Akshar Shrivats

IB Computer Science HL 4B

Potter

22 March 2021

KNN Project - An Analysis in Flight Delays

Abstract:

This report analyzes data published by the FAA which records 5.8 million domestic flights from airports across the USA. Inspired by my pursuit of a private pilot's license, I wondered what factors caused commercial flights to be delayed the most. To answer this question, I analyzed and visualized data to find trends between factors such as Airlines, Arrival Times, Airports, and Delay Classifications. This allowed me to create a KNN model with 92% accuracy that could predict if a flight was going to be delayed for more than 15 minutes or cancelled all together. After analysis, it was determined that the single most important factor in predicting if a flight is to be delayed is checking if it had a delay on arrival.

Introduction:

Travelers around the world are faced with disappointment, anger, and confusion daily as their flights are delayed or cancelled due to adverse weather, maintenance issues, or airport logistics problems. My dataset tracks date, day of week, airline, origin and destination, departure time, departure delay, time until wheels up. Cancellation status, and the form of delay. I know that some of these factors will be more heavily weighted than others - specifically, the day of week (the busiest flight days are Mondays and Thursdays) and type of delay (Weather, Late Aircraft, Airline, Security, or Air System). My KNN model aims to predict if a flight will get delayed based on factors including departure airport, airline, weather conditions, ATC status, and more.

Recently, I have begun working toward my Private Pilot's License and learned a lot about airport procedure and aviation protocol. Even for my current training flights, I have to generate full scale weather reports and often have to delay or cancel my flights due to adverse AIRMETS (AIRman's METeorological Information) or NOTAMs (Notice To Airmen). This includes information regarding everything from wind speeds and direction to thunderstorm warnings, unlit towers, and missile testing. Commercial flights have to follow the same briefing structure as any pilot, so it will be interesting to see how predictable a delay is and test it on my future training flights to determine if they will actually get off the ground. I plan on reviewing data published by the FAA regarding commercial flights in 2015. First, doing an exploratory data analysis to see correlations between factors, then generating a KNN model to test new data.

Methodology and Technical Approach:

This project will be utilizing the K-Nearest-Neighbors Algorithm and the Euclidean distance formula to create a classification model. The first step in this process, however, will be normalizing the data such that it may fall within some normal boundary. After normalization, we will increase the weight of some specific factors by multiplying their Euclidian distance by a scalar. This will cause any deviation in their value to have an increased impact on the final classification of a specific data point. The equations I will use to complete my model are listed below:

$$Normalization\ Formula = \frac{(X - X_{minimum})}{(X_{maximum} - X_{minimum})}$$

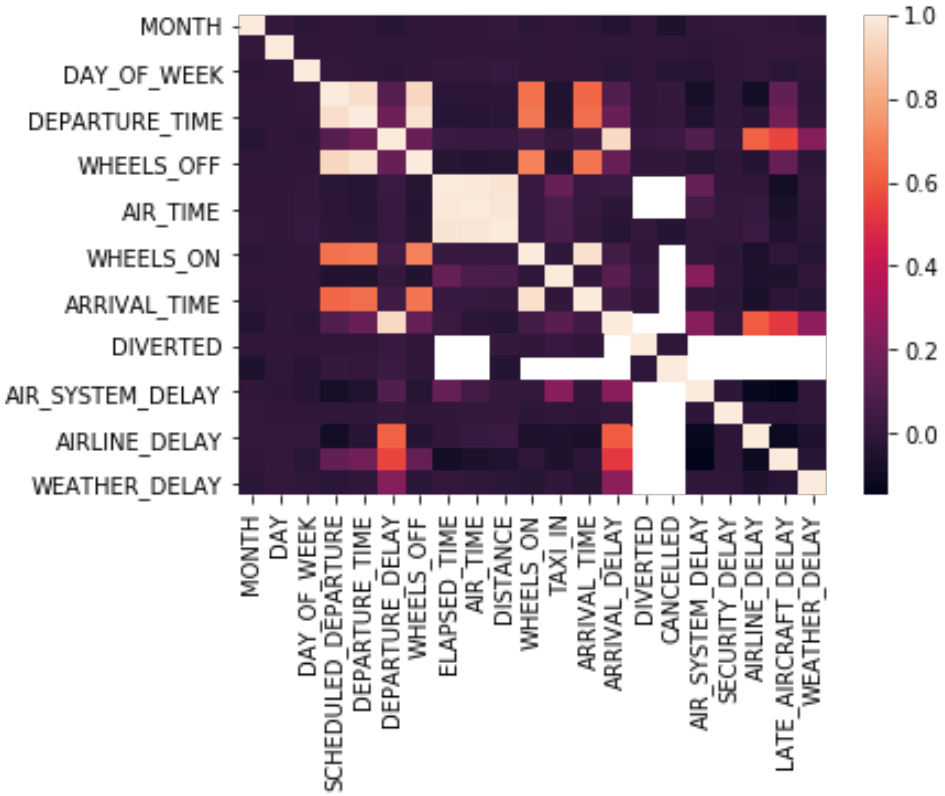
Where X represents the training data given to the algorithm (flights.csv in this project)

