Flight Big Data Analysis: Turning Data Into Insights

CMPE 297 Big Data Project Report

Team-1

Advisor – Prof. Weider D. Yu

Poojitha Amin
011811306
poojitha.amin@sjsu.edu

Sneha Vadakkemadathil
011810721
sneha.vadakkemadathil@sjsu.edu

Darshit Thesiya
011424647
darshit.thesiya@sjsu.edu

Vikas Miyani
011410152
vikas.miyani@sjsu.edu

Abstract— **The world of business is undergoing a data driven revolution using analytics to guide decision-making. In the recent days, big data is beginning to have a major impact on air travel with more data being created through the increase in the number of planes departing and landing, plane sensors and the passengers on board; the opportunities to use this data will only increase. It provides innovative companies with the opportunity to improve major aspects of their business, from using the data to improve customer retention through to making planes safer, more reliable and scheduled on time. The issue of airline delays, measured by the number of late arrivals, as a percent of total operations, has been of increasing importance in recent years, as more of the US population chooses air travel as the preferred mode of transportation. This analysis majorly focusses on airline delay and its causes. The applications for big-data analysis in airspace system performance and safety optimization have high potential because of the availability and diversity of airspace related data. The variability of underlying data warehouse can be leveraged using virtualized cloud infrastructure for scalability to identify trends and create actionable information. This project discusses the concepts of big data, big data definitions and the approach towards developing and using big data infrastructure for analyzing aviation data. It also introduces the data sources, nature of data collected, data loading, the tools used for the visualizations.**

Keywords—Big Data; Delay Prediction; Flight Dataset; Data Analysis;

I. INTRODUCTION

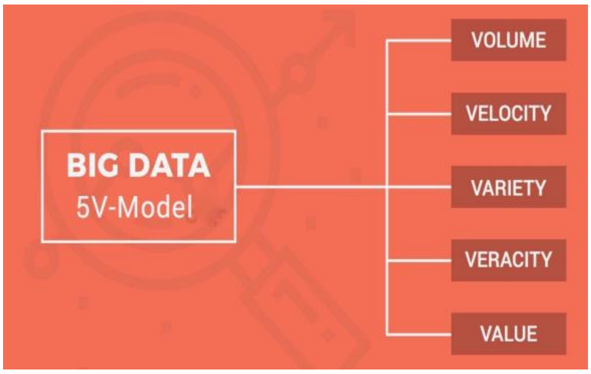
Airlines, airports, aircraft manufacturers, suppliers, governments and others in the global aviation space depend on data for operational planning and execution. Complex and concurrent data sets create immense technical and human challenges in collecting, sorting, and mining aviation databases. Aviation data sets exceed the capabilities of desktop computing. Big-data analytics provides the aviation industry scalability, extensibility, and query capability through cloud based database architecture.

Once the data is gathered, the next challenge we had is analytics. Aviation data sets go beyond desktop capability, requiring time-consuming manual slicing of data. Application of big-data analytical methods, data warehousing and software solutions for fast-response data mining can address these problems. The unprecedented variety, longevity, and detail of aviation data we acquired, generated, stored and managed provided unique capabilities and value to our analysis.

II. BIG DATA AND ITS 5V'S

The term “big-data” can be defined as data that becomes so large that it cannot be processed using conventional methods. The size of the data which can be considered to be Big Data is a constantly varying factor and newer tools are continuously being developed to handle this big data. In order to make sense out of this overwhelming amount of data it is broken down into five V’s: Velocity, Volume, Value, Variety and Veracity.

The big data used for flight data analysis is also broken down into five dimensions and each dimension is explained to understand its significance and impact.



A. Volume

In the big data pyramid, volume is the base. The vast amounts of data we have becomes so large in fact, that we can no longer store and analyze data using traditional database technology. The data in this project includes historical and near real-time data and since the data is collected from the year 1987, the volume sky rockets to millions of records. We have a base of 40 GB of processed data. Collecting and analyzing this data is clearly an engineering challenge of immensely vast proportions.

B. Velocity

Velocity refers to the speed at which vast amounts of data are being generated, collected and analyzed. Not only must it be analyzed, but the speed of transmission, and access to the data must also remain instantaneous to allow for near real-time access. Big data technology allows us now to analyze the data, without ever putting it into databases.

C. Variety

Variety is defined as the different types of data we can now use. Flight data is available from multiple sources. In this project, we have made use of new and innovative big data technologies to allow data to be harvested, stored, and used simultaneously.

D. Veracity

Veracity is the quality or trustworthiness of data. Gleaning volumes of data are of no use if the quality or trustworthiness of data is not accurate. To focus on quality, it was important to set metrics around what type of data we have collected and from what sources.

