Flight Delay Prediction For Aviation Industry Using Machine Learning

Exploratory Data Analysis :

The same way how the data cleaning and preprocessing was done in two separate notebooks, the EDA was done in two as well, however the difference here is that the visualizations done on each of the EDAs were done with different libraries. The first was done using matplotlib and Seaborn, and the second with plotly.

On the first EDA notebook, the following questions were addressed:

1. Total Number of Flights by Airline
2. Number of Delayed Flights by Airline
3. Percentage of Delayed Flights by Airline
4. Total Minutes Delayed by Airline
5. Average Delay Time by Airline
6. 30 Most Common Destination (Cities)
7. Worse and Best months to travel
8. Is there a Better day of the month to travel?
9. Best weekday to avoid delays
10. Impact of Delays (Departure vs Arrival Delay)
11. Most Popular Destinations with Average Arrival Delays
12. Number of Destination by Airline
13. Recommended airlines based on lowest delay times

You will notice that each one of these questions were addressed and discussed individually and afterward, put together to answer question 13.

Again, I won't go through all of them here, but just share a few interesting findings:

Total Number of Flights by Airline: The plot from Figure 3 talks by itself, therefore, it is quite easy to interpret. Basically stating that the top 5 airlines in terms of number of flights are:

- SouthWest Airlines
- Delta Airlines
- American Airlines
- SkyWest Airlines
- United Airlines

With no additional comments about this, I will come back to this list after looking at other plots.

