Doyen Rémy
Nagaraj Rudrappa Manju

Group Assignment Business Reporting Tools

NYC Flights data: Technical Report

As airline analysts we were asked to investigate different aspects of delays from the three different New York City airports: J.F. Kennedy airport, LaGuardia Airport and the Newark Liberty International Airport.

To perform this analysis, the first thing we have done was the data preparation. Indeed, we received several databases and a preprocessing of the datasets was needed in order to be able to use them.

1) Preprocessing of the data

Firstly, since we were interested about the reasons of the delays, we modified the negative delays (meaning that the airplane either landed or took off in advance) equal to 0 so that these early landings or takes off will have less impact on the different computation.

```
PROC SQL;
     create table nyc.Flights as
         select year,month,day,dep_time,sched_dep_time,
           arr_time,sched_arr_time,
           arr_delay, CASE when arr_delay <=0 then 0 else arr_delay END as arr_delay1,
           carrier,flight,tailnum,origin,dest,air_time,distance,hour,minute,time_hour,
           dep_delay, CASE when dep_delay <=0 then 0 else dep_delay END as dep_delay1
          from nyc.Flights ;
QUIT;
```

Still on this table, we then, took care about the different date and hour present in our flight database. We converted it in the desired format.

```
data nyc.flights(rename=(date=departure_Date));
set nyc.flights;
date= mdy(month,day,year);
format date date10.;
sched_dep_time=input(cats(sched_dep_time,"00"),hhmmss.);
sched_arr_time=input(cats(sched_arr_time,"00"),hhmmss.);
hour=input(cats(hour,"00"),hhmmss.);
format sched_dep_time time5.;
format sched_arr_time time5.;
format hour time5.;
run;
```

Doyen Rémy
Nagaraj Rudrappa Manju

We derived a variable named time_of_the_day since we wanted to understand if the time of the day can impact the delay of a plane. To do so, we used the following SQL statement:

```
(CASE when hour(sched_dep_time) >=  12 and hour(sched_dep_time) <  17 then "Afternoon"
      when hour(sched_dep_time) >=  17 and hour(sched_dep_time) <=  22 then "Evening"
      when hour(sched_dep_time) >= 6 and hour(sched_dep_time) <  12 then "Morning"
      else "Night"
 END) as time_of_the_day
```

Also, we derived a variable season to analyze the delays as per the seasonality of the year:

```
(CASE when month  >= 3 and  month  <= 5 then "Spring"
      when month  >= 6 and  month  <= 8 then "Summer"
      when month  >= 9  and  month  <= 11 then "Autumn"
      else "Winter"
 END) AS Season
```

Then, we also wanted to understand of the saturation of the airports could explain the delay. We considered that the saturation of the airports is correlated to the number of flights. So we created a table flightpermonth as following:

```
PROC SQL;
     CREATE TABLE nyc.flightpermonth AS
     SELECT count(flight) as Count_Flight, month, origin, AVG(arr_delay1) as AVG_Arr_delay
     FROM nyc.flights
     GROUP BY 2,3;quit;
```

Since we wanted to know the best / worst routes, we decided to use a spider chart in Tableau. To do so, we needed another table constructed by using a nunion statement as follow:

```
proc sql;
 create table nyc.flights_airport_both as
 select "origin" as route_identifier,*
 from nyc.flights_airport_origin
 union all
 select "dest",*
 from nyc.flights_airport_destination;
 run;
```

Finally, we a succession of join, we merged the tables together.

Doyen Rémy
Nagaraj Rudrappa Manju

2) Construction of the dashboards

The different dashboards we created have two main objectives:

- Understand what the current situation of the 3 New York airports is, according to the airlines, routes, …
- To discover what are the possible reasons for delay

The first dashboard is named 'Situation Airport'. With this dashboard, the user is now able to discover which of the three airports has in average the more departure and arrival delay. Moreover, in this first dashboard, the relationship between arrival and departure delay is plotted. We notice that this relation is linear, and the coefficient is almost equal to 1. That's why, for the following dashboard we decided to focus on the arrival delay. We indeed, considered that the arrival delay was, from a passenger point of view more important than the departure delay. From this first dashboard we can observe that the worst airport in term of average number of delay is the Newark Liberty International Airport.

The second dashboard is used to understand, which airlines are in average responsible of the most of delays, in general and by airports. We can see that the ExpressJet Airlines has an important number of flights and more average delay that the other big airlines companies. Especially for the Newark Liberty International Airport.

The third dashboard 'Delay by Routes' should show which are the worst or the best routes per airlines and per origin airport. We can easily see that the Newark Liberty International Airport is responsible of the 7 routes that have in average the more arrival delay.

The dashboard seasonality can be used to understand when the worst period of the day / year for the average arrival delay are. This dashboard allows us to know that in average, the arrival delay is higher in winter and in summer. Also, we know that the average arrival delay is higher during the evening.

Doyen Rémy
Nagaraj Rudrappa Manju

The fifth dashboard studies the link between the weather condition and the average arrival delay according to the month of the year. We can observe several things on this dashboard. Firstly, the three months with the most average arrival delay are June, July and December. And two thing may maybe explain this situation: the humidity (the higher the humidity, the higher the average arrival delay) and the temperature (when the temperature is really high or low, the average arrival delay looks to be higher).

Finally, the last dashboard is named 'Technical Dashboard'. The aim of this dashboard is to understand if the manufacturer can explain the delay. With this dashboard, we notice that the Brazilians manufacturer Embraer has in average a lot of arrival delay. The user can also go deeper and know which model is responsible for this delay.