Homework 4

Victoria Easton

2022-09-21

homework 4

For this homework, we will continue to develop our first project. Please expand your written proposal to the start of a report and submit this as a **PDF**.

1. Introduction (5 points)

You should have a introduction section with three subsections

- Opening: This is typically the first paragraph that introduces the larger problem the paper is targeting. What is the context, why is it important?
- Background: What information does the reader need to understand the specific work the authors did, why is it important, and what will it contribute to the larger issue?
- Challenge: What are the specific hypotheses/questions/goals of the current work?

2. Data Section (5 points)

Include a written description of the data, including the variables and how the data were collected.

Also create a series of figures that that illustrate how the predictor variables explain the response. Make sure to include appropriate labels, titles, shading of points, etc. . .

Introduction

Everyday 2.9 million people fly across the US's 5.3 million square miles of domestic airspace. Every year the Federal Aviation Administration (FAA) handles 16.4 million flights throughout the US. However, with all this flight traffic comes delays. There are not many people left who would say they have never been delayed on a flight before. According to the FAA, the biggest cause of flight delays is weather, which accounts for about half of flight delays. The other half of flight delays are caused by volume, equipment, and runways. The FAA estimated that the total yearly cost of flight delays was $33 billion in 2019, $18 million of which was loss to the passenger. Another $2.4 million dollars is estimated to have been loss by passengers avoiding future air travel due to the delays. Flight delays are a big part of air travel and cost companies and people lots of money. The burning question is whether these flight delays can be predicted.

In order to answer this burning question, we can fit a multiple linear regression model with a flight's delayed time as the response. This will help predict flight delays so pasagengers and arilines can have more information.

The goal of this paper is to see what aspects of flying that can predict flight delays at arrival. More specifically does scheduled departure time, departure delay time, flying time, distance, origin/destination of flight, and aircraft carrier have and effect how delayed a flight is from it's original landing time.

Data

The data set came from kaggle.com but was originally from The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) website. BTS tracks information on domestic flights by large airlines and makes them avaliable to the public. The full data set contains information on flight arrival and departure details for all commercial flights within the USA, from October 1987 to April 2008. Since this is a very large data set, the data set obtained from kaggle.com contains only data from the year 2008.

A lot has changed since 2008 and to have the dataset be more relatable to today's airline's, the carriers that no longer operate have been removed from the dataset. Flights that also did not have an arrival delay time were also removed from the dataset. This created a new dataset with information on 612,923 flights. The arrival delays ranged from -67 minutes to 1,525 minutes. This means a flight originally that departed late was still able to land over an hour earlie and the longest delays landed over 24 hours after its originally scheduled time. Flights had a departure delay anywhere from 6 minutes late to 1,521 minutes late. The majority of flights leave between 11:45am and 6:30pm with the majority of flight landing between 1:30pm and 8:20pm. Flight's flying time time range from 15minutes to 727minutes for an average distance of 798.6 miles.

Carrier abbreviations:
AA -> American Airlines Inc.
AS -> Alaska Airlines Inc.
B6 -> JetBlue Airways
DL -> Delta Air Lines Inc.
F9 -> Frontier Airlines Inc.
HA -> Hawaiian Airlines Inc.
OO -> SkyWest Airlines Inc.
UA -> United Air Lines Inc.
WN -> Southwest Airlines Co.
YV -> Mesa Airlines Inc.

Source: https://www.faa.gov/air_traffic/by_the_numbers

```
Flight %>%
    ggplot(aes(y=`ArrDelay`, x=`Distance`, color= UniqueCarrier)) +
    geom_jitter(alpha = .2)+
    facet_wrap(.~UniqueCarrier)+
    labs(title="Satterplot of Flight's Arrival Delay Based on the Distance Traveled",
         subtitle = "by Airline in 2008",
        caption = "Source: https://www.kaggle.com/datasets/giovamata/airlinedelaycauses
        *Carrier Abbreviations Above",
        fill = "Carrier",
        x="Distance (miles)", y="Arrival Delay (min)")+
    theme(plot.caption = element_text(face = "italic"))
```

