Analysis of United States Flight Data: Delay and Cancellation

NYU Tandon School of Engineering

CS-GY 6513: Big Data, Spring 2022

Professor Juan Rodriguez

Date *May 17th, 2022*

U.S. Flight Delay and Cancellation Analysis

Introduction

The United States has a massive air transportation system, with over 80 airports handling millions of passengers each year. With the nation featuring several major cities spaced hundreds of miles apart, safe and efficient transportation is a crucial component of infrastructure, supporting commerce and livelihoods. The industry yields an extensive and comprehensive corpus of data. Our project aims to get a better picture of the landscape and derive some practical insights.

In this exploratory data analysis, we evaluate the on-time performance of scheduled flights originating and arriving at airports in the United States. We explore various potential factors that contribute to delays and cancellations, investigating whether some airlines are more reliable than others and what times of the year are likely to experience more disruptions, for example. Visualizations and observations are included to help reveal potential trends. Accompanying this report are files containing the code and queries we ran to generate our results and visualizations. We conclude with a summary of our findings and a proposal for an extension to our project: an application that would enhance our capability to process flight record data in the future.

Data Sets and Technologies Used

https://community.amstat.org/jointscsg-section/dataexpo/dataexpo2009
https://www.kaggle.com/datasets/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018

We performed a historical analysis using two separate data sets from 1987-2018. The first major data set was collected from the American Statistical Association (amstat). From the site:

> The data consists of flight arrival and departure details for all commercial flights within the USA, from October 1987 to April 2008. This is a large dataset: there are nearly 120 million records in total, and takes up 1.6 gigabytes of space compressed and 12 gigabytes when uncompressed.

The second major data set was found on Kaggle and is originally from the Bureau of Transportation Statistics. It consists of 8 gigabytes of airline travel data. This data was analyzed using Spark, NumPy and Pandas. Spark was needed since the distributed nature allowed us to analyze these large data sets.

Preliminary Data Analysis

Before proceeding with more detailed analysis, we first aimed to develop a high-level understanding of the data. In particular, we needed to determine the significance of the fields available - what they describe and how they represent it - since this would inform not only how we interpret the results but also *how we work with the data* to obtain those results. We imported a small sample of records from the **2018.csv** file obtained from Kaggle into a pyspark SQL context in a Jupyter notebook. Our preliminary querying on this sample yielded some surprising findings about the data.

Please refer to the file titled *preliminary_data_analysis(scheduled_flights_sample).ipynb*

As one would expect, most of the fields in the data set pertain to time. There are *points* in time describing when some event occurred and *durations* (amounts) of time, with the minute as the unit of precision. The first major surprise is in the data type used: both categories are represented as float64. This makes sense for durations of time; the field value is simply how much time some process took, in minutes. However, a point in time refers to a time of day using two digits for the hour and two digits for the minute in 24-hr time; for example, a value of 1512.0 refers to 3:12 pm on the specified date (which is indicated in a different field). Under this scheme, any possible value (time of day, precise to the minute) has a unique representation, but numerical calculations are unintuitive; decimal notation is used even though there are 60 minutes in an hour.

Another consideration is the *standardization* of data. The data sets are centrally maintained by federal agencies under regulation, rather than relying on unaffiliated local entities. Accordingly, the data take on consistent, predictable values; in particular, airports are referred to (exclusively) by their 3-letter International Air Transport Association (IATA) codes, rather than

by geographical location or name (which could be subject to abbreviation). This adherence to nomenclature consistency prevents the inadvertent creation of multiple variations of what are actually the same value. Lastly, time data are by nature objective; there is no room for interpretation when recording or describing the minute some event took place.

There was one surprising finding related to data standardization, however: the time point indicating when an event occurred is expressed in the time zone where the event occurred. If, for example, we wanted to determine the order in which various flights took off, we could not simply sort the records according to the **WHEELS_OFF** field, since those values could be from different time zones. Furthermore, since a flight can of course span multiple time zones, the time point values within the *same* record could be from different time zones. Before discovering that local time zone recording was used, we expected that all time point values would be standardized to a single time zone.

The data are quite comprehensive, documenting the various events and stages of the flight process (e.g. taxiing between the gates and the runways in addition to being in the air), as well as the scheduled and actual time points/durations. In fact, it seems that some of the data are redundant because some fields can be derived from others. As an example, it is intuitive that one can determine a flight's departure delay simply by finding the difference between the *scheduled* and *actual* departure times. Yet the data set includes these 3 fields:

- **CRS_DEP_TIME**, the scheduled departure time
- **DEP_TIME**, the actual departure time
- **DEP_DELAY**, the amount of time the flight's departure was delayed

Demonstration: predicting **AIR_TIME** *value based on* **WHEELS_OFF** *and* **WHEELS_ON**

U.S. Flight Delay and Cancellation Analysis

To further explore the issue of data redundancy, we decided to see whether we could reliably predict the value of a particular field based on the values of other existing fields. For this demonstration, consider these fields:

- WHEELS_OFF, the point at which the plane took off (wheels off the runway)
- WHEELS_ON, the point at which the plane landed (wheels on the runway)
- AIR_TIME, the duration of the plane's time in the air (between when it took off and when it landed)

Intuitively, the duration of the plane's time in the air should be the difference between the WHEELS_ON and WHEELS_OFF values. As mentioned earlier, however, these are decimal representations of clock time, so we first need to devise a function to handle this arithmetic:

```
#define function to find difference, in minutes
def subtract_float_time(timeStart, timeEnd):
    hoursPart_Start = int(timeStart) // 100
    hoursPart_End = int(timeEnd) // 100
    minutesPart_Start = int(timeStart) % 100
    minutesPart_End = int(timeEnd) % 100
    hours_diff = (hoursPart_End - hoursPart_Start) % 24
    timeDifferenceInMinutes = float(60*hours_diff + (minutesPart_End - minutesPart_Start))
    return timeDifferenceInMinutes
```

Given two float values representing time points, the subtract_float_time function returns the amount of time between the points, in minutes. Since this function is to be applied to the columns of a dataframe, we used Spark's user-defined function (udf) capability:

```
# create udf to be applied to 2 dataframe columns
udf_subtract_float_time = udf(subtract_float_time, returnType = FloatType())
```

Then, we used a query to generate a view of a few records containing the 3 fields of interest along with our calculated field:

```
# compare time subtraction to AIR_TIME
dfFlights.filter(col("WHEELS_ON").isNull() == False).\
    select('ORIGIN','DEST','WHEELS_OFF','WHEELS_ON',\
           udf_subtract_float_time(col("WHEELS_OFF"), col("WHEELS_ON")).alias("ON_minus_OFF_inMinutes"), \
          'AIR_TIME').\
    toPandas().head(10)
```

]:

	ORIGIN	DEST	WHEELS_OFF	WHEELS_ON	ON_minus_OFF_inMinutes	AIR_TIME
0	EWR	DEN	1527.0	1712.0	105.0	225.0
1	LAS	SFO	1118.0	1223.0	65.0	65.0
2	SNA	DEN	1345.0	1631.0	166.0	106.0
3	RSW	ORD	1611.0	1748.0	97.0	157.0
4	ORD	ALB	703.0	926.0	143.0	83.0
5	ORD	OMA	2259.0	1.0	62.0	62.0
6	IAH	LAS	801.0	854.0	53.0	173.0
7	DEN	CID	1329.0	1554.0	145.0	85.0
8	SMF	EWR	2247.0	627.0	460.0	280.0
9	RIC	DEN	1611.0	1748.0	97.0	217.0

We now examine some sample records, performing a few manual calculations to see whether our function returned the expected result:

- Consider the row in the above view with index 1. The **WHEELS_OFF** and **WHEELS_ON** fields indicate that the flight took off at 11:18am and landed at 12:23pm. It was in the air for 65 minutes (1 hour and 5 minutes), and indeed our custom **ON_minus_OFF_inMinutes** field has a value 65.0, matching the **AIR_TIME**.
- Now consider the row with index 0. The difference between the indicated **WHEELS_OFF** and **WHEELS_ON** is 1 hour and 45 minutes, and our custom field indeed has the value 105.0. However, the **AIR_TIME** is 225.0. This record represents a flight that departed from Newark (EWR) and arrived at Denver (DEN), "going back" 2 time zones. Thus, comparing the time points is not sufficient; we have to adjust the difference by adding 2 * 60 minutes = 120 minutes to achieve the true **AIR_TIME** of 225.0.

