CS 503 – Data Visualization Assignment 3

Yi Ren (002269013), Wentao Lu (002276355)

Our research question

In context of this particular dataset, we are interested in these questions:

- From the dataset, we see that the real departure time rarely matches the scheduled departure time, some flights take off earlier (negative delay) while others take off later (positive delay). So, is the delay time at take-off related to the origin airport? Which airport has the most flights delayed?
- Similar question to the first one but in terms of seasonality, which naturally comes up because many delays are usually caused by busy crowds. So, does weekday impact departure delays? Do flights delay more often on weekends?
- What about delay at landing? Is it related to the flight distance? In common sense, the longer the distance, the more unexpected things could happen during the flight, so we want to study the relationship between the two.

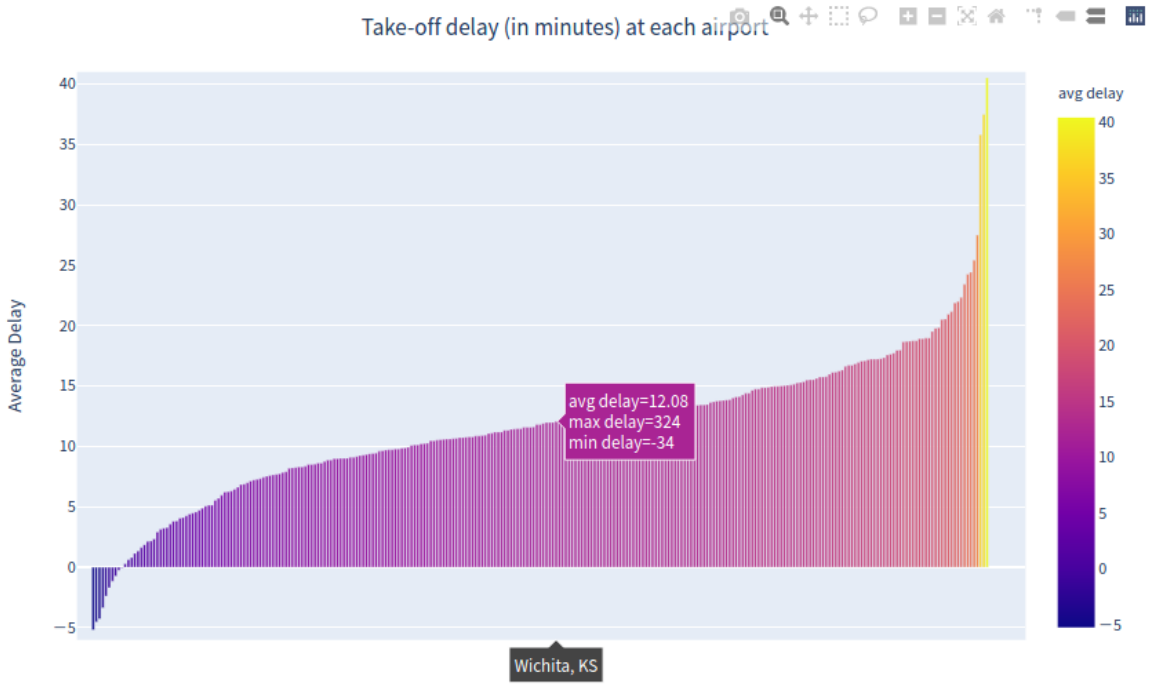
EDA and data preprocessing

In brief, here's what we did to preprocess the data (the attached Jupyter notebook has very detailed step-by-step code and explanations).

- remove columns that are irrelevant to our research question
- select the subset of data that have not been cancelled
- our data is pretty complete except that 0.29% percent of the `ARR_DELAY` column is missing. This is only a small portion, so we reasonably filled in those missing values using the median value.