E. Value

Value refers to the worth of data being extracted. Having endless amounts of data does not always translate into having high value data. The key is to ensure that the data being collected can be turned into value, as quickly and as cost effectively as possible.

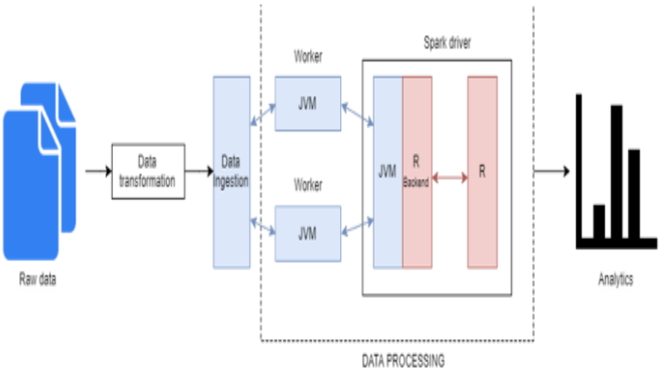
III. USING BIG DATA TO SOLVE PROBLEMS

The massive datasets big data aims to collect, cultivate and, ultimately, mine, can be a source of valuable information for governments, companies and everything in between. In fact, big data is being sought as a solution to all kinds of problems that extend well beyond the tech realm, over even the business realm. Pattern recognition and anomaly detection can identify fraud, waste, and abuse in healthcare domain. Credit card fraud detection in milliseconds helps financial services firms protect their customers' security while reducing loss due to fraud. Big data helps instantly predict market trends and customer needs.

Big data analytics can play the biggest role is dealing with the unpredictable problems, the airline industry faces daily. With hundreds of planes, thousands of flights, and millions of employees and passengers, there is now too much data and too many variables for humans to sort through fast enough to fix problems or even prioritize potential threats. From big events like hurricanes and snowstorms to smaller disruptions like air traffic control delays, mechanical failures, or even lines of thunderstorms, analytics are necessary to streamline the system. While much of this activity today is mostly reactive, the next step will be for aviation to proactively avoid some of the delays, congestion, and inefficiencies that annoy passengers and keep the global industry at single-digit profit margins. Through this project, we are aiming to solve this problem by analyzing the influence of various factors that cause a delay in the airline operations. We are using predictive analytics to look at what has happened to date, and to reliably forecast what will happen next by applying statistical models. Predictive analysis improves agility, accuracy and helps in dealing with absent data.

IV. APPLICATION ARCHITECTURE

The application architecture is premised on a skill set of developing reliable and scalable data pipelines. The architecture includes components for data cleaning, transformation, ingestion, data processing, analytics and visualization. The below diagram shows the stack diagram of our project architecture.



The data pipeline takes raw data and converts it into insight. Data transformation includes cleaning of raw data and conversion to parquet file format. Parquet is a format that can be processed by a number of different systems it also doesn't lock into a specific programming language. The architecture consists of an R to JVM binding on the driver that allows R program to interact with Spark. Spark SQL provides inbuilt support for parquet format. Spark streaming processes the incoming data streams in-memory, performing aggregations, count and simple checks. Spark Streaming is able to process 100,000-500,000 records/sec. Along the way decisions are made on what happens to the data, how it is stored, what tools are used to process the data and eventually the manner of providing access to the outside world.

V. BIG DATA TOOLS USED

Following is the list of tools and technologies used in building our project:

A. R

R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical and graphical techniques, and is highly extensible. It provides the below features:

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,

- graphical facilities for data analysis and display either on-screen or on hard copy,
- a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

Below is a list of R packages used in the project:

- dplyr

  A fast and consistent tool for working with data frame like objects, both in memory and out of memory.
- sparklyr

  It is an R interface to Apache Spark, a fast and general engine for big data processing. The package supports connecting to local and remote Apache Spark clusters, provides a 'dplyr' compatible backend, and provides an interface to Spark's built-in machine learning algorithms. It provides functions like:

  a. filter() to select cases based on their values.
  b. arrange() to reorder the cases.
  c. select() and rename() to select variables based on their names.
  d. summarise() to condense multiple values to a single value.
  e. sample_n() and sample_frac() to take random samples.

- data. table

  It provides an extension of data.frame package. It helps with fast aggregation of data, fast ordered joins, fast add/modify/delete of columns by group using no copies at all, list columns, a fast and friendly file reader and parallel file writer.

- ggplot2

  All ggplot2 plots with a call to ggplot, supplying default data and aesthetic mappings, specified by aes. we then add layers, scales, coords and facets.

- maps

  This package is used for display and plotting of maps.

