CS 6630 Final Project Process Book

Visualization for Flight On-time Performance of the United States

Run Li, Yulong Liang, Zhi Wang

Contents

1. Introduction
   - Basic info
   - Background and Motivation
   - Objectives
2. Data Process
   - Data collection
   - Data cleaning up
   - Data Aggregation
3. Visual Design
   - Description
   - Prototype
4. Implementation

1. Introduction

In this project, we will explore an instance of visualization of the on-time performance data of all the **domestic flights** in the **United States** in **2016**.

➔ Basic Info

➔ Background and Motivation

➔ Objectives

1.1 Basic Info

Project title:
Visualization for Flight On-time Performance of the United States

Group members:
Run Li, u0879939, u0879939@utah.edu
Yulong Liang, u1143816, u1143816@utah.edu
Zhi Wang u0761669, u0761669@utah.edu

Project repository link:
https://github.com/zhiwang93/CS6630Project

In this project, we will explore an instance of visualization of the on-time performance data of all the **domestic flights** in the **United States** in **2016**.

1.2 Background and Motivation

Air travel in the United States has seen a steady rising after the period of post-9/11. By the end of 2016, there were over 5,116 public airports and a total number of 6,676 commercial aircraft in the U.S., which serve more than 2,500,000 passengers everyday.

In 2016, U.S. airlines carried an all-time high number of passengers — 823.0 million systemwide with 719.0 million domestic and 214 million international, which is 3.1 percent more than the previous record high 798.2 million reached in 2015. Moreover, U.S. carrier enplanements are predicted to grow 2.5 percent per year before 2037 according to the Federal Aviation Administration (FAA) Aerospace Forecast.

Despite the rapid growth of aviation industry, the flight on-time performance in U.S. is still unsatisfying: while the percentage of delayed flights fluctuated between 16.7 and 24.1 in the recent 10 years, the average length of delays has increased since 2010 and reached 58.9 minutes in 2015.

1.2 Background and Motivation

Although the Department of Transportation (DOT) requires all U.S. airlines to report on operations to and from only the 29 major airports, all the reporting airlines provide their entire domestic data. These data are published on the website of DOT's Bureau of Transportation Statistics (BTS) for public access.

BTS also summaries and provides monthly reports on the on-time performance of domestic flights. These statistical reports are inclusive and precise but may be too professional and obscure for the public to retrieve information. Moreover, these reports are separated from each other which prevent the readers to have an integrated insight into the data.

1.3 Objective

In this project, we will explore an instance of visualization of the on-time performance data of all the domestic flights in the United States since 2002. We expect to give people an intuitional view and interactive experience of the data, which can make it easier for the public to discover not only the on-time performance in terms of different regions, airports, carriers, months and time slots but also the relationship between distribution of flights and their spatial and temporal conditions. With this tool, passengers can make better decisions on the selection of airports, flight carriers, departure time and whether to buy a delay insurance or not; airlines can gain a comprehensive and comparative perspective on their operation and aviation authorities can explore the overall performance and make policies to promote the development of the entire aviation industry.

1.3 Objective

In this project, we expected to show:

- The connection of each airport to other airports.
- The distribution of flights of each airport in each time period.
- A comparison between planned departing time and actual departing time.
- A comparison between planned arriving time and actual arriving time.
- The rate of diversion and cancellation.
- The temporal and spatial distribution of delayed flights.
- The percentage of causes of delay.
- A view of the time evolution of the flights.

With these visualizations, we will show how complicated the aviation system is and to reveal some relationship between the distribution of flights and their spatial and temporal conditions. We will see which area has the highest density of flights, which airport is the busiest, at what time we may expect a delayed flight, etc.

2. Data Process

Our data are from assorted sources. We have collected records of more than 5 million flights and information of more than 10 thousand airports.

