Flight Delay Prediction

Springboard Capstone Project 2

Problem Statement

Determine how many minutes early or late your flight will depart from its destination.

The Data

Bureau of Transportation Statistics tracks flight data and delays:

https://www.transtats.bts.gov/Tables.asp?DB_ID=120&DB_Name=Airline%20On-Time%20Performance%20Data&DB_Short_Name=On-Time

NOAA has weather information for download

https://www.ncdc.noaa.gov/cdo-web/datasets

Wrangling

The data was surprisingly clean.

Decision to use only departure data for Washington Dulles International Airport (IAD)

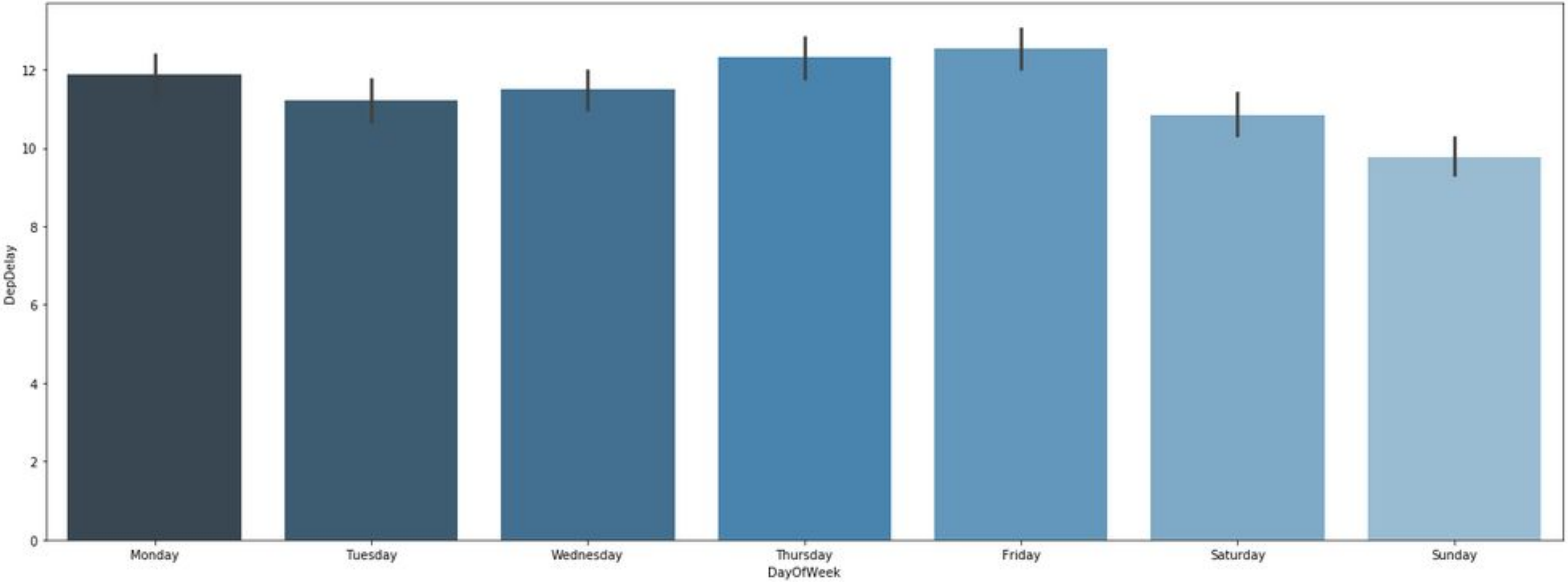
Flight and weather data merged by date

Trends

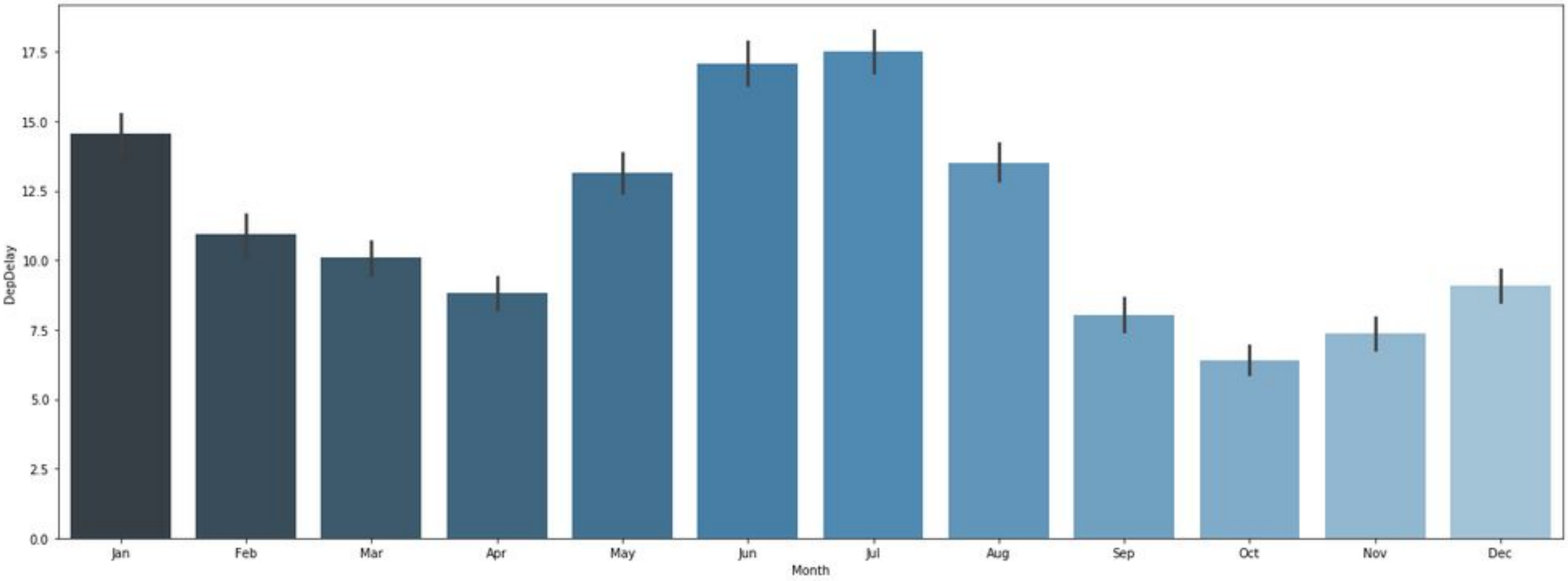
Distinct trends were evident when inspecting the data.

Creating data visualizations really brought these out.

