier

4. In our final step, we filter the airports data.

```
SELECT faa, name AS 'Airport Name' , lat AS latitude,lon AS longitude
FROM Airports
WHERE faa in (SELECT dest
              FROM flights)
              OR (SELECT origin
                  FROM flights);
```

We did query part 4 times, but we have trouble loading files in tableau, therefore we came with the last method which has been explained above. first, we create one big table of all data, second, we create two tables, weather and all the rest into one table, third we create three tables weather, flights, and airports. But as we discussed earlier, we had trouble loading data into tableau then we came to the final solution.

Exporting the queries out of SQL

We exported the queries from the previous Data Cleaning step out of SQLite and saved it as a CSV file.

We did this by Tools < Export < Query results < we copy-pasted the queries < select: Null values as nothing and then we saved it as a CSV file.

Data Aggregation

In general, raw data are not aggregated, it is the job of a data scientist to aggregate raw data to do some interesting analysis and ultimately provide insights from these analyses. Furthermore, depending on the audience, these analyses will have to be easily interpretable and non-bias. Some of the most used aggregation functions are:

- SUM, sum simply aggregates the data by adding all the values within the raw data. In some cases, SUM could be useful because of its easiness to interpret. In our case, we didn't use SUM because one of the downsides of using SUM is that it is sensitive to how many values are there within the data. For example, when grouping our data, groups that have more data will have higher SUM, this is of course makes our analysis non apple to apple comparison.
- COUNT, count will display how many instances are available within the raw data. Much like SUM, one of the downsides is that it is sensitive to how many values are there within the data. In our case, we used COUNT in order to support our analysis, but we didn't use it in our visualization. COUNT is also useful when aggregating non numeric data.
- AVERAGE, average is the result of adding all the values within the raw data divided by the number of instances within the data. In general, AVERAGE is the most common data aggregation method used. It is easy to interpret for general audience, and it is not sensitive to how many values are there within the data. The downside of using AVERAGE is that it is very sensitive to outliers.

- MEDIAN, median takes the middle value of a sorted data. MEDIAN is usually used as an alternative to AVERAGE, because it is not sensitive to outliers. The downside is that MEDIAN is not as easy to understand as AVERAGE, general audience will immediately understand what an AVERAGE is, but might not so in the case of MEDIAN.
- MIN/MAX, minimum/maximum will take the lowest/highest value from the data. The downside with MIN/MAX is that it only considers a single value from the whole dataset.

In our case of flights data, where delays and earlies are displayed in minutes, we used AVERAGE in most of our visualization for easiness of interpretation. We also used MEDIAN, COUNT, and MIN/MAX in order to support our analysis.

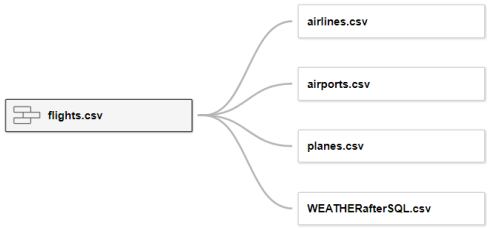
Visualizations and analysis in Tableau

In the Tableau program, we added the CSV files by uploading them as a text file.

In the Tab "Data Source" we first made sure that the connection and data interpreter were correctly set:

- On the right top, we clicked on Connection < Extract. This is to make sure that other people can open the file and extract the data source
- On the left we clicked on Data interpreter < Automatically filters the data, this is to make sure that Tableau already filters the data.
- Here we change the name of columns that we create for the separation of delays and early arrivals, (pure arr delay: the amount of time departing after schedule, pure dep delay: the amount of time arriving after schedule, pure early arrival: the amount of arriving before schedule, pure early dep: the amount of departing before schedule). then, we turn all the numbers to positive for better calculation and presentation, since the tableau will interpret the negative numbers as less value, then we create the new column of total pure delay and total pure early. (sum of arrival and departure).
- For showing the routes and connection of two points on the map we use tableau and make a union between using two flights table, then we create three columns to rout identifier, route location, and rout order. And use the route order formula to join this table to the airports.

In the Data Source itself, we made the relations between the tables. Our Data Model has 5 tables in total: airlines, airports, planes, weather and flights.

