**Flight Delay Prediction and Analysis**

**Executive Summary**

By analyzing historical flight data and identifying patterns leading to delays using the dataset of 2015 Flight Delays and Cancellations, we develop a model to predict future delay occurrences and discover the flight delay rates of different airports.

The information we find in our analysis could provide significant insights for both business organizations and individuals in decision making, such as airplane company could optimize flight scheduling and improve customer satisfaction with accurate delay information, airport management organization could discover potential issues which caused the higher delay rate and develop solutions, individuals could use the information to do better travel planning and mitigate flight delay risks.

**Data Description**

1. Data source:

<https://www.kaggle.com/datasets/usdot/flight-delays/>

1. What the data are (what is measured by each variable, in what units, whether you will treat it numerical or categorical, etc.)

flights.csv:

YEAR:Year of the Flight Trip, numerical

MONTH:Month of the Flight Trip, numerical

DAY:Day of the Flight Trip, numerical

DAY\_OF\_WEEK:Day of week of the Flight Trip, numerical

AIRLINE:Airline Identifier, categorical

TAIL\_NUMBER:Aircraft Identifier, categorical

ORIGIN\_AIRPORT:Starting Airport, categorical

DESTINATION\_AIRPORT:Destination Airport, categorical

TAXI\_OUT(minutes):The time duration elapsed between departure from the origin airport gate and wheels off, numerical

SCHEDULED\_TIME(minutes):Planned time needed for the flight trip, numerical

ELAPSED\_TIME(minutes):AIR\_TIME+TAXI\_IN+TAXI\_OUT, numerical

AIR\_TIME(minutes):Time duration between wheels\_off and wheels\_on time, numerical

DISTANCE(miles):Distance between two airports, numerical

TAXI\_IN(minutes):The time duration elapsed between wheels-on and gate arrival at the destination airport, numerical

ARRIVAL\_DELAY(minutes):ARRIVAL\_TIME-SCHEDULED\_ARRIVAL, numerical

airports.csv:

IATA\_CODE: Location Identifier, categorical

AIRPORT: Airport's Name, categorical

CITY: City Name of the Airport, categorical

STATE: State Name of the Airport, categorical

COUNTRY: Country Name of the Airport

LATITUDE: Latitude of the Airport, numerical

LONGITUDE: Longitude of the Airport, numerical

1. Sample size (*n*) and number of variables (*k*)

flights.csv:

Number of observations: 5819079

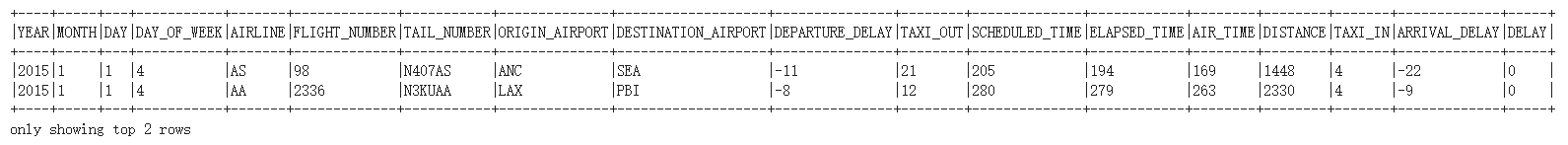
number of variables:15 (select 15 from 31 variables)

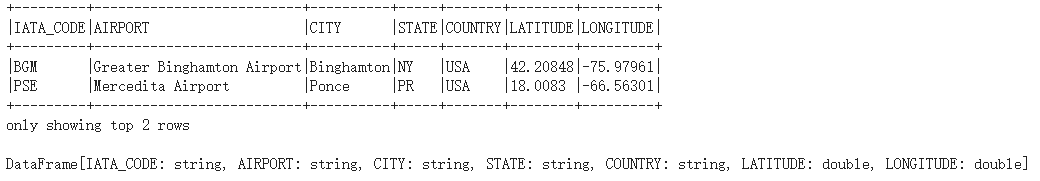
Airports.csv:

Number of observations: 322

number of variables:7

1. A small sample of observations from the data (a few observations to demonstrate the above points).





1. Why the data are of interest.

The 2015 Flight Delays and Cancellations dataset is valuable for several reasons. It helps airlines and airports identify patterns and prediction of delays, enabling proactive measures to mitigate future issues. Understanding these factors can improve customer satisfaction and loyalty, leading to a competitive advantage in the industry. PySpark's advanced machine learning algorithms enable predictive analytics, allowing airlines to make informed decisions about flight schedules, crew allocation, and maintenance planning. This data also aids in risk management and planning, driving innovation and service improvement. Graphframe and network analysis provide a unique perspective on network relationships, revealing systemic issues that disproportionately affect performance. Understanding delays and cancellations can help develop sustainable practices and aid in regulatory compliance. By leveraging PySpark's capabilities, stakeholders can derive actionable insights from the dataset, contributing to improved passenger experience, environmental sustainability, and economic viability.

# **Research Questions**

Using this dataset, we plan to investigate about 1) predict whether a flight will be delayed or not based on the information we have, 2) discover the flight delay rates for different airports. To dig deeper in this area, we should solve several problems as follows:

1. Determine the relevant factors that contribute to the arrival delay of flights.

Lots of factors can contribute to the delay of a flight, and in the dataset, there is too much information which is irrelevant to solving the question. After we learned about the meaning of each factor, we chose the factors that would give the most informative answers for our further question using our domain knowledge.

1. Predict whether a flight will be delayed or not.

By analyzing historical flight data, including flight date, airline company, aircraft type, departure and arrival airports, departure delay or not, distance of the flight, flight time etc., we could discover the pattern of delayed flights. Thus predict future delay occurrences using real-time data to make business decisions and improve customer satisfaction. Accurate flight delay predictions empower airlines and airports to proactively manage resources, adjust schedules, and communicate with passengers, enhancing overall operational resilience and customer satisfaction.

1. Discover the airports that have the most in and out flights.

We will Assess inbound and outbound flight volumes at different airports to consider the importance levels of airports. The larger flight volumes an airport possesses, the more important it is . Thus if any unforeseen situation occurs, more flights related to that airport would be affected. Besides, to calculate the flight delay rate of different airports, it is important to take into account how important the airport is. We will only take into consideration the airports with more than 10000 flights departing each year.

Insights into airport traffic dynamics inform route planning, infrastructure investments, and marketing strategies, helping airlines and airports optimize network connectivity and capitalize on emerging market opportunities.

1. Find out the airports with the relatively higher flight delay rates.

We will use the capacity of the airport and the number of delayed flights to work out the delay rate of each airport. This question is important because monitoring airport performance metrics enables stakeholders to identify areas for improvement, implement best practices, and enhance the overall passenger experience, driving loyalty and competitive advantage. Besides, it can give insights to airline companies to optimize their flight routines, for example, if a departure airport has a higher delay rate than a nearby airport, the airline companies can change the origin airport to one which has a lower delay rate to improve customer satisfaction and loyalty.

**Methodology**

1. Data Preparation

Our flight dataset contains 5819079 observations and is 564 MB in size. It is a relatively large dataset. If we choose to manually upload it to Google collab, it will take a lot of time. Therefore, we choose to utilize Kaggle API to directly fetch the dataset from kaggle. After we get the Kaggle API from kaggle account settings, we initialize the environment of KAGGLE\_USERNAME and KAGGLE\_KEY. Next, we can use “kaggle datasets download” to get the zipped datafile from the dataset. After that, we used ZipFile to unzip the dataset and get the csv files we needed.

1. Data Cleaning

In the original dataset “flights.csv”, flights are described according to 31 variables. Since the DEPARTURE\_TIME and ARRIVAL\_TIME variables don't contain the dates and we are only focusing on the ARRIVAL\_DELAY in our predictive modeling, we choose to drop unwanted variables like SCHEDULED\_ARRIVAL, ARRIVAL\_TIME, WHEELS\_OFF ,'WHEELS\_ON , SCHEDULED\_DEPARTURE, DEPARTURE\_TIME and so on. Moreover, we looked into the missing data and found that only a small proportion (105k, which makes 2% of total observations) of the observations have missing ARRIVAL\_DELAY data. These data cannot be filled with calculation of other variables related to flight time since these are also missing for these rows. Therefore, we chose to drop all the observations with missing data “ARRIVAL\_DELAY”.

