28/10/2019

Ieseg – Msc in Big Data Analysis

BRT GROUP ASSIGNMENT

Technical report

Omar ABDELGELIL

Jules MOTYL

Winghao ZONG

**General remarks**

We decided to approach the data set following a general method:

1. Explore general relations, Discover large trends

We look at airlines, airports, weather…  
Do we see any kind of pattern in delay associated with those?  
IF yes-> why? onto step 2.

1. Dive into a specific domain and make hypothesis to explain the first observed pattern

Investigate more precise variables (numbers of flights, number of gate, temperature, delay...) for each category and discover what influence they have on the pattern we first observed.

This process should then continue until we can answer our hypothesis or until the dataset cannot give us more precise answers.

The exploration of the dataset was first done in SAS, to create smaller and more understandable tables from bigger ones. Then, with the help of Tableau, we identified the patterns and then came back to SAS to enrich the table we created with new variables or to create new ones.  
This has been a continuous process between observing the results in tableau and inputting new tables from SAS to gain a better understanding of the relations between the different variables and delays.

One of the most used table has been this one:

/\* Delay per flight Company \*/

**proc** **SQL**;

create table Delay\_per\_Airline as

SELECT A.name, case when F.air\_time <**90** then "Short Flight"

when F.air\_time <=**210** then "Medium Flight"

when F.air\_time >**210** then "Long Flight"

else "unknown" end as Length\_of\_Flight,

sum(F.dep\_delay) as Total\_DepDelay, sum(F.arr\_delay) as Total\_ArrDelay, sum(F.air\_time) as Total\_Time\_Spend\_in\_Air,

sum(F.dep\_delay)/sum(F.air\_time) as Ratio\_DepDelay\_Airtime, sum(F.arr\_delay)/sum(F.air\_time) as Ratio\_ArrDelay\_Airtime,

count(F.time\_hour) as Number\_of\_Flight

FROM Air.Flights F, Air.Airlines A

WHERE F.carrier = A.carrier

Group by **1**,**2**;

**quit**;

**run**;

/\* exporting the tables created \*/

**PROC** **EXPORT** DATA = Delay\_per\_Airline

OUTFILE = "C:\Users\jmotyl\Documents\GitHub\group assignment\DelayPerAirlines.csv"

DBMS = DLM REPLACE ;

DELIMITER = "," ; /\* séparator \*/

**run**;

It first was much more simple and got enriched one step at a time, making sure to keep it smart enough to allow for diverse analysis, but also to keep it always meaningful and understand the relation between variable by adding them one by one.  
The last 7 lines are just here to export the table to a CSV file that we then imported into Tableau.

**Omar’s Story**

When analyzing the flights performances and potentially what factors may be deducted to affect the flights delays, temperature had to be at the front line.

We can see here when analyzing the average temperature per month throughout the year with both the average delay time and the average number of delayed flights for 2013 per month that there is a positive correlation there. This means that the higher the temperature goes, the more delay time occurs. This maybe due to the fact that at summer (Period of year where temperature goes higher) people tend to have holidays and therefore the traffic at airports during this period of the year is at it’s peak. A possible recommendation for that might be to hire seasonal ground workers where they could help with the baggage transportation for example. Or airports which operates just until midnight throughout the year could work 24/7 for this period of time and therefore more flights could be scheduled.

The Following part of the story somehow align with the recommendation previously proposed. The analysis between the year of the plane being manufactured and the departure delay showed that the changes in the manufacturing processes didn’t change a lot throughout the years in a way that could reduce the delay. So the only logical reason in our opinion that could justify the graph’s shape might be due to the increase in population and the more connected the world is becoming and thus more traffic on airports. Also the airlines with the lowest operating flights showed more departure delays and one reason for that might be because those airlines runs as a charter service and therefore if a delay happened in a previous flight then the following flight might be delayed as well to give more room for passengers and bag gages connections.

Moving on to the following part of the story, we analyzed the airports performances. It is shown that although LGA airport did not have the most number of flights operating from its base; it still had the most delay compared to the other two airports. One reason could be that this airport is not as equipped as the other two airports in terms of number of gates, number of staff or employees or even IT infrastructure that may lead to more waiting lines at check-in and boarding passes counters.

The map shown in the next part of the story shows the worst 40 routes from NYC in terms of delay. We can see that the length of the flight is not a major factor when it comes to delays. And this could bring us to the point that usually based on reports and articles, the major factor related to flights delay is usually related to the flight preparations before taking off more than weather. Things like fueling the plane, technical maintenance checkups, packed gates and thus the plane have to wait more time on the ground until one is available.

Next part of the story focused on analyzing the number of engines and type of engines used by planes and showed that the majority of the delay came from planes who used a single 4 cycle engine or a double turbo shaft engines. Moreover American Airlines showed the most delays when operating with such engines.

An interesting correlation as well might be the relationship between number of seats in an airplane and the departure delay. From the dataset, we couldn’t deduct a clear relationship there but if more data provided, this might be useful if a positive correlation is founded as then an airport could try to optimize their administrative procedures for passengers as the more seat number the more the delay might be.

**Story 4**

This section mainly analyzes the impact of different weather conditions on flight departure delay.

We connected the weather and flight tables in the database together so that each flight has a corresponding weather condition, such as wind speed, temperature, humidity, etc.

In order to assess the impact of weather conditions more clearly and more accurately, we took the approach of averaging the delays for each interval. For example, in the interval of 10-20 wind speed, we collect the delay of the flight taking off in this interval, and then calculate the average delay. At the same time, we also do the same calculation for each integer interval so we can avoid ignoring some important information.

For example, here is one of the pieces of code.

/\*Ranges of Humidities and Delay\*/

**proc** **sql**;

create table HumidityEffects as

select count(f.time\_hour) as NumberOfFlights,sum(f.dep\_delay)/count(f.time\_hour) as AverageDepDelay, Humidity from

(

select origin, time\_hour,

case when humid > **0** & humid < **10** then "0-10"

when humid >= **10** & humid < **20** then "10-20"

when humid >= **20** & humid < **30** then "20-30"

when humid >= **30** & humid < **40** then "30-40"

when humid >= **40** & humid < **50** then "40-50"

when humid >= **50** & humid < **60** then "50-60"

when humid >= **60** & humid < **70** then "60-70"

when humid >= **70** & humid < **80** then "70-80"

when humid >= **80** & humid < **90** then "80-90"else

"90+" end as Humidity from groups.weather)

a,

groups.flights f,

groups.weather w

where f.origin = w.origin

and f.time\_hour = w.time\_hour

and w.origin = a.origin

and w.time\_hour = a.time\_hour

group by Humidity;

**quit**;