Visualization of Domestic Flight Performance in the United States Process Book

Danyang Chen, Wenwan Yang, Wenshuai Ye

April 17, 2015

1 Overview and Motivation

With the significant increase in the number of flights each year, more and more passengers get stuck at the airport due to delay. According to the investigation conducted in 2014, the aggregate minutes of delay in reaching final destinations amount to over 80 millions. Inspired by this situation, we desire to identify airports and airlines of best and worst performance so that people can choose airlines based on their needs. In this process book, we will cover detailed project objectives, description of our data, exploratory analysis we have done, our design evolution, implementation, and improvement and future works.

2 Project Objectives

At first, we aimed to answer the following questions through data visualization:

- How much time will the flights be delayed in a specific airport?
- How much time will the flights be delayed for a specific airline?
- Whats the most on-time airport in U.S.?
- Whats the most on-time airline?

As we started to dig into the data and implement our designs to working prototypes, we found that our visualization can address more questions and gain insights on:

- How are the airports distributed in U.S.?
- The comparison of the performance between two airports or airlines.
- Will airports with long departure delay always be associated with long arrival delay?
- Are there any airports or airlines which usually depart or arrive earlier than scheduled? (Surprisingly, the answer is yes. The details are introduced in the Evaluation section)

3 Data

There are three types of data used in the visualization: topological json data loaded to construct a US map, airport location data used to map the cities, and detailed flight data for exploration and main visualization.

a **topological json data**
us.json: This is a json file containing the us map info. The original file was downloaded and extracted from United States Census Bureau[1] as an shp file. With topojson installed, we converted the file to a json file.

b **airport location data**
airport.csv: This is a csv file containing the latitude and longitude of each airport. We downloaded the file from a third party website[2].

c **detailed flight data**
*ontime_*2014*xx.csv*: These are csv files extracted and downloaded from Bureau of Transportation Statistics, United States Department of Transportation[3]. Each file contains information on year, month, flight date, carrier, airline id, flight number, original airport id, original airport, original city, original state, destination airport id, destination airport, destination city, destination state, departure time, departure delay time, arrival time, arrival delay time, carrier delay time, weather delay time, security delay time, late aircraft delay time, and national aviation system (NAS) delay time of flights. For this project, we only investigate flights operated in 2014. *ontime_avg_*2014.*csv*: Because we are only interested in the average performance of flights given a certain year, month, carrier, departure airport, and arrival airport, the *ontime_*2014*xx.csv* files are processed to calculate the averages in python. We merged the 12 averages files into one, outputing the *ontime_avg_*2014.*csv* file.

4 Exploratory Data Analysis

We made a draft map with airports attached to it as dots initially to see how the map fits on the screen and how the dots fit on the map. It turns are that there are too many airports (21490), and filtering out some of them may be a plausible way to make the visualization better. We then calculated the total number of routes (summarize by month, carrier, original airport, and destination airport) and found that this number is 73544, confirming that we do need to remove airports that do not contain any route from the map.

5 Design Evolution

a **Show the information of the specific route when clicking on any particular route.** So far we have the US map with each dot representing an airport. We considered to implement the interactive that when the user click on any particular route. The details about the specific route will be shown which include the average departure delay time, average arrival delay time and the rank of the route.

[1] www.census.gov/geo/maps-data/data/cbf/cbf_nation.html
[2] ourairports.com/data
[3] www.transtats.bts.gov

b **Select data of different years.**
We considered to use multiple years data and provide user the option to choose the data visualisation of one specific year.

c **Show bar chart informing the cause of delay.**
Carrier-delay, weather-delay, nas-delay, security-delay and late-aircraft-delay are the top five five reasons for the delay of the airplane. We considered to make a bar chart to compare the delay time of each reason.

d **Visualize data by different states.**
We considered to give user the option to view the data by different state. The data visualisation could compare the average delay time of the airports in different states.

e **Add the data of the economic status of each airline and visualize the relationship between the economic status and the average delay time.**
the economic status and the average delay time. This is the feature followed by visualizing data by different states. After we compared the average delay time by states, we found some states always have longer or shorter time than other states. The economic status of the state would be a factor for that because the richer the state, the busier that the airport will be.

6 Implementation

So far we have finished data cleansing and preparation, and implemented a working prototype of our design based on the real data. As shown in the following screenshot, on the left corner of the page lies a map of the United States, in which each gray dot represents an airport. The bar chart on the right of the map shows the average departure (in blue) and arrival (in pink) delay time of each airport in minutes. The user can sort the airports by name, departure delay time, and arrival delay time. In the design studio of the peer review, we received the feedback that there are too many airports and the bar chart is too long to navigate. We took the advice, and added the select state filter to improve the user experience. Thus, the user can select a specific state and clearly see the performance of the airports in which he or she is interested.

The line chart below the map shows the average delay time in minutes of each airline. Each line represents an airline. The gray block on the right is an optional feature to be implemented: to show bar chart informing the cause of delay.

7 Improvements and feature work

We will continue to implement our design, complete the line chart of airlines to add label denoting the objects, and enable interaction and animation between the multiple charts. For example, when the user clicks an airport, the corresponding point on the map will be highlighted, the flight routes connecting to that airport will be shown, and the airlines line chart will also be updated to only show the per airline information in the selected airport, etc. We have received several feedbacks from both our TF and the peer review session. We will also work on the suggestions and improve our design.

- Find a way to make the visualization clearer. There are too many airports both on the map and in the bar chart. Instead of plotting all of the airports, we are thinking to only show selected airports with significant traffics.