1. Graph Analysis

The flight and airport dataset can be analyzed using the Pyspark Graphframe. We need to build the vertices and edges from the data file according to the Graphframe naming convention. The vertices in our project are the airports. We renamed the IATA\_CODE to id in our airport dataframe to get the vertices. For the edges, the src and dst are the ORIGIN\_AIRPORT and DESTINATION\_AIRPORT of the flight dataframe.

After we created the dataframe, we are able to perform filtering, groupBy and other queries on the nodes and edges to gain insights about the airports and the delay rates. In addition, outdegrees and indegrees can be calculated to see the rank of incoming and outgoing flights of airports. Moreover, we utilized the pagerank algorithm to rank the airports and created the network diagram of the Graphframe.

1. Predictive Modeling Preparation

A binary target variable, DELAY, is created to classify flights as delayed(1) if ARRIVAL\_DELAY is more than 15 minutes, and not delayed(0) otherwise. And categorical features such as AIRLINE, TAIL\_NUMBER, ORIGIN\_AIRPORT, and DESTINATION\_AIRPORT are indexed using ‘StingIndexer’ and encoded with ‘OneHotEncoder’ to prepare them for use in machine learning models. A ‘VectorAssembler’ is utilized to combine all the features into a single vector, ‘ all\_feacture’, which serves as the input for the predictive models.

1. Model Training and Evaluation

The dataset is split into training and testing sets to evaluate the model's performance on unseen data. We are going to use three different machine learning models that are trained and evaluated. First, Random Forest Classifier: A Random Forest model is trained using the features vector. Ramdo Forest is chosen for its ability to handle non-linear data and provide a robust model by averaging multiple decision trees. Second, the Logistic Regression model is also trained to predict flight delays. This model is beneficial for understanding the influence of different features on the probability of a flight being delayed. Next, A Decision Tree model is trained as it provides clear decision relies and is easy to interpret. This model can also capture non-linear patterns in the data. For each model’s performance is evaluated using the accuracy metric, calculated as the number of correctly predicted instances in the testing set. As a result, the accuracy of each model is reported, providing insights into which model performs best on the task of predicting flight delays based on the given features.

**Results and Finding**

**Flight Delay Prediction**

In our flight delay prediction analysis, we employed three distinct models: Logistic Regression, Decision Tree, and Random Forest. Our approach involved randomly partitioning the dataset, with 75% of observations allocated to the training data and the remaining 25% to the validation data. After training each model on the training data, we evaluated their performance on the validation set.

Among the models tested, the Logistic Regression Model exhibited the highest accuracy, achieving approximately 85%. Overall, our findings indicate the potential effectiveness of logistic regression for flight delay prediction.

|  |  |  |
| --- | --- | --- |
| Logistic Regression | Decision Tree | Random Forest |
| 0.849773404860478 | 0.8356746181957612 | 0.8211343261024866 |

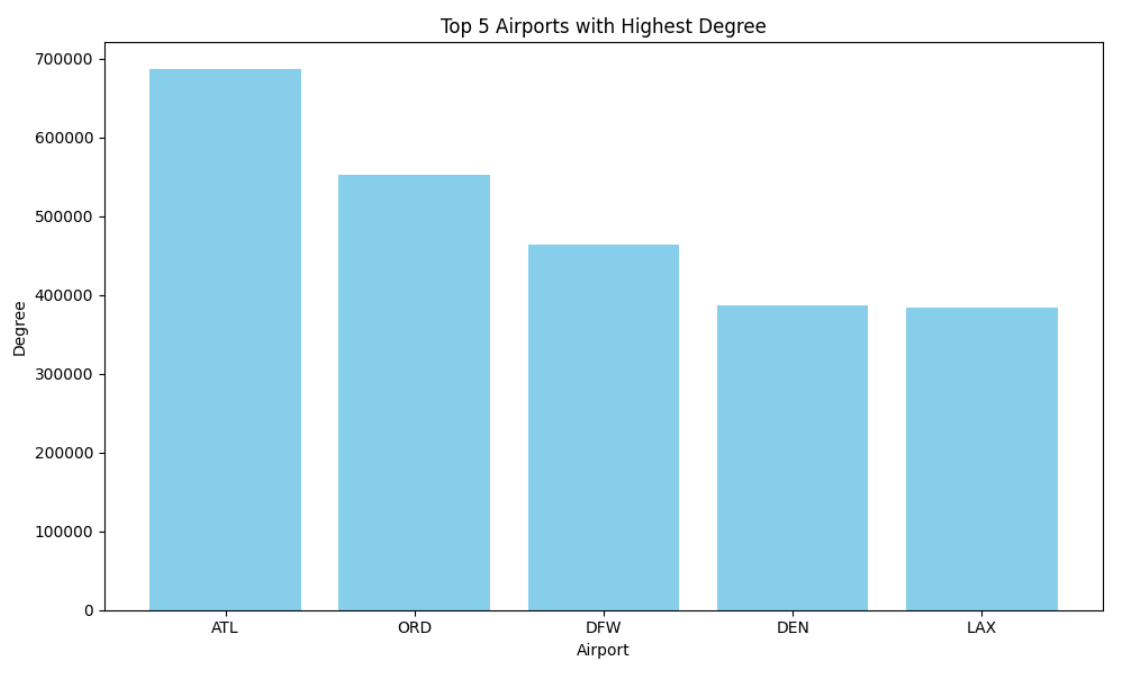
Using this model, and given the real-time data, we could predict the flight delay occurrences more confidently. This capability enables airlines and airports to proactively manage operations, allocate resources efficiently, and communicate with passengers effectively. By anticipating potential delays, airlines can adjust schedules, optimize crew assignments, and mitigate disruptions, ultimately enhancing the overall travel experience for passengers.

