Interactive Flight Data

Dashboard

IE6600 Computation and Visualization

Spring 2024 SEC 03

PROJECT 4

FINALREPORT

GROUP 3:

Lokhi Nalam (002649847)

Pooja Arumugam (002872003)

Sathvik Ramappa (002847460) Arya Lokesh Gowda (002249418)

Part 1:
Introduction

The freedom and convenience of air travel have revolutionized the way we connect with people and explore new destinations. However, navigating the ever-changing world of airfare pricing can be a daunting task for both seasoned travelers and those embarking on their first journeys. Understanding the factors that influence airfare can empower individuals to make informed decisions and optimize their travel budgets.

The airline industry serves as a vital component of the global economy, facilitating travel and commerce across vast distances. Understanding the dynamics of airfare pricing is crucial for both consumers and industry stakeholders alike. The aviation sector stands as a cornerstone of modern global connectivity, facilitating the movement of people and goods with unprecedented efficiency. Central to the functioning of this intricate network is the pricing of airfares, a dynamic interplay of market forces, regulatory frameworks, and consumer behavior. In this report, we delve into the Consumer Airfare Report Table 1A, which provides comprehensive insights into airfare trends across all U.S. airport pair markets.

Airfare pricing is a multifaceted phenomenon influenced by an array of factors, ranging from operational costs and fuel prices to demand-supply dynamics and competitive pressures. By delving into the granular details of airfare data, we endeavor to unravel the intricacies of this complex ecosystem and shed light on the underlying drivers of price fluctuations. The dataset offers a detailed analysis of airfare pricing dynamics, encompassing a wide array of variables such as route popularity, seasonal fluctuations, and competitive market forces. By examining historical data and trends, we aim to uncover patterns and insights that can inform consumer decision-making, industry strategies, and policy formulation.
Through thorough analysis and interpretation of the data presented, we seek to address key questions surrounding airfare trends, including factors influencing price variations, the impact of market competition on ticket costs, and the implications for consumer welfare.

By shedding light on the complex interplay of factors shaping airfare pricing, this report aims to provide valuable insights for travelers, industry professionals, and policymakers navigating the dynamic landscape of the airline industries.

Part 2: Dataset Selection and Confirmation

The primary aim of this project is to gather air fare and have the government ready to use data for creating plans for airport competition. The dataset lists airport markets where the origin or destination airport is an airport that has other commercial airports in the same city. The fare and traffic data is developed to provide a basis for the requisite analysis. A section of the dataset provides real life examples of concerns and or information pertaining to U.S. airline service to various U.S. cities, at the end of the project we may gain valuable insights into domestic airfare patterns and potentially identify cost-effective travel options or research trends in airline competition.

We will carefully examine the dataset before starting our research to make sure the data is reliable and intact. This entails looking at the data's origin, the methods used to obtain it, and any biases or constraints that might have an impact on its validity.

We will verify that the dataset's variables align with both our analytical approach and our research objectives. This means confirming the existence of important information needed for our study, like geographic identifiers, demographic characteristics, and poverty rates.

To gain a fundamental understanding of the dataset, we will utilize exploratory data analysis (EDA) to identify patterns, trends, and outliers. This will help us assess how rich and valuable thedata is for our study.

Part 3: Data Acquisition and Inspection

Before proceeding with analysis, we conducted a comprehensive inspection of the dataset to understand its structure, content, and quality. This inspection encompassed the following steps:

- Variable Identification: We identified the variables included in the dataset, paying close attention to key variables such as poverty rates, demographic characteristics, and geographic identifiers.
- Data Types and Formats: We examined the data types and formats of each variable to ensure compatibility with our analysis tools and methods. This involved identifying numerical, categorical, and text variables, as well as date formats.
- Missing Values and Outliers: We assessed the presence of missing values and outliers within the dataset, as these can impact the integrity and validity of our analysis. Strategies for handling missing data were considered, such as imputation or exclusion, based on the extent and nature of missingness.
- Data Quality Checks: We performed data quality checks to identify any anomalies or inconsistencies in the dataset. This included examining summary statistics, frequency distributions, and cross-tabulations to detect potential errors or discrepancies.

	tbl	Year	quarter	citymarketid_1	citymarketid_2	city1	city2	airportid_1	airportid_2	airport_1	airport_2
0	Table1a	2021	3	30135	33195	Allentown/Bethlehem/Easton, PA	Tampa, FL (Metropolitan Area)	10135	14112	ABE	PIE
1	Table1a	2021	3	30135	33195	Allentown/Bethlehem/Easton, PA	Tampa, FL (Metropolitan Area)	10135	15304	ABE	TPA
2	Table1a	2021	3	30140	30194	Albuquerque, NM	Dallas/Fort Worth, TX	10140	11259	ABQ	DAL
3	Table1a	2021	3	30140	30194	Albuquerque, NM	Dallas/Fort Worth, TX	10140	11298	ABQ	DFW
4	Table1a	2021	3	30140	30466	Albuquerque, NM	Phoenix, AZ	10140	14107	ABQ	PHX
...	...	...	...	...	...	...	...	...	...	...	...
242097	Table1a	2023	3	35412	31703	Knoxville, TN	New York City, NY (Metropolitan Area)	15412	12953	TYS	LGA
242098	Table1a	2023	3	35412	32467	Knoxville, TN	Miami, FL (Metropolitan Area)	15412	11697	TYS	FLL

Part 4: Data Cleaning and Preparation

Column Datatypes:

- Following data inspection, we proceeded with data cleaning and preparation to ensure that the dataset is ready for analysis. This involved addressing missing values, correcting errors, standardizing variable formats, and performing any necessary transformations or preprocessing steps.

