Airline Data Analysis

Manish Kolla
Data Mining (CS 4740)
Department of Computer Science
Atlanta, GA, United States
maneeshkolla@gmail.com

***Abstract*— This study discusses in detail the airports which suffer with frequent departure and arrival delays, during what days of the week and month the delays are maximum, and the possible reasons that are causing these delays. The study was conducted by analyzing a dataset of over 600,000 flight records from 2008. The dataset included information on the departure and arrival times of flights, the airports involved, the airlines operating the flights, and the reasons for any delays.**

The study found that Thursdays and Sundays of the week had the highest rates of departure and arrival delays.

The study identified the following as the possible reasons for the delays: Carrier, NAS, Weather, Aircraft Delay. The study also found that data visualization and grouping were essential to the analysis. Data visualization helped to identify patterns in the data that would not have been apparent otherwise. Grouping helped to shorten the data set and make it easier to analyze.

I. Introduction

The reason I have chosen airline data is, during the end of my CS 1302 class, my professor gave us this airline data set and instructed us to plot graphs and understand the data after highlighting the importance of airline data analysis. During the process of this analysis, I found this data set very interesting and insightful. Being in a introductory class like CS 1302, I couldn't do much of data analysis, but now after taking the Data Mining (CS 4740) with Professor Jingyu Liu, I feel more confident that I can do some analysis with this dataset. This is my first data science class I have taken so far, and throughout the course of this semester, I have learned a lot of data mining algorithms and methods. Diving into the airline data analysis, it is crucial for the aviation industry as it helps airlines to make informed decisions about their operations, customers, and market strategies. For the data I have chosen it's primarily being analyzed for information about the flight delays and the possible factors that might affect it alongside analyzing what days of a month or week the delays occur frequently. These are my final derivatives of this project.

Code Link:
https://colab.research.google.com/drive/1IXT_DuTOInwduw-053naFvYXQ5gR3-8z?usp=sharing

I. Material and Methods

Throughout this project, I have followed the four steps which are.

1. Understanding the data
2. Cleaning the data
3. Data Mining
4. Analyzing the conclusions

During the entire course of this project, the active python libraries used are:

1. Pandas- In managing and analyzing the dataset.
2. bz2- Conversion of bz2 compression to csv file
3. Matplotlib- Plotting graphs
4. NumPy- Finding optimal number of clusters.
5. Seaborn- Plotting statistical graphs and heatmaps.
6. Sklearn- Clustering method and silhouette score

1. Understanding the data

I retrieved this dataset names 2008.csv.bz2 from Harvard Dataverse (The Data Exposition: 2009). Initially the file was compressed using the bz2 compression for space compatibility. Using pandas library, I have uncompressed to csv file.

Characteristics of this data:

Data size: 624059(rows) x 30(columns)

The 30 attributes include: Year, Month, Day of week/Month, Actual Departure and Arrival, CRS Departure and Arrival, Departure and Arrival delay, Airtime, distance, Unique carrier, flight number, tail number, Origin, Destination, Taxi In, Taxi out, Cancellation Codes, Diversion, Delays Caused by Weather, Security, Carrier, NAS, aircraft.

Out of which I am mainly focusing on.

1. Month
2. Day of the week
3. Day of the Month
4. Departure Delays
5. Arrival Delays
6. Airtime
7. Origin
8. Destination
9. Delays Caused by Weather, Security, Carrier, NAS, aircraft.

In the entire data set, there is no unique parameter, all the attributes and features are repetitive making it harder for classifications. One interesting finding of this dataset is, for the feature Month only January and February were given due to which analyzing the delays based on months is not possible.

