Capstone Project:
Predicting Departure Flight Delays

Julie Vovchenko

March 2020

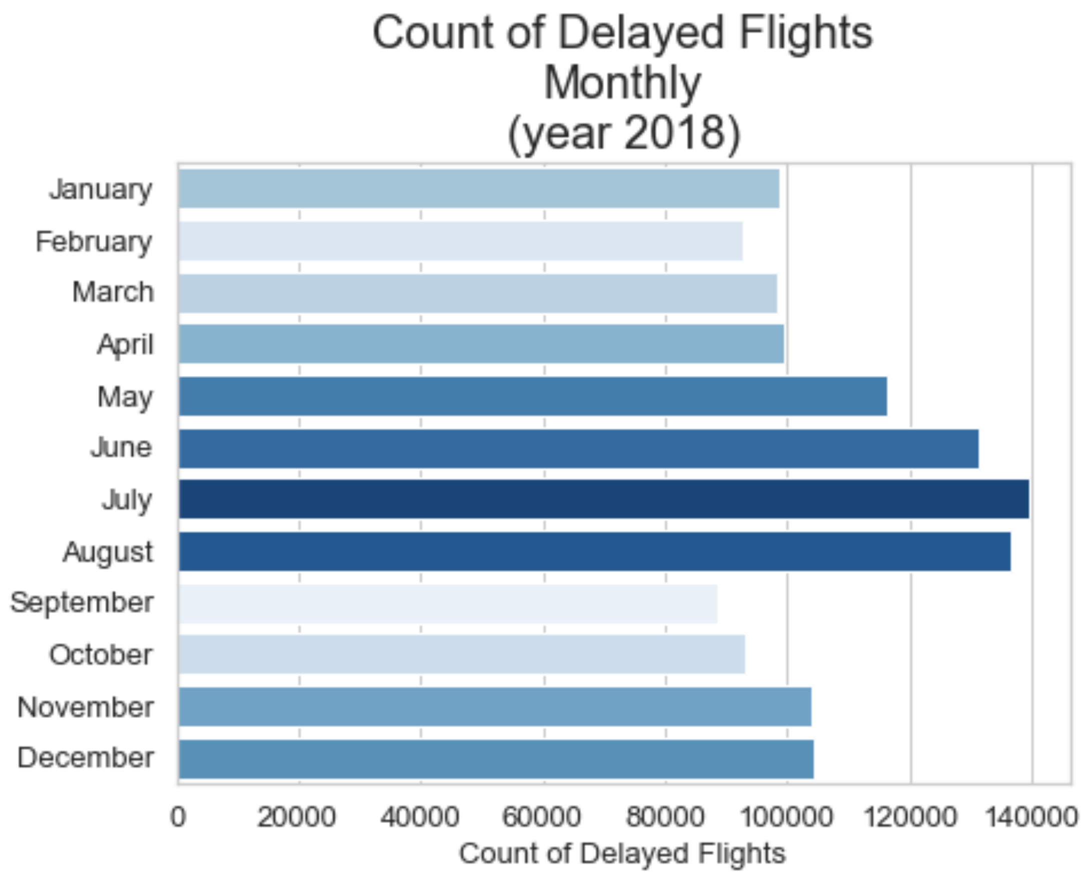
Problem Statement

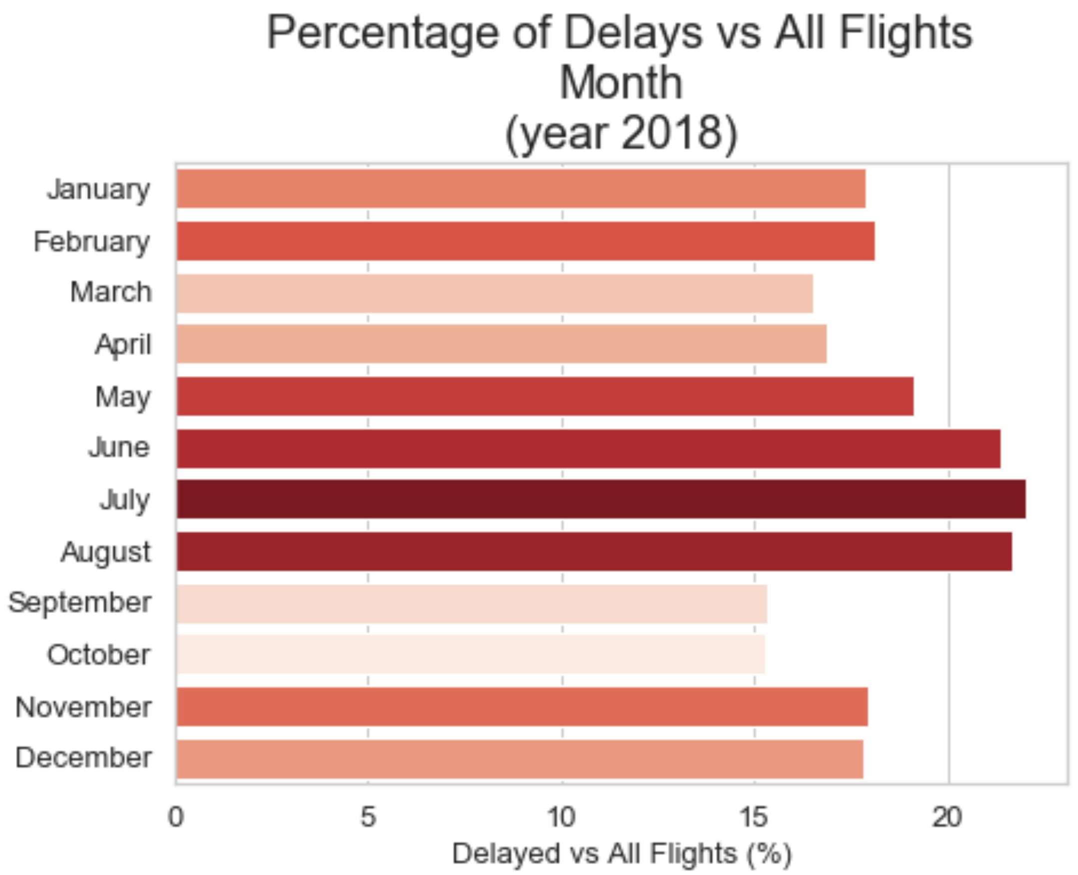
https://www.businesstraveller.com/business-travel/2017/06/28/summer-holiday-flights-face-delays-says-bbc-report/

Sometimes we catch ourselves in situations that we are concerned **if our flight will be delayed**, especially if there is a connecting flight. Many of us travel with our entire family or are traveling for a business trip and every second counts.

