ORD Flight Delays EDA

Aden Ramirez

February 18, 2023

Abstract

ORD is one of the biggest and busiest airports in the United States suporting hundreds of thousands of flights a year. The data that will be analized November 2021 - November 2022 flights at O'Hare international Aiport obtained from the Bureau of Transportation Statistics. This exploratory data analysis will explain te process of cleaning the data and its methods. Then will explore some basic stats on delays, and how the airport itself is used and delayed. Last a couple intuitions of some possible variable relationships will be visualized.

1 Important Terms

This contains some terms and information that will be used throughout the EDA that may not be common knowledge to help you understand.

1.1 General

Delay - A flight delay is defined by the Federal Aviation Administration (FAA) as a flight arriving 15 minutes or longer after scheduled arrivale time

1.2 Governing Bodies

FAA - Federal Aviation Administration is (FAA) is the governing body of all aviation in the U.S from aircraft certification to air traffic control

DOT - Department of Transportation (DOT) is a governing body that regulates all transportation, in this case they offer many unique identifiers and are the parent body to the Bureau of Transportation Statistics (BTS) that the data is sourced from.

IATA - International Air Transport Association (IATA) is an organization that makes a standard for transport type aircraft and in our case create the airport codes referenced throughout this research.

ICAO - International Civil Aviation Organization (ICAO) is a United Nations body that has standards for worldwide air travel. They are less important in this research as we look at domestic travel, they do play in role in how the U.S forms their regulations.

1.3 Airline Theory

Hub and Spoke - This is a major theory of how to operate and airline, in short it means the airline has many big airport hubs that fly to other hubs. Each hub will then fly the passenger to thier final destination.

Point to Point - This is another rivaling theory to hub and spoke, this means the airline will just fly the airport to the destination with not stop.

2 The Data

This data was collected from the Bureau of Transportation Statistics (BLS) as part of the online Airline On-Time Arrival Performance Data. I selected one year of data to begin, choosing the most recent as of Janurary 2023. The data range spans all U.S. Carriers flying domestically over the course of

November 2021 - November 2022. The data has many fields totaling to 120 columns. BLS provides an excellent readme that explains each data field, though I will highlight some very important ones to my research here.

Note: _ is not shown in these data types as they appear in the .CSV

2.1 Basics

There are alot of columns that are basic information that would be expected of most sets, there is alot of options in this set with plenty of formating opertunity.

Year - Year (4 digit)
Quarter - Quarter of the year (1-4)
Month - Month
DayofMonth - Day date of the month
DayofWeek - Day in words, such as 'Monday'
FlightDate - Flight Date Aggregation (yyyymmdd)

2.2 Indentifiers

There are many indentifiers included that are not entirely useful for my use as a DOT Marketing ID, but worth having should there be NA data columns that can be cross referenced. There is also multiple ways to indentify *anything* in aviation, so there are international Air Transport Association(IATA) and US identifiers in this set. The most important for my analysis are:

Marketing Airline Network - These are common codes for airlines someone would see on their ticket, such as UA for United Airlines

Tail Number - Aircraft Tail Number, can uniquely identify a register aircraft

Flight Number Operating Airline - Flight number as you would see on a ticket for example: UA1234

2.3 Flight Information

2.3.1 Route Information

All the information pertaining to the route of the flight:

Origin - Origin Airport in IATA standard (3 Letter) Example: ORD or PHX

OriginCityName - Origin Airport City (There are special cases such as CLT)

OriginState - Origin State Code

OriginStateFips - Origin State FIPS code, this will allow for adding state geometry in maps

OriginStateName - Origin State Name string

Dest - Destination Airport in IATA standard

DestCityName - Destination Airport City

DestState - Destination State Code

DestStateFips - Destination State FIPS code, this will allow for adding state geometry in maps

DestStateName - Destination State Name string

2.3.2 Delay Information

All the information pertaining to the timing of the flight, and delay if applicable. There is much more, but this is a good focus for the analysis and EDA. Extensive diversion information is not included, as it is noted, but not the focus of the research.

