Group 16
Nour Azar
Ahmad Omar Nakib
Mario Serrano

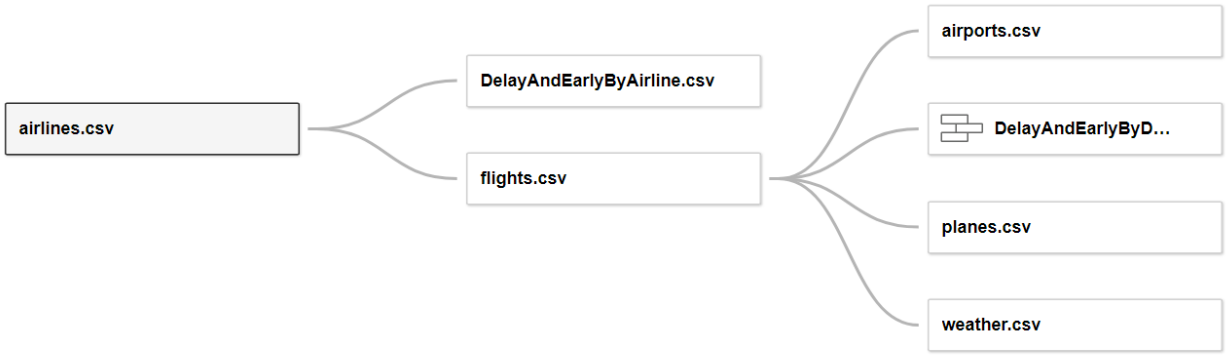
Group Assignment Business Reporting Tools

NYC Flights Airport Delays

According to Statista, the number of flights increased steadily from year 2000 until 2019 (before the pandemic) to reach 38.9 million flights, increasing the chances that more flight delays will occur. Flight delays constitute a burden both on the airline and the passenger, and according to Berkley News more than half of that cost is borne by the passengers. In fact, the collective impact of these costs on the US economy was estimated at 32.9 billion dollars. These delays are caused by many factors, including the weather conditions, security issues, technical issues with the aircraft etc. In this paper, we will specifically examine the flight delays in New York airports in 2013 and evaluate the reasons behind them by analyzing the dataset obtained.

In addition to the tables provided for us from the dataset, we created two tables by executing queries on SQL lite, we called the first DelayAndEarlyByAirline and the second DelayAndEarlyByDestination. Later in this report we will explain how we created these tables and for which purpose.

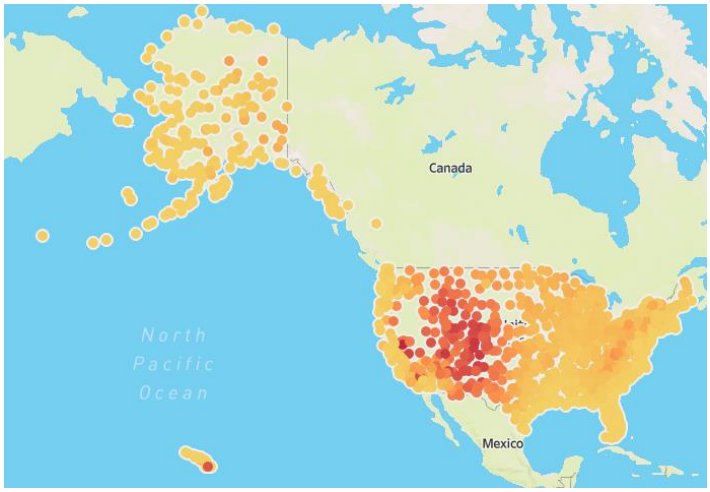
In order to create a relationship between the tables, we created connections between them as shown below:



- Airlines.csv to Flights.csv: Carrier reference column
- Airlines.csv to DelayAndEarlyByAirline.csv: Airline name column
- Flights.csv to airports.csv: Origin = Faa and Dest = Faa
- Flights.csv to delayAndEarlyByDestination.csv: Dest = Destination
- Flights.csv to planes.csv: Tailnum = Tailnum
- Flights.csv to weather.csv: Time Hour = Time Hour, Year = Year, Month = Month, Day = Day, Origin = Origin