By utilizing a public dataset, provided by the **Bureau of Transportation Statistics**, on local flights in the United States from 2018, we plan to predict whether a flight will be delayed by 15 minutes or more.

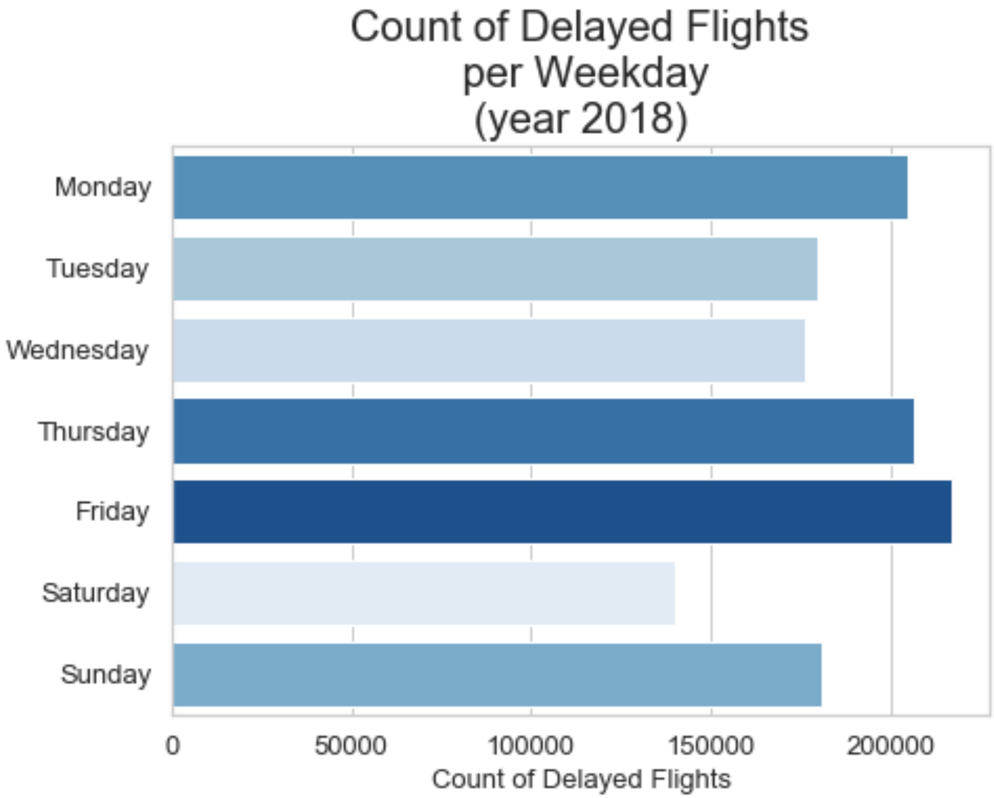
Exploratory Data Analysis - Month

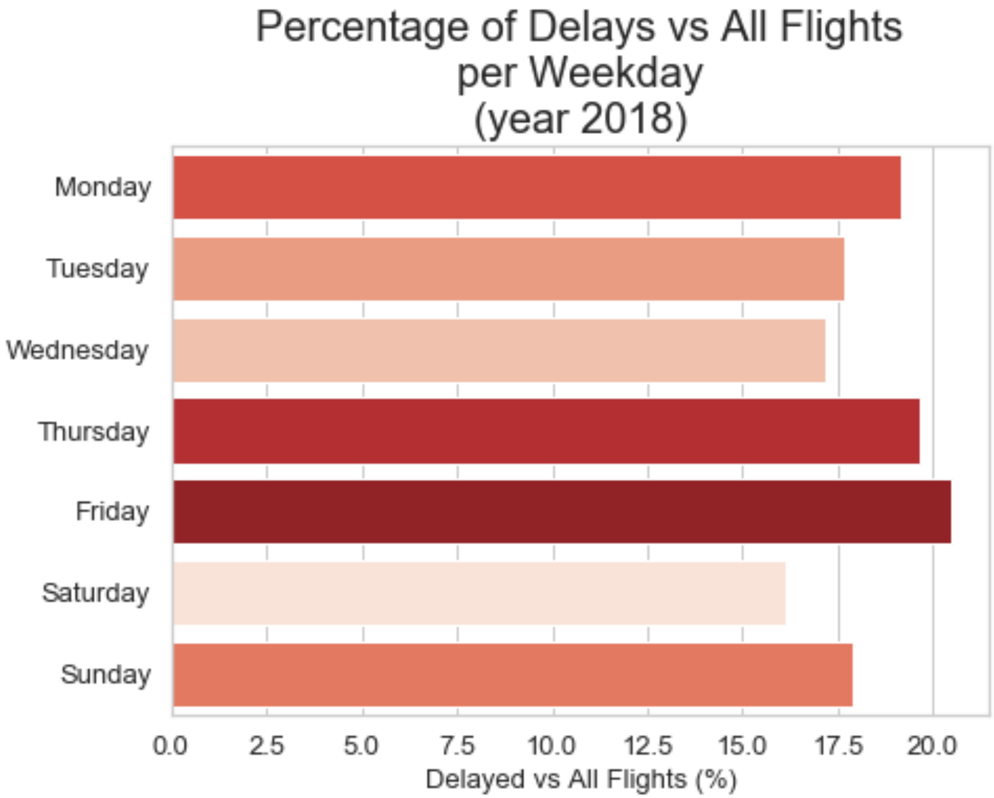




- February may appear to be the best month to travel in raw count of all delays
- Percentage of delayed flights vs all flights in that month indicates October as the best month to travel

