New York — Chicago

Flight Price Predictor

Aida Sarinzhipova, Bradley Stoller, Devi Mahajan, Cassandra Maldonado, Kirthi Rao, & Kyler Rosen

Passenger Name:
Professor Navarro

Flight Number:
ADSP31006

Seat:
IP02

Gate:
UChicago

Date:
March 12, 2025

Agenda

1. Business Case & Problem Statement
2. Modeling Approach & Methodology
3. Data Understanding
4. Exploratory Data Analysis (EDA)
5. Preprocessing & Feature Engineering
6. Modeling Results & Analysis
7. Recommendations & Future Work

Business Case & Problem Statement

Business Case & Problem Statement

Travelers often struggle with knowing when to book flights to secure the lowest prices. Airline pricing algorithms are complex, and last-minute purchases often result in higher costs.

By using time series forecasting techniques and machine learning models, we aim to:

- Identify price trends over time
- Predict future ticket prices based on historical data
- Provide insights for consumers to make cost-effective purchasing decisions

Integration into Commercial Platforms:

- The predictive model can be integrated with travel booking websites, airline pricing tools, and consumer apps to help users make cost-effective flight purchases.
- It can serve as a dynamic price alert system, notifying travelers when fares are likely to rise or fall.

Modeling Approach & Methodology

Modeling Hypothesis & Assumptions

Modeling Hypothesis:

- Airline ticket prices tend to increase as the departure date approaches
- Fluctuation patterns are influenced by the time remaining until departure
- Historical price data analysis helps identify these patterns
- Predictive models can be developed to forecast future price trends

Key Assumptions:

- **Non-Stationarity:** Means and variances change over time
- **Seasonality:** Holidays, weekends, and major events affect prices
- **Correlation:** Price is influenced by multiple factors (e.g. travel duration, distance, seats remaining, etc.)
- **External factors:** Some demand fluctuations not in the data

Proposed Models

Traditional Time Series Models:

- ARIMA
- SARIMA
- Prophet

Machine Learning Models:

- Decision Tree
- Random Forest
- XGBoost

Deep Learning Models:

