Individual Learner

US Domestic Flight Delays

Ian Gledhill

03/06/2019

Introduction

Overview

This project will look at the delayed flight data recorded in the US for the 3 months from January to March 2018 and use data analysis to create a model to predict whether a flight would be delayed from its scheduled departure time. The data is provided by the US Department of Transport and covers 16 carriers operating from 336 airports.

Dataset

Source

The data is provided by:

https://www.transtats.bts.gov

BUREAU OF TRANSPORTATION STATISTICS U.S. Department of Transportation 1200 New Jersey Avenue, SE Washington, DC 20590 855-368-4200

The page to download the data used in this project can be found here:

https://www.transtats.bts.gov/Fields.asp?Table_ID=236

Sample Size for Analysis

The DoT record more than 500,000 domestic flights per months. The period covered by this project contains 30,000 data points. Each data point has 33 measurements.

Shape of the Data

The data that the DoT make available falls into the 11 groupings shown below (the number of measurements in each group are shown in brackets). Not all the groupings are relevant to this project, which is only concerned with delayed departures; for example, the arrival performance at the destination airport is not relevant. The groupings that are considered useful are shown in bold.

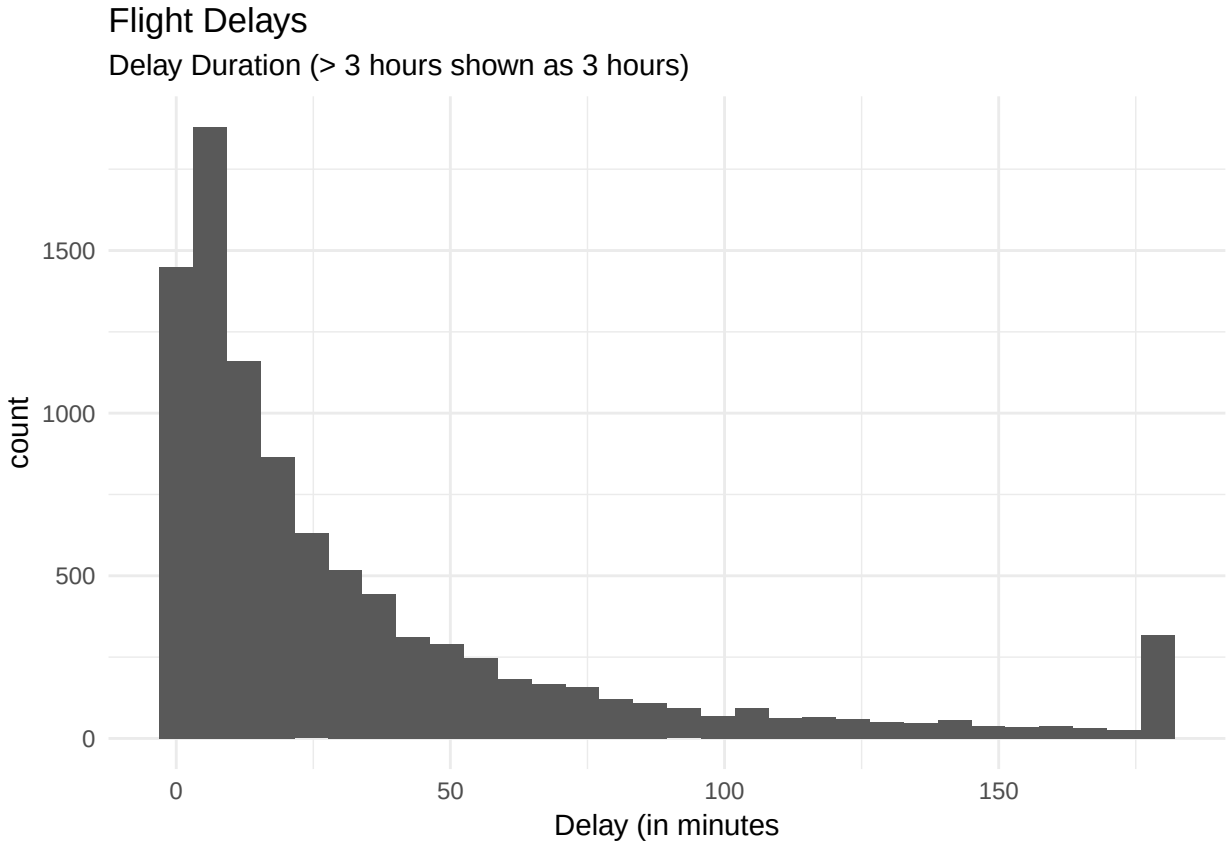
- **Time Period** (6 catagorical)
- **Airline** (5 catagorial)
- **Origin** (9 catagorical)
- Destination (9 catagorical)
- **Departure Performance** (3 catagorical, 6 continuous)
- Arrival Performance (3 catagorical, 6 continuous)
- **Cancellations and Diversions** (3 catagorical)
- **Flight Summaries** (3 catagorical, 3 continuous)
- **Cause of Delay** (5 catagorical)
- **Gate Returns** (3 continuous)
- Diverted Airport Information (4 x sub-groups reprenting each diversion with 4 catagorical and 3 continuous)