(This demonstration and other parts of our preliminary analysis can be found in the notebook file named "scheduled flights sample (preliminary data analysis).ipynb".)

This demonstration reveals some of the considerations behind the convention choices we've discussed in this preliminary analysis. Our attempt to predict the true duration of time spent in the air using existing fields, which was already complicated by the use of float data type to represent clock time, was unsuccessful because it did not account for time zones. We could

have added logic to our function to handle this, but it is much more practical to simply use the provided fields (**AIR_TIME** in this example). Even if some fields could be derived easily from others without these kinds of errors, it is still helpful to have a comprehensive set of fields available so that a data scientist can choose the ones that are most directly applicable to her analysis, depending on the purpose of the investigation. That is, even in a set of fields where some can be derived from others, no one field is necessarily more important, central, or integral than others; they should be held in equal standing, so it is appropriate to include all of them, at the expense of memory/storage space.

This might explain the choice to record time points in the local time zone, which we described earlier as unexpected. Using a single standardized time zone would perhaps make calculations and comparisons simpler, but the inclusion of a comprehensive set of fields in the data set obviates such calculations. Instead, time can be represented according to the local time zone, which is more natural and consistent with how people keep time.

Our final consideration in this preliminary analysis is the handling of missing data. Given the rigor and standardization we've discussed, it is not surprising that this was not a major problem. The fields that are necessary to even describe a scheduled flight (e.g. date, carrier, origin, and destination) are present for virtually all records. The major reason for "missing" data relates to *canceled* flights. Each record represents a scheduled flight with scheduled times, but if a flight was canceled, then the fields pertaining to *actual* times are blank or NaN, since the flight did not take place. In order to analyze delayed flights, one can simply exclude canceled flights from consideration by querying on the CANCELLED field. However, to do so is to disregard potentially-significant data. If we are evaluating the reliability of an airline carrier or airport according to whether it fulfills its scheduled flights on time, then it seems inappropriate to

disqualify canceled flights, since a cancellation is arguably more disruptive than a delay. Our audience should observe this caveat when interpreting the results of flight delay analysis.

In summary, our preliminary analysis allowed us to understand the advantages and limitations of the data available. We encountered some surprising and unintuitive conventions regarding the representation of time, but after further exploration, we can infer some reasons for these choices. While there is some redundancy in the data set, its comprehensiveness and standardization make it appropriate for an exploratory data analysis.

U.S. Flight Delay and Cancellation Analysis

Exploratory Data Analysis

#	Name	Description	Notes
1	Year	1987-2008	
2	Month	1-12	
3	DayofMonth	1-31	
4	DayOfWeek	1 (Monday) - 7 (Sunday)	
5	DepTime	actual departure time (local, hhmm)	
6	CRSDepTime	scheduled departure time (local, hhmm)	
7	ArrTime	actual arrival time (local, hhmm)	
8	CRSArrTime	scheduled arrival time (local, hhmm)	
9	UniqueCarrier	unique carrier code	
10	FlightNum	flight number	
11	TailNum	plane tail number	
12	ActualElapsedTime	in minutes	
13	CRSElapsedTime	in minutes	
14	AirTime	in minutes	
15	ArrDelay	arrival delay, in minutes	A flight is counted as "on time" if it operated less than 15 minutes later the scheduled time shown in the carriers' Computerized Reservations Systems (CRS).
16	DepDelay	departure delay, in minutes	
17	Origin	origin IATA airport code	
18	Dest	destination IATA airport code	
19	Distance	in miles	
20	TaxiIn	taxi in time, in minutes	
21	TaxiOut	taxi out time in minutes	
22	Cancelled	was the flight cancelled?	
23	CancellationCode	reason for cancellation (A = carrier, B = weather, C = NAS, D = security)	
24	Diverted	1 = yes, 0 = no	
25	CarrierDelay	in minutes	Carrier delay is within the control of the air carrier. Examples of occurrences that may determine carrier delay are: aircraft cleaning, aircraft damage, awaiting the arrival of connecting passengers or crew, baggage, bird strike, cargo loading, catering, computer, outage-carrier equipment, crew legality (pilot or attendant rest), damage by hazardous goods, engineering inspection, fueling, handling disabled passengers, late crew, lavatory servicing, maintenance, oversales, potable water servicing, removal of unruly passenger, slow boarding or seating, stowing carry-on baggage, weight and balance delays.
26	WeatherDelay	in minutes	Weather delay is caused by extreme or hazardous weather conditions that are forecasted or manifest themselves on point of departure, enroute, or on point of arrival.
27	NASDelay	in minutes	Delay that is within the control of the National Airspace System (NAS) may include: non-extreme weather conditions, airport operations, heavy traffic volume, air traffic control, etc.
28	SecurityDelay	in minutes	Security delay is caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.
29	LateAircraftDelay	in minutes	Arrival delay at an airport due to the late arrival of the same aircraft at a previous airport. The ripple effect of an earlier delay at downstream airports is referred to as delay propagation.

The following columns had missing values which were handled:

- DepTime, ArrTime

U.S. Flight Delay and Cancellation Analysis

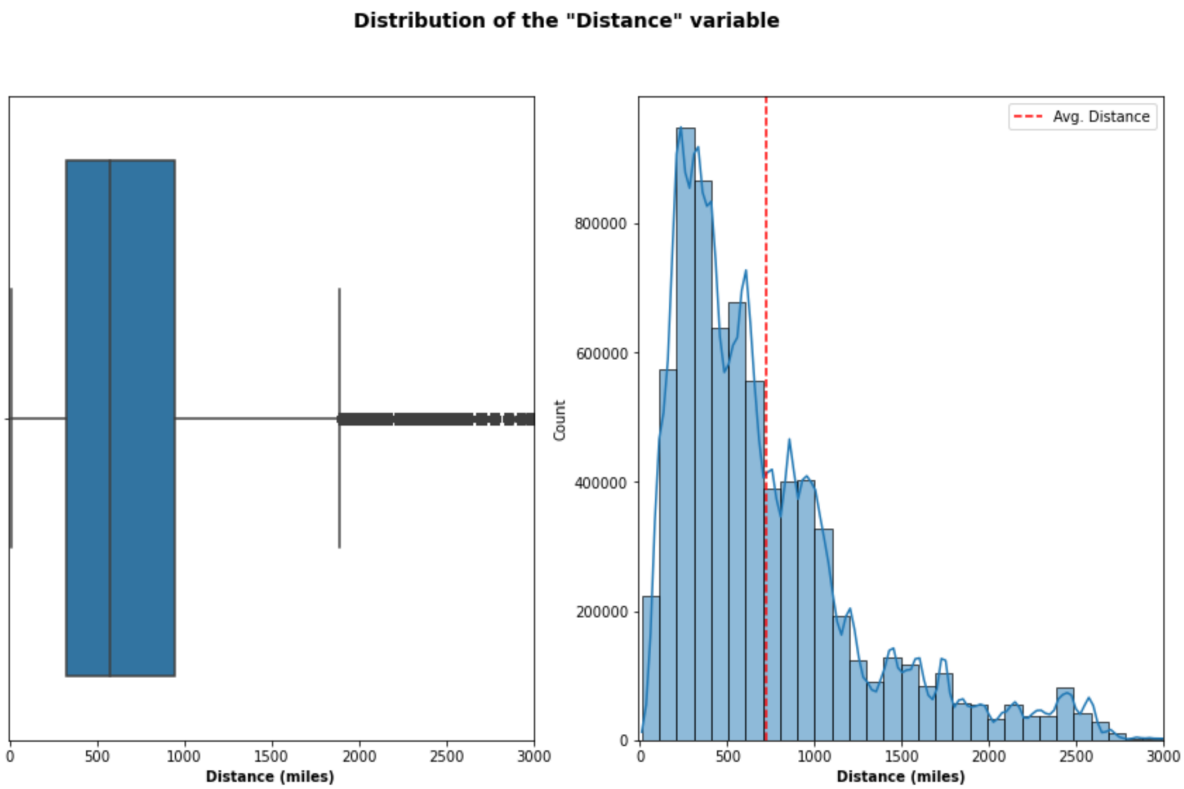
- ActualElapsedTime, CRSElapsedTime
- AirTime, ArrDelay, DepDelay
- TaxiIn, TaxiOut
- CarrierDelay, WeatherDelay, NASDelay, SecurityDelay and LateAircraftDelay