2. Cleaning the data

Using pandas library and replacing all the missing values with the mean value of that column, in order to prevent any data loss alongside removing all the null values. Initially there were null values in every attribute which caused a lot of outliers in the boxplot, but after removing the null values and replacing the missing values with its mean value of that column, it was able to get rid of most of its outliers. Since

the data set is very large, in order to make solid evidence, we have to remove the airports with minimal amount of data for better supporting evidences. After removing the airports with minimal amount of data we were able to narrow down the data to 428,000 rows with 67 destinations and 62 origins. The reason for not removing the outliers is, columns such as Departure Delay, Arrival Delay might have outliers and when removed we might lose data with maximum/minimum delays. In order to avoid valuable data loss, we are not removing the outliers.

3. Data Mining Approaches

In this section, I have done a deeper analysis of the delays and made conclusions based on airports with higher departure and arrival delays and the leading causes for the delays. This section can be broken into four parts.

1. Basic Statistical Analysis
2. Data Visualization
    a. Departure/Arrival Delays w.r.t individual airports
    b. Departure/Arrival Delays w.r.t air routes
    c. States with higher departure/arrival delays
    d. Factors affecting these delays.
3. Finding the relationship between airtime and departure/arrival delays using K means Clustering.
4. Maximum/minimum Delays with respect to Day of Week and Month

Basic Statistical Analysis

Before I did further deep analysis into data mining, I have calculated the correlation among all the variables and made some anecdotal conclusions which are:

1. Long hour flights are likely to be scheduled in the early mornings compared to at night.
2. As the distance increases the arrival delay decreases and departure delays increase
3. Weekends are the best time to fly to avoid any kind of arrival/departure delays.
4. Month ends tend to have less departure and arrival delays.
5. Flights scheduled in the beginning of the week are likely to get more cancelled compared to the ones over the weekend.

Based on the mean values of categorical data

1. Daniel K. Inouye International Airport- Hawaii (HNL), Jacksonville International Airport-Florida (JAX), Ronald Reagan Washington National Airport- Virgina (DCA) have mean departure and arrival delays less than 5 minutes.
2. O'Hare International Airport- Illinois (ORD) and San Francisco International Airport- California (SFO) have mean departure and arrival delays greater than 25 minutes.

During the process of using data visualization, I have encountered a lot of issues such as run time crashing, longer executions, and axis overlapping. In order to get rid of them, I have combined the destination and origination airports and made a new attribute for air route, this helps in analyzing which air routes have maximum number of delays and minimum number of delays. After concatenating the airports, we got it narrowed down to 2421 unique air routes.

Data Visualization and Results

I have taken several parts datasets by considering largest delays and least departure delays for better understanding about airports which have airlines departing early and late arrivals and also airports with early arrivals. This method helped in understanding the frequent routes which have early departures yet having late arrivals and possible causes for it. Without using the method, we cannot analyze airports which have early departures and arrivals and the air routes associated with them. Since the data set is very large, I have considered only the smallest and largest 1000 rows when analyzing for early departures and arrivals. Initially, I have plotted the graphs by taking the least and maximum number of departure delays. Which helped in faster executions and overcoming axis overlapping. Since arrival delays are directly caused by departure delays, departure delays require more analysis compared to arrival delays. Below are the conclusions based on airports with high and low departure delays and the specific reasons for those delays. Most of these plots are plotted using matplotlib. When the lowest 1000 rows with least departure delays have been plotted, according to my finding,

Airports with early departure of greater than 40 minutes (*Figure 1*)

1. Atlanta Airport- Atlanta (ATL)
2. Charlotte Douglas International Airport-North Carolina (CLT)
3. Dulles International Airport- Virgina (IAD)
4. Harry Reid International Airport- Nevada (LAS)
5. Harry Reid International Airport-Texas (SAT)
6. Los Angeles International Airport- California (LAX)

