Flight Delays

Ben Kean
2/17/2019

problem

Most of us have been there before. We are on the way to the airport for a vacation, business trip or a long weekend out of town and *it* happens. A text appears on your phone. It is from a strangely formatted number and obviously written by a bot. Upon realizing this you know all too well what has happened: **Your flight has just been delayed.** This is NOT the way you wanted to start your trip.
Is it possible that this could have been avoided? I believe that there are ways to lessen your chances of getting a delayed flight. In this study, I looked at ways in which I could domestic flight data to predict which flights should be avoided. I believe that important elements to consider when choosing a flight are the departure city, time of day, the size of the airport (in terms of flights) and time of year.

client

These days, flight delays seem to be unavoidable. Flight delays cost business & leisure travelers millions of dollars each year in added travel costs, lost revenue and missed opportunities. **If there was a way for these travelers to predict which flights to avoid, they could save themselves time, money and stress.** They could select higher performing flights and better connection cities. I believe that anyone who travels via airline, can could benefit from this study. However, I am focusing on the business traveler as the ‘client’.

data

The data used for this project will be from the Kaggle website. It is sourced from the Department of Transportation’s Bureau of Labor Statistics and contains flight data for the year 2015.

Link: https://www.kaggle.com/usdot/flight-delays#flights.csv

approach