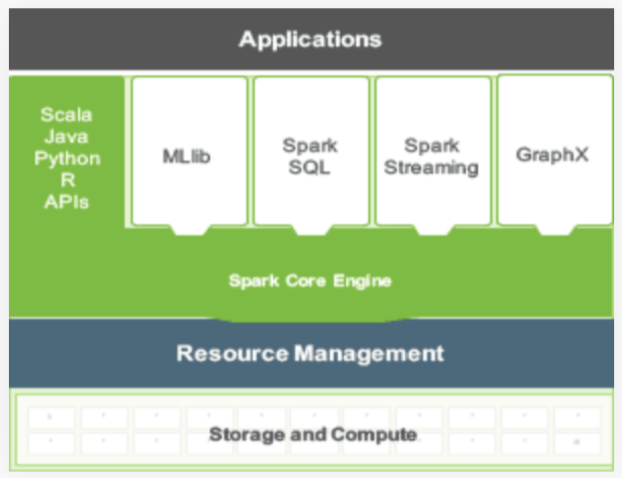
- yarrr

  This package contains a mixture of functions and data sets. The pirate plot function creates a pirate plot, a transparent plot for displaying continuous data as a function of 1, 2, or 3 discrete variables.

B. *Apache Spark*

Apache Spark is a fast, in-memory data processing engine with elegant and expressive development API's to allow data workers to efficiently execute streaming, machine learning or SQL workloads that require fast, iterative access to datasets. With in-memory processing, we can increase the processing speed. It uses the concept of a Resilient Distributed Dataset (RDD), which allows it to transparently store data on memory and persist it to disc only it's needed. This helps to reduce most of the disc read and write – the main time-consuming factors – of data processing.

We have used Spark's power to enrich the data and derive insights. Spark comes packaged with higher-level libraries, including support for SQL queries, streaming data and graph processing. Spark provides fault tolerance and real-time stream processing.

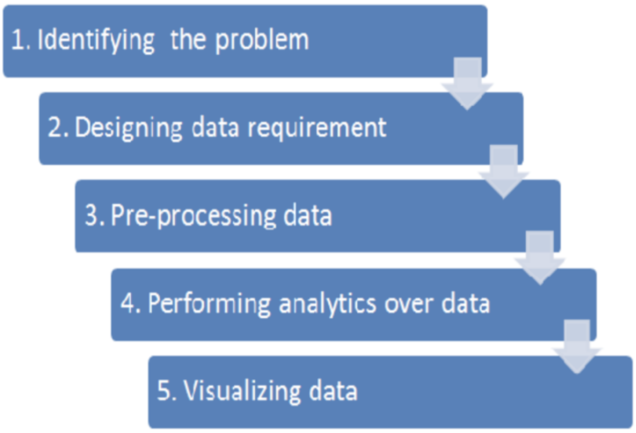


C. *Apache Zeppelin*

Apache Zeppelin, an open source data analytics and visualization platform. It is a web based notebook that enables interactive data analytics including Data Ingestion, Data Discovery, and Data Visualization all in one place. The notebook is integrated with distributed, general-purpose data processing systems such as Apache Spark. Zeppelin interpreter concept allows any language/data-processing-backend to be plugged into Zeppelin. Apache Zeppelin has an interactive interface that allows instant display of results of the analytics. We have made use of Apache Zeppelin to create our dashboard with the Spark interpreter for running R queries.

VI. FLIGHT DATA ANALYSIS

Big data analytics refers to the process of collecting, organizing and analyzing large sets of data to discover patterns and useful information. In this project, we performed analysis on the enormous flight data set provided by the United States Department of Transportation. We conducted our analysis by systematically following different phases of big data analytics life cycle.


1. Identifying the problem
2. Designing data requirement
3. Pre-processing data
4. Performing analytics over data
5. Visualizing data

A. *Identifying the problem*

Without delving into details about the nature of the challenge and existing sources of data, it is difficult to determine for sure if the problem is a Big Data problem. Following the initial analysis on the flight dataset provided by U.S transportation department, we concluded that the problem is a Big Data problem. We looked into the dataset to find unique characteristics of a Big Data problem and applied the 5 Vs screening question before coming to a conclusion. Our analysis plan includes complex questions of the data, such as determining cause and effect of airline delay. Even though the data is collected from a single system, it consists of many different variables with unclear relation to one another.