- RNN
- LSTM

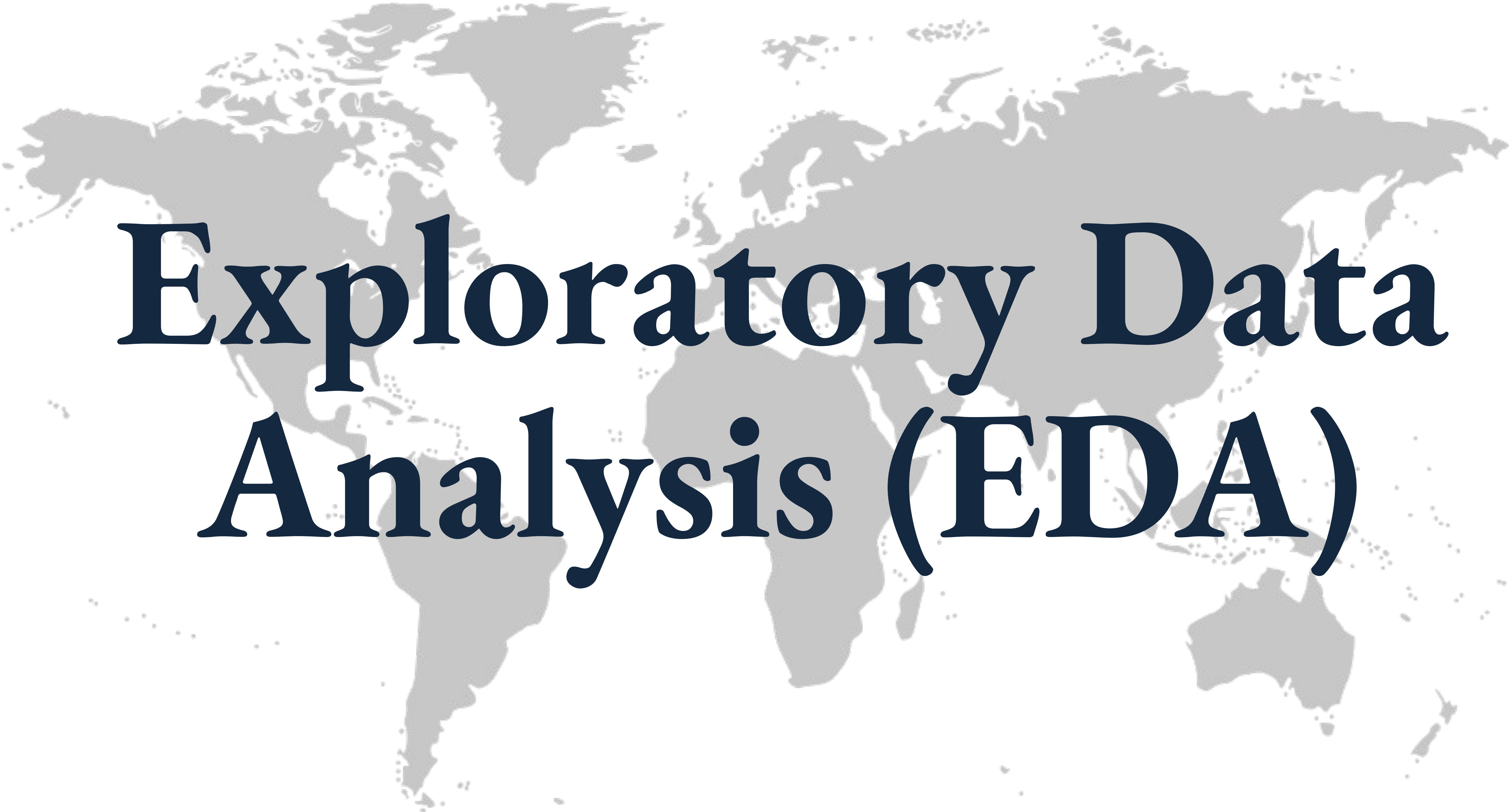
Data Understanding

Data Description & Properties

Dataset: Flight Price Dataset from Kaggle

- **82 million flight itineraries** with **46 features** related to the flight itineraries
- Each row is a purchasable ticket on Expedia between **April 16 and October 5 of 2022**
- Flights were to/from the following airports: ATL, DFW, DEN, ORD, LAX, CLT, MIA, JFK, EWR, SFO, DTW, BOS, PHL, LGA, IAD, OAK