$$Euclidian\ Distance = \sqrt{w_1(x_1 - x_1')^2 + \cdots + w_n(x_n - x_n')^2}$$

Where x represents a specific row of data from the training set and w represents the weight placed on the column's distance. A standard value for w is 1.

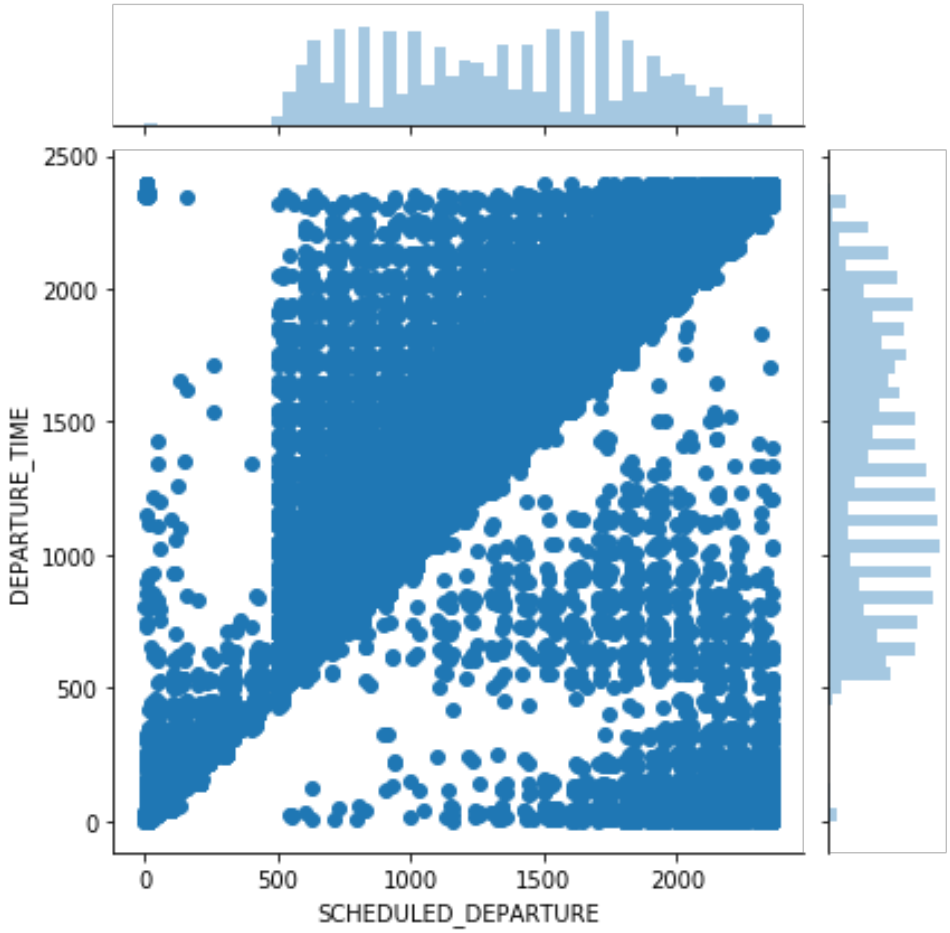
Experimental Process:

I began my experimental process with some exploratory data analysis to find any obvious trends in the data. This began with constructing a correlation matrix for the important factors in my training data. This meant that I needed to get rid of factors that played no obvious role - such as Tail Number, Year, or Aircraft ID. The resulting correlation matrix is:



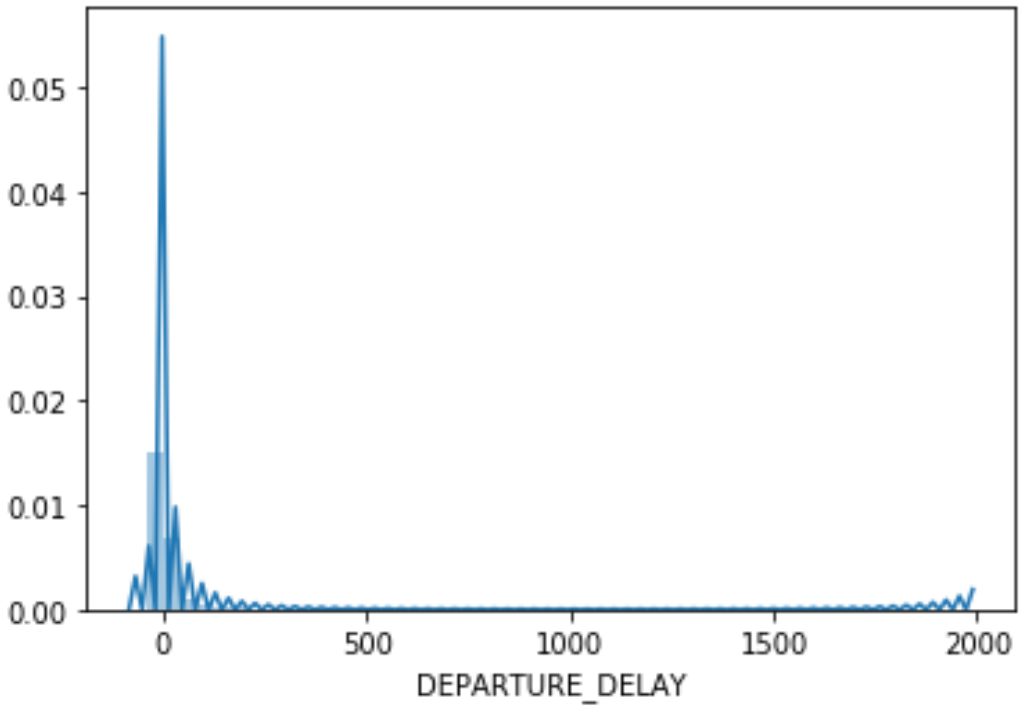
This showed me that some of my initial predictions about the trends in the data were incorrect. We can see from the “MONTH”, “DAY”, and “DAY_OF_WEEK” sections of the matrix that they have little to no correlation whatsoever with any other factors present in the data. We can also see that a majority of delays are registered as “AIRLINE_DELAY”, meaning that the airline you choose to fly on may have a large role in determining if you get to your destination on time. We can also see a strong correlation between “ARRIVAL_DELAY” and “DEPARTURE_DELAY”. These values were measured as positive or negative deviations from

the scheduled arrival or departure time. Next, I created a joint plot that plotted the connection between “SCHEDULED_DEPARTURE” and “DEPARTURE_TIME”. Ideally, there would be a 1:1 relationship between these two variables signifying that all flights will leave on time. The joint plot also creates a histogram so we can get an idea of when the majority of departures are scheduled. This chart is plotted below:

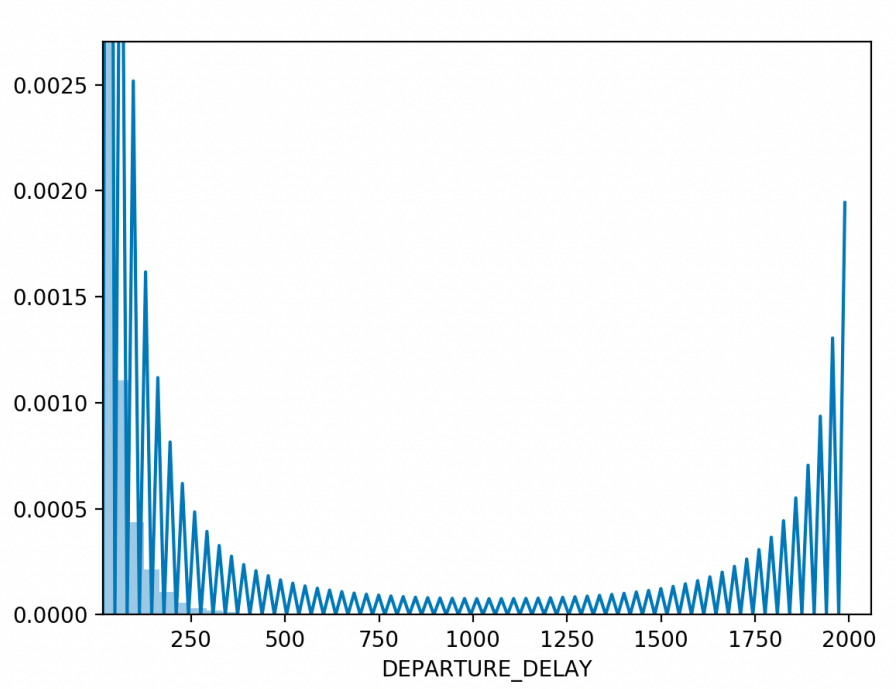


If there is one thing that this plot is not, it’s a straight line. Instead, it paints a clear picture of the rampant delays and ridiculous wait times that plague the commercial airline industry. One thing to note when analyzing the data is that it is measured in “Zulu time” or UTC and in 24-hour format. Hence, if a flight departing near midnight is delayed, it will revert back to a “low” time number. That can explain the cluster of data points located on the bottom left of the graph. We