Handling Missing Values:

- We began by addressing missing values within the dataset. Depending on the extent and nature of missingness, we employed various strategies such as imputation, deletion of rows or columns with excessive missing values, or treating missing values as a separate category, where applicable.

Dealing with Outliers:

- Outliers can distort analysis results; hence, we carefully examined variables for outliers and decided on appropriate strategies to handle them. This may involve removing outliers based on statistical criteria or transforming variables to reduce the influence of outliers.

Standardizing Variable Formats:

- Ensuring consistency in variable formats is essential for accurate analysis. We standardized variable formats such as dates, categorical variables, and numerical variables to ensure uniformity and compatibility across the dataset.

Addressing Data Integrity Issues:

- We reviewed the dataset for any data integrity issues, such as duplication or inconsistencies, and took corrective actions as necessary. This may involve merging duplicate records, resolving discrepancies between variables, or validating data against external sources.

Encoding Categorical Variables:

Categorical variables were encoded into numerical format, where appropriate, using

techniques such as one-hot encoding or label encoding. This transformation facilitates the inclusion of categorical variables in analytical models and algorithms.

Feature Engineering:

- We conducted feature engineering to derive new variables or transform existing ones to enhance the predictive power of the dataset. This may include creating new composite variables, scaling numerical variables, or extracting relevant features from existing variables.

Normalization and Scaling:

- Numerical variables were normalized or scaled to ensure comparability and improve the performance of analytical models. Common techniques include min-max scaling, z-score normalization, or robust scaling, depending on the distribution and range of values.

Handling Imbalanced Data:

- If the dataset exhibits class imbalance in categorical variables, we applied techniques such as oversampling, under sampling, or synthetic data generation to balance the distribution and prevent biased model outcomes.

Partitioning Data:

- We partitioned the dataset into training, validation, and test sets to facilitate model training, evaluation, and validation. This partitioning ensures that models are trained on a subset of data, validated on another subset, and tested on a separate unseen subset.