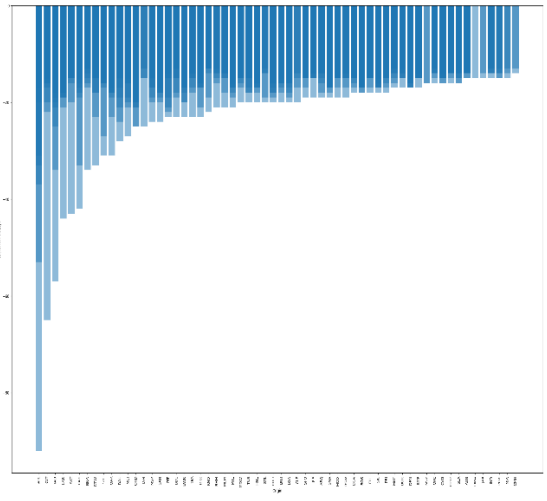

Figure 1

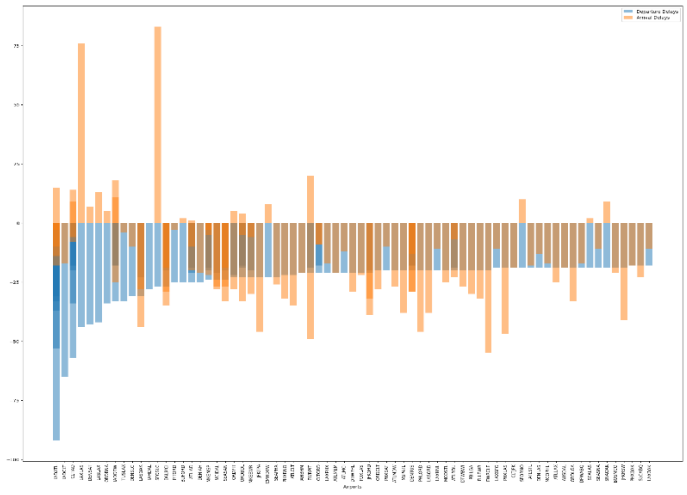

Figure 2

Air routes with frequent early departures are (*Figure 2*):

1. Daniel K. Inouye International Airport- Hawaii (HNL) to Los Angeles International Airport- California (LAX)
2. San Diego International Airport- California (SAN) to Seattle-Tacoma International Airport- Washington (SEA)
3. to Phoenix Sky Harbor International Airport- Arizona (PHX) to Los Angeles International Airport- California (LAX)
4. Oakland International Airport- California (OAK) to John Wayne Airport- California (SNA)
5. San José Mineta International Airport- California (SJC) to Portland International Airport- Oregon (PDX)
6. Dulles International Airport- Virgina (IAD) to Charlotte Douglas International Airport-North Carolina (CLT)
7. Harry Reid International Airport- Nevada (LAS) to Seattle-Tacoma International Airport- Washington (SEA)
8. John Wayne Airport- California (SNA) to Seattle-Tacoma International Airport- Washington (SEA)
9. Harry Reid International Airport- Nevada (LAS) to Portland International Airport- Oregon (PDX)
10. Harry Reid International Airport- Nevada (LAS) to Phoenix Sky Harbor International Airport- Arizona (PHX)

There are airports which depart flights early yet having high arrival delays, which are (*Figure 2*):

1. Los Angeles International Airport- California (LAX)
2. Dulles International Airport- Virgina (IAD)
3. San Francisco International Airport- California (SFO)
4. Austin-Bergstrom International Airport- Texas (AUS)
5. Boston Logan International Airport- Massachusetts (BOS)
6. Salt Lake City International Airport- Utah (SLC)

Air routes from

1. Harry Reid International Airport- Nevada (LAS) to Los Angeles International Airport- California (LAX)
2. Salt Lake City International Airport- Utah (SLC) to San Francisco International Airport- California (SFO)

Tend to be having early departures yet late arrivals. The common thing in these two air routes is the destination state which is California. California tend be having higher arrival delays.

