For assignment 2, I use 2015 flight delay and cancellation dataset from Kaggle. Details about this dataset and preprocessing steps can be found in “Overview of dataset” section in my Streamlit app and Jupyter notebook. With this dataset, I want to enable users to look at the relationship between delay/cancellation and origin airport/operated airline. This question should be intuitive to many people: Do flights from a certain airline are more likely to delay or cancel? Do flights from some particular airports are more likely to delay?

In order to better serve users’ interests, I put those two questions into two separate pages. In this way, users can clearly see what question can my app answers and they don’t need to scroll down to find the question they are interested in.

For delay vs. origin airport question, I visualize the data over a US map. Each circle is marked on an airport, and the color of the circle indicates the delay rate of that airport. I set darker (red) to higher delay rate airports and lighter (yellow) to lower delay rate airports, which follows human intuition. Users can also hover over the points to check the airport’s name and the actual delay rate (if they are not satisfied by color differentiation).

For delay/cancellation vs. airline question, my goal is to visualize a clear comparison of cancelled/delayed flight count between different airlines. My initial plan was to use a bar chart to visualize it. It didn’t work out well because normal flights (not delayed or cancelled) dominate the dataset, so it’s hard for users to see the difference of cancelled/delayed flight count by human vision (bar too short to see difference). Thus, I changed gear to visualize a dataframe and use highlighting to indicate which airline has worst performance. This method works out well because the dataframe only has a few rows and columns and users should feel fairly easy to interpret it.

I did this assignment solo. Thus, pre-processing of dataset and visualize it in Streamlit are all my own work. I spent about 5 hours looking for a dataset and pre-processing it. Then I spent another 5 hours visualizing data over US map. This part took longer than I expected because there are a lot of map visualization tools available, but I need a one that can differentiate locations by delay rate, and also have hover-over effect to show detailed information. It took me a while to find the right package to use and implement it. Later I spent 4 hours visualizing dataframe. This visualization is much easier than map, but I tried an alternative option first (bar chart), which took me extra time. In the end, I spent about 1 hour cleaning up file, writing this document.