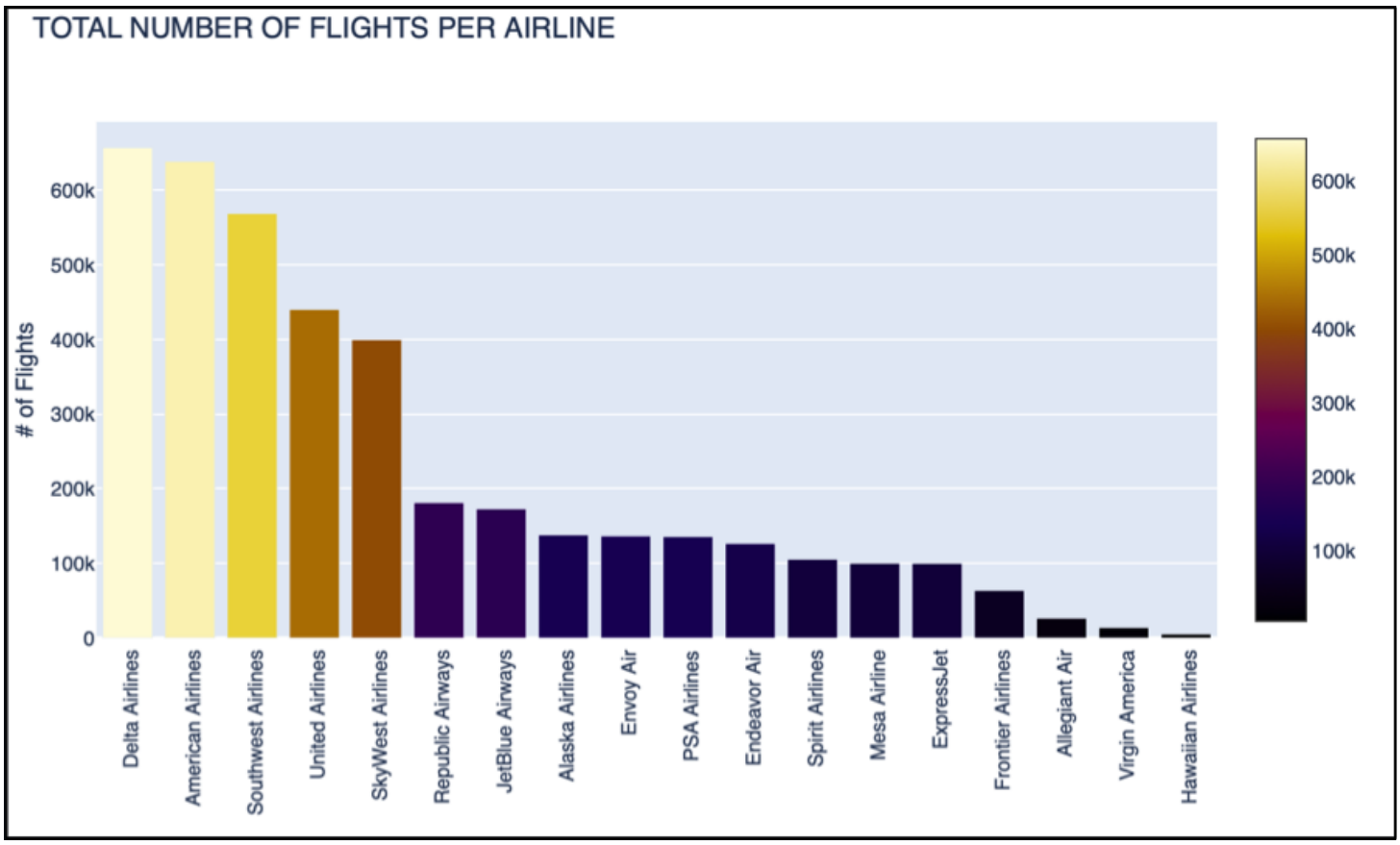

Figure 1. Total number of flights by airline sorted is descending order.

Percentage of Delayed Flights by Airline: It seems normal to think that the most flights you have the more likely it is that you will end up having more delayed flights. It's simple math right? For example, lets assume a fix percentage of delayed flights such as 30%, well 30% of 100 is 30, whereas 30% of 1000 is 300. We translate that into flights, and there is a huge difference with a ratio of 10:1 in terms of numbers, but the percentage remains the same.

Now according to this dataset, the average of delayed flights in the US for 2018 was 37.52%, which is the red horizontal line on plot from Figure 4. I know that in the introduction I mentioned a 20% of flights within the US being delayed, but that number if overall for the 58 airlines that operate domestic US flights, whereas my dataset only looks at 18 airlines which I am assuming are the major carriers.

You as the airline don't want to be above that red line/threshold, you want to be as far as possible below it. If you pay attention to Delta Airline, they are top 5 in terms of number of flights, but they are dead last in terms of delay percentage. It is quite interesting the relationship that they have managed to achieve.

Another interesting observation is that SouthWest Airlines and American Airlines are two of the other top 5 in terms of number of flights and they are both above that threshold that we want to avoid.