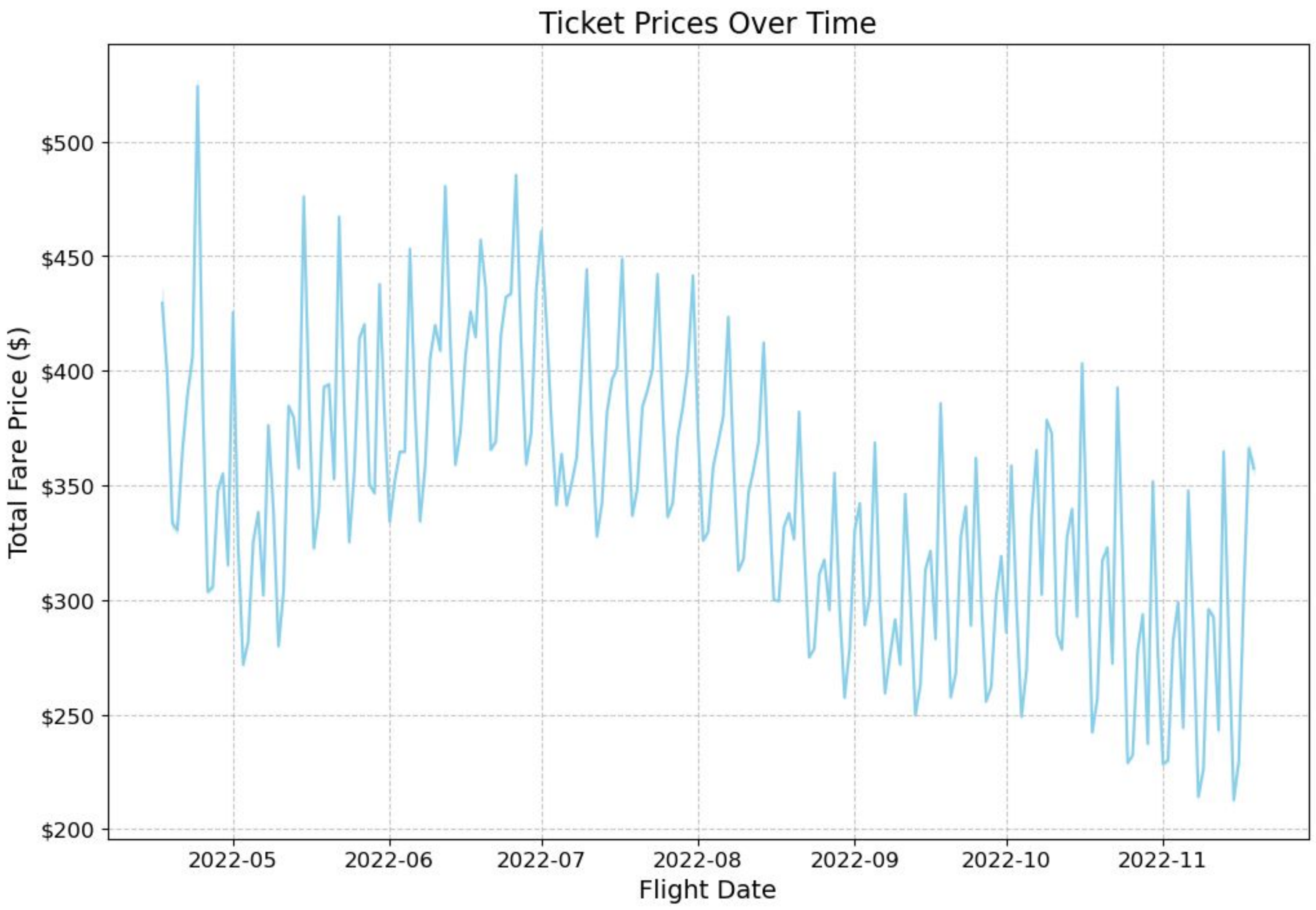
Key Variables	Description
totalFare	Total fare of ticket at time of search date.
searchDate	Date price is being checked.
flightDate	Date of departure.
startingAirport	Flight origin airport.
destinationAirport	Flight destination airport.
totalTravelDistance	Distance from origin to destination airport.



Exploratory Data Analysis (EDA)