CRSDepTime - CRS Departure Time (local time: hhmm)

DepTime - Actual Departure Time (local time: hhmm)

DepDelay - Difference in minutes between scheduled and actual departure time

DepDelayMinutes - Difference in minutes between scheduled and actual departure time

DepDel15 - Departure Delay Indicator if the flight is at least 15 minutes delayed

DepartureDelayGroups - Departure Delay intervals, every (15 minutes from ≤ 15 to > 180)

DepTimeBlk - CRS Departure Time Block, Hourly Intervals

TaxiOut - Taxi Out Time, in Minutes

WheelsOff - Wheels Off Time (local time: hhmm)

WheelsOn - Wheels On Time (local time: hhmm)

TaxiIn - Taxi In Time, in Minutes

CRSArrTime - CRS Arrival Time (local time: hhmm)

ArrTime - Actual Arrival Time (local time: hhmm)

ArrDelay - Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.

ArrDelayMinutes - Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.

ArrDel15 - Arrival Delay Indicator, 15 Minutes or More (1=Yes)

ArrivalDelayGroups - Arrival Delay intervals, every (15 minutes from ≤ 15 to > 180 minutes)

ArrTimeBlk - CRS Arrival Time Block, Hourly Intervals

Cancelled - Cancelled Flight Indicator (1=Yes)
CancellationCode - Specifies The Reason For Cancellation
Diverted - Diverted Flight Indicator (1=Yes)
CRSElapsedTime - CRS Elapsed Time of Flight, in Minutes
ActualElapsedTime - Elapsed Time of Flight, in Minutes
AirTime - Flight Time, in Minutes
Flights - Number of Flights
Distance - Distance between airports (miles)
DistanceGroup - Distance Intervals, every 250 Miles, for Flight Segment
CarrierDelay - Carrier Delay, in Minutes
WeatherDelay - Weather Delay, in Minutes
NASDelay - National Air System Delay, in Minutes
SecurityDelay - Security Delay, in Minutes

3 Data Cleaning

3.1 Aggregation

Collecting the initial one year (11/21-11/22) data from BTS yields twelve seperate .zip files containing a .csv file and .html documentation readme. The first step to using this data is to concatanate all this information together into one usable format. I accomplish this with a Python script using the *Pandas* library. This script pulls in all .csv files into a list of Pandas Dataframes. Once these are Pandas Dataframes it is quite simple with all variables matching, they get concatanated. Now with a single monolithic Dataframe, it is written to a new .csv file.

3.2 Subset Files

This research is focusing on a single airport, Chicago O'Hare International Airport. The data sourced is for all U.S Airline Carriers that fly domestically. The first step in cleaning this data is to properly subset the information into a smaller, less bloated file to reference. My approach will produce to files as output.

First, one main file will use the *Pandas* library in Python to pull in our single monolithic csv file into a Dataframe. From here it is quite simple to mask only the data we want. I filter by either the *Origin* or *Dest* to be the airport of interest ORD. Once this mask is applied, the values are verfied to be an expected value, and once again exported using the next masked dataframe to a seperate .csv file.

Second, a second file that contains only variables being heavily used, to help with operation speed. All the variables listed in section Figure 2 will be maintained. The process will follow the same process as section Figure 3.2. The output will be a much more compact file to operate on, to aid speed for the EDA and future model development.

3.3 NA Values

Now that the data has been subset into what will be used to explore, the next concern is NA values. There are some columns where this is acceptable, but other such as out delay times that are not. The first step is finiding all the columns that have an NA value and then determining what the value should be replaced with. Many of the delay related columns have many NA values, the delay types (Weather, NAS, Security, and Late Aircraft) are null if they are not applicable. These columns get their NA values set to 0, as in no delay minutes. This takes care of the NAs that could corrupt data. The last big concerning item is flights that have no delay value, either arrival delay or arrival delay minutes, which could not be creted. Any NAs of these are dropped. This operation dropped *17004* data points. This is signifigant, but still far less than 5% of all the data. This now leaves the data with *582721* observations.