For accuracy, we used interpolation for the columns regarding Dep. and Arr. time on only the delayed flights

Some of the aspects of data we were interested in finding out about:

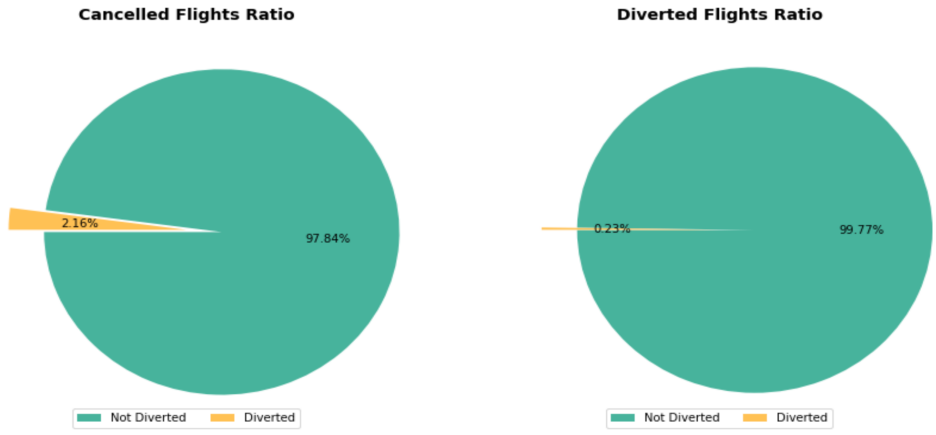
- What are the factors affecting the flight's cancellation?
- When do airline cancellations happen?
- Are cancellations related to delays?
- Are there any other interesting patterns?

U.S. Flight Delay and Cancellation Analysis



During the three years it has been observed that:-

- We have a right-skewed distribution, with a mean value around 750 miles.
- More than 75% of the flights that haven't been canceled were less than 1000 miles in distance.

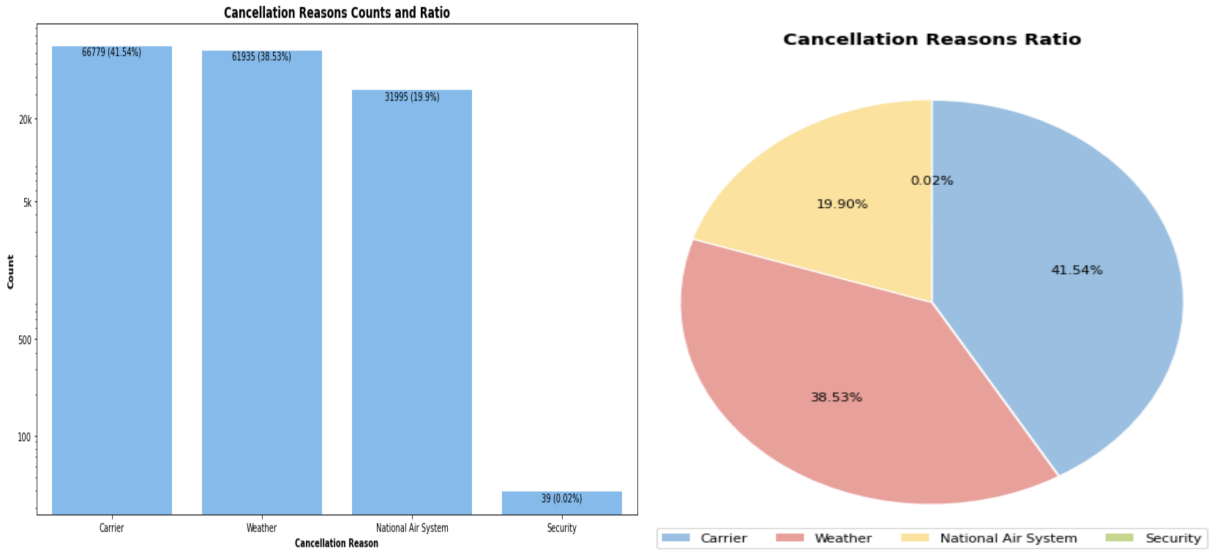


The pie charts generated for all the three years (1994,2006,2007) show that:-

U.S. Flight Delay and Cancellation Analysis

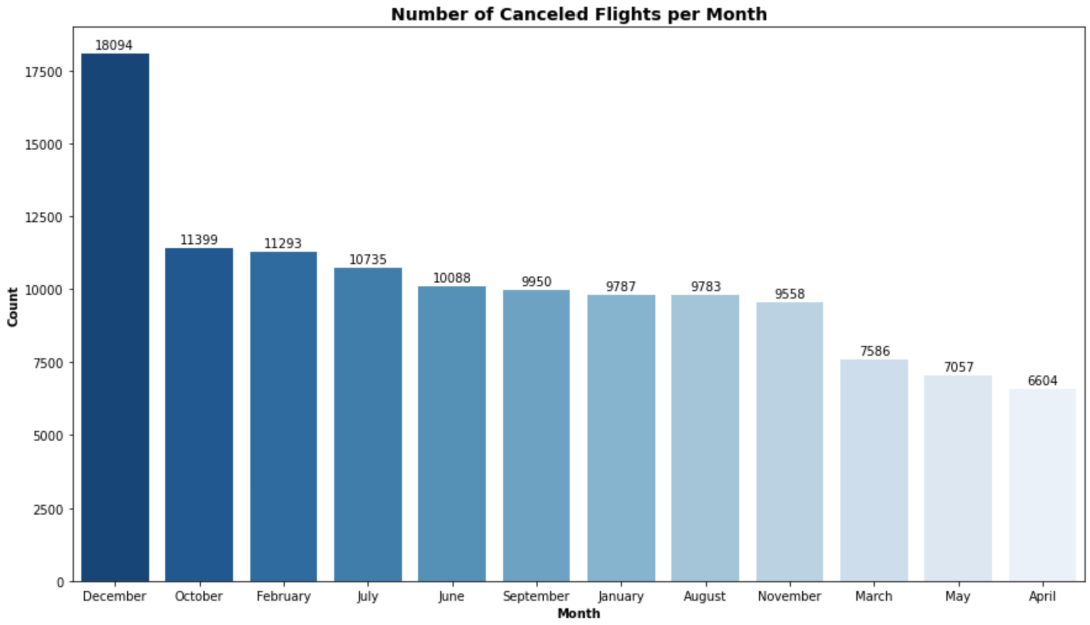
- Around 1.29% to 2.16% of the flights *(for the above mentioned years)* have been canceled.
- Only 0.23% of the flights were diverted.

