Montclair State University

Department of Computer Science

Masters Project Report

An Analysis of US Airline Delay and Cancellation Data

Yash Bulsara

Advisor: Dr. Stefan Robila

Department of Computer Science

Montclair State University

Montclair, NJ

Contents

[Acknowledgment 5](#_Toc153050299)

[Abstract: 6](#_Toc153050300)

[1. Introduction 7](#_Toc153050301)

[1.1Background 7](#_Toc153050302)

[1.2 Motivation 7](#_Toc153050303)

[1.3 Project Goal: 8](#_Toc153050304)

[1.4 Objective: 8](#_Toc153050305)

[2. Related work 9](#_Toc153050306)

[3. Datasets: 11](#_Toc153050307)

[4.0 Project Overview 12](#_Toc153050308)

[4.1 Importing Libraries 13](#_Toc153050309)

[Analysis 1: First I found the top 10 flight carriers by the amount of flights in the USA. 14](#_Toc153050310)

[Analysis 2: Find the reasons for the delay over the years to compare the data. 17](#_Toc153050311)

[Analysis 3: Find the average arrival delay by month for the last 5 years. Here I use ploty to plot data. 19](#_Toc153050312)

[Analysis 4: Find average arrival delays by destination for the last 5 years. Here I show a delay in minutes and destination by city. 21](#_Toc153050313)

[Analysis 5. find the most unreliable day to fly for the last 3 years. 24](#_Toc153050314)

[Analysis 6. The busiest airport in the USA in 2015. 25](#_Toc153050315)

[Analysis 7: The most popular city where people are boarding flights and also the least popular city for the same year 2015. 27](#_Toc153050316)

[Analysis 8: Comparing the arrival and departure delays for the flights operated in 2015 by each month. 28](#_Toc153050317)

[Analysis 9: Comparing several flights operating each day of the week during the year 2015. 30](#_Toc153050318)

[Analysis 10. Find the average arrival and departure delays by airlines 32](#_Toc153050319)

[Analysis 11: Find the average aircraft speed between departure and destination of airlines 35](#_Toc153050320)

[Analysis 12: Find the cancellation rate by airline companies as well as by cities. 37](#_Toc153050321)

[5.0 Project Setup 40](#_Toc153050322)

[5.1 Python Libraries 40](#_Toc153050323)

[6.0 Challenges 40](#_Toc153050324)

[7.0 Future work 41](#_Toc153050325)

[8.0 Conclusion 41](#_Toc153050326)

[References: 42](#_Toc153050327)

[Appendix: 43](#_Toc153050328)

Table of Figures:

[Figure 1 Research by David, Ranjit, and Patrick 9](#_Toc153049940)

[Figure 2 Importing Libraries 13](#_Toc153049941)

[Figure 3 Appending CSV Files 13](#_Toc153049942)

[Figure 4 Top 10 Carriers by amount of flights 15](#_Toc153049943)

[Figure 5 Reasons for Flight Cancellations 16](#_Toc153049944)

[Figure 6 Delayes per type for 2014 17](#_Toc153049945)

[Figure 7 Delays per type for 2017 18](#_Toc153049946)

[Figure 8 Delays per type for 2018 18](#_Toc153049947)

[Figure 9 Average arrival delay over the years 19](#_Toc153049948)

[Figure 10 Means Arrival Delays 2014 20](#_Toc153049949)

[Figure 11 Mean Arrival Delays 2018 21](#_Toc153049950)

[Figure 12 Mean Arrival by Destination 2014 22](#_Toc153049951)

[Figure 13 Mean Arrival by Destination 2016 23](#_Toc153049952)

[Figure 14 Mean Arrival by Destination 2018 23](#_Toc153049953)

[Figure 15 Top 10 Destinations by average Departure delay for Year 2022 24](#_Toc153049954)

[Figure 16 Most Unreliable day to fly 25](#_Toc153049955)

[Figure 17 Origin Airport Distribution 26](#_Toc153049956)

[Figure 18 Origin City Distribution 28](#_Toc153049957)

[Figure 20 Delay By Months 30](#_Toc153049958)

[Figure 21 Delay by day of week 32](#_Toc153049959)

[Figure 22 Mean Arrival and Departure Delay 34](#_Toc153049960)

[Figure 23 Mean Speed by Airlines 36](#_Toc153049961)

[Figure 24 Cancellation Rate by Airlines 39](#_Toc153049962)

[Figure 25 Cancellation Rates by Cities 39](#_Toc153049963)

# Acknowledgment

I would like to thank my advisor for this project Dr. Stefan Robila for his support and guidance which helped me to complete this project. The project helped me to learn new skills which will help me in my future career too. And also I Would like to thank all Montclair State University professors for helping and guiding me throughout my Master’s Program.

# Abstract:

Flight delays are gradually increasing and more financial difficulties and customer dissatisfaction with airline companies. To overcome this problem I am using exploratory data analysis to visualize the data so that airline companies know in which area they need to work to improve their efficiency. The Kaggle dataset Airline Delay and Cancellation Data, 2009 - 2018 provides a thorough analysis of airline performance and operational disturbances throughout the ten years from 2009 to 2018. This dataset is a useful tool for investigation, examination, and learning about the dependability and punctuality of airlines during this momentous time in aviation history. It includes vital details regarding individual flights, such as flight numbers, airports of departure and arrival, scheduled and real timings of departure and arrival, and the different kinds of delays these flights encountered, expressed in minutes. Additionally, the collection offers information on flight cancellations and their causes. This information can be utilized by analysts, and industry participants to identify trends, patterns, and causes that are responsible for flight interruptions. It offers a chance to look into how weather, air traffic, airline-specific operations, and other factors affect delays and cancellations. This dataset is a useful resource for comprehending how the airline industry is evolving and refining tactics to improve passengers' overall travel experiences. My goal is to analyze this information and visualize it so that companies can easily understand it. So I analyzed the dataset and visualized the information such as average delay by destination and also by year. Apart from that I found aircraft speed, the airline with the most operating flights, the busiest airport, cancellation rate by airlines.

# 1. Introduction

## 1.1Background

One of the biggest fears that dissatisfied people face is flight delays. Many flights are delayed every year, resulting in high costs for airlines and passengers. The airline's reputation is at risk and passengers' time and money are at stake. One of the best performance indicators of the aircraft is the suspension. Weather, traffic or other unforeseen circumstances are possible, but there are also causes that can be resolved through a streamlined process. Therefore, flight delay information is important to understand how travel is carried out.

The data I work on provides an important tool for investigating the reliability and timeliness of aircraft operations during a critical period in the aviation industry. It contains a lot of information such as flight number, departure and arrival airport, scheduled departure and arrival time, information regarding a particular flight such as departure and arrival time. The data also provides information on various types of flight delays, including departure delays, arrival delays, taxi departure delays and taxi arrival delays, all measured in minutes. The file also contains details of flight cancellations, including the reasons listed for each cancellation. Because of its ability to discover patterns, patterns, and other information, the data is useful to aviation industry researchers, analysts, and stakeholders.

At this critical time for the aviation industry, the following information provides an important tool to control the reliability and timeliness of aircraft operations. It contains useful information, including general personal details such as flight number, departure and arrival airports, scheduled departure and arrival times, and actual time. The data also provides information on the number of delays experienced in flights, including departure delays, arrival delays, taxiing and slow taxiing; All recorded in minutes. The database I use also contains information on flight cancellations, including a list of the reasons for each cancellation. This information is important to researchers, analysts, and stakeholders because it allows the search for patterns, patterns, and other information. .[1]

## 1.2 Motivation

In this review I will use different data search techniques to make the most of the data, I use Python for analysis. I use different Python libraries to analyze data. Flight delays can occur for many reasons such as weather, technical errors, poor management and more. Once the cause of the delay is found, airlines can work to improve service or a flight schedule can be created through analysis. This file contains flight data from 2009 to 2018. I also have some additional data to review along with flight data from 2015, and I also downloaded flight data from 2022 from the OST[2] website. I will review and compare models every year.

## Project Goal:

## • Check for suspensions and cancellations over the years.

## • Evaluate the performance of different airlines and airports. >• Evaluate the impact of factors such as weather, air traffic, and aircraft maintenance on flight operations.

## • Develop predictive models to predict potential delays and cancellations.

## • Evaluate the effectiveness of operational strategies to minimize disruption.

## 1.4 Objective:

The objective of this project is to analyze flight data and find some important information that can be useful to people or aircraft. Thanks to this review, airlines can save money by changing some of their operations. Travelers can also save valuable time by planning their trip properly. Therefore, the main target of this study is passengers, airport staff, travelers, airlines, etc. Using the data, delays at certain airports on certain days can be predicted and passengers can plan their trips accordingly. Thanks to this analysis, we can see the airports with the most delays, so authorities can take steps to reduce delays, such as adding a runway to the airport or taking some important measures. Finally, airlines can use this analysis to plan flights to minimize delays

# 2. Related work

In the research paper by David Phillips, Tanujit Dey, and Patrick Steele, they analyzed the US air transport network they used regression, simulation, and network optimization techniques.[3]

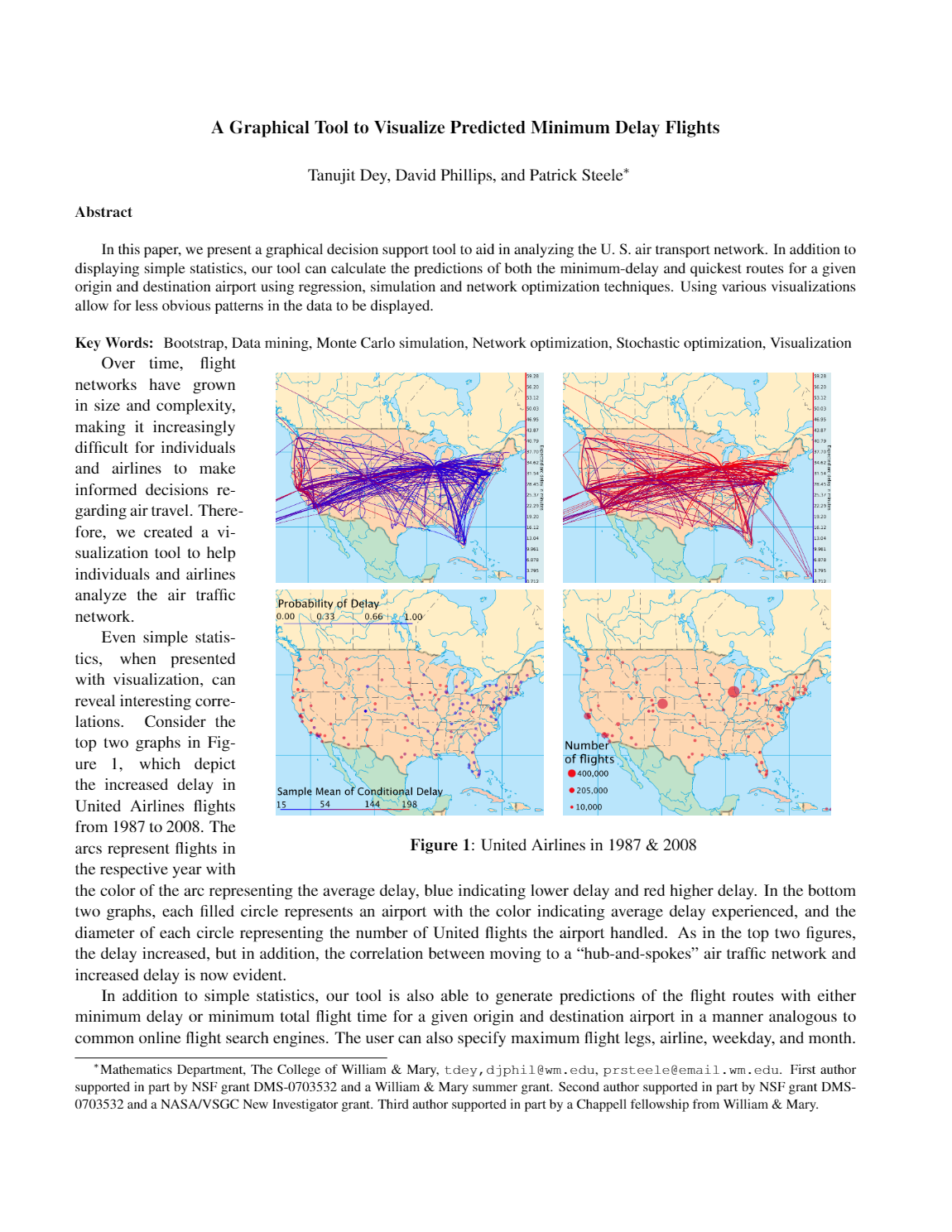


Figure Evaluation of US Transport Network [3]

Figure 1 shows the increase in United Airlines flight delays from 1987 to 2008. The arcs indicate dates on the corresponding dates. colors of the year and colors. The two graphs at the bottom show each filled circle as an airport, with the color designating the typical delay encountered and the diameter of each circle signifying the quantity of United Flights the airport handled.