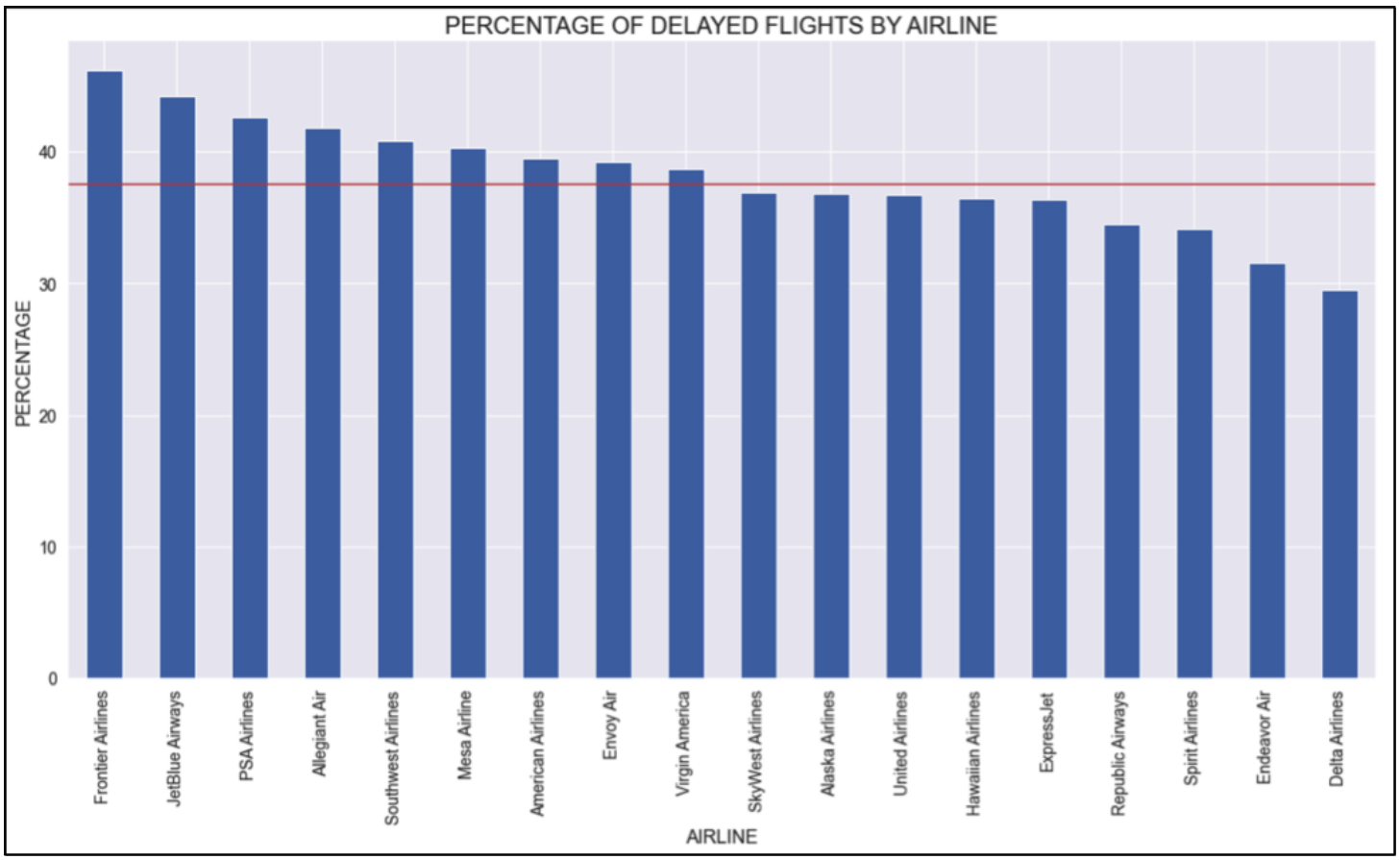

Figure 2.Percentage of delayed flights by airline

Most Popular Destinations with the largest arrival delay: Because there are a total of 358 destination airports within 341 cities, I decided to focus only on the top 30.

Chicago, Atlanta, New York, Dallas-Fort Worth and Denver are the top 5 destination, with Chicago being number 1, but interesting enough it has a pretty high average of annual delays, so if you are traveling to Chicago, there is a high chance that your flight will be delayed. Atlanta in the contrary, is the second most popular destination and with a very low delay at arrivals. New York and Dallas-Fort Worth aren't great, and Denver is just within the average.

Out of the top 15 destinations, the city with the most delays is by far Newark, where you are almost guaranteed to arrive late. Others cities that have very negative records are San Francisco, Orlando, Boston, Philadelphia, Ft. Lauderdale, Tampa and Chantilly.

