CAP5771

INTRODUCTION TO DATA SCIENCE

Team Members:

Name: Harsh Gupta

UFID: 21828165

Name: Muthukumaran Ulaganathan

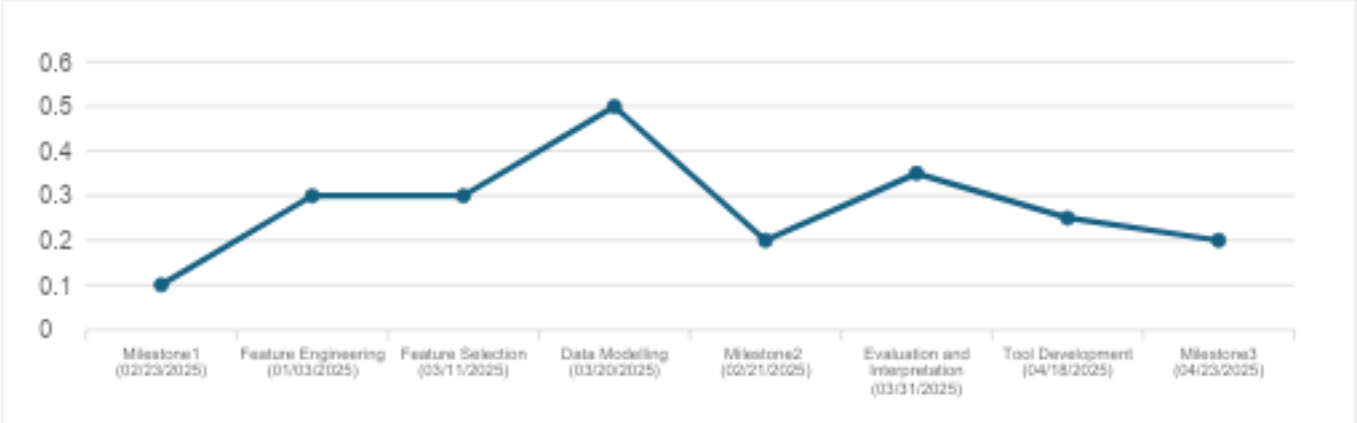
UFID: 28225998

Introduction:

This report focuses on the data preprocessing and exploratory data analysis (EDA) steps for a flight delay prediction project. The project utilizes several datasets that provide various insights into airline operations, flight delays, and related factors. The primary datasets include the U.S. International Air Traffic data, which captures flight data between U.S. and international airports, and the Flight Delay and Cancellation dataset, which contains detailed flight performance metrics, such as departure and arrival delays. Additional datasets such as Storm Events data and Consumer Airfare Reports are used to analyze external factors, including severe weather events and airfare trends, that may contribute to delays.

The aim of preprocessing and EDA stages is to clean, transform, and explore the data to understand patterns and relationships between different features. The datasets include a variety of attributes like flight times, carrier details, airport information, and delay causes. This report outlines the steps taken to prepare the data for modeling, which includes handling missing values, transforming variables, and performing initial statistical analysis. The insights gained from this process are essential for building a predictive model that can accurately forecast flight delays based on historical data.

Project Timeline:



Datasets report:

Name	Source	Attributes	Description	Contributor
U.S. International Air Traffic data(1990-2020)	Kaggle	'data_dte', 'Year', 'Month', 'usg_apt_id', 'usg_apt', 'usg_wac', 'fg_apt_id', 'fg_apt', 'fg_wac', 'airlineid', 'carrier', 'carriergroup', 'type', 'Scheduled', 'Charter', 'Total' 'data_dte', 'Year', 'Month', 'usg_apt_id', 'usg_apt', 'usg_wac', 'fg_apt_id', 'fg_apt', 'fg_wac', 'airlineid', 'carrier', 'carriergroup', 'type', 'Scheduled', 'Charter', 'Total'	Departures: Data on all flights between US gateways and non-US gateways, irrespective of origin and destination. Each observation provides information on a specific airline for a pair of airports, one in the US and the other outside. Three main columns record the number of flights: Scheduled, Charter, and Total. Passengers: Data on the total number of passengers for each month and year between a pair of airports, as serviced by a particular airline.	Harsh
Flight Delay and Cancellation Dataset (2019-2023)	Kaggle	'FL_DATE', 'AIRLINE', 'AIRLINE_DOT', 'AIRLINE_CODE', 'DOT_CODE', 'FL_NUMBER', 'ORIGIN', 'ORIGIN_CITY', 'DEST', 'DEST_CITY', 'CRS_DEP_TIME', 'DEP_TIME', 'DEP_DELAY', 'TAXI_OUT', 'WHEELS_OFF', 'WHEELS_ON', 'TAXI_IN', 'CRS_ARR_TIME', 'ARR_TIME', 'ARR_DELAY', 'CANCELLED', 'CANCELLATION_CODE', 'DIVERTED', 'CRS_ELAPSED_TIME', 'ELAPSED_TIME', 'AIR_TIME', 'DISTANCE', 'DELAY_DUE_CARRIER', 'DELAY_DUE_WEATHER', 'DELAY_DUE_NAS', 'DELAY_DUE_SECURITY',	This dataset is a collection of flight performance metrics for commercial airlines. It includes information on scheduled and actual flight operations, such as flight dates, carrier codes, flight numbers, and airport details (origin and destination). The dataset captures crucial operational data including scheduled and actual departure/arrival times, delays (departure and arrival), taxi times, and elapsed flight time. Additionally, it provides detailed breakdowns of	Muthu

		'DELAY_DUE_LATE_AIRCRAFT'	delay causes—such as carrier-related, weather-related, and security delays—as well as indicators for cancellations and diversions.	
Storm Events data	National Centers for Environmental Information	'BEGIN_YEARMONTH', 'BEGIN_DAY', 'BEGIN_TIME', 'END_YEARMONTH', 'END_DAY', 'END_TIME', 'EPISODE_ID', 'EVENT_ID', 'STATE', 'STATE_FIPS', 'YEAR', 'MONTH_NAME', 'EVENT_TYPE', 'CZ_TYPE', 'CZ_FIPS', 'CZ_NAME', 'WFO', 'BEGIN_DATE_TIME', 'CZ_TIMEZONE', 'END_DATE_TIME', 'INJURIES_DIRECT', 'INJURIES_INDIRECT', 'DEATHS_DIRECT', 'DEATHS_INDIRECT', 'DAMAGE_PROPERTY', 'DAMAGE_CROPS', 'SOURCE', 'MAGNITUDE', 'MAGNITUDE_TYPE', 'FLOOD_CAUSE', 'CATEGORY', 'TOR_F_SCALE', 'TOR_LENGTH', 'TOR_WIDTH', 'TOR_OTHER_WFO', 'TOR_OTHER_CZ_STATE', 'TOR_OTHER_CZ_FIPS', 'TOR_OTHER_CZ_NAME', 'BEGIN_RANGE', 'BEGIN_AZIMUTH', 'BEGIN_LOCATION', 'END_RANGE', 'END_AZIMUTH', 'END_LOCATION', 'BEGIN_LAT', 'BEGIN_LON', 'END_LAT', 'END_LON', 'EPISODE_NARRATIVE', 'EVENT_NARRATIVE', 'DATA_SOURCE'	The Storm Events dataset provides detailed information on severe weather occurrences in the United States. It includes event timing, location (state, county, geographic coordinates), event type, and impact metrics such as injuries, fatalities, and damage estimates. This dataset is used to analyze weather patterns and assess the effects of severe storms	Harsh
Airline Fleets	Kaggle	'Parent Airline', 'Airline', 'Aircraft Type', 'Current', 'Future', 'Historic', 'Total', 'Orders', 'Unit Cost', 'Total Cost (Current)', 'Average Age'	This dataset includes details on parent airlines, individual airline brands, aircraft types, current operating fleets, future orders, unit and total aircraft costs, and the average age of the fleets	Harsh
Consumer Airfare Report	Data.Gov	'tbl', 'Year', 'quarter', 'mkt_fare', 'citymarketid_1', 'citymarketid_2', 'city1', 'city2', 'carairlineid', 'car', 'carpax', 'carpaxshare', 'caravgfare', 'fareinc_min', 'fareinc_minpaxsh', 'fareinc_max', 'fare_inc_maxpaxsh', 'fare_inc_x3paxsh',	This dataset includes metrics such as market fare, year, quarter, city market identifiers, cities, airline identifiers, passenger metrics, fare increments, and geocoded market details—supporting analysis of domestic short-haul airfares.	Muthu

		'Geocoded_City1', 'Geocoded_City2', 'tbl5pk'		
Airline IATA code	**Wikipedia**	'IATA', 'ICAO', 'Airline', 'Call sign', 'Country/Region', 'Comments'	Contains mapping of airline codes and names	Muthu

Data Integration and Schema Analysis:

To ensure consistency and usability, multiple datasets were cleaned, transformed, and merged based on relevant keys such as date, state, and IATA code. The following steps were performed:

1. Airline Delay Data Processing

- Extracted the date from the existing timestamp column.
- Filtered the dataset to include only records from 2020, as the original dataset was too large for efficient processing.