- Throughout the data cleaning and preparation process, we maintained comprehensive documentation of the steps undertaken, including any transformations, modifications, or decisions made. This documentation ensures transparency, reproducibility, and accountability in the data preparation process.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 242102 entries, 0 to 242101
Data columns (total 23 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   tbl             242102 non-null  object
 1   Year            242102 non-null  int64
 2   quarter         242102 non-null  int64
 3   citymarketid_1  242102 non-null  int64
 4   citymarketid_2  242102 non-null  int64
 5   city1           242102 non-null  object
 6   city2           242102 non-null  object
 7   airportid_1     242102 non-null  int64
 8   airportid_2     242102 non-null  int64
 9   airport_1       242102 non-null  object
 10  airport_2       242102 non-null  object
 11  nsmiles         242102 non-null  int64
 12  passengers      242102 non-null  int64
 13  fare            242102 non-null  float64
 14  carrier_lg      240571 non-null  object
 15  large_ms        240571 non-null  float64
 16  fare_lg         240571 non-null  float64
 17  carrier_low     240499 non-null  object
 18  lf_ms           240499 non-null  float64
 19  fare_low        240499 non-null  float64
 20  Geocoded_City1  206749 non-null  object
 21  Geocoded_City2  206749 non-null  object
 22  tbl1apk         242102 non-null  object
dtypes: float64(5), int64(8), object(10)
memory usage: 42.5+ MB
```

- RangeIndex: Indicates the range of indices for the DataFrame, which is from 0 to 242101, inclusive.
- Data columns: Shows that there are a total of 23 columns in the DataFrame.
- Last 3 columns were dropped from the data frame. We addressed missing values in certain columns by either filling missing string values with 'Unknown'. filling missing numeric values with the mean of the column.
- We defined specific data types for selected columns to ensure consistency and accuracy in our analysis.
- We converted columns to their designated data types as per our predefined mappings.

From this summary, we can see that the dataset contains information such as year, quarter, city markets, airports, number of passengers, fare information, carrier details, and some geographical information among others. The specific details of what each column represents would require referring to the dataset's documentation or context.

Part 5: Exploratory Data Analysis (EDA)

FLIGHT DATA DASHBOARD:

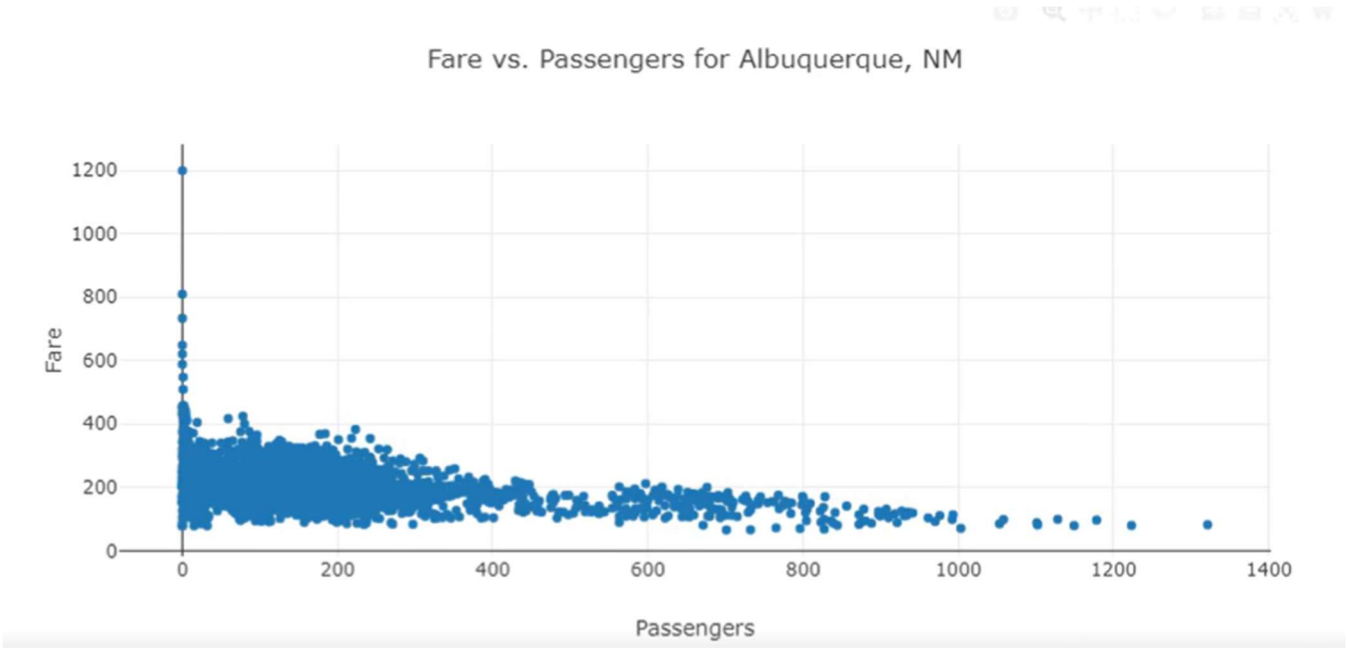