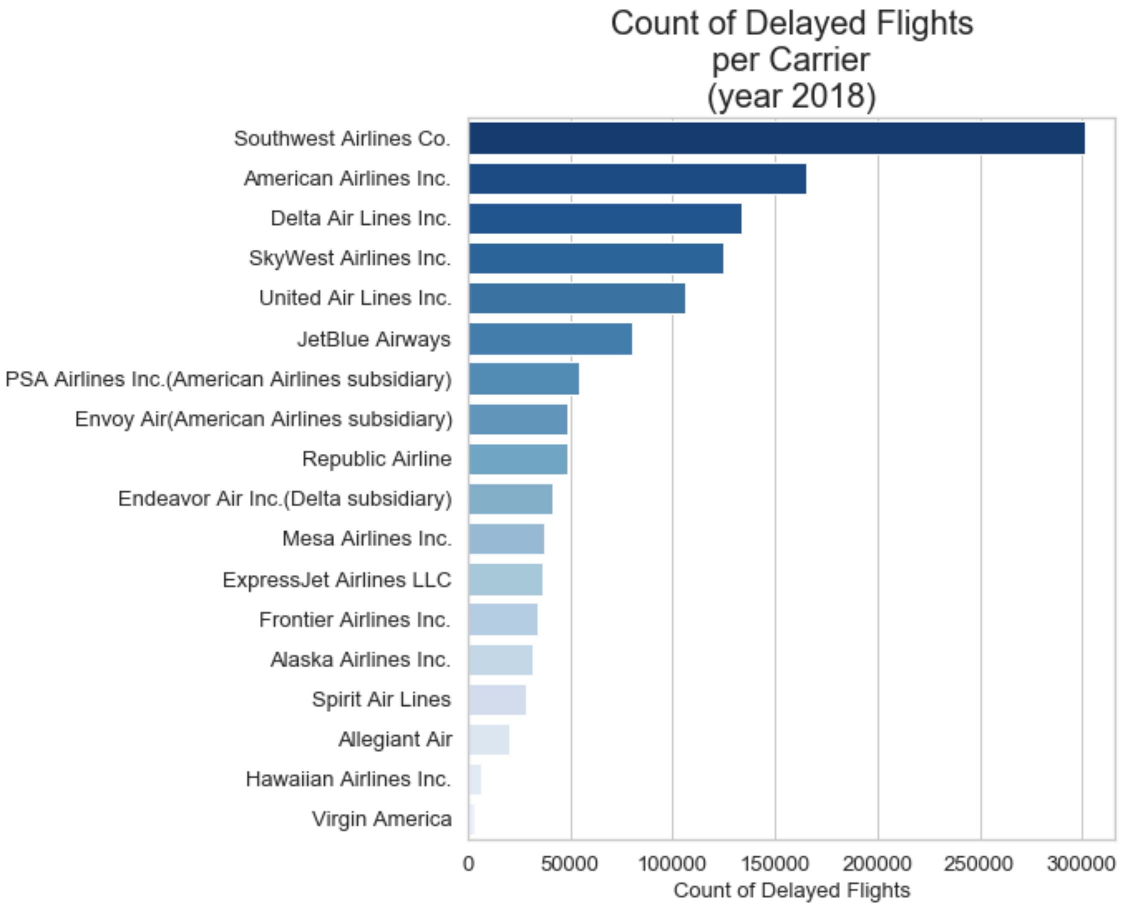
Exploratory Data Analysis - Weekday

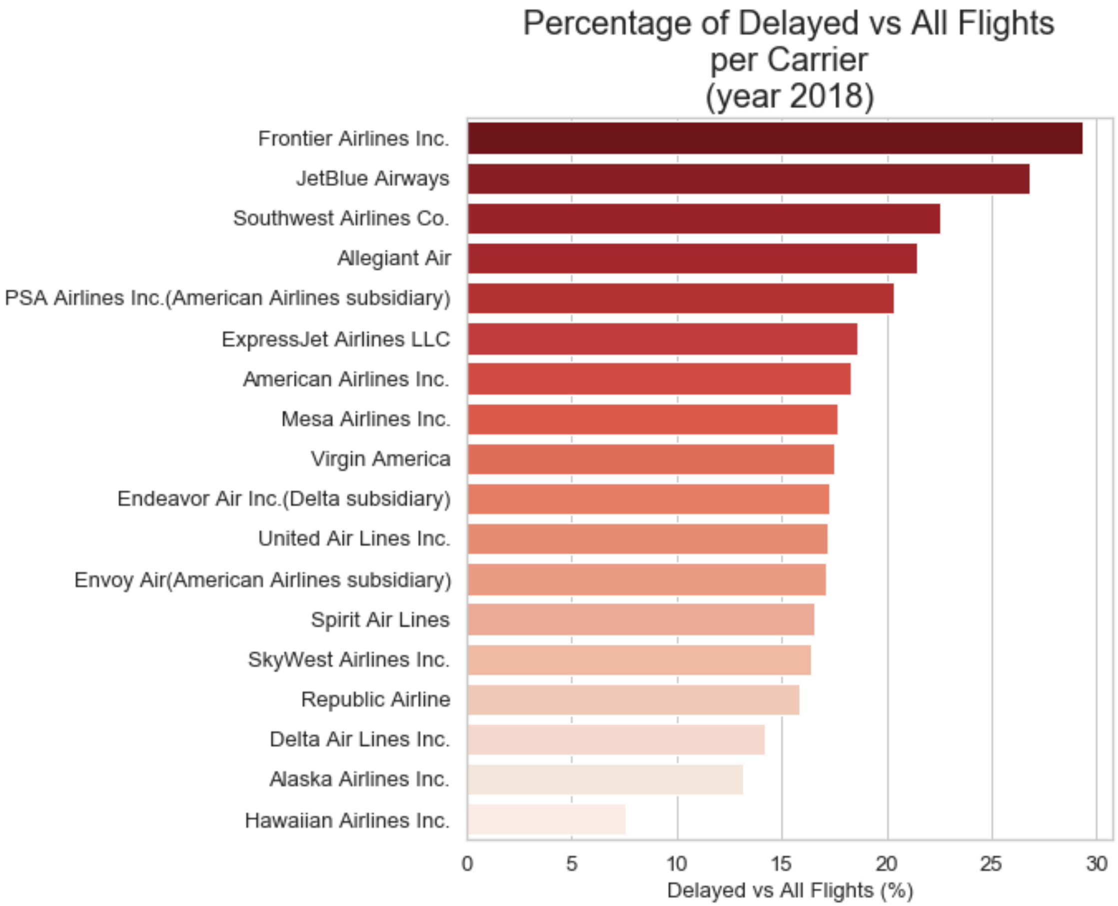




- Saturday is the best weekday to travel based on both raw delay count and percentage of all delays vs all flights on that weekday