In this paper, they explained about their tool how they made their tool, and using different algorithms, they analyzed the data.

A similar dataset has been used for several different data

mining tasks in a Coursera specialization course

called “IBM Data Analyst”:[4]

* Yearly number of flights canceled
* Average delay time by airline
* Monthly average delay for each type based on
* airline
* Yearly number of flights delayed based on the arrival and departure state

In the Research Paper By Yushan Jiang, Yongxin Lui, Dahai Liu, and Houbing Song, Named,” Applying Machine Learning to Aviation Big Data for Flight Delay Prediction”, They applied machine learning algorithms for flight delay prediction.

The goal of this research is to forecast flight delays through the analysis of two datasets from 2016: Quality Controlled Local Climatological Data and Airline On-Time Performance Data. Pre-processing data, feature engineering, and model building utilizing both deep learning neural networks and conventional machine learning models are all part of the research. The goal of the study is to investigate the many variables that affect aircraft delays, such as operational problems, weather, and airport traffic. To enhance operational effectiveness and decision-making in the aviation sector, the authors stress the significance of precise and timely flight delay prediction..[5]

In the Research Paper by Sanket Biswas, Riyanka Kundu, Somnath Rakshit, Priti Gupta, and Subhas Barman, Named, “A statistical approach to predict flight delay using gradient boosted decision tree”, they predicted flight delays using gradient boosted decision tree.

The paper addresses the unpredictable nature of departure and arrival delays owing to a variety of circumstances and explores the application of Gradient Boosted Decision Trees to predict flight delays. The suggested methodology focuses on features including the day of the week, carrier, origin and destination airport IDs, scheduled departure and arrival times, and actual departure and arrival delays to anticipate delays in passenger flights using regression-based models. The model achieves high accuracy in forecasting flight delay patterns through training and evaluation utilizing Mean Absolute Error, Root Mean Squared Error, and Coefficient of Determination. The model's potential uses for airline agencies, passengers, and logistics in aircraft transportation are highlighted in the paper. The document also identifies the model's shortcomings and recommends expanding its application to additional airports in the future. All things considered, the study shows how well the Gradient Boosted Decision Tree model works to anticipate flight delays.[6]

# 3. Datasets:

The datasets I used were taken from Kaggle. The datasets used are “Airline Delay and Cancellation Data, 2009 – 2018” and “2015 Flight Delays and Cancellations”.[7]

 "Airline Delays and Cancellations Data 2009-2018" is a comprehensive and detailed report that provides a comprehensive overview of the performance of various airlines and airports across the country. United States of America. The data covers a broad period from 2009 to 2018 and is an important tool for aviation industry data scientists, analysts and researchers looking to learn more about the complex world of flight operations.[8]

The data set has a number of important features that capture many aspects of flight data. This feature provides a detailed view of the profile, including the month and year the profile was closed, carrier code and name, and airport code and name. From important information such as arrivals, arrival times, arrival delays, cancellations, to detailed information about delays such as flight delays, air delays, National Airspace System slow, security slow, aircraft slow.

1. The first dataset contains 10 dataset files containing information about Airline delays and cancellations between 2009 to 2018. This dataset has data on flights between 2009 and 2018. This dataset has the following information.

*FL\_Date: Information about the flight from the departure point. The data set is in yy/mm/dd format.*

*OP\_carrier: This information relates to the identity of the aircraft. In other words, it has to do with the name of the airline.*

*OP\_rier\_FL\_Num: This data uniquely identifies a particular flight.*

*Background: This relates to the home airport of the flight. The configuration file contains the airport code of the origin airport.*

*DEST: Information regarding the destination airport of a particular flight.*

*CRS\_dep\_time: This refers to the time a particular flight leaves its base. It is given in hour and minute format. For example, if the departure time is 15:10, it will be marked as 1510 in the data set.*

*DEP\_Time: This information refers to the departure time from the original airport.*

*DEP\_Delay: This information represents the total delay time of the flight. It is the time difference between DEP\_Time and CRS\_dep\_time. It is usually measured in minutes, but in some cases can be measured in hours.*

*TAXI\_OUT: This information is about the time it takes for the aircraft to go from the airport exit gate to the runway and take off. This time will vary depending on the density of the airport.*

*Wheels\_Off: This information is about the time the flight spends in the air. In some cases, it may take more time than planned due to different reasons.*

1. The second dataset also taken from Kaggle which is similar to this dataset is “ 2015 Flight Delays and Cancellations”[9]

The "Flight Delays" Transportation Information set from the US Department of Transportation provides detailed information about domestic flights. Within the country. This information displays the main time zone and includes important features such as flight information (date, carrier, flight number, departure, destination), airport information, arrival and departure delays, cancellation information, and weather information. Due to its richness, these data are an important tool for analyzing and understanding the complexity of the slow plane. The 2015 Flight Delays dataset includes additional information that helps deepen the profile, such as airline name, IATA code, day of the week, and airport city and state.

1. I downloaded the 2022 flight data from the OST website which has huge data with various types of information. This data available on the OST website includes airline information such as airline IATA code, Flight number, Tail number, and unique code, and also this data includes origin and destination airport information such as airport ID, airport name, city name, and state name. Apart from this information, the data includes the departure and arrival performance of flights. Additionally, the data also have details about the cancellation and diversion of flights and the reasons for cancellation or delay. From all this information I downloaded some of these data to perform my analysis. [2]

# 4.0 Project Overview

Using a Python notebook to visualize the data analysis results was the method I used to create this project. Python notebooks facilitate the sharing of documents that combine explanation text, equations, computational output, graphics, and other multimedia resources with live code. Numerous data science tasks, such as exploratory data analysis and data transformation and purification, are carried out using these notebooks. It is also a type of interactive computing. Users run code in this environment and observe what happens in an iterative dialogue between the data scientist and the data to edit or repeat.

## 4.1 Importing Libraries

Here I am Importing Libraries such as pandas, Numpy, and Matplotlib.

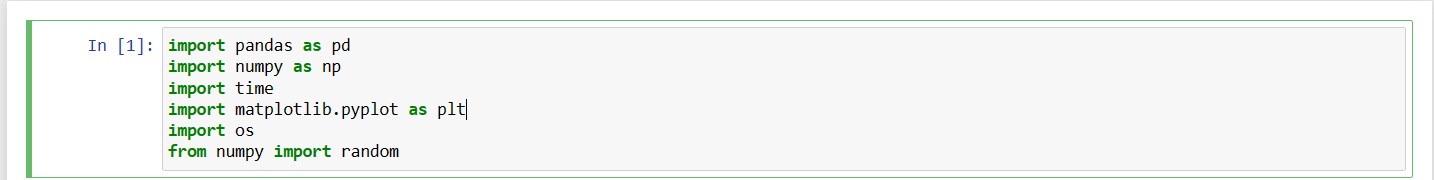


Figure Importing Libraries

I am Importing all CSV files and appending them together.



Figure Appending CSV Files

## Analysis 1: First I found the top 10 flight carriers by the amount of flights in the USA.

Code:

carriers\_flight\_count\_df = analysis\_df.groupBy(F.col('OP\_CARRIER')).count().orderBy(F.col('count').desc())

top\_10 = carriers\_flight\_count\_df.limit(10).toPandas()

top\_10 = top\_10.rename(columns={'OP\_CARRIER':'Carrier'})

top\_10

Using this code I found my output and I visualized it using a chart so It can be easily understood.

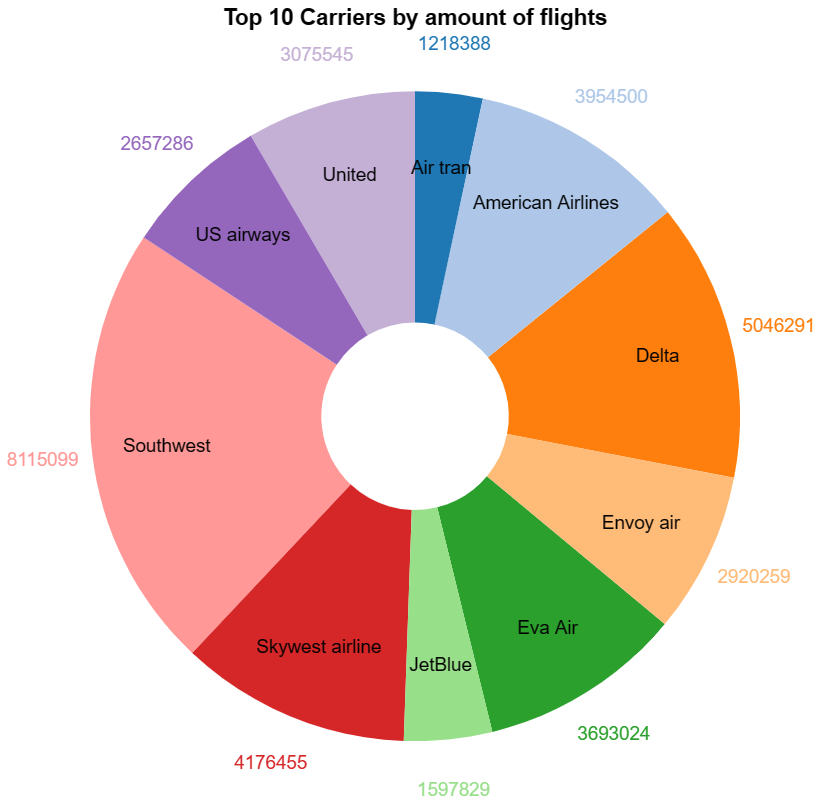


Figure Top 10 Carriers by amount of flights

Results:

This Visualization shows the airlines with the most operating flights. Using this dataset airport authority can provide more space to those airlines so it can avoid delays. Apart from that customers can also decide their airline to fly.

* After This I found reasons for cancellations of flights, I also visualized that in a chart[10]

Visualization:

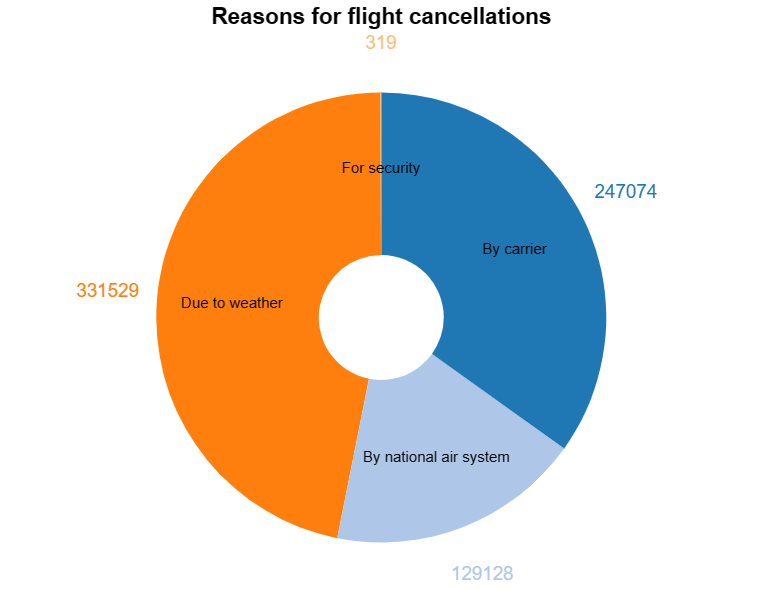


Figure Reasons for Flight Cancellations

Results:

In the chart, it is visible that weather conditions contribute most to delays and the second reason is by the airline companies for reasons such as poor management or some fault in the plane.

To avoid this delay authorities can plan an alternate way to minimize weather delays. By building modern infrastructure, the delays can be reduced. The second major reason is delays by airline companies, which can be avoided by proper training of staff and regular maintenance of aircraft.

