United States flights reports Data exploration

Data set(explanation and approach)

Before I start, I would like to make a few points. My initial thought was to analyse more years and more into depth but I experienced a series of problems from downloading to handling the large files. After many crushes of my system and many failed download attempts, I downloaded the 2008-year file from the following url as advised by various mentors in Udacity Knowledge:

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7

As mentioned above, I experienced a lot of crushes during many different plotting attempts, so I came up with a somewhat simpler analysis. I hope it will be enough as today is 13.09.2020 and my deadline for passing all 5 projects (this is the last) is 17.09.2020. Having said that, thank you in advance, and I move on with the analysis.

I started my analysis by wrangling and cleaning the selected dataset of United States flights reports for the year 2008.

The dataset contains flights reports in the United States, including carriers, arrival and departure delays, and reasons for delays. It consists of 2389217 flight reports in 24 columns. The majority of the variables are numeric apart from the carrier, the origin and destination airports and the cancelation code.

There are some issues however with this dataset. Months variable contains only 4 months and the arrival and departure delays variables contain negative values which can be logically explained up to a point. Departure delays high negative values translating in very early departures is something that needed to be addressed. Since there is no sufficient explanation on this variable to understand how we came up with these values, I decided to create sub sets filtering these values. I decided to use negative values no less than -60 for arrival delays, which seems like something that could happen from time to time, early arrivals that is, and no less than -5 for departure delays, because early take offs are not something very usual to my experience. I also use a subset with no negative values at all to apply logarithmic transformations to the data, without errors. Lastly, specifically for flight cancellations I use the complete dataset, as the number of cancellations is very small and when filtering as described above, all the cancelled flights are filtered out.

Below is a brief explanation of the data set columns I used:

"UniqueCarrier" column contains 20 airlines,

"Origin" column contains 291 origin airports,

"Dest" column contains 293 destination airports,

“Cancelled” column contains values from 0 to 1 for not cancelled or cancelled flights,

"CancellationCode" column contains four cancelation codes:

- A for CarrierDelay,
- B for WeatherDelay,
- C for NASDelay ,
- and D for LateAircraftDelay

“DayofWeek” column contains numeric values from 1 to 7 which we convert to categorical data type with the actual days of the week from Monday to Sunday in that order,

“DayOfMonth” column contains days of the month from 1 to 31,

“CRSDepTime” column contains the scheduled departure time in numerically with the first two digits expressing hours in 24h format and the two last expressing minutes. I did not convert tis variable as I only used it to create three different timeframes which were stored as categorical datatype,

“AirTime” column contains the duration of the flight in minutes,

“Distance” column contains the distance between Origin and Destination Airports in miles

“ArrDelay” column contains the arrival delay, in minutes and

“DepDelay” column contains the departure delay, in minutes

The main feature(s) of interest in the dataset were the flight cancellations and flight delays and I tried to find what impacts them.

Exploratory Analysis

I started the exploratory part of the analysis with univariate analysis, by looking at some general characteristics about the flights. I looked at the number of flights for each airline. First by far is Southwest Airlines.

Then I looked at the airports to find the top 5 busiest as origin and destination airports. Same five were first in both categories and with the same order:

1. William B. Hartsfield-Atlanta International Airport
2. Chicago O'Hare International Airport
3. Dallas/Fort Worth International Airport
4. Denver International Airport
5. Los Angeles International Airport

Next, I looked at days of the week and month. Here something that stands out is that Saturday and the last day of the month have the lowest number of flights.

After that I examined the flight cancellations. I plotted a pie chart to see visually the proportion of cancelled to not-cancelled flights. The cancelled flights are about 2.7%.

Then I examined the reasons of cancellations. The most common reasons of cancellation were Airline related and Weather related, with NAS related reasons coming third and the rarest reason of cancellation were Security related.

After the cancellations I examined arrival and departure delays distributions.

Both distributions were long tailed and highly skewed to the right, even after zooming in the most common flight delays time. There are a lot of flights with few minutes of delay and very few with a large number of minutes.

I then applied a logarithmic transformation to have a better look. After the transformation the graph appeared unimodal and the distribution normal centred roughly around 20 minutes for both arrival and departure delays

I then proceeded with the bivariate analysis/exploration.