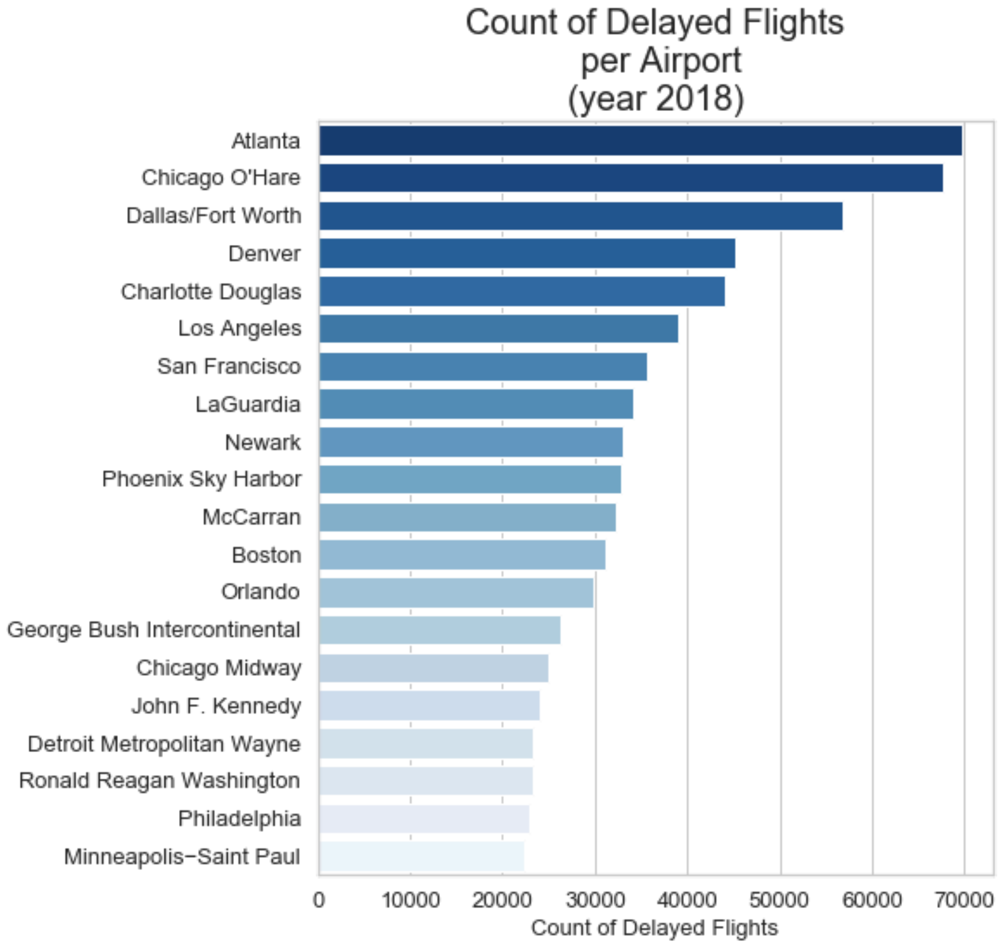
Exploratory Data Analysis - Carrier

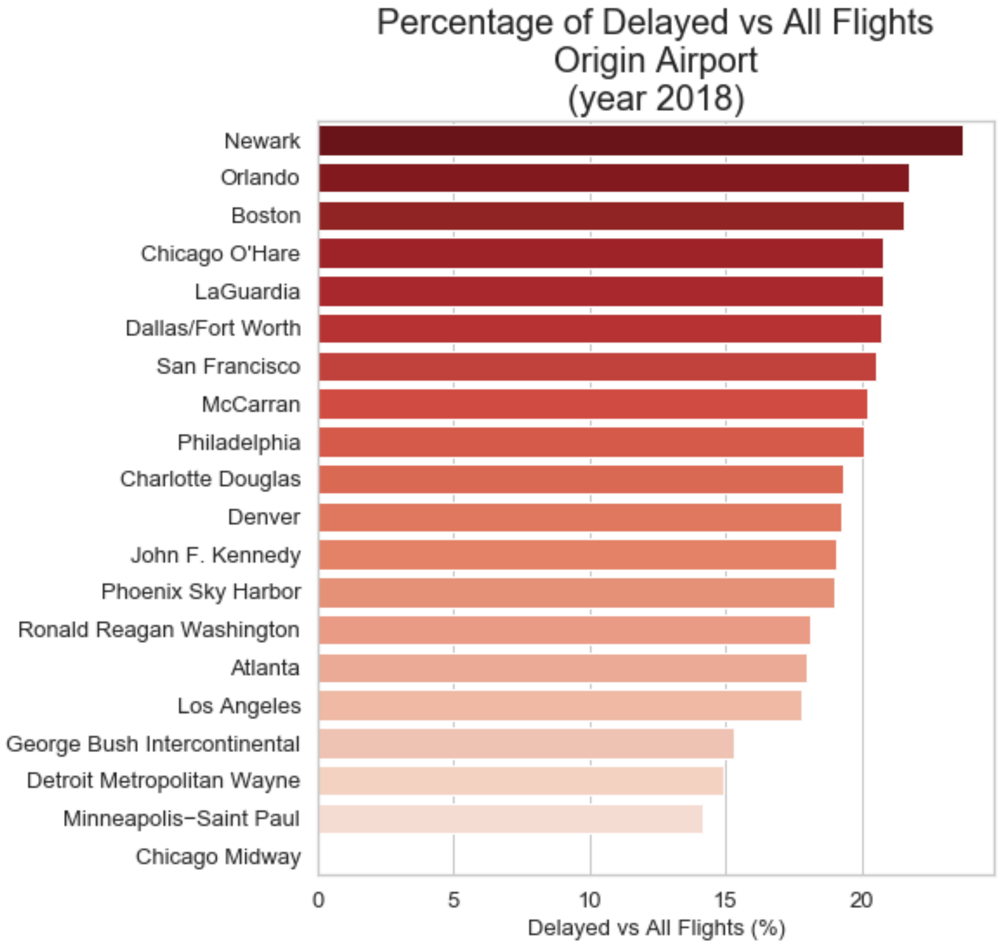




- Raw count shows top airlines with delays as Southwest, American Airlines, Delta
- Frontier Airlines' delayed flights are almost a third out of all flights they have