B. *Data Collection*

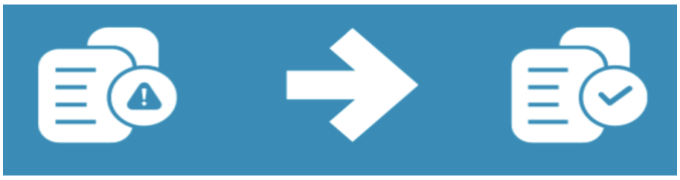
Data collection is the process of gathering information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. It is the first step in big data analysis. Data collection plays the most important role in the Big Data cycle. Consequences from improperly collected data include inability to answer research questions accurately and distorted findings resulting in wasted resources. This data analysis project is to explore what insights can be derived from the enormous domestic airline data set provided by the United States Department of Transportation. The airline data set consisting of details of all commercial flights within the USA dating from 1987 till date is obtained from the Research and Innovative Technology Administration (RITA) which coordinates the U.S. Department of Transportation research programs. The data was obtained from 29 commercial airlines and 3,376 airports, and consists of nearly 360 million records. This accounts for more than 40 GB of uncompressed data storage.

C. *Identifying the Scope of Analysis*

Assessing the scope of analysis or identifying the needs will help determine where exactly to focus on the data cleaning efforts. This makes it a crucial step to be performed between data collection and data cleaning. We set the scope of our analysis by focusing on the overall objectives of our project and finding answers to questions such as – what do we want to get out of the data? and what questions do we want to answer? After identifying the objectives of our project, we established the key data necessities. Then we compared the list of objectives with the actual dataset to see what pieces of data can be used. This helped us to identify the areas that need to be fixed during data cleansing.

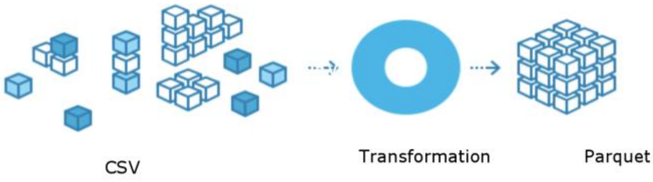
D. *Data Cleaning and Organization*

Data cleaning is the process of detecting, diagnosing, and editing faulty data. One of the biggest issues with the big data is inaccurate, erroneous, or otherwise incomplete data. While this may seem like a tiny blip in the realm of big data and analytics, it is possibly the biggest barrier in enhancing overall data insight capabilities. The first step in data analysis after data acquisition is to improve data quality. This is the most critical step in the data value chain; even with the best analysis, inaccurate data will generate wrong results. Analyzing big data will result in erroneous conclusions unless necessary steps are taken to validate and clean the data. Airline data set needs to be pre-processed prior to analysis for reliable integration and effective analysis results. This step includes removing irrelevant data points and retaining only the required ones with respect to our analysis. We also cleansed the airline data to get rid of NA values and recode a few columns for purpose of better understanding.



E. *Converting Data to a Partitioned Parquet File*

After pre-processing, we organized the data to parquet compressed file format. Parquet is a compressed columnar file format.



We chose parquet format because our data is huge in volume and parquet improves efficiency in terms of storage and performance. The two main advantages of a columnar format are that queries will de-serialize only that data which is actually needed, and compression is frequently much better since columns frequently contained highly repeated values.

Columnar file formats greatly enhance data file interaction speed and compression by organizing data by columns rather than by rows. A partition is a subset of the data that all share the same value for a particular key. Parquet files can create partitions through a folder naming strategy.

F. Data Attributes

The major data attributes that we focused on for analysis are:

- Year: The year in which the flight took place.
- Month: The month in which the flight took place.
- Day of Month: The numerical day of the month in which the flight took place.
- Day of week: The day of the week in which the
- flight took place.
- Departure Time: Actual departure time of the flight.
- Arrival Time: Actual arrival time of the flight.
- CRS Departure Time: Scheduled departure time.
- CRS Arrival Time: Scheduled arrival time.
- Carrier: The initials of the airline corresponding to
- the flight.
- Origin: The airport code for the airport that the flight
- is departing from.
- Destination: The airport code for the airport that the flight is departing to.
- Airport: Airport name for a code.
- City: The city the airport belongs to.
- State: The state the airport belongs to.
- Cancelled: was the flight cancelled?
- Cancellation Code: reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
- Arrival Delay: arrival delay, in minutes.
- Departure Delay: departure delay, in minutes.

G. Data Analysis

Data Analysis is the process of systematically applying statistical and/or logical techniques to describe, illustrate and evaluate data. It is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the help of specialized systems and software. This stage can be iterative in nature, especially if the data analysis is exploratory, in which case analysis is repeated until the appropriate pattern or correlation is uncovered. Exploratory analysis helps to understand the causes of an observed event, such as airline delay, the nature of the data and the key features in the data needed for analysis. We performed analysis by manipulating the data in a number of ways such as plotting it out and finding correlations. Data analysis was performed in R software environment. R provides an integrated suite of software facilities for data manipulation, calculation and graphical display. We used graphical techniques like scatter plots, box plots, multiple-line graphs and quantitative techniques like mean and median.