Moreover, real-time flight delay predictions empower travelers to make informed decisions, such as adjusting travel plans, exploring alternative routes, or allowing extra time for connections. This proactive approach minimizes inconvenience and stress associated with unexpected delays, contributing to a smoother and more enjoyable travel experience.

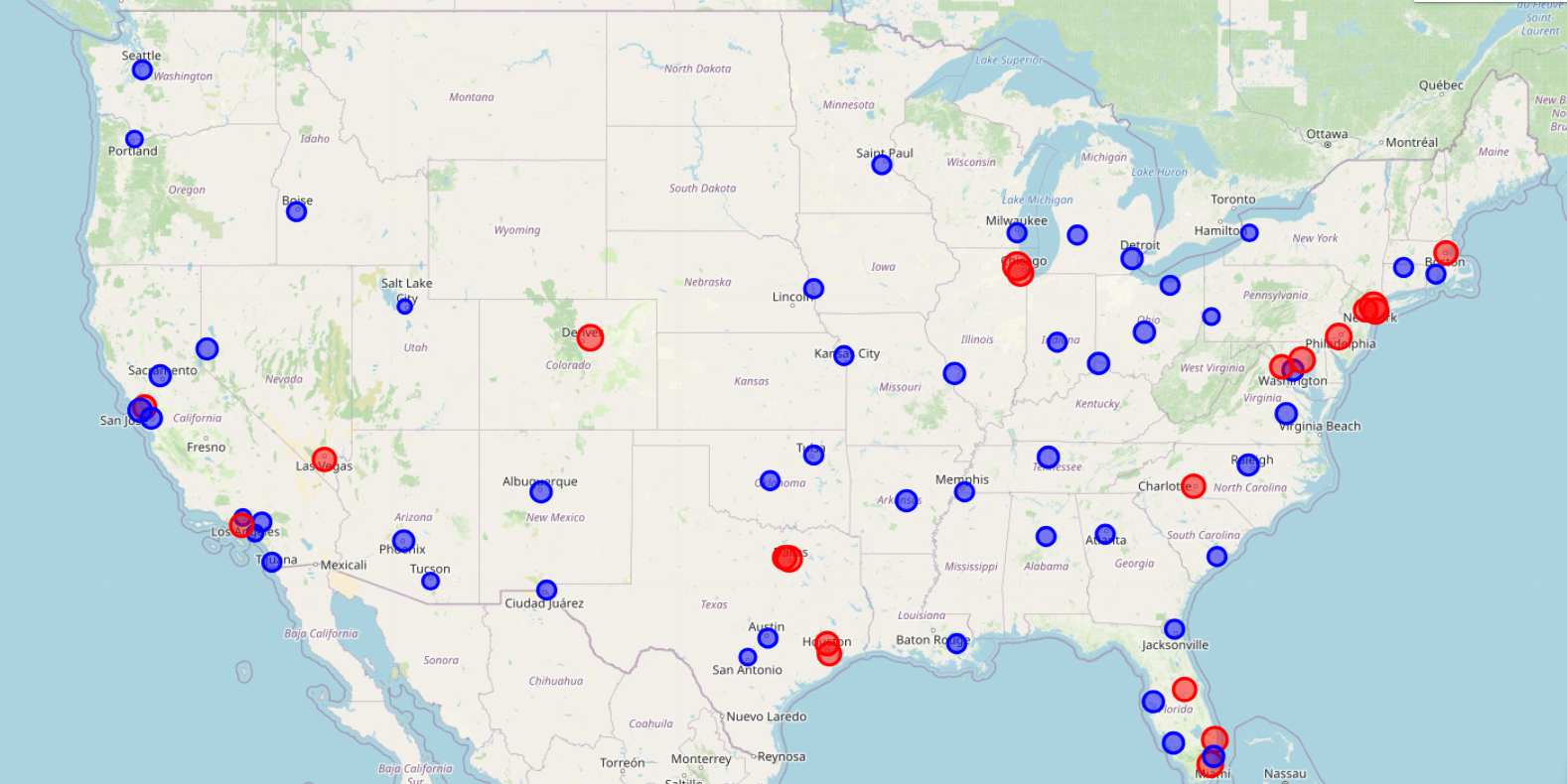
**Airport Delay Rate Analysis**

Given the extensive size of our dataset comprising millions of records, it is essential to ensure that our findings provide valuable insights. To achieve this, we prioritize airports based on their capacity, which is typically indicated by the volume of departure and arrival flights. Initially, we assess the importance of airports by analyzing their 'outdegree', representing the number of departure flights.

As mentioned before, we define a flight as 'late' if it arrives 15 minutes or more after the scheduled arrival time. And we will only focus on large airports(with departure flights exceeding 10,000 in the dataset) to do the airport delay rate analysis. This criterion ensures a sufficient sample size for robust analysis and meaningful insights into delay patterns.



With the geographical information, we create visual representations of airport locations on a map (red labels represent the airports with top quarter of the delay rates) which could provide a comprehensive understanding of what the analysis could help in decision making in both the business world and daily life.



From the map above, we can discover that in the Great DMV Area, we have 3 airports which are Baltimore-Washington International airport(BWI), Ronald Reagan Washington National Airport(DCA) and Washington Dulles International Airport(IAD), Notably, BWI and IAD exhibit higher delay rates compared to DCA.

For airline companies operating in this region, this insight provides an opportunity to optimize flight operations by potentially reallocating departure airports from BWI or IAD to DCA. By leveraging DCA's comparatively lower delay rate, airlines can enhance operational efficiency and improve customer service by minimizing the likelihood of flight delays. Such strategic adjustments can lead to more reliable flight schedules and heightened customer satisfaction.