Exploring the common reasons for cancellation (2006/2007):



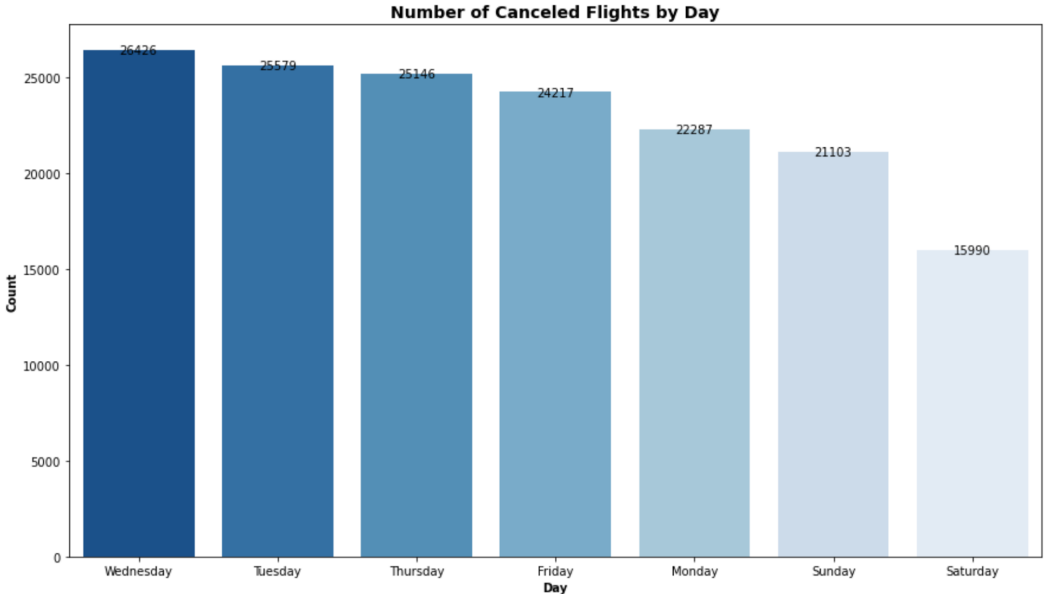
From the above charts it was apparent that most flight cancellations were due to carrier and weather reasons with around 31-39% and 41-46% each. Followed by National Air System with ranging around 20-23%. And the security reasons had the least effect.

U.S. Flight Delay and Cancellation Analysis



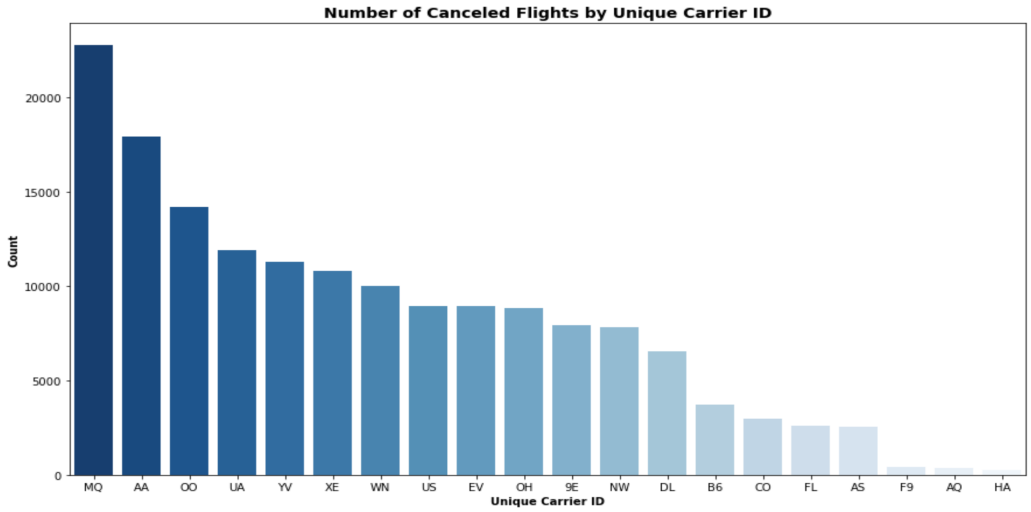
For our observed time duration(1994/2006/2007),from the above barchart,we were able to infer that most of the canceled flights were around the months of December to February,probably due to the bad weather because it is winter during these months of the year.

Analyzing the relation between day of the week and canceled flights (1994/2006/2007)

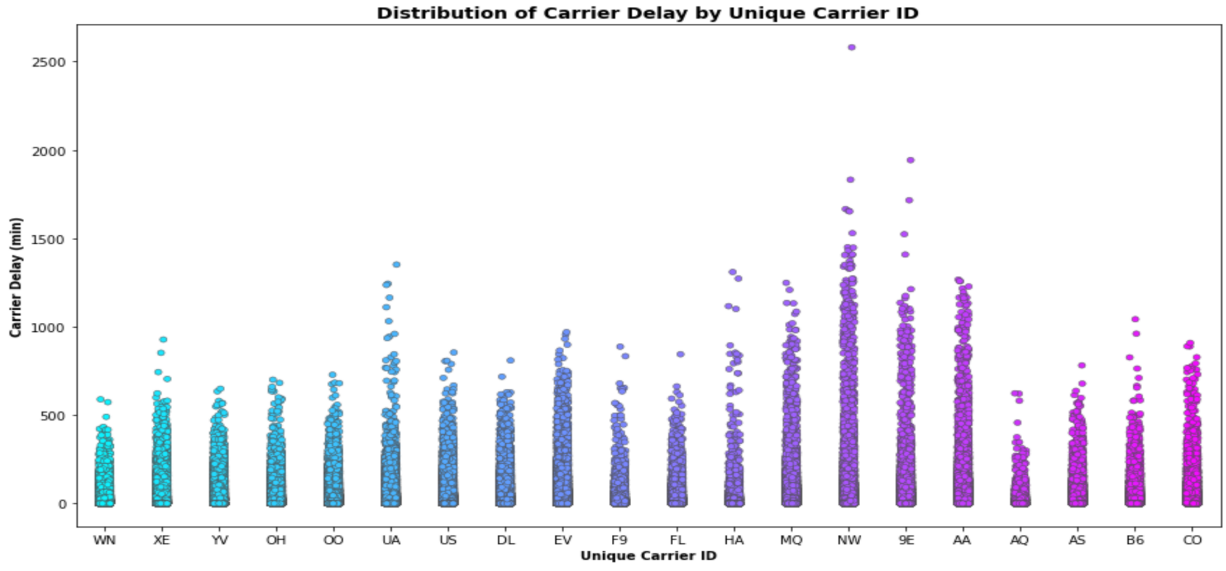


U.S. Flight Delay and Cancellation Analysis

From the data gathered from the charts it's seen that most flights would be canceled during weekdays (Tuesday to Friday) and it's much better to travel during the weekends.

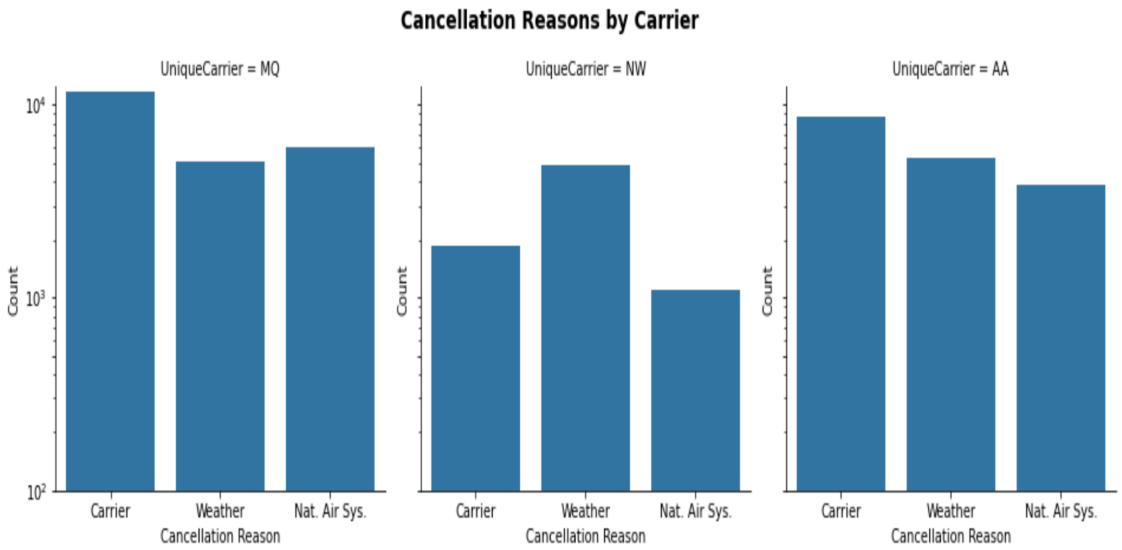


During 2006 and 2007, it would seem Envoy Airlines (MQ) and for the year 1994 it is US Airways for which we can give the award to most canceled flights of more than around 19000 flights canceled over the span of the entire year.