## Analysis 2: Find the reasons for the delay over the years to compare the data.

Code:

# Different types of delays by year [11]

delays = ['CARRIER\_DELAY', 'WEATHER\_DELAY', 'SECURITY\_DELAY', 'LATE\_AIRCRAFT\_DELAY']

for i in range(len(csv\_list\_delay)):

mean2 = csv\_list\_delay[i][['CARRIER\_DELAY', 'WEATHER\_DELAY', 'SECURITY\_DELAY', 'LATE\_AIRCRAFT\_DELAY']].mean()

X = np.arange(len(delays))

fig = plt.figure()

plt.bar(X - 0.2, mean2, color = 'b', width = 0.40, label = 'Mean')

plt.xticks(X, delays, rotation = 45)

plt.title('Delays per type for ' + year[i], fontsize=16)

plt.ylabel('Minutes', fontsize=14)

plt.xlabel('Delay', fontsize=14)

plt.legend()

plt.show()

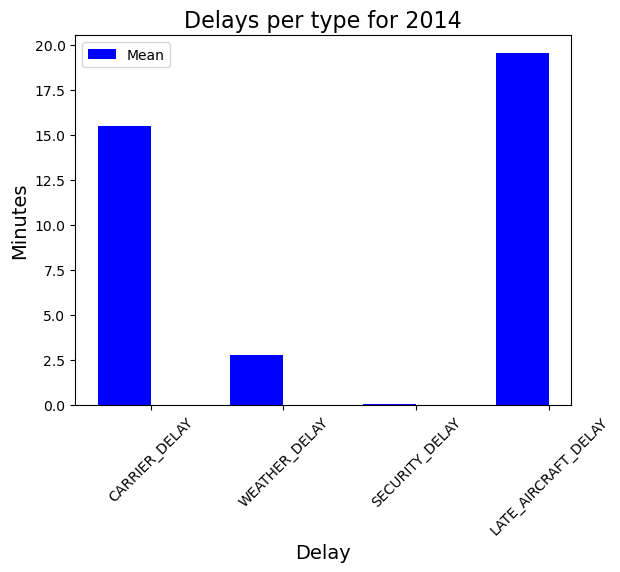


Figure Delayes per type for 2014

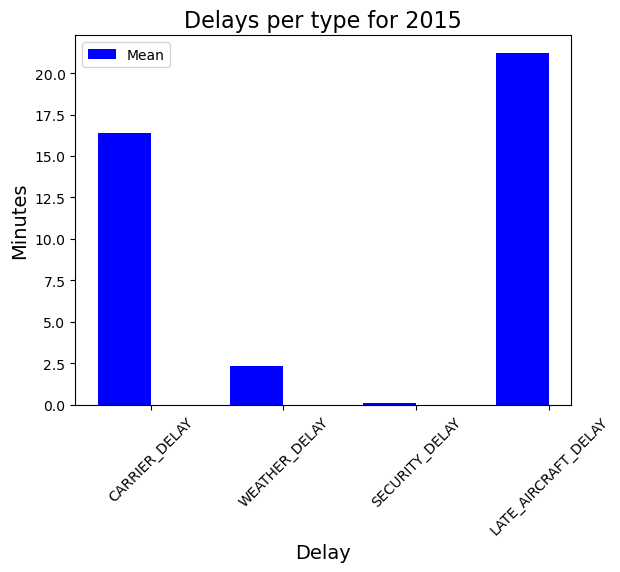


Figure Delayes per type for 2015

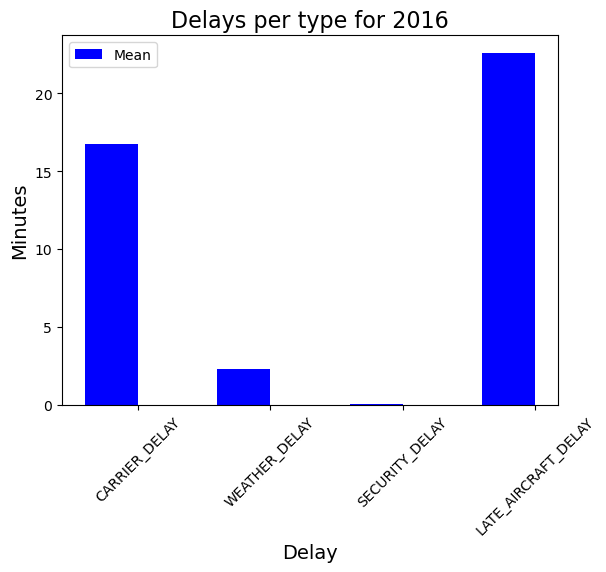


Figure Delyes per type for 2016

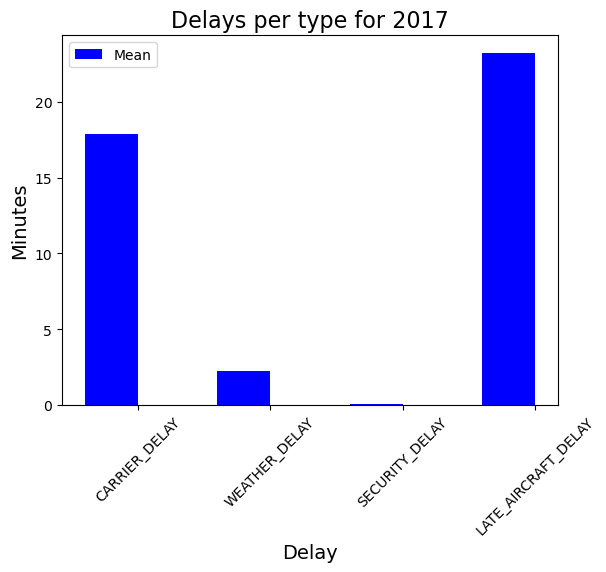


Figure Delays per type for 2017

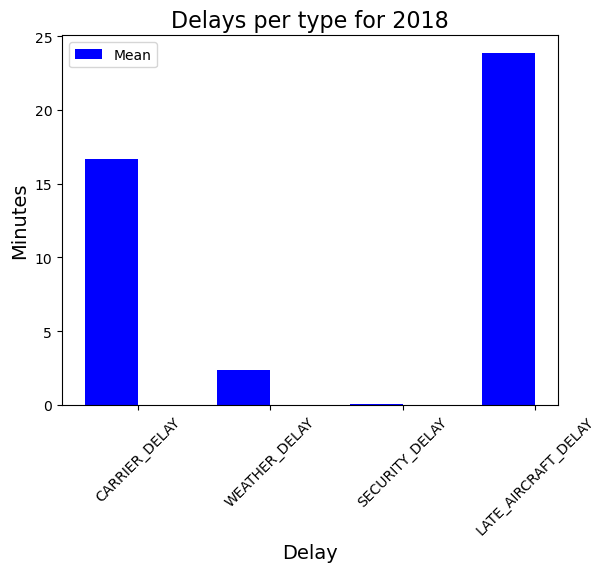


Figure Delays per type for 2018

Results:

The analysis shows reasons for delaying flights over the years. This shows that the reasons contribute almost the same over the years so all the issues need to be addressed equally to avoid delays. Here In the document, I included graphs for 2014, 2017, and 2018 and I combined average delay for all years.

The above Results show the delays by months for all the years by its delay type. Now I am combining these results to show the average delay across all years. This shows that the average delay increased over the years.

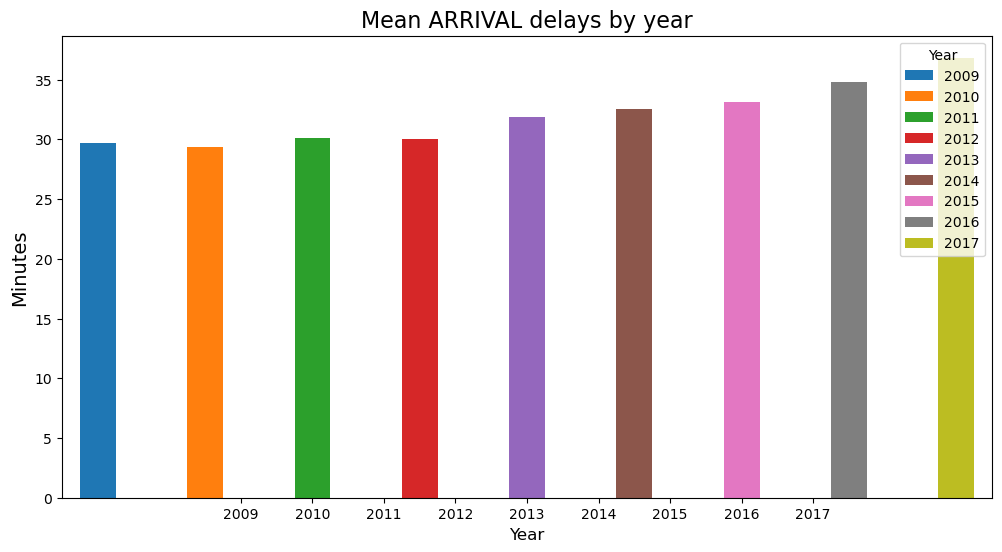


Figure Average arrival delay over the years

## Analysis 3: Find the average arrival delay by month for the last 5 years.

Here I used ploty to plot data.

# Delay per month (2014-2018), mean and median value

for i in range(len(csv\_list\_delay)):

csv\_list\_delay[i]['FL\_DATE\_month'] = pd.to\_datetime(csv\_list\_delay[i]['FL\_DATE']).dt.month

plt.figure(figsize=(25, 12)).subplots\_adjust(hspace = 0.5)

plt.subplot(2, 2 ,1)

csv\_list\_delay[i].groupby('FL\_DATE\_month').ARR\_DELAY.mean().plot.bar()

plt.title('Mean ARRIVAL delays by month for ' + year[i], fontsize=16)

plt.ylabel('Minutes', fontsize=14)

plt.xlabel('month of the year', fontsize=14)

plt.subplot(2, 2 ,2)

csv\_list\_delay[i].groupby('FL\_DATE\_month').ARR\_DELAY.median().plot.bar()

plt.title('Median ARRIVAL delays by month for ' + year[i], fontsize=16)

plt.ylabel('Minutes', fontsize=14)

plt.xlabel('month of the year', fontsize=14)

plt.show()

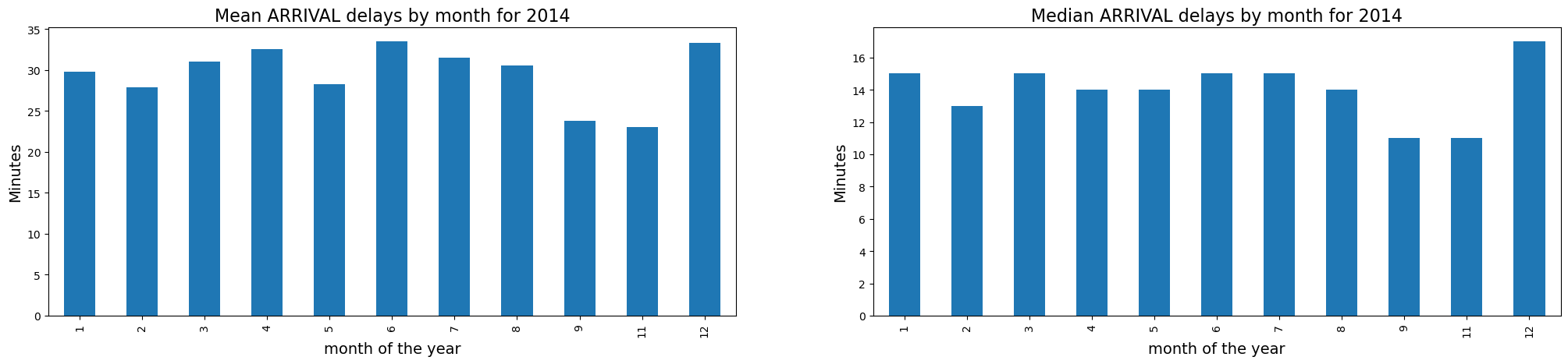


Figure Mean and Median Arrival Delays 2014

Results:

Here I found the average arrival delay by months for the year 2014. Here the data shows December has the highest arrival delays. The reason can be weather or management issues. This data will help authorities find the cause of the delay and rectify it. Apart from that the passenger can also plan their journey using this visualization.

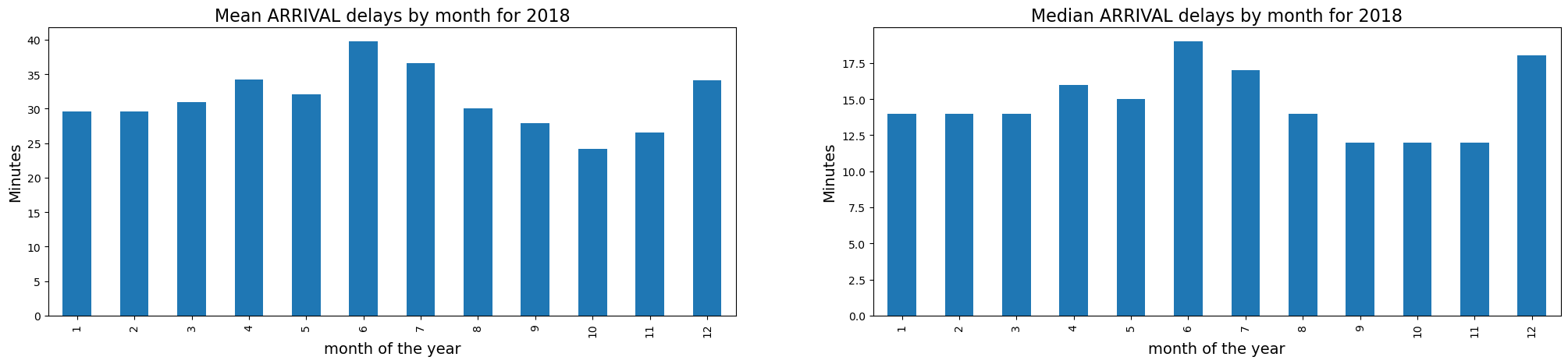


Figure Mean and Median Arrival Delays 2018

Results:

Here I found the average arrival delay by month for the year 2018. Compared to 2014 the arrival delays the delay time is increased comparatively. It shows that over the years air traffic has increased, and the industry is getting new passengers so airports need to expand their infrastructure to manage more planes.

