1. Data Sources

The dataset we have chosen for this report is airline delays datasets which contains information on airline delays in the year 2019 and 2020. Its includes data such as the number of flights arriving at the airport, the number of delays flights, reasons for delays and total delay times. Analyzing this dataset can provide valuable information for businesses in the aviation industry, airports, and airlines. From this dataset, we can analyze the historical trends in flight delay overtime which in the case is 2019 and 2020. This helps us to generate the percentage of flight delays between these years. Additionally, we can compare delay metrics across different airlines, airports, and regions to identify performance and variations and outliers. Determine which carriers or airports have the highest delay rates and assess the potential factors driving these differences. Moreover, we can investigate the main reasons behind flight delays which could include weather-related issues, air carrier delays and more.

1.1 Data Acquisition Automation

To automate the data acquisition process, we purpose utilizing web scraping techniques. From the research it was found that the dataset which we downloaded from Kaggle was collected from US government websites(https://www.bts.gov/). So, by leveraging tools like Scrapy in Python, we can programmatically extract the required data from the government website. The US government website provides monthly reports of airline delays, with which we could fetch the monthly flight delay report without manual intervention, ensuring up to date and accurate data for our visualization.

2. Exploratory Data Analysis

In this step, we conducted data exploration, understanding, and transformation on the airline delay dataset. From data exploration, we found out that these datasets have 3351 rows and 21 columns which contain data ranging from period like year and month, name of carrier, airport code and airport name and lastly the number of flights that were delay and their reasons.

2.1 Handling Missing Values

For transformation, we started by checking for missing values in the dataset. Upon inspection, we found that 8 rows contained null values. To ensure data integrity, we proceeded to remove these null values from the datasets.

3. Generating Insights

By performing different analysis, we generate insight from this airline delay dataset. These insights provide valuable information for good decision making and operational improvements in the aviation industry and help them minimize the overlay airline delays.

3.1 Percentage of Delayed Flights

In this observation, we generated the overall percentage of flight delays in 2019 and 2020. From the analysis, it was found that in 2019, every 5th flight arrived with a delay (20.29%), while in 2020, every 10th flight arrived with a delay (11.72%). So, in total, there was 143,073 hours of delay in 2019, whereas there was only 42,678 hours of delay in 2020.

The difference between the two years is extremely significant, amounting to 100,395 hours of delay in total.

3.2 Carrier Performance Analysis

For the second observation, we analyzed the carrier performance using data such as "arr_del15", "carrier_delay", and other delay metrics. This analysis helped us generate information regarding their performance and potential areas for improvement. It was found that Southwest Airlines Co has the highest average delay value of around 7500 minutes across various delay metrics. This indicates that flights operated by this carrier experience longer delays on average compared to other airlines. Conversely, Allegiant Air has the lowest average delay value among the airlines, at around 1000 minutes. This suggests that flights operated by Allegiant Air experience relatively shorter delays compared to other airlines.

3.3 Total Delayed Flights vs. Total Delayed Flights by Airport

From the third observation, which compared total delayed flights against total delayed flights by airport, it was found that DFW has the highest total delayed flights, approximately 570,000. Airlines operating at DFW may face operational constraints due to frequent delays, impacting on their schedule adherence and overall customer satisfaction. On the other hand, IAH has the lowest total delayed flights, approximately 210,000. This comparatively lower number of delayed flights suggests better operational performance and efficiency at IAH. Airlines operating at IAH may benefit from smoother operations, improved schedule adherence, and reduced disruptions compared to airports with higher delay rates. The significant disparity in total delayed flights between DFW and IAH underscores the importance of effective airport management and operational planning.

3.4 Distribution of Delay Reason

From this fourth observation, it was found that more than half of all delays were caused by the late arrival of flights (57.4%). The National Aviation System is the second-largest cause of delays at airports (34.2%). Weather conditions cause fewer delays than expected (8.1%). Security issues are an extremely rare cause of flight delays (0.2%)

4. Deploy to real world

The following analysis on the airline delay datasets has provided us with a valuable insight that can be used in various real-world scenarios.

- Airlines can use our analysis to identify patterns and trends in delay causes, allowing them to implement targeted strategies to improve operational efficiency.
- By understanding the factors contributing to delays, airlines can proactively communicate with passengers and manage expectations during travel disruptions.
- In future, our analysis can serve as a foundation for developing predictive analytics modal that can forecast future delay occurrences based on historical data and external factors such as weather patterns and air traffic volume.