3.4 Miscellaneous

Outliers in this dataset can get large. I decided to explore with a data subset where all delays are *6 hours (720 min.)* or less (still including non-delays). This decision is not final as a better consideration for what an outlier is should be condicted in this set. This does simplify understanding the data, and making the point through visualiztion easier for the purpose of this EDA. With this data mask the observations left is now *582045*.

4 Exploratory Data Analysis

To begin exploring the data there is some basic information to gather to learn about this data and begin to show some areas that may be worth investigating further with machine learning models.

ORD Amount of Flights: 582045
ORD Amount of Delays: 106187
ORD Percent of Flights Delayed: 18%
ORD All Flights Delay Mean: 3 minutes

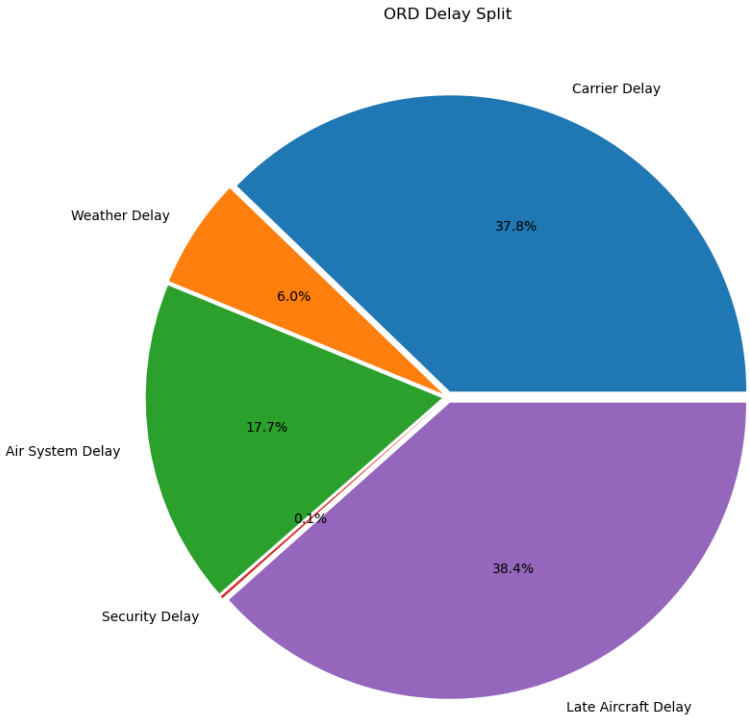

Figure 1: ORD Delays by Type

As seen in Figure 1 the two biggest delays are by carrier and late aircraft. The research questions are going to dive much deeper in how these delays are created and interacted with by airline and airport staff. Prediciting and eliminating these delays could say millions of hours a year.

ORD All Flights Delay Median: -8 minutes (8 minutes early!)
ORD Delayed Flights Mean: 67 minutes
ORD Delayed Flights Median: 42 minutes

To begin breaking down the delays that occur, there are five categories officially made by the FAA in this data set tracked. Carrier Delay, Late Aircraft Delay, National Air System Delay, Security Delay, and Weather Delay.