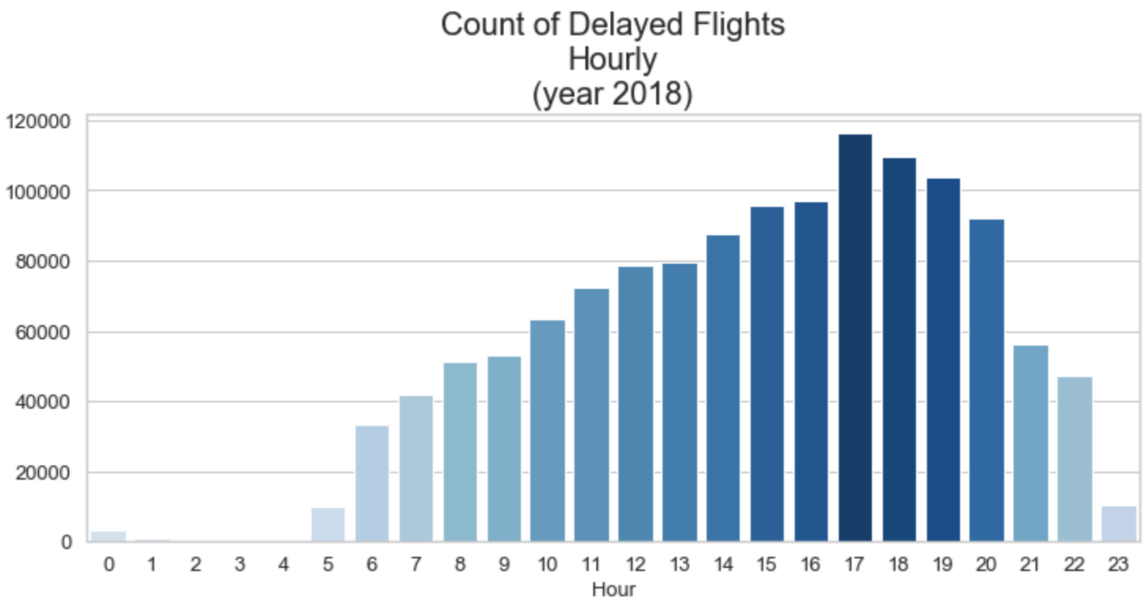
Exploratory Data Analysis - Origin Airport

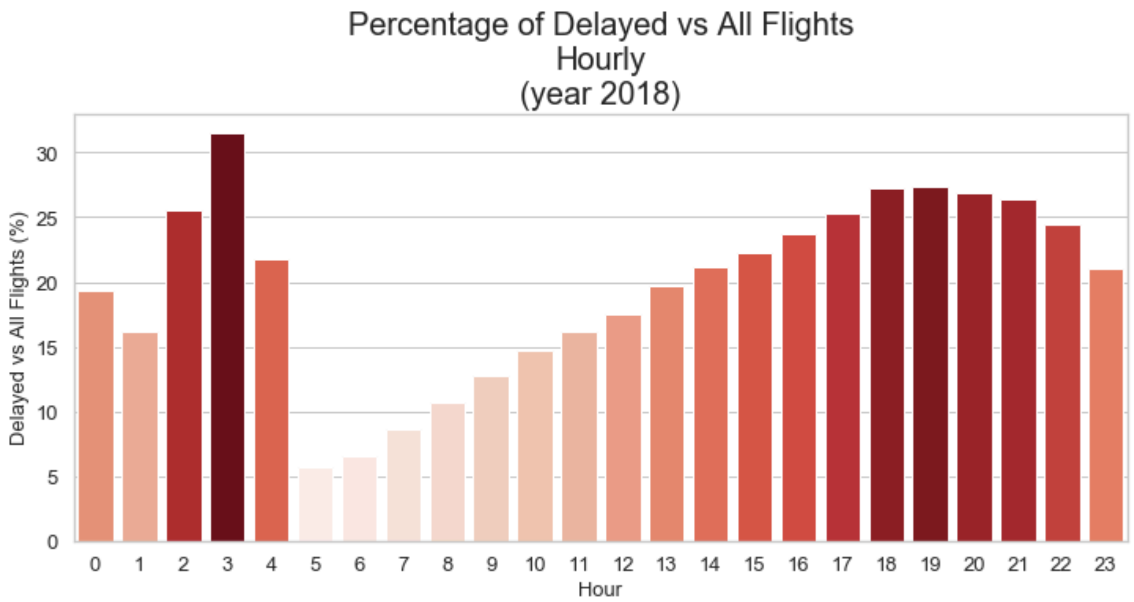




- Top airports in raw count of delays: Atlanta, Chicago, Dallas
- Out of all flights Newark has almost 25% delayed flights