I tried to find out if the numerical variables of our dataset are correlated in any way. I used scatterplots to do that.

More specific, I examined arrival and departure delays which, as expected, showed a very strong correlation and airtime and distance which showed a very strong correlation too.

Then I examined arrival and departure delays relation with distance and airtime. Contrary to my expectation neither airtime nor distance seem to affect any kind of delays a lot.

Since the scatterplots did not show any strong relation between the variables of interest (arrival and departure delays) and the other two numeric variables, I moved on further exploring the flight cancellations, starting from the relation between airline and flight cancellation.

I plotted both the number and the average of cancelled flights in a bar chart.

Southwest Airlines appeared first in the number of cancelled flights; however, we've seen earlier that it is the Airline with the biggest number of flights by far. So, I thought it would be better to look at the averages where it ranks below the midpoint.

On the other hand, Mesa Airlines is first in the averages of cancelled flights table, while their number of flights is relatively low. Pinnacle Airlines is a similar case, with high average ranking and lower flight count, which is certainly not a good sign for their image. American Eagle and American Airlines have both high ranking in both tables and Hawaiian, Aloha and Frontier Airlines, appear to be the airlines with the lowest flight cancellations' averages, but very low number of flights too.

Then I looked at the days of the week, plotting again both the number and the average of cancelled flights per day in a bar chart.

Tuesday and Friday appeared to be the days with the highest cancellation averages while at the same time weekends seemed to have the lowest.

Then, following the same logic, I plotted both the number and the average of cancelled flights per day of the month in a bar chart.

This time we noticed the last day of the month showing the lowest count of cancelled flights but at the same time the highest percentage too. The low number of cancellations can be explained by the low number of flights we have seen earlier in this analysis. One other thing that we can notice is that the second half of the month seems to show relatively lower percentages of cancellations.

Last examined variables for flight cancellation numbers and averages, were origin and destination airports.

William B. Hartsfield-Atlanta International Airport was first at number of cancellations but third in average cancellation with a lower percentage of more than half the percentage of the second airport of Dallas/Fort Worth International Airport. On the other hand, Chicago O'Hare International Airport, is first at averages and second at number of cancellations with a relatively high number.

Next, I looked at the arrival and departure delays trying to find out how much, airlines, airports, day of month and day of week impact the flight delays.

We looked at the variables in the same order we did in the flight cancellations exploration. This time we looked at delays' averages.

Regarding the airlines' variable, we noticed that the two ends of the graphs, are the same in both Arrivals and Departures. Mesa Airlines, in the higher end, seems to be the Airline that experiences the most delays out of all airlines, and on the other hand in the lower end, Hawaiian and Aloha Airlines seem to have the lowest delays out of all the airlines.

Looking at the days of the week variable, we find Fridays and Thursdays once again in the first two places, with Monday a close third. The departure delays are more evenly spread across days while arrival delays show a slightly different image. Wednesdays and Saturdays are in relatively better place with slightly lower arrival delays than the other days.

As for the days of the month, departure delays appear slightly higher throughout, but generally the variation of delay is similar to both departure and arrival delays. The last day of the month shows the longest delays while the first and fourth days are second and third longest.

Looking at the airports variable Chicago O'Hare International Airport apart from being first at flight cancellations' averages, is first again in average delays as well. This time things look even worse. The average delay compared to the other airports is far greater, with arrival delays being more than double and departure delays almost double. The other four airports are very close to one another with Los Angeles and Denver International airports showing the shortest delays of the five.

Next, I wanted to explore what impact has in delays the time of the flight. I engineered a new variable, dividing the day in three parts, by order of popularity. I categorized it as follows:

1. 8:00 - 16:00 most popular,
2. 16:00 - 00:00 second most popular and
3. 00:00 - 08:00 the least popular.

I also engineer one additional variable in order to categorize delays from Minor to Major as follows:

1. Minor delays: less than 5 minutes,
2. Significant delays: from 5 to 60 minutes and
3. Major delays: from 60 minutes and more.

After writing the necessary code to engineer the variables I plotted a clustered bar chart to explore the relation between the time of the travel and the delays.

As expected, there was a definite relation there. Minor delays occur much more often between 08:00 and 16:00 and significant delays too, but the relation here is not that strong. The interesting thing to observe here is that major delays between 16:00 and 00:00 seem to occur just as much as between 08:00 and 16:00, if not more.

