Airfare Data Analysis and Prediction

Milestone: Project 1

Group 10

Nishant Upadhyay

Aditya Kumar

Priyansh Nileshbhai Vagadiya

upadyay.nis@northeastern.edu

kumar.aditya1@northeastern.edu

vagadiya.p@northeastern.edu

Signature of student: Nishant Upadhyay (33.33%)

Signature of student: Aditya Kumar (33.33%)

Signature of student: Priyansh Nileshbhai Vagadiya (33.33%)

Submission Date: 02/15/2025

Project Report: Airfare Data Analysis and Prediction

Course: IE6600 Computation and Visualization for Analytics
Semester: Spring 2025
Project: Data Analysis and Visualization with a Focus on Static Visualizations and Statistical Analysis
Dataset: Consumer Airfare Report (data.gov)
Machine Learning Model Used: XGBoost Regressor

1. Introduction:

The objective of this project is to analyze airfare data, perform statistical analysis, and create static visualizations using Matplotlib. Additionally, a predictive model is developed using machine learning to estimate airfare prices. The dataset used for this project comes from data.gov and contains detailed fare information across various airline markets.

2. Data Acquisition and Inspection:

2.1. Dataset Overview:

The dataset includes 14,881 entries with 21 columns. Key attributes in the dataset include:

- Market Fare (mkt_fare): The cost of airfare in different markets.
- Year and Quarter: Temporal information for tracking trends.
- Passenger Share (carpaxshare): The proportion of passengers traveling with a specific airline.
- Average Fare (caravgfare): The mean fare per airline in a given market.
- Fare Distribution (fareinc_min, fareinc_max): The minimum and maximum airfare in different markets.

2.2. Data Inspection:

Initial analysis revealed:

- No duplicate rows were found.
- Missing values in columns Geocoded_City1 and Geocoded_City2, which were dropped.
- Categorical variables (city names, airline names) were encoded numerically for analysis.