H. Data Visualization

Data visualization help understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization. These illustrations deviate from the use of hundreds of rows, columns and attributes toward a more artistic visual representation of the data. R Programming offers a satisfactory set of inbuilt function and libraries (such as ggplot2, leaflet, lattice) to build visualizations and present data. We used R software coupled with zeppelin for providing visualization to our analysis results.
Zeppelin has a built-in query and visualization tool. Another feature of Zeppelin is the support of dynamic forms to integrate user input into the notebooks. For example, we could develop an interactive query where the airport and day of the week are configurable by the user.

VII. *business use-cases*

On-time performance of airlines schedule is a key factor in maintaining current customer satisfaction and attracting new ones. However, flight schedules are often subjected to irregularity. Due to the tight connection among airlines resources, these delays could dramatically propagate over time and space unless proper recovery actions are taken.

A. Delay Cause Analysis

Our analysis model can be used by airline companies to assess the following:

- The major reasons for arrival and departure delays.
- The distribution of various factors causing the delay.
- Airport’s schedule performance and how it affects arrival delay.
- The relationship between arrival delay and factors such as time of a day, distance, weather and months of a year.

This would help the companies to improve future performance by taking actions such as improving air traffic management or ability of airplanes to operate in poor weather conditions.

B. Performance Analysis

Our analysis model can help the government to find insights about the following:

- On-time performance trend of large domestic airlines every year.

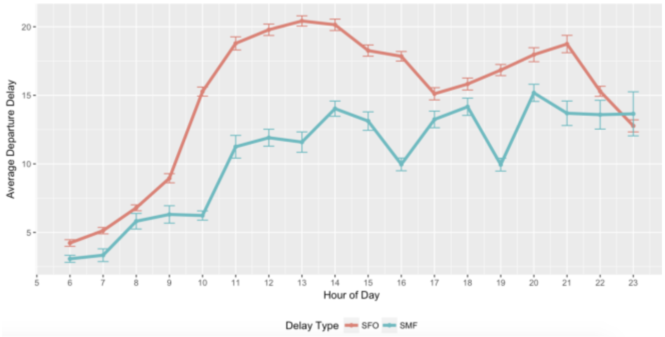
- Performance of different airports and its impact on overall delay causes.
- The average delay for each type of delay, for example weather, security, carrier and NAS (National Aviation System) delays.

VIII. *statistical analysis and visualizations*

A. Average delay pattern over the course of a day

The plot below shows that the delay at San Francisco airport is higher than the Sacramento airport. The graph also reveals the following trends:

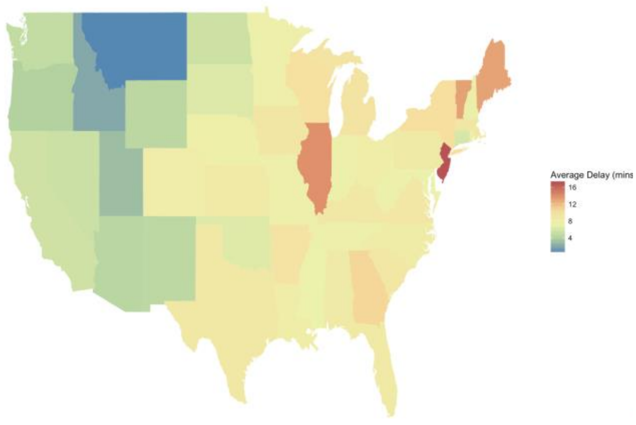
- Early morning flights are mostly on time
- As the day progresses and air traffic increases the average flight delay increases and it is at its peak during the noon time.



B. Average flight arrival delay by state

Next, we look at whether the airport is responsible for flight delays. Some airports are much busier and appear to have more frequent flight delays than the others. We created a plot for the States in the US where each state in colored based on its average flight delay.

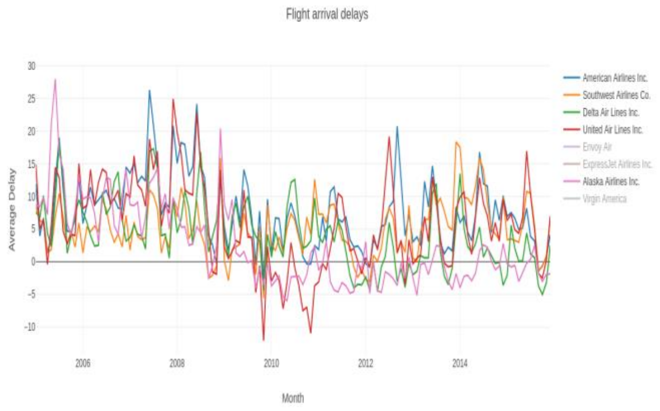
- Maximum flight delay is observed in the State New Jersey.
- Minimum flight delay is observed in States like Montana and Idaho.

