NY Airport Flight Delays in 2013

Data Preparation

In order to analyze the flight data, we first read in all of the data tables into SAS for further processing like this:

```
DATA group.airlines;
	INFILE
'C:\Users\iboishin\Documents\GitHub\BusinessReportingTools\Group_assignment\a
irlines.csv' DLM=',' DSD FIRSTOBS=2;
	INPUT obs carrier $ airline_name :$27.;
run;
```

Then, we manipulated some of the data in order to fill in as many null values as we could, like this:

```
DATA group.flights;
	INFILE
'C:\Users\iboishin\Documents\GitHub\BusinessReportingTools\Group_assignment\f
lights.csv' DLM=',' DSD FIRSTOBS=2;
	INPUT obs year month day dep_time sched_dep_time dep_delay arr_time
sched_arr_time arr_delay carrier $ flight tailnum $
		origin $ dest $ air_time distance hour_sched_dep minute_sched_dep
time_hour $19.;
	IF arr_delay = . THEN arr_delay = arr_time - sched_arr_time;
run;
```

While we could not fill in all of the missing data, we reduced the missing data in the `arr_delay` column from 9430 to 8713, which leaves us with about 2.5% of that column being missing in the end.

Once we had all of the data as clean as possible, we merged it all together based on the flights table with this code:

```
PROC SQL;
CREATE TABLE group.Basetable as
	SELECT *, f.distance / (f.air_time/60) as Mi_per_hr
	FROM group.Flights as f LEFT JOIN group.Airlines as al
		ON f.carrier = al.carrier
		LEFT JOIN group.Airports as air
		ON f.dest = air.faa
		LEFT JOIN group.Planes as p
		ON f.tailnum = p.tailnum
		LEFT JOIN group.Weather as w
		ON f.origin = w.origin AND f.time_hour = w.time_hour
	ORDER BY obs;
QUIT;
```

Since we knew that all of the other data tables are simply additional information for each of the flights, we double checked that our merged datatable did not have more rows than the original flights data table as that would have meant that something went wrong in the merge. In addition, we added a flight

speed column based on the *distance* and *air_time* columns from the original flights database in order to test whether flight speed had any influence over the flight delay.

After all of the data tables were properly merged into one datatable, we removed the following unnecessary columns:

- `speed` - Data was only available for 963 instances of the table. Given that this is less than 0.3% of the total data, the amount of known values are too small to make a generalization on the whole.
- `faa, carrier` - The *faa* column is essentially a duplicate of the *dest* column. Thus, it is redundant. The carrier column is in a similar situation in regards to the airline name column.
- `year` - There is only one value of year for the whole database, making this variable redundant as well.
- `dep_time, sched_dep_time, arr_tim, sched_arr_time` - All of these values were necessary in order to determine the delay value. However, the specific departure and arrival times are less pertinent when it comes to creating generalizations on aggregate data.

Since we did not believe that these values would be of much us in our analysis of flights delay, we removed them in order to have the most concise database possible for importing into tableau later. We removed the aforementioned columns using this code:

```
PROC SQL;
ALTER TABLE group.basetable
DROP speed, faa, year, dep_time, sched_dep_time, arr_tim, sched_arr_time,
carrier ;
QUIT;
```

Finally, before moving to Tableau, we wanted to get a better understanding of the data. So, we created a couple summary charts to give us an idea of what figures we will be looking to make in Tableau. For example:

```
PROC SQL;
CREATE TABLE group.Airlines_delay as
      SELECT carrier, airline_name as Airline, ROUND(AVG(dep_delay),0.1) as
Dep_Arr_Delay, ROUND(AVG(arr_delay),0.1) as Avg_Arr_Delay,
               ROUND(MAX(dep_delay),0.1) as Max_Dep_Delay,
ROUND(MAX(arr_delay),0.1) as Max_Arr_Delay,
               ROUND(MIN(dep_delay),0.1) as Min_Dep_Delay,
ROUND(MIN(arr_delay),0.1) as Min_Arr_Delay,
               ROUND(AVG(distance),0.1) as Avg_Distance,
ROUND(AVG(air_time),0.1) as Avg_Air_Time, count(*) as Nr_Flights
      FROM group.Basetable
      GROUP BY 1, 2;
QUIT;
```

We used the *basetable* in Tableau in order to do some exploration of the data and the potential causes of the flight delays. Once we had determined the charts that we would like to make, we coded each of those charts in SQL with the intention of having one table per figure in Tableau, like such:

```
PROC SQL;
CREATE TABLE group.delay_over_time_airline_speed as
SELECT month, airline_name, avg(dep_delay) as avg_dep_delay, avg(arr_delay)
as avg_arr_delay, avg(Mi_per_hr) as avg_speed
FROM group.basetable
GROUP BY 1, 2;
QUIT;
```

In the end though, we thought it smarter to keep the *baseable* as the data source for most of the charts in Tableau for two reasons:

1. A lot of the columns between the newly created tables, notably those dealing with departure and arrival delay, were duplicated several times. This means that more data than necessary would be imported into Tableau thus slowing down the program and causing it to crash more often.

