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 4 carriers operating from 146 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 35 measurements.

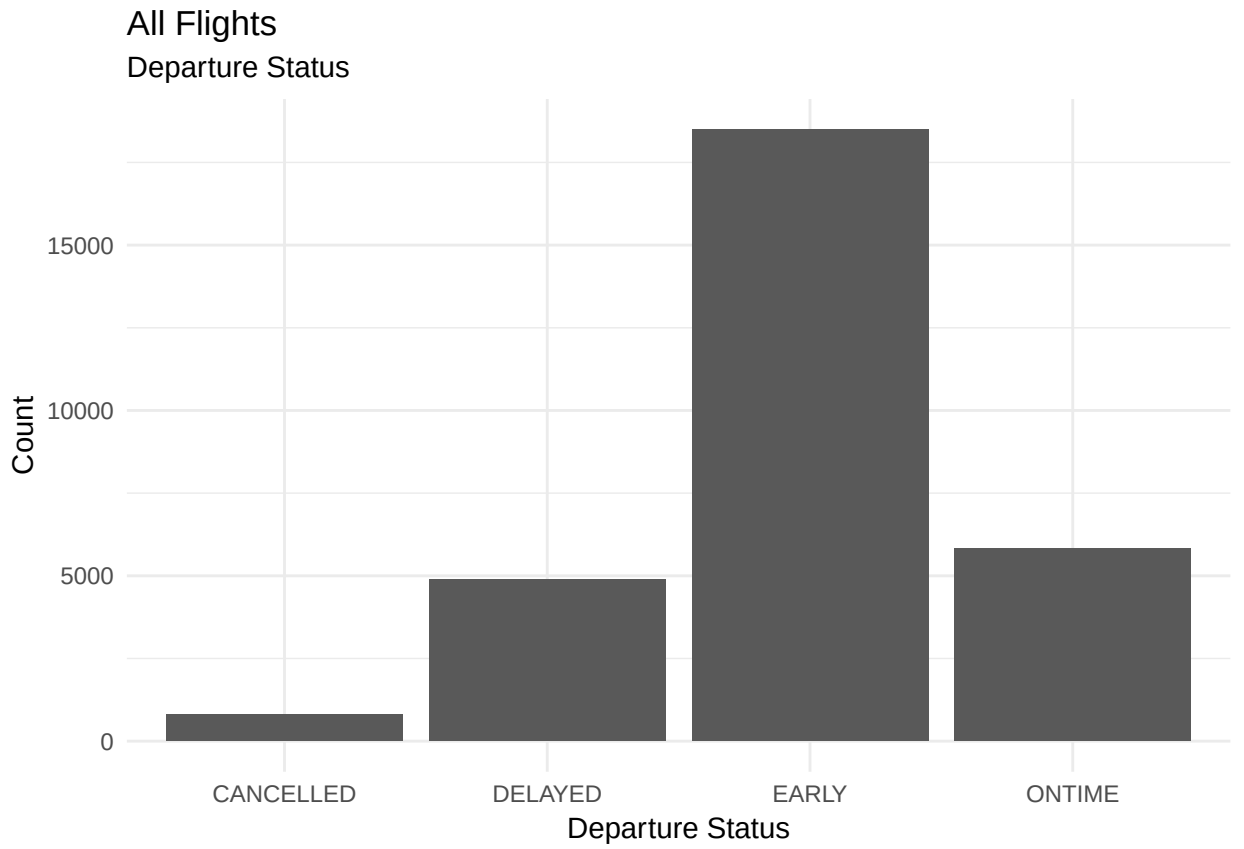
Shape of the Data

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.