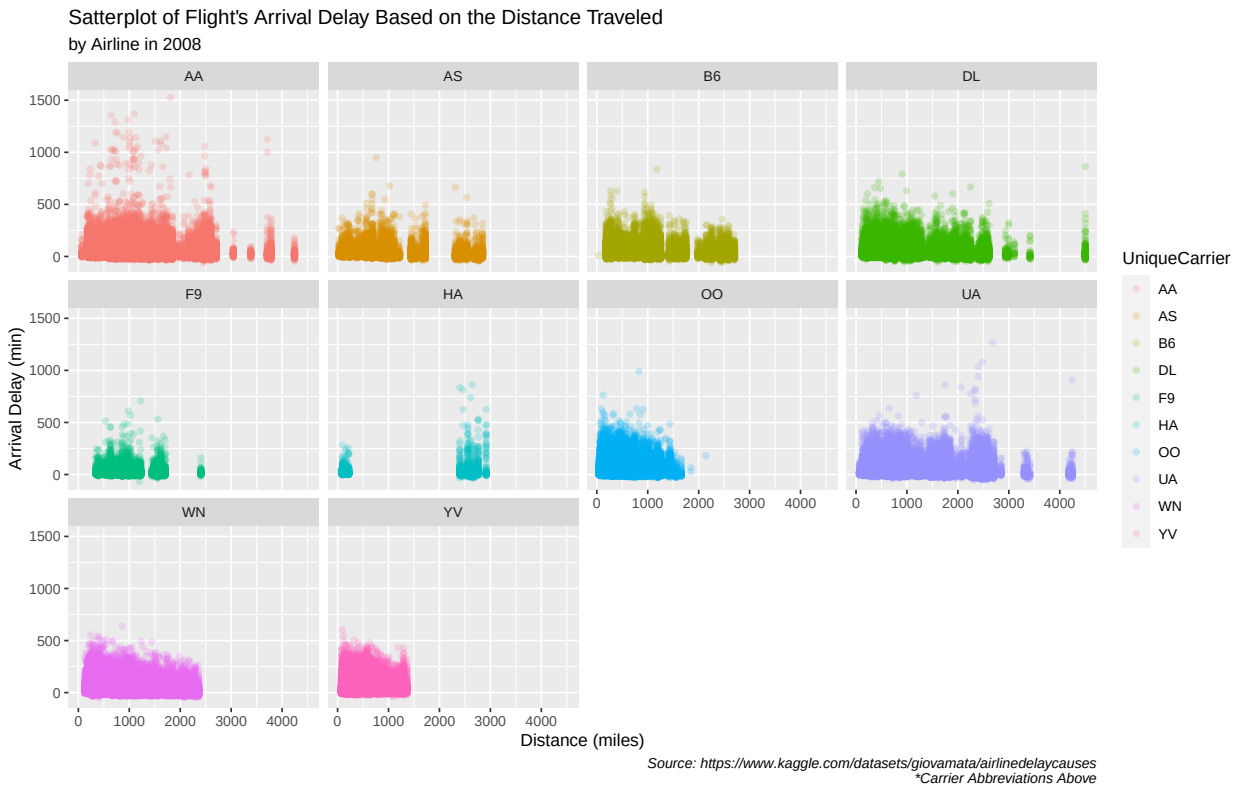

Figure 1: There is not much of an overal trend in arrival delays as the distance of the flight increases. However, the carriers that do not travel far tend to have shorter arrival delays. A closer look at American Airlines shows there are many flights that did not travel ver far with very long arrival delays.

```
Flight %>%
    ggplot(aes(y=`ArrDelay`, x=`DepDelay`)) +
     geom_jitter(alpha=.1) +
     xlim(0,500)+
     geom_smooth(method = "lm")+
    labs(title ="Scatterplot of Flight's Arrival Delay Based on Their Departure Time in 2008",
         caption = "Source: https://www.kaggle.com/datasets/giovamata/airlinedelaycauses",
         x= "Departure Delay (min)", y= " Arrival Delay (min)")+
    theme(plot.caption = element_text(face = "italic"))
```