The next set of graphs were plotted with considering largest departure delays from the top 1000. And the findings are:

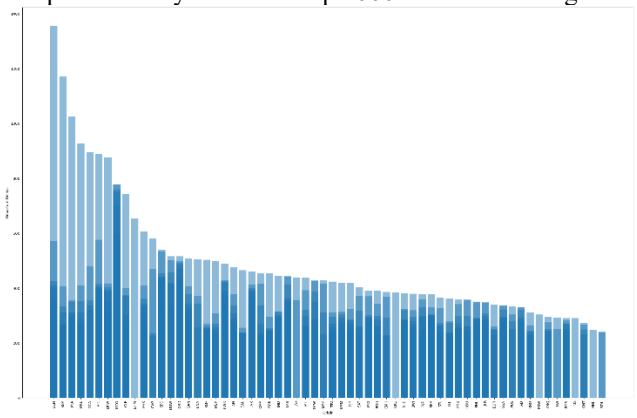

Figure 3

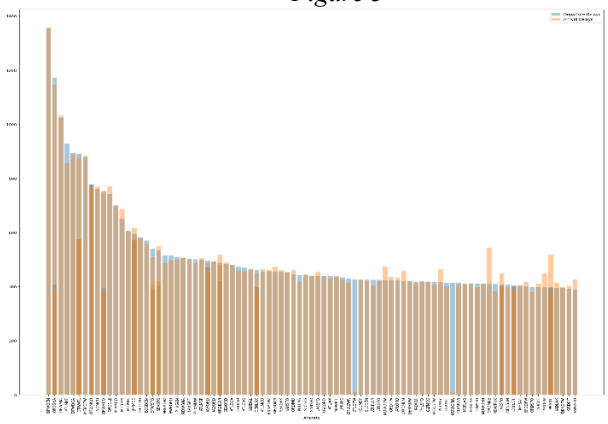

Figure 4

Airports with high departure delays (*Figure 3*):

1. Orlando International Airport- Florida (MCO)
2. Denver International Airport- Colorado (DEN)
3. San Francisco International Airport- California (SFO)
4. Chicago Midway International Airport- Illinois (MDW)
5. Atlanta Airport- Atlanta (ATL)
6. Dallas/Fort Worth International Airport- Texas (DFW)
7. O'Hare International Airport- Illinois (ORD)
8. John Wayne Airport- California (SNA)
9. Miami International Airport- Florida (MIA)
10. General Mitchell International Airport- Wisconsin (MKE)

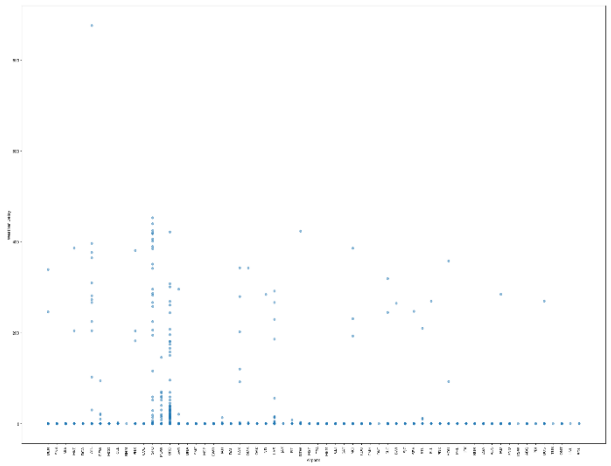

Figure 5

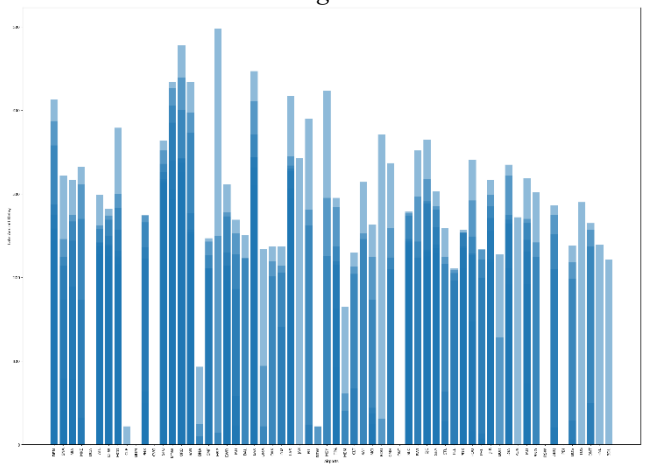

Figure 6

Similar airlines also tend to have higher arrival delays. According to these conclusions we can say that airports in Florida, California, and Illinois tend me having higher departure delays. Airlines especially flying from