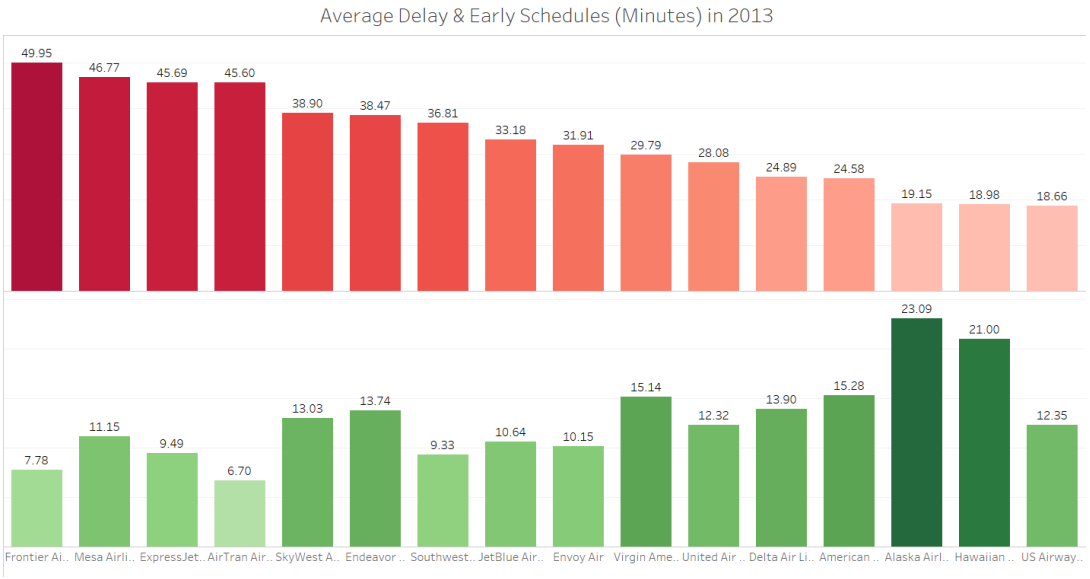


Below, we give an overview of the worksheets, dashboards, and story that we made. We always first explain how we made the visualization and then show a picture of the end results. Below every visualization, we give our analysis.

Worksheet 1: Evaluating the delays and early arrivals for the different airlines:

In the first worksheet 'Delays & Earlies per Airline', we calculate the average amount of delays and early arrivals per airline company in 2013 (remember from the first cleaning data step that we only have data from 2013). Note that the amount of delay and of early arrivals is noted in minutes. The original data is in minutes, so no data transformation is done within the time units. We use a bar chart here because we want to compare 2 values (average delays and average early arrivals) across different subgroups of data (the different airline companies).

In the columns, we put the names of the airlines and in the rows, we put the total pure delays and total pure early. Pure delay means it's just positive values (delays) that are retained, pure early means it's just negative values (earlies) that are retained, these negative values are transformed to positive value to adjust calculation. We changed the measured value to average and adjusted it from "standard" to "fit width". We renamed the title of the histogram via right-click on the top of the histogram.



The airlines that perform the worst are Frontier Airlines and AirTran Airways Corporation. Frontier Airlines has the most delays and doesn't have many early arrivals. AirTran Airways Corporation also has a lot of delays (it takes 4th place) and it has the lowest average early arrivals, 6.70 minutes on average.

The airline company with the most average delays, Frontier Airlines, doesn't necessarily have the lowest average of early arrivals. The best performing airline company in this bar chart is Alaska Airlines with the most average amount of early arrivals and the third least average amount of delays. An important side note is that Hawaiian Airlines is the second-best performing airline company, which is no surprise

because, in 2020, Hawaiian Airlines was named the most punctual airline in the U.S. for the 16th straight year (Fortune magazine, 2020).

Source:

ASSOCIATED PRESS, 'Hawaiian Airlines named most punctual airline in U.S. for 16th straight year', internet, Fortune, 2020-02-23, (https://fortune.com/2020/02/23/hawaiian-airlines-most-punctual-airline/).

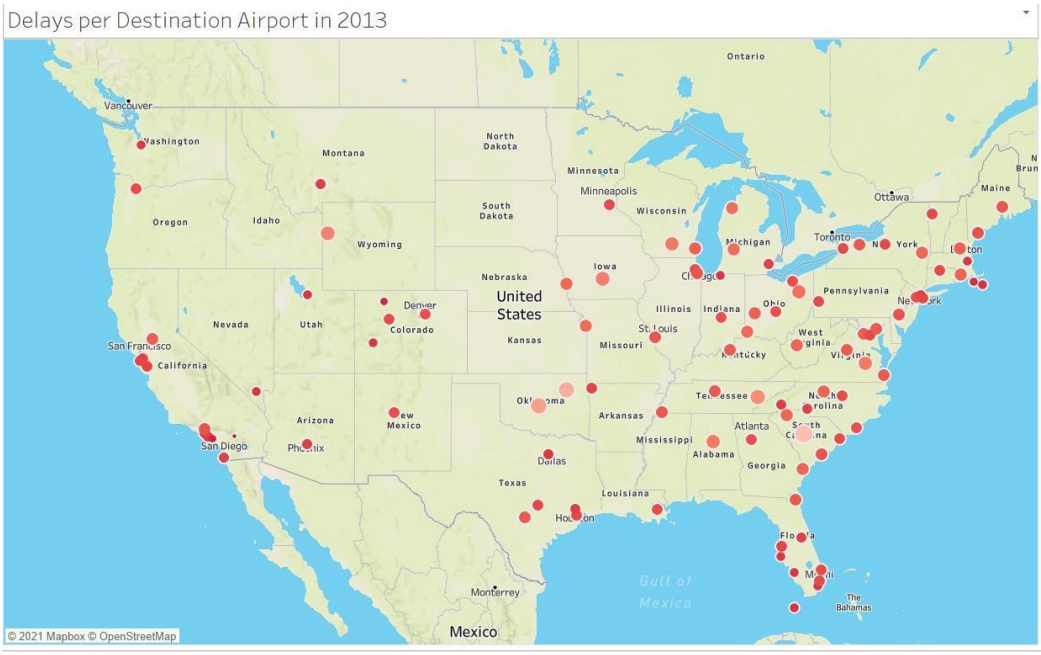
Worksheet 2: "Evaluating the delays depending on the destination airports

