Data Visualization Project – Flight Data

**Link to the Workbook**

Version 1

<https://public.tableau.com/profile/henrik.steffen#!/vizhome/DataVisualizationProject_sub_v1/Story_v1?publish=yes>

Version 2

<https://public.tableau.com/profile/henrik.steffen#!/vizhome/DataVisualizationProject_sub_v2/Story_v2?publish=yes>

**Summary**

For this project I looked at flight data for the years 2007 and 2008, which comprises a combined 14.4 million flights (roughly 7M per year). Looking at the distribution of arrival and departure delays, one can see that the modes of the distribution are actually on the negative side, meaning that most flights arrive/depart earlier than scheduled. Nonetheless, the average arrival/departure delay is around 10 min, probably due to the fact that the distribution is very long tailed with maximum delays of around 2600 min. The emergence of delay, for the most part, seems to be carrier-dependent, as there is little to none correlation to factors such as number of flights per airport, number of flights handled per carrier or distance covered. Carrier delay and NAS delay are the types of delay that most significantly contribute to the overall delay, while weather delay and security delay only play minor roles in most cases. There is also a temporal component to delays with peaks in the average arrival delay in the middle of the year and towards the end of the year.

**Design**

I decided to show the distribution of Arrival and Departure Delay as histograms of 10 min bins. Since these distributions are extremely long tailed I decided to show only delays up to 300 min (the maximum is at ~2600 min) to focus on the peaks of the distribution. I also included a reference line for the average delay.

I plotted the average arrival delay for each carrier, the distance per carrier, the number of flights per carrier and airport as bar diagrams in descending order, since they allow for an accurate distinction of differences between carriers. I decided to color the carrier with the highest average arrival delay in orange, and to color this carrier in orange in other plots (number of flights, distance) as well to make it easily visible if there is a relationship between these factors.

For visualizing the airports handling the most flights, I decided to show a map of the airports as circles with the size of the circle indicating the number of flights originating from a given airport. I only included airport with more than 100,000 flights in order to remove all the smaller airports cluttering the map and focus on the biggest airports. The size of the circle is not optimal to determine the absolute number of flights per airport but one can easily see differences between airports and the top airports are clearly visualized on the map.

I then plotted the average departure delay per airport on a map in order to visualize whether there is a geographical pattern. The average departure delay is encoded both by the size of the circle and the color of the circle (the circle represents the airport). I chose this double encoding since the differences in average departure delay are quite small. There are only four levels for the color encoding, which allows only for a crude grading but in combination with circle size, I think the airports with the highest departure delay can clearly be distinguished.

I visualized the correlation between average departure delay and number of flights per airport or between late aircraft delay and number of flights per airport as scatterplots. I also included a trendline to more clearly show the correlation.

I plotted the different kind of delays and the average arrival delay per month as line diagram since this is most suitable to show changes over time. For the plot of the different kind delays, I showed the plots for 2007 and 2008 next to each other in order to emphasize that these values can greatly differ from year to year. In all cases I assigned the colorblind palette to more easily distinguish between the different colors.

Lastly, I visualized the composition of the different kinds of delays per carrier as stacked bar diagram. This allows for a quite accurate assessment of the ratio of each delay type.

**Feedback**

On Dashboard 2:

* Selecting on the map filters on the plots for Number of flights per carrier and Average arrival delay but does not highlight the airport in the bar diagram for number of flights per airport
* I have included highlighting from the map in the bar diagram for number of flights
* Selecting an airport in the bar diagram for number of flights per airport does neither filter on the other two bar diagrams (as does selecting an airport on the map) nor does it highlight the airport on the map
* I have included these actions

Scatterplots:

* For the scatterplots on dashboard 3 and 5 a lot of dots are overlapping, making it hard to distinguish individual dots of they are close together
* I reduced opacity of the dots to make it easier to see when dots are overlapping and I also changed the color to a dark grey in order to increase the contrast to the background

**Resources**

https://community.tableau.com/thread/155956

https://www.tutorialgateway.org/tableau-rank-calculation/

https://community.tableau.com/thread/127052