1. Atlanta Airport- Atlanta (ATL)
2. San Francisco International Airport- California (SFO)
3. O'Hare International Airport- Illinois (ORD)
4. Chicago Midway International Airport- Illinois (MDW)

These following airports tend me having higher delays due to weather changes (*Figure 5*).

Airlines flying from these airports (*Figure 7*):

1. Orlando International Airport- Florida (MCO)
2. San Francisco International Airport- California (SFO)
3. Newark Liberty International Airport- New Jersey (EWR)
4. O'Hare International Airport- Illinois (ORD)

Have delays due to the National Aviation System (NAS). NAS Delay is caused mainly by air traffic at certain busy airports and sometimes due to non-extreme weather conditions for takeoff. The rest other airports mentioned above with higher departure delays are due to their carrier or late aircraft delays (*Figure 6*).

And air routes with frequent higher departure delays are (*Figure 4*):

1. San Francisco International Airport- California (SFO) to Los Angeles International Airport- California (LAX)
2. Los Angeles International Airport- California (LAX) to San Francisco International Airport- California (SFO)
3. San Francisco International Airport- California (SFO) to Harry Reid International Airport- Nevada (LAS)
4. San Francisco International Airport- California (SFO) to Seattle-Tacoma International Airport- Washington (SEA)
5. Harry Reid International Airport- Nevada (LAS) to San Francisco International Airport- California (SFO)
5. Newark Liberty International Airport- New Jersey (EWR) to O'Hare International Airport- Illinois (ORD)
6. O'Hare International Airport- Illinois (ORD) to LaGuardia Airport- New York (LGA)
7. O'Hare International Airport- Illinois (ORD) to Newark Liberty International Airport- New Jersey (EWR)
8. Dallas/Fort Worth International Airport- Texas (DFW) to O'Hare International Airport- Illinois (ORD)
9. San Francisco International Airport- California (SFO) to San Diego International Airport- California (SAN)

From the above air routes with maximum departure delays, we can conclude that flights flying from or to California or Illinois have the maximum number of departure and arrival delays.

