

Visual Analytics - Final Project

Engineering in Computer Science

Prof. Giuseppe Santucci

TA: Marco Angelini

Due Date: 10/06/2018

Valerio Colitta - 1656690

Davide Spallaccini - 1642557

Flights And Delays In The United States

Introduction The airline industry is one of the fields that continuously produces a large amount of complex data regarding the large number of flights that take place every day. In this work we concentrate only on flights of a single year in the United States. Even considering a single nation and a single year the number of flights is really high, so in this work we aim at providing a visualization of this data that helps in better understanding patterns in the number of flights in the different states and in the delays that were recorded across time.

Data We got our data from the United States Department of Transportation, and in particular from the Bureau of Transportation Statistics (BTS) website. As the independent statistical agency within the Department of Transportation (DOT), the Bureau of Transportation Statistics (BTS) is a politically objective supplier of trusted and statistically sound baseline, contextual, and trend information used to shape transportation policy, investments, and research across the U.S. and abroad. BTS is the preeminent source of statistics on commercial aviation, and transportation economics. In particular we used the On-Time Performance table (available at ...) under the Airline and Airports subsection of the website. This table contains on-time arrival data for non-stop domestic flights by major air carriers, and provides such additional items as departure and arrival delays, origin and destination airports, flight numbers, scheduled and actual departure and arrival times, cancelled or diverted flights, taxi-out and taxi-in times, air time, and non-stop distance.

It is a very large dataset, that contains more than 5 millions records of flights in the United States. The main features are:

- **ARR_DELAY** : minutes of delay on the arrival for the flight
- **DEP_DELAY** : minutes of delay on the departure for the flight
- **AIR_TIME** : minutes of air-time
- **DISTANCE** : distance in miles covered by the airplane
- **CARRIER_DELAY** : Carrier delay in minutes
- **NAS_DELAY**: National Air System delay, in minutes
- **LATE_AIRCRAFT_DELAY** : Late aircraft delay in minutes
- **ORIGIN_STATE_ABR** : State of departure abbreviation

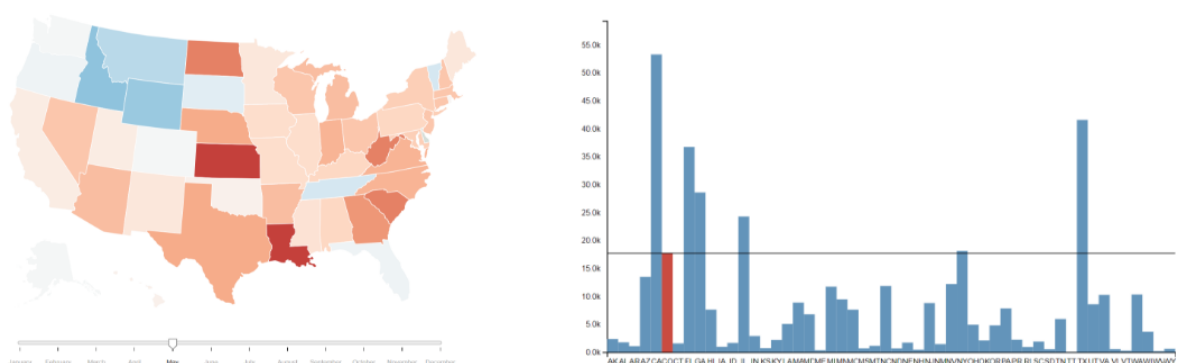
Because of the huge amount of data, we couldn't just load them via d3 or javascript; we tried, but the time to load it was just too much. Therefore we decided to do some preprocessing on the data, before loading them.

Flights were grouped by **month**, **state** and counted to use those values in the USA map. The same happened for the DD scatter plot

In addition to that, we randomly sampled about a thousand records, in order to plot the single instances on the parallel coordinates, and perform the PCA on them.

State Patterns The first visualization elements we are going to describe have the goal to display for each state the amount of airline departures. In particular we want to visualize how many flights leave each state during the different months of the year and how these numbers vary from state to state. For the first task we drew a map of the USA and colored each state according to the number of flights for the selected month with a heat map. The color of each state in this case is independent from the color of the other states, it only depends on the mean of flights for that state during the year. A *hot* color tells that in that month the number of flights was higher than the average while a *cold* color that the number of flights was respectively low. A slider under the map allows the user to change the month under consideration. On the other hand, in order to visualize how the number of flights differs from state to state in a quantitative fashion we display a bar chart with a bar for each state whose height is proportional to the number flights in that state. When you click on a bar the bar becomes red and a line appears cutting all the other bars at the height of the selected one, highlighting the differences; if then you over the mouse on other bars the difference number (positive or negative) is shown.

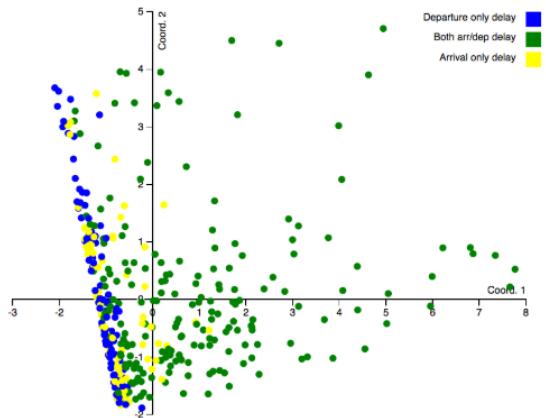
We found interesting that some patterns appear really clear in the map, for example in February almost everywhere the number of flights is below the average, on the contrary in August this number is far over the average probably highlighting the departures due to holidays. In July the same observations hold, and in particular northern countries tend to have higher increase of flights. An exception to the summer excitement is Florida, probably because it is one of the destination for holiday travels and local population usually stays in the country.