3. Data Cleaning and Preparation:

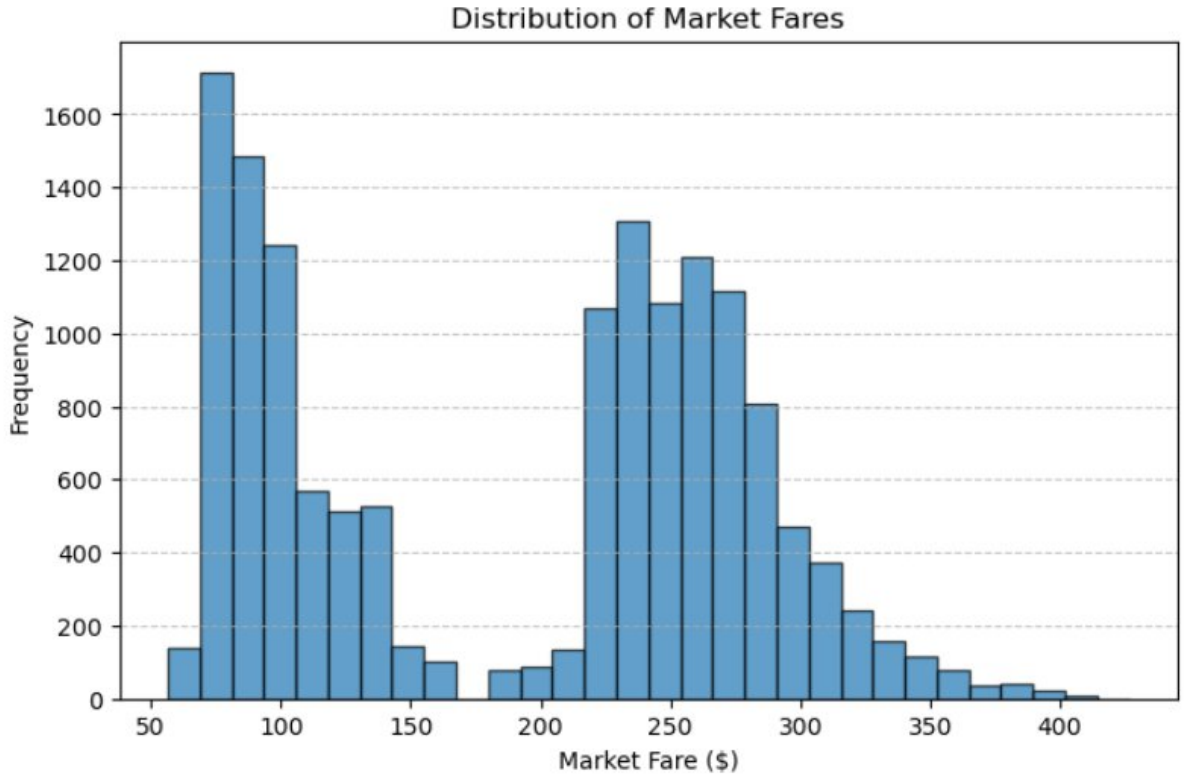
To ensure data quality,

- Missing values in numerical columns were replaced with median values.
- Categorical variables (cities, airlines) were converted to numerical codes.
- Dataset was normalized and structured for analysis.

4. Exploratory Data Analysis (EDA) and Visualization

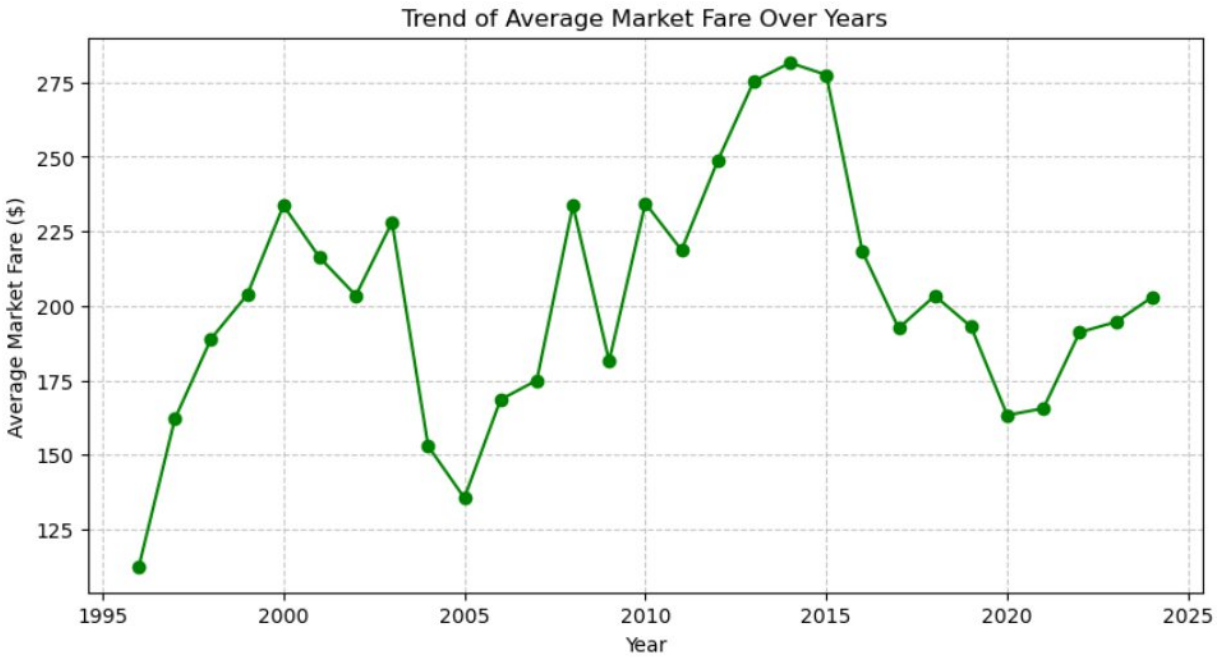
Static Visualizations (Matplotlib):