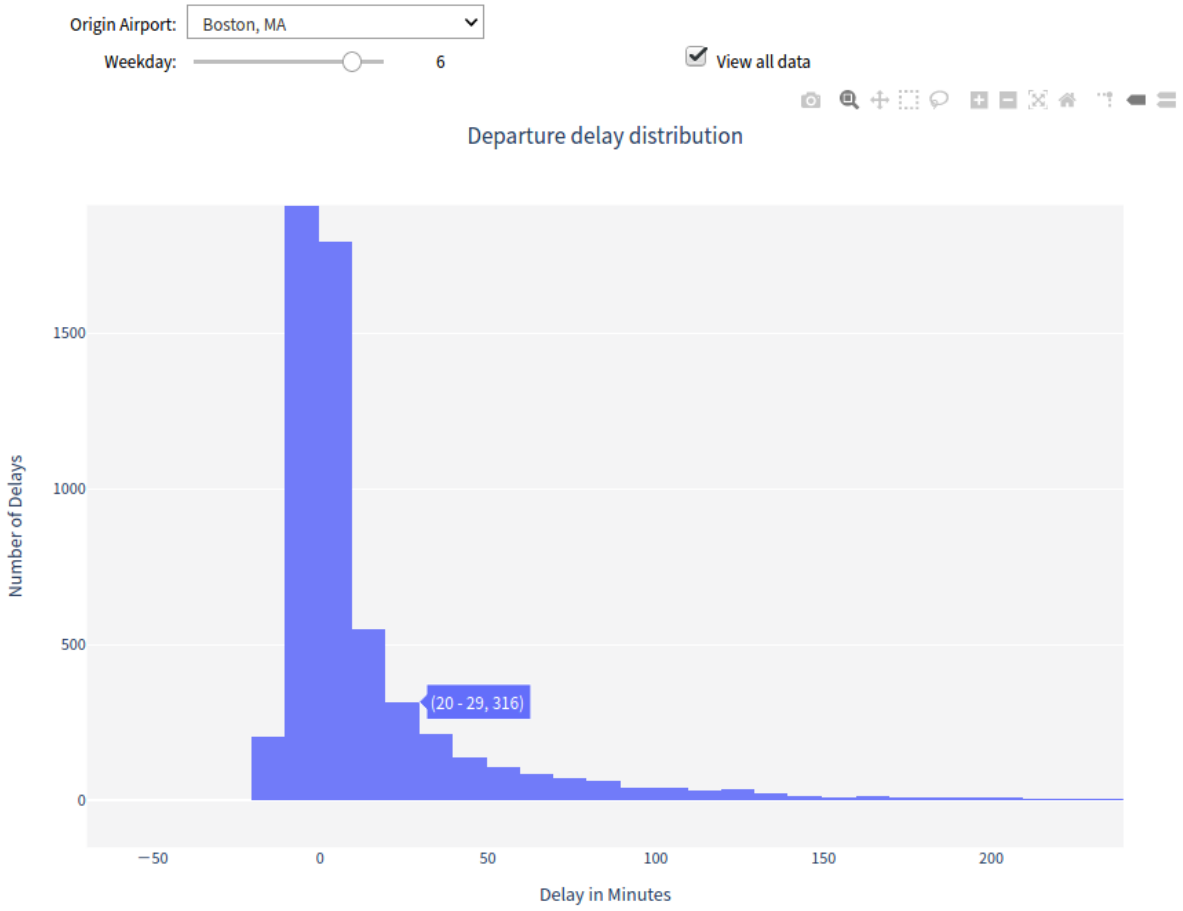
Visualization analysis

To answer our research questions, we start by looking at the summary statistics of departure delay for each origin airport. In the figure below, we create a barplot that shows the mean departure delay for each airport, sorted in ascending order. By hovering the mouse on a specific bar, we can see the origin airport city as well as the minimum, average and maximum delay time in minutes. The plot uses a color scale so that higher average delay has a brighter color. From this figure, we can easily identify how likely flights from a given airport are going to delay, for example, the highest bar shows that flights from Adak Island in Alaska delay over 40 minutes on average.



The barplot looks nice and simple, but this is a high-level overview since it is solely based on the summary statistics, not the whole dataset. For example, we only have few observations for some specific airport, so these high-level statistics may be biased and unreliable. It would be nicer if we could also look at each airport more closely.

For this purpose, we are going to plot the distribution for each airport using the whole dataset, and then make the plot interactive that allows the user to click and select a different airport. Besides, we also make the `DAY_OF_WEEK` column selectable to compare between weekdays and weekends, the user can choose to see data for a given weekday, or view the data in its entirety for the December (since we only have data for December, 2009), this is how it looks.