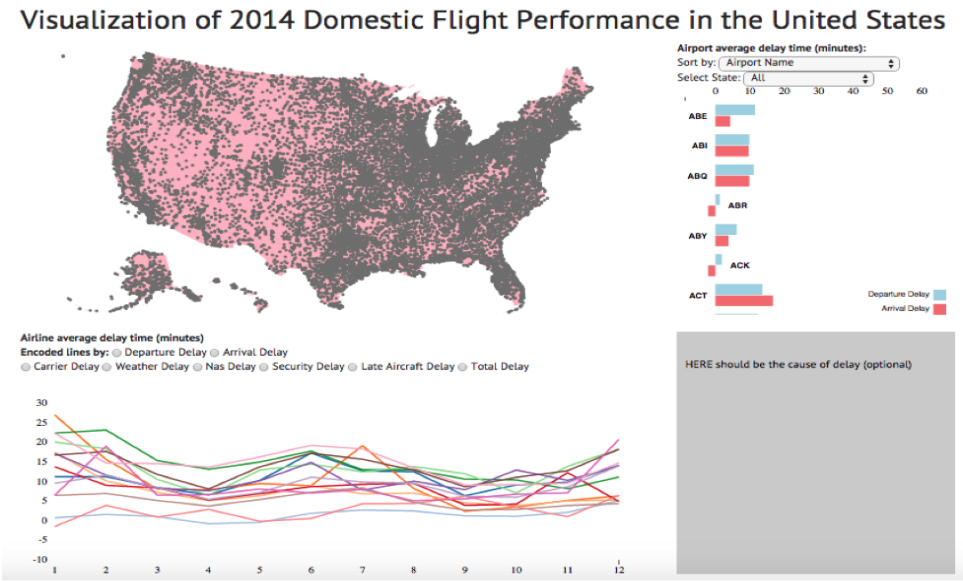

Figure 1: Visualisation

- Think again about the color. People usually regard red as bad and green as good. The pink color in the airport bar chart is likely to mislead the user to the wrong assumption that the arrival delay is always longer than the departure delay. Besides, currently each chart on the page has their own colors. As we continue our implementation, the colors should be reasonable and not confusing.
- Add comparison to the mean. We can add the overall average delay time as reference in both the bar chart and the line chart, so that when the user looks into the details, he or she will not lost the general picture and compare the airport or airline to others average performance.

8 Evaluation

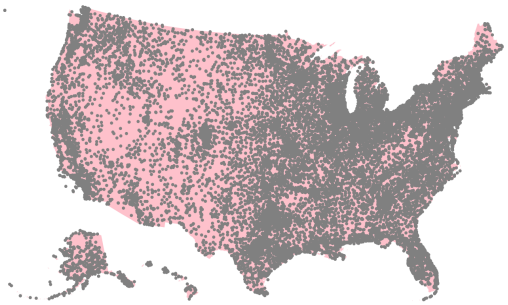

Figure 2: US Map

The first data visualisation is a US map with each dot representing an airport. As we can see from the US map, the eastern part of the United States has more airports than the western part. and the costal city has more airports than the inland city.

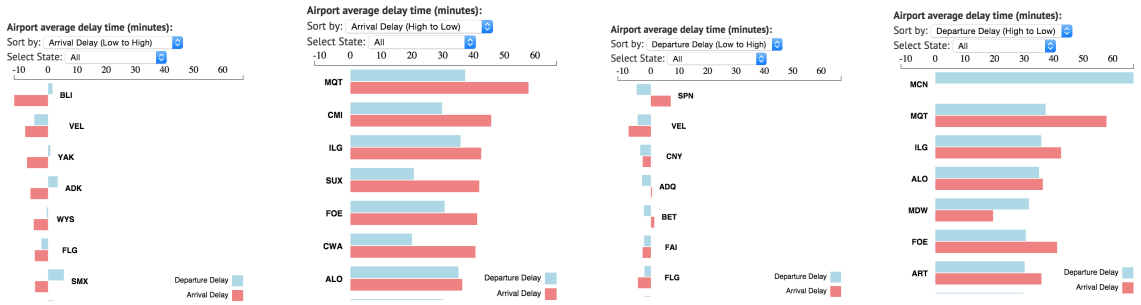

Figure 3: From left to right, shortest arrival delay, longest arrival delay, shortest departure delay, longest departure delay

The second data visualization is a barchart comparing the airport average delay time. *MQT, CMI, ILG, SUX, FOE* are the five airports that have the longest arrival delay.

Likewise, from the bar chart, we find that *BLI, VEL, YAK, ADK, WYS* are the five airports that have the shortest arrival delay.

From the other two figures, we also notice that *MCN, MQT, ILG, ALO, MDW* have the

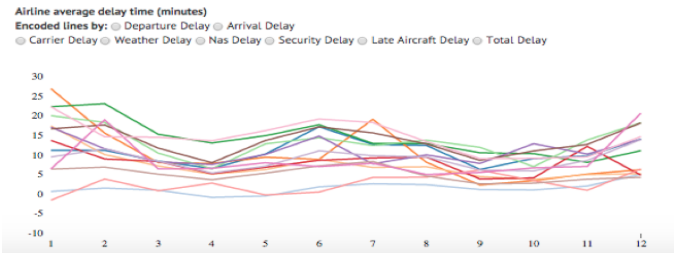

Figure 4: line chart

longest departure delay and *SPN, VEL, CNY, ADQ, BET* have the shortest departure delay.

The line chart compare the total delay time of different airlines. *AS* has the shortest delay time. *B6, F9, MQ, OO* usually have longer delay time than other airlines.

9 Next Steps

Overall, we believe we are in pretty good shape in this visualization project. The next step can be divided into three segments: adding one more chart at the bottom right corner, designing user-friendly interaction elements, and optimizing some of the visualization elements such as color.