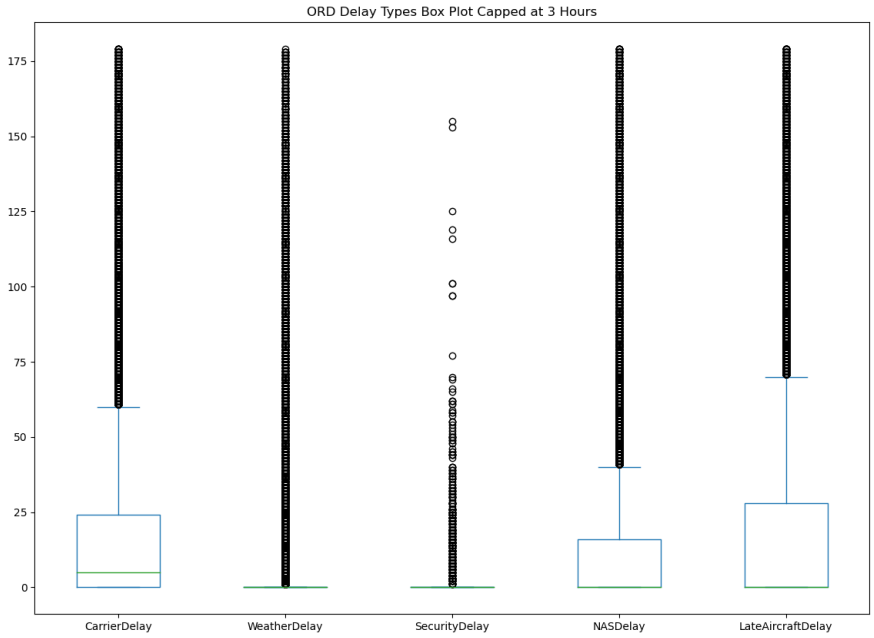

Figure 2:
In Figure 2 This box plot is capped at the highest delay group in the data, 180 minutes or 3 hours. The large amount of flights are delayed for short amounts of time, but you can see the two largest areas cover much higher ranges of delays.

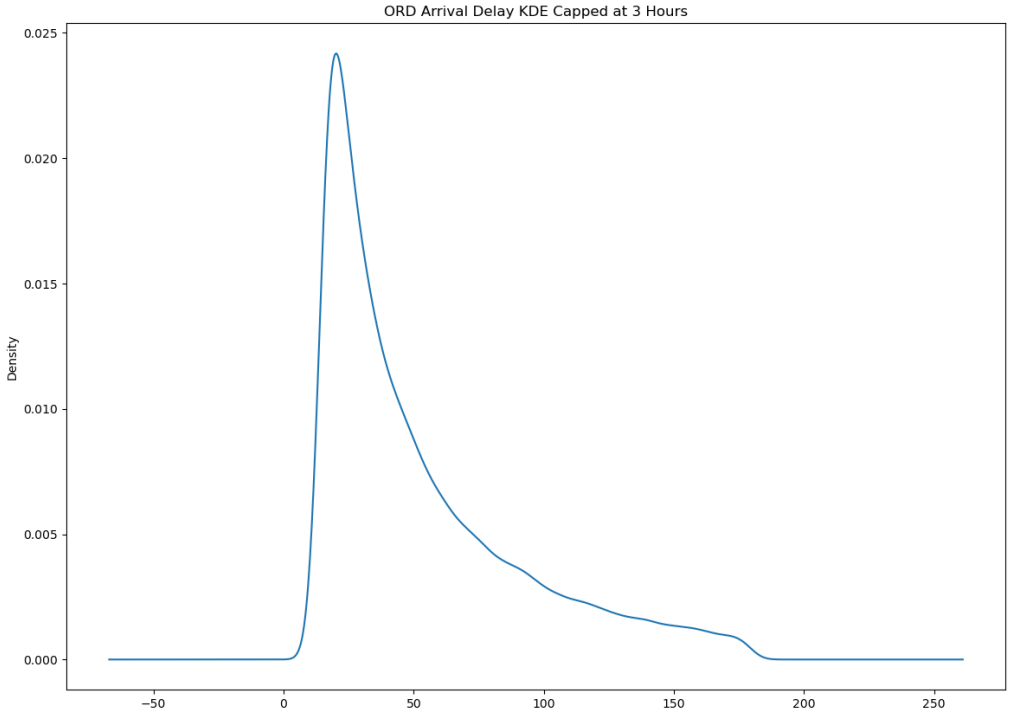

Figure 3:
In Figure 3 looking at arrival delays, we can see the highest density peaks well before the hour mark sharply decreasing to a non zero before the max included amount (180 minutes).