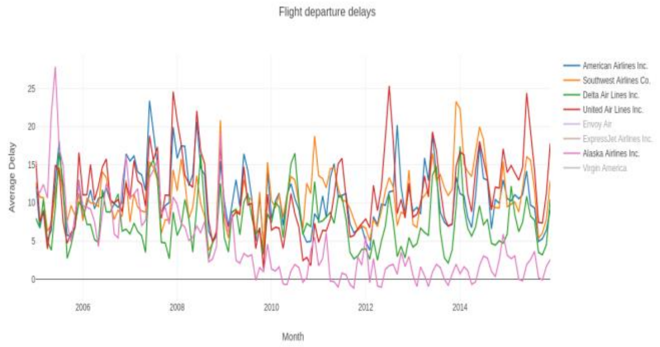


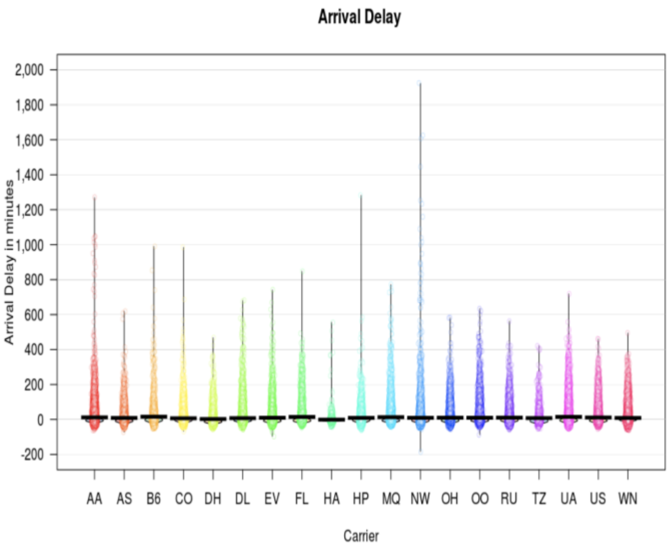
C. Average delay comparison by carrier

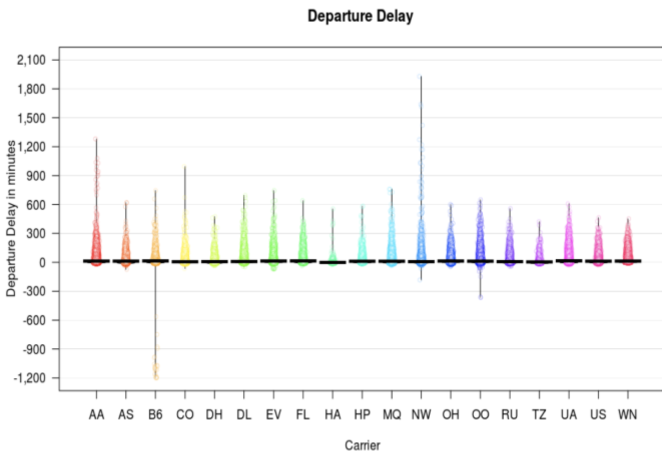
Flight performance may also be impacted by the quality of services provided by the airlines. Certain airlines are consistently good with on time performance.

- Leading the way with on time performance is Alaska airlines, followed by Delta airlines.
- Poor on-time performance is demonstrated by some airlines like American Airlines and United Airlines. This implies growing congestion at its service routes, signaling that it may need to boost capacity or add more routes to assuage this problem.