The code demonstrates how to create an interactive dashboard using Dash to visualize flight data. Users can select a city to view the correlation between fare and number of passengers, as well as the average fare by carrier for that city.

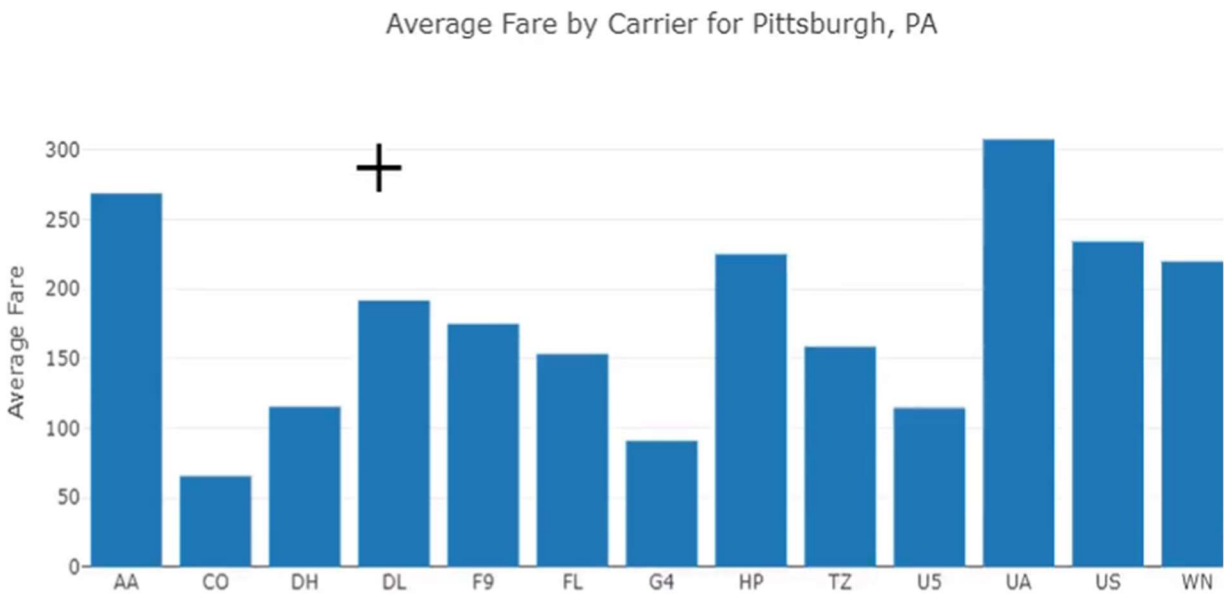
A visualization of a scatter plot, which is a type of graph that displays the relationship between two variables. In this case, the x-axis represents the number of passengers, and the y-axis represents the fare. Each data point represents a single flight.

The plot shows a positive correlation between fare and passengers. This means that as the number of passengers on a flight increases the fare also tends to increase. There are a few possible explanations for this. One possibility is that airlines charge more for flights that are in high demand. Another possibility is that airlines charge more for flights that are longer or that travel to more desirable destinations.

It is important to note that the correlation does not necessarily mean that there is a causal relationship between fare and passengers. It is also possible that other factors, such as the time of year or the day of the week, could influence both the fare and the number of passengers on a flight.

.





update_fare_by_carrier function:

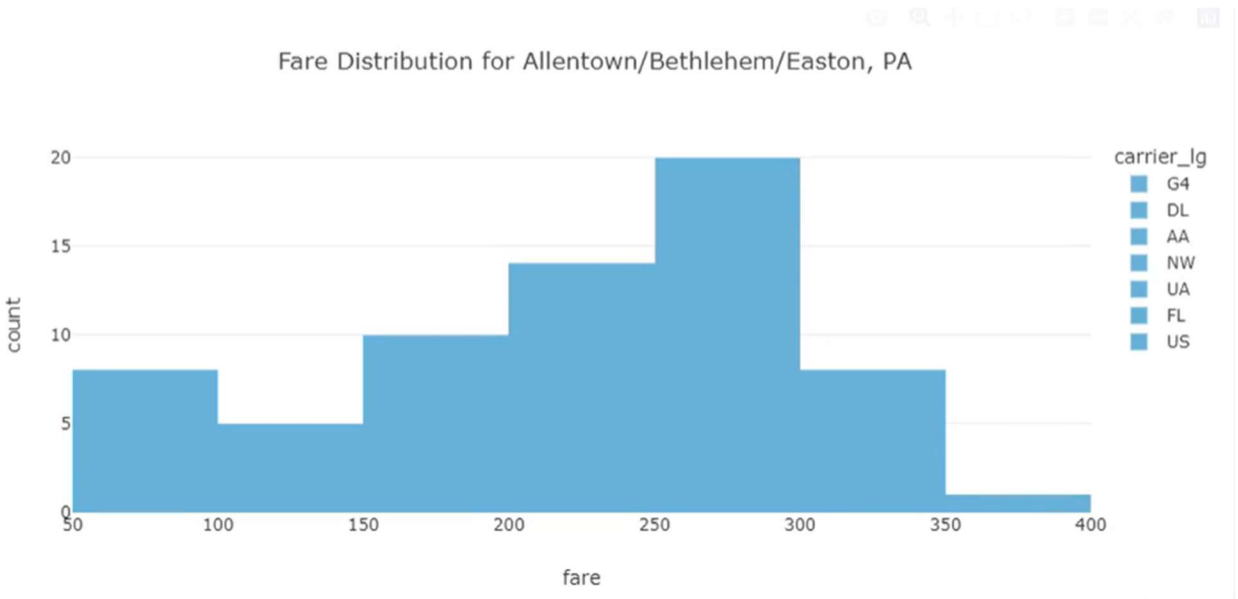
- This function is called whenever the user selects a city from the dropdown menu in the dashboard.
- It takes the selected city as input (selected_city).
- Filters the data frame (df) to include only flights departing from the chosen city (df[df['city1'] == selected_city]).
- Groups the data by carrier ('carrier_lg') and calculates the average fare ('fare') for each carrier group. This is done using the groupby function followed by mean.
- Resets the index to create a regular dataframe (reset_index()).
- Creates a bar chart ('type': 'bar') with:
    - X-axis representing the carrier names ('carrier_lg').
    - Y-axis representing the average fare for each carrier ('fare').
- The chart title and axis labels are set based on the selected city (f'Average Fare by Carrier for {selected_city}').

The image the visual output of the update_fare_by_carrier function. It displays a bar chart that compares the average fare by carrier for a specific city (which is most likely the first chosen city from the dropdown menu since the code sets the default value to the first unique city in df['city1']).