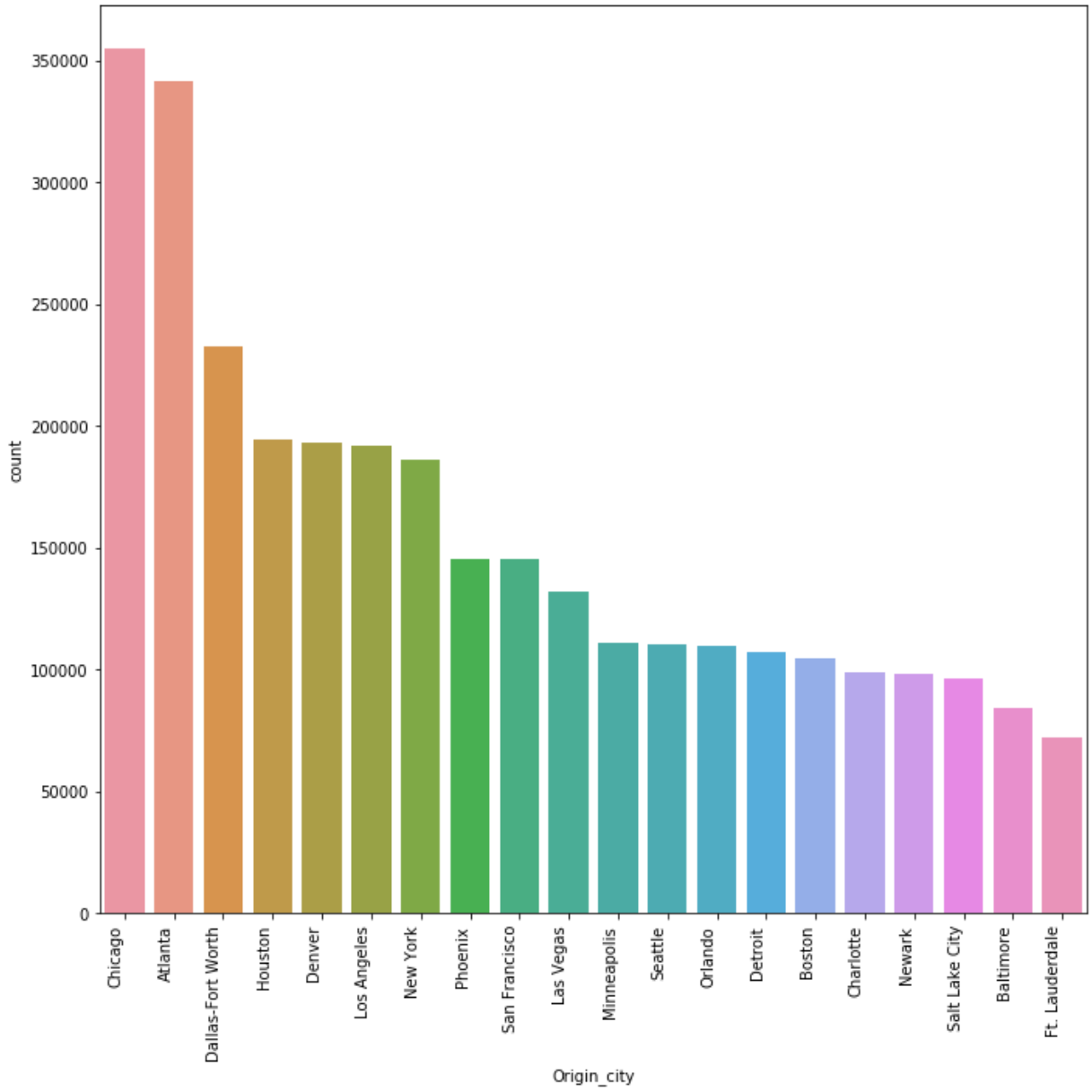
can also see that the distribution of delays throughout a given day remains fairly consistent, only spiking in the morning and evening hours. I also used the distplot functionality within seaborn to plot the distribution of the “DEPARTURE_DELAY” variable.



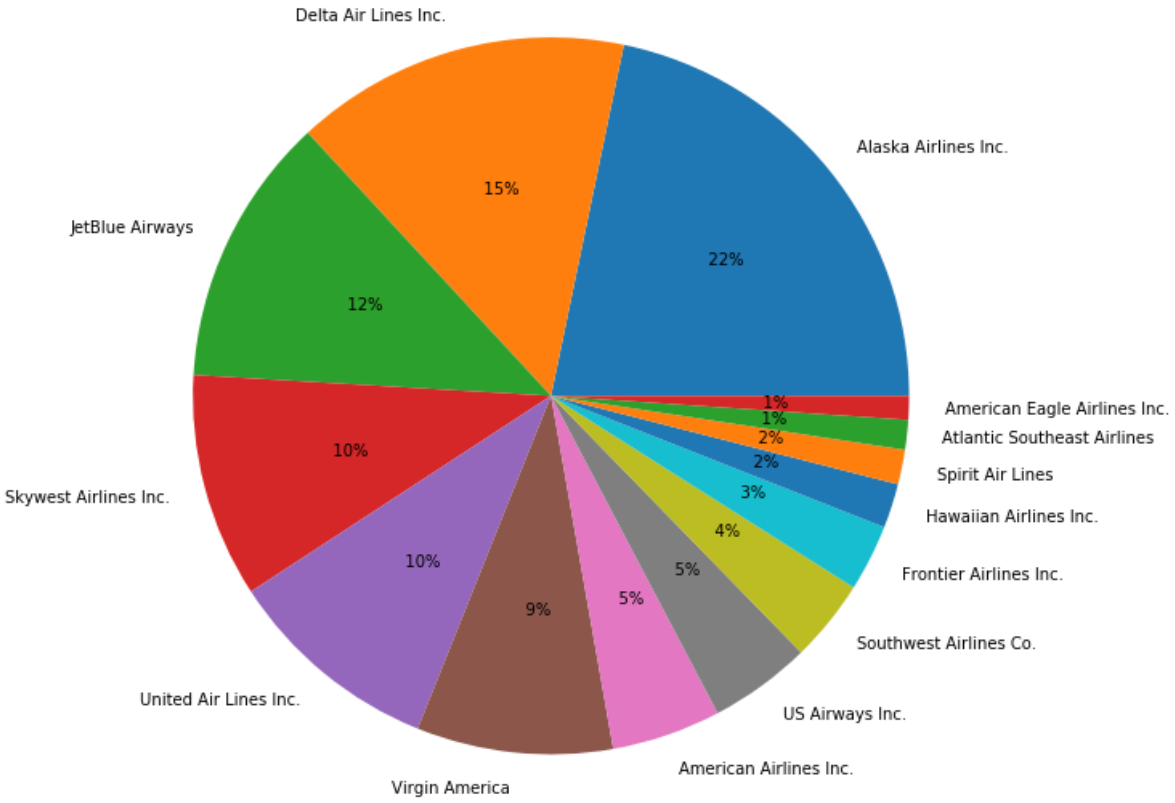
This graph showed that the vast majority of flights were actually on time - yet if we narrow our delay window to greater than 15, we can see that flights are susceptible to extremely long delays.