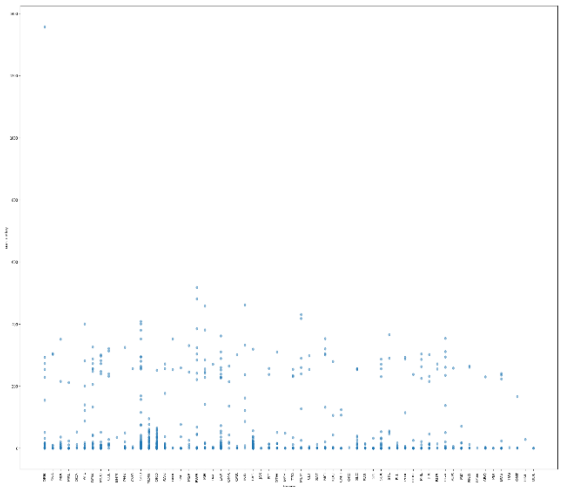

Figure 7

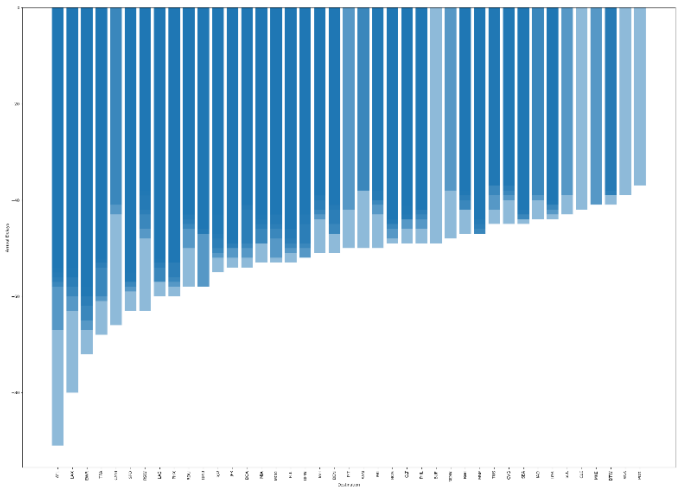

Figure 8

According to the analysis of airports with higher arrival delays these airports tend to have higher arrival delays and are the ones with higher departure delays. But the airports with least arrival delays i.e. early arrivals are (*Figure 8*):

1. Newark Liberty International Airport- New Jersey (EWR)
2. John F. Kennedy International Airport- New York (JFK)
3. Atlanta Airport- Atlanta (ATL)
4. O'Hare International Airport- Illinois (ORD)
5. San Francisco International Airport- California (SFO)
6. Los Angeles International Airport- California (LAX)
7. Harry Reid International Airport- Nevada (LAS)
8. Tampa International Airport- Florida (TPA)

Many of the above airports also tend to have higher departure delays yet they also have early arrivals. For every individual airport there will be 100's of flights operating each day and these are the overall statistics throughout the first two months of the year.

K means Clustering

In order to find a relation between airtime and delays, I have used K means clustering between the attributes Airtime and Departure and arrival delays. During the process of K means, I have chosen the number of clusters based on values generated by silhouette scores and also the elbow method for better accuracy of clusters. The silhouette method has automatically generated the optimal number of clusters based on the maximum score, but for the elbow method, I have manually calculated the slopes of line and chose the one with huge drop. The optimal number of clusters in both the methods turned out to be the same clusters, which is 2. Below is the graph of K means when plotted with maximum (*Figure 9*) and minimum (*Figure 10*) airtime and departure delays.

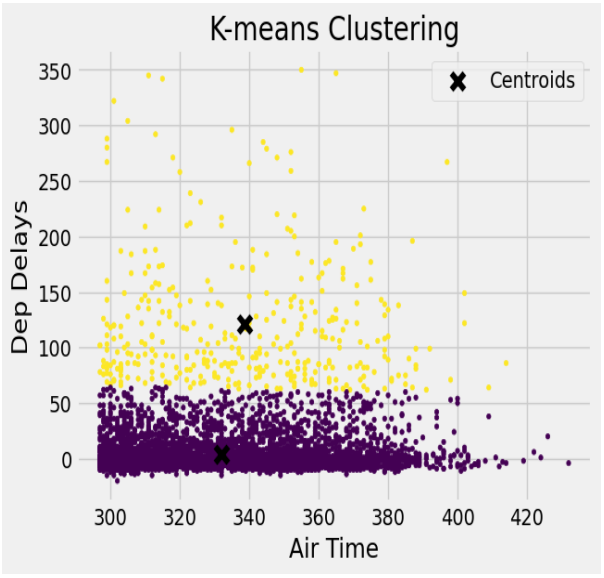

Figure 9

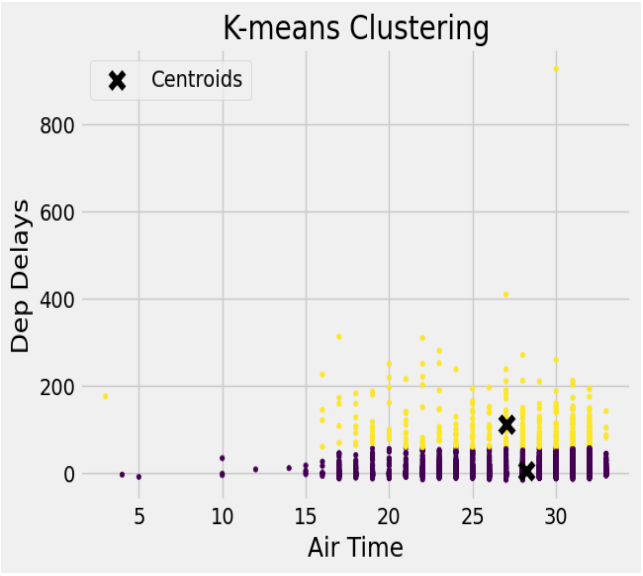

Figure 10

Below are the K means clustering plots plotted with maximum (*Figure 11*) and minimum (*Figure 12*) airtime and arrival delays.