In the second worksheet 'Delay per destination airport', we display all the delays per destination airport in 2013 on a map.

We used a map here because this is the best way to visualize all the destination airports in this case. On top of that, the map is interactive so for instance, the user can zoom in on the southwest part of the US.

In the columns, we put lon or the longitude and in the rows, we put the lat or the latitude.

We have put pure_arr_delay in the filter so that the airports that do not have any delays are not shown on the map. We used the average pure_arr_delay and dragged this into details (so that the details of the information of a specific airport are shown when a user goes over the boll), into size (so that the bolls of airports with a greater average of arrival delay are larger) and finally also into color (we edited the color to red to have the bolls clearly displayed on the map). On the right bottom of the map, there was "1 NULL" displayed, we clicked on this and then filtered the null values do not have any null values on the map.



This map is a great way to analyze all the airports in a particular region. Let's say for instance, that we analyze all the airports around San Francisco. There are 4 airports in total around that large American city. We can conclude, by simply going over all the 4 bolls, that Sacramento International Airport has the highest amount of average delay (21.19 minutes) and the other 3 airports (San Francisco International, Norman Y Mineta San Jose International Airport, and Metropolitan Oakland International Airport) all have an average delay of around 15 minutes. So, when we would fly to San Francisco, we would have the highest probability of a delay if we land at Sacramento International Airport. This analysis is something to keep in mind if you would ever plan a trip to San Francisco!

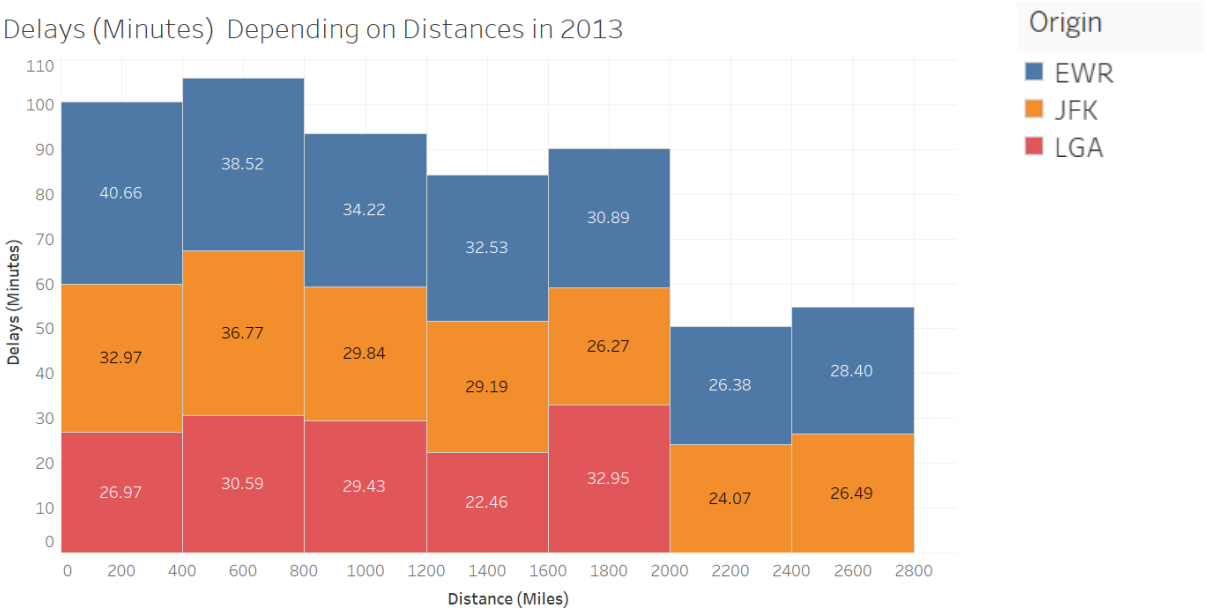
Worksheet 3: "Evaluating the delays depending on distances"

On the third worksheet 'Delays Depending on Distances', we also evaluate the delays (as on the previous worksheet) but now we look at the relationship between delays and the distances of the flight routes.