## Analysis 4: Find average arrival delays by destination for the last 5 years.

Here I showed a delay in minutes and destination by city.

Code:

# Delay per Destination (2014-2018), mean and median value

for i in range(len(csv\_list\_delay)):

plt.figure(figsize=(25, 12)).subplots\_adjust(hspace = 1)

plt.subplot(2, 2 ,1)

mean\_origin = csv\_list\_delay[i].groupby('DEST').ARR\_DELAY.mean().sort\_values(ascending=False)

mean\_origin[:20].plot.bar()

plt.title('Mean ARRIVAL delays by Destination for ' + year[i], fontsize=16)

plt.ylabel('Minutes', fontsize=14)

plt.xlabel('City', fontsize=14)

plt.subplot(2, 2 ,2)

median\_origin = csv\_list\_delay[i].groupby('DEST').ARR\_DELAY.median().sort\_values(ascending=False)

median\_origin[:20].plot.bar()

plt.title('Median ARRIVAL delays by Destination for ' + year[i], fontsize=16)

plt.ylabel('Minutes', fontsize=14)

plt.xlabel('City', fontsize=14)

plt.show()

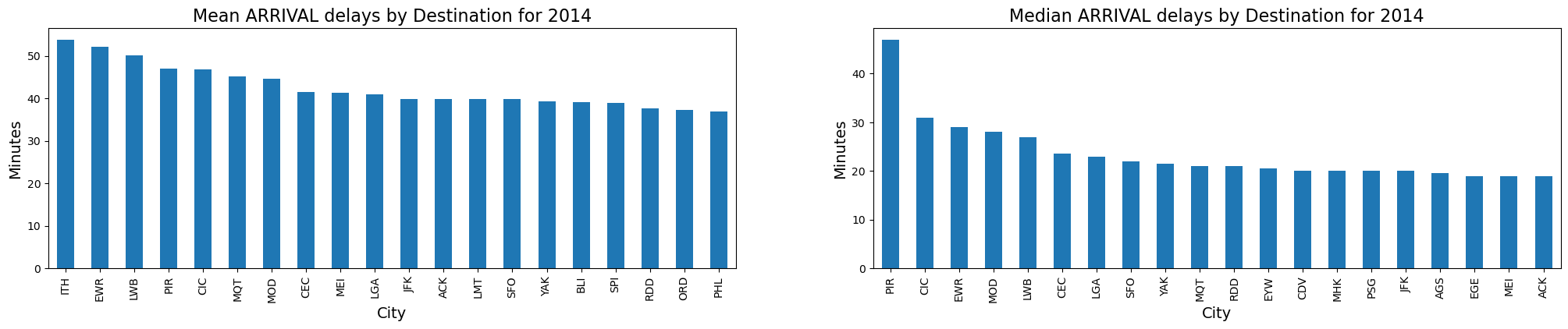


Figure Mean and medianArrival by Destination 2014

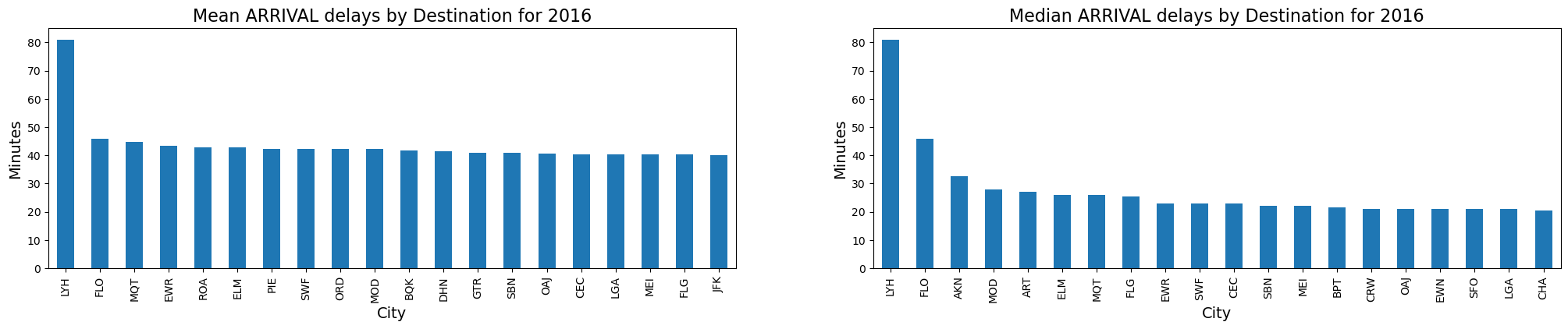


Figure Mean and Median Arrival by Destination 2016

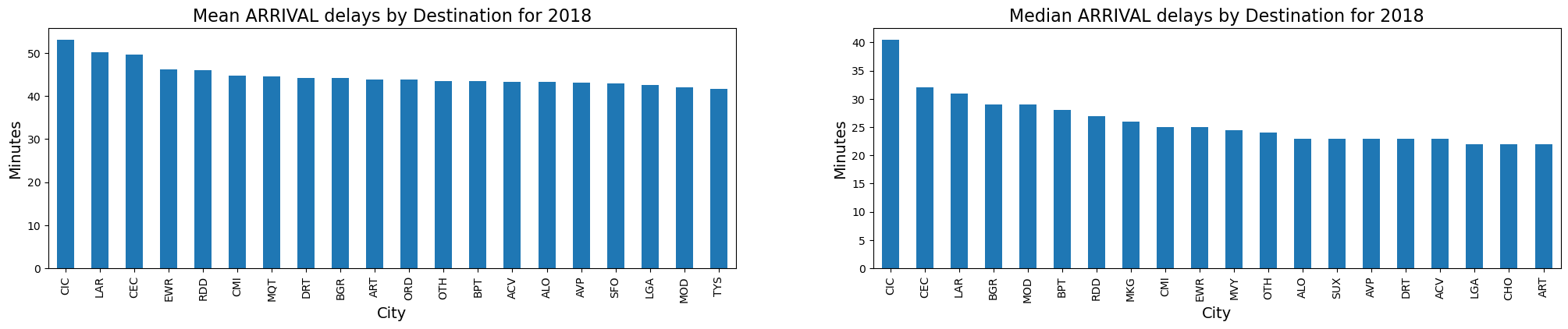


Figure Mean and Median Arrival by Destination 2018

Results:

Here are average arrival delays by destination for the last few years. This represents that over the years the the destination has changed for most delayed arrivals of airplanes. This data helps to reschedule flights or make additional runways at a particular airport.

Apart from this, I analyzed 2022 data which I downloaded from the OST website. Here are the top 10 Destinations by average departure delay for the year 2022.[2]

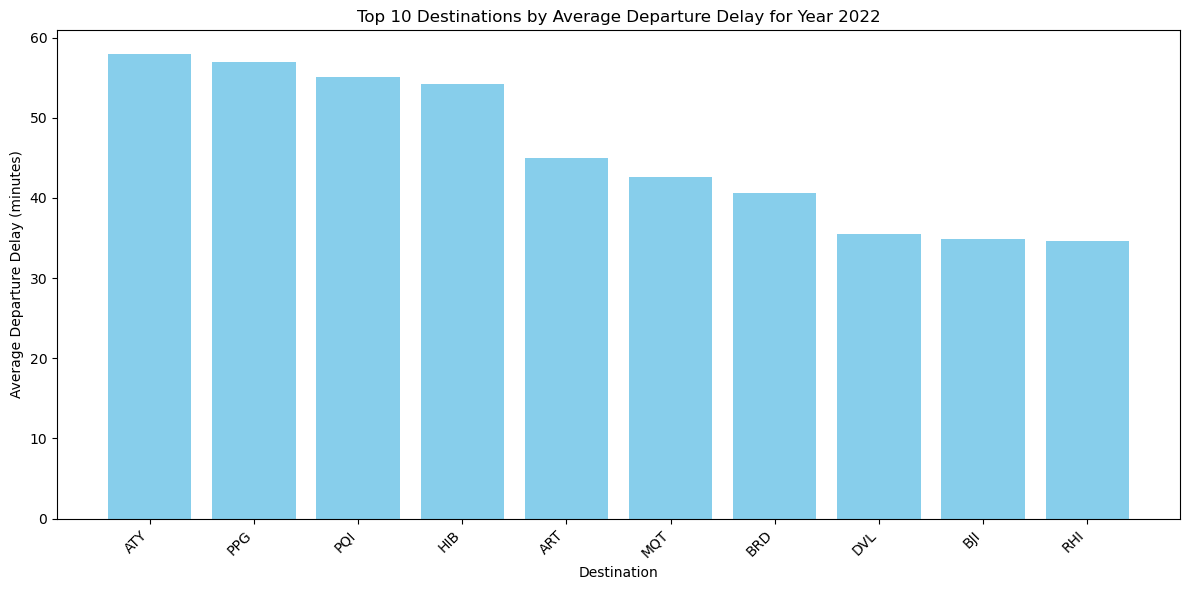


Figure Top 10 Destinations by average Departure delay for Year 2022

## Analysis 5. find the most unreliable day to fly for the last 3 years.

For that, I took datasets of the last 3 years and performed analysis using the origin of the flights.

Code:

#most unreliable day to fly in last 3 years

df = pd.read\_csv('Final\_project/2018.csv')

df = df[df['ORIGIN'].isin(['ATL', 'LAX', 'ORD', 'DFW',' DEN',' JFK', 'SFO',' SEA', 'LAS', 'MCO'])]

df\_2017 = df.append( pd.read\_csv('Final\_project/2017.csv'))

df\_2017 = df\_2017[df\_2017['ORIGIN'].isin(['ATL', 'LAX', 'ORD', 'DFW',' DEN',' JFK', 'SFO',' SEA', 'LAS', 'MCO'])]

df = df.append(df\_2017)

df\_2016 = df.append( pd.read\_csv('Final\_project/2016.csv'))

df\_2016 = df\_2016[df\_2016['ORIGIN'].isin(['ATL', 'LAX', 'ORD', 'DFW',' DEN',' JFK', 'SFO',' SEA', 'LAS', 'MCO'])]

df = df.append(df\_2016)

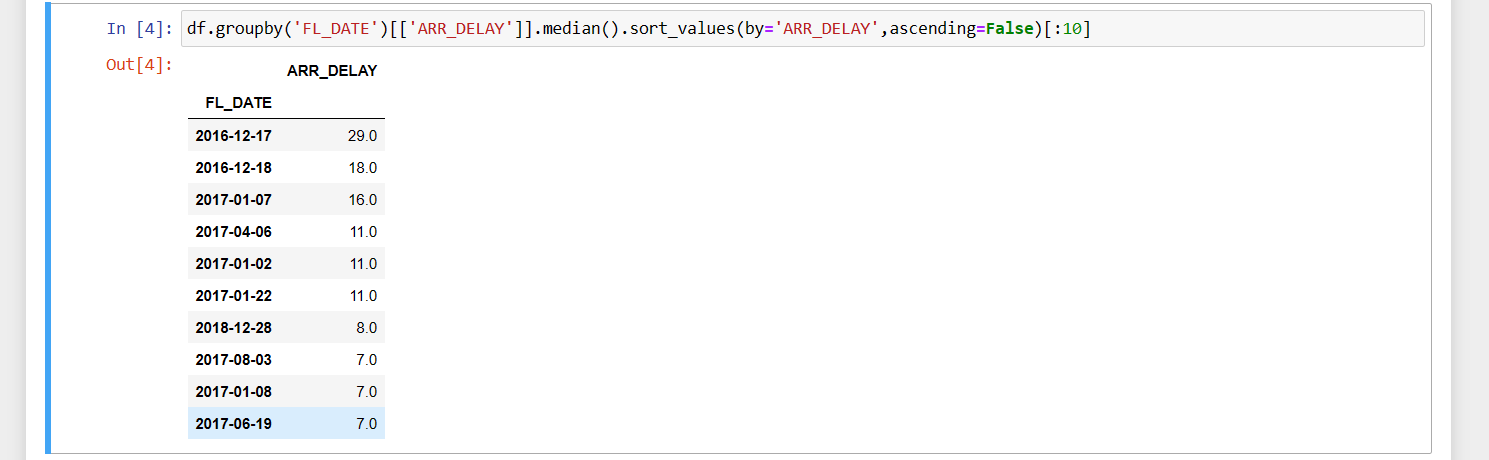


Figure Most Unreliable day to fly

Results:

The analysis shows that 17th December 2016 was the most unreliable day to fly which has had maximum delay for the last 3 years. The analysis also shows that most of the delays occur in winter due to the weather conditions. This analysis only shows the days that had the most average delay per flight for a particular day. This analysis is important to learn from that day.