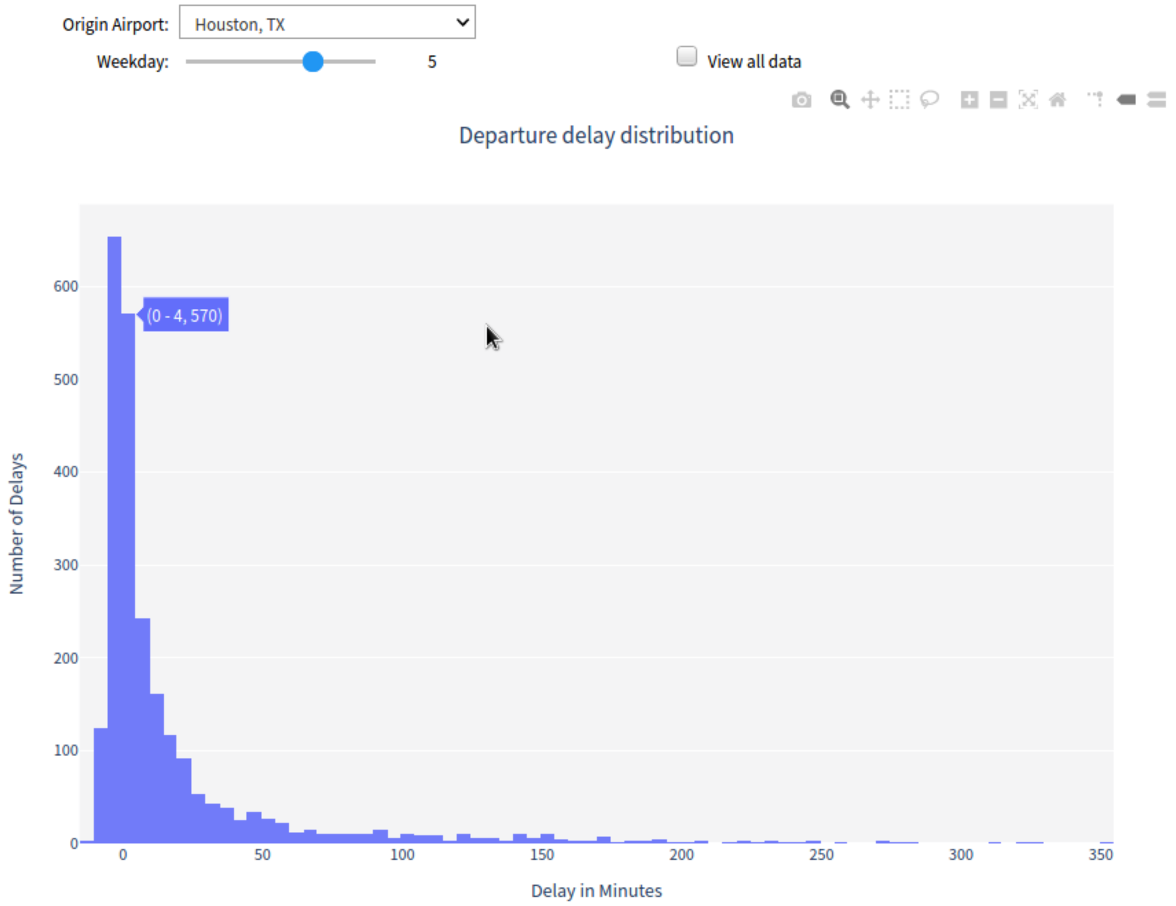


Origin Airport: Boston, MA

Weekday: 6

☑ View all data

In this figure, we are free to choose any airport to see its detailed delay distribution. By clicking the "view all data" checkbox, the user sees the whole distribution, and it's also possible to uncheck the box and use the slider to look at a given weekday. If we hover over a single bar, it will show the delay time range in that bin and how many counts are in the bin.