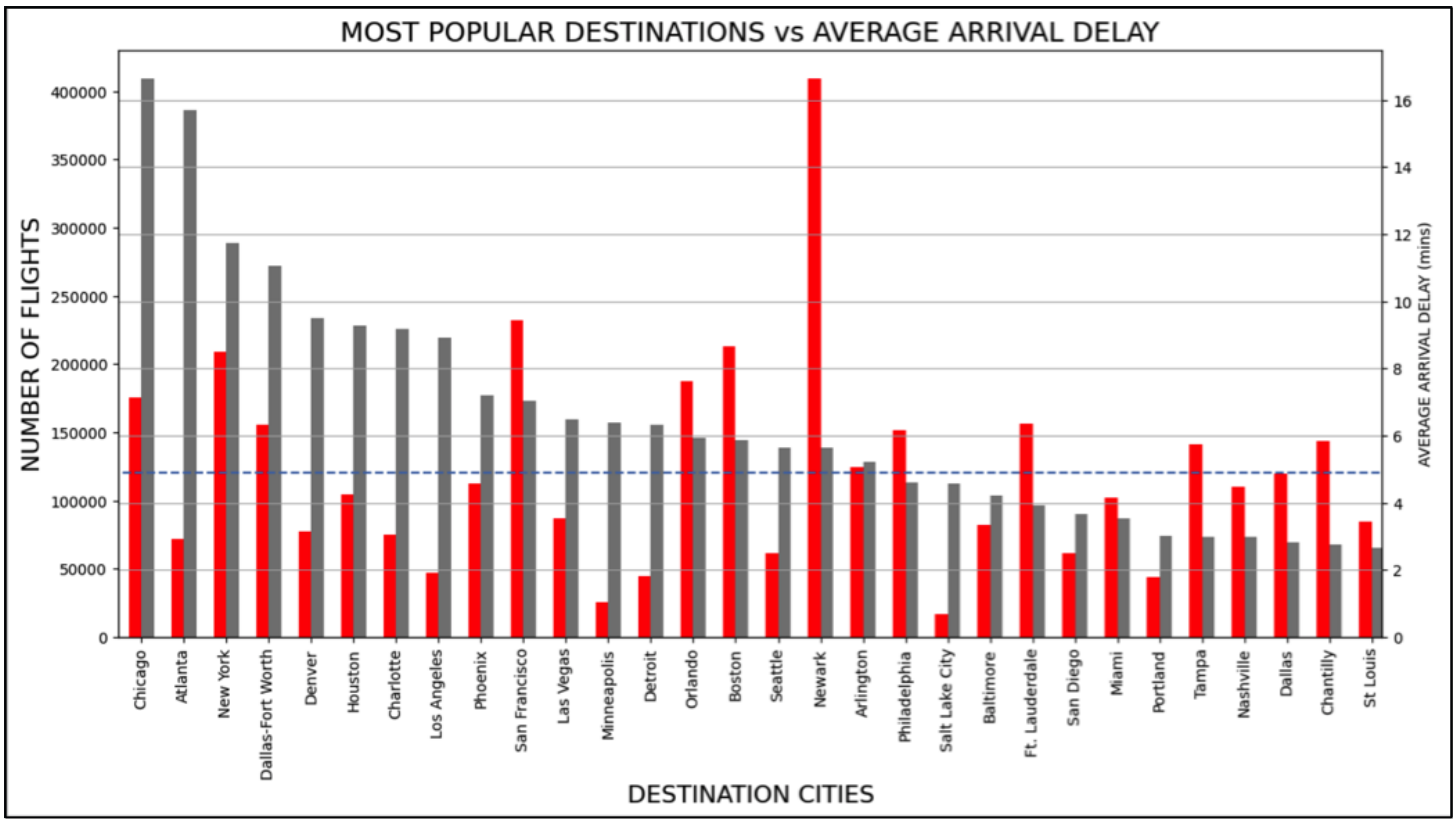

Figure 3. Most popular destinations (cities) with their average arrival delay (min)

Now the plot on Figure 5 compares the most popular destinations again with the average departure delays, with the dashed line being the average. So again, you would want to be below that threshold, but in this case we are talking about cities and multiple airlines at the same time.

If we look at Chicago, we can see that it has quite a high average departure delay, but combining this information with the one from Figure 4, we can infer that flights going to Chicago try to compensate for late departures by reducing the elapse time, and in average it seems as they succeed. With regards to Atlanta, it still is in a good position by being the second most popular destination, with low arrival delay and still with an average delay below the average. I am not sure if this is related to the arrival or departure airports, the weather in this area, or why exactly this happens, and in order to explain it, I would need some additional data which I don't have and that goes beyond the scope of this project anyways, but perhaps is something that can be added later on.

Once again Newark is in bad shape by having the highest average of departures delayed. Orlando and Boston and two others that combined with Figure 4, puts them in bad position. And then you can see the cities which are in pretty bad shape going way above the threshold, such as Philadelphia, Baltimore, Ft. Lauderdale, Miami, Tampa, Nashville and Dallas. Reasons for this? again not enough data nor time to find out.