1- Airports:

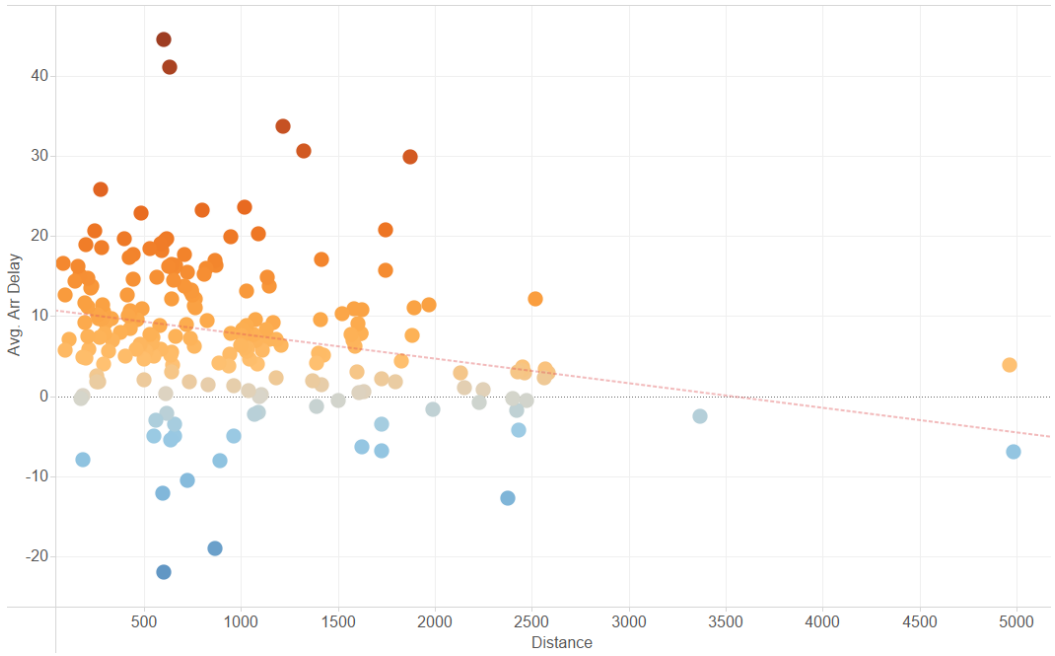
We examined the different airports in the United States that were listed in the dataset, we defined the longitude as the independent variable and the latitude as the dependent variable and we obtained the following map:



The colors differ according to the average altitude of the airport (in feet): the higher the altitude, the darker the color.

2- Distance:

From the Flights table, we extracted two variables: the distance between the departure and the destination, and the arrival delay for each flight. We defined the distance as the independent variable and the average arrival delay as the dependent variable and accordingly we generated a scatter plot.



We noticed that there is a weak negative correlation between the variables in the graph above. We can say that most of the delays happen during shorter flights, as longer flights might be able to correct the delay during the travel time. Usually, shorter flights have are more frequent, which turns out in a busier schedule for the airports, hence more delays.

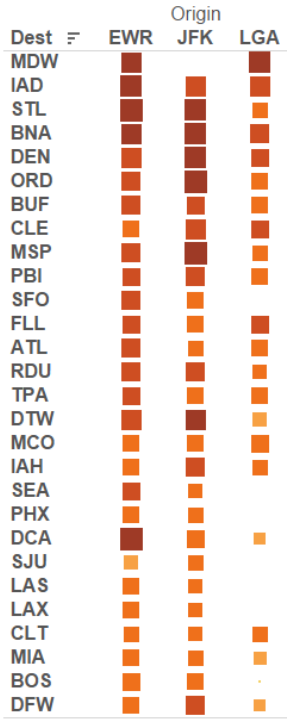
3- Engine

From the Flights table and the Planes table, we extracted two variables: the engine type and the departure delay. We defined the average departure delay as the dependent variable and the engine type as the independent variable and accordingly we generated a scatter plot.

The darker the color, the higher number of engines for each engine type. We noticed that the engines 4 Cycle and Turbo-shaft associate with the highest average of departure delays.