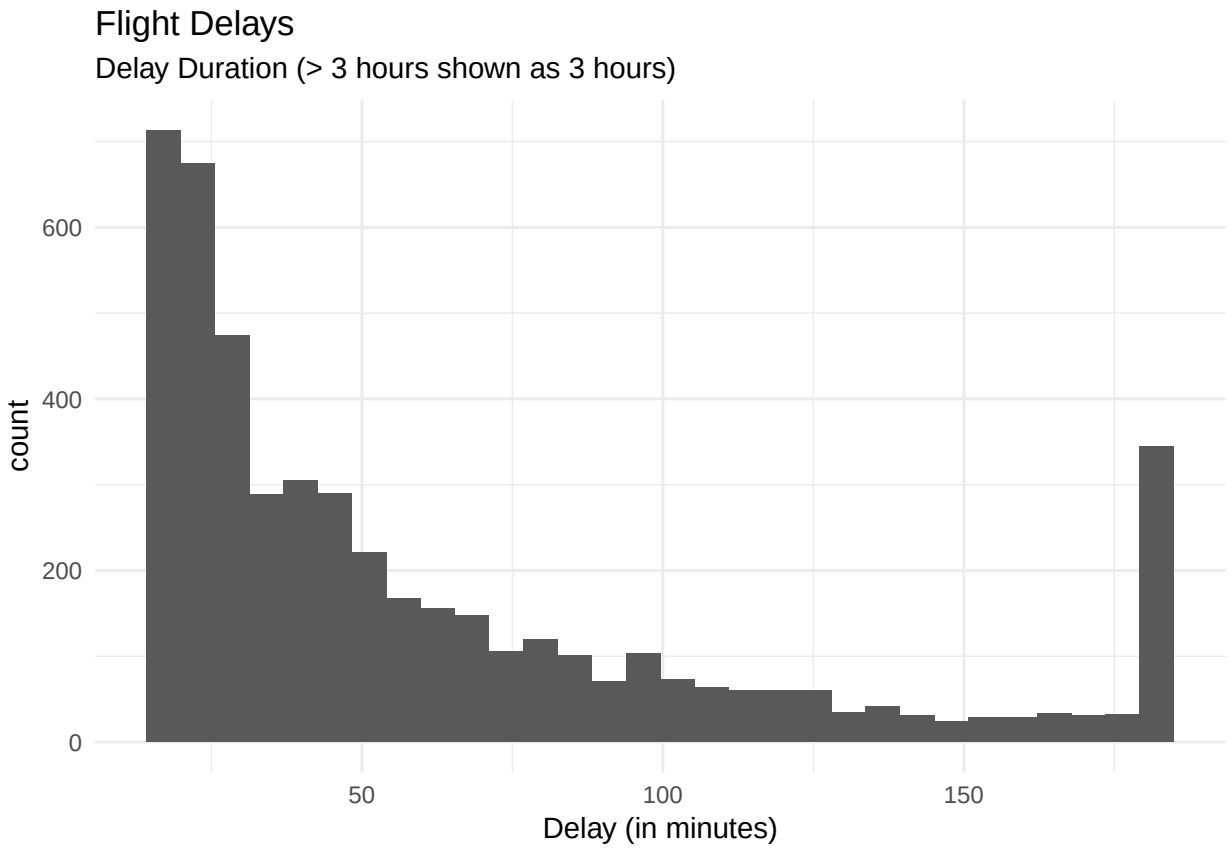
- **Time Period** (6 catagorical)
- **Airline** (5 catagorial)
- **Origin** (9 catagorical)
- Destination (9 catagorical)
- **Departure Performance** (3 catagorical, 6 continuous)
- Arrival Performance (3 catagorical, 6 continuous)
- **Cancellations and Diversions** (3 catagorical)
- **Flight Summaries** (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 deomestic departures from US airports in the period covered by the data. The break down of these flights by departure status is show below:

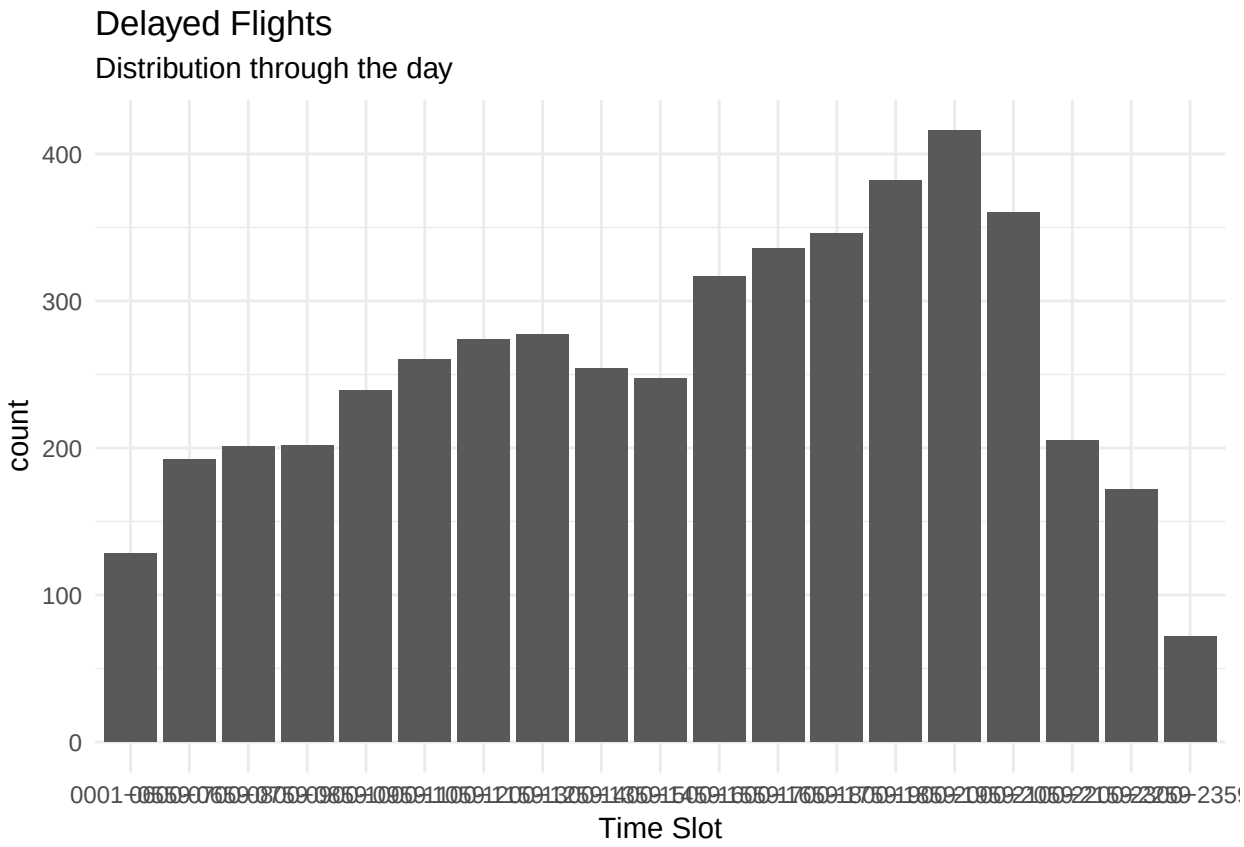


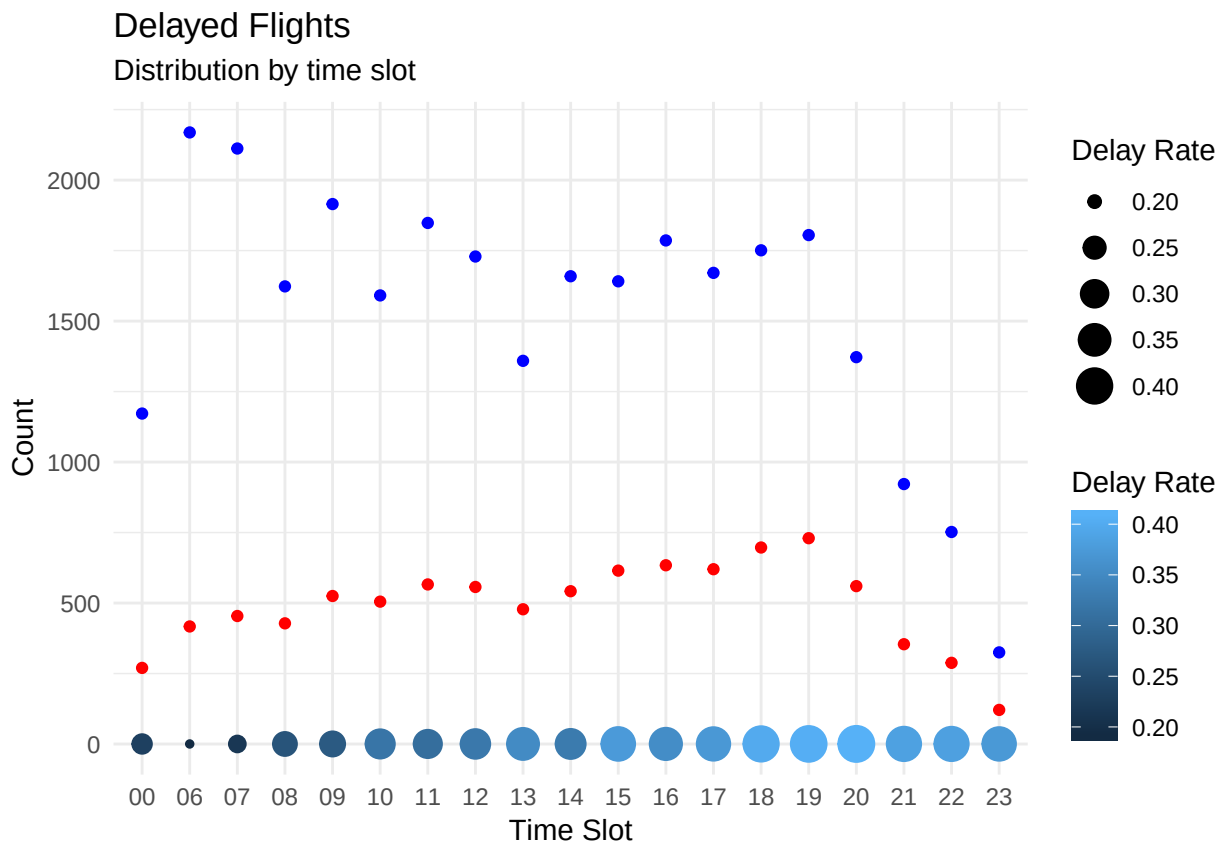
Of the *active flights*: * 61.6% of flights departed ahead of their scheduled time * 19.4% of flights departed at their scheduled time * 16.3% of flights departed after their scheduled time