- Data collection
- Data cleaning up
- Data Aggregation

2.1 Data Collection

We collect our data for the project from several different sources:

Geography data:
Census Bureau: www.census.gov

Airport data:
Federal Aviation Administration (FAA): www.faa.gov
OpenFlights: www.openflights.org
OurAirports: www.ourairports.com

Flight operation data:
Buerau of Transportation Statistics: www.bts.gov.
Real-time flight dynamic data: FlightAware: www.flightaware.com

2.1 Data Collection

We collect our data for the project from several different sources:

Census Bureau:
www.census.gov

Geography data data

Federal Aviation Administration (FAA):
www.faa.gov
OpenFlights:
www.openflights.org
OurAirports:
www.ourairports.com

Airport data

Bureau of Transportation Statistics:
www.bts.gov.

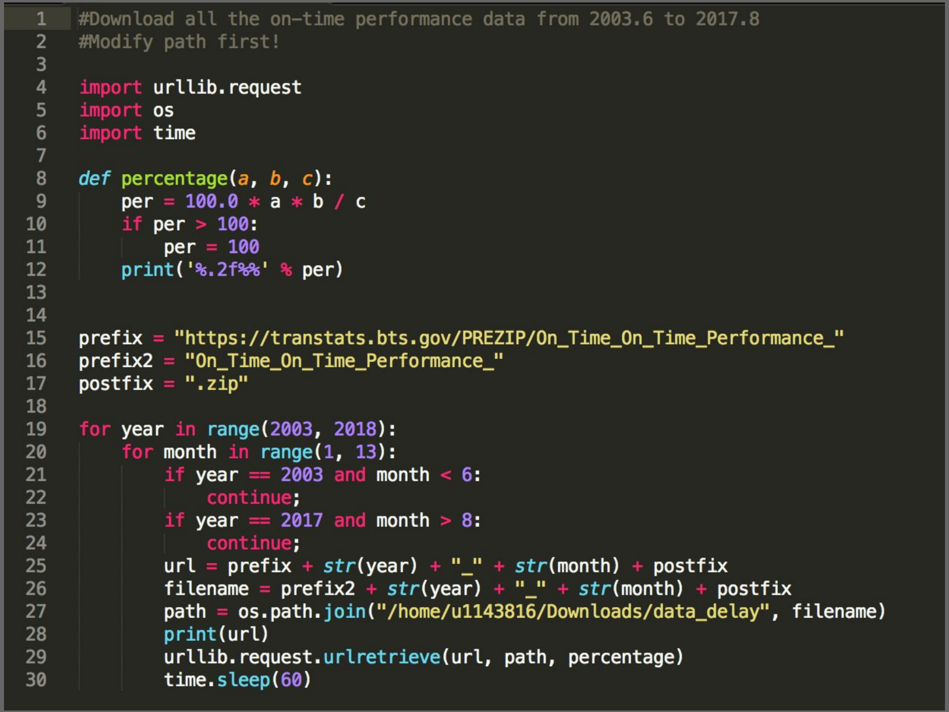
Flight operation data

FlightAware:
www.flightaware.com

Real-time flight dynamic

2.1 Data Collection

To avoid repetitive work, we wrote a simple crawler using python to download pre-zipped data from BTS.



```
#Download all the on-time performance data from 2003.6 to 2017.8
#Modify path first!

import urllib.request
import os
import time

def percentage(a, b, c):
    per = 100.0 * a * b / c
    if per > 100:
        per = 100
    print('%.2f%%' % per)


prefix = "https://transtats.bts.gov/PREZIP/On_Time_On_Time_Performance_"
prefix2 = "On_Time_On_Time_Performance_"
postfix = ".zip"

for year in range(2003, 2018):
    for month in range(1, 13):
        if year == 2003 and month < 6:
            continue;
        if year == 2017 and month > 8:
            continue;
        url = prefix + str(year) + "_" + str(month) + postfix
        filename = prefix2 + str(year) + "_" + str(month) + postfix
        path = os.path.join("/home/u1143816/Downloads/data_delay", filename)
        print(url)
        urllib.request.urlretrieve(url, path, percentage)
        time.sleep(60)
```

2.2 Data Cleaning up

The original data are .csv file format which do not need further process.

However, the original files are elaborate so that there are many fields which will not be analyzed in this project, such as the city information, airline information and other statistics.

Hence, we separated the original huge data tables into small ones with fewer fields, to avoid meaningless time cost on data loading. We store all the new datasets in .csv format in the ‘dataset’ folder.