In the columns we put distance, and, in the rows, we put the average of the total_pure_delays. We converted the field distance from discrete (this is the default in Tableau) to continuous. The fields in green are continuous and the fields in blue are discrete. We also changed added bins to distance via creating and adding the bin to 400.

We filter by origin because this data only has 3 origins, and distance is always related to origin - destination. We also dragged origin into color to change the colors that are displayed in the histogram. After that, we made sure that the different parts of the histogram are put on fit width.

As a final step, we excluded the outliers, by going over them with our mouse and clicking on the right mouse < exclude, this to make the visualization shorter and easier to interpret.



As we can see from the graph, we can conclude that higher distance routes don't have higher delays and that flights with shorter distances tend to have the most delays. We do have to add an important reflection here that there can be some sort of bias. Maybe shorter distance flights have a more probability of delay because there are more shorter flights than longer flights?

Worksheet 4: "Evaluating changes in delays over time"

On the fourth worksheet, 'Delays Over Time', we display both the arrival and departure delays in minutes for every month in 2013.

For this analysis, we used a line graph, because we want to display changes over a short period of time (in this case 1 year) and because we compare changes over the same period for more than one group (her we have 2 groups, namely both arrival and departure delays).

In the columns, we put time hour and we converted this to month. In the rows, we have put the average of the pure_arr_delay and pure_dep_delay.

We dragged 'names' into color to change the color of our 2 lines to orange and blue and we synchronized the axis to make sure that both pure_arr_delay and pure_dep_delay have the same intervals.

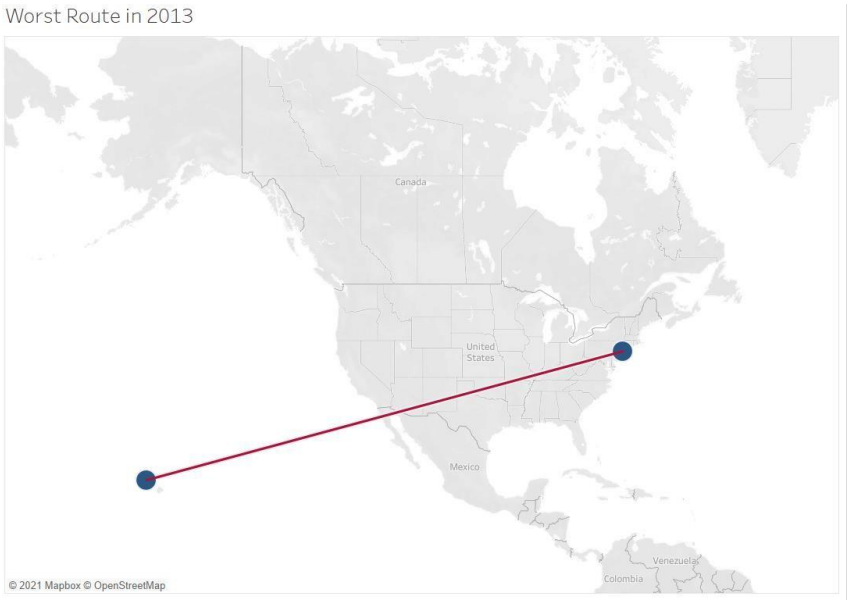
This line graph can tell us what the worst month is to fly in the skies. We first did some internet research and found out that in the US, December, due to a flood of holiday travelers and inclement weather (CBS News, 2008) and July and August (Uniglobe, 2020) are said to be the months with the highest airlines. Our line graph confirms our previous internet research: the months with the most delays are July and December. What is however interesting is that there is already a peak in June.

Sources:
ANONYMOUS, 'All About Flight Delays', internet, Uniglobe, 2020-01-27, (https://www.uniglobe.com/blog/all-about-flight-delays)
HAEBER, J., 'Which Month Grabs the Most Flight Delays?', internet, CBS News, 2008-02-06, (https://www.cbsnews.com/news/which-month-grabs-the-most-flight-delays/)

Worksheet 5: Plot the worst routes or the routes with the most delays

In this worksheet, we display the worst route in 2013.
We filter the result based on maximum delay and choose the highest number then we find the route with the highest amount of delay, for arrival, departure and total is JFK to HNL.

We changed the background to light to not let the user be distracted by other things on the graph via Format < Maps < Background < Light.