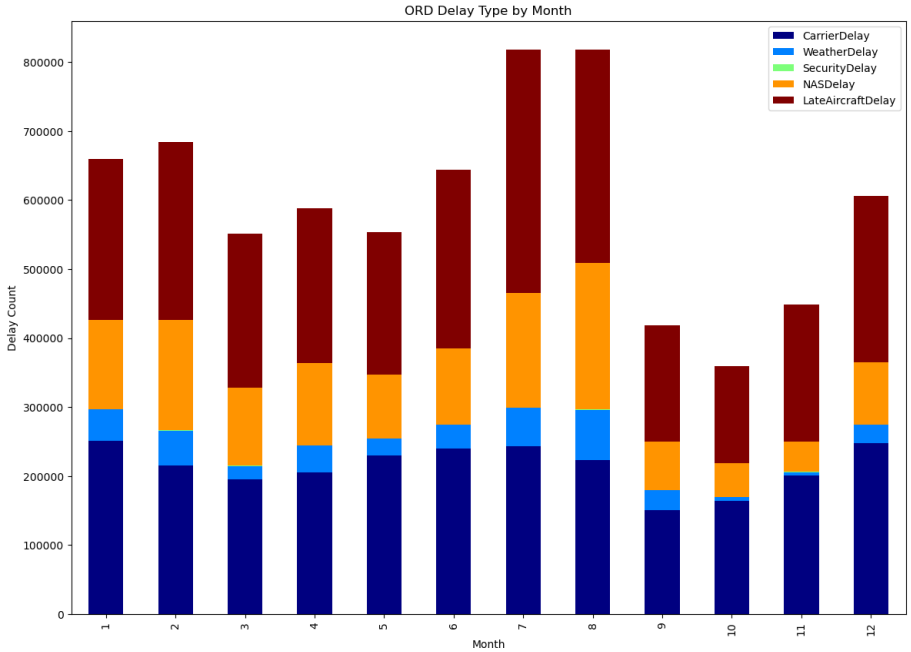

Figure 4:
In Figure 4 this reflects similar to the pie chart, of note is the summer months have a much higher delay count. The busy travel season could be contributing to this, and worth further investigation.