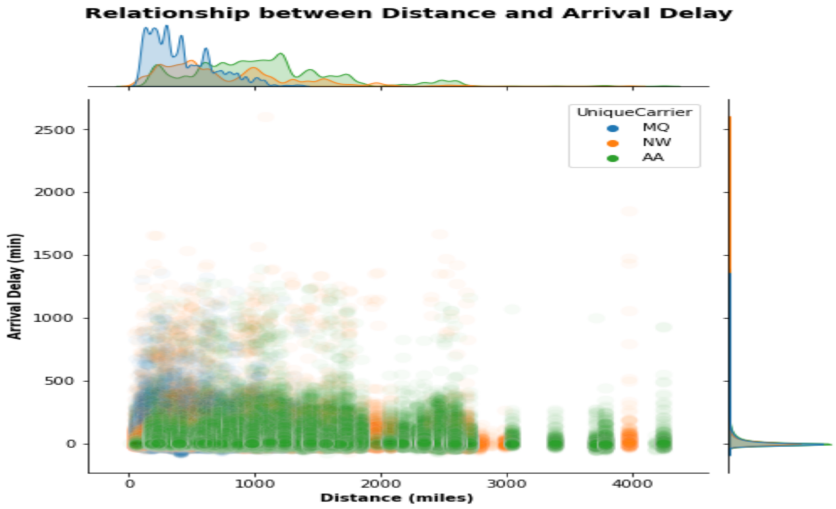


U.S. Flight Delay and Cancellation Analysis

As seen from the above distribution,for the duration of 2006-2007,The Northwest Airlines Inc. (NW) had the most delayed flights due to carrier delay, followed by Endeavor Air (9E) ,American Eagle Airlines Inc. (MQ) and American Airlines Inc. (AA).



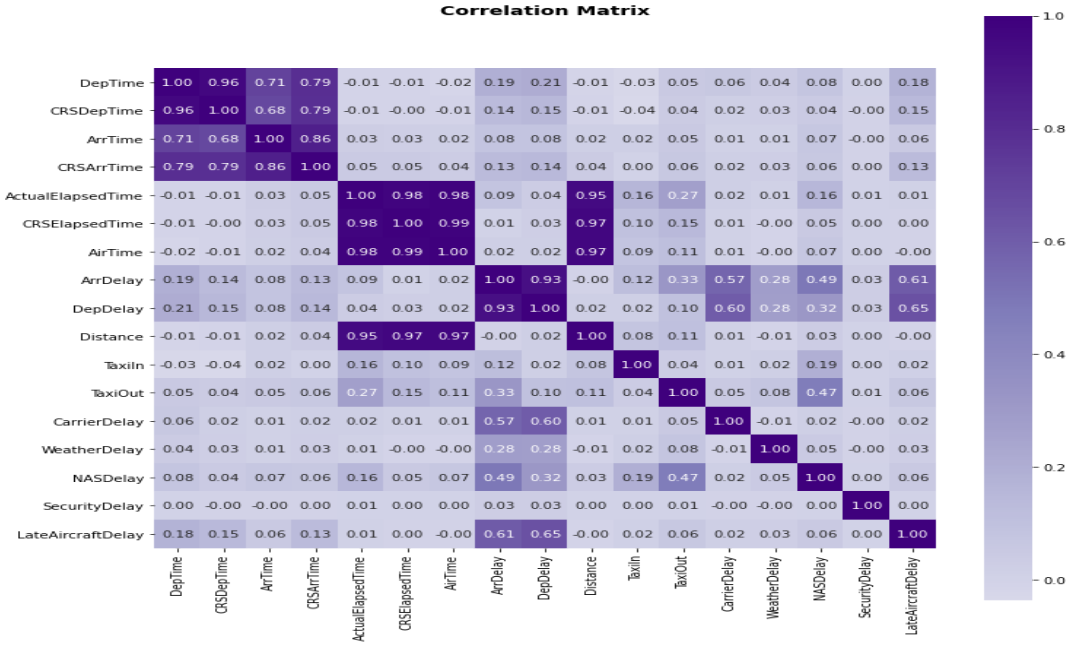
As observed among all the carriers, American Airlines Inc. and Northwest Airlines Inc. had most canceled flights which occurred due to carrier and weather reasons,while on the other hand, American Eagle Airlines Inc. 's had most of its canceled flights due to weather reasons.



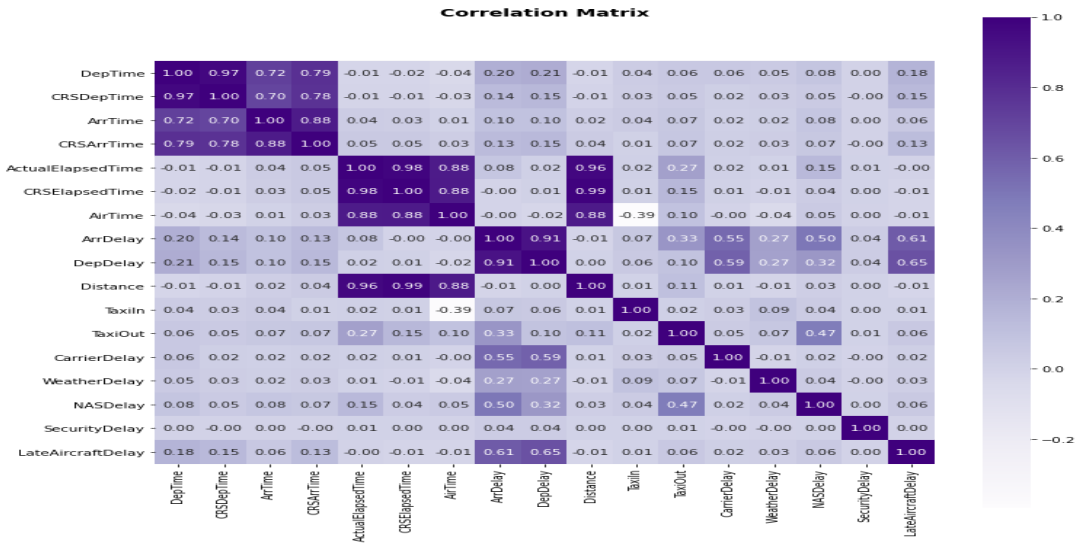
U.S. Flight Delay and Cancellation Analysis

For the years 2006 and 2007,the observations matches with what we have found before, but for American Eagle Airlines Inc.'s shortest flights arrived more later than the longest ones than the other carriers

For Year 2007:



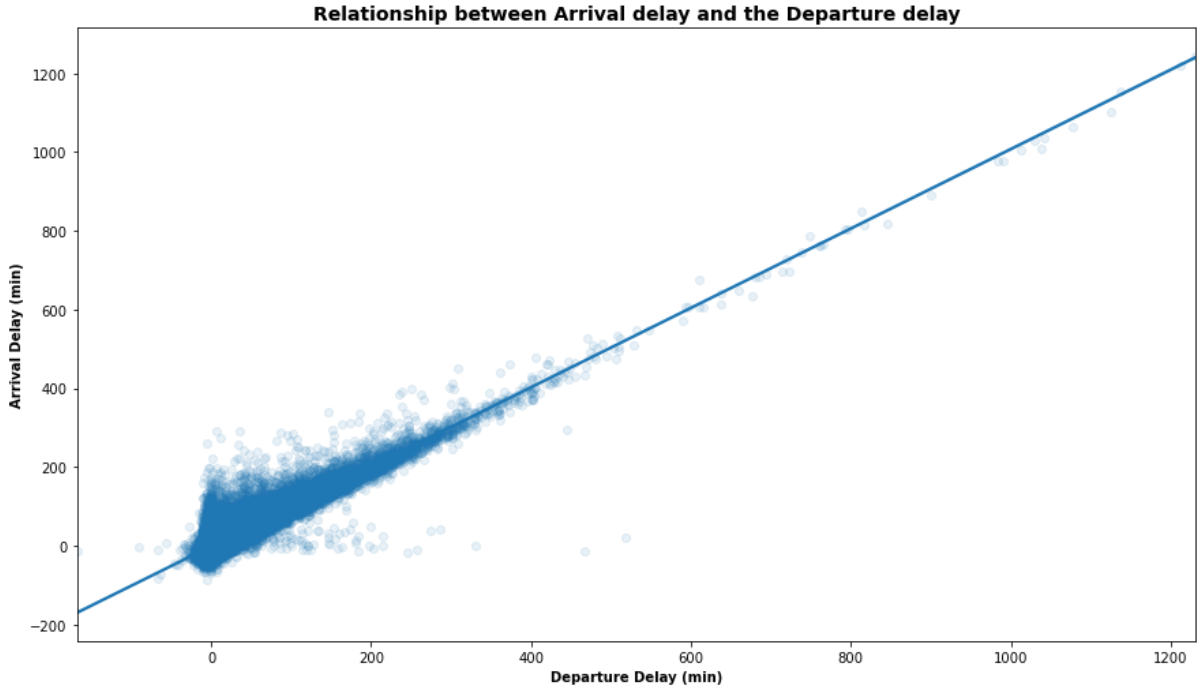
For 2006:



U.S. Flight Delay and Cancellation Analysis

The correlation matrices reveal some significant relationships among the variables that are consistent with our intuitions. For example, Distance was highly correlated with ActualElapsedTime, CRSElapsedTime, and AirTime (0.95 to 0.99). ArrDelay and DepDelay were highly correlated (0.91-0.93), which makes sense because a flight which arrived late most likely left late and hence would be considered delayed on departure. Low correlation values (close to 0) are observed between variables that we would not expect to be related; for example, Distance (between origin and destination airports) and TaxiIn (the amount of time elapsed between the plane's landing and its arrival at the destination gate) have a low correlation.
We further explored some of these relationships:

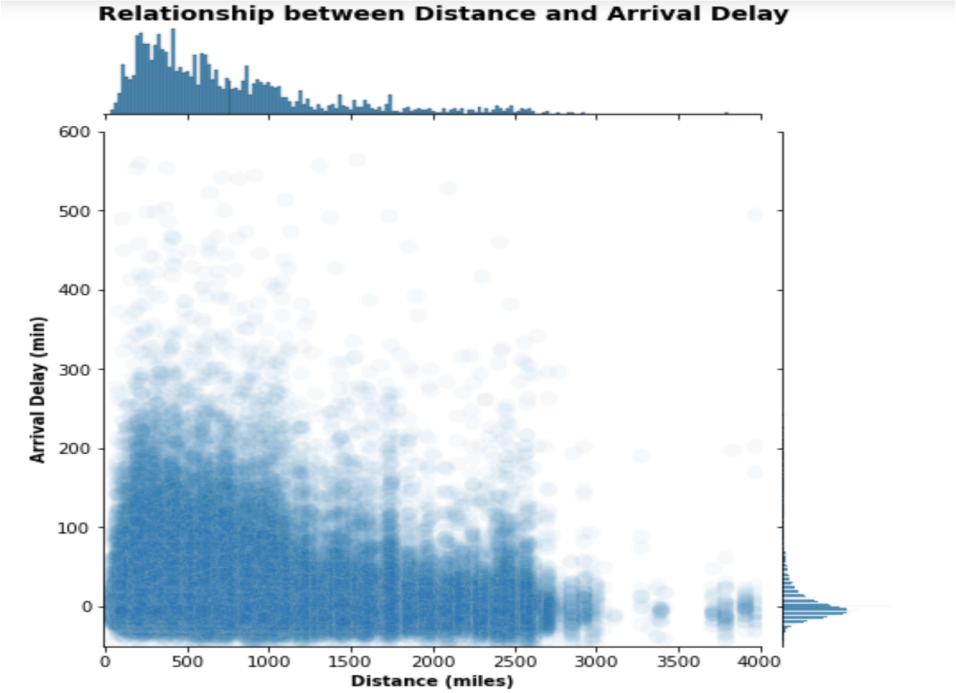
Relationship between arrival delay and departure delay for the year 2007 (similar results were found for other years):



As expected, a positive correlation. If the flight took off late, it would likely arrive late.

U.S. Flight Delay and Cancellation Analysis

Exploring the relationship between arrival delay and distance



From the above chart and using the correlation matrix (r = - 0.001 for year 2006), it turns out there is a negative relationship between distance and arrival delay as the shorter flights are the most delayed flights on arrival. This was observed for all the three years we used.

Finding the best time to book a flight

First, we need to convert some columns:

```
def convert2Time(var):
    """ Convert records to proper time format """
    try:
        if var != 0 or var != np.NAN:
            if str(int(var))[:-2].zfill(2) == '24':
                hh = '00'
            else:
                hh = str(int(var))[:-2].zfill(2)

            mm = str(int(var))[-2:].zfill(2)
            time = f"{hh}:{mm}"
        return time
    except Exception as e:
        pass
```

U.S. Flight Delay and Cancellation Analysis

```
# Extract/Convert time
airlines_copy['DepTime'] = airlines_copy['DepTime'].apply(convert2Time)
airlines_copy['CRSDepTime'] = airlines_copy['CRSDepTime'].apply(convert2Time)
airlines_copy['ArrTime'] = airlines_copy['ArrTime'].apply(convert2Time)
airlines_copy['CRSArrTime'] = airlines_copy['CRSArrTime'].apply(convert2Time)

airlines_copy[['DepTime', 'CRSDepTime', 'ArrTime', 'CRSArrTime']].head()
```

	DepTime	CRSDepTime	ArrTime	CRSArrTime
0	12:32	12:25	13:41	13:40
1	19:18	19:05	20:43	20:35
2	22:06	21:30	23:34	23:00
3	12:30	12:00	13:56	13:30
4	08:31	08:30	09:57	10:00