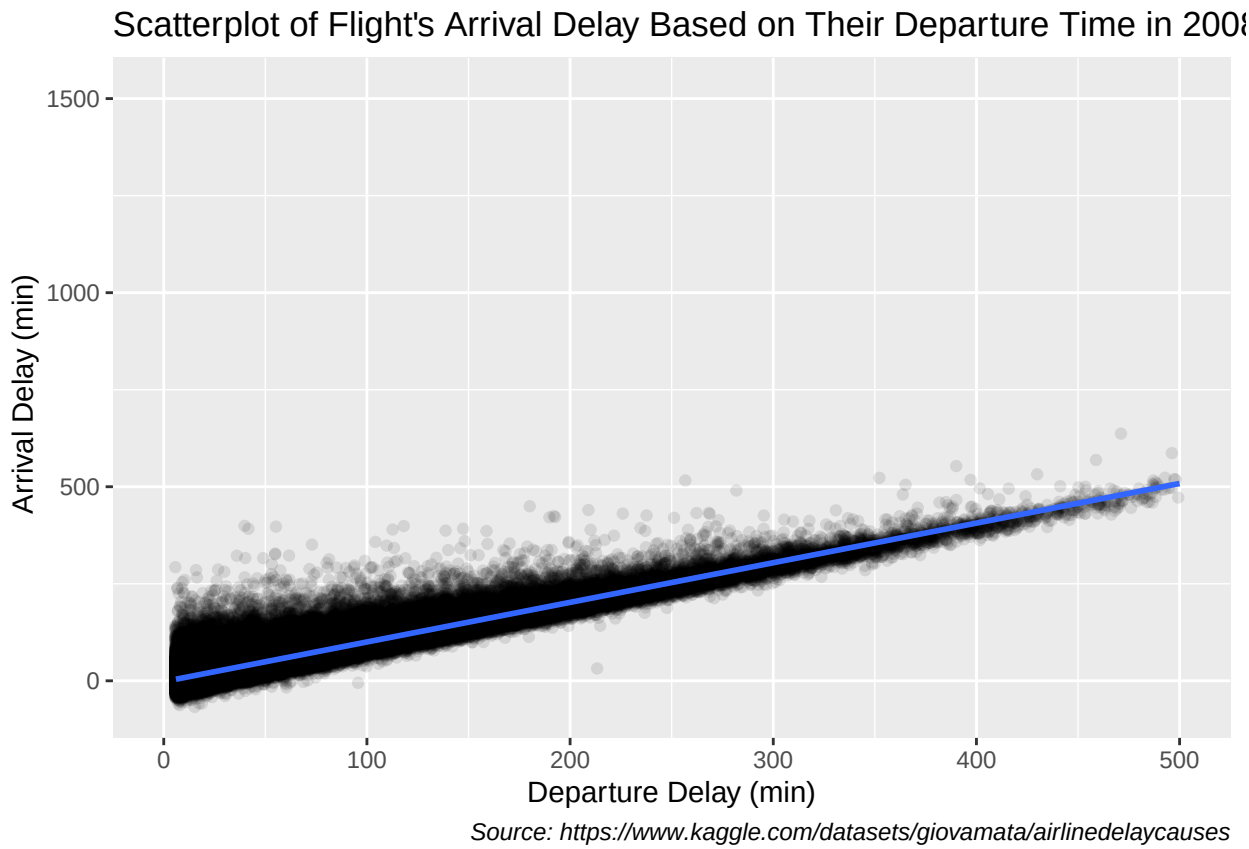

Figure 2: The Figure shows the arrival delays for flights that had less than a 500 minute departure delay. As the departure delay increased the arrival delay tended to also increase. Howvever this figure shows that for the flights delayed less than about 3.5 hours there are more flights that landed more than 3.5 hours later than landed earlier than 3.5 hours behind scheudle. Above a 3.5 hour departure delay, the arrival delay is more similar to departure delay.