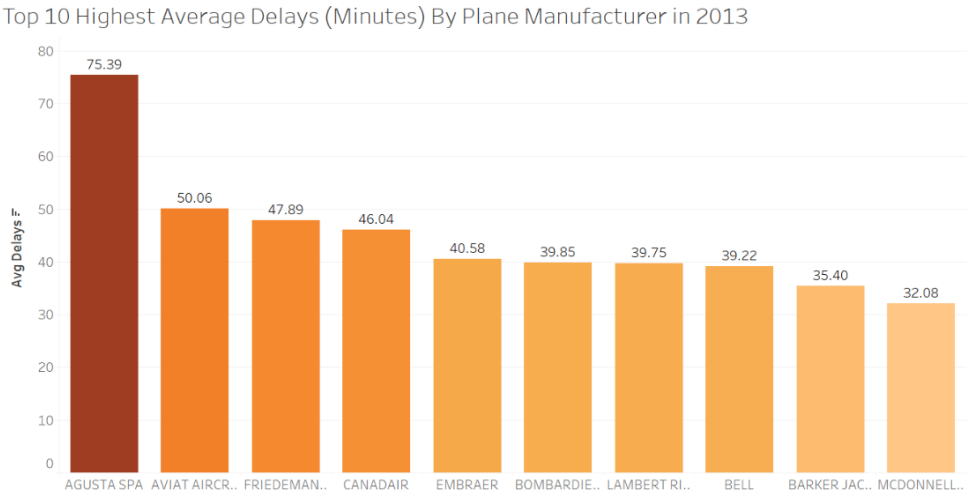
We can conclude from our visualization/story that the worst route or the route with the most delay is the route from JFK, John F. Kennedy International Airport, which is in New York to HNL, or Daniel K. Inouye International Airport, which is in Honolulu on the island of O'ahu in the state of Hawaii. This route is a long-distance route because it takes 8 hours and 6192.58 km in total. This finding is rather a surprise because we calculated in worksheet 3 the delays depending on distances and found out that shorter-distance routes have a higher probability of delay.

Worksheet 6: Which manufacturer has the most delays?

In the sixth worksheet, we display which manufacturer of the airplanes has the most delays.

In the columns we put the manufacturer, and, in the rows, we put the total_pure_delays. We used average because count is sensitive to total values i.e. boeing will have the larger probability of having a delay because it is the most used manufacturer.

We excluded null values and removed some manufacturers who had very small bars. These manufacturers only had little delays, so it is no use to take them into our analysis. Therefore, in the graph below, these are the top 10 highest delays airplane manufacturers displayed.

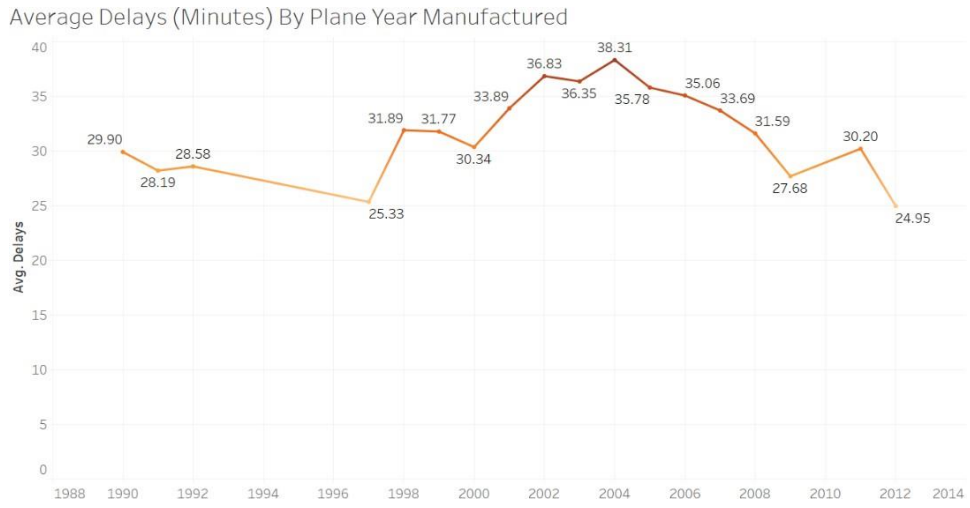


From the graph above, we can see that AGUSTA SPA has the highest average delay in 2013.

Worksheet 7: Do newer planes perform better?

In the seventh worksheet, we display a line graph that evaluates whether the year of the plane manufactured affects the delay

In the columns we put the plane year manufactured, and, in the rows, we put the average total_pure_delays.