a. Histogram of Market Fare Distribution



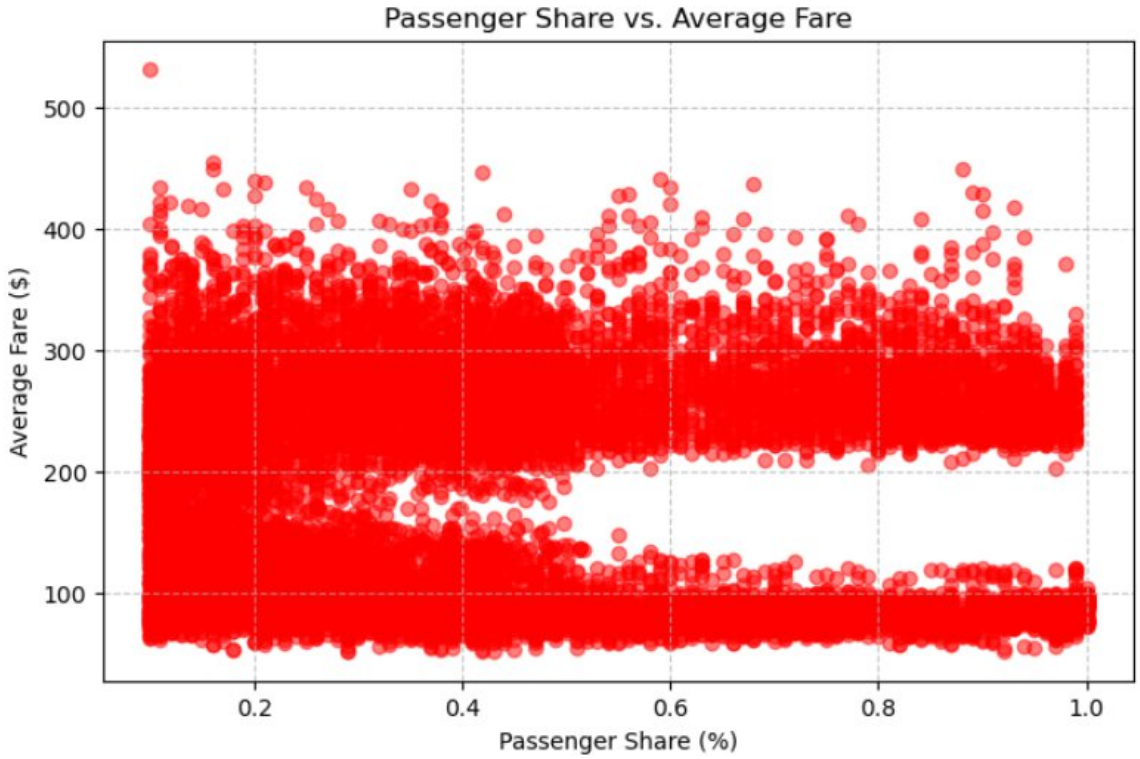
The histogram showed that airfare prices are right-skewed, meaning most fares are lower-priced with some high-cost outliers.

b. Line Plot: Trend of Average Market Fare Over the Years



A fluctuating trend was observed, indicating external factors such as fuel costs and economic conditions affecting airfare pricing.

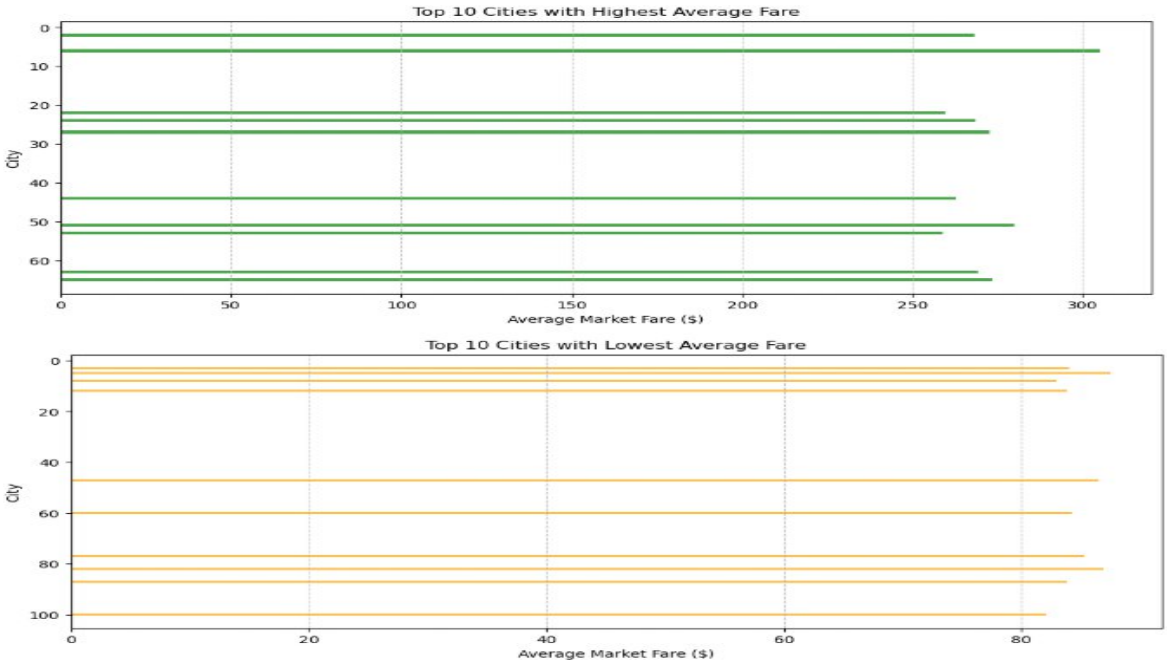
c. Scatter Plot: Passenger Share vs. Average Fare