2. Importing tables that are meant to plot a specific chart prevents the user from digging deeper into the charts with filters of secondary variables.

As a result, in order to minimize the size of the source data and to improve the functionality of our final dashboards, we believe it wiser to simply use the basetable that we created earlier.

Story Explained

Delay by Airline

In the Delay by Airline page, we wanted to display the variation in departure and arrival delays between the different airlines. We made the sheet rather static to give the user a soft introduction to the purpose of the dashboard: the various reasons for flight delays. It is evident from this chart that airlines typically make up some of the delay departure that they start with. This chart also shows which airlines are more efficient with their organization and more capable of arriving on time.

Departure Delays by Plane Type

We are analyzing departure delay using boxplot of an airline with reference to plane type that specific airline is using, in order to find connection between average airline delay and model of airplane that airline is using. Through the boxplots, the user can easily see the average departure delays by airline as well as the spread about that average. Thus, it gives the user a feel for the accuracy and reliability of the different airlines. By clicking on any airline name in the box plot, the user can take note of the plane types that that airline uses.

Departure Delay Partially Made Up

We are analyzing average departure and arrival delays over different months of a year with reference to airline popularity (i.e. the number of flights) and the average speed. The overall trend is very intuitive: the highest average delays happen over the holidays (summer and Christmas). The tree map on the bottom acts as a filter for this chart, allowing the user to explore these trends for the specific airlines.

Delay throughout the day

The fourth page offers an interactive chart that examines the typical amount of delay over the various hours of the day. It shows that in general, there are no delays at the beginning of the day and then the delays rise steadily. This makes sense because as delays occur throughout the day, they end up piling on and becoming rather significant. It isn't until the end of the day when the delays start to come back down a little, likely due to the fact that there are less flights at the end of the day. The chart can further be drilled into by filtering by Airline Name, Origin Airport and Destination Airport.

Delay by Weather Conditions

The two weather conditions that seemed to have the largest effect on the amount of delay were wind speed and precipitation. In both cases, as the measurements when up, so did the amount of delay. It should be noted that the wind chart relationship is most representative up to a wind speed of about 26 mph. After that, the amount of data points used to make the aggregate drops to below 1000 and may thus simply be due to chance. Yet, we found them important to get the whole picture so we didn't filter the data to exclude them.

Delay by Airports

For this dashboard we used two sheets, one with the delay by origin for each airport, and the other with the delay by destination for every airport as well. In the first sheet, we show the departure delays in a boxplot to see how it is spread over the different flights. We can see that the results are similar for the different origins, concentrating half of the observations approx. Between 80 minutes and 300 minutes, with an average median of almost 200 minutes.

On the second sheet, we put the destinations in the USA map. The bigger the point the more number of flights the airport received during the year. We can also see the average arrival delay per destination, just looking at the color. Intuitively, stronger red color means more delay on average and green color points are actually airports that have negative delays on average.

Finally, in the dashboard we filter by origin to make it interactive, so the user can see for every origin airport the different destinations, with the number of flights they received from that airport and the average arrival delay.

Best and Worst Routes

We used two different sheets for this dashboard, one with the best and worst routes and the other one with the routes by distance. For the first one, first we created a table *routes_ordered* in SQL, where basically we group by origin and then by destination, and use the average arrival and departure delay, to finally calculate the overall average delay (average between arrival and departure delay). Secondly, we

used a where condition to filter by number of flights for every route to be more or equal than 10, to make a fairer comparison (some routes have just 1 flight). Finally, we used this table on tableau to get the top 5 destinations per origin (best and worst), ordered by average delay. For this we created an index filter (a variable with the index function) and then we fixed it to 5.

In the second sheet, we create a table to see the distance between the different routes. Lastly, we created the interactive dashboard where just by clicking on any origin or destination from the best or worst routes tables, we can see the distance for that route or routes (the last option in case we select the origin as there are 5 destinations per origin per table).

Delay by Manufacturer

On this page we are analyzing average departure and arrival delay per manufacturer of an airplane and we come to the conclusion that Agusta SPA has the highest average arrival and departure delay