Individual travelers can also benefit from this information by proactively mitigating flight delay risks. By opting to depart from DCA instead of BWI or IAD for the same destination, travelers can reduce their exposure to potential delays and enhance their overall travel experience.

**Conclusion** This comprehensive analysis of the 2015 Flight Delays and Cancellations dataset has unlocked pivotal insights into the dynamics of flight delays, offering valuable perspectives for airlines, airports, and passengers alike. Through meticulous data processing, pattern identification, and the development of predictive models, our study not only forecasts future delays with a notable degree of accuracy but also elucidates the varying delay rates across different airports.

The predictive models we've established serve as a testament to the power of data-driven decision-making in the aviation industry. Airlines can harness this information to refine operational strategies, optimize flight schedules, and enhance passenger communication, thereby significantly improving customer satisfaction. Similarly, airport management can leverage these insights to pinpoint systemic issues contributing to delays, formulating targeted interventions to bolster efficiency and service quality.

For travelers, the findings of this study demystify the often unpredictable nature of flight delays, empowering them with knowledge to make informed travel decisions, effectively manage their time, and reduce stress associated with air travel.

In essence, our exploration of the 2015 flight data sheds light on the multifaceted issue of flight delays, offering a blueprint for enhanced operational resilience, customer satisfaction, and overall industry performance. As we move forward, the continued application of such analytical methodologies promises not only to mitigate the inconvenience of delays but also to elevate the air travel experience for all stakeholders involved.