To explore deeper the same relation, I plotted the same data on a heatmap as well. This way I could better see how major delays compare from the most to second most popular travel timeframe.

After some necessary summarization and pivoting work, I was able to plot the data in the heatmap.

Viewing the heatmaps with the annotations we could see clearly both visually and numerically that more major delays were observed during the timeframe from 16:00 to 00:00 than the most popular one from 08:00 to 16:00. An interesting and rather unexpected find!

Since airlines, origin and destination airports, day of the week and of the month, all seemed to impact flight delays I wanted to proceed with the multivariate exploration to look at multiple relations at once.

I looked at the relation between airlines and destination airport and the average arrival delay, and between airlines and origin airport and the average departure delay.

After the necessary summarization and pivoting the generated heatmap showed that Chicago O'Hare International Airport had the most high-average delays and William B. Hartsfield-Atlanta International Airport the second most high-average delays. However, the single highest delay average was in William B. Hartsfield-Atlanta International Airport too and more specific from American Eagle Airlines. The highest arrival delay on the other hand was observed at Chicago O'Hare International from Mesa Airline.

The heatmap makes obvious that delays are affected by the combination of airline and airport as for example Mesa Airlines average arrival delay is almost 0. This is a quite large difference from the highest observed average arrival delay of over 36 minutes we've seen earlier. We can easily spot similar observations all over the table. Another example is the other highest average we've seen earlier, the one for departure delays from American Eagle Airlines with almost 40 minutes at William B. Hartsfield-Atlanta International Airport when at Denver International Airport their average is 0 and at Los Angeles International Airport less than 2,5.

Next, I examined the relation of the different times of flights with the different airports in a new heatmap.

There is a very interesting observation out of this heatmap. There are significantly higher delay averages in all five airports during the second most popular travel timeframe between 16:00 and 00:00 than the most popular travel timeframe between 00:00 and 08:00.

I decided to looking at the average delays from airport to airport during the three timeframes in a new heatmap too.

The same thing was observed once again, out of this new heatmap. Higher average delays during the second most popular timeframe at the vast majority of airports as well.

The two last plots made us wonder whether the initial assumption that the timeframe between 08:00 and 16:00 was indeed the most popular one is true.

To find that out we plotted a last heatmap to see the number of flights per timeframe.

Finally, the assumption proved correct since we could clearly see that most flights took place during the timeframe from 08:00 - 16:00.

Explanatory Analysis (Data Visualisation)

The key insights that will be conveyed in the explanatory report will be the factors linked to higher (and in some cases the lower) flight delays as well as higher flight cancellations (and in some cases the lower). More specific I will name the Airlines, Origin and Destination Airports, Week days and Days of Month, as well as the times of the day with the highest and lowest Cancellations and Delays:

Flight Cancellations

Certain Airlines:

- Mesa Airlines (Higher)
- Frontier Airlines (Lower), with a mention about Hawaiian and Aloha Airlines which are ranked very low

Certain Week Days:

- Friday
- Tuesday

Certain Month Days:

- The 4th
- The 6th
- The 31st

Certain origin and destination Airports:

- Chicago O'Hare International Airport

Flight Delays

Certain Airlines

Arrival Delays:

- Mesa Airlines (Higher)
- Aloha Airlines (Lower), with a mention about Hawaiian Airlines which is ranked very low too

Departure Delays:

- Mesa Airlines
- Hawaiian Airlines (Lower), with a mention about Aloha Airlines which is ranked very low too

Certain Week Days:

Arrival Delays:

- Fridays

- Tuesdays

Departure Delays:

- Fridays

Certain Month Days:

Arrival Delays:

- 31st
- 4th

Departure Delays:

- 31st
- 4th

Certain origin and destination Airports

Arrival Delays:

- Chicago O'Hare International Airport and

Departure Delays:

- Chicago O'Hare International Airport and

Certain timeframes:

Arrival Delays:

- 8am-4pm
- 4pm-12pm

Departure Delays:

- 8am-4pm
- 4pm-12pm

List of resources:

- ***Stackoverflow.com***
- ***pandas.pydata.org/docs***
- ***matplotlib.org***

- ***seaborn.pydata.or***