Overview of the Data

There are 30,000 departures recorded in total for the period covered by the data. Of these 29,196 have recorded departure times, which this report refers to as *active flights*. The other 804 have no departure recorded, mostly because the flights were cancelled.

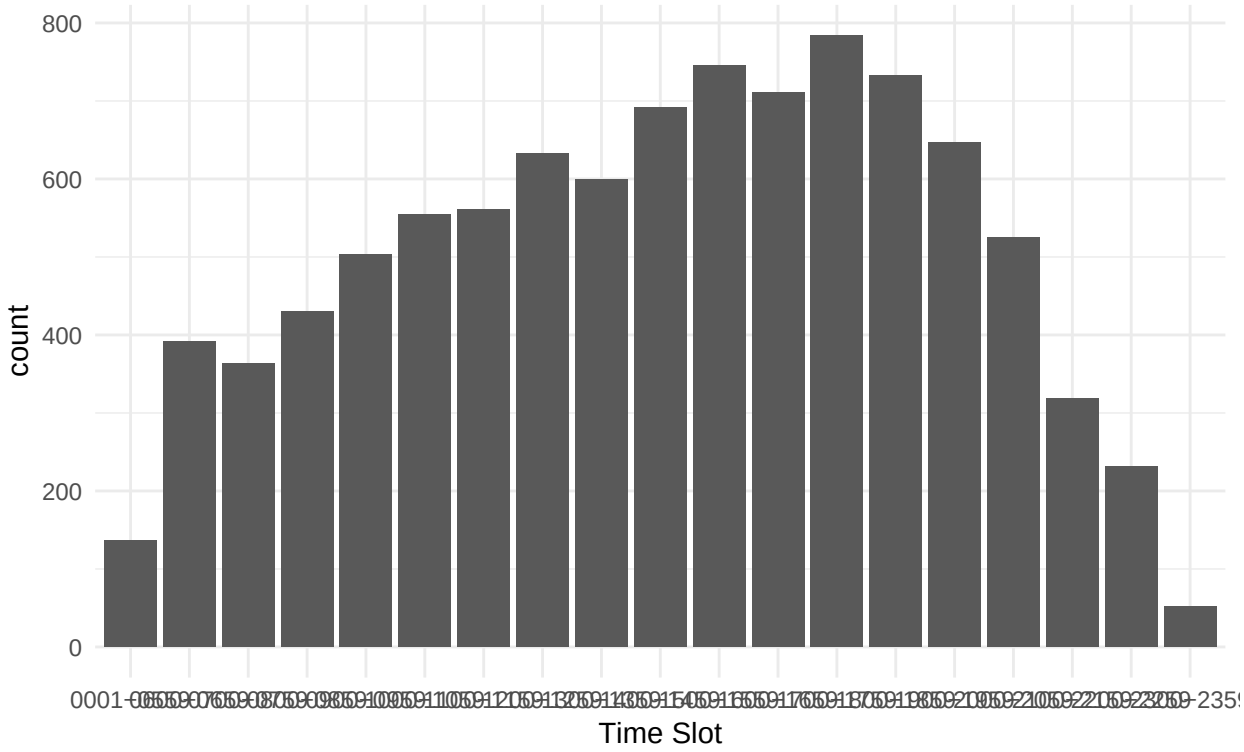
Of the *active flights*: * 62.7% of active flights departed ahead of their scheduled time * 4.40% of active flights departed at their scheduled time * 32.9% of active flights departed after their scheduled time