For this study, I focused on which time of day the delay occurred, and at which airport/region the delay occurred. I thought it would be necessary to consider these two variables together due to environmental issues that affect specific airports (e.g. afternoon thunderstorms in Miami). Time of year was also considered to account for weather that only occurs during certain times of the year. In addition to the challenges that weather can pose, I wanted to see if avoiding certain flight times can improve one's chance of avoiding a flight delay, despite the time of year.

Data Wrangling

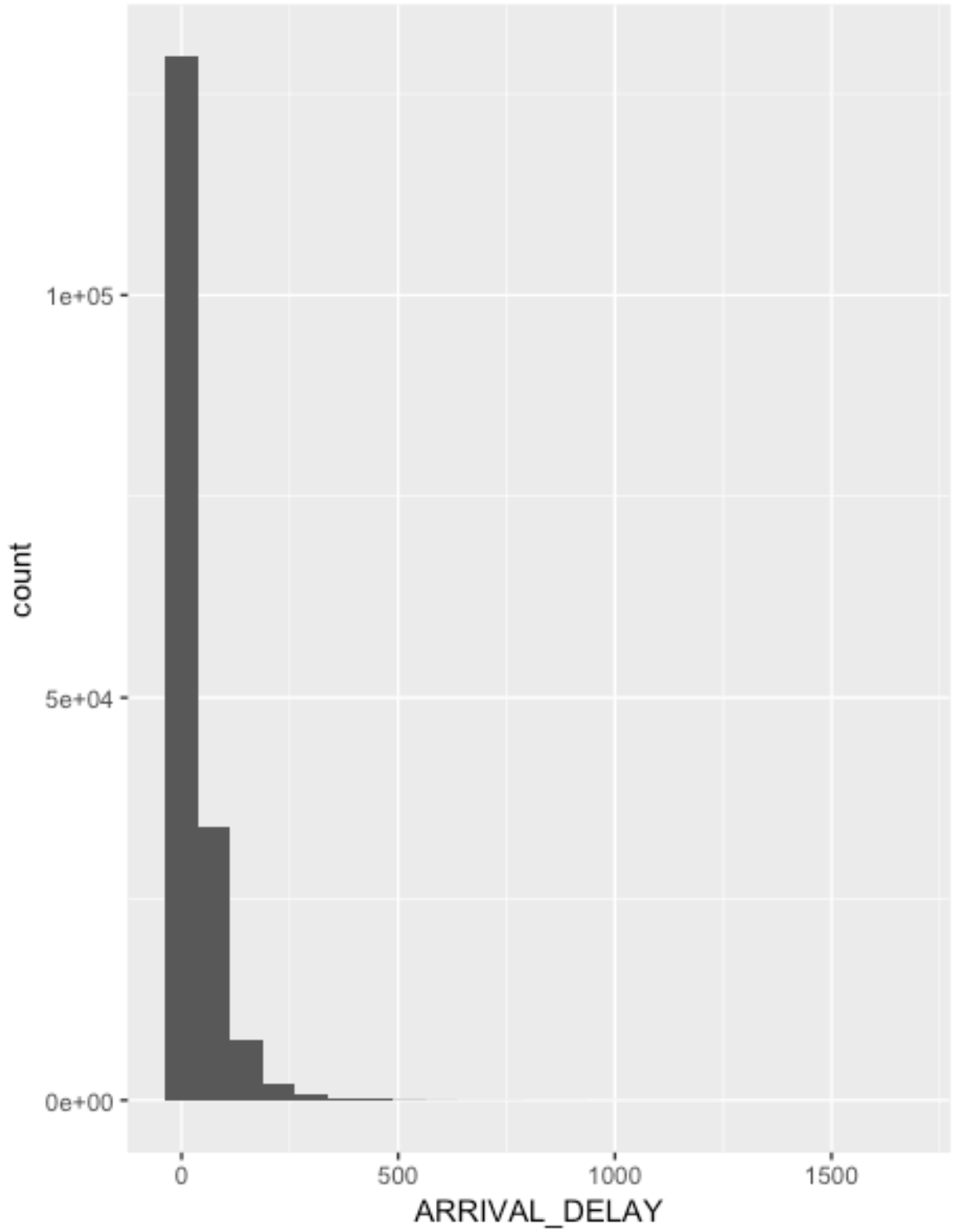
All of the data was stored in CSV files. There were three files in total: the flight data, the airport data and the airline data. The airport data can be joined to the flight data via a three character airport code. The airline data can be joined to the flight data via a two byte character code.
General cleaning that I performed on the data:
-Removed rows from the flight data that had null values for the departure delay
-Removed rows from the flight data that had null values for the arrival delay
-Converted the 4 character flight time filed to an actual time
-Created a new field called flight hour by converting the first two bytes fo the flight time
-Created boolean fields for the departure delay fields where 0 was no delay and 1 was a delay
-Applied a semi join to the flight data and the airport data to eliminate invalid airports -Joined the flight data to the airport data
-Created a time of day field (morning, afternoon, evening, overnight) based on the flight hour -Created a season field based on meteorological season