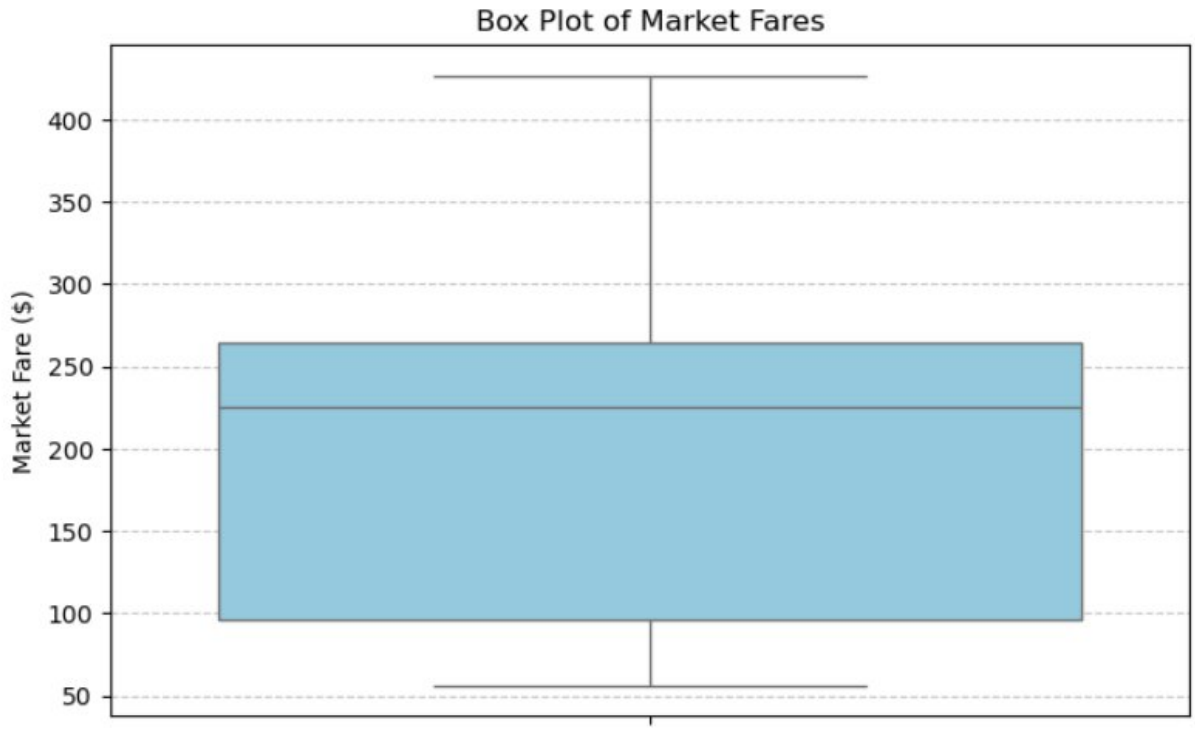
A weak positive correlation was found, suggesting that airlines with higher passenger shares tend to have competitive pricing.

d. Bar Charts: Top 10 Cities with Highest and Lowest Fares

Cities with higher competition had lower fares, while monopolistic markets had higher fares.