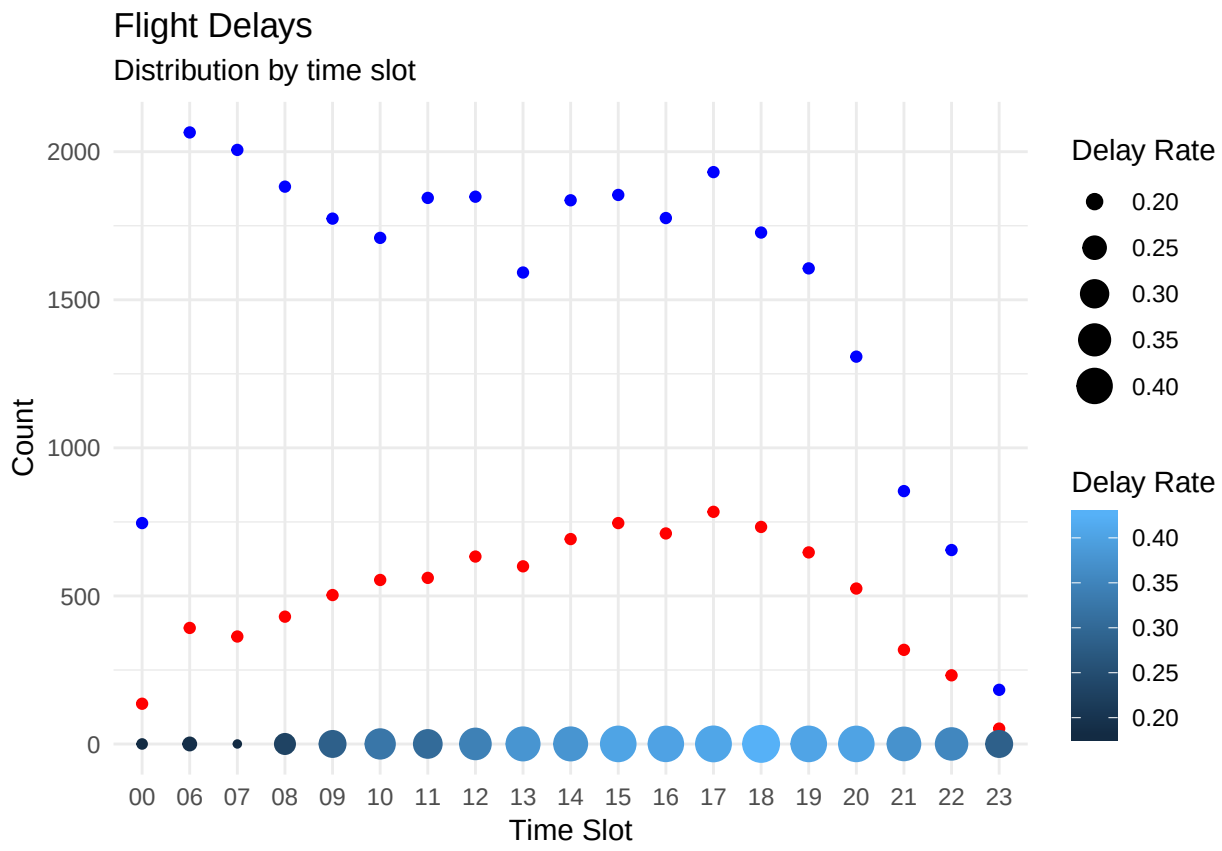
This project is only considering *delayed* departures, i.e. where the recorded actual departure time is after the scheduled departure time.