Delays The other visualization elements we devised concern flights delays. We created three views:

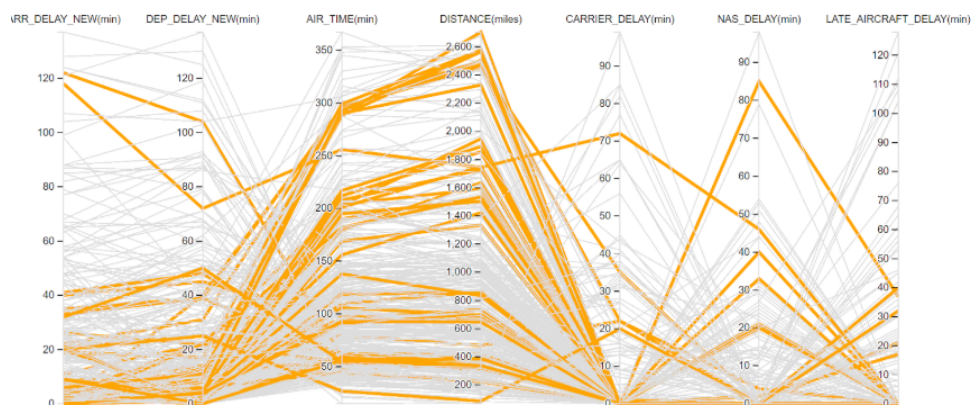
- PCA plot
- Parallel Coordinates
- DD scatter

PCA plot This plot displays in 2D the records sampled in the dataset. In order to project them in 2D we applied PCA via a python script. Out of all the sampled elements we took only the ones that had some delay at the arrival or at the departure.



Once plotted we categorized them in three classes (Departure delay only, Arrival delay only, Both Arrival/Departure delay). An interesting result, was to notice that flights form clusters, and thus there is a good chance that they are close each other also in higher dimensions.

Parallel Coordinates The same data used in the PCA plot were used in the parallel coordinates. PC are a standard method to visualize high dimensions on a 2D environment. For this view we decided to use almost only numeric features that concerned delays. This was made because they were the only numeric values available.

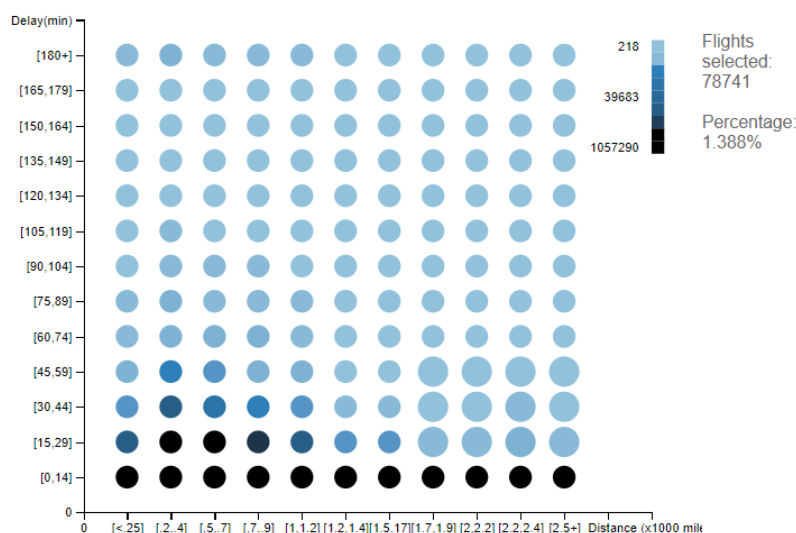


To give a better user-experience, axis can be moved around and interval of interests can be selected.

This view is also coordinated with the USA map. Whenever we want to know what are the flights departed from one of the states, we can filter those by clicking on the state in the map. The lines regarding these flights will be highlighted in orange.

DD scatter The last view in the second group is the DD scatter. It displays relationships between flight delays and their distances. We decided to use a scatter plot where on the x-axis there are all the possible intervals of distance, while on the y-axis the intervals for the delays. Each cell represents the **number of flights whose distances and delays lie in that intervals**.

It is colored according to a heatmap, in order to distinguish the quantities.



Group of point can be selected, and the exact number, together with the percentage out of all flights they occupy is displayed. This is useful in order to "query" the dataset. For example, thanks to this view, we noticed that during the whole year of 2017, the number of flights that did not delayed more than 15 minutes is the 81%, while those for which the delay was greather than **three hours** are just the 1%, that means about 60.000 flights. Another data extracted thanks to this view, is that most of the flights tend to cover about 500 - 1000 miles.

Lastly we notices that the delays tend to decrease with the increase of the flight's distance.

Why using a scatter plot ? We opted for this, over a correlation matrix for two main reasons. First, it works on two different features and then it is not symmetric. A correlation matrix is symmetric.

Second reason, we wanted it to brushable, to have some interaction with it. In a correlation matrix this would be possible, but in a limited way

Conclusions (Slides for this project are available [here](#)).