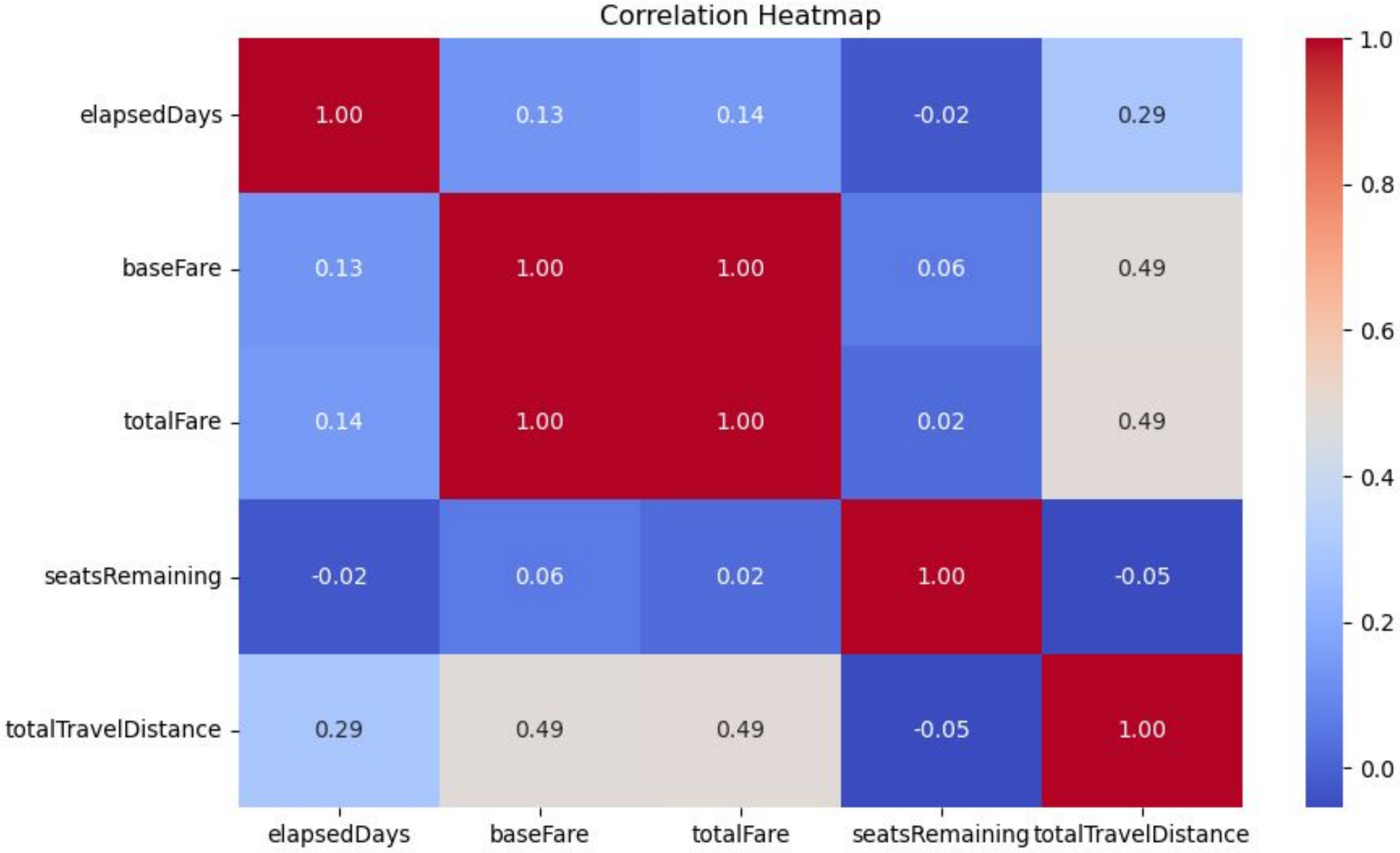
Fare Changes

- Prices exhibit fluctuations, with a slight downward trend leading up to flight dates
- The variability suggests potential seasonal or demand-driven changes in fare pricing
- Useful for predicting how fares change over time



Correlations

- Total fare is strongly correlated with base fare
- Elapsed days has a weaker correlation, indicating last-minute bookings may not be as influential
- Pricing strategies are largely independent of remaining seats, meaning last-minute sales do not heavily affect cost