Application of Statistics

Summary Statistics

Below are summary statistics of the data when a delay occurred. If a delay did not occur on a flight, then it was not included in the summary below.

Below the summary statistics is a histogram showing the distribution of the delay times. Clearly, I am dealing with a left-skewed data distribution.
Min. 1st Qu. Median Mean 3rd Qu. Max. 1.00 6.00 15.00 33.12 39.00 1636.00



Probability of having a delay

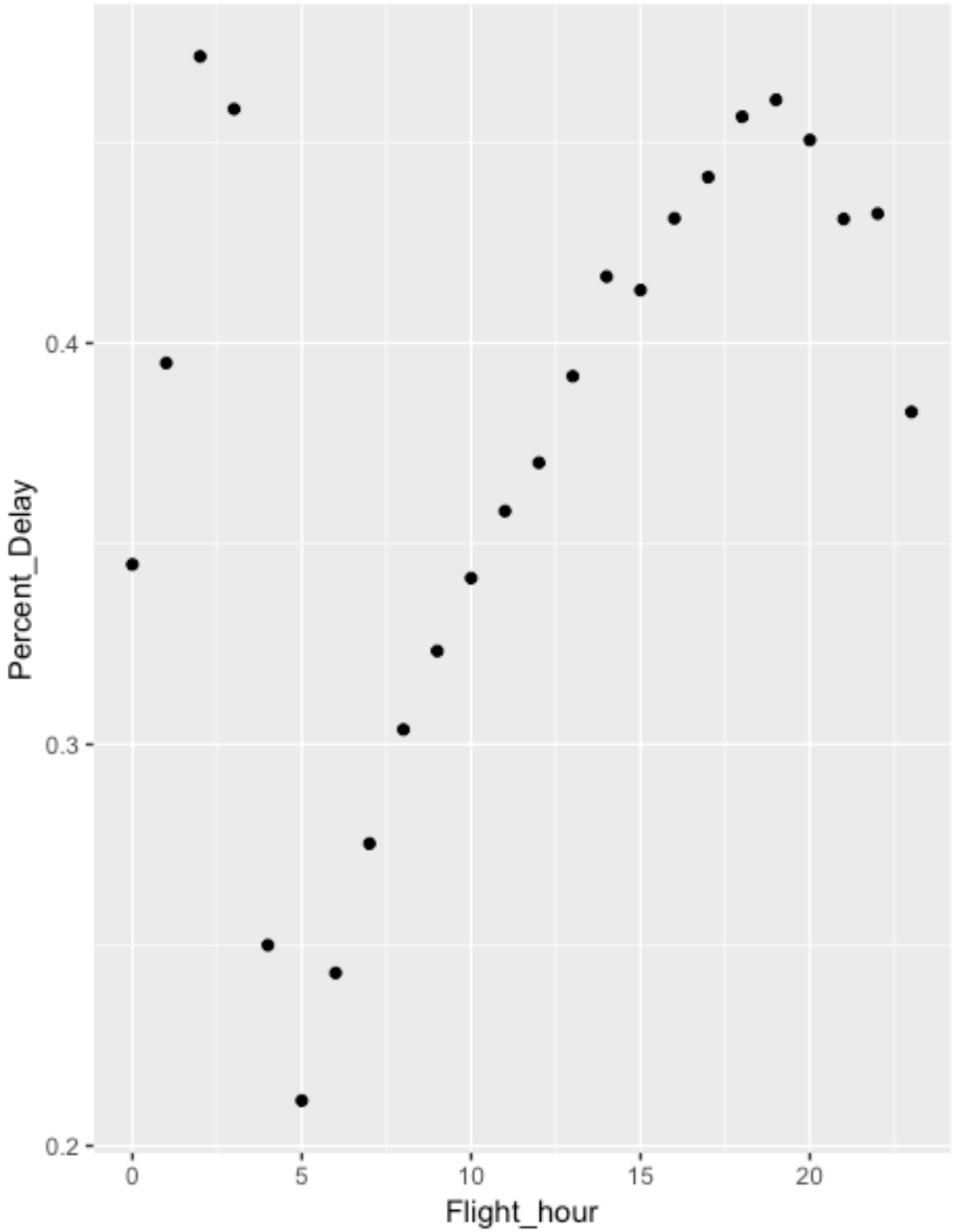
The first data point I wanted to look at was percentage of flights that are late (delayed) upon arrival. Initially, I thought this would be 10-15%. However, according the data that I looked at, around 37.4% of flights are delayed - over 150% higher than I expected! A slightly higher number of flights were delayed for Departure. Departing flights were delayed around 39% of the time for the sample I looked at.
At a rate of 37.4%, a person would have the following expected chances of having a delay:
P(delay given n number of flights) = 1-(1-0.374)^5 1 flights - 37.4%
2 flights - 60.8%
3 flights - 75.5%
4 flights - 84.6%
5 flights - 90.4%

Probability of delay given an event

The next data point I looked at was determining the probability of an arrival delay given that a departure delay occurred.
P(arrival delay|departure delay)
The results were not surprising. When there was a delay on the departure end of flight, there was likely a delay on the arrival side. The probability of a delay on arrival give there was a delay on departure was observed was 71.1%. Where there was not a delay on departure, you were only expected to be delayed 16.3% of the time. This expected outcome seemed obvious, but I wanted to review it because it might be a good indicator to use in a regression model.