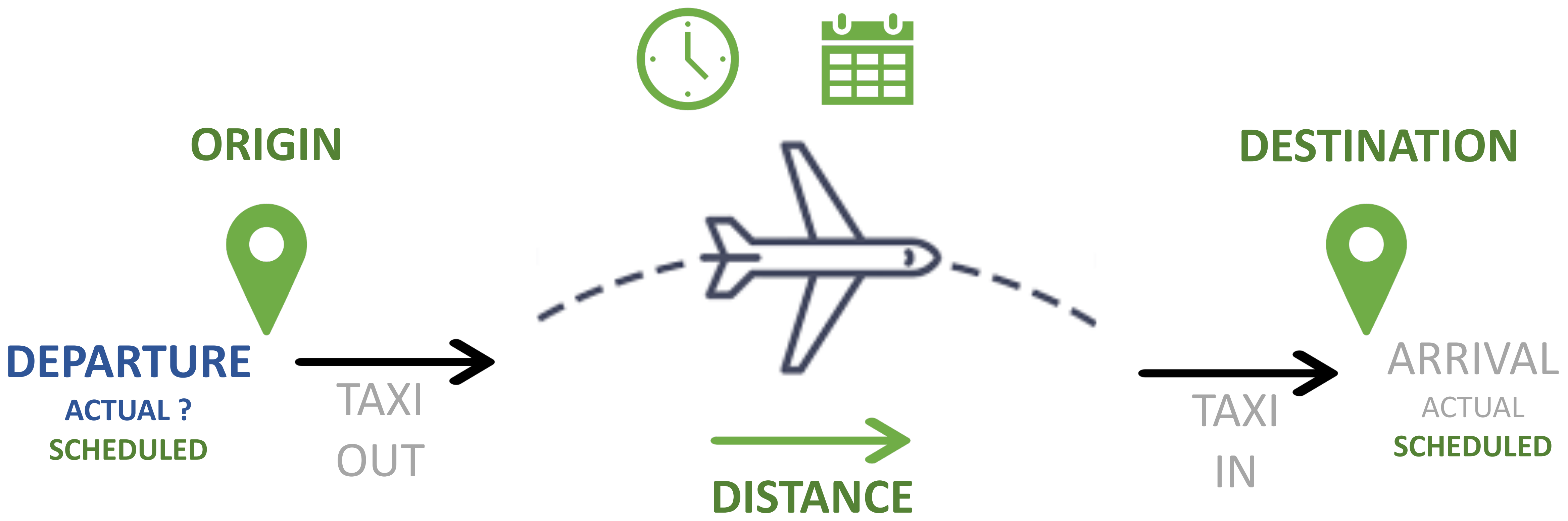
Exploratory Data Analysis – Departure Hour





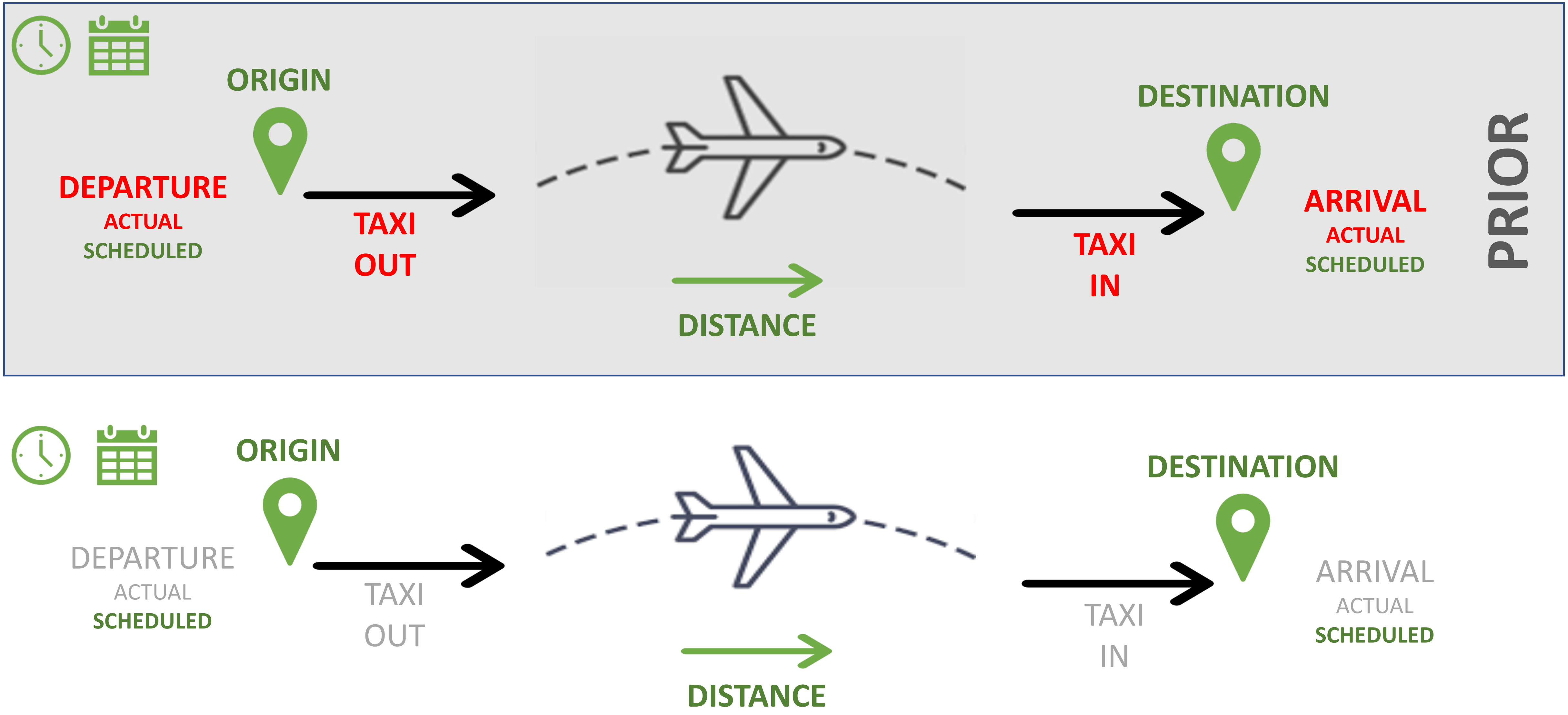
- Early morning flights have a very small sample size of all flights
- Important to have both raw count of delays and percentage, having only one could be misleading
- Top departure hour with delays are 6pm-9pm, after work

Feature Engineering: Current Flight Features



- Features that happened after departure are incorrect to use
- Using only features that are known in advance (scheduled)