2. Joining Airline Delay Data with International Air Traffic Data

- Merged using IATA code and date to align flight delays with international air traffic patterns.

3. Incorporating Storm Events Data

- Extracted the state from existing columns in the storm events dataset.
- Merged with the previously combined dataset on date and state to integrate weather-related impacts.

4. Cleaning and Preparing Airline Fleet Data

- Dropped rows where essential numerical columns (Current, Total, Average Age) were missing.
- Filled missing values in categorical columns (Parent Airline, Airline, Aircraft Type) with "Unknown".
- Replaced NaNs in numerical columns (Orders, Future) with 0 to ensure proper aggregations.
- Converted cost-related columns (Unit Cost, Total Cost (Current)) to numeric by removing currency symbols and commas.
- Converted arrival delay values (ARR_DELAY) to absolute values to standardize delay interpretation.

5. Merging Airline Fleet Data with Airline Codes

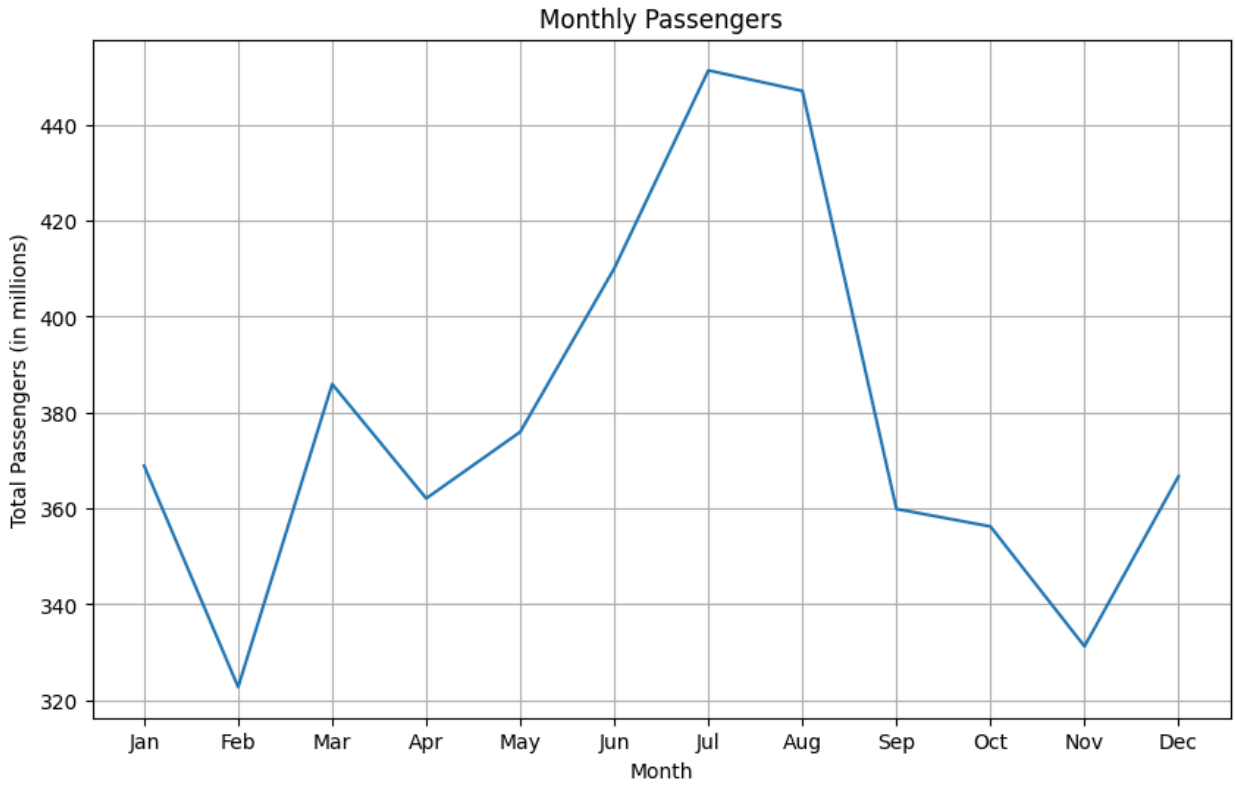
- Merged airline fleet data with airline codes dataset on carrier name to obtain corresponding IATA codes.
- This ensured a consistent mapping for later integration with airline fare/pricing data.

6. Final Merging with Airfare Data

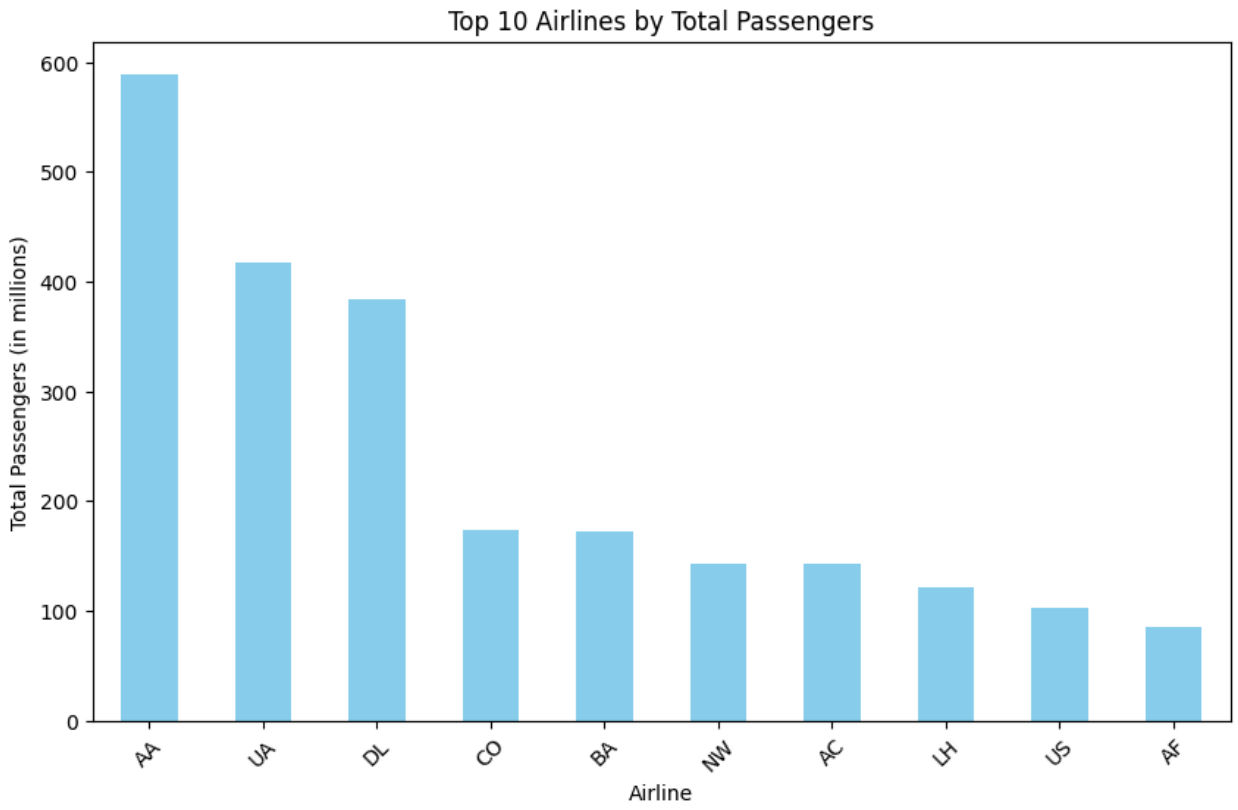
- Used IATA code and state as primary keys to join the airfare dataset with the previously merged dataset (airline delays, storm events, international air traffic).
- This created a unified dataset containing airline operations, delays, weather conditions, international traffic, fleet details, and fare information for deeper analysis.