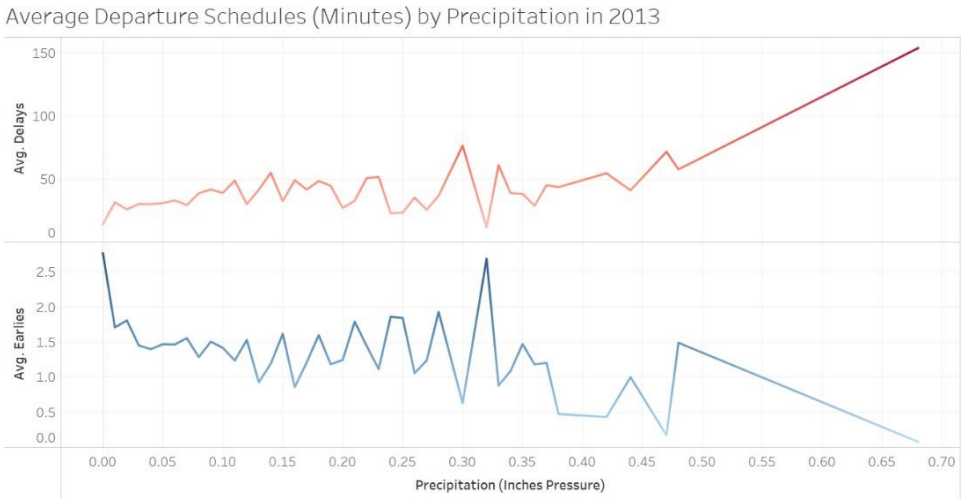


Interestingly, older planes (more than 15 years old) perform better in terms of delay compared to the less old planes (less than 10 years old). The trend is increasing at first, and then decreasing later. We might conclude that newer planes (less than 3 years old) perform better compared to the older ones, but it is not always the case. It is also worth mentioning that less old planes (less than 10 years old) have a significantly higher number of flights in this dataset, so the probability of having a delay is larger compared to the older ones.

Worksheet 8: Does rain (precipitation) affect delays?

In the eighth worksheet, we display a line graph that shows the correlation between departure schedules (delays and earlies) and precipitation.

In the columns we put the precipitation, and, in the rows, we put the average departure delays & earlies. We only take into account departure schedules because the weather data is only recorded in the origin airports.

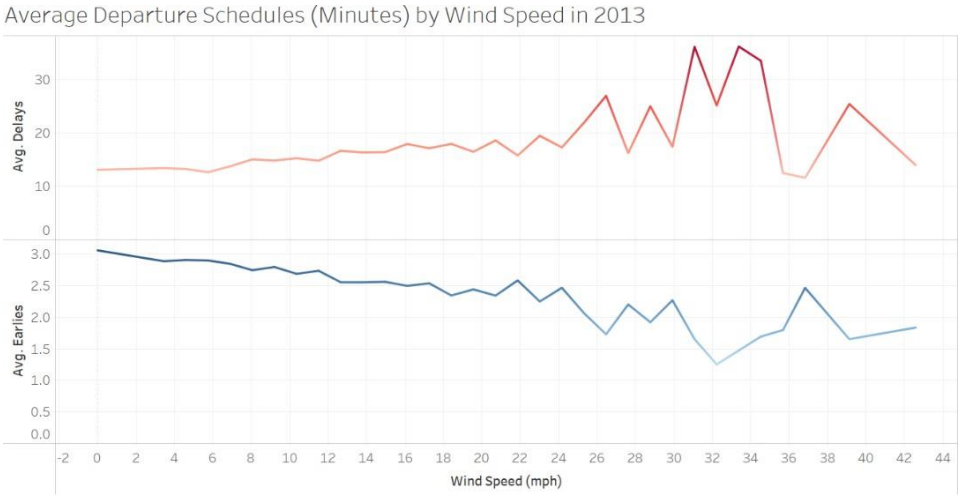


In a common sense, rain (precipitation) will affect delays i.e. a flight will be delayed due to bad weather. Our data confirm this, we take a look at both delays and earlies and it can be seen clearly that there is an increasing trend for the delays, and a decreasing trend for the earlies.

Worksheet 9: Does wind speed affect delays?

In the ninth worksheet, we display a line graph that shows the correlation between departure schedules (delays and earlies) and wind speed.