Feature Engineering: Same Day Prior Flight Information



Feature Engineering: Same Day Prior Flights Total Count

Departure Time

12:30 am — PRIOR — 1st flight of the day

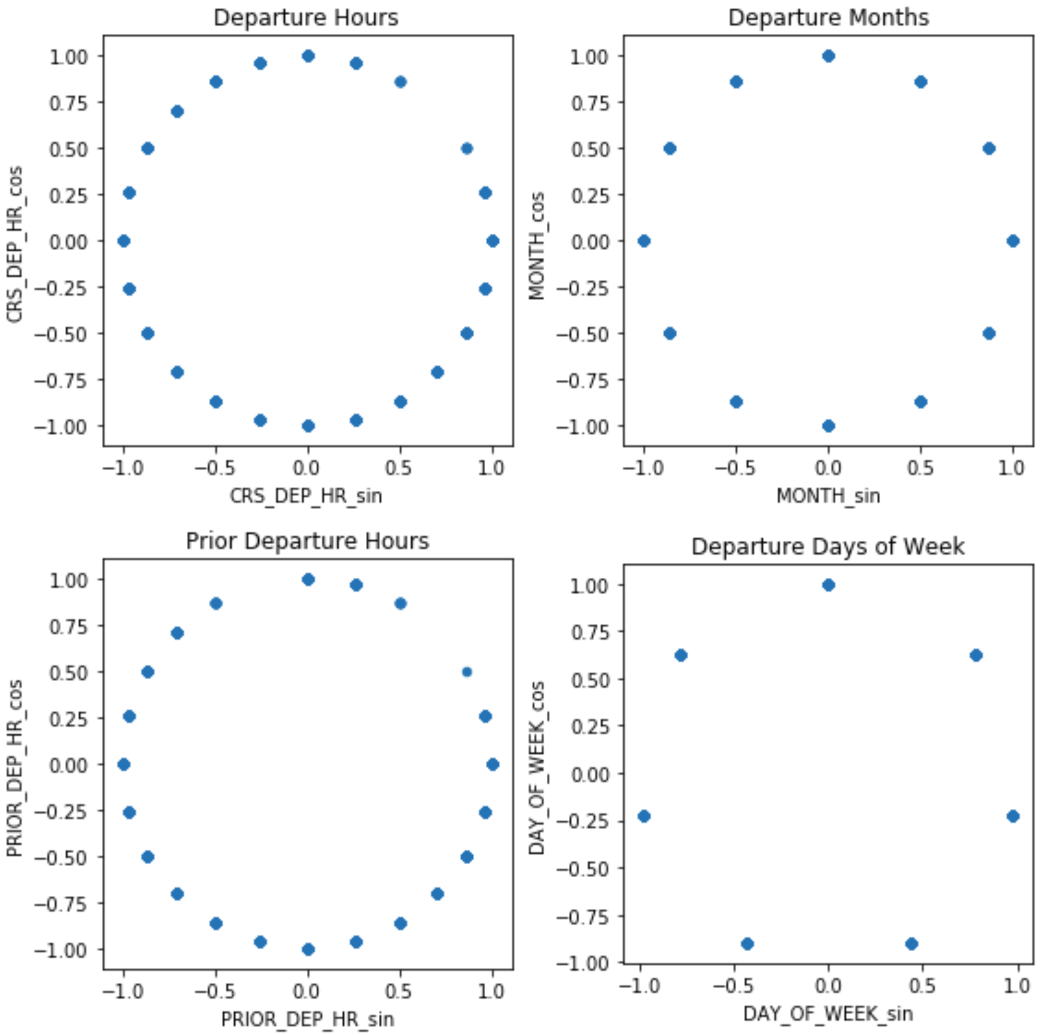
08:00 am — PRIOR — 2nd flight of the day

01:00 pm — PRIOR — 3rd flight of the day

TOTAL COUNT OF PRIOR FLIGHTS SINCE MIDNIGHT = 3

05:00 pm — Current Flight

Feature Engineering: Cyclical Features



<u>Cyclical Features:</u>

- Hours (0-23)
- Days of a Week (Mon – Fri)
- Months (Jan – Dec)
- Days of a Month (0 – 30)