4- Routes

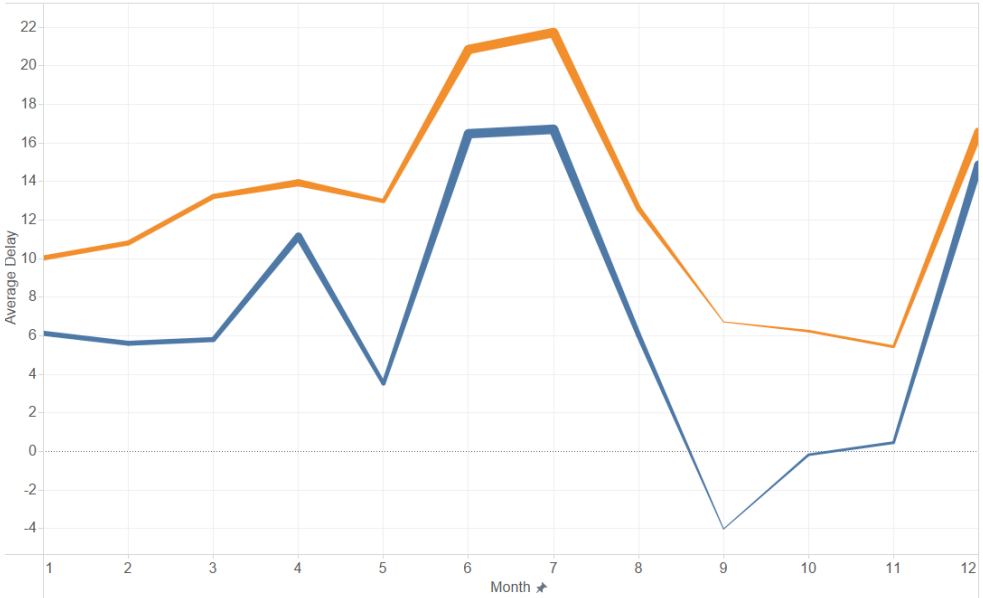
From the Flights table, we extracted two variables: the departure airport and the destination airport. We defined the departure airport as the independent variable and the destination airport as the dependent variable and accordingly we generated the following graph:



The darker the color and the bigger the shape, the higher average of departure delays. The list is ordered from highest average time of delay to lower, which means that some of the worst routes are from Newark Airport to Chicago Midway, Dulles International and St. Louis Lambert International.

5- Average delays per month

From the Flights table, we extracted three variables: the departure delay, the arrival delay and the month. We defined the departure delay and the arrival delay as the dependent variable and the month as the independent variable and accordingly we generated the following graph:



The red line represents the average departure delay while the blue line represents the average arrival delay. We noticed from the graph that for all months, the average departure delay is always higher than the average arrival delay, and that the delays increase in summer (June and July) and on Christmas holidays (December).

6- Delays per Airlines

We created a graph that includes all the airlines in the dataset and their respective total number of delays, and we obtained the following:

We noticed that United Airlines Inc. have the highest number of delays, followed by ExpressJet Airlines Inc., with only a slight difference between them.