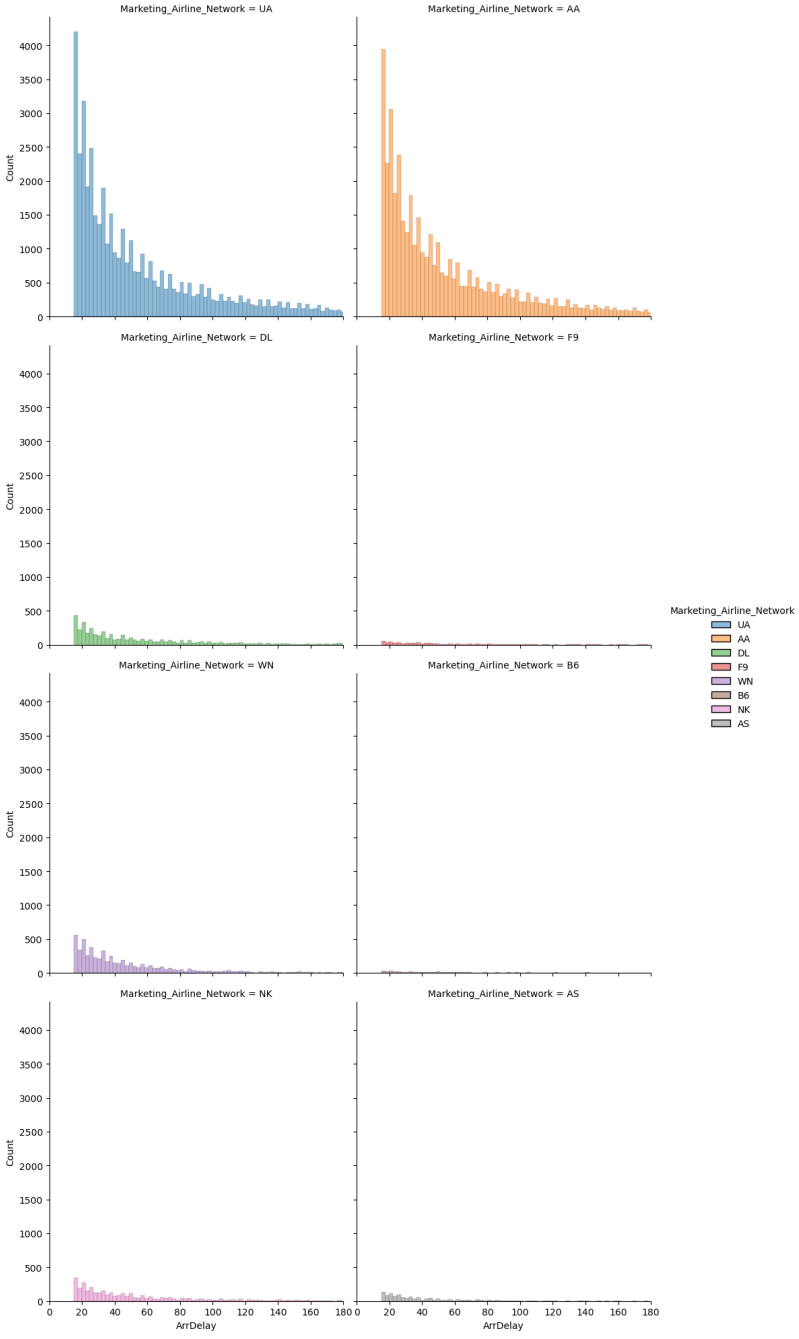

Figure 5:
In Figure 5 this shows the amount of delays distibution by each operating airline. United and American have the largest, which can be expected as ORD is a hub airport for American Airlines, and United Airlines who both fly a hub and spoke model.

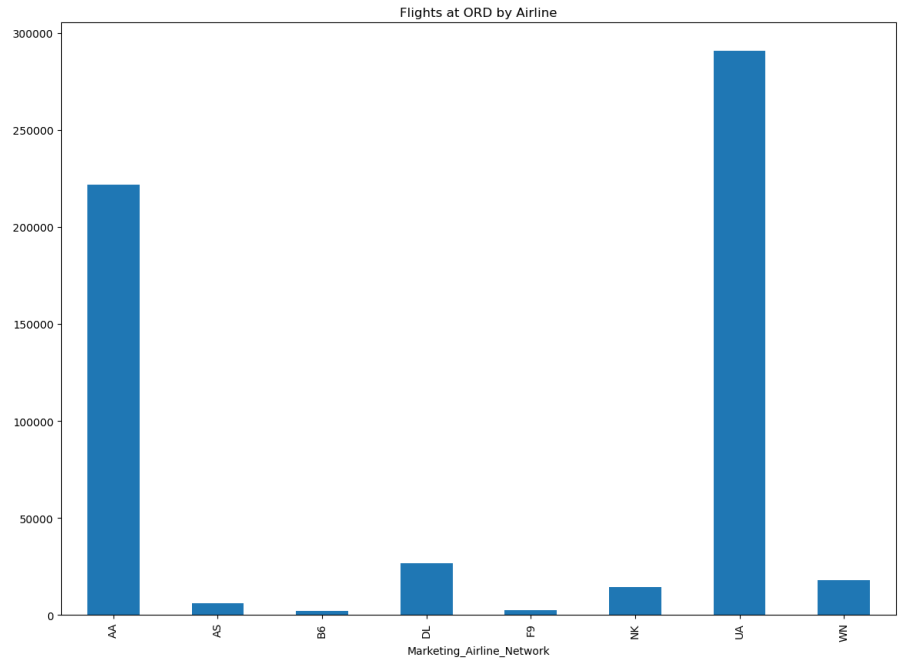

Figure 6:
In Figure 6 To show the previous figure, American and United fly the most to ORD, though there is a signifigantly larger amount operated by United. In Figure 5 there is not as large a difference in their delay distributions.