EDA report:

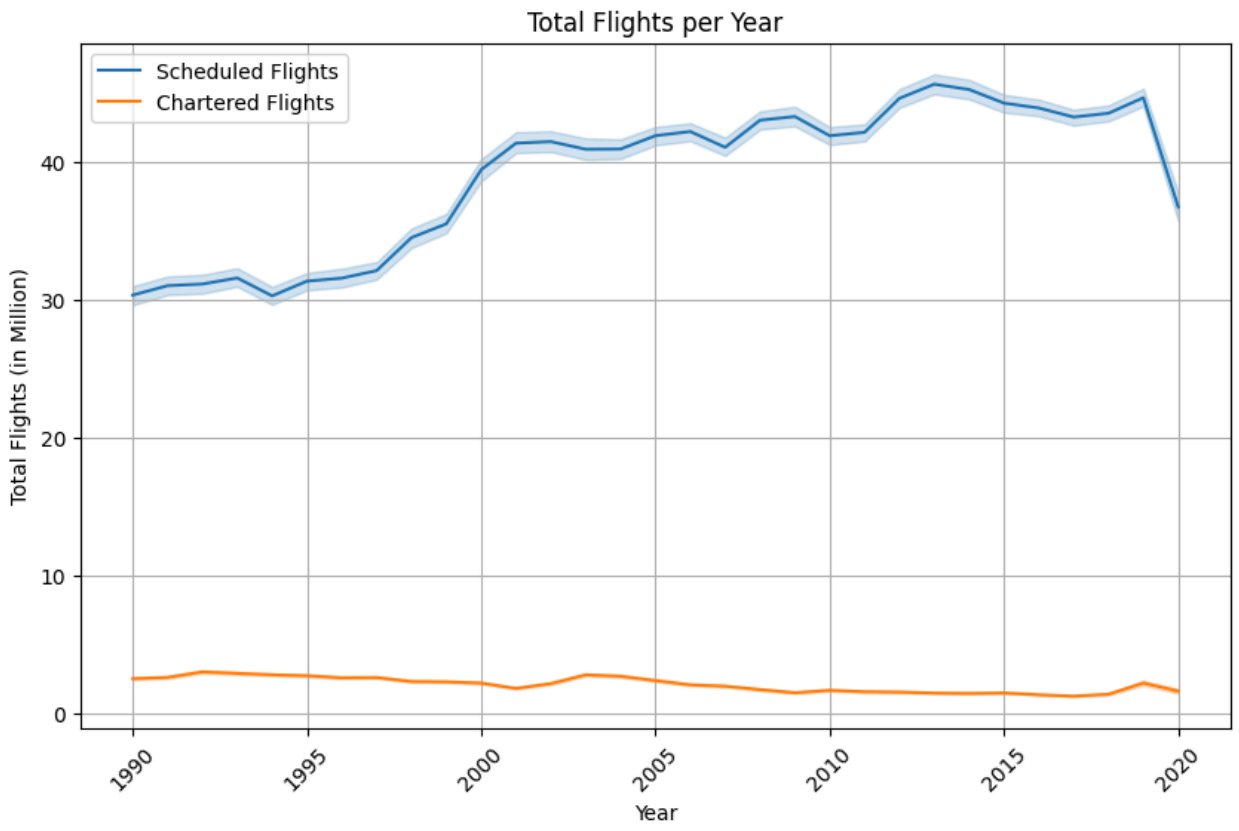
DATASET: U.S. International Air Traffic data(1990-2020)



- Passenger numbers peak during summer (July-August), indicating a seasonal travel trend. The lowest numbers are in February and November.

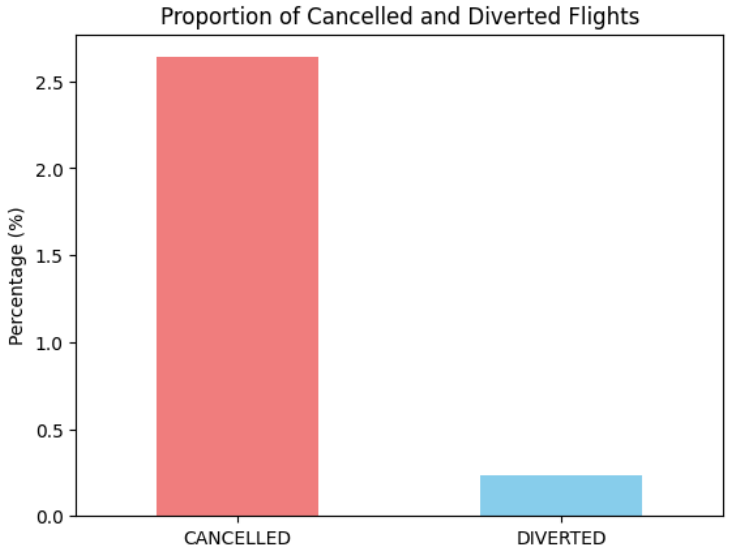


- Airline 'AA' dominates in total passengers, followed by 'UA' and 'DL', suggesting market concentration among these carriers

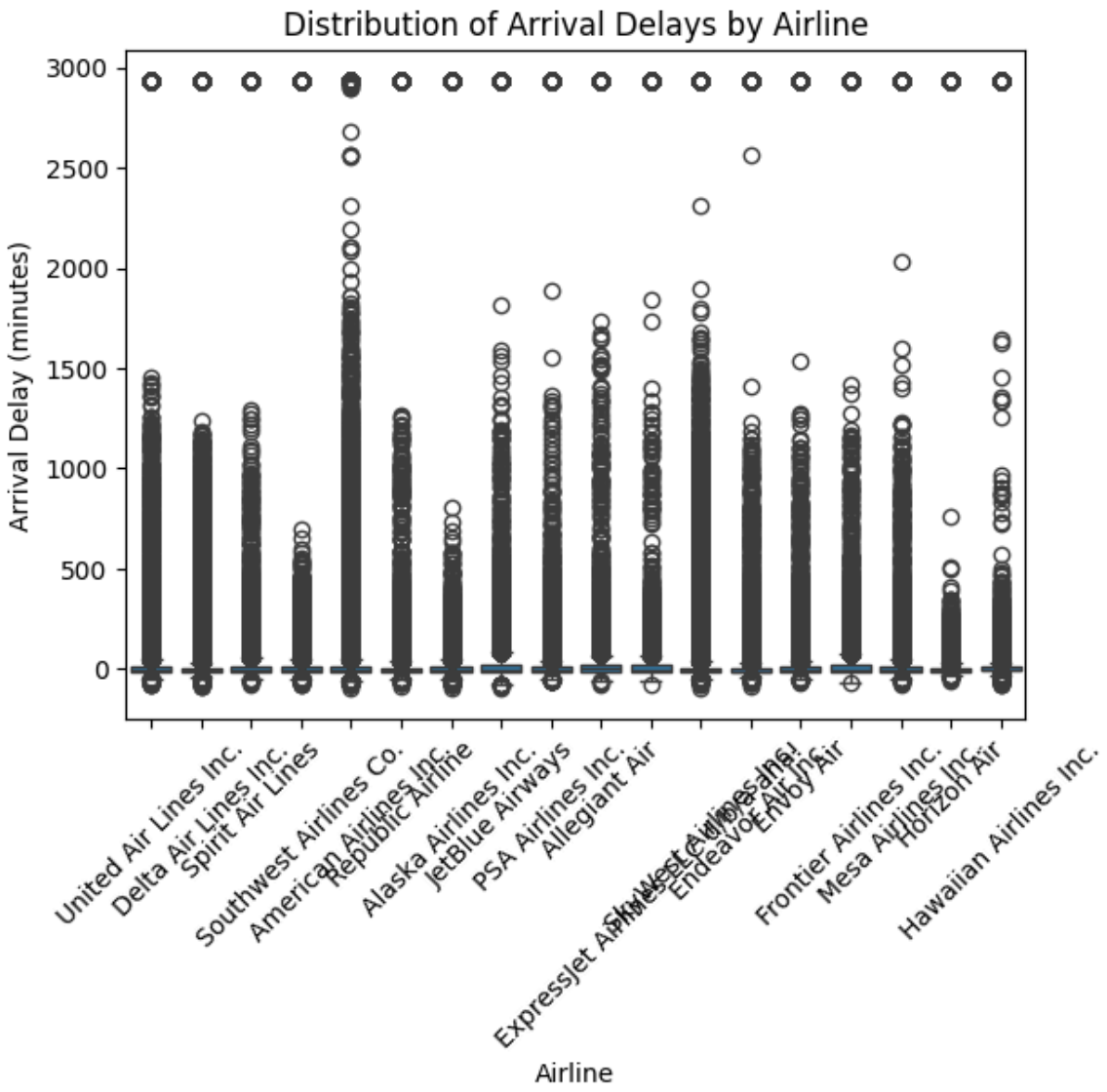


- Scheduled flights show consistent growth until a sharp decline post-2020, likely due to COVID-19. Chartered flights remain negligible throughout the years.