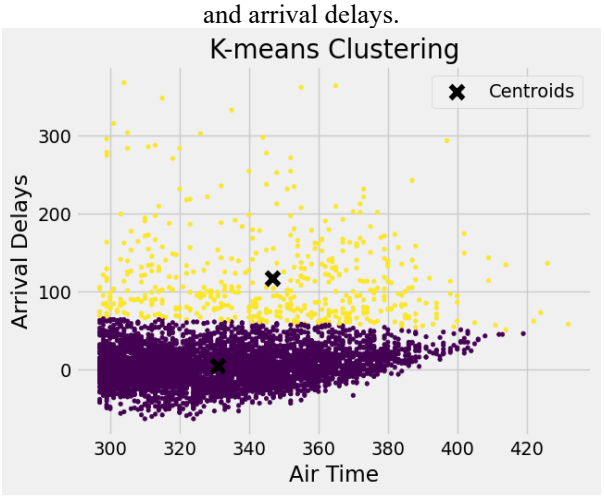

Figure 11

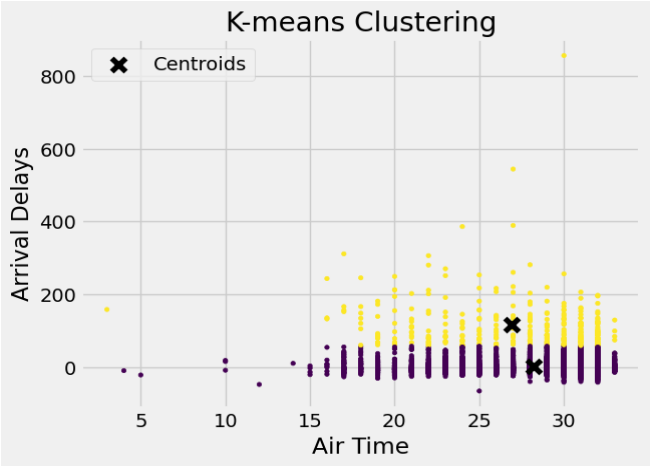

Figure 12

In these four plots, we can find that each of the departure delays is similar to their arrival delays plots, which is due to their very high correlation as calculated in the heatmaps. As we can observe, as the airtime (the distance between airports increases) both the arrival and departure delays decreased. Alongside shorter airtime there are very fewer delays. Most of the delays are for airlines with airtime in between 300-400 minutes of fly time. Aircraft with departure and arrival delays of less than 50 minutes tend to be different from the ones with delays greater than 50 minutes. One of the reasons for dissimilarity can be delays that are less than 50 minutes are often caused by minor issues such as weather, air traffic control, or late arriving passengers. These issues can often be resolved quickly, and the flight can still depart and arrive relatively close to its scheduled time. On the other hand, delays greater than 50 minutes are often caused by more serious issues such as mechanical problems, crew scheduling issues, or security issues. These issues can take longer to resolve and may even result in the flight being cancelled or delayed for several hours. And another possible consequence of having delays greater than 50 minutes for passengers might miss the connecting flights, important business meetings, or family events. In some cases, passengers may even have to spend the night in an airport or hotel, adding additional expenses and inconvenience.

This is the main reason for aircraft with delays of less than 50 minutes being different from the ones with greater than 50 minutes due to impact faced by the passengers.