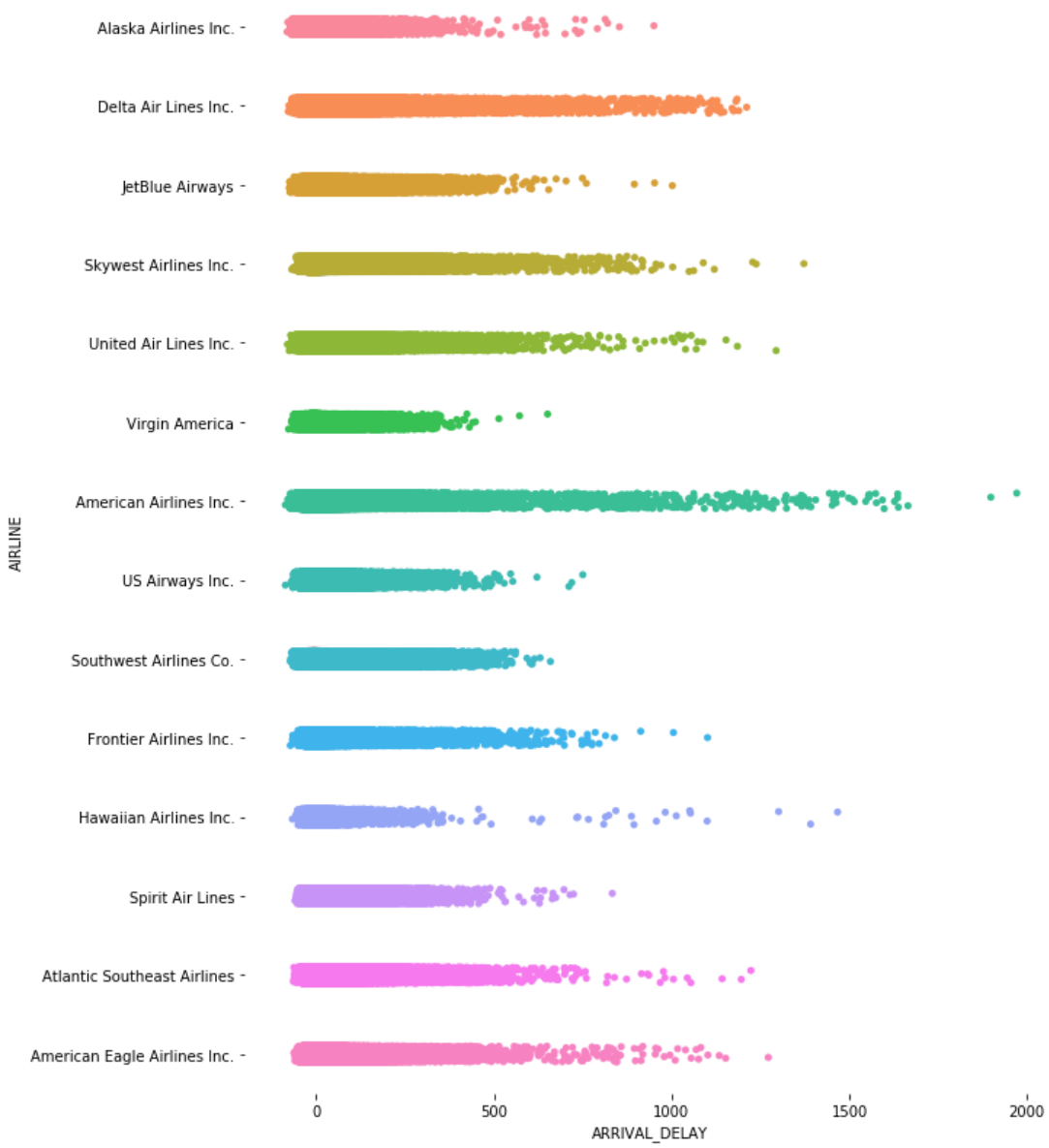
As the final part of my data analysis I wanted to gain a greater understanding of the Airport and Airline distributions present within my data. This began with a histogram showing the distributions of flights from the top 20 airports in the data:



The distribution of airlines present:



And an analysis of the arrival delays for all airlines in the data:



This data showed that the airline with the longest delays was American Airlines. Which came as no surprise to me, just based on my personal experience.

Finally, after gaining a thorough understanding of my data and the context surrounding its individual factors, I felt prepared to create my KNN model. In my training data, I used the factors “SCHEDULED_DEPARTURE”, “DEPARTURE_DELAY”, “AIR_TIME”,

“SCHEDULED_ARRIVAL”, “ARRIVAL_DELAY”, “AIR_SYSTEM_DELAY”, “SECURITY_DELAY”, “AIRLINE_DELAY”, “LATE_AIRCRAFT_DELAY”, and “WEATHER_DELAY”. My model was designed to classify a flight as “CANCELLED” or “DELAYED” if the departure time exceeded 15 minutes from the schedule. Unfortunately, I was unable to use categorical factors such as Airport or Airline simply due to my distance function. Using Euclidian distance, I am inherently bound to only using factors that contain numerical values as it is impossible to measure the distance to a category using that formula. My first KNN model produced only gave me an 88.6% effectiveness as all the factors were normed between 0 and 1 with no weights. However, after looking at my correlation plot, I could see what factors corresponded most with the end departure time. To account for this, I boosted the weight of the “ARRIVAL_DELAY”, “AIR_SYSTEM_DELAY”, and “AIRLINE_DELAY”. After recreating my model with these weights, I obtained a 92% accuracy.

Conclusion:

Many factors play into determining if a flight will become delayed; from security to air traffic control you never know what time the wheels will truly be off the ground. Yet the problem of airline delays is one that negatively affects not just the end customer, but airlines as well. The more time a plane spends on the ground, the less time it is spending making them money. This report highlighted a few reasons that a departure delay might occur. Determining effectively that the most important factors in predicting airline delays are the airline that the plane is registered to, air system delays resulting from high airport traffic or emergency situations, and most importantly an arrival delay. The Airline a passenger travels on will determine many aspects of the flight experience. Different access to mechanical repairs, food delivery, staff availability, or airport timings based on Airline could mean entirely different take-off times, even for an

identical flight. Air system delays were most prevalent for flights taking off from America's busiest airports such as Atlanta or Los Angeles. These delays are often due to long taxi times and queues for take-offs, prolonged holds at the gate for maintenance and tug availability, and emergency landings or aborted takeoffs from the airport. Finally, the most impactful factor and most heavily weighted factor in my KNN model is the arrival delay. A flight that arrives late automatically has less time to clean, refuel, and restaff before its next flight. These will automatically put the Airline team under pressure and most often result in the next flight being delayed as well. Moving forward, the fastest way to tell if your flight has been delayed is to check if the plane has been running on schedule for the entire day.