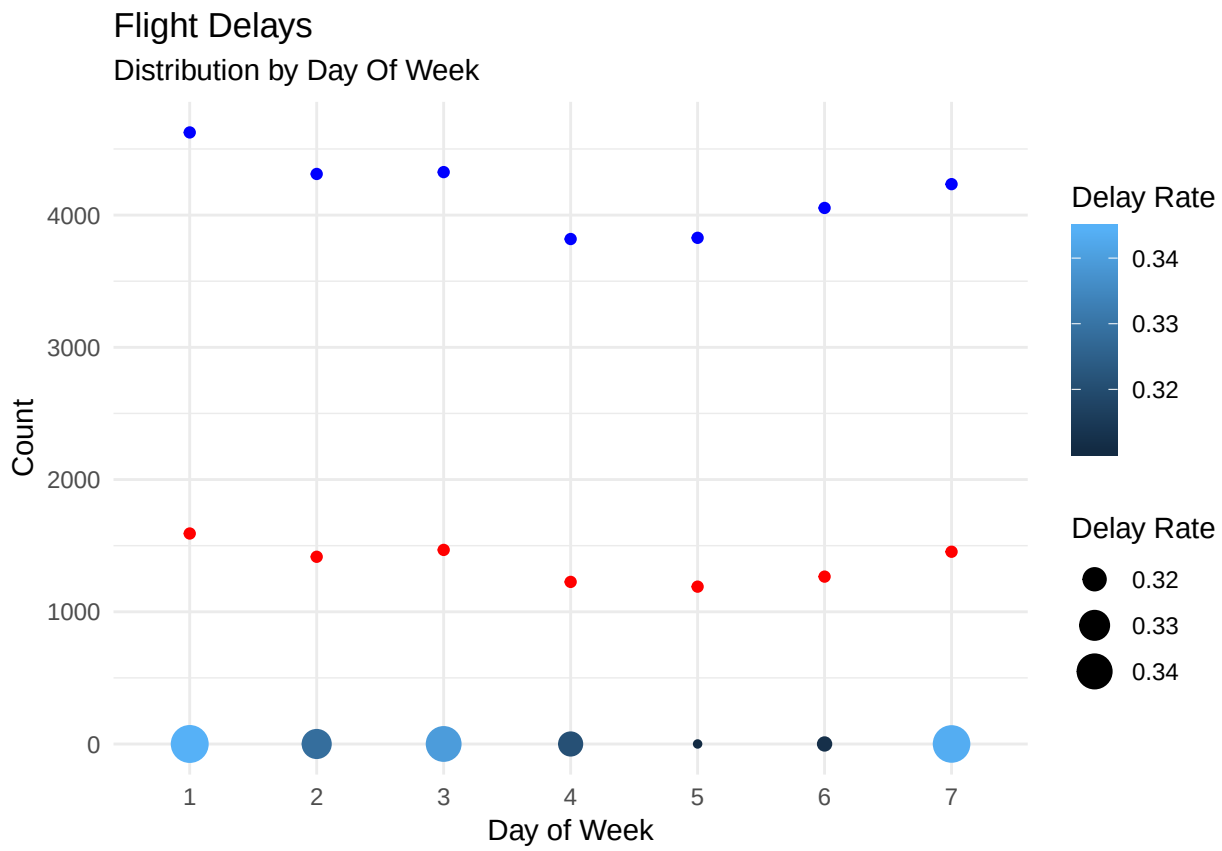


We can see that most flight delays are less than 50 minutes.

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```


Flight Delays
Distribution through the day
800
600
400
200
0
count
0001
Time Slot


Flight Delays
Distribution by time slot
Count
2000
1500
1000
500
0
00 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Time Slot
Delay Rate
0.20
0.25
0.30
0.35
0.40
Delay Rate
0.40
0.35
0.30
0.25
0.20



Training and Test Data Partitions

Goal

The goal of this project is to create a prediction model that would estimate the probably deviation from the scheduled departure time for any domestic flight from any US airport covered by the DoT data.

Approach

Data Cleansing

Cancelled Flights

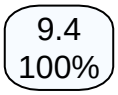
The aircraft will either leave early, leave on time or be delayed. The expected departure time is given by *CRS_DEP_TIME* and the number of minutes relative to this time of the actual departure is given by *DEP_DELAY*. In some case the *DEP_DELAY* will not be available (and be given a value of N/A in the data). This is normally due to the flight being cancelled but can be not available even though the flight is not recorded as being cancelled.

Modelling Techniques

The data set is quite large and has many features. This makes it computationally expensive. To aleviate this Principal Component Analysis (PCA) will be used to reduce the number of features.

An initial attemp to use SVM (*Support Vector Machines*) proved to be too computationally expensive to be useful. CART (*classification and regression trees*) has been adopted.

After training the model the best CP (*complexity parameter*) was found to be 0.2.




```
##  .outcome
##       9.4 null model
```

Results

Conclusion

Bibliography and References