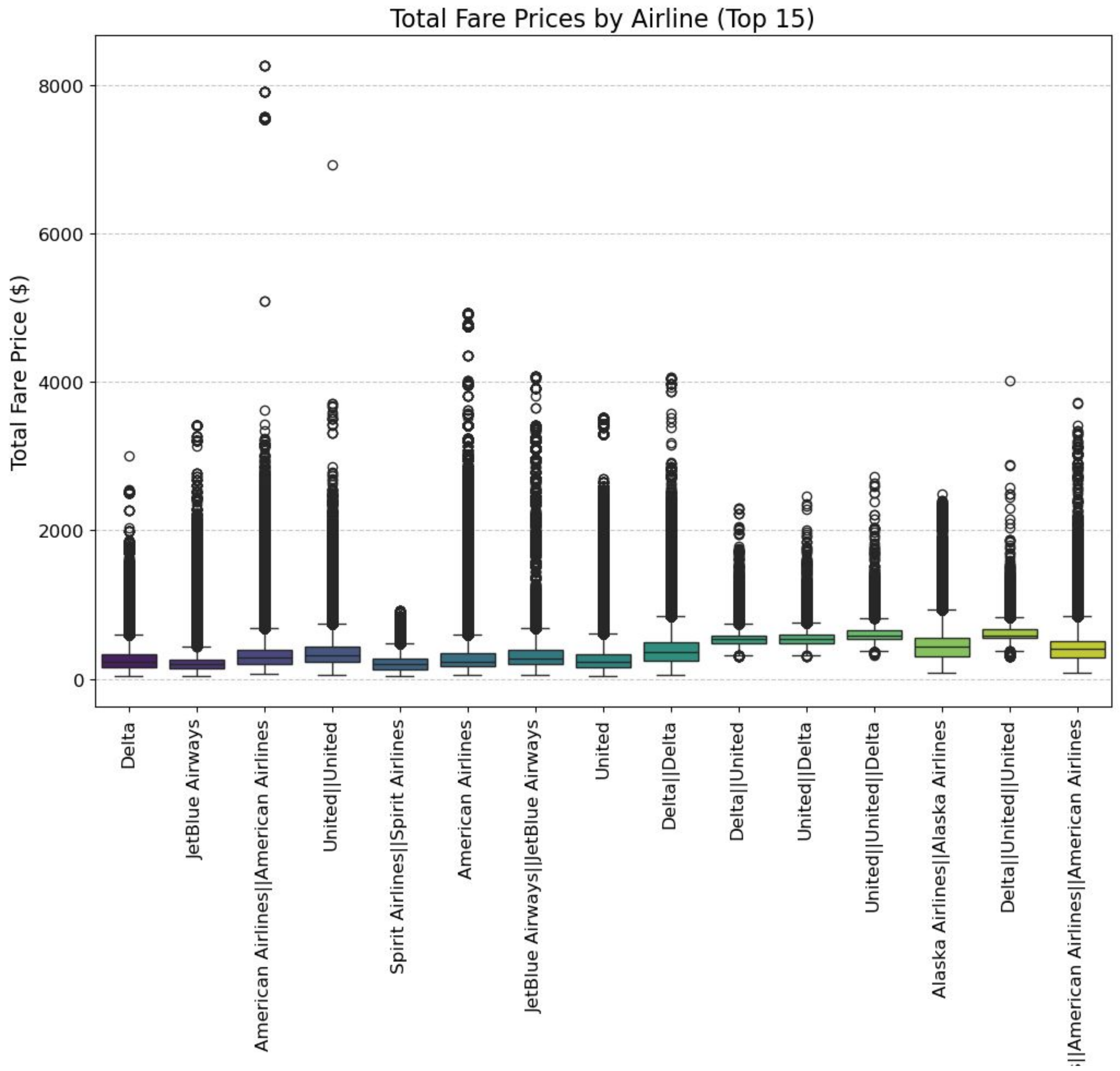


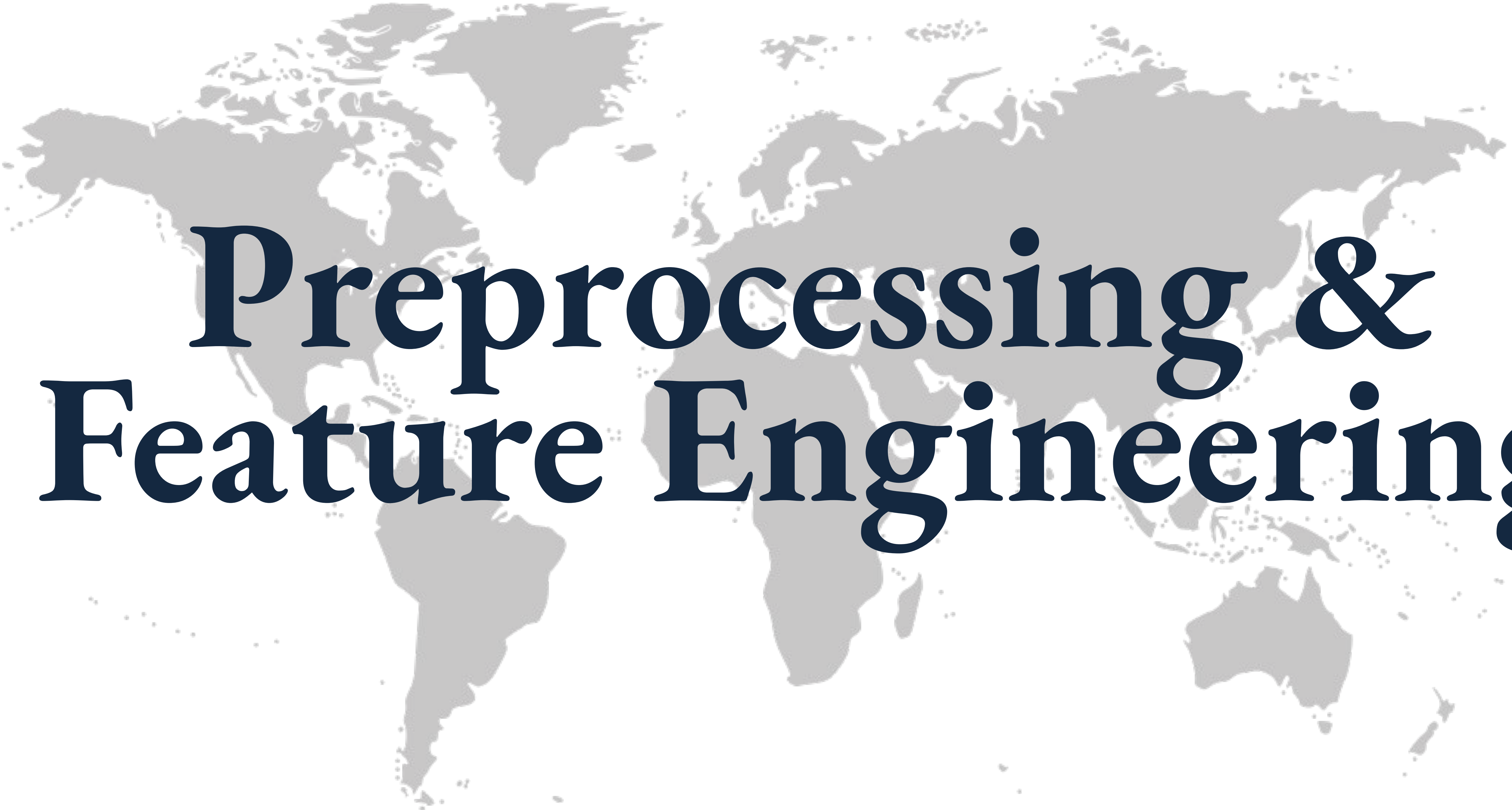
Correlation Heatmap

	elapsedDays	baseFare	totalFare	seatsRemaining	totalTravelDistance
elapsedDays	1.00	0.13	0.14	-0.02	0.29
baseFare	0.13	1.00	1.00	0.06	0.49
totalFare	0.14	1.00	1.00	0.02	0.49
seatsRemaining	-0.02	0.06	0.02	1.00	-0.05
totalTravelDistance	0.29	0.49	0.49	-0.05	1.00

Boxplot of Ticket Prices

- Ticket prices show significant variation across airlines, but airline-specific pricing strategies are less predictive than originally assumed.
- Premium airlines tend to have higher price volatility, while budget airlines maintain a narrower price range.
- Feature engineering excluded airline code due to its limited impact on model performance, shifting focus to more predictive variables like travel distance and departure time.





Preprocessing & Feature Engineering

Preprocessing & Feature Engineering

1. **Derived Features:**
   - daysToDeparture (difference of searchDate and flightDate)
   - Binned seatsRemaining into categories (low, medium, high availability)
   - Extracted temporal features (e.g., month, weekday, hour of departure)
   - Created a holiday indicator → Flights near holidays flagged for price impact.
2. **Missing Data Handling:**
   - Imputation of missing totalTravelDistance values
3. **Sequence-Based Additions:**
   - Lag prices from T-1 through T-7 were added to each row

Modeling Results & Analysis

Model Comparison

Model	MAE	RMSE	MAPE	R^2
SARIMA	$143	210	44%	-0.48
Prophet	$121	168	34%	-0.31
Decision Tree	$64	114	23%	0.65
Random Forest	$58	105	20%	0.70
XGBoost	**$56**	**104**	**19%**	**0.71**
RNN	$110	190	39%	0.25
LSTM	$91	163	31%	0.44