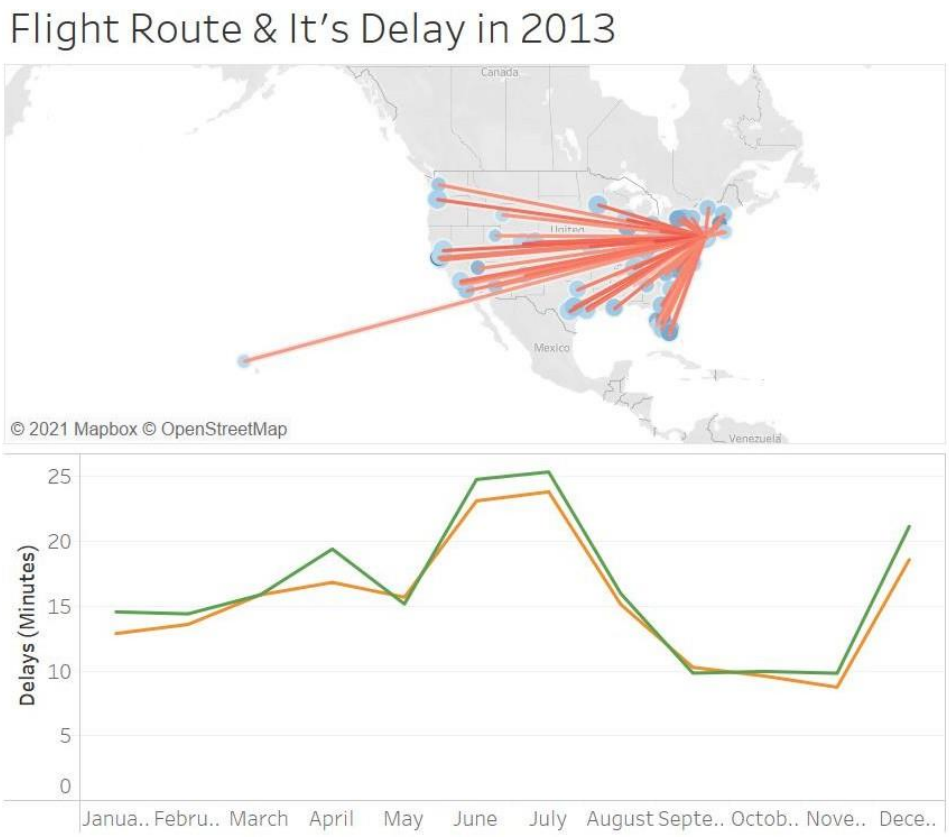
In the columns we put the wind speed, and, in the rows, we put the average departure delays & earlies. We only take into account departure schedules because the weather data is only recorded in the origin airports.



We take a look at both delays and earlies and our data shows that there is an increasing trend for the delays, and a decreasing trend for the earlies.

Dashboard 1: Flight Route & It's Delay in 2013

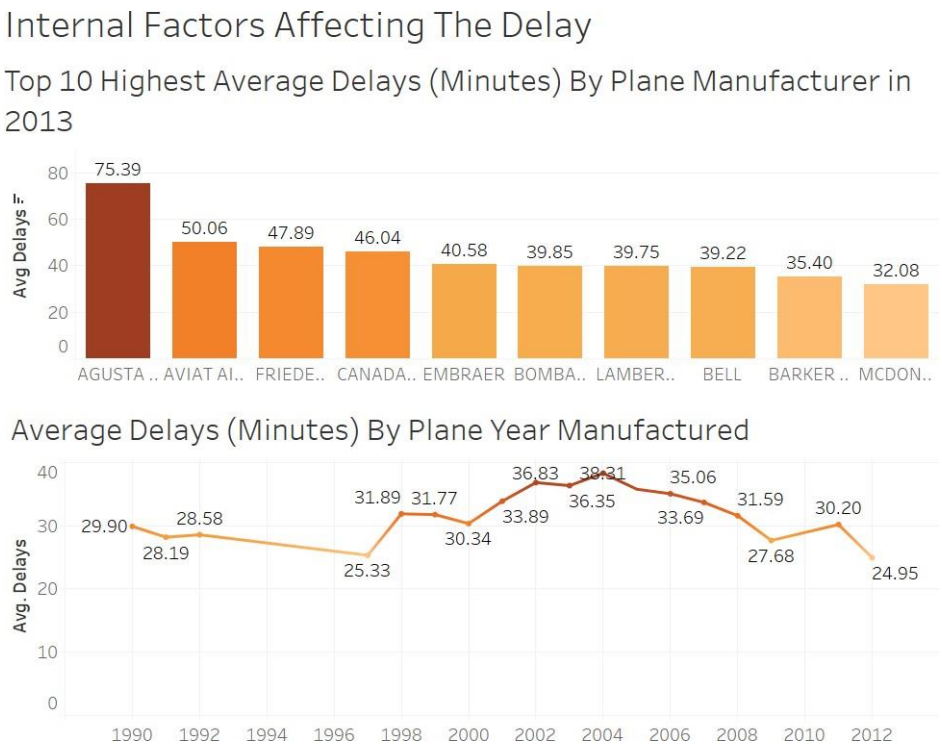
In this dashboard, we combine the flight route worksheet with the delays over time worksheet. The goal is to evaluate the delays over time within a certain route and also to see which route has the highest average delay within a certain month.



This dashboard provides insight on the route with the highest delay and its evolution within months. For example, if a route has a higher average delay, we can delve deeper into its evolution within different months to see whether if the delays are seasonal. It might be common to have seasonal delays (for example during holidays), but if certain route continuously has high delays, one might want to delve even deeper to see if there's any problem within the origin or the destination airport.