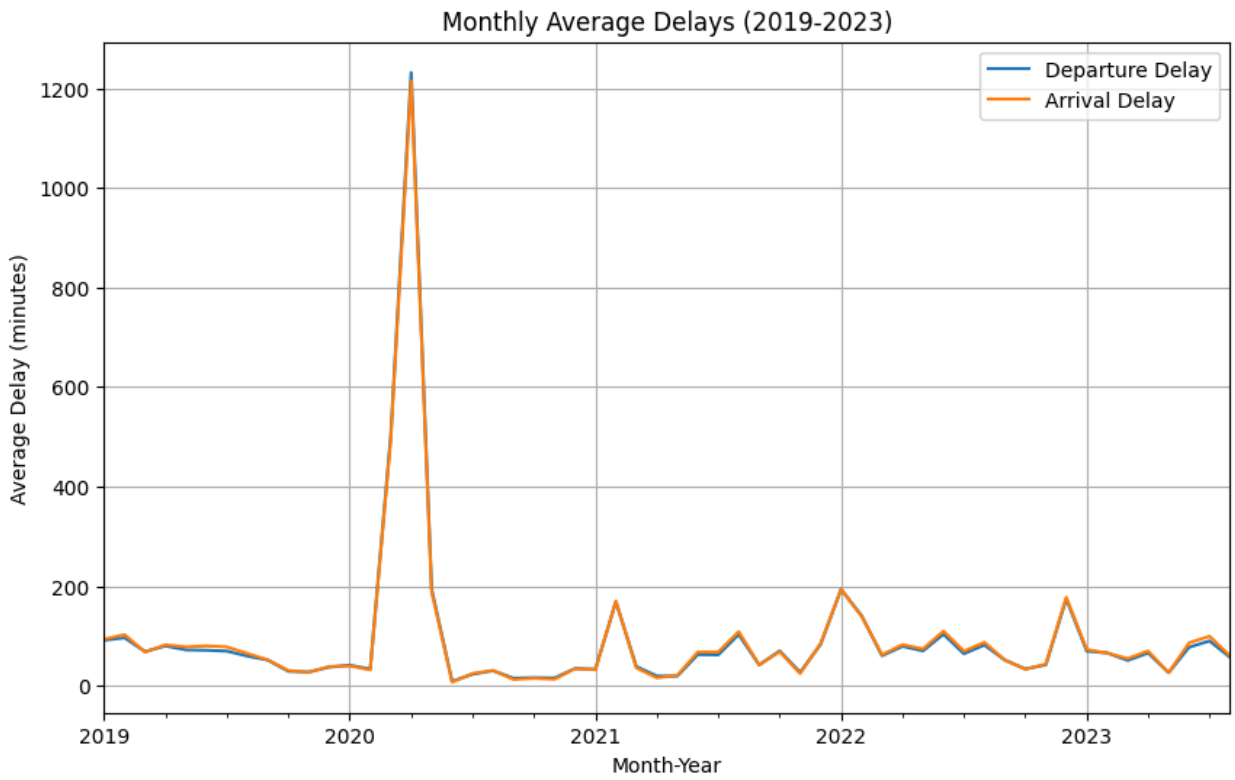
DATASET: Flight Delay and Cancellation Dataset (2019-2023)



- Flight cancellations occur at a significantly higher rate than diversions, indicating that operational issues more often result in cancellations rather than rerouting.

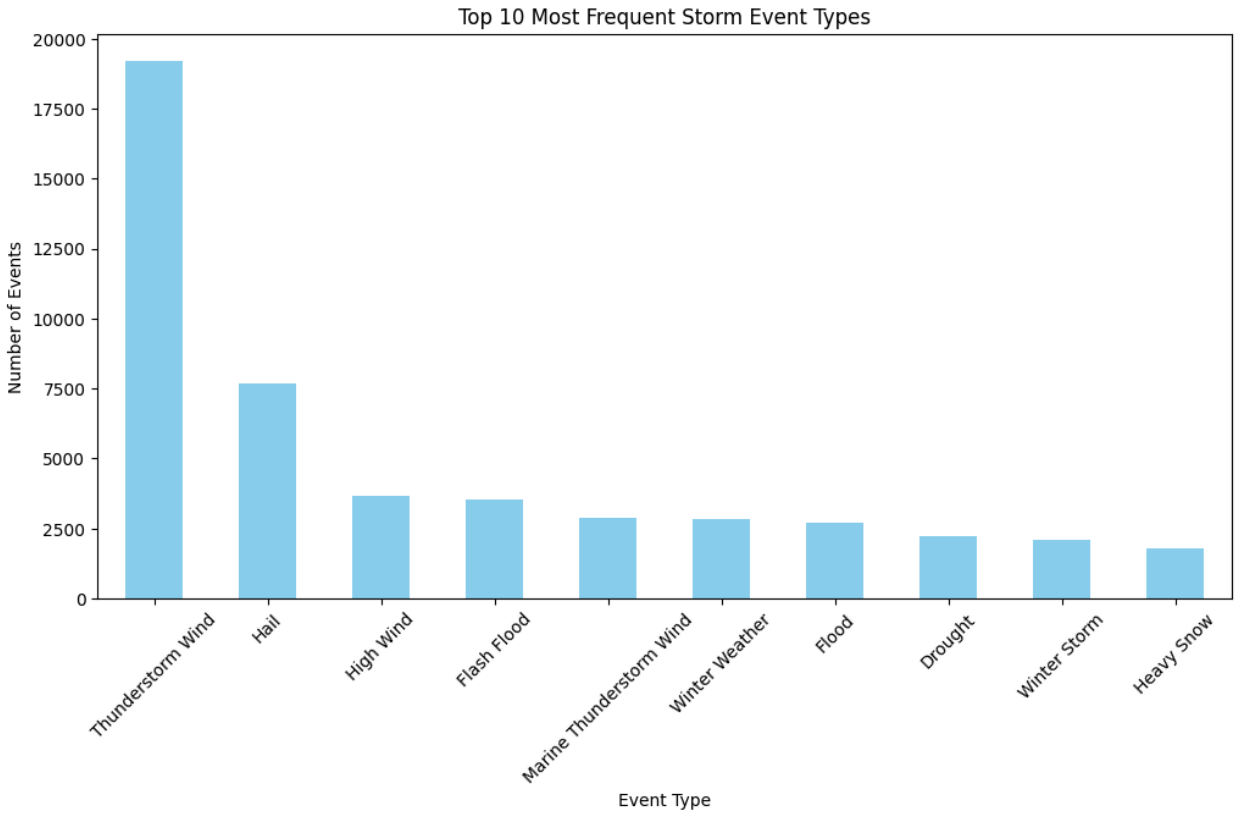


- Most airlines experience a wide range of arrival delays, with significant outliers suggesting occasional extreme delays affecting certain carriers more than others.