7- Delay and early Destination

We created a new table from a query executed using SQLlite, where we selected from the Flights table the destination airport column, the count of early and on-time arrivals and the count of late arrivals, sorted by the destination airport. We then calculated the percentage of the early/on-time arrivals to the total flights per airport.

```
-- Checking the count of delays by airline

select carrier,airlines.name, count(*) as delays from flights
left join airlines on flights.carrier = airlines.carrier
left join airports on airports.faa = flights.origin
left join planes on planes.tailnum = flights.tailnum
where flights.dep_delay <= 0
group by flights.carrier , airlines.name
order by delays desc;

-- Checking the count of early and on time arrivals by airline, manufacturer

select airlines.name, manufacturer,count(*) as early_and_on_time from flights
left join airlines on flights.carrier = airlines.carrier
left join airports on airports.faa = flights.origin
left join planes on planes.tailnum = flights.tailnum
where flights.dep_delay <= 0
group by airlines.name, manufacturer
order by early_and_on_time desc;

-- Checking the null data

select * from flights
left join airlines on flights.carrier = airlines.carrier
left join airports on airports.faa = flights.origin
left join planes on planes.tailnum = flights.tailnum
left join weather on weather.time_hour = flights.time_hour
where planes.manufacturer is null;

-- the count

select airlines.name, planes.manufacturer, count(*) as delays from flights
left join airlines on flights.carrier = airlines.carrier
left join airports on airports.faa = flights.origin
left join planes on planes.tailnum = flights.tailnum
where flights.dep_delay > 0
group by airlines.name, planes.manufacturer
order by delays desc ;


--departure delay <= 0

select flights.dest, count(flights.dep_delay) early_and_on_time from flights
left join airlines on flights.carrier = airlines.carrier
left join airports on airports.faa = flights.origin
left join planes on planes.tailnum = flights.tailnum
where flights.dep_delay <= 0
group by flights.dest
order by early_and_on_time desc;

--departure delay >= 0


select flights.dest, count(flights.dep_delay) early_and_on_time from flights
left join airlines on flights.carrier = airlines.carrier
left join airports on airports.faa = flights.origin
left join planes on planes.tailnum = flights.tailnum
where flights.dep_delay > 0
group by flights.dest
order by early_and_on_time desc;
```

The following graph illustrates what we obtained after defining the independent variable as the destination airport and the dependent variable as the percentage of late and on time arrivals.

The blue color represents the percentage of on-time flights and the red color represent the percentage of delayed flights.

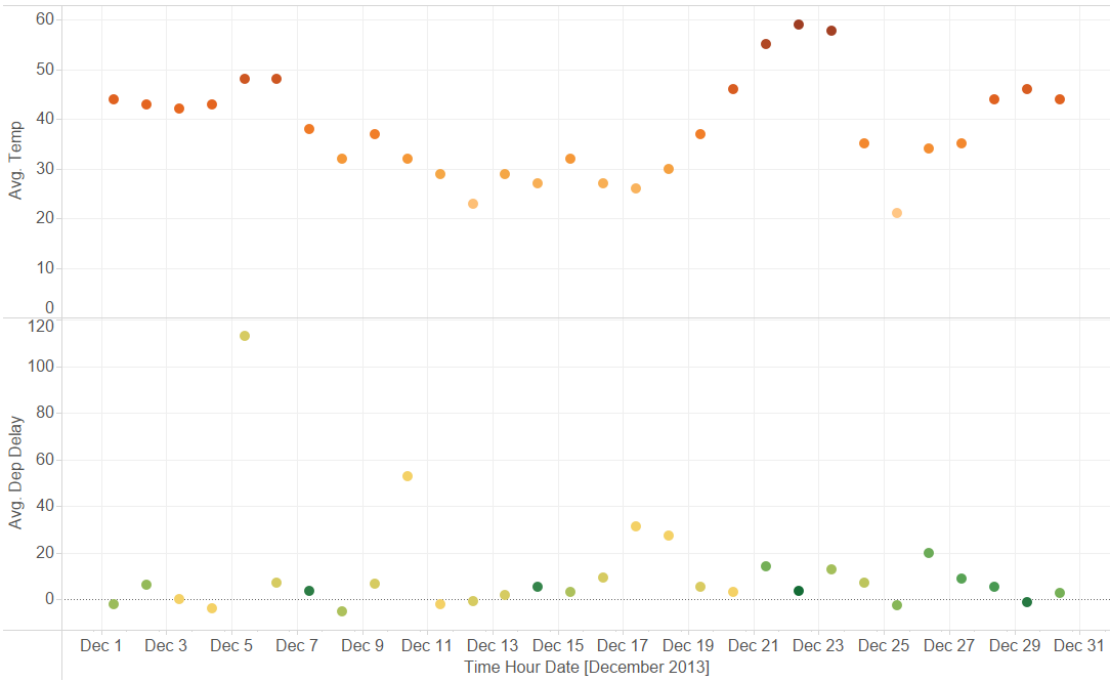
The line in the middle separates the graph into two parts: the part above includes the late and on time percentages that are above 50% and the part below includes the percentages below 50%. We notice that the late percentage is lower than the on-time/early percentage for all airports except for MDW (Chicago Midway airport) and Den (Denver aiport) where the late arrivals exceed the on-time and early arrivals, especially for MDW where the difference is almost 10%.