2.3 Data Aggregation

We have three main types of datasets:

- **Map data**
  For generating US map, no need to modify.
- **Airports and Flights Aggregated Data**
  For drawing circles (indicating airports) and lines (indicating flights on the map as well as showing brief information when clicked. Need to do aggregation of bts data then join with airport information data.
- **Airports and Flights Statistical Data**
  When the reader clicked a particular airport, corresponding statistical data will be loaded to generate charts underneath the US map.

2.3 Data Aggregation

Airports information

We choose to select IATA Code, Name, City, Country, Longitude and Latitude for each airport then we started to collect those so that we can use the airport data to locate the positions on the map with the longitude and latitude of each airport and link it with the bar charts.

Flights information

For the each flights we have its Origin, Destination, Flight counts, Departure delay, Arriving delay and Cause of Delay. For analyzing the causes of flight delay, carrier delay, weather delay, NASA delay, security delay and late aircraft delay are calculated respectively. We will focus on using flights data sets for generating the bar chart flights for each airport.

3. Visual Design

We use multiple features in order to have a diverse design.

- Description
- Prototype

3.1 Description

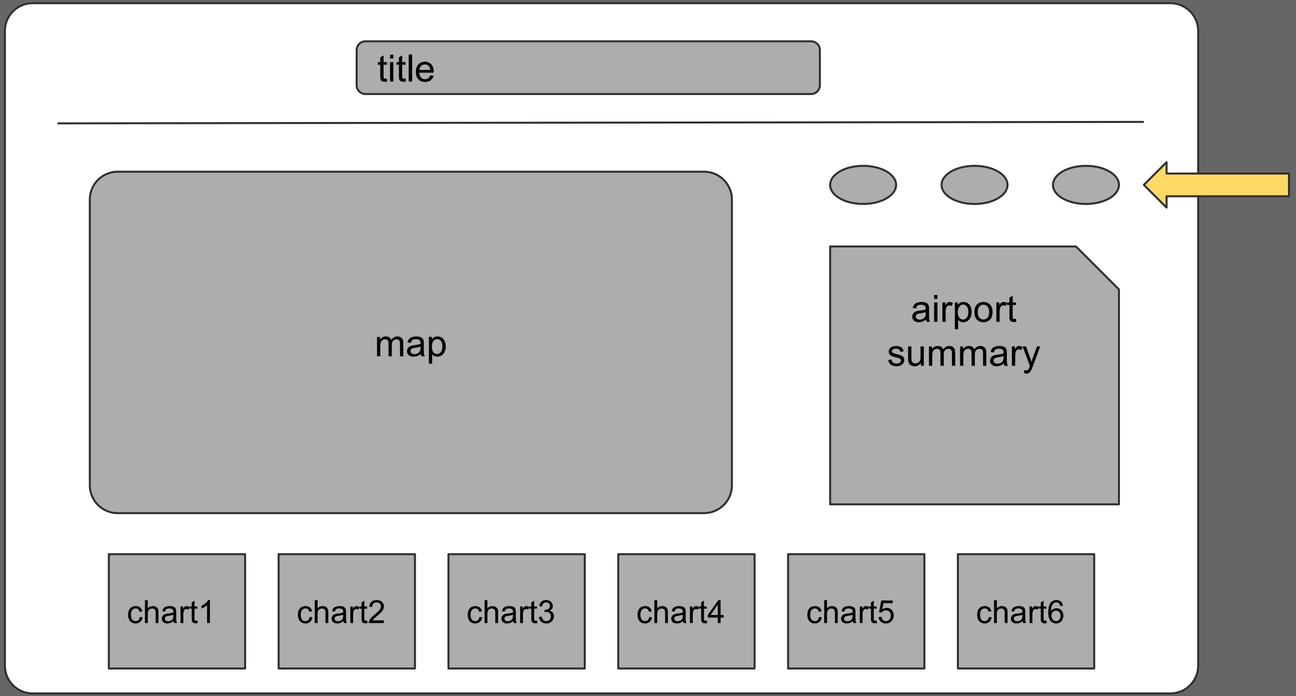
Our document tree is in the following structure:

```
..
/dataset
        airportconnection.csv
        January.csv
        ...
/js
        Script.js
        Usmap.js
        ...
/css
        Style.css
HTML.html
```

The HTML.html file is the final display. All the designs and features will be shown here. The /dataset folder is to store all the data we need. /js folder is to store all our logical coding work and /css folder is to store our color schemes.