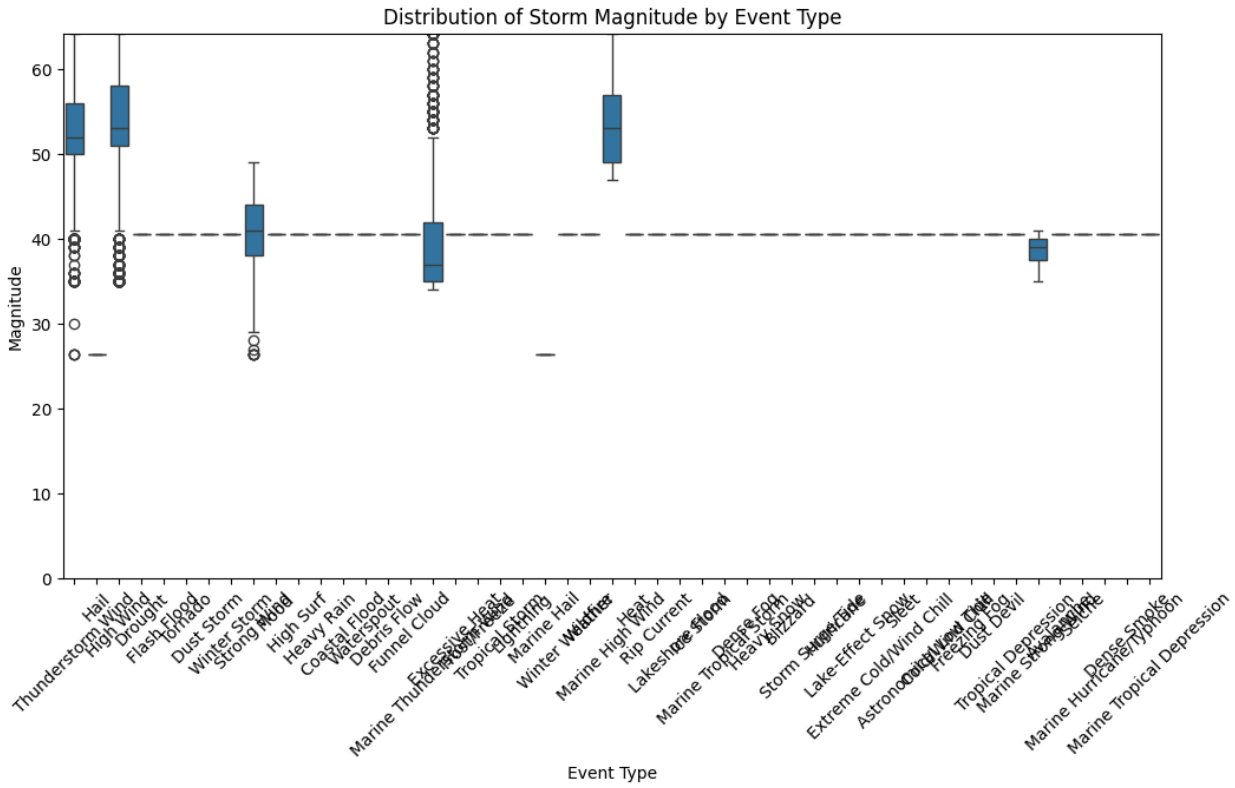


- A major spike in delays occurred in early 2020, likely due to the COVID-19 pandemic, while delays remained relatively stable in the following years with periodic fluctuations.

DATASET: Storm Events data

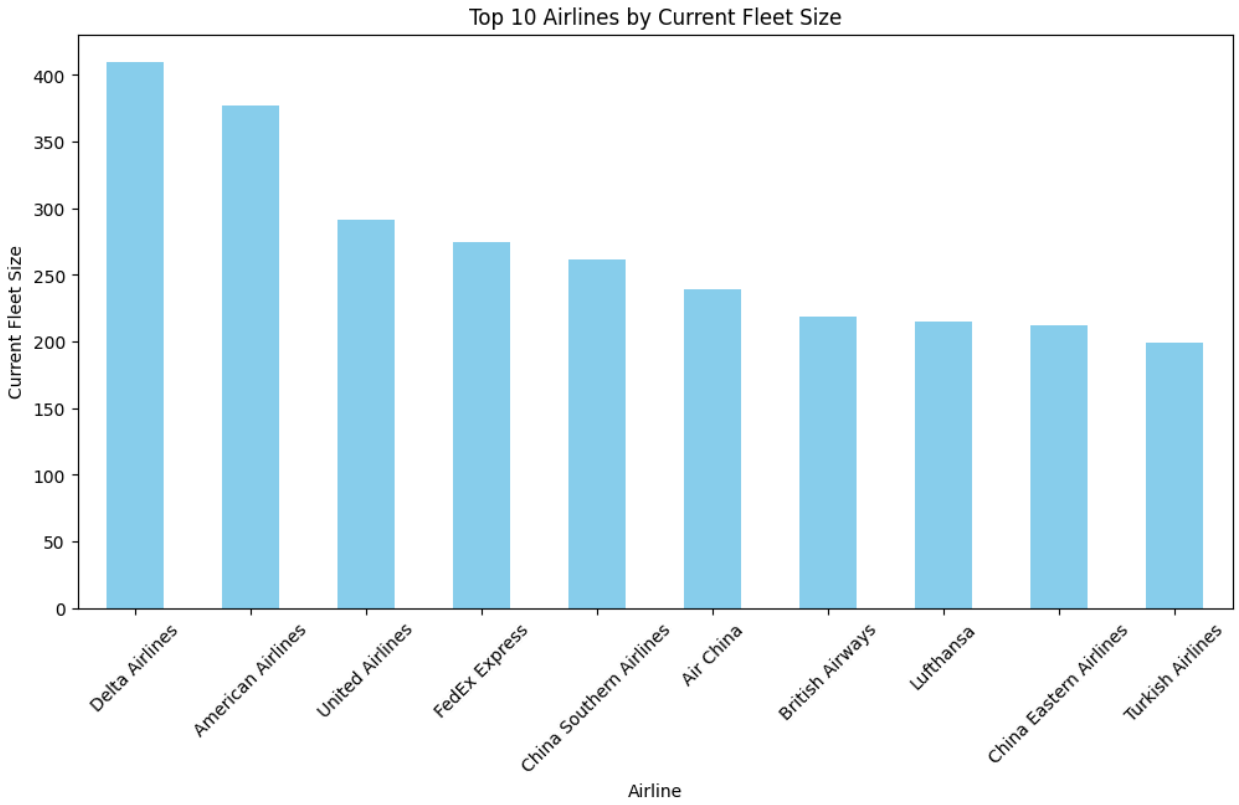


- Thunderstorm Wind is by far the most common storm event, followed by Hail and High Wind, indicating that severe winds are the predominant weather hazard.

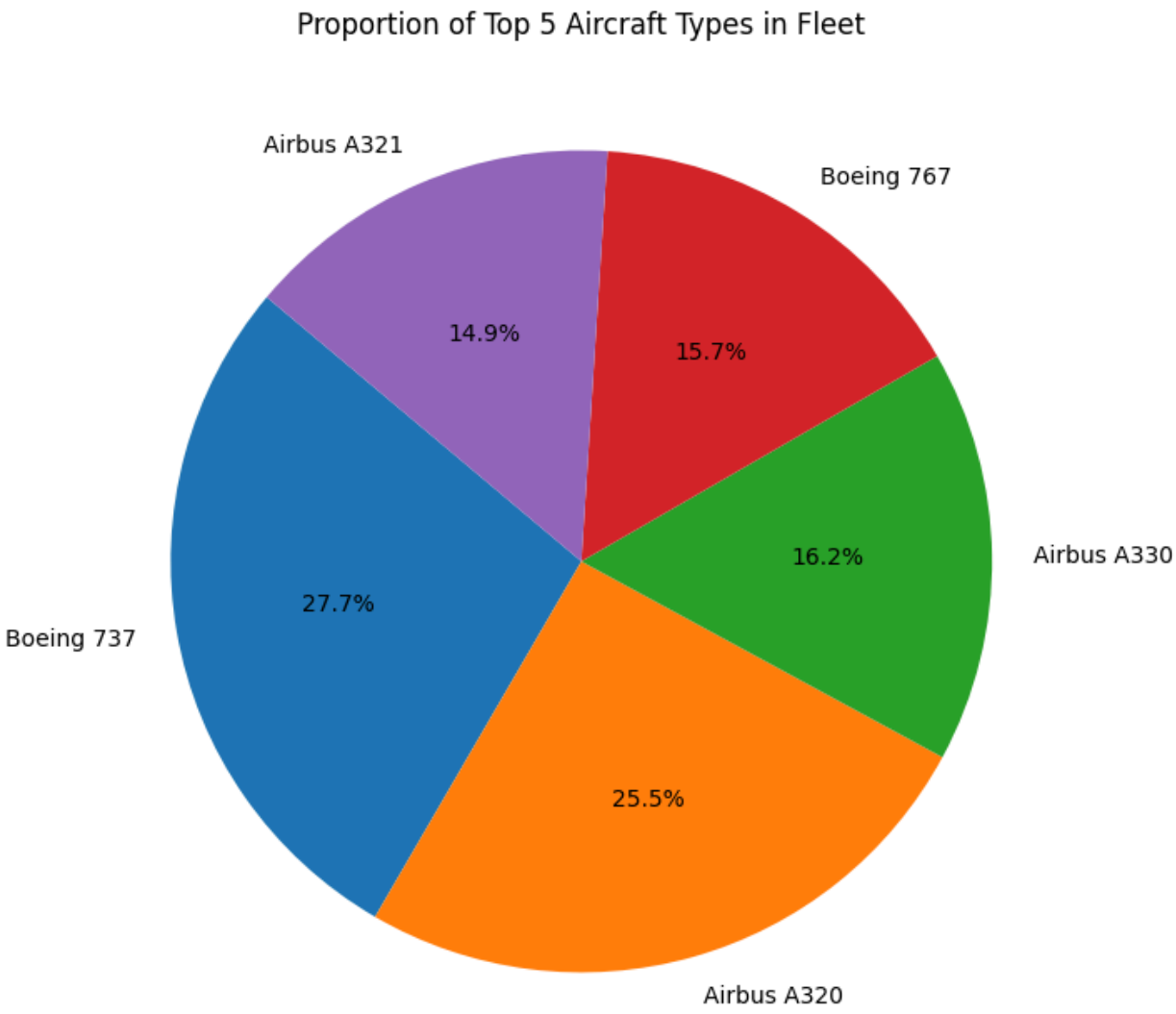


- Storm magnitudes vary significantly across event types, with some events exhibiting extreme outliers, suggesting that certain storms can be exceptionally severe compared to others.