8- Airline manufacturer

We created a new table from a query executed using SQLlite, where we selected the airline name, the manufacturer name, the count of departure delays and the count of Early and on time flights by using the left join function to join both the Flights table and the Airlines table. We then calculated for each airline the percentage of the early/on-time flights to the total early/on-time flights and the percentage of delays to the total delays for all airlines. The generated a graph after defining the independent variable as the sum of percentage of delayed flights and the dependent variables as the airline and manufacturer name.

9- Temperature:

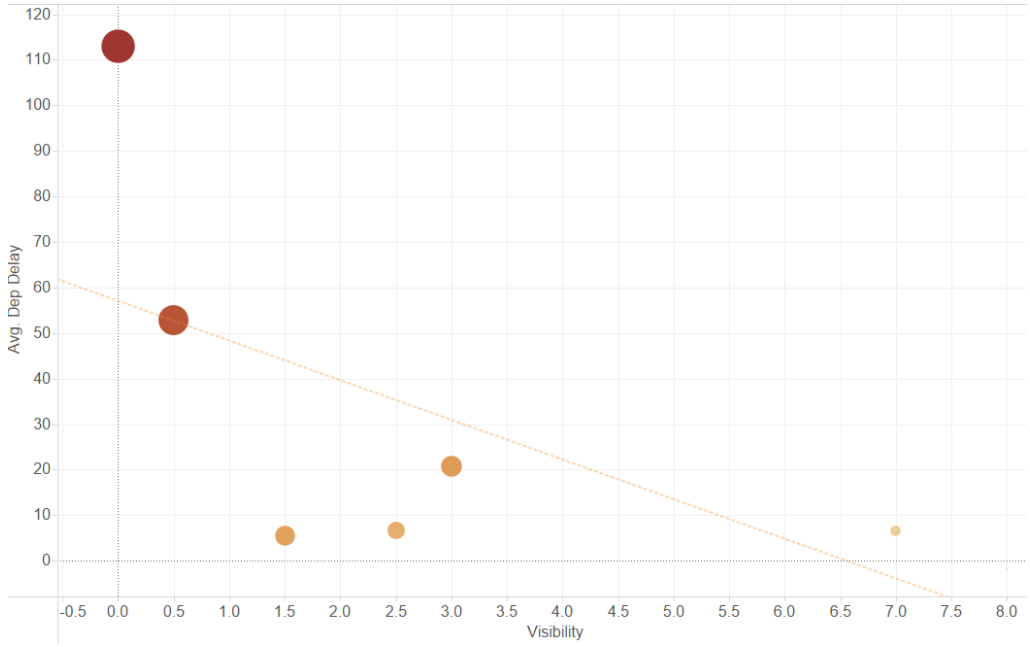
From the Flights table and the weather table, we extracted three variables: the departure delay, the temperature, and the time hour date. We defined the departure delay and the temperature as the dependent variables and the time hour date as the independent variable and accordingly we generated the following two graphs:



We can play around with this graph because it has multiple filters. The upper graph shows the average temperature per hour, day and month. The lower graph shows the average departure delay on the same hour, day and month as the temperature. We haven't noticed any correlation between the temperature and the number of delays.

10- Visibilty:

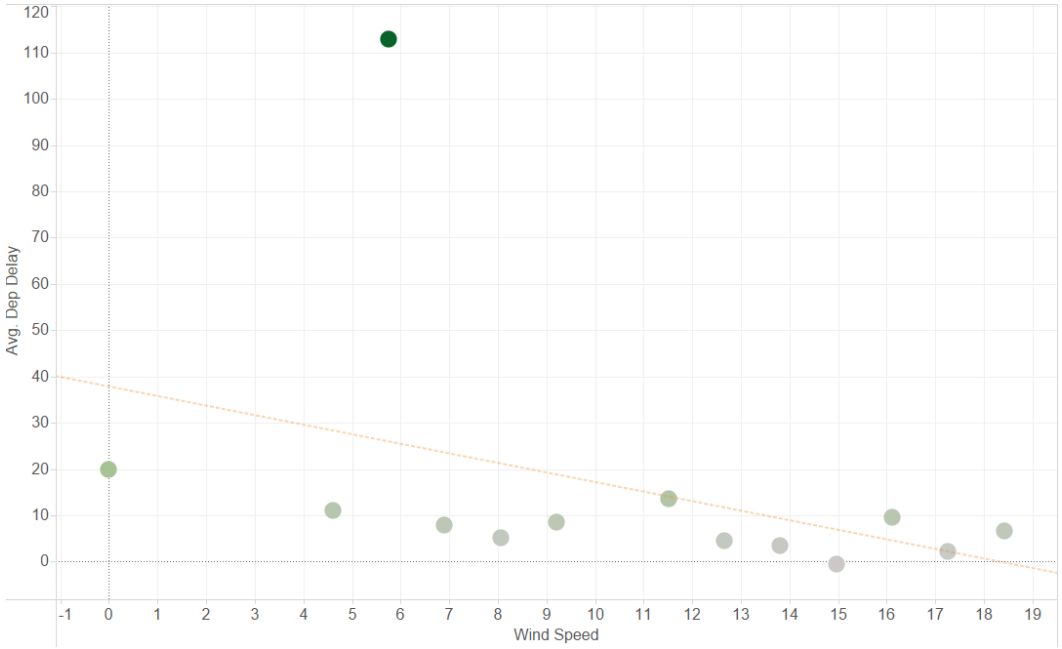
From the Flights table and the weather table, we extracted two variables: the departure delay and the visibility. We defined the departure delay as the dependent variable and the visibility as the independent variable and accordingly we generated the following graph:



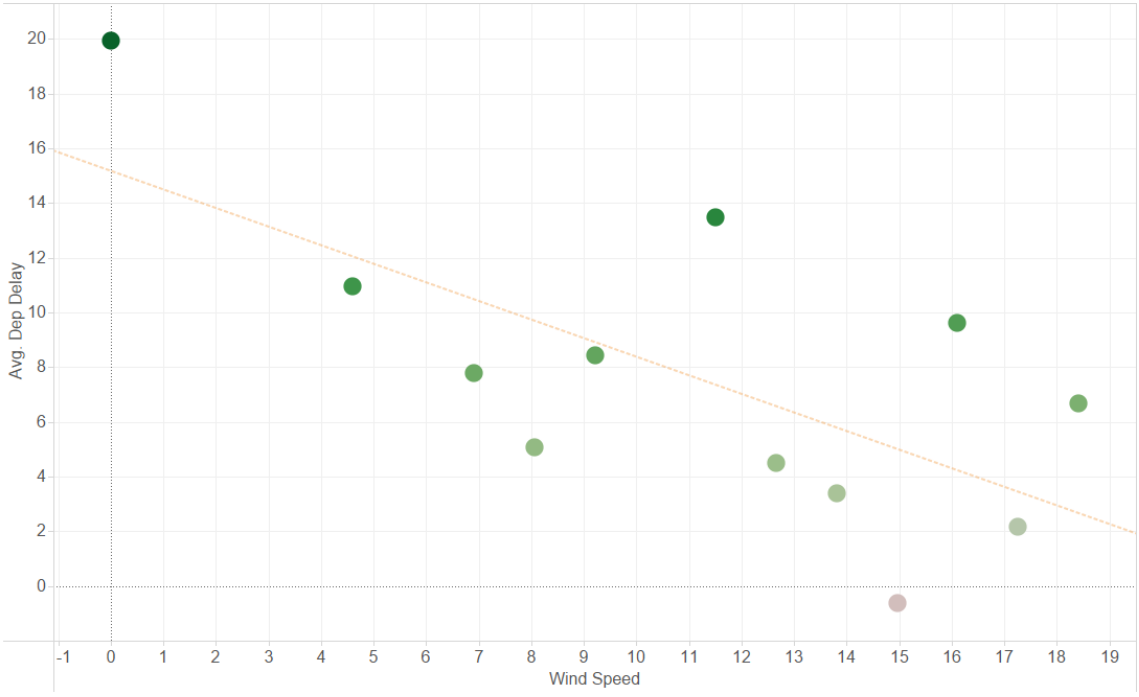
We noticed from the graph above that there is a negative correlation between the visibility and the average departure delays: the less the visibility the higher the departure delays.