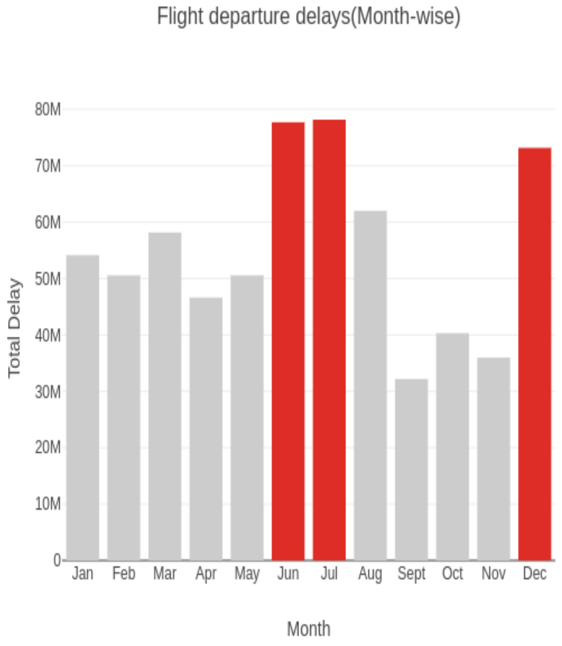


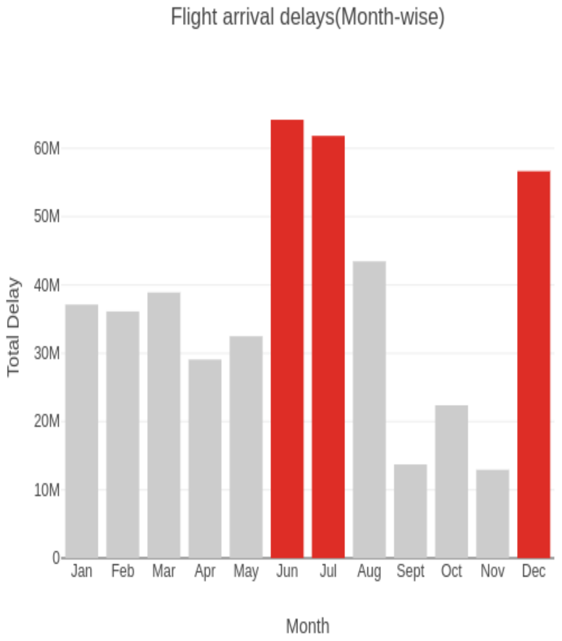


D. Average Flight delay (Month wise)

In order to get a seasonal pattern in delays we plotted the graph across the year. This shows that there is a peak in delays during the months of June, July and December. The delays generally fall down during the beginning of the year. The rationale as to why there is such high delays might be the following:

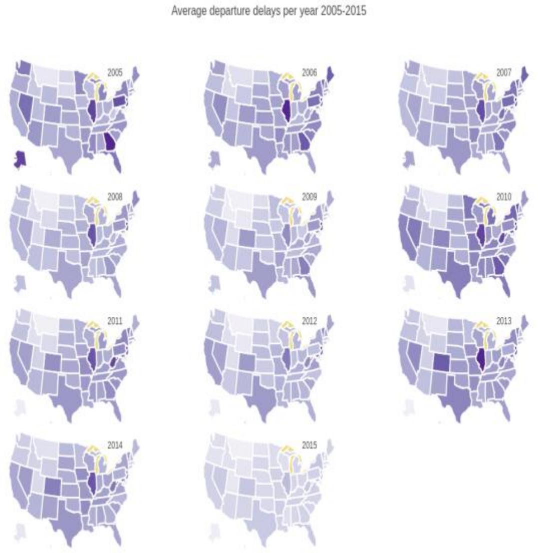
- Repair work after a hectic winter season
- A lot of passenger traffic for summer and winter vacation.

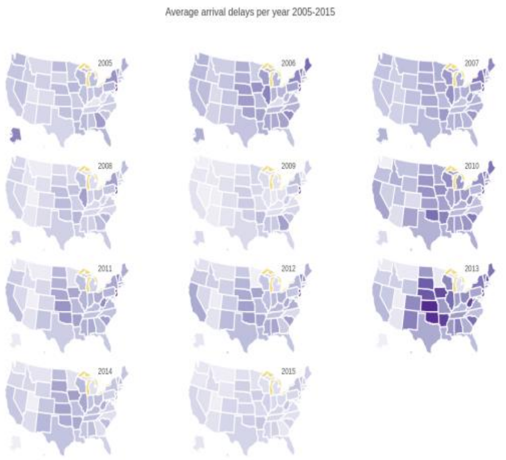




E. Year wise average delay for states

The plot below shows the year wise average departure and arrival delays of the States in US.





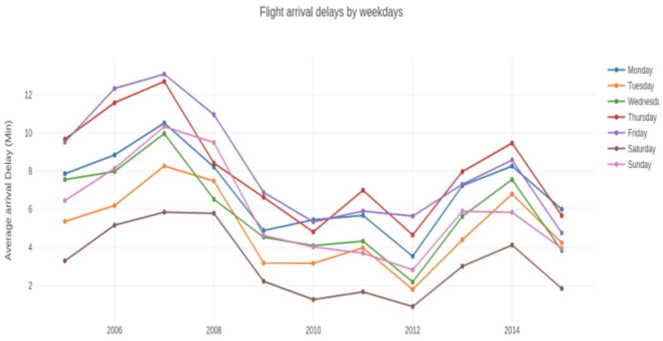
F. *Best days of the week to fly*

When it comes to analyzing the days of the week with minimum delays, Saturday reigns supremacy in this category and Friday is the worst day to travel.

- According to our data analysis, it is clear that there are fewer delays on Saturday and any other day of the week. This is because people avoid booking Saturday flights because they assume the crowds at the airport would be the same as any other museum, beach or park on a Saturday.
- Friday and Thursday, are the days when most flights are scheduled, meaning there is more opportunity for the delays to begin cascading throughout the airport.