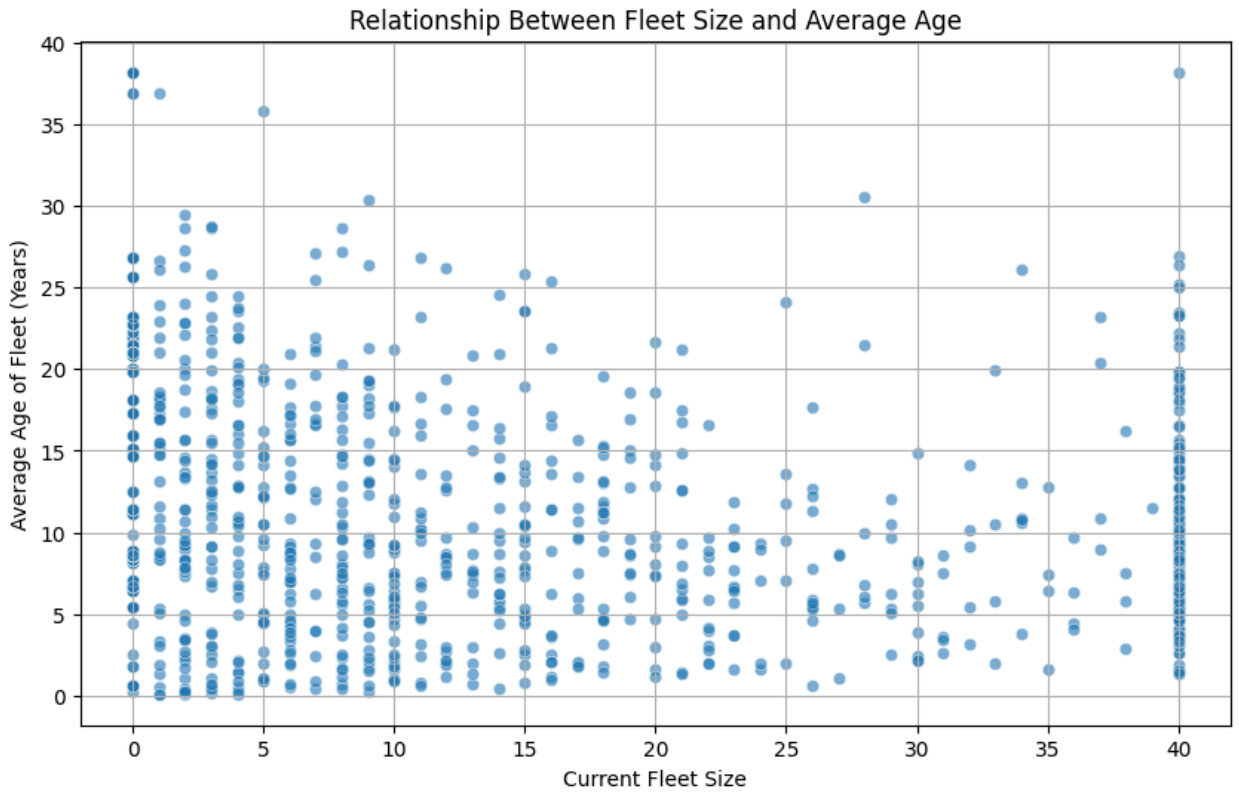
DATASET: Airline Fleets



- Delta Airlines has the largest fleet, followed closely by American Airlines and United Airlines, indicating dominance in commercial aviation

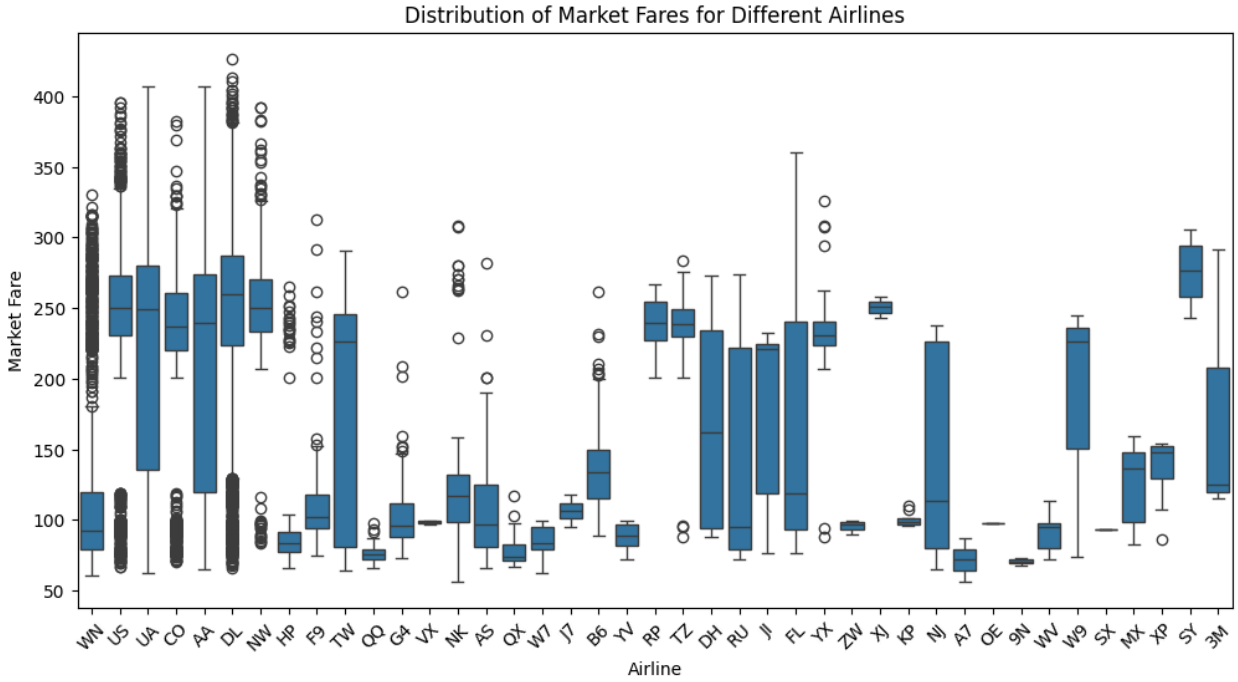


- Boeing 737 and Airbus A320 make up the majority of the fleet, showing preference for efficient, high-capacity aircraft.