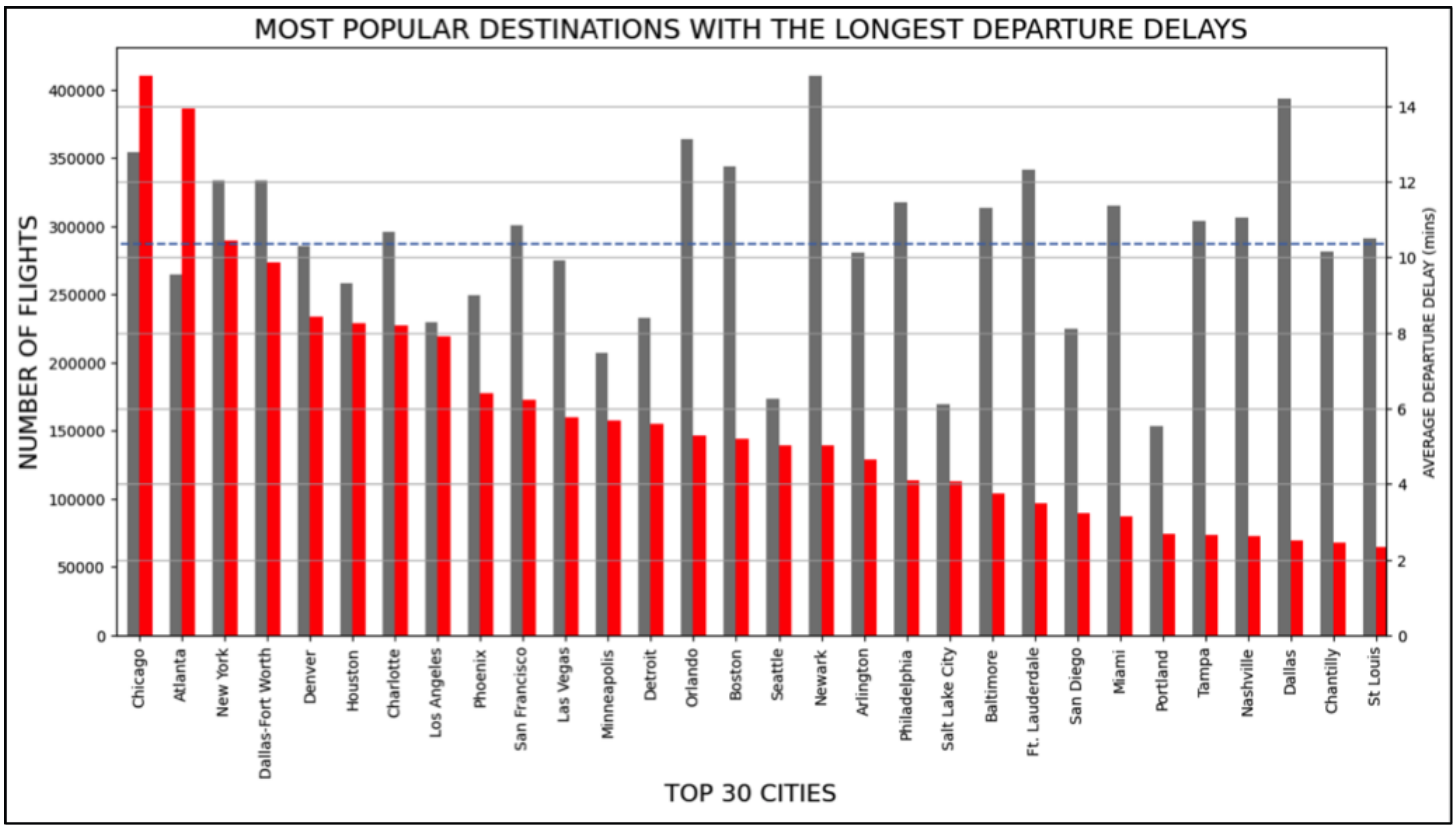

Figure 4. Most popular destinations (cities) with longest average departure delays (min)

Number of destinations by Airlines The plot from Figure 7 is the last one that I will comment on this introductory README. Here you see the number of destinations per airline and once again it's interesting because it shows as highlighted on that plot, that Delta Airlines is the third with most destination. Remember, that it is also top 5 in terms of number of flights, it has the lowest percentage of delayed flights, and it is in negative with regards to the total delayed minutes. It seems as they perform quite well from this pack of 18 airlines so it is the one that I would recommend based on this information for the year 2018. Now this might have changed, I really could say. What I could do and add it later on to this project, is extend the study to all the files cover the 10 years available and that way see if this is a one year trend, or if it is really a historical one, which in that case, it will become more solid to make such a recommendation, but for now I will have to live with what I have.