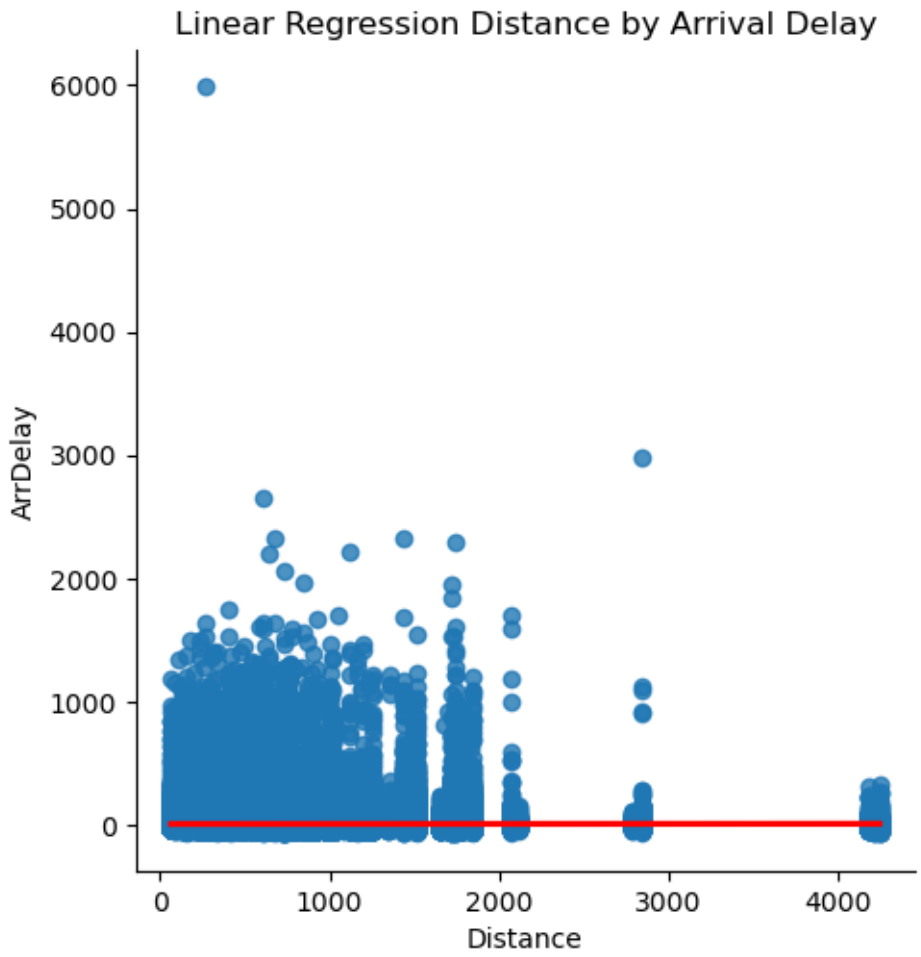

Figure 7:
In Figure 7 One of the relationships I wanted to briefly look at to further think about in the coming day was distance of the flight and the arrival delay. This is a very stange visual I want to look into more, there is almost a regonizeable cluster for certian distance airports. Theis has no relationship between just these two variables.