* Now I am using another data set which is 2015 flight delays and cancellations.

## Analysis 6. The busiest airport in the USA in 2015.

Code:

dff = df['AIRPORT\_x'].value\_counts()[:10]

label = dff. index

size = dff. values

colors = ['sky blue, '#FEBFB3', '#96D38C', '#D0F9B1', 'gold', 'orange', 'light grey',

'light blue','light green,'aqua']

trace = go.Pie(labels=label, values=size, marker=dict(colors=colors),hole = .2)

data = [trace]

layout = go.Layout(

title='Origin Airport Distribution'

)

fig = go.Figure(data=data, layout=layout)

py.iplot(fig)

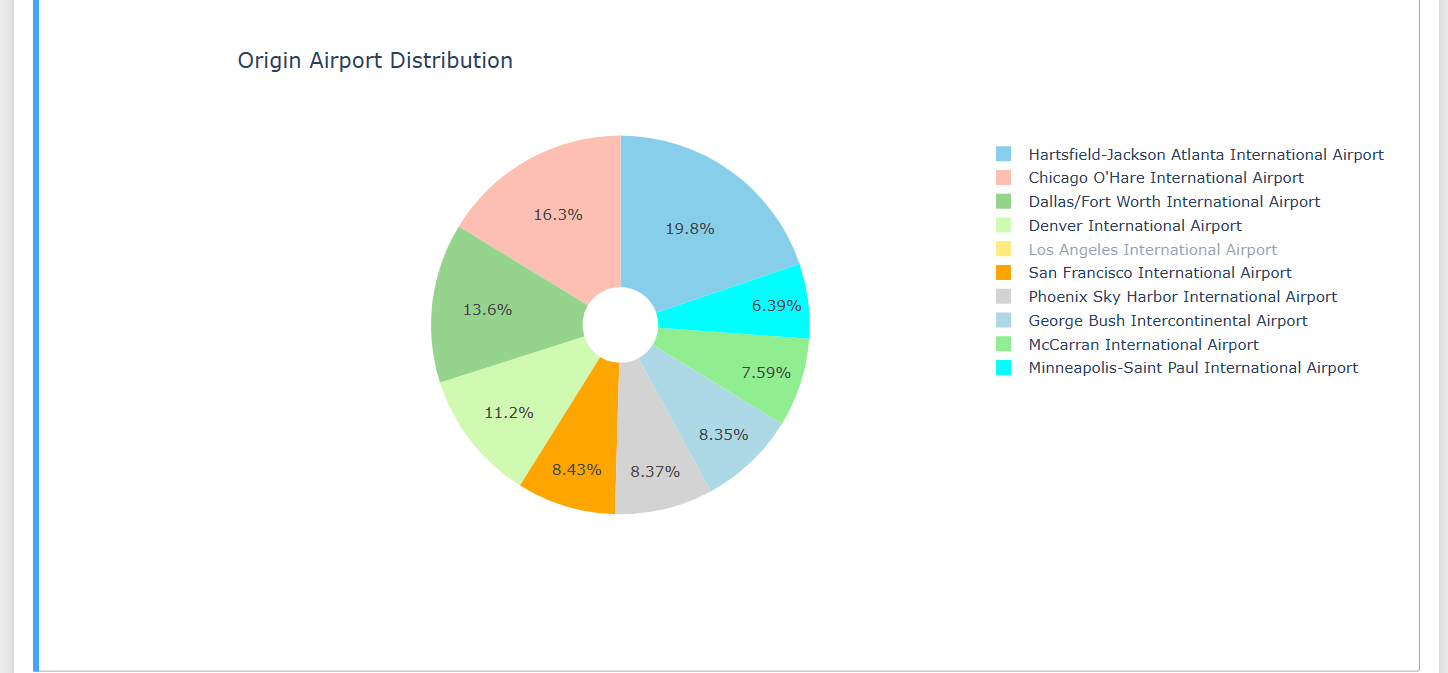


Figure Origin Airport Distribution

Results:

The analysis shows that Atlanta airport is the busiest airport in the USA with a total share of more than 19% followed by Chicago airport with more than 16% contribution. Saint Paul airport is the least busy airport of all. This analysis is helpful for the government to build infrastructure for particular airports or to build new airports.

## Analysis 7: The most popular city where people are boarding flights and also the least popular city for the same year 2015.

Code:

dff = df.CITY\_x.value\_counts()[:10]

trace = go.Bar(

x=dff.index,

y=dff.values,

marker=dict(

color = dff.values,

colorscale='Jet',

showscale=True

)

)

data = [trace]

layout = go.Layout(

title='Origin City Distribution',

yaxis = dict(title = '# of Flights')

)

fig = go.Figure(data=data, layout=layout)

py.iplot(fig)

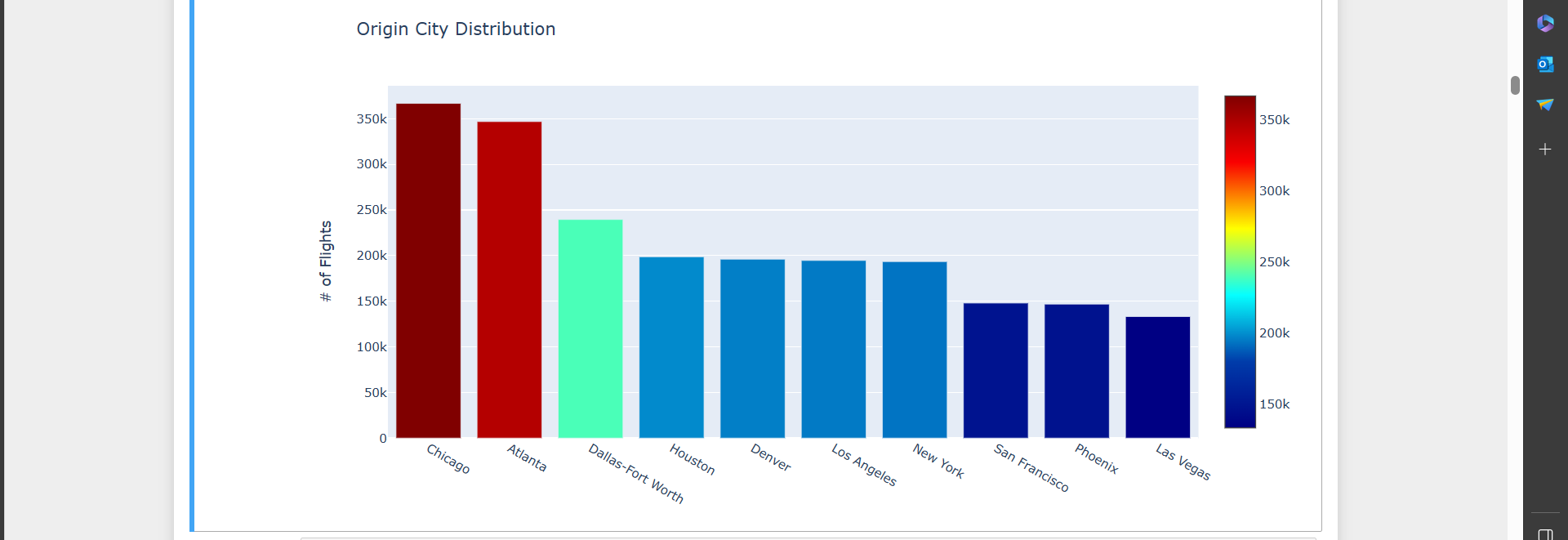


Figure Origin City Distribution

Results:

The analysis shows that Chicago is managing more flights compared to other cities in the USA. Using this dataset passenger can decide which city to board their flight and also government can take proper steps to manage air traffic.

## Analysis 8: Comparing the arrival and departure delays for the flights operated in 2015 by each month.

Code:

df['dep\_delay'] = np.where(df.DEPARTURE\_DELAY>0,1,0)

df['arr\_delay'] = np.where(df.ARRIVAL\_DELAY>0,1,0)

dff = df.groupby('MONTH').dep\_delay.mean().round(2)

dff.index = dff.index.map(month)

trace1 = go.Bar(

x=dff.index,

y=dff.values,

name = 'Departure\_delay',

marker = dict(

color = 'aqua'

)

)

dff = df.groupby('MONTH').arr\_delay.mean().round(2)

dff.index = dff.index.map(month)

trace2 = go.Bar(

x=dff.index,

y=dff.values,

name='Arrival\_delay',

marker=dict(

color = 'red'

)

)

data = [trace1,trace2]

layout = go.Layout(

title='% Delay (Months)',

yaxis = dict(title = '%')

)

fig = go.Figure(data=data, layout=layout)

py.iplot(fig)

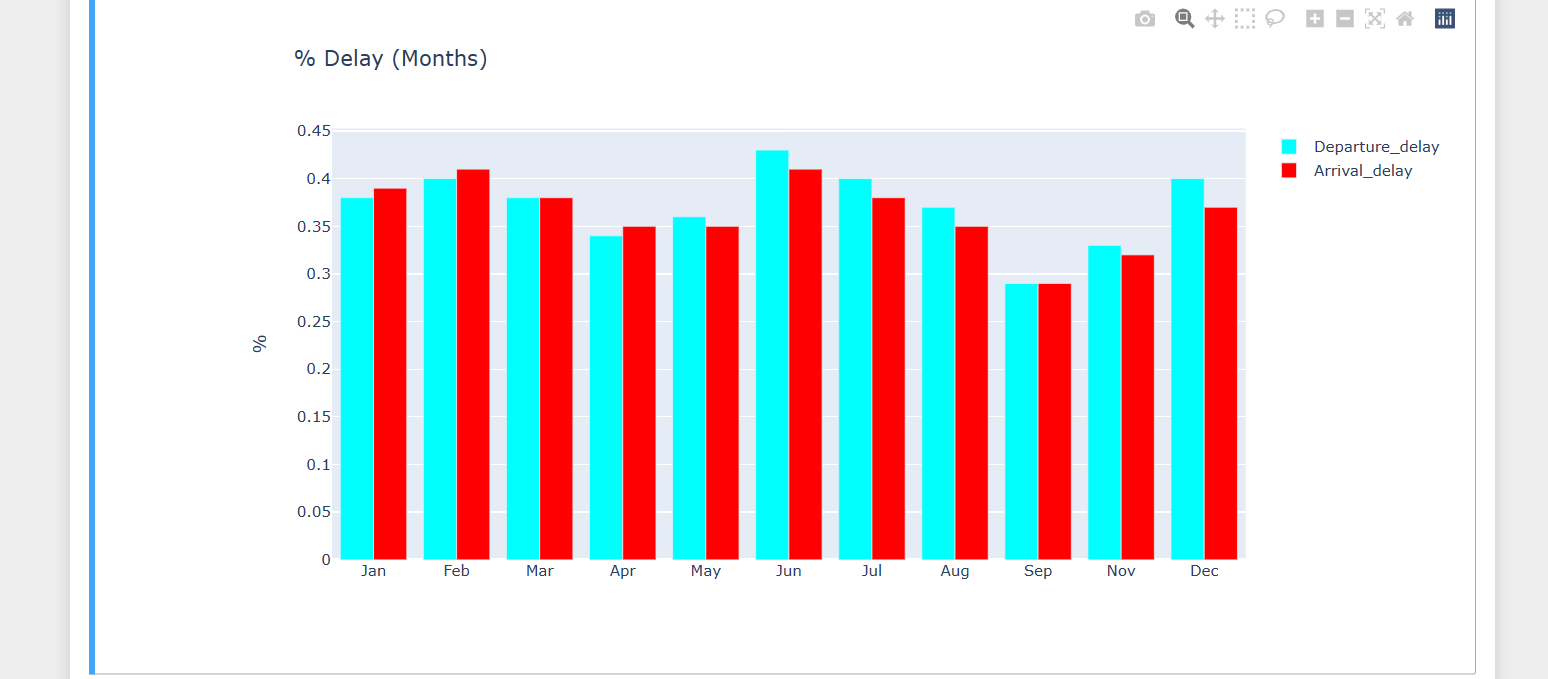


Figure Delay By Months

Results:

This analysis shows some interesting results as for the first few months of the year the arrival delay is greater than the departure delay and for the rest of the year the departure delay is greater than the arrival delay. There can be multiple reasons for that such as weather conditions, shortage of staff, airport conditions, etc. By rectifying the issues for the delay the delay can be reduced.