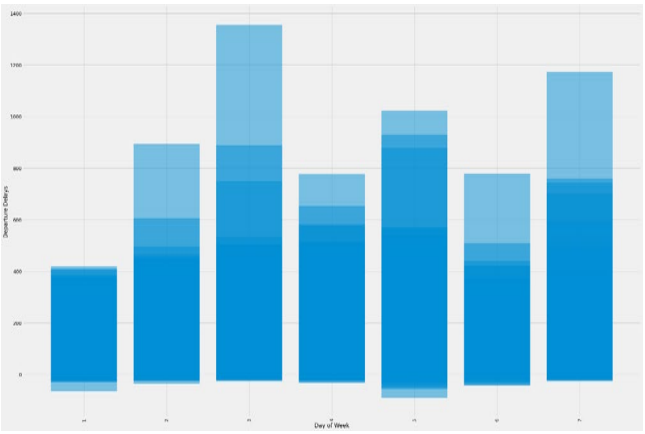

Figure 13

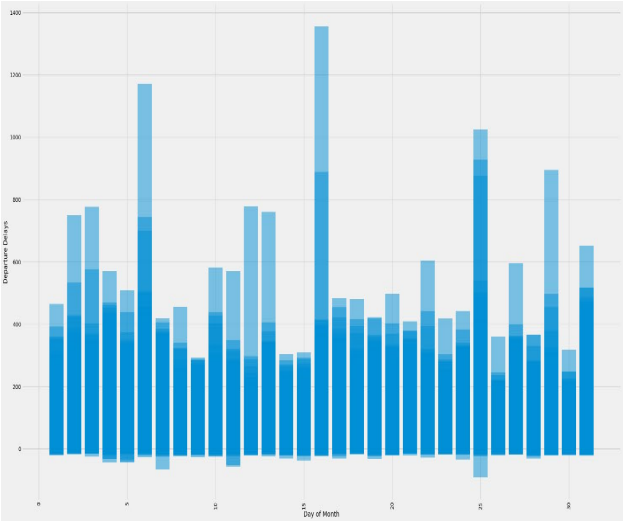

Figure 14

Delays with respect to Day of Week and Month

When the average departure delays are calculated with entire data set by grouping the day of the week and day of the month attributes and a selected amount of data set, the following observation have been made, these results have also been verified by plotting bar charts respectively (*Figure 13,14*).

Weekdays with early departures/arrivals:

1. Mondays (1): -16 minutes
2. Fridays (5): -17.3 minutes

Weekdays with less departure/arrival delays:

1. Saturdays (6): 9.44 minutes
2. Wednesdays (3): 10.11 minutes

Weekdays with high departure/arrival delays:

1. Thursday (4): 14.9 minutes
2. Sunday (7): 14.3 minutes

When the same method is used for the Day of the Month, it shows that the 6th, 16th and 12th day of the month tend to have maximum average of departure delays.

Summary and Conclusion

Throughout this project, I have encountered several errors and problems due to large amount of data set, regardless I have come up with several different approaches such as categorizing data based on maximum, minimum delays which eventually made my analysis easier in understanding the data and finding the airports/ air routes with maximum and minimum number of delays. Being it a airline data, which requires more analysis rather than clustering the data, I have primarily focused on analyzing and sub sectioning the data set by analyzing smaller data sets with feature selection by least and highest departure/arrival delays. The data and finding the results from them. In discovering the relationship between airtime and delays, I have used K means Clustering to see the similarity of points and came to discover an interesting renown finding. In the later part, I

have found the weekdays and what days in a month have the most and least number of delays based on grouping the data. All the conclusions and results have been mentioned in the earlier sections under the Data visualization and results along with the substantial proofs.

REFERENCES

[1] 2008, "2008.csv.bz2", Data Expo 2009: Airline on time data, https://doi.org/10.7910/DVN/HG7NV7/EIR0RA, Harvard Dataverse, V1

[2] 2008, "Data Expo 2009: Airline on time data", https://doi.org/10.7910/DVN/HG7NV7, Harvard Dataverse, V1