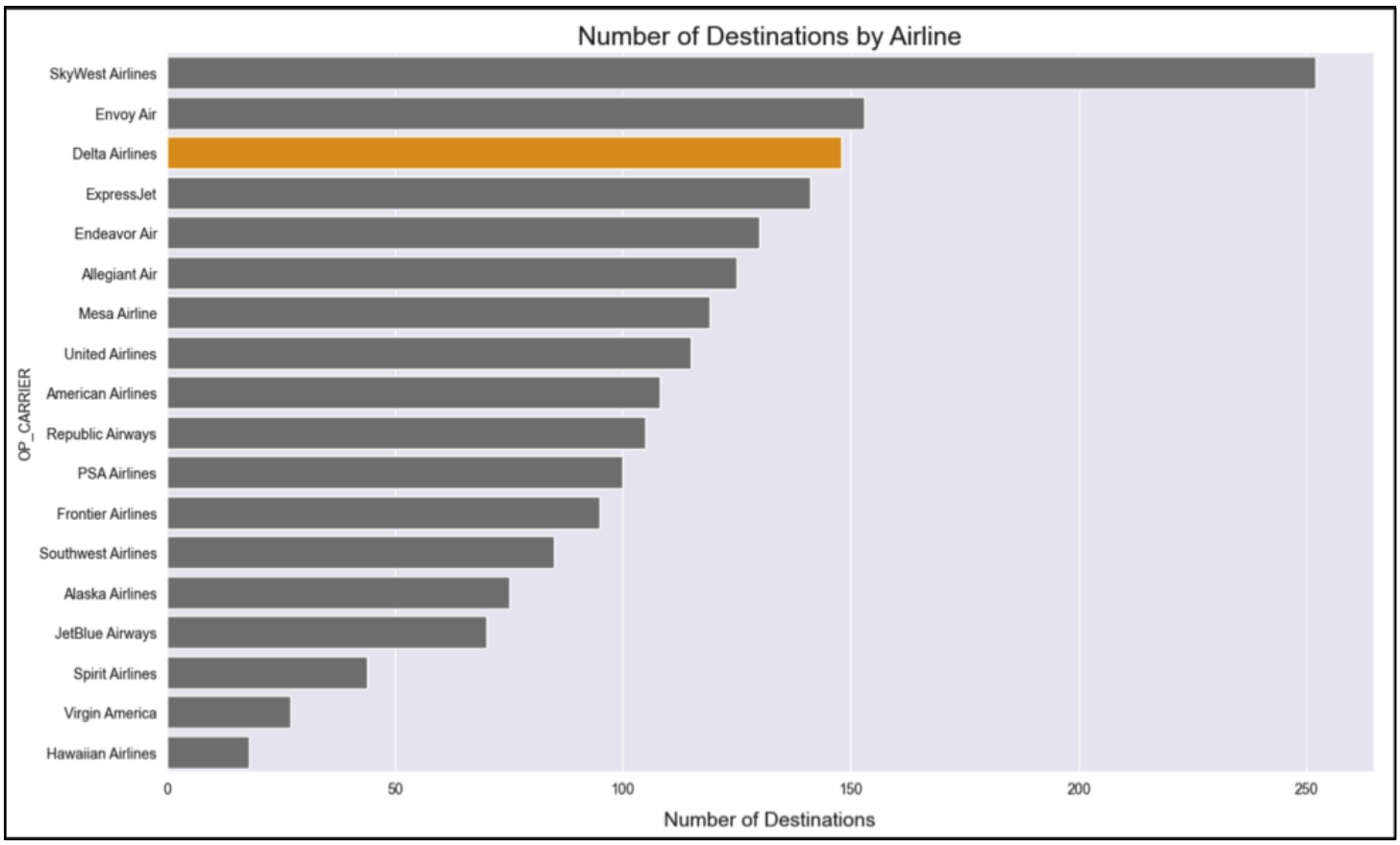

Figure 5. Number of destinations by airline