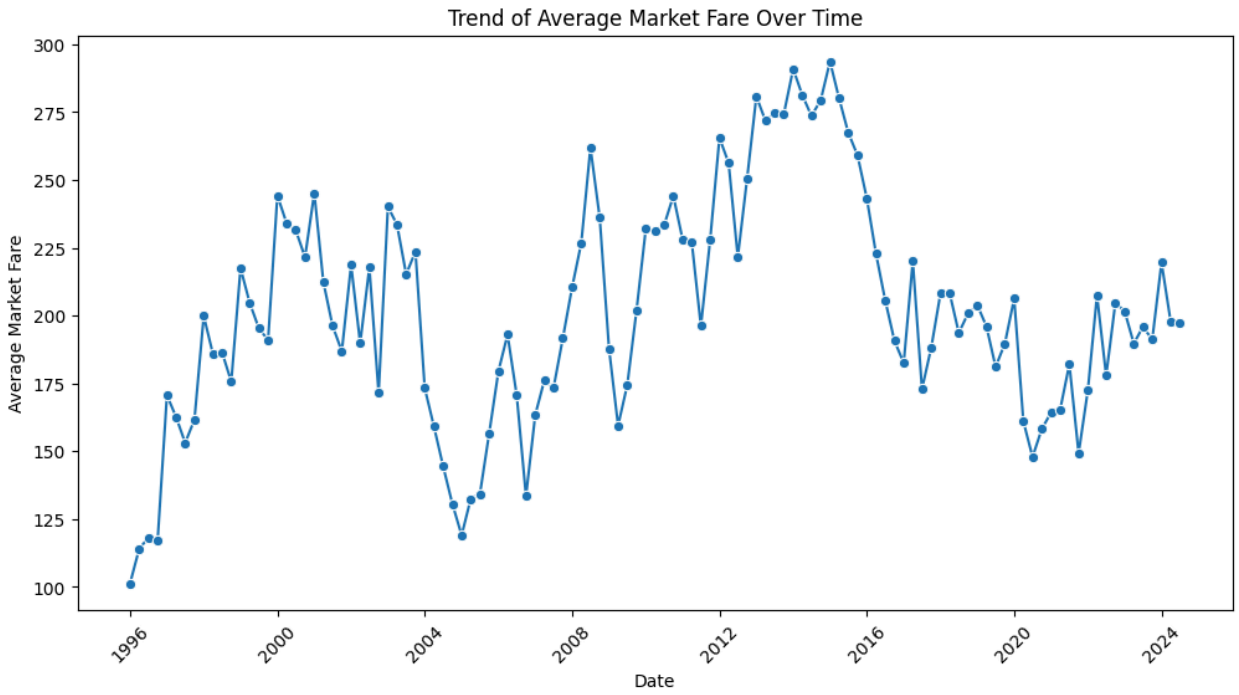


- No clear correlation, but older fleets tend to have smaller sizes, suggesting newer aircraft dominate larger fleets.

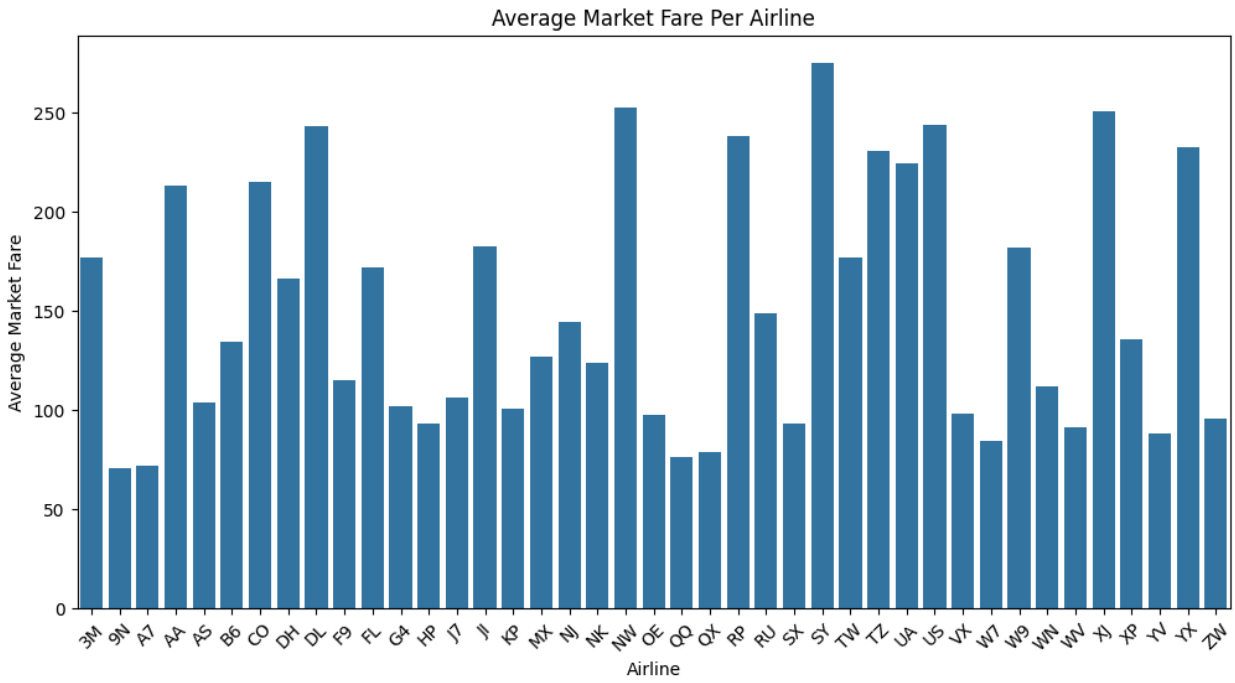
DATASET: Consumer Airfare Report



- Airlines exhibit a wide range of fare distributions, with some having large variations (wider boxes and whiskers), indicating fare flexibility.
- Several airlines have significant outliers, suggesting occasional high-priced flights due to demand surges or premium services.
- Some airlines have a tight range of fares, possibly reflecting standardized pricing strategies.



- The average market fare has seen fluctuations over time, with peaks and dips.
- There was a general upward trend until around 2015, followed by a decline, possibly due to changes in airline policies, fuel prices, or competition.
- The recent years show a stabilization with smaller fluctuations, suggesting market adjustments.



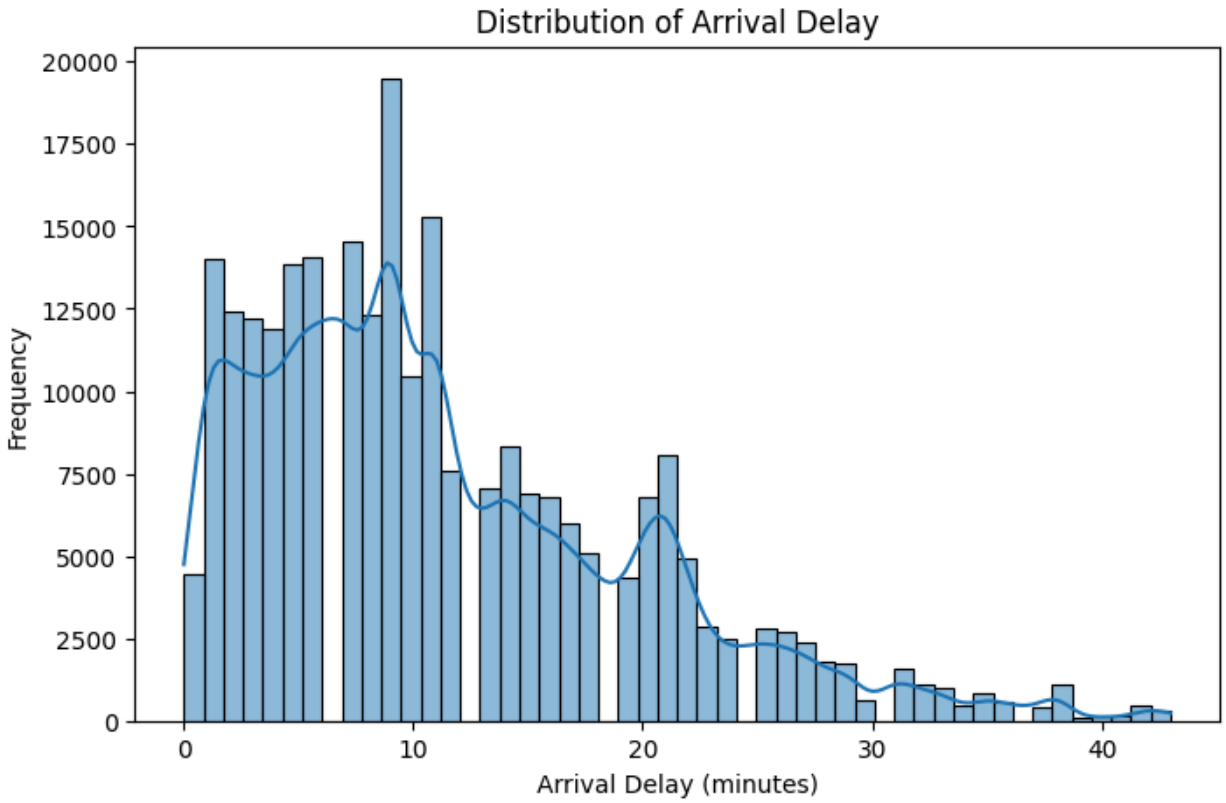
- There is significant variation in market fares across different airlines.
- Some airlines have consistently higher fares, possibly due to premium services or longer routes.
- A few airlines have much lower average fares, indicating a focus on budget travel.