Each bar in the chart represents a carrier, and the height of the bar corresponds to the average fare for that carrier. The X-axis shows the abbreviated carrier names, and the Y-axis shows the average fare. The title at the top of the chart indicates that it shows the "Average Fare by Carrier" for the chosen city.

- Users can analyze the correlation between fare prices and passenger numbers for specific cities by viewing scatter plots dynamically updated based on user selections.
- This can help us understand how airlines adjust fares based on local demand and competition.
- This assists users in making informed decisions related to travel planning, pricing strategies, and airline selection.

FLIGHT STATS DASHBOARD:



- The bar chart would likely show the average fare for each carrier departing from a specific city
- The Y-axis would represent the count value
- The X-axis would represent the average fare, likely displayed in the currency used in the original data (USD, EUR, etc.).
- Each bar would correspond to a particular carrier, and the height of the bar would represent the average fare for that carrier's flights departing from the selected city.

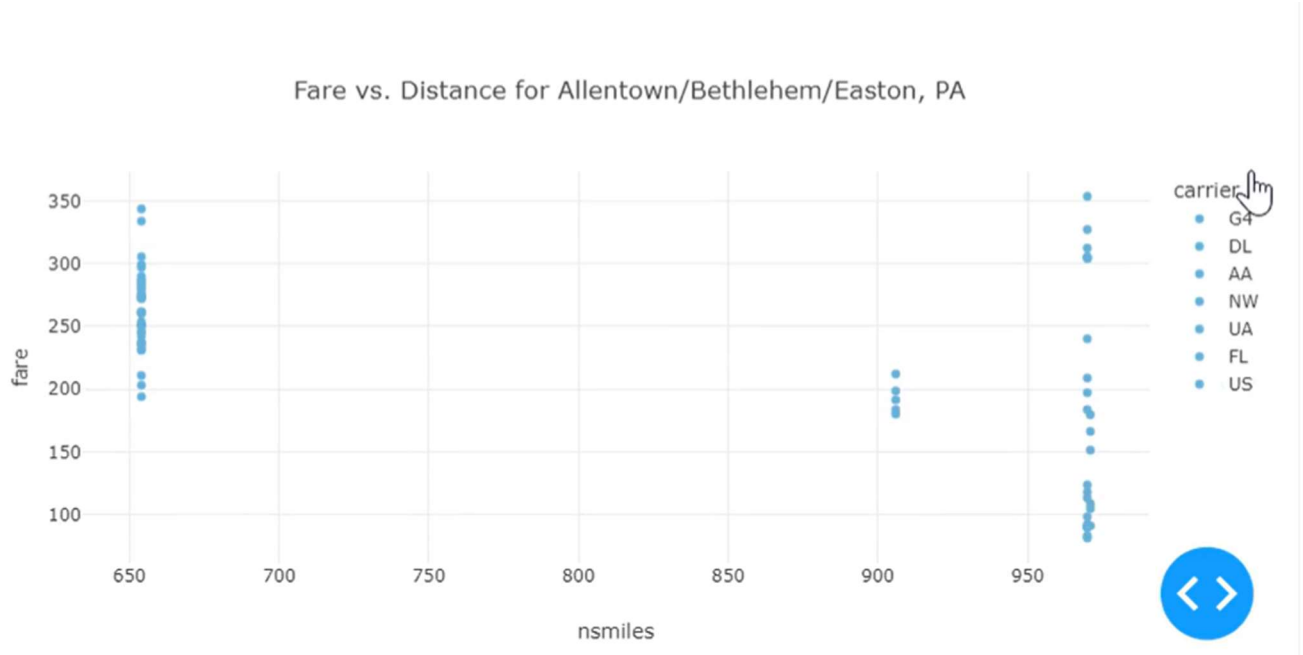
Passengers by Carrier: Illustrates the total number of passengers served by each carrier for a selected city, offering insights into carrier performance and market share. This graph assists in understanding passenger preferences and identifying dominant carriers in specific city markets.

For instance, the bar with the shortest height would represent the carrier with the lowest average fare for flights departing from that city. Conversely, the tallest bar would indicate the carrier with the highest average fare.

.

The X-axis could represent different carriers, and the Y-axis could represent the average fare or total number of passengers for each carrier departing from that city (or all cities).

Fare Histogram: Displays the distribution of fares for a selected city, providing insights into the fare ranges and frequency of occurrence. It helps users understand the fare distribution pattern, aiding in fare analysis and comparison across different cities.

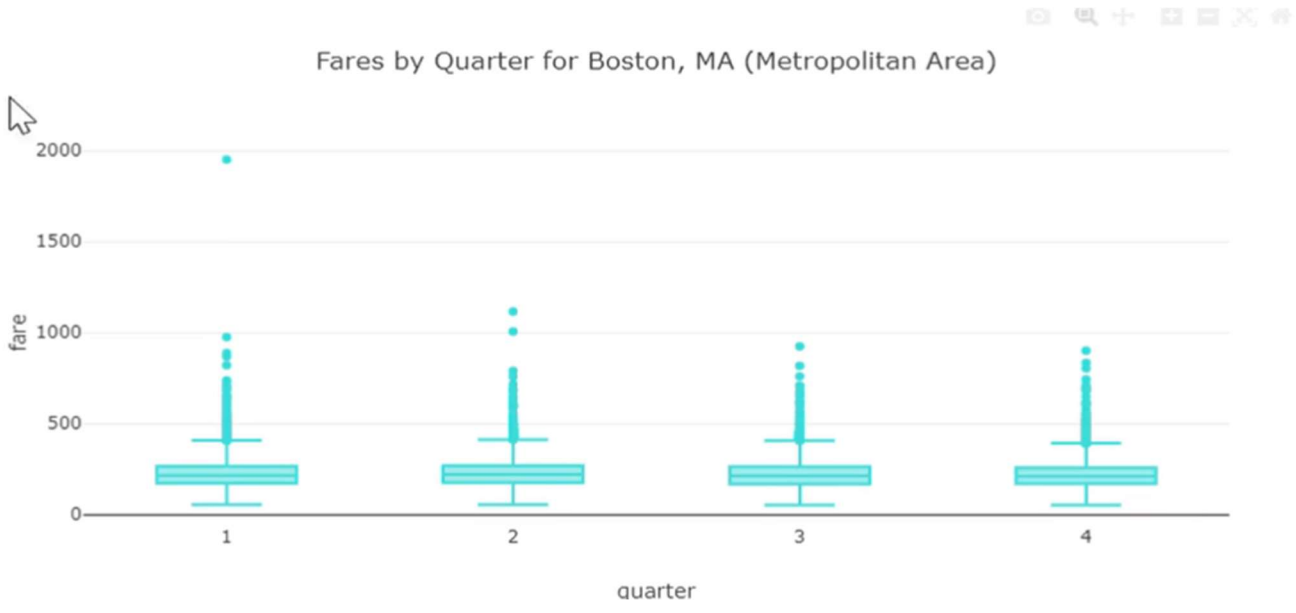


A scatter plot for fare vs. distance would typically show:

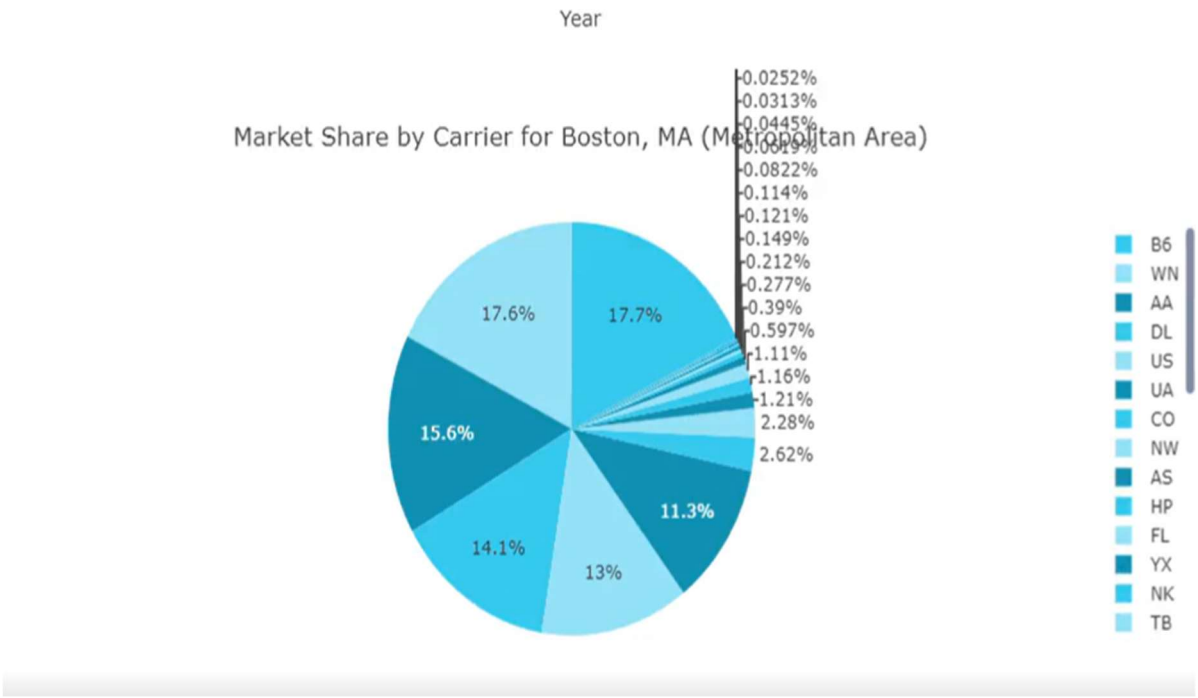
- Flights with a higher number of passengers (X-axis) might be spread across a wider range of distance (Y-axis). This is because larger aircrafts can offer a wider variety of fares (e.g., economy, business, first class).
- There might be a general trend where flights with more passengers (X-axis) tend to have slightly higher distances(Y-axis) on average.