e. Box Plot of Market Fares



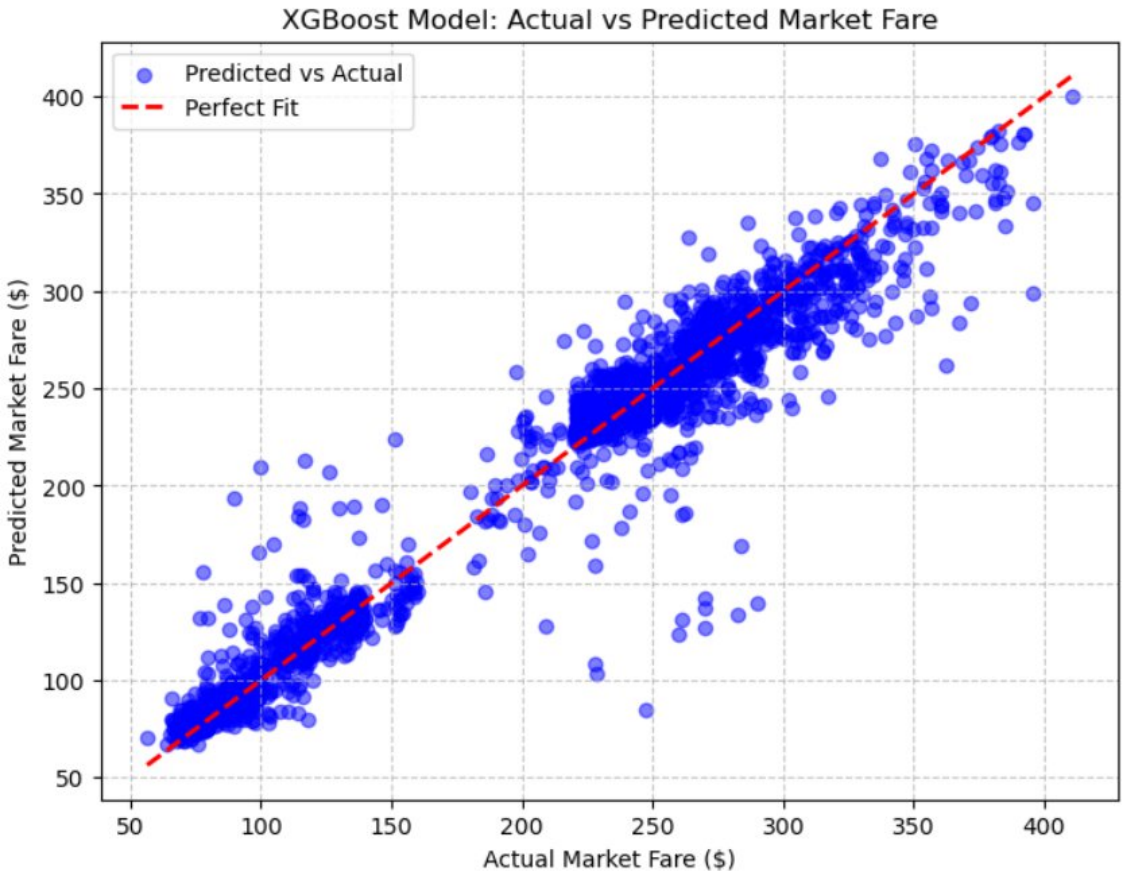
5. Statistical Analysis:

Key Statistical Findings:

- Mean Market Fare: $191.42
- Median Market Fare: $225.69
- Standard Deviation: $87.37
- 95% Confidence Interval for Mean Fare: ($190.01, $192.82)
- Correlation Between Passenger Share and Average Fare: 0.033 (Weak correlation)

6. Machine Learning Model: XGBoost Regressor

To predict market fares, we used XGBoost Regressor model



Model Performance:

- Mean Absolute Error (MAE): $9.77
- Mean Squared Error (MSE): $287.01
- Root Mean Squared Error (RMSE): $16.94
- R-squared Score (R^2): 0.9628 (96.28% accuracy)

Insights:

- The model performed exceptionally well, explaining 96.28% of the variance in airfare prices.
- The low RMSE and MAE indicate precise fare predictions.
- Factors such as year, quarter, airline share, and fare distribution played a crucial role in prediction accuracy.

7. Conclusion and Recommendations

Key Findings:

- Market fares vary significantly over time and location.
- The passenger share weakly affects pricing, suggesting competition influences fares.
- The dataset is skewed with outliers, indicating irregular airfare pricing patterns.
- The XGBoost model successfully predicts fares with high accuracy.

Recommendations:

- ❖ Airlines should strategically adjust pricing based on demand trends.
- ❖ Additional factors (e.g., fuel prices, seasonality, customer demand) could further improve predictions.
- ❖ Future work can involve time-series forecasting models (e.g., ARIMA) for long-term fare trend predictions.