## Analysis 9: Comparing several flights operating each day of the week during the year 2015.

Code:

dayOfWeek={1:'Monday', 2:'Tuesday', 3:'Wednesday', 4:'Thursday', 5:'Friday',

6:'Saturday', 7:'Sunday'}

dff = df.DAY\_OF\_WEEK.value\_counts()

dff = dff.to\_frame().sort\_index()

dff.index = dff.index.map(dayOfWeek)

trace1 = go.Bar(

x=dff.index,

y=dff.DAY\_OF\_WEEK,

name = 'Weather',

marker=dict(

color = dff.DAY\_OF\_WEEK,

colorscale='Jet',

showscale=True

)

)

data = [trace1]

layout = go.Layout(

title='# of Flights (Day of Week)',

yaxis = dict(title = '# of Flights'

)

)

fig = go.Figure(data=data, layout=layout)

py.iplot(fig)

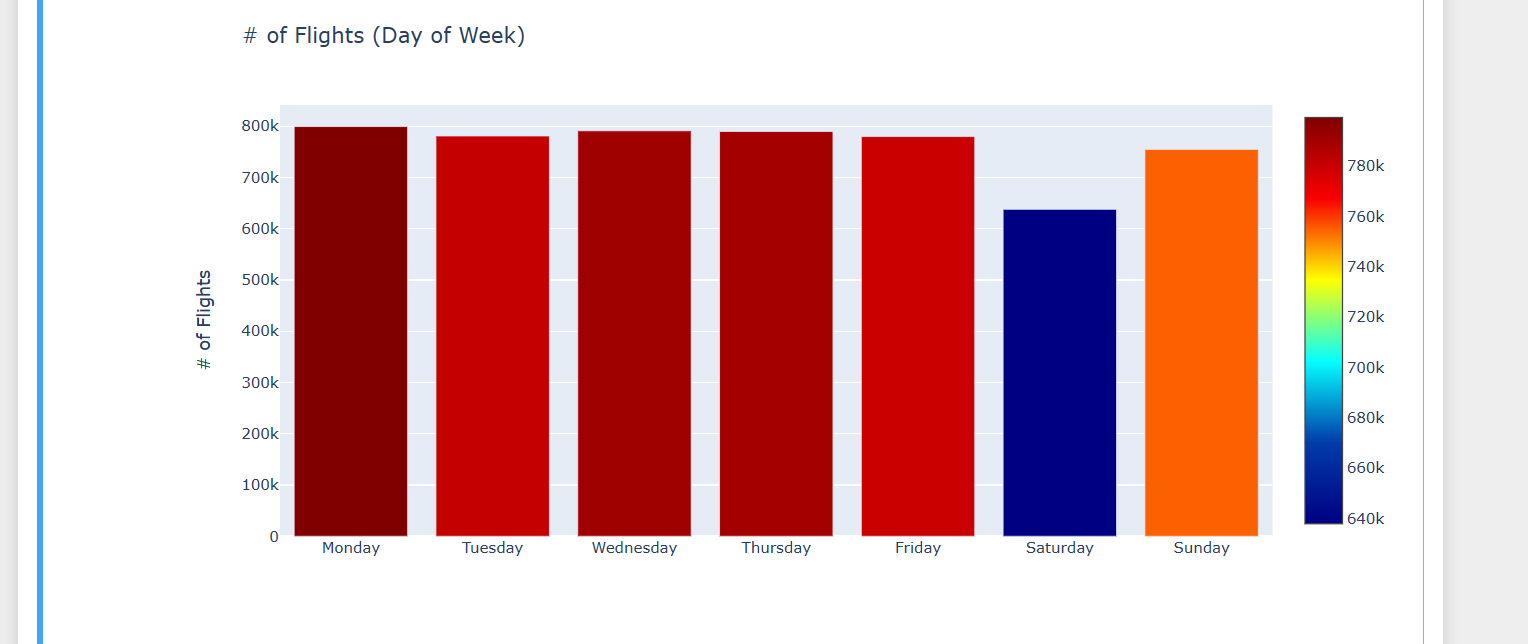


Figure Delay by day of week

Results:

The analysis shows that Saturday is the least busy day as data shows that several flights operating on Saturday is less than on other days of the week while Monday was the busiest day. This analysis is useful for non-frequent travelers as they can plan their journey on Saturday to avoid delays and to avoid long queues at the airports.

## Analysis 10. Find the average arrival and departure delays by airlines

df['DEP\_ARR\_DIFF'] = df['DEPARTURE\_DELAY'] - df['ARRIVAL\_DELAY']

dff = df.groupby('AIRLINE').DEP\_ARR\_DIFF.mean().to\_frame().sort\_values(by='DEP\_ARR\_DIFF',

ascending=False).round(2)

trace = go.Bar(

x=dff.index,

y=dff.DEP\_ARR\_DIFF,

marker=dict(

color = dff.DEP\_ARR\_DIFF,

colorscale='Jet',

showscale=True

)

)

data = [trace]

layout = go.Layout(xaxis=dict(tickangle=15),

title='Mean (Departure Delay - Arrival Delay) by Airlines',

yaxis = dict(title = 'minute')

)

fig = go.Figure(data=data, layout=layout)

py.iplot(fig)

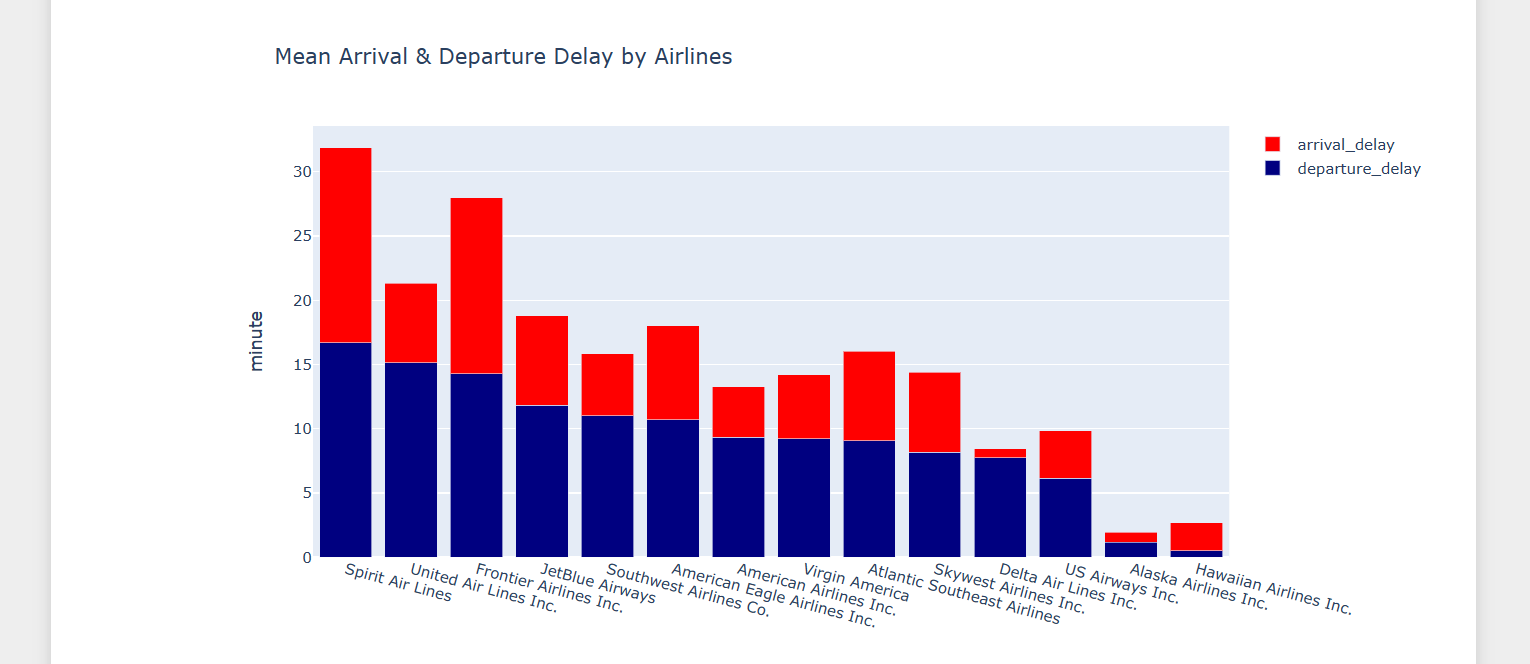


Figure Mean Arrival and Departure Delay

Results:

The analysis shows the average departure and arrival delay by a particular airline. There are many reasons for departure delays such as airline management issues, runway issues, weather issues, etc. Same for arrival delays the reasons can be weather issues, ATC issues, late departure issues, etc.

By this airline companies can rectify their issues and they can make efforts to solve their issues to improve their service and business.

## Analysis 11: Find the average aircraft speed between departure and destination of by airlines

Code:

df['SPEED'] = 60\*df['DISTANCE']/df['AIR\_TIME']

dff = df.groupby('AIRLINE').SPEED.mean().to\_frame().sort\_values(by='SPEED',

ascending=False).round(2)

trace = go.Scatter(

x=dff.index,

y=dff.SPEED,

mode='markers',

marker=dict(

sizemode = 'diameter',

sizeref = 1,

size = 30,

color = dff.SPEED.values,

colorscale='Jet',

showscale=True

)

)

data = [trace]

layout = go.Layout(xaxis=dict(tickangle=-20),

title='Mean Speed by Airlines',

yaxis = dict(title = 'Speed')

)

fig = go.Figure(data=data, layout=layout)

py.iplot(fig)

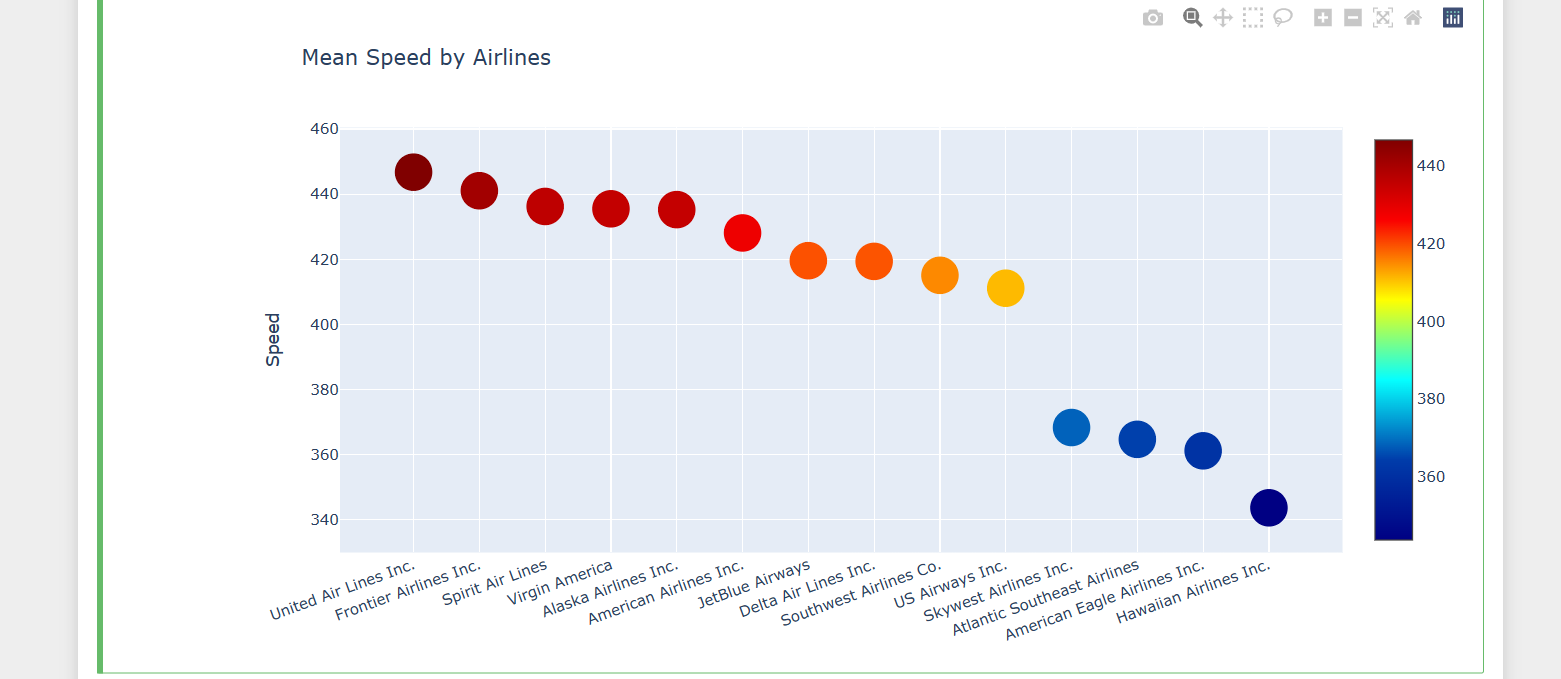


Figure Mean Speed by Airlines

Results:

The analysis shows the average speed of aircraft by the airlines. Here, United airlines operate their aircraft at the highest speed with an average speed of more than 450mph followed by Frontier Airlines with an average speed of around 440 while Hawaiian Airlines operates at the lowest speed compared to all other airline companies with an average speed of around 350mph. The speed of the aircraft depends on many factors such as the condition of the aircraft, the area where the aircraft is flying and the distance covered by a particular aircraft. This analysis can be very useful for airline companies to compete against their rivals.