- A blue circle in the upper right corner of the chart. This could represent a flight offered by a major carrier (blue) that has a high number of passengers (X-axis) and a high distance (Y-axis). Conversely, a blue dot in the bottom left corner might represent a budget carrier (blue) with a low number of passengers (X-axis) and a lower distance (Y-axis).
- **Fare vs. Distance Scatter Plot**: Shows the relationship between fare and distance for flights originating from a selected city, helping users analyze fare trends in relation to travel distance. This graph aids in understanding fare dynamics based on travel distance, supporting route planning and fare optimization strategies.

INSIGHTS INTO AIR TRAVEL DATA:

In some variations, individual data points (fare values for each flight) might be overlaid on the box plots, providing a more granular view of the data distribution.



The lines extending from the box represent the range of the remaining data points. They typically extend to the most extreme values within 1.5 times the interquartile range (IQR) from the quartiles. Data points beyond these whiskers are considered outliers and are often plotted as individual points.
A wider box indicates a larger spread of fares within that quartile. This suggests more variability in fares for flights within that fare range.
This visualization offers a more detailed picture of fare distribution compared to a regular box plot. It reveals how the "spread" and "typical fare" vary across different fare ranges.
This graph displays the distribution of fares across different quarters for the selected city, offering insights into seasonal fare variations.



Shows the market share distribution of different airlines (carriers) for flights departing from Boston, Massachusetts (assuming the data focuses on departing flights). Here's a breakdown of the elements and interpretation:

- **Slices:** The pie chart is divided into slices, each representing a different carrier operating flights departing from Boston.
- **Slice Size:** The size of a slice corresponds to the market share of the corresponding carrier. A larger slice indicates a higher market share, meaning that carrier offers a larger portion of the total flights departing from Boston compared to other carriers.
- **Percentages:** The pie chart likely includes labels displaying the percentage value for each slice. This value represents the proportion of total passengers that flew with that specific carrier out of all passengers departing from Boston on flights included in the data.

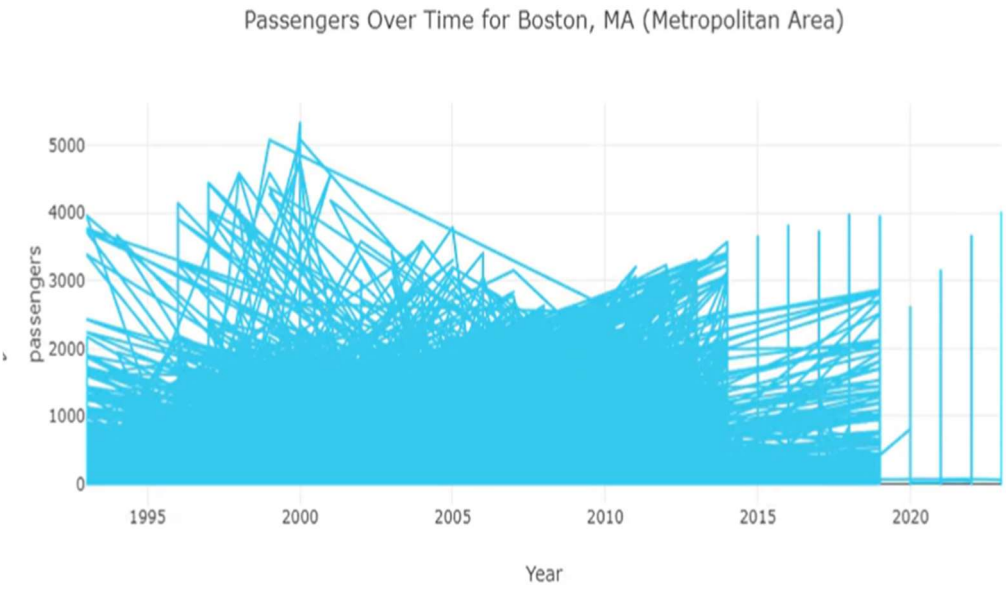