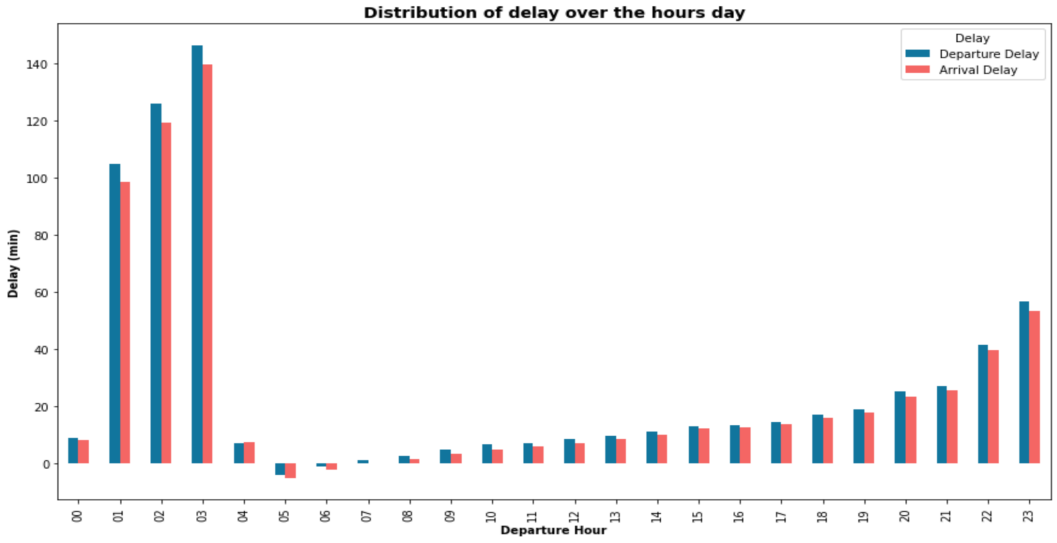
```
# Extract the hour
airlines_copy['DepHour'] = airlines_copy['DepTime'].apply(lambda x: x[:2])
airlines_copy['ArrHour'] = airlines_copy['ArrTime'].apply(lambda x: x[:2])

airlines_copy[['DepTime', 'CRSDepTime', 'ArrTime', 'CRSArrTime', 'DepHour', 'ArrHour']].head()
```

	DepTime	CRSDepTime	ArrTime	CRSArrTime	DepHour	ArrHour
0	12:32	12:25	13:41	13:40	12	13
1	19:18	19:05	20:43	20:35	19	20

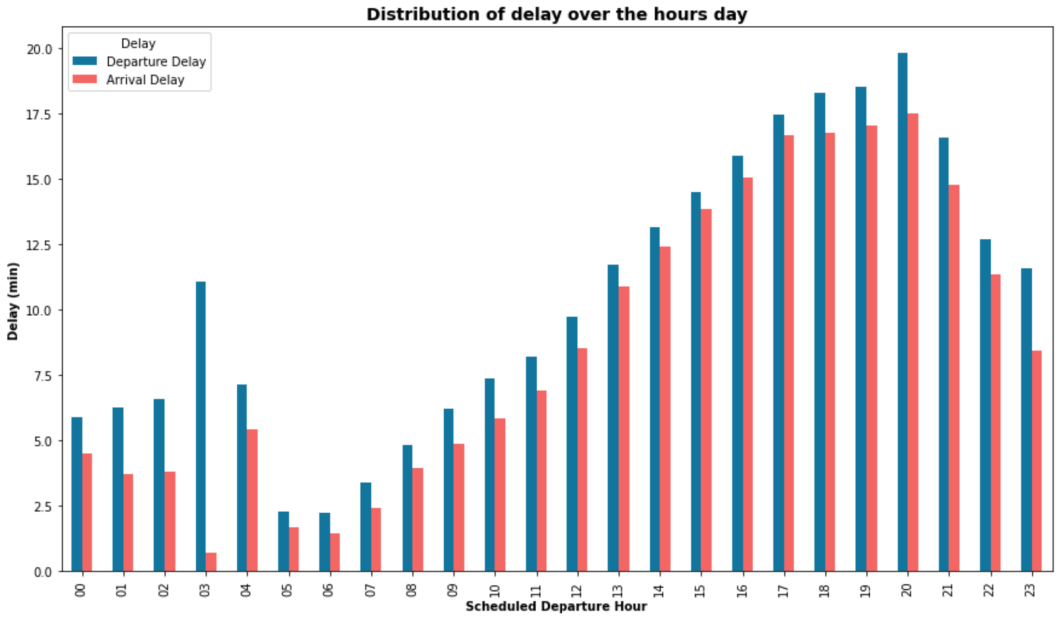
Plotting the distribution of delay over the hour of the day compared to:

1. Departure hour



U.S. Flight Delay and Cancellation Analysis

2. Scheduled departure hour



From the above, it appears the early hours of the day *(05:00 AM to 8:00 AM)* are the best hours to book a flight as they have the least delays.

Based on the previous findings, and after investigating and plotting the correlation matrix between the variables, we deduced that:

- Most of the flights were not diverted (only 0.24%).
- The most canceled flights were due to the carrier and weather reasons.
- Northwest Airlines Inc. followed by American Airlines Inc. has the most delayed flights and American Airlines Inc. has the most canceled ones.
- We have found a negative correlation between the distance and the arrival delay, in which we observed that the shortest flights in distance are the most delayed on arrival.

U.S. Flight Delay and Cancellation Analysis

- And if you want our advice on when to book your flight? from the above charts it turns out the early hours of the morning have the least delays over the hours of the day.

Notebooks used for EDA:-

Analyzing Year 1994.ipynb

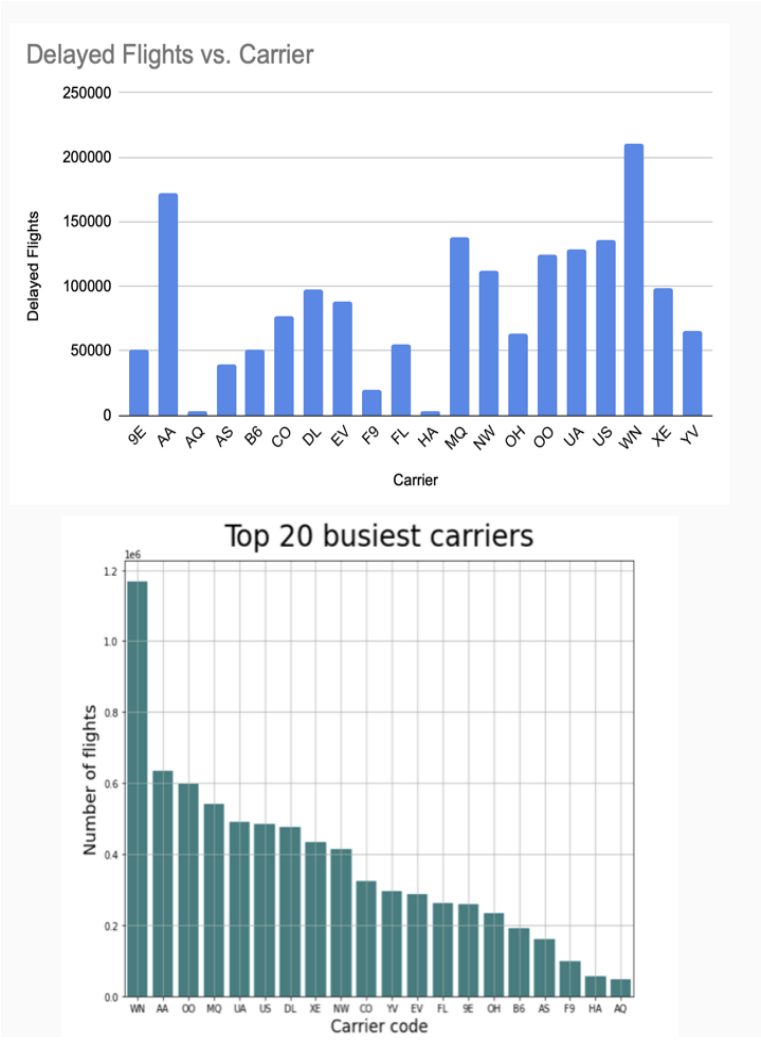
Analyzing Year 2006.ipynb

Analyzing Year 2007.ipynb

Analysis: 1994/2001/2007:

For the Amstat data set we wanted to view progression over time so we picked specific years to analyze. An issue we had was that some of the years were missing information, so a full view of the data was not possible. We chose to visualize it at set intervals. For 1994/2001/2007 we attempted to determine which airports and carriers were busiest and how many delayed flights there were per carrier.