11- Wind Speed

From the Flights table and the weather table, we extracted two variables: the departure delay and the wind speed. We defined the departure delay as the dependent variable and the wind speed as the independent variable and accordingly we generated the following graph:



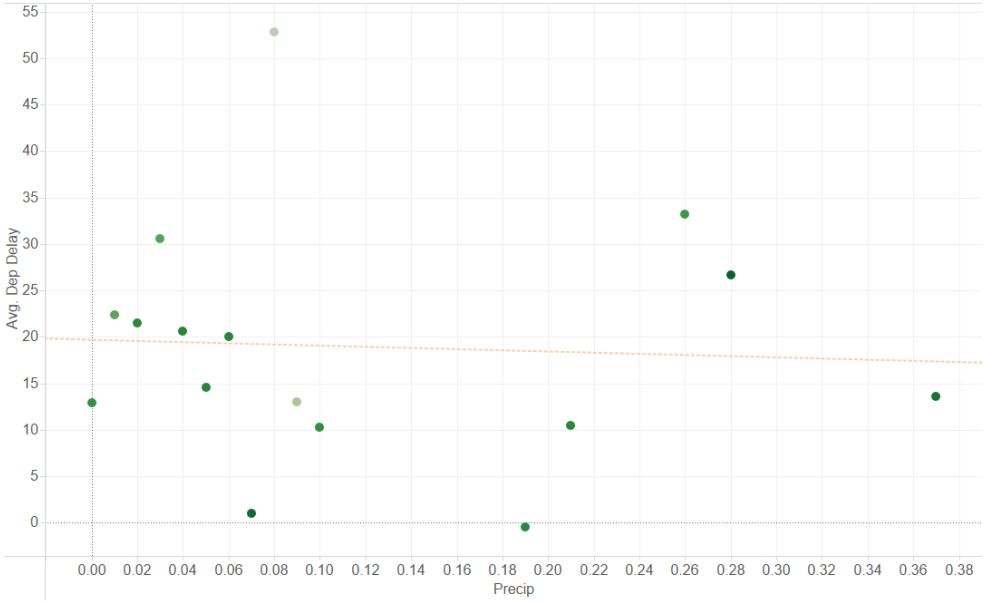
We noticed from the graph above that there is an outlier, and when we exclude it we obtained the following graph:



We noticed that there is a negative correlation between the wind speed and the average departure delays: the less the wind speeds the higher the departure delays. However, we found that this observation is ambiguous since normally the higher the wind speed, the more dangerous the flight is, thus the more it is likely to get delayed. Therefore, we think that the data related to the wind speed should be revised and accordingly we will not rely on it.

12- Precipitation

From the Flights table and the weather table, we extracted two variables: the departure delay and the precipitation. We defined the departure delay as the dependent variable and the precipitation as the independent variable and accordingly we generated the following graph:



We couldn't find any trend or relationship between the precipitation and the average departure delays, therefore we conclude that the precipitation doesn't really affect the flight delays.

13- Conclusions

After analyzing the graphs created and show in Tableau we can make some conclusions about the flight delays. There seems to be a correlation between most of the aggravated weather conditions and flight delays. Most of the weather plots show, for example, that when there is more rain, the average delay is higher. The same applies for less visibility, where less visibility means higher average delay.

Also, it is important to mention that most of the flights get delayed when the airports are more busy than usual. This is shown when looking at the average delay by month, and it appears that the higher average delay happens during the summer months and Christmas holidays.

One last thing is that shorter flights tend to have more delays than longer flights. This happens, because shorter flights have more frequency than longer flights, which result in a busiest schedule for the airport and hence a higher probability of delay.