## Analysis 12: Find the cancellation rate by airline companies as well as by cities.

Code:

dff = df.groupby('AIRLINE')[['CANCELLED']].mean().sort\_values(by='CANCELLED',

ascending=False).round(3)

trace1 = go.Scatter(

x=dff.index,

y=dff.CANCELLED,

mode='markers',

marker=dict(

symbol = 'star-square',

sizemode = 'diameter',

sizeref = 1,

size = 30,

color = dff.CANCELLED,

colorscale='Portland',

showscale=True

)

)

data = [trace1]

layout = go.Layout(xaxis=dict(tickangle=20),

title='Cancellation Rate by Airlines', yaxis = dict(title = 'Cancellation Rate'

)

)

fig = go.Figure(data=data, layout=layout)

py.iplot(fig, filename="age")

dff = df.groupby('CITY\_x')[['CANCELLED']].mean().sort\_values(by='CANCELLED',

ascending=False)[:10].round(3)

trace2 = go.Scatter(

x=dff.index,

y=dff.CANCELLED,

mode='markers',

marker=dict(symbol = 'diamond',

sizemode = 'diameter',

sizeref = 1,

size = 30,

color = dff.CANCELLED,

colorscale='Portland',

showscale=True

)

)

data = [trace2]

layout = go.Layout(xaxis=dict(tickangle=20),

title='Cancellation Rate by Cities',

yaxis = dict(title = 'Cancellation Rate'

)

)

fig = go.Figure(data=data, layout=layout)

py.iplot(fig)

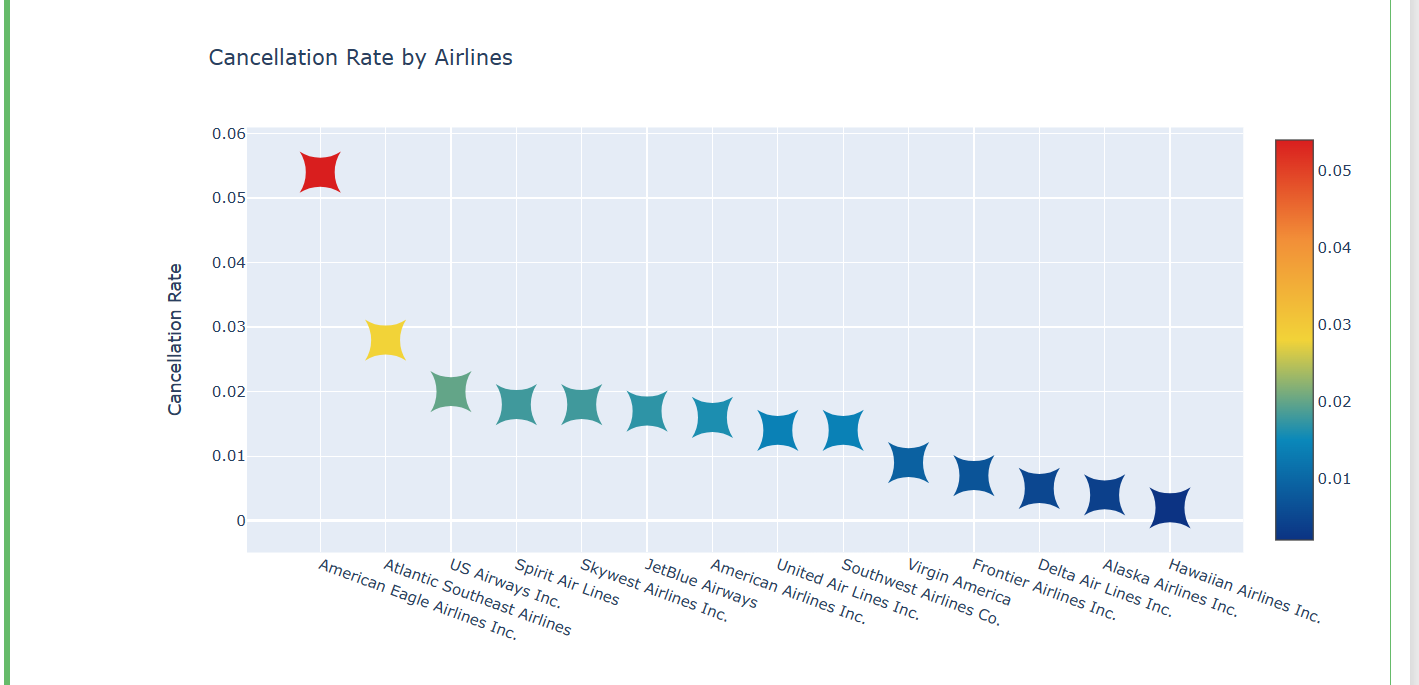


Figure Cancellation Rate by Airlines

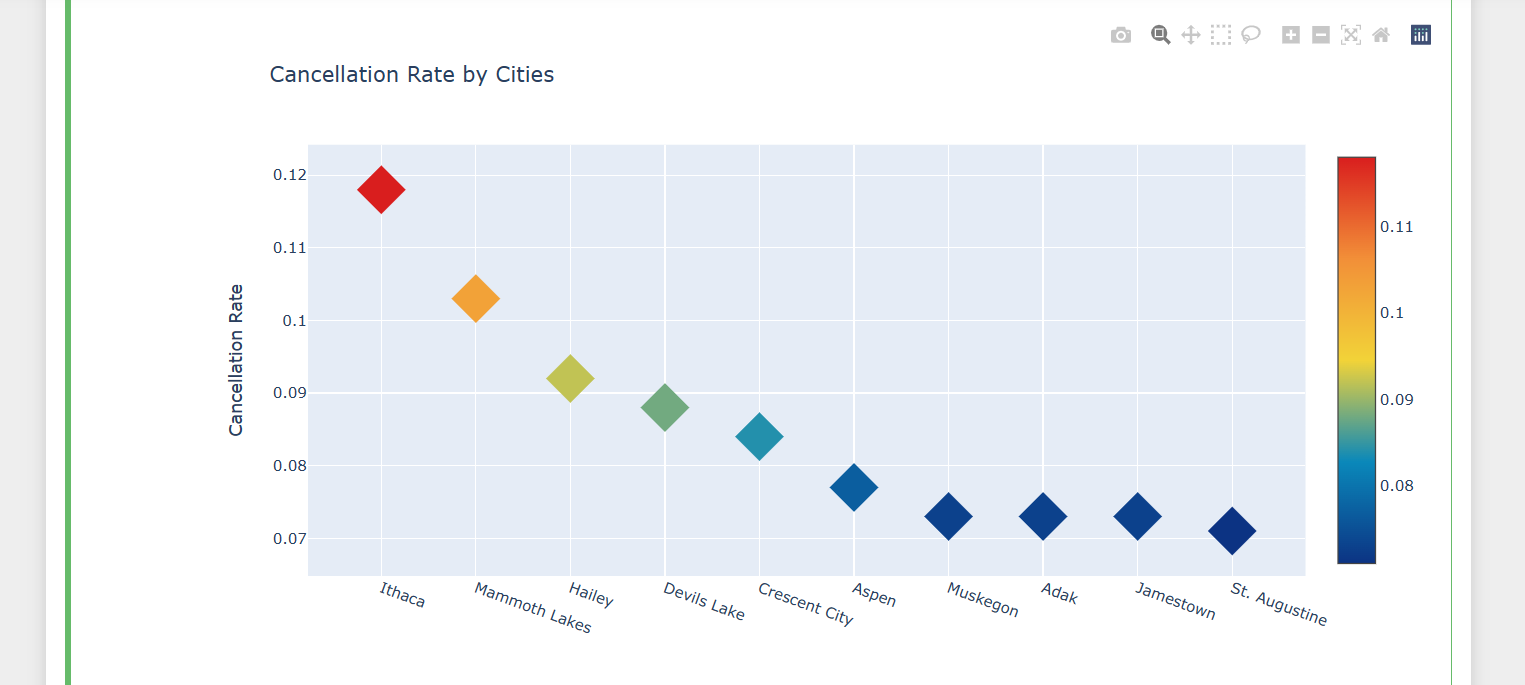


Figure Cancellation Rates by Cities

Results:

This analysis shows the flights canceled by airlines and in which city the most flights were canceled in the year 2015. There can be many reasons for cancellation of flights such as weather issues, poor management of airlines technical issues of aircraft, etc. American Eagle Airlines has the highest rate of cancellations compared to other flights while Hawaiian Airlines is most punctual with their operation. The airlines that are poor in their operating of flights can use this analysis and they can focus on the regular operation of their flights to enhance the customer experience and improve their business.

# 5.0 Project Setup

This notebook can be run by utilizing an environment that can run Python Notebooks. This is the underlying technology that provides a convenient way to combine Python code with markdown text into a single canvas. The advantage of a Python notebook is that it allows the user to selectively run and modify parts of code easily, without needing to run the program in its entirety. In addition, users can embed formatted text and figures making it easy for others to read and modify code directly. I used a notebook using an anaconda environment.

Jupiter notebook is available on the anaconda environment, I installed the anaconda environment and ran the notebook on that environment. And I installed all required packages in the Anaconda command interface.[12] Here I am attaching my github link where I uploaded my code.[13]

## 5.1 Python Libraries

Several Python libraries were utilized as part of the project. Numpy library was imported to analyze numerical data, and Pandas was used to perform data science tasks. NumPy aims to provide an array object faster than traditional Python lists[14]. I also used the pandas library. Pandas allow us to analyze big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant.[15] Matplotlib library was used to plot the data. Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible. [16]

# 6.0 Challenges

The First and major challenge of the project was the size of the dataset. This was a huge dataset. It was a difficult task to manage this much data. Apart from that the data was not properly organized and had many null values. It takes too much memory to perform analysis on such a big dataset. Also, the column names were in codes and the dataset values were also in code. For the analysis to be understood by everyone the names have to be renamed. Also, Any inconsistency in the way the data is nested would result in no data being transmitted from the CSV files to the database. The analysis became frustrating at first because the system crash issues were there because of the size of the dataset.

# 7.0 Future work

In this project, I analyzed the dataset and visualized the data so it can be understood by passengers, airline companies, and also government. They can use it to save their time, and money, improve efficiency, and increase their business. In future the additional datasets can be used which are available on OST website to get most recent data. In the future, a web application can be made so that anyone can put their filters and sorting to find the information they are looking for.

# 8.0 Conclusion

Researchers and analysts can use this data to find patterns, analyze trends over time, and measure the performance of individual airports and airlines. A comprehensive assessment of changes affecting flight time is provided by key indicators such as arrivals, on-time arrivals, late arrivals, cancellations and delays, affecting many things such as flight delay, weather delay, security delay and aircraft delay.

This analysis is a great tool for anyone involved in the aviation industry as it provides practical advice on how to improve operational efficiency and contributes to discussions on how to improve travel weather. This analysis provides a solid basis for mining data on the complexity of the aviation industry, or even checking the performance of a particular carrier, measuring the difference in time or understanding the effects of weather conditions on delays.

# References:

[1] S. V. S. A. MALLADI, “Analysis and Visualization of Airline Flight Delays Data: Unveiling Patterns and Insights,” Medium. Accessed: Dec. 06, 2023. [Online]. Available: https://medium.com/@msvs.akhilsharma/analysis-and-visualization-of-airline-flight-delays-data-unveiling-patterns-and-insights-eab86e639905

[2] “OST\_R | BTS | Transtats.” Accessed: Dec. 09, 2023. [Online]. Available: https://www.transtats.bts.gov/Homepage.asp

[3] T. Dey, D. Phillips, and P. Steele, “A Graphical Tool to Visualize Predicted Minimum Delay Flights,” *Journal of Computational and Graphical Statistics*, vol. 20, Jun. 2011, doi: 10.2307/23110488.

[4] “Data Analysis with R,” Coursera. Accessed: Dec. 06, 2023. [Online]. Available: https://www.coursera.org/learn/data-analysis-with-r

[5] Y. Jiang, Y. Liu, D. Liu, and H. Song, “Applying Machine Learning to Aviation Big Data for Flight Delay Prediction,” in *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, Aug. 2020, pp. 665–672. doi: 10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00114.

[6] S. Manna, S. Biswas, R. Kundu, S. Rakshit, P. Gupta, and S. Barman, “A statistical approach to predict flight delay using gradient boosted decision tree,” in *2017 International Conference on Computational Intelligence in Data Science(ICCIDS)*, Jun. 2017, pp. 1–5. doi: 10.1109/ICCIDS.2017.8272656.