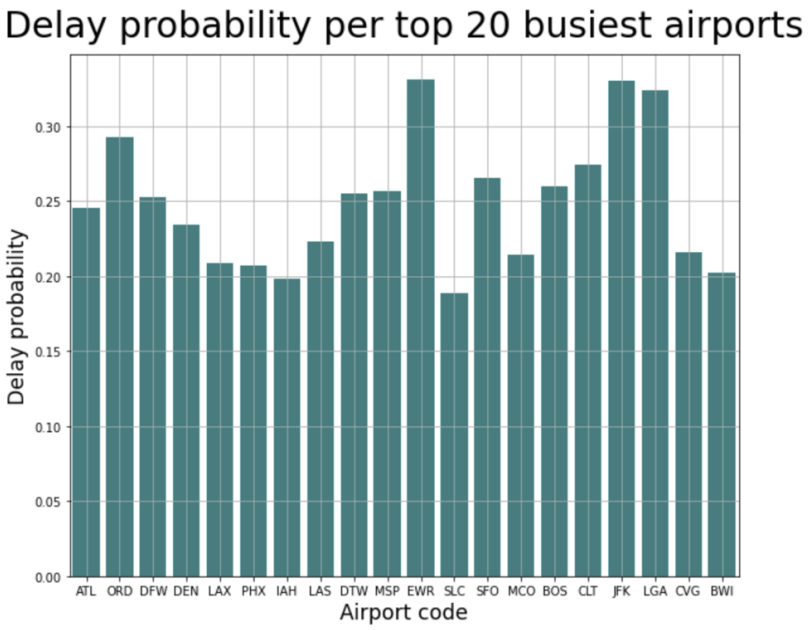
U.S. Flight Delay and Cancellation Analysis



2007

These charts are for 2007. From this we can see that WN was the busiest carrier by far but also had the highest number of delayed flights. Close to 10% same with AA but they had a much higher percentage of canceled flights. These charts were tracked for 3 separate years 1994, 2001, and 2007. Using this information we can easily see which airline runs an efficient business and which ones to avoid.

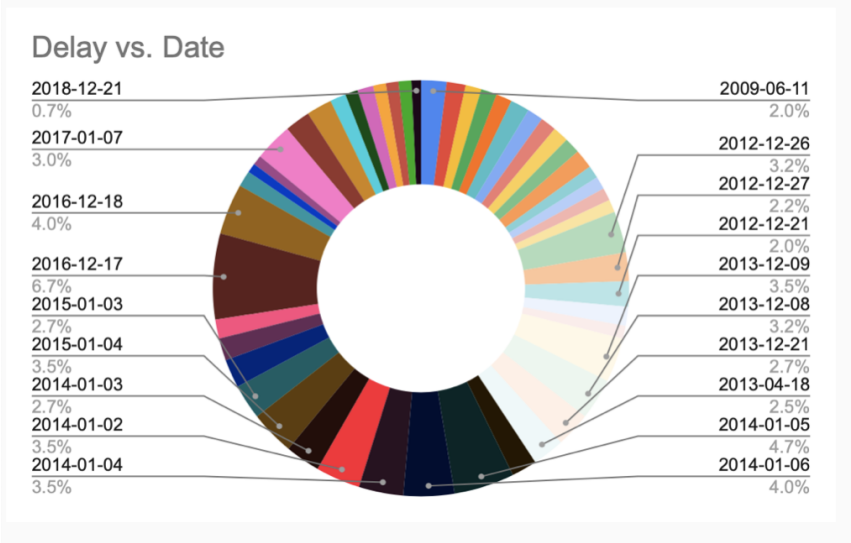
U.S. Flight Delay and Cancellation Analysis



2007

We can see that all airports have a baseline delay probability. This would typically be based on mechanical issues, personnel issues or climate challenges. But certain airports tend to be outliers with much higher probabilities. In this case we learned that EWR, SFO, JFK LGA and a few others have gotten worse as the years progressed in our data set. Seattle airport lost its top 20 busiest listing so was not seen here. CLT(Charlotte) actually decreased its probability significantly so it was one of the few airports that improved. I would want to analyze why that occurred, possibly due to loss of its local population or businesses moving out of that region.

U.S. Flight Delay and Cancellation Analysis



2008-2018-Kaggle

For the Kaggle data set we wanted to calculate another metric which was the most unreliable day to fly. This metric was only calculated for the last 10 years as that is much more relevant than historical data. We learned that the Holiday and Winter months had a much higher frequency of having days with the most delayed flights. Around Christmas and during January/February. The days where travel was bad also decreased as the years progressed through this dataset.

2008-2014 had more occurrences than 2014-2018. With respect to Christmas we concluded that since so many flights are occurring during this period any weather systems would cause a large impact on the flight schedule.

Notebooks Used for Analysis:-

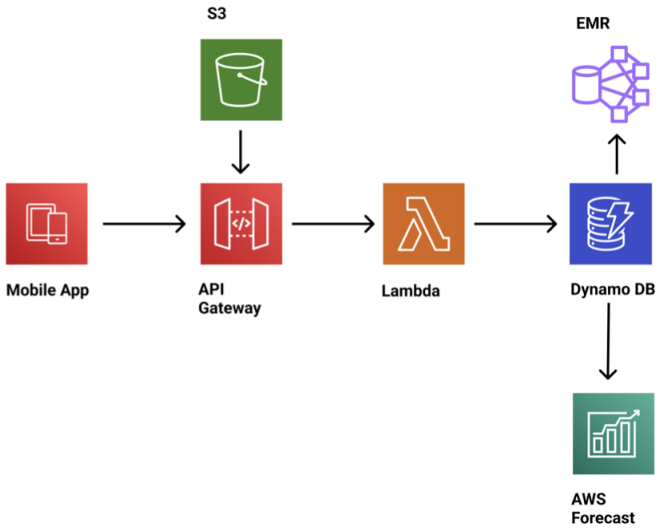
BigDataProjectFinal-1994.ipynb

BigDataProjectFinal-2001.ipynb

BigDataProjectFinal-2007.ipynb

most-unreliable-day-to-fly.ipynb

Further Big Data Development



For continued analysis we designed a schematic of a mobile application deployed on the AWS infrastructure. With this deployed, we can run scripts continuously to analyze real time data provided by flying customers nationwide. First we would develop a mobile app which would be hosted on a S3 bucket. This application would connect using the API Gateway to the DynamoDB. The Lambda scripts would push data from the mobile application to the noSQL database. Data being supplied by the mobile users would not be relational because not all categories would be input due to lack of information or users who do not want to input all data. Because of this a noSQL database was chosen. The AWS EMR product allows us to run Spark on the DynamoDB instance. This would allow us to produce analysis that we produced so far using historical data. This constant data stream will provide us a longer understanding of air travel tendencies. AWS Forecast can be implemented on this data to provide prediction services from our historical data and the data stream. This will allow customers to choose their flight dates and plans their schedules accordingly. From a business perspective we can sell this data to other Airlines so that they can predict travel and plan their future supply needs accordingly. The

first way to improve this would be to implement a data stream using a service like Kafka or AWS Kinesis. This push approach would allow us to analyze our data in real time. The second way to improve this would be to distribute our application using the various availability zones that AWS provides. This way the application and its database will be available to all users nationwide at a closer proximity reducing data transaction times.

Conclusion

Overall we would like to conclude that while travel delays have changed through the past few decades the one factor that has been consistent is the delay associated with weather. It has also been observed that it's much better to travel during weekends in order to prevent being affected by flight cancellations. Be an early bird! Book your flights which are scheduled for departure early in the morning,as they are the least susceptible for delays. Most of the flights observed were primarily canceled due to weather and carrier reasons. Travel during the holiday months has been notorious and unpredictable. Planning ahead and having travel insurance would be a great way to mitigate risks. For summer travel besides the price hikes travelers have nothing to worry about since inclement weather during the summer doesn't affect air travel nearly as much as it does during the winter months.

Resources

https://www.faa.gov/data_research

https://www.kaggle.com/datasets/flashgordon/usa-airport-dataset

https://github.com/varunrao10/Domestic_Airport_Airline_Delay_Pred

https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-2018

https://www.kaggle.com/code/antgoldbloom/most-unreliable-day-to-fly