The pie chart has a large blue slice labeled "UA (15.6%)" and a smaller green slice labeled "AA (11.3%)". This would indicate that United Airlines (UA) captures a larger portion (15.6%) of the market for departing flights from Boston compared to American Airlines (AA) (11.3%) based on the data used to create the chart.

This chart provides a clear visualization of the market share held by different carriers operating in the selected city, aiding in understanding the competitive landscape of the aviation market.

The line chart provides a visual representation of how the number of passengers has changed over time for Boston. By examining the trends and data points, you can gain insights into historical passenger volumes and potential factors that might have influenced them.

The line starts relatively low in 1995, increases steadily until around 2000, and then plateaus or slightly declines. This could suggest that passenger volume at airports in the Boston area grew over a decade or so, and then reached a stable level or experienced a slight decrease.

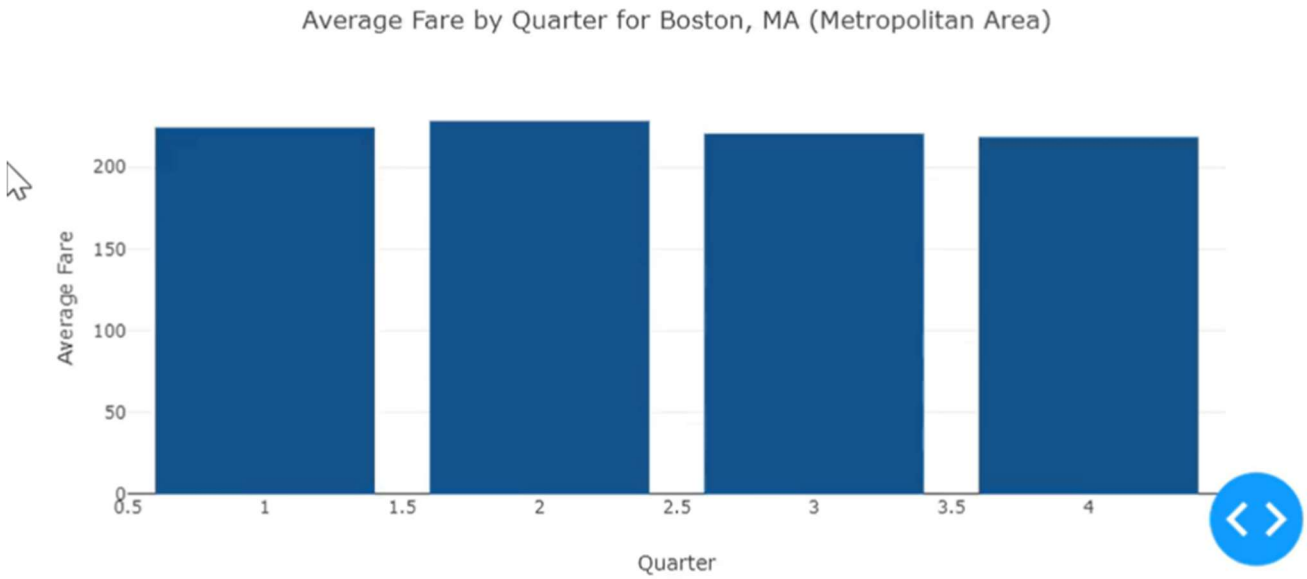
By illustrating the trend of passenger numbers over time, this visualization helps identify patterns and trends in air travel demand for the chosen city.



This axis likely represents the average fare, possibly in dollars (USD). The scale likely starts at a point and increases by intervals.

Each bar corresponds to a specific quarter. The height of the bar represents the average fare for that quarter.

Imagine the bar for Q2 (second quarter) is the highest, followed by Q1(first quarter), and then Q3(third quarter) and Q4 (fourth quarter) have the lowest average fares. This could suggest a seasonal trend where average fares are highest in the summer (Q2) and possibly spring (Q2), and then lower in other parts of the year.



Showing the average fare for each quarter, this visualization offers an overview of fare fluctuations throughout the year, supporting strategic pricing decisions for airlines.

CONCLUSION:

To sum up, This report has explored various aspects of flight data for Boston, Massachusetts (metropolitan area). By analyzing the data and using interactive visualizations, we gained insights into passenger volume trends, market share distribution among carriers, the relationship between fare and passenger numbers, and potential seasonal patterns in average fares.