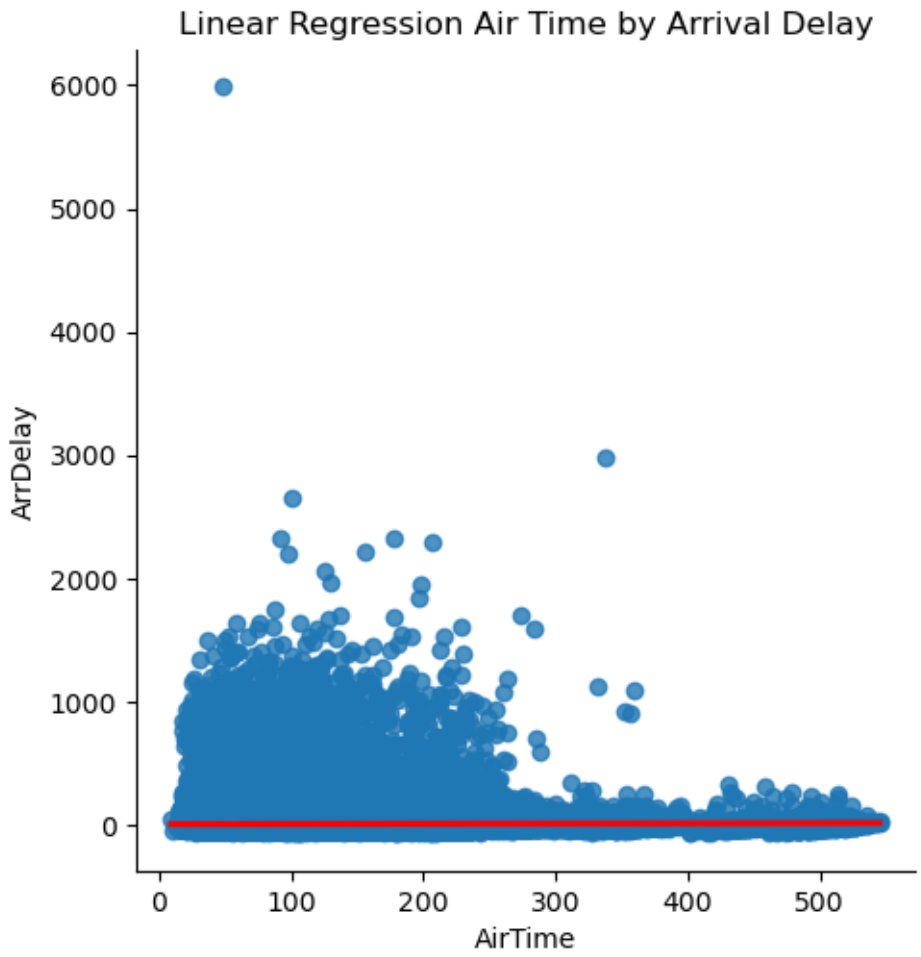

Figure 8:
In Figure 8 The other relationship I wanted to preview was airtime by arrival delay, the thinking being more time in the air laves possibility to make up time flying faster or catching good wind. This very simple linear regression shows not simple relationship, though at a glance looks like more airline has lower delays, but there is obviously a more complex relationship.

5 Conclusion

This exploratory data analysis cleaned the data to be more usable, indentified some areas the data could be improved, and explored some basic relationships. Moving forward to modelling some this EDA will prove valuable to understanding the data, and helping work towards the data being easily workable for machine learning models.

5.1 Data Improvement

During the EDA there is some spots of the data that can be improved and expanded. For future visualizations it would be helpful to have a nicer way of showing information such as airline names or even a map showing where flights go. There are supporting tables provided by BTS that would assist in expanding tha data. Another consideration is how to deal with outliers when building models there is a large range, that shouldn't eliminate all high delays. This should be solved mathmatically and researched what the FAA considers long delays, or a point of where delay is considered a new flight. Working with the data could also be made easier with some table work, on the names, and adding new information that is used more than once.

5.2 Relationships and Methods

In the EDA only a couple areas were explored, and both showed the problem is much more complex than a linear regression. Thinking ahead I think the possibility of a logisitic regression is there that could provide a good predictor or a delay. Another area I did not consider earlier is trees that could prove promising. There may be the need for some dimension reduction as these complex relationships may start adding uo in variable size. Lastly clustering may still be a good option where there are splits on more focused data.