Probability of delay given hour of departure

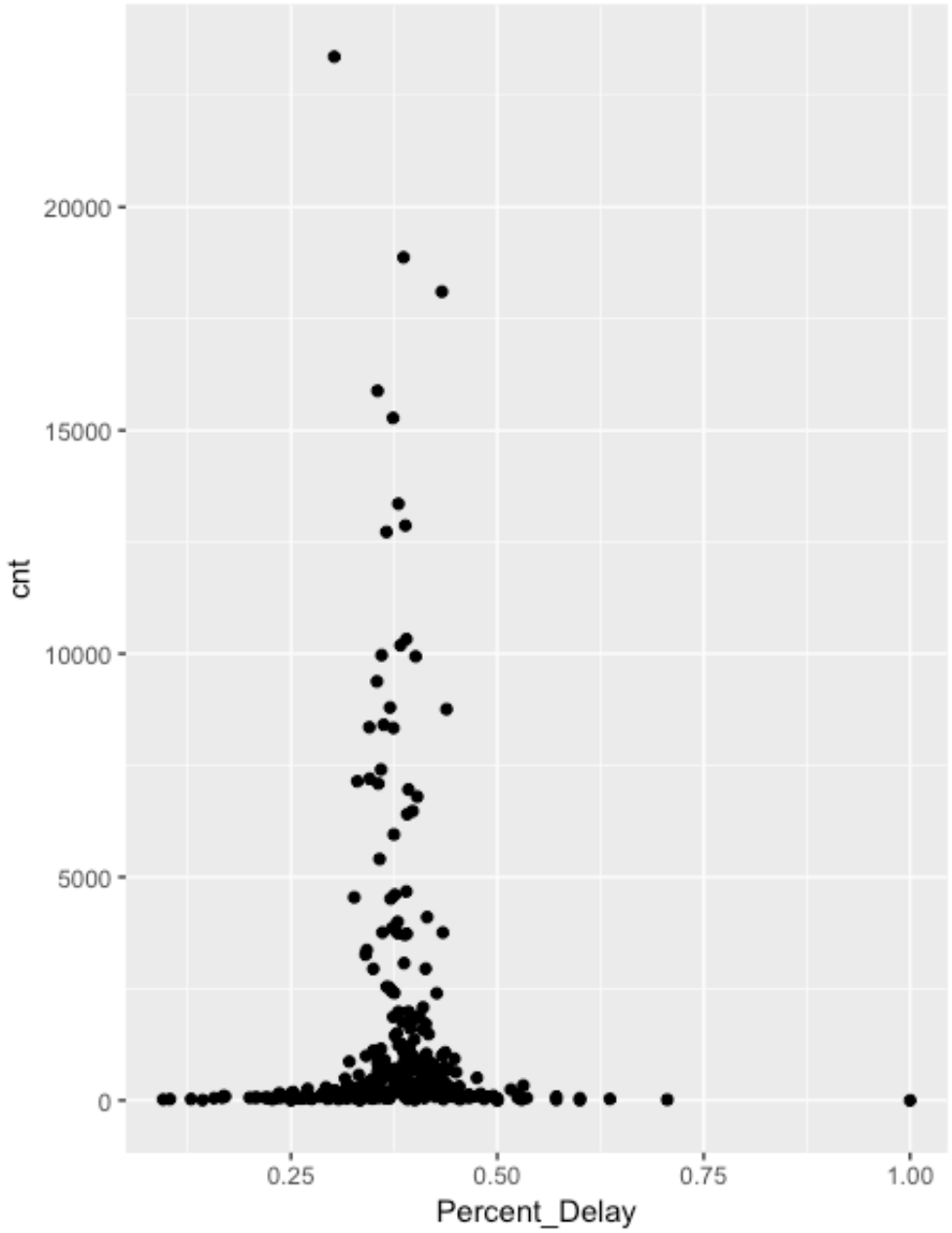
This statistic is the one of the main themes of my analysis. I expected that the results for the morning would show a lower probability of delay. My expectations generally follow what was shown in the data. The results show that early morning hours, that is 4 or 5 AM, show a lower expected probability of delay (~25%) than later in the day and those hours after midnight (+40%).



Delay by flight hour

Size of Airport

I wanted to see if the size of an airport was correlated to the probability of a delay. I expected busier airports would have more delays. However, initial results do not support this idea. It seems that regardless of size, most of the larger airports seems to hover in the 37% range according to the graph.