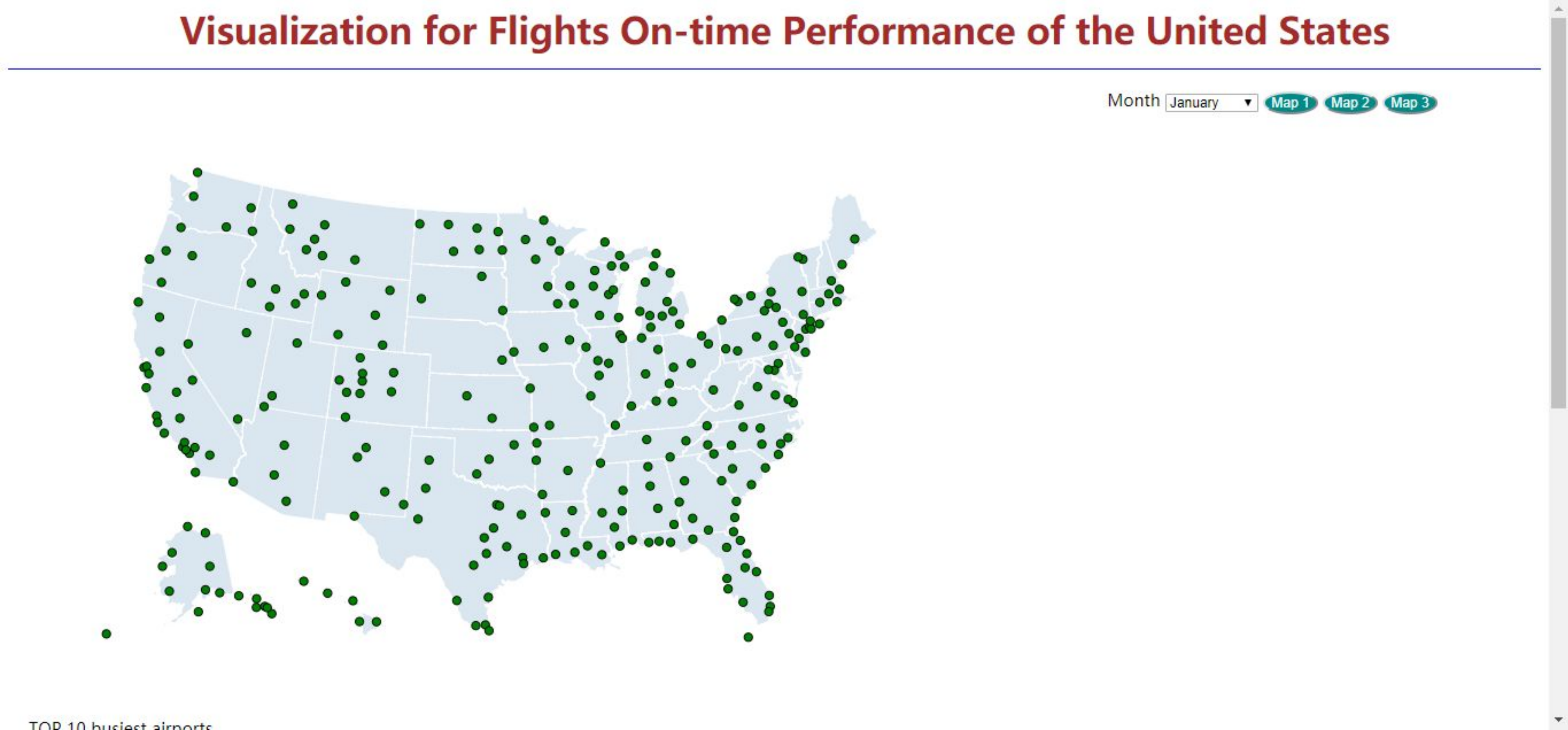
3.2 Prototype

The final look of our website will be similar to this frame:



Map selector

Current Website View

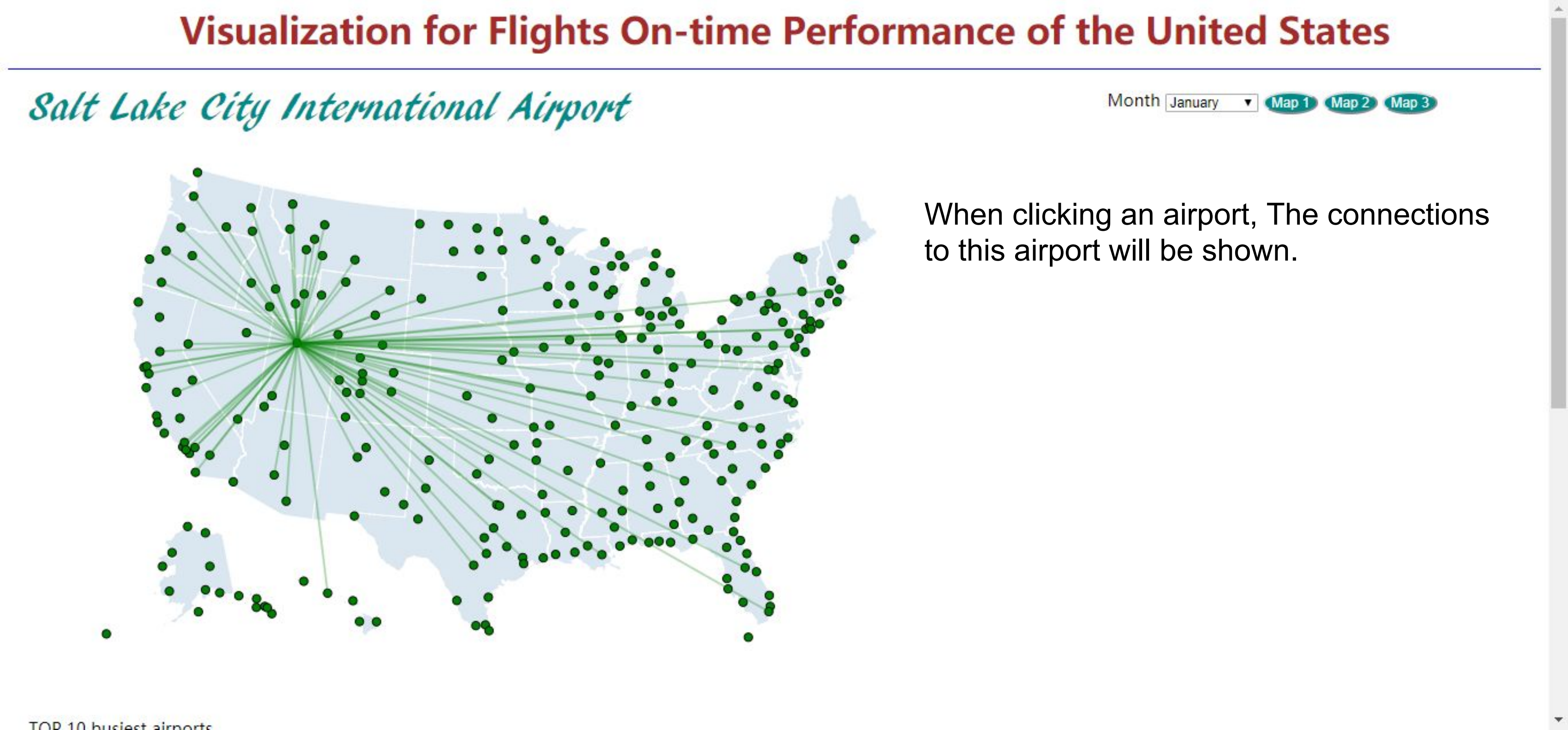


Visualization for Flights On-time Performance of the United States

Month January

TOP 10 busiest airports

Current Website View



Visualization for Flights On-time Performance of the United States

Salt Lake City International Airport

TOP 10 busiest airports

When clicking an airport, The connections to this airport will be shown.

More is coming...