Individual Contributions:

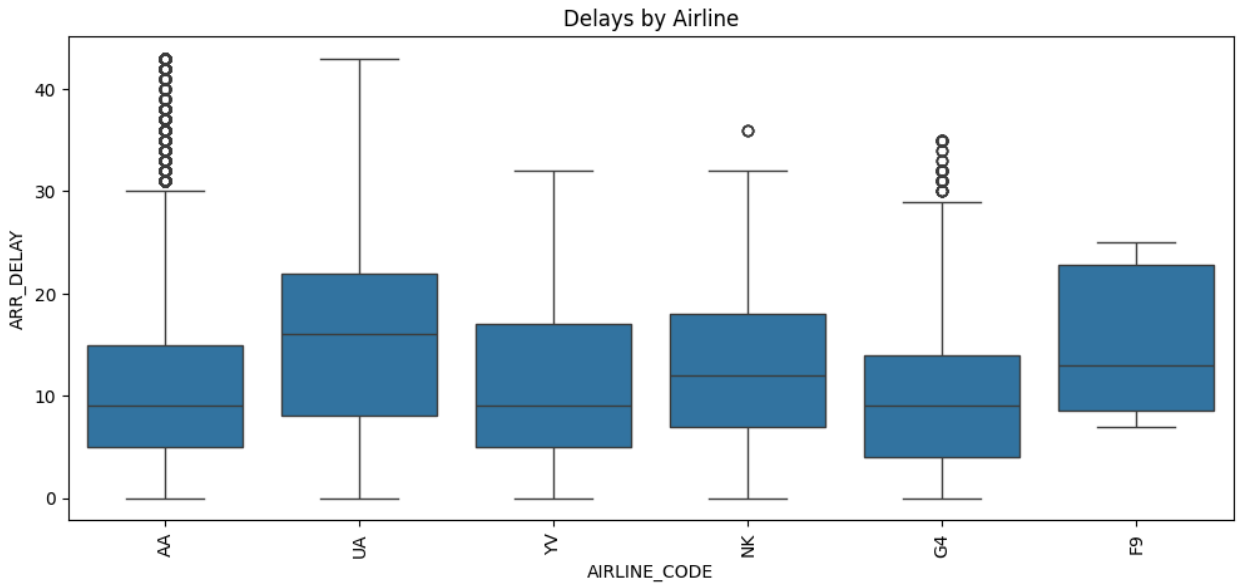
- **Harsh Gupta:**
  - Did Preprocessing and EDA on the following datasets
    - **Airline Fleets**
    - **Storm Events data**
    - **U.S. International Air Traffic data(1990-2020)**
- **Muthu**
  - Did Preprocessing and EDA on the following datasets
    - **Consumer Airfare Report**
    - **Flight Delay and Cancellation Dataset (2019-2023)**
  - Scraped the following dataset from wikipedia and cleaned it

    - **Airline IATA code**
- Worked on Data Integration and Schema Analysis

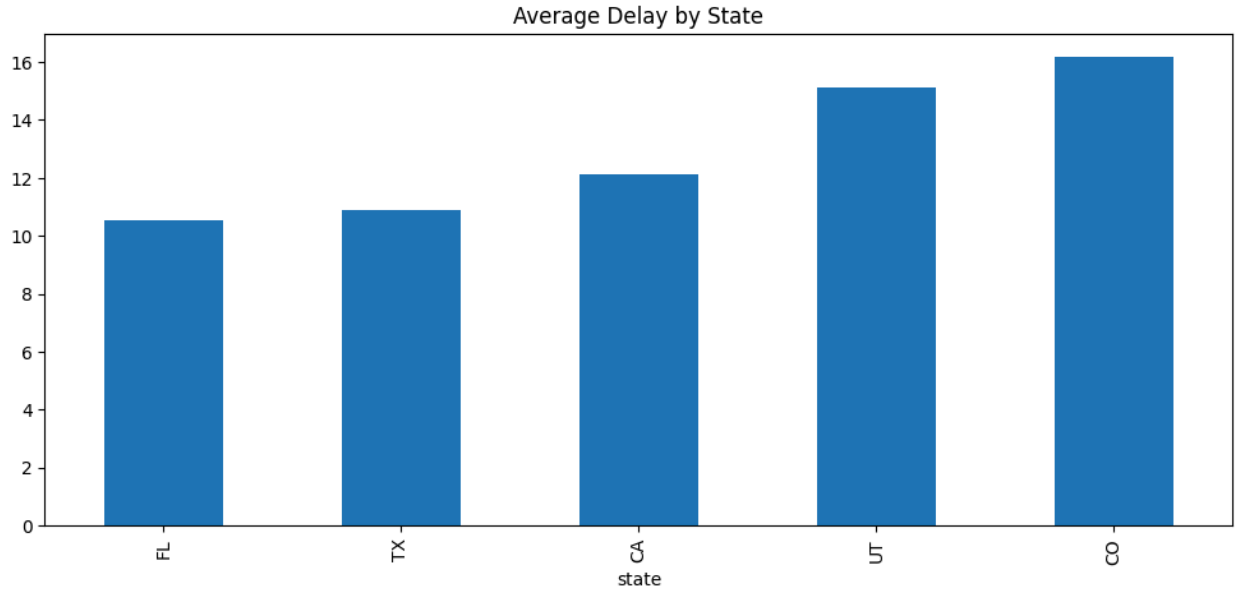
EDA after merging data into a single dataframe:



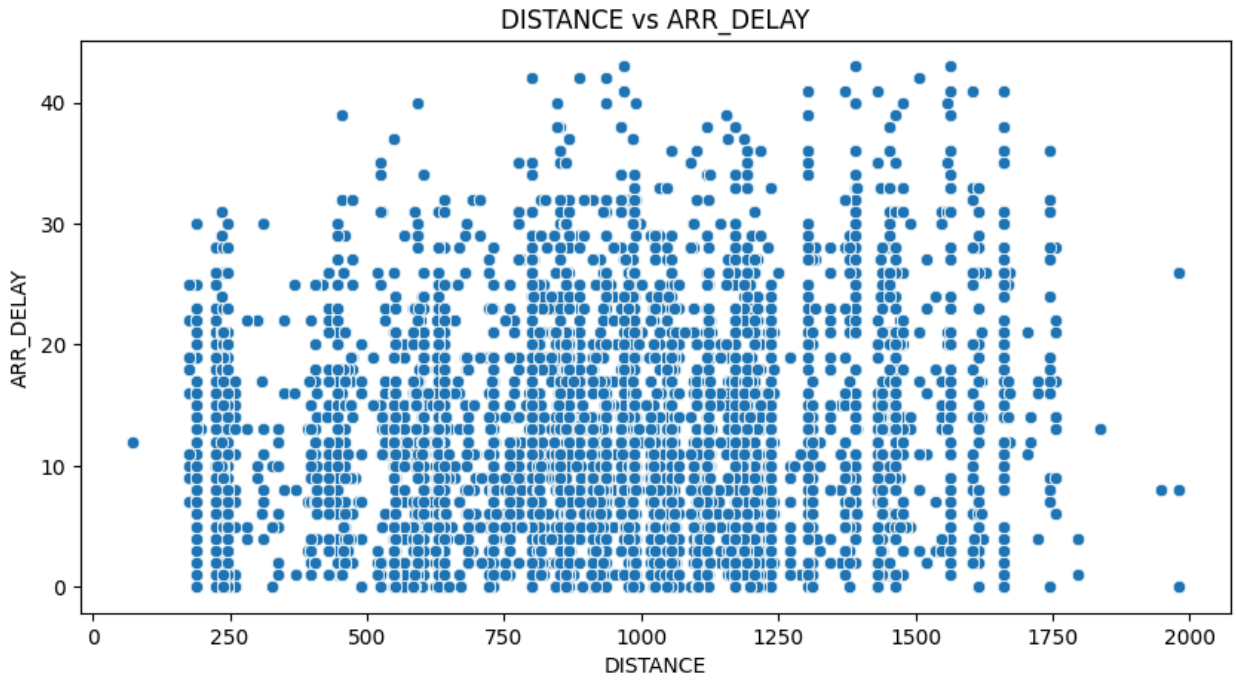
Most flights experience small delays (under 15 minutes), with frequency decreasing as delays increase. There's a noticeable peak around 10 minutes.



Airlines show varied delay distributions, with some having significant outliers. AA and G4 seem to have more extreme delays, while UA and F9 have broader delay distributions.



The states with the highest average flight delays are Utah (UT) and Colorado (CO), while Florida (FL) and Texas (TX) have relatively lower delays.



There is no clear correlation between flight distance and arrival delay, as delays appear scattered across all distances. This suggests that factors other than distance, such as airport congestion or airline operations, may have a greater impact on delays.