[7] “Airline Delay and Cancellation Data, 2009 - 2018.” Accessed: Dec. 06, 2023. [Online]. Available: https://www.kaggle.com/datasets/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018

[8] “RITA | BTS | Transtats.” Accessed: Dec. 06, 2023. [Online]. Available: https://www.transtats.bts.gov/DataIndex.asp

[9] “2015 Flight Delays and Cancellations.” Accessed: Dec. 06, 2023. [Online]. Available: https://www.kaggle.com/datasets/usdot/flight-delays

[10] “Data Visualization in jupyter notebook,” GeeksforGeeks. Accessed: Dec. 06, 2023. [Online]. Available: https://www.geeksforgeeks.org/data-visualization-in-jupyter-notebook/

[11] “Pyplot tutorial — Matplotlib 3.8.2 documentation.” Accessed: Dec. 06, 2023. [Online]. Available: https://matplotlib.org/stable/tutorials/pyplot.html

[12] “Jupyter :: Anaconda.org.” Accessed: Dec. 06, 2023. [Online]. Available: https://anaconda.org/anaconda/jupyter

[13] “Yashbulsara/Final-project,” GitHub. Accessed: Dec. 12, 2023. [Online]. Available: https://github.com/Yashbulsara/Final-project

[14] “Introduction to NumPy.” Accessed: Dec. 09, 2023. [Online]. Available: https://www.w3schools.com/python/numpy/numpy\_intro.asp

[15] “Pandas Introduction.” Accessed: Dec. 09, 2023. [Online]. Available: https://www.w3schools.com/python/pandas/pandas\_intro.asp

[16] “Matplotlib — Visualization with Python.” Accessed: Dec. 09, 2023. [Online]. Available: https://matplotlib.org/

# Appendix:

import pandas as pd

import numpy as np

import time

import matplotlib.pyplot as plt

import os

from numpy import random

csv\_path\_files\_list = []

for i in range(2009,2018):

csv\_path\_files\_list.append('Final\_project/'+ str(i) +'.csv')

csv\_list = []

for file in sorted(csv\_path\_files\_list):

csv\_list.append(pd.read\_csv(file).assign(File\_Name = os.path.basename(file)))

csv\_list\_delay = []

for csv\_list\_year in csv\_list:

csv\_list\_delay.append(csv\_list\_year[(csv\_list\_year['ARR\_DELAY'] > 0)])

del csv\_list

# Delay per month, mean and median value

for i in range(len(csv\_list\_delay)):

csv\_list\_delay[i]['FL\_DATE\_month'] = pd.to\_datetime(csv\_list\_delay[i]['FL\_DATE']).dt.month

plt.figure(figsize=(25, 12)).subplots\_adjust(hspace = 0.5)

plt.subplot(2, 2 ,1)

csv\_list\_delay[i].groupby('FL\_DATE\_month').ARR\_DELAY.mean().plot.bar()

plt.title('Mean ARRIVAL delays by month for ' + year[i], fontsize=16)

plt.ylabel('Minutes', fontsize=14)

plt.xlabel('month of the year', fontsize=14)

plt.subplot(2, 2 ,2)

csv\_list\_delay[i].groupby('FL\_DATE\_month').ARR\_DELAY.median().plot.bar()

plt.title('Median ARRIVAL delays by month for ' + year[i], fontsize=16)

plt.ylabel('Minutes', fontsize=14)

plt.xlabel('month of the year', fontsize=14)

plt.show()

# Delay per Destination, mean and median value

for i in range(len(csv\_list\_delay)):

plt.figure(figsize=(25, 12)).subplots\_adjust(hspace = 1)

plt.subplot(2, 2 ,1)

mean\_origin = csv\_list\_delay[i].groupby('DEST').ARR\_DELAY.mean().sort\_values(ascending=False)

mean\_origin[:20].plot.bar()

plt.title('Mean ARRIVAL delays by Destination for ' + year[i], fontsize=16)

plt.ylabel('Minutes', fontsize=14)

plt.xlabel('City', fontsize=14)

plt.subplot(2, 2 ,2)

median\_origin = csv\_list\_delay[i].groupby('DEST').ARR\_DELAY.median().sort\_values(ascending=False)

median\_origin[:20].plot.bar()

plt.title('Median ARRIVAL delays by Destination for ' + year[i], fontsize=16)

plt.ylabel('Minutes', fontsize=14)

plt.xlabel('City', fontsize=14)

plt.show()

# Different types of delays by year

delays = ['CARRIER\_DELAY', 'WEATHER\_DELAY', 'SECURITY\_DELAY', 'LATE\_AIRCRAFT\_DELAY']

for i in range(len(csv\_list\_delay)):

mean2 = csv\_list\_delay[i][['CARRIER\_DELAY', 'WEATHER\_DELAY', 'SECURITY\_DELAY', 'LATE\_AIRCRAFT\_DELAY']].mean()

X = np.arange(len(delays))

fig = plt.figure()

plt.bar(X - 0.2, mean2, color = 'b', width = 0.40, label = 'Mean')

plt.xticks(X, delays, rotation = 45)

plt.title('Delays per type for ' + year[i], fontsize=16)

plt.ylabel('Minutes', fontsize=14)

plt.xlabel('Delay', fontsize=14)

plt.legend()

plt.show()

# Delays by City

city\_by\_delay = df.groupby('ORIGIN').ARR\_DELAY.sum().sort\_values(ascending=False)

plt.figure(figsize=(20, 6))

city\_by\_delay[:15].plot.bar()

plt.title('Delays by City', fontsize=16)

plt.xlabel('City', fontsize=14)

plt.ylabel('Hours', fontsize=14)

plt.show()

#most unrealiable day to fly in last 3 years

df = pd.read\_csv('Final\_project/2018.csv')

df = df[df['ORIGIN'].isin(['ATL','LAX','ORD','DFW','DEN','JFK','SFO','SEA','LAS','MCO'])]

df\_2017 = df.append( pd.read\_csv('Final\_project/2017.csv'))

df\_2017 = df\_2017[df\_2017['ORIGIN'].isin(['ATL','LAX','ORD','DFW','DEN','JFK','SFO','SEA','LAS','MCO'])]

df = df.append(df\_2017)

df\_2016 = df.append( pd.read\_csv('Final\_project/2016.csv'))

df\_2016 = df\_2016[df\_2016['ORIGIN'].isin(['ATL','LAX','ORD','DFW','DEN','JFK','SFO','SEA','LAS','MCO'])]

df = df.append(df\_2016)

import pandas as pd

import matplotlib.pyplot as plt

# Assuming you have the Kaggle flight data CSV file

kaggle\_data\_path = 'Final\_project/2022.csv'

# Load the data into a Pandas DataFrame

flight\_data = pd.read\_csv(kaggle\_data\_path)

# Convert the 'FL\_DATE' column to datetime

flight\_data['FL\_DATE'] = pd.to\_datetime(flight\_data['FL\_DATE'])

# Filter data for the year 2022

flight\_data\_2022 = flight\_data[flight\_data['FL\_DATE'].dt.year == 2022]

# Group by destination and calculate the average departure delay

average\_delay\_by\_destination = flight\_data\_2022.groupby('DEST')['DEP\_DELAY'].mean().reset\_index()

# Select the top 10 destinations based on average departure delay

top\_10\_destinations = average\_delay\_by\_destination.nlargest(10, 'DEP\_DELAY')

# Display the result

print(top\_10\_destinations)

# Create a bar graph for the top 10 destinations

plt.figure(figsize=(12, 6))

plt.bar(top\_10\_destinations['DEST'], top\_10\_destinations['DEP\_DELAY'], color='skyblue')

plt.xlabel('Destination')

plt.ylabel('Average Departure Delay (minutes)')

plt.title('Top 10 Destinations by Average Departure Delay for Year 2022')

plt.xticks(rotation=45, ha='right') # Rotate x-axis labels for better readability

# Show the plot

plt.tight\_layout()

plt.show()

dff = df.CITY\_x.value\_counts()[:10]

trace = go.Bar(

x=dff.index,

y=dff.values,

marker=dict(

color = dff.values,

colorscale='Jet',

showscale=True

)

)

data = [trace]

layout = go.Layout(

title='Origin City Distribution',

yaxis = dict(title = '# of Flights')

)

fig = go.Figure(data=data, layout=layout)

py.iplot(fig)

dff = df.AIRLINE.value\_counts()[:10]

trace = go.Bar(

x=dff.index,

y=dff.values,

marker=dict(

color = dff.values,

colorscale='Jet',

showscale=True)

)

data = [trace]

layout = go.Layout(xaxis=dict(tickangle=15),

title='Airline distribution',

yaxis = dict(title = '# of Flights'))

fig = go.Figure(data=data, layout=layout)

py.iplot(fig)

df['dep\_delay'] = np.where(df.DEPARTURE\_DELAY>0,1,0)

df['arr\_delay'] = np.where(df.ARRIVAL\_DELAY>0,1,0)

dff = df.groupby('MONTH').dep\_delay.mean().round(2)

dff.index = dff.index.map(month)

trace1 = go.Bar(

x=dff.index,

y=dff.values,

name = 'Departure\_delay',

marker = dict(

color = 'aqua'

)

)

dff = df.groupby('MONTH').arr\_delay.mean().round(2)

dff.index = dff.index.map(month)

trace2 = go.Bar(

x=dff.index,

y=dff.values,

name='Arrival\_delay',

marker=dict(

color = 'red'

)

)

data = [trace1,trace2]

layout = go.Layout(

title='% Delay (Months)',

yaxis = dict(title = '%')

)

fig = go.Figure(data=data, layout=layout)

py.iplot(fig)

dff = df.groupby('AIRLINE')[['CANCELLED']].mean().sort\_values(by='CANCELLED',

ascending=False).round(3)

trace1 = go.Scatter(

x=dff.index,

y=dff.CANCELLED,

mode='markers',

marker=dict(

symbol = 'star-square',

sizemode = 'diameter',

sizeref = 1,

size = 30,

color = dff.CANCELLED,

colorscale='Portland',

showscale=True

)

)

data = [trace1]

layout = go.Layout(xaxis=dict(tickangle=20),

title='Cancellation Rate by Airlines', yaxis = dict(title = 'Cancellation Rate'

)

)

fig = go.Figure(data=data, layout=layout)

py.iplot(fig, filename="age")

dff = df.groupby('CITY\_x')[['CANCELLED']].mean().sort\_values(by='CANCELLED',

ascending=False)[:10].round(3)

trace2 = go.Scatter(

x=dff.index,

y=dff.CANCELLED,

mode='markers',

marker=dict(symbol = 'diamond',

sizemode = 'diameter',

sizeref = 1,

size = 30,

color = dff.CANCELLED,

colorscale='Portland',

showscale=True

)

)

data = [trace2]

layout = go.Layout(xaxis=dict(tickangle=20),

title='Cancellation Rate by Cities',

yaxis = dict(title = 'Cancellation Rate'

)

)

fig = go.Figure(data=data, layout=layout)

py.iplot(fig)

df['SPEED'] = 60\*df['DISTANCE']/df['AIR\_TIME']

dff = df.groupby('AIRLINE').SPEED.mean().to\_frame().sort\_values(by='SPEED',

ascending=False).round(2)

trace = go.Scatter(

x=dff.index,

y=dff.SPEED,

mode='markers',

marker=dict(

sizemode = 'diameter',

sizeref = 1,

size = 30,

color = dff.SPEED.values,

colorscale='Jet',

showscale=True

)

)

data = [trace]

layout = go.Layout(xaxis=dict(tickangle=-20),

title='Mean Speed by Airlines',

yaxis = dict(title = 'Speed')

)

fig = go.Figure(data=data, layout=layout)

py.iplot(fig)

import pandas as pd

import matplotlib.pyplot as plt

# Create an empty list to accumulate data

all\_mean\_delays = []

# Extract unique years

unique\_years = set()

# Extract mean delays by year

for i in range(len(csv\_list\_delay)):

csv\_list\_delay[i]['FL\_DATE\_year'] = pd.to\_datetime(csv\_list\_delay[i]['FL\_DATE']).dt.year

unique\_years.update(csv\_list\_delay[i]['FL\_DATE\_year'])

mean\_delays = csv\_list\_delay[i].groupby('FL\_DATE\_year').ARR\_DELAY.mean()

all\_mean\_delays.append(mean\_delays)

# Combine data and plot a bar graph

combined\_mean\_delays = pd.concat(all\_mean\_delays, axis=1)

combined\_mean\_delays.columns = sorted(unique\_years) # Sort the columns

combined\_mean\_delays.plot(kind='bar', figsize=(12, 6), width=4.5)

plt.title('Mean ARRIVAL delays by year', fontsize=16)

plt.ylabel('Minutes', fontsize=14)

plt.xlabel('Year', fontsize=12)

tick\_positions = range(len(combined\_mean\_delays))

plt.xticks(tick\_positions, combined\_mean\_delays.index, rotation=0)

plt.legend(title='Year')

plt.show()