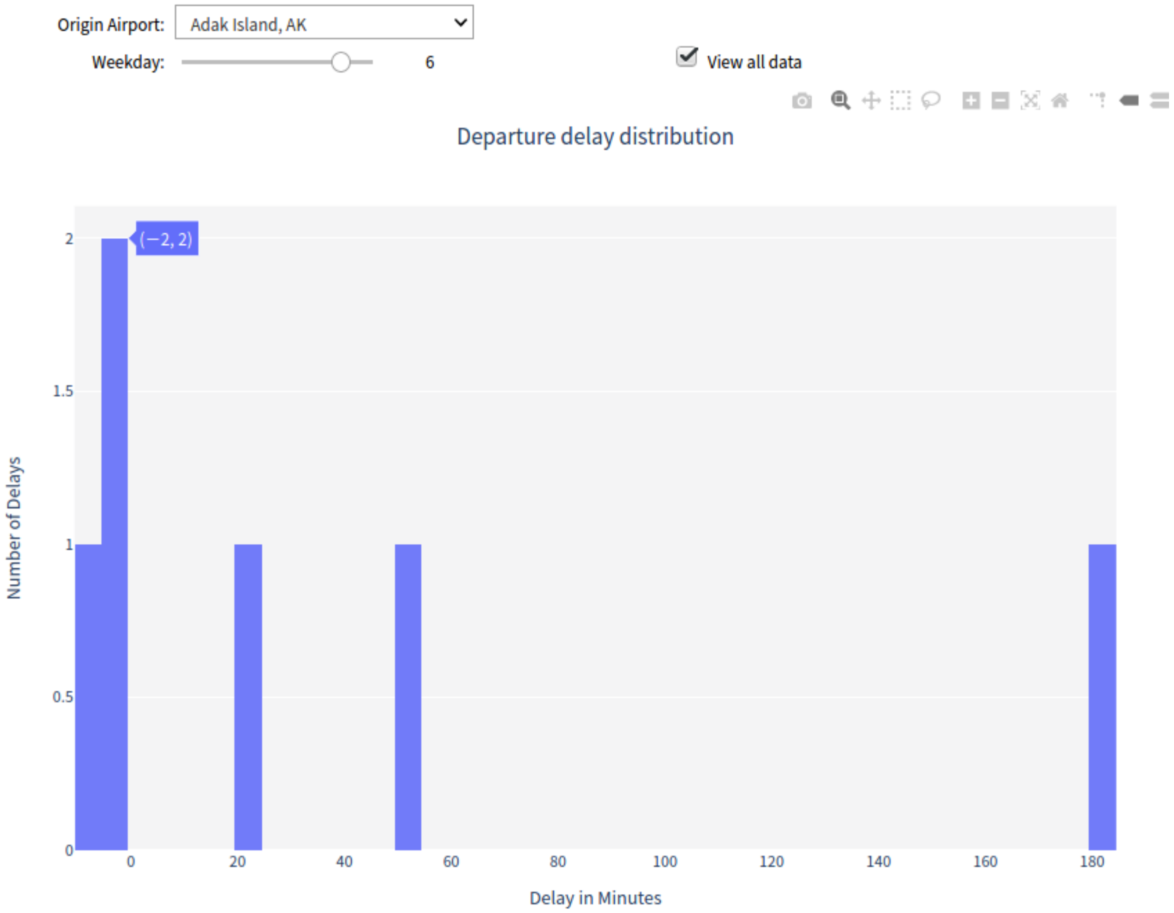


Origin Airport: Houston, TX

Weekday: 5

□ View all data

For example, this screenshot shows that 570 flights from Houston IAH airport in Texas took off within 4 minutes of the scheduled time. The user can also drag the mouse on the horizontal axis to slide the bar plot, or select a random region to zoom in, this makes the graph much more flexible.