<u>Each cyclical feature converted into 2:</u>

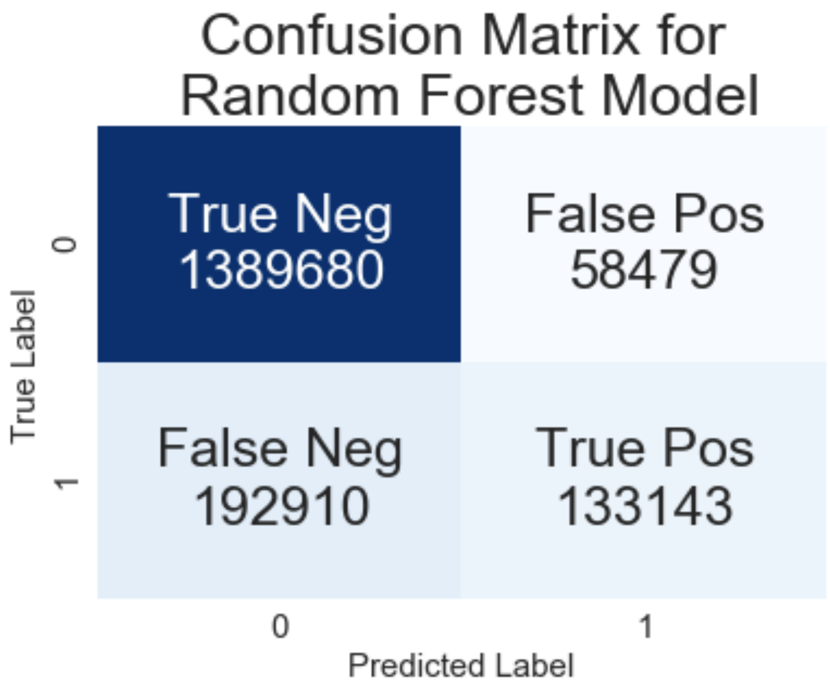
- sine
- cosine

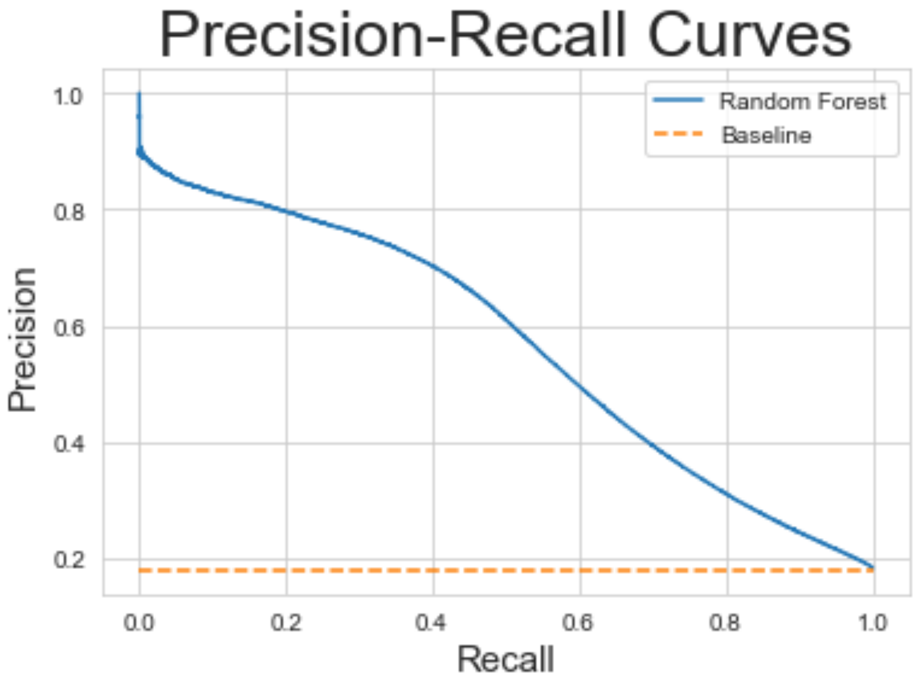
Modeling

Unbalanced Data	**18%** - departure delayed, **82%** - departure on-time
Scoring Method	Average Precision Score (it is the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight)

Models	Average Precision Train	Average Precision Test
Baseline Model	0.18	0.18
Logistic Regression	0.493	0.492
Random Forest	0.646	0.561
AdaBoost	0.497	0.495
Feed-Forward Neural Network	0.543	0.544

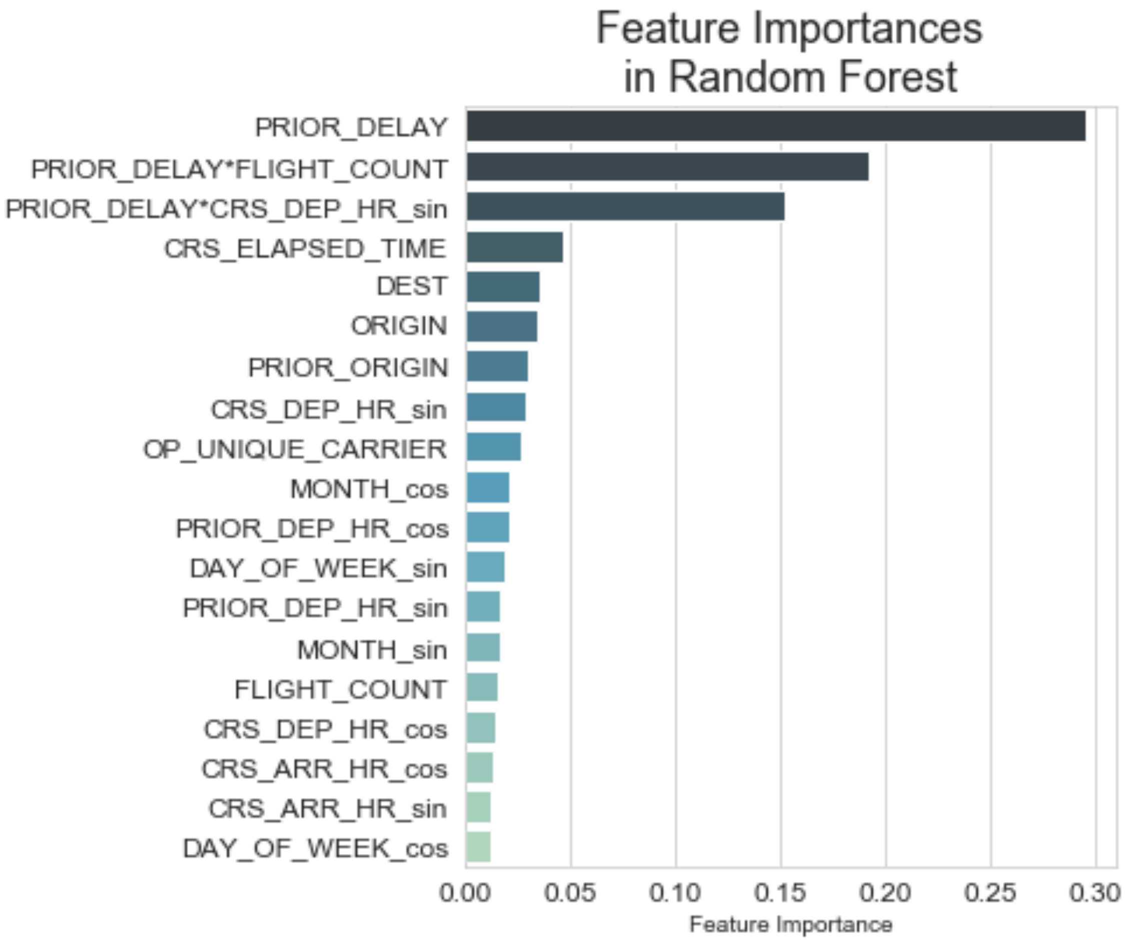
Confusion Matrix and PR-curve





- Our Random Forest Model has higher average precision score (0.568) vs the baseline model
- Precision-Recall curve shows the trade-off between precision and recall for different thresholds.
- Baseline represents of proportion of the positive class – 18%.

Conclusion and Recommendation



- Random Forest, average precision score 0.57, meaning that out of all positive predictions our model identified 57% were actually positive.
- Feature Importance vs coefficients
- Feature Importance explain the predictive power of the features.

Next Steps

1. Identify the significance of each feature by using stats models.

2. Collect daily weather data, like wind speed and precipitation rate, for each origin and destination location for each flight. Here are some resources: noaa.gov, weather.gov

3. Get data about each plane used for the flight (flightradar24.com). Data helps to identifying flight delays that occur due plane malfunction: manufacture quality or age of the plane.