XGBoost performed best in all four of the error metrics

RNN & LSTM struggled due to high price variance

Performance & Evaluation

XGBoost Feature Importance

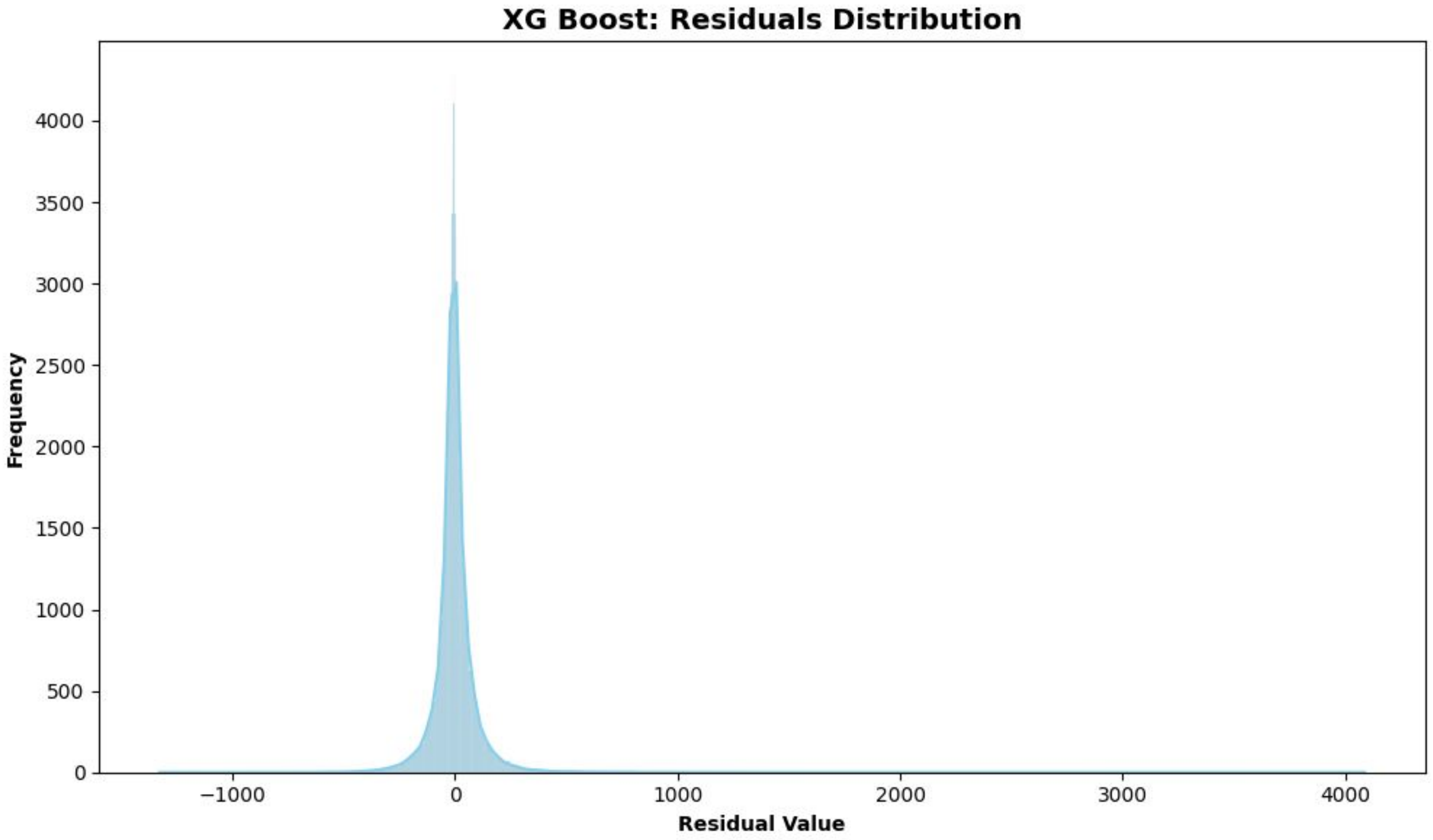
Hyperparameter tuning applied with GridSearchCV (288 fits)

Best parameters found:

- colsample_bytree: **1.0**
- gamma: **0**
- learning_rate: **0.05**
- max_depth: **12**
- min_child_weight: **3**
- n_estimators: **200**
- Cross-validation RMSE: **60.81**