Dashboard 2: Internal Factors Affecting the Delay

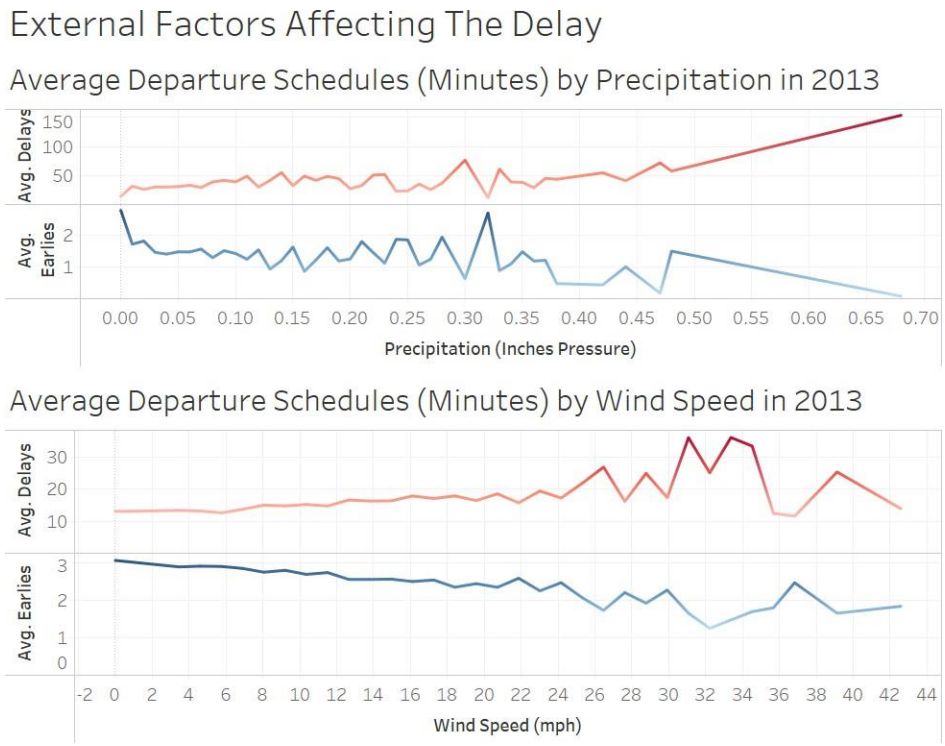
In this dashboard, we combine the delays by plane manufacturer worksheet and the delays by plane year manufactured. The goal is to see the correlation between the internal factors of a flight (Planes & Engines) and the delays.



This dashboard provides insight on the correlation between internal factors (in this case Plane Manufacturer and the year the plane was manufactured) and the delays. While correlation does not imply causation, it is worth mentioning that these 2 variables play a factor in affecting the delay.

Dashboard 3: External Factors Affecting the Delay

In this dashboard, we combine the delays by rain (precipitation) worksheet and the delays by wind speed. The goal is to see the correlation between the external factors of a flight (Weather) and the delays.



This dashboard provides insight on the correlation between external factors (in this case precipitation and wind speed) and the delays. This dashboard also confirms a common sense that bad weathers are affecting delays in a flight. Another thing worth mentioning is that while weather is a common cause of delays, another factor could play a role as well.

What have we learned after these visualizations and analyses?

While doing this group assignment we have come to the following interesting insights, which we will use in our future careers as data analysts/scientists:

1. The first step, data cleaning is an important step, but also a very time-consuming step. One-third of our time was spent on the data cleaning part.

2. As a data analyst/scientist one must be aware that the graphs/visualizations that we make can have a biased undertone and not fully represent the truth.

3. We also learned the following interesting and very useful tips from our data analysis which we can implement ourselves when we book our next vacation and pass on to friends and family:
    a. Don't fly with Frontier Airlines – they have the most average delays. On the contrary, the best airlines to fly with, in terms of punctuality, are Alaska Airlines and Hawaiian Airlines.
    b. Have a look at our map to see for each specific region (for instance the San Francisco region) which airport performs best.
    c. Shorter distance routes tend to have the most delays. So, it's better to stay in the sky longer, than to wait longer in the airport lounge.
    d. Do not fly in June, July, August, or December. From September till November is the best time to be in the skies because then there is the lowest number of delays.
    e. The worst route or route with the most delays in the US is from JFK in New York to Daniel K. Inouye International Airport, HNL, in Honolulu, Hawaii. We strongly recommend not fly on this route.
    f. If you step on an airplane that is made by Agusta Spa, then the chances are high your flight will have a delay. Therefore, it could be useful to call the airline company and ask who has manufactured the airplane.