Origin Airport: Adak Island, AK

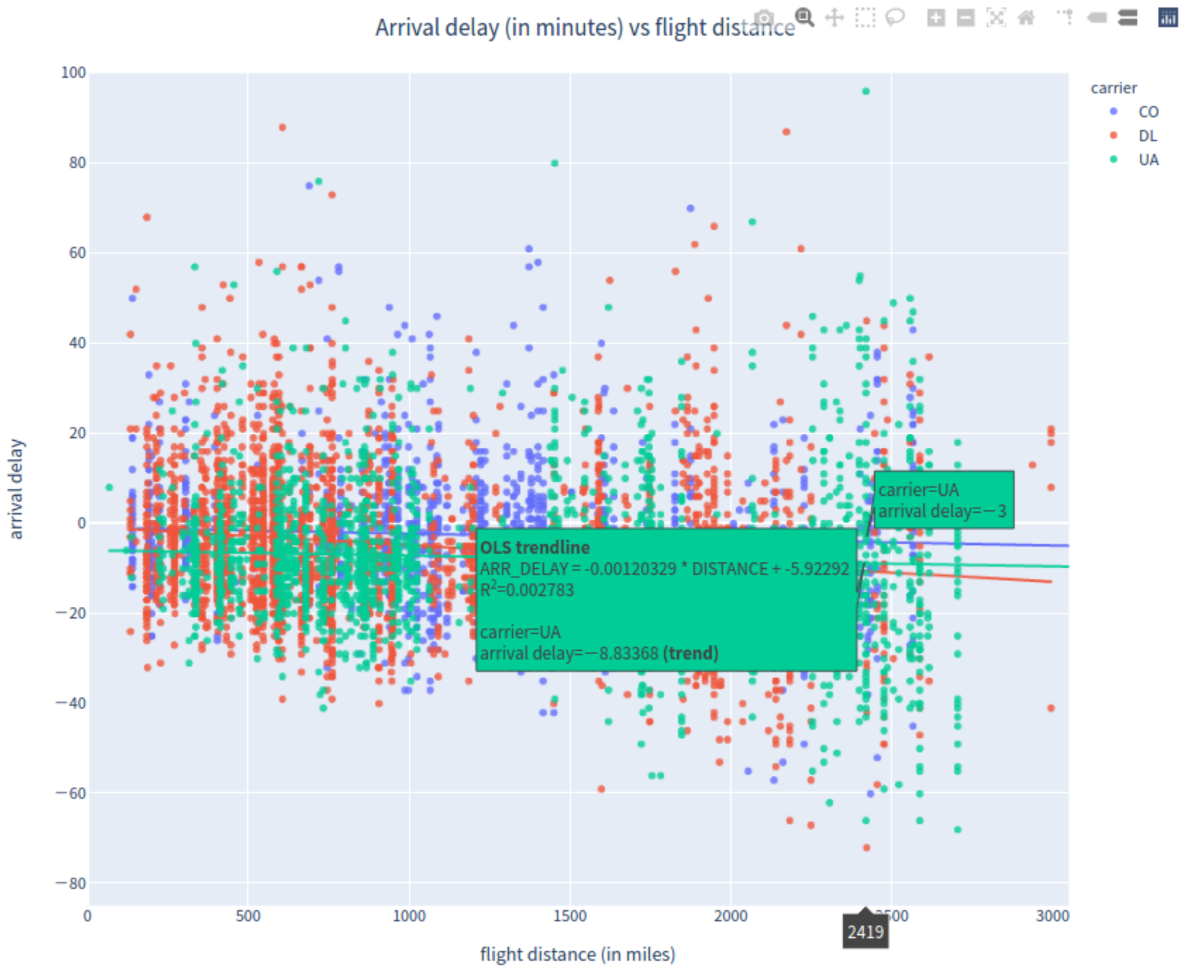
Weekday: 6

☑ View all data

If we look at the origin airport of Adak Island in Alaska, now we see why its average delay is so high. This airport only has 6 observations, in fact 3 of them depart earlier than schedule, 2 are moderately delayed, but there's 1 outlier that delays more than 3 hours which makes the average delay very large, so now we know that this is not a really lazy airport. In general, the graph shows that over 95% of flights from most airports departure within an hour of the schedule, which is totally acceptable. Besides, the distribution on weekdays are very close to that on weekends, showing no significant variation.

Finally, let's look at the relationship between arrival delay and flight distance. For this problem, we are only going to look at flights that took off on time, so that we can rule out the contribution of departure delay. To simplify our problem, let's just focus on the following major airline carriers:

- CO: Continental Airlines
- DL: Delta Airlines
- UA: United Airlines



In this figure, we make a scatter plot that draws arrival delays against flight distances. Each `(distance, delay)` pair in the plot is a small data point, whose color represents the aircraft carrier, specified in the legend. It's clear from the figure that arrival delays randomly line on both sides of 0, for a given distance, we have many positive delays as well as negative, so there's no clear pattern between the two. We also draw the regression line for each carrier, whose slope is almost 0, showing no linear relationship at all.

This figure is also interactive, as we can hover over any data point to see the exact carrier, delay and distance, as well as the trendline equation and its R-squared value.

Despite the result that distance does not seem to impact arrival delays, we do observe some differences between carriers. For short-distance flights less than 1000 miles, we can see that red (Delta airline) and green (United airline) are similar in data size, but green data points are more closely centered around 0, so United airlines delay less frequently in this case. Besides, we can see that as distance increases, the deviation of data also goes up, the spread of data becomes wider. This makes sense because if the airplane flies faster than expected, it is more likely to arrive much earlier in long distance than in short distance, a longer distance and flight time would always make the speed difference more noticeable, as is the case with marathon versus sprint.