XGBoost	
MAE	56.86
RMSE	104.65
MAPE	19.73 %
R^2	0.71

Residual Error Analysis



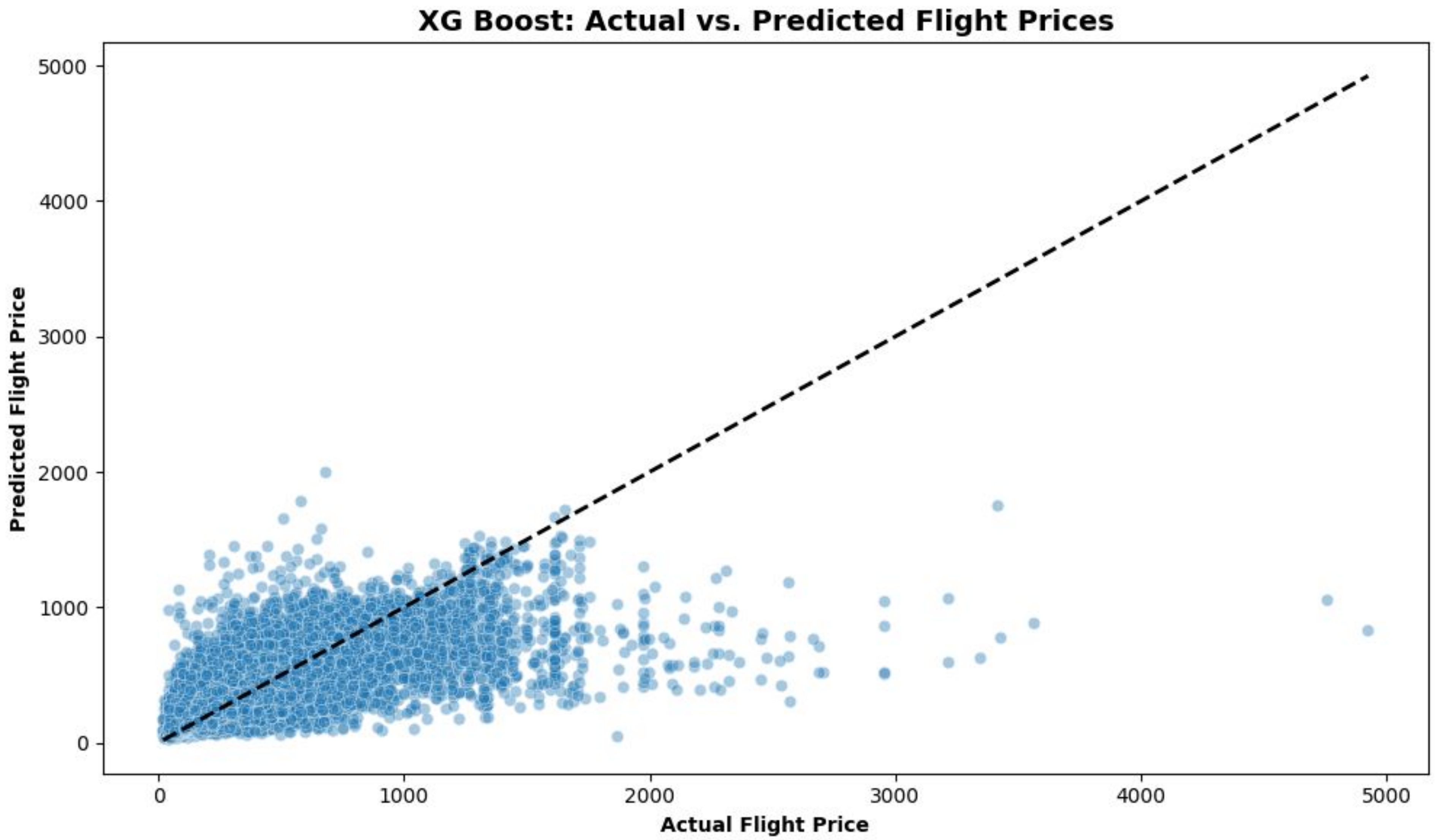
Residual distribution is largely normal (Gaussian) with a mean of 0

Some extreme outliers caused by last-minute flight fluctuations

Actual vs. Predicted Value Analysis

XGBoost captures the most variation in price, but struggles as fares increase