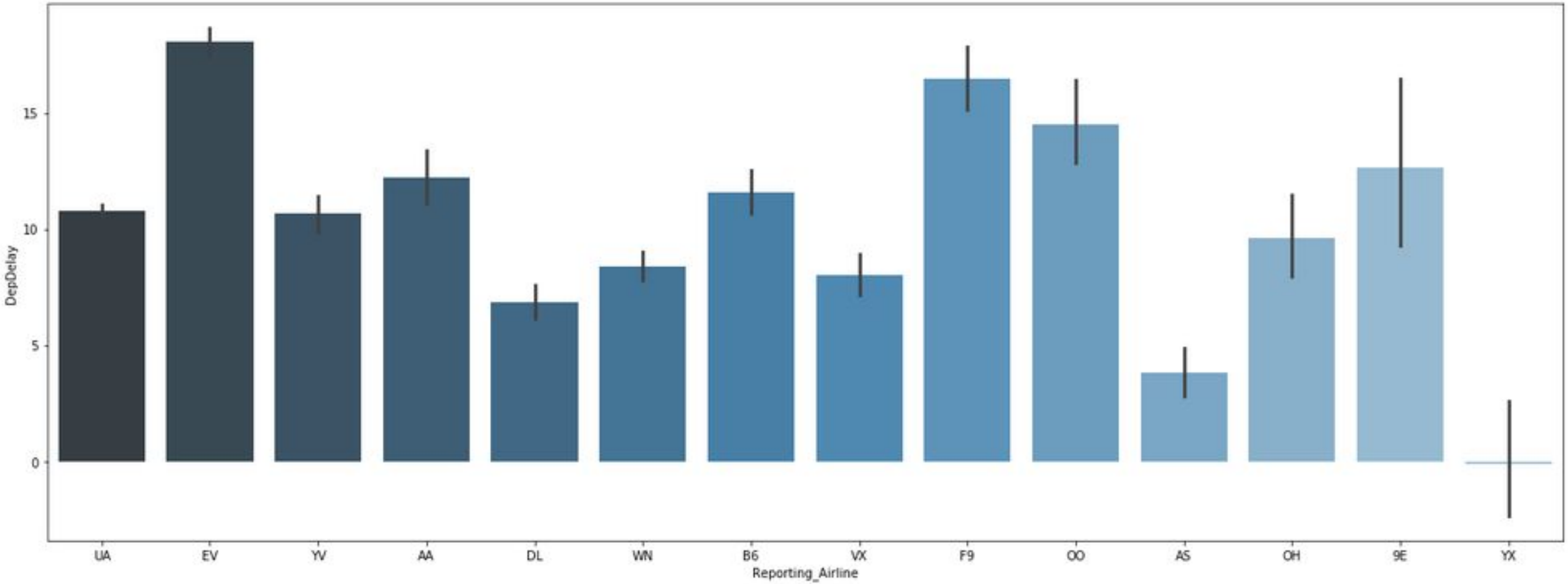
Departure Delay by Day of Week



Departure Delay by Month



Departure Delay by Airline

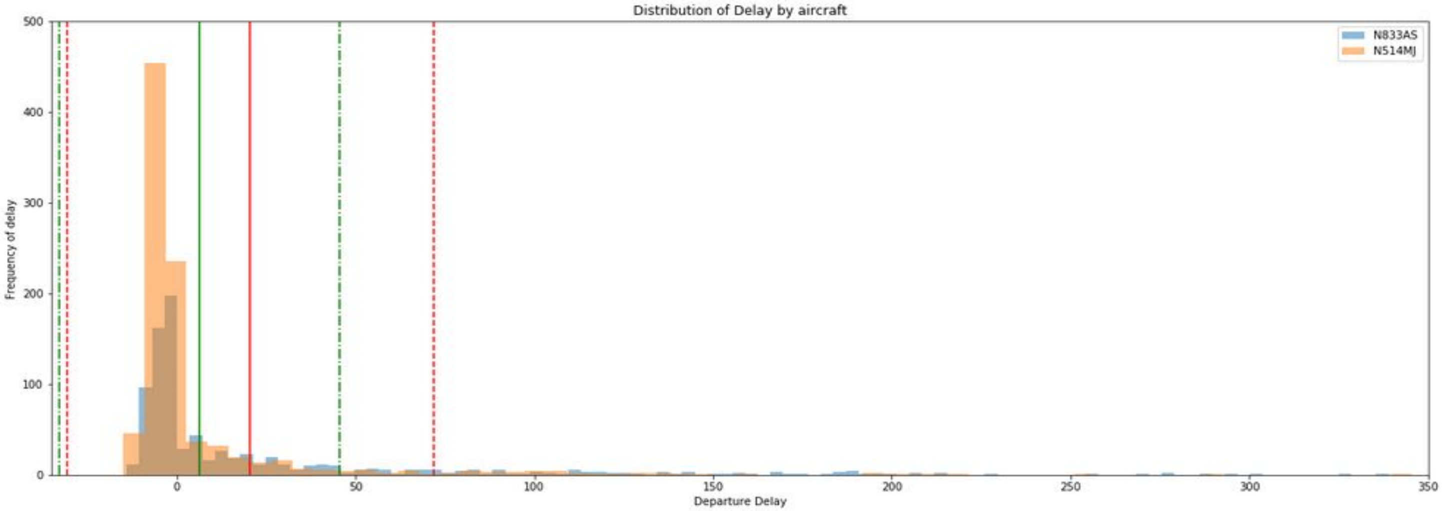


Tail number

Tracks individual aircraft

There were too many to keep in the Machine Learning Model, but they did show interesting trends

Distribution of Delay for 2 Aircraft



Of particular interest here is the overall shape of the data. The data is normalized around zero delay, but it has an incredibly long and thin tail.

Data Preparation for Machine Learning

A heatmap was used to help remove any data that we too highly correlated

Pandas “get dummies” was used to change categorical data to “One Hot” encoding

X data was scaled to remove undue influence of scale on the learning algorithms

Machine Learning Models

A number of models were used

R-Squared was used to interpret quality of fit

Machine Learning Results

Here are results:

1. 0.076 Random Forest

2. 0.064 Catboost

3. 0.033 Lasso

4. 0.033 Ridge

5. -0.004 K-Nearest Regression

6. -0.083 SVM

Next Steps

With the non-normalized distribution it was going to be difficult to get a good prediction of delays down to the minute.

Instead of using regression, try framing the problem statement as a classification problem with on-time (negative or 0 delay), slight delay (less than 15 minutes), or significant delay (greater than 30 minutes).