Airport size

Machine Learning
Logistic Regression

The first machine learning technique I looked at was logistic regression. The main goal I was hoping to achieve was to determine whether or not a flight was delayed. Therefor, my dependent variable was a field I created Departure_Delay_BOOL that is a boolean where 0 represents no delay and 1 represents a delay. The independent variables I used to test for a delay were:

Flight_hour: the 2-byte hour for the scheduled departure. This was converted to a factor.
State: this is the US State where the departure airport is located.
season: the meteorological season. For example Jun, Jul & Aug are summer Dec, Jan, Feb are winter, etc.
Day_of_Week: the day of the week the flight departed.
Taxi_Out: the amount of time it takes for a airplane to taxi into position for departure

The baseline model, where it was predicted that there were no delays gave an accuracy of around 61%. Based on this, I was hoping to get at least 70% accuracy. However, despite all the combinations I tried, I was not able to get better than 62%. Or, a 1% increase in performance over the baseline.

Initially, I classified *all* departure delays greater than 0 as a delay. However, this creates two issues in the analysis. The first issue is that not all delays are posing issues and need to be tracked. A delay of 1 minute was being treated the same as a delay of 1 hour. These are two different severity levels and should be treated different. The next issue is that the factors I was testing were trying to predict different kinds of delays. Season and to some extent Flight_hour and State are mainly factors to consider for weather delays while Day_of_Week and Taxi_Out are more related to the airlines operational delays or air_system_delays. After splitting up the delays and identifying attributes separately, I was able to get overall delay predictions of 62%, however, overall model accuracy was only around 50-55%.

Linear Regression

The other ML technique I attempted was linear regression. I wanted to take another angle and assess arrival delay to see if I would have better results. The following elements are the independent variables that I compared to arrival delay:

ELAPSED_TIME: the amount of time the flight took en route + Taxi time
DISTANCE: the length, in miles, of the flight

Both variables I picked turned out to be significant with 3 starts each from the R summary. However, results from this comparison were quite low and only about 5% of the variation is explained by the model.

Recommendations

1 Travelers should aim to leave earlier in the morning.

Based on my findings, there are a couple of reasons travelers should aim to leave earlier in the morning.
Travelers will have less of a chance of getting delayed if they leave earlier in the morning.

Morning (05 - 10) : 29%
Afternoon (11 - 16): 39%
Evening (17 - 22): 44%
Overnight (23 - 04): 37%

If a traveler is delayed, the severity of the delay will be less compared to other times of the day. The average delay for the time of day is as follows:
Morning (05 - 10) : 28.7 min
Afternoon (11 - 16): 33.6 min
Evening (17 - 22): 37.6 min
Overnight (23 - 04): 30.1 min

2 Days of the week

In terms of delays, the day of the week was quite close and there are not actionable results one should take. The average arrival day was between 31 and 36 minutes with Friday having the smallest average delay and Monday having the largest. When looking at frequency of delays, Thursday had the most arrival delays with a delay around 40% of the time. Tuesday and Wednesday were only delayed 36% of the time.

3 Travelers should consider the following airports, if possible*:

The airports listed here were those that were listed in the top 100 for frequency of delays were Honolulu, Los Angles, and New York-Laguardia. These airports had delayed flights 40% of the time.

In terms of best performing airports out of the top 100, Salt Lake City, Atlanta, Portland, Detroit and Chicago-Midway were all under 35% in terms of delay frequency with Salt Lake City and Atlanta at only around 30%. Surprisingly, Atlanta is the business Airpot in the world!
*note, only the top 100 airports in terms of volume were considered. This was due to small airports having very frequent delays and low volume (e.g. St Cloud regional airpot was delayed +60% of the time but only had 48 flights)

4 Overall Airline Performance

This chart shows the airlines with the highest average delay and which airlines should be avoided.

AIRLINE	avg
Frontier Airlines Inc.	42.06258
Spirit Air Lines	41.44998
American Eagle Airlines Inc.	39.74338
United Air Lines Inc.	39.65803
JetBlue Airways	38.89763
Atlantic Southeast Airlines	35.70021
American Airlines Inc.	34.61215
Skywest Airlines Inc.	33.00361
Delta Air Lines Inc.	32.58150
Virgin America	30.85803
Southwest Airlines Co.	29.82274
US Airways Inc.	27.41993
Alaska Airlines Inc.	22.80630
Hawaiian Airlines Inc.	15.51550