The scatter plot generated through the interactive dashboard revealed a positive correlation between the number of passengers and fare. This suggests that flights with a higher passenger capacity might offer a wider range of fares, including potentially higher fares for business or first-class seating. The pie chart, likely focusing on departing flights from Boston, provided a snapshot of the market share distribution among different carriers. This information can be valuable for understanding the competitive landscape of airlines operating in the Boston area.

Furthermore, the line chart depicting passenger volume over time offered a historical perspective. It showed a potential growth trend in passenger numbers, followed by a period of stability or slight decline. This trend analysis can be helpful for understanding past passenger volumes and potentially forecasting future air travel demand in Boston. Finally, the bar chart representing average fare by quarter hinted at a possible seasonal pattern, with potentially higher fares during peak travel seasons like summer.

Overall, the data and visualizations presented in this report provide valuable insights for stakeholders in the Boston air travel industry. Airlines can leverage this information to understand passenger preferences, optimize pricing strategies, and potentially adjust their offerings to cater to different passenger segments. Additionally, policymakers and airport authorities can use these insights to plan for future infrastructure needs and ensure the Boston area remains a well-connected air travel hub.

Key Points to Consider for Further Analysis:

- While this report focused on data for Boston, similar analyses could be conducted for other cities or regions to identify potential similarities or differences in passenger volume trends, market share distribution, and fare patterns.
- The data visualizations presented here offer a starting point for further exploration. Additional data sources could be incorporated to provide a more comprehensive picture of the air travel landscape in Boston, such as information on specific routes, flight durations, or airline on-time performance.
- Combining insights from passenger volume trends with data on economic indicators or travel industry trends could provide valuable context for understanding the factors influencing air travel demand in Boston.

By building on the foundation established in this report and incorporating these additional considerations, a deeper understanding of the air travel landscape in Boston can be achieved. This knowledge can be instrumental in making informed decisions that benefit airlines, policymakers, airport authorities, and ultimately, the traveling public.