In order from the least busy to most busy, there are typical delay percentages in airports:

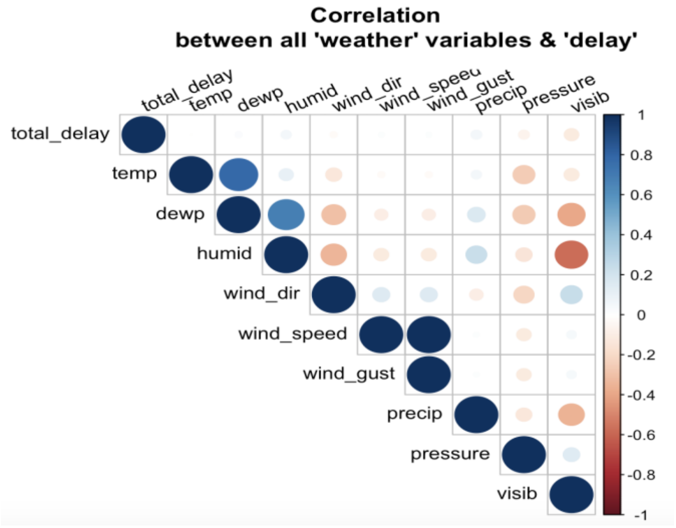
1. Saturday - 18.11%
2. Thursday - 21.69%
3. Wednesday - 23.59%
4. Sunday - 24.13%
5. Monday - 25.69%
6. Tuesday - 26.60%
7. Friday - 29.75%



G. *Correlation between weather and delay*

Flight delays are often linked to weather conditions. For this we merged the flight and weather data sets based on the columns like origin, year, month, day and hours.

- All the weather variables are involved in the delay caused.
- Total delay increases with increase in humidity and temperature. During summer, there is a higher humidity resulting in the occasional rains which is why there is a higher delay during that time.
- We can also use this information to predict the delay based on the weather variable.
- Weather variables include temperature, humidity, wind direction, wind speed, precipitation, pressure and visibility.



IX. FLIGHT DELAY PREDICTION

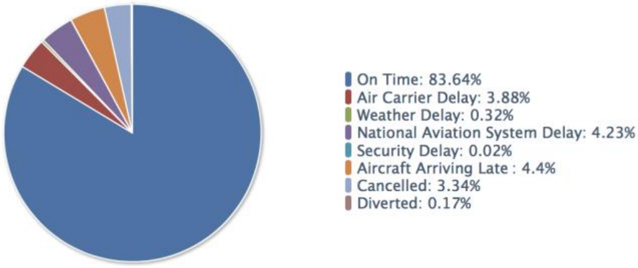
Flight delay hurt airlines, passengers and airports. Their prediction is crucial during the decision-making process for all players of commercial aviation. Predicting the delays accurately becomes all the more difficult because of the complexity of the air transportation system, the number of methods for prediction and the volume of data generated in the aviation industry. We used historical and predictive analysis to get valuable insights from airline dataset. Predictive analytics looks at what has happened to date, and applies statistical models to reliably forecast what will happen next. Our analysis on the flight data will help in making the following predictions.

- Identifying flights that are likely to be delayed.
- Identifying the months in which the probability for flight delays are more.
- Identifying the airlines that are more likely to perform better in future.
- Identifying the airports that are likely to have more flight delays in future.

X. SUMMARY AND CONCLUSION

Flight delays are important as it may increase the costs to customers and operational costs to airlines. It is crucial during the decision-making process for every player in the air transportation system. Each minute reduced per flight could save US$1.2 million in annual crew cost and US$5 million in annual fuel savings for a mid-sized airline. Multiply the savings across hundreds of airlines around the world and the potential savings are huge, besides greater efficiency for airlines and convenience for passengers. With how popular flying has become to every day person, trying to travel in the most efficient way has become extremely important. With the data available from United State Bureau of Transportation and Statistics, we were able to analyze it and come up with best and worst ways to travel the country.

In conclusion, we were able to find reasonable answers for each one of our questions. We found some carriers have much longer delays than others. We found a strong correlation between month, travel day and departure delay as well as finding which day and month is best to travel on. Another interesting find was the impact of weather variables on the delay. The graph below concludes are analysis on various factors impacting the delay in flights.



ACKNOWLEDGMENT

We thank Professor Weider D. Yu for his support, infinite insight and astute guidance in completing this project successfully. Working under him was a wonderful learning experience.

REFERENCES

[1] https://www.transtats.bts.gov
[2] https://www.r-project.org

[3] https://zeppelin.apache.org/docs/0.7.3