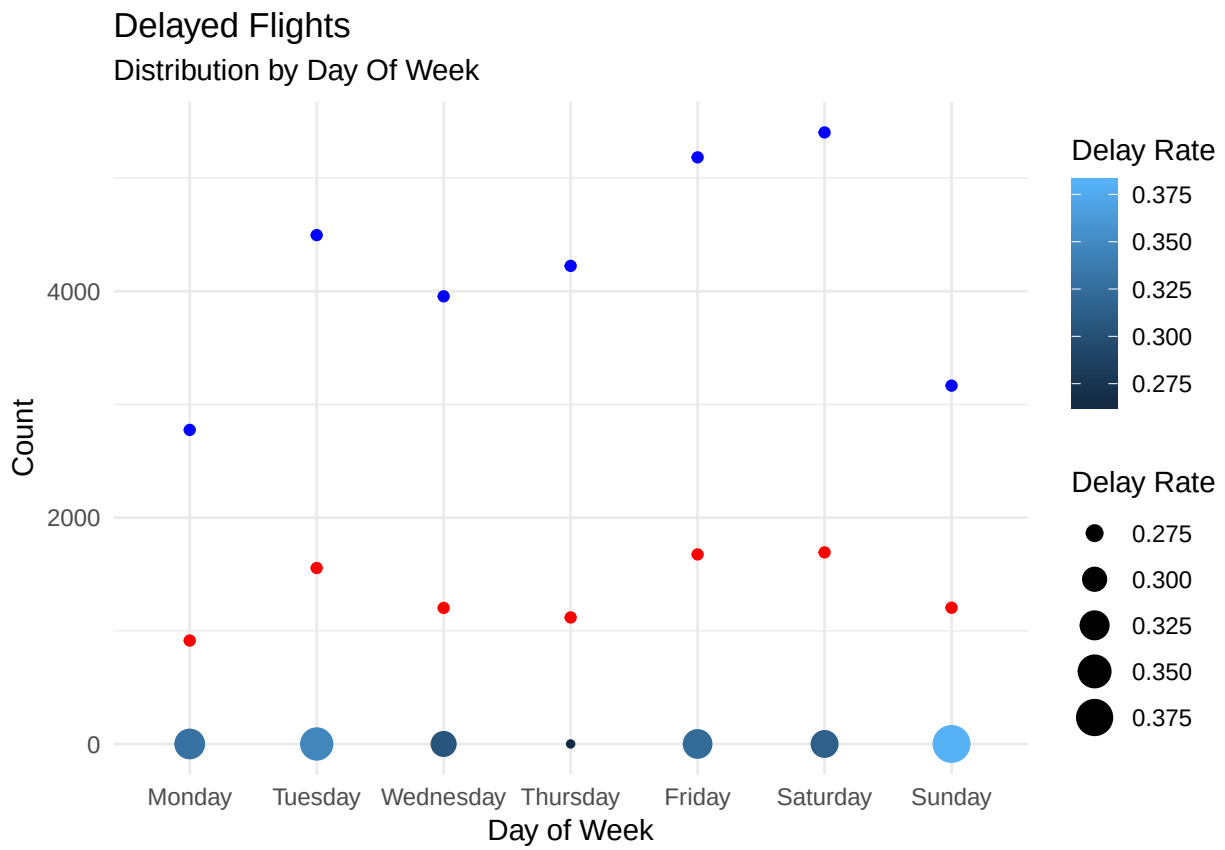
There are quite a small number of outliers which have exceptionally long delay times. These have been included in the 180 minutes bin to make the histogram more readable. We can see that most delays are within 30 minutes.

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


Delayed Flights
Distribution through the day
count
400
300
200
100
0
Time Slot



This shows the number of departures in an hour slot (in blue) and the number of delays (in red).



This shows the number of departures by day (in blue) and the number of delays (in red).

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

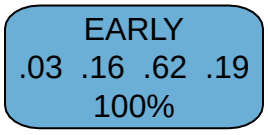
The aircraft will either leave early, leave on time or be delayed. The expected departure time is given by *CRS_DEP_TIME* and the number of minutes relative to this time of the actual departure is given by *DEP_DELAY*. The number of flights cancelled is 808. As these planes don't have a departure time the *DEP_DELAY* for the flight will not be available and will therefore be given a value of N/A in the data.

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  CAN DEL EAR ONT
##     EARLY [.03 .16 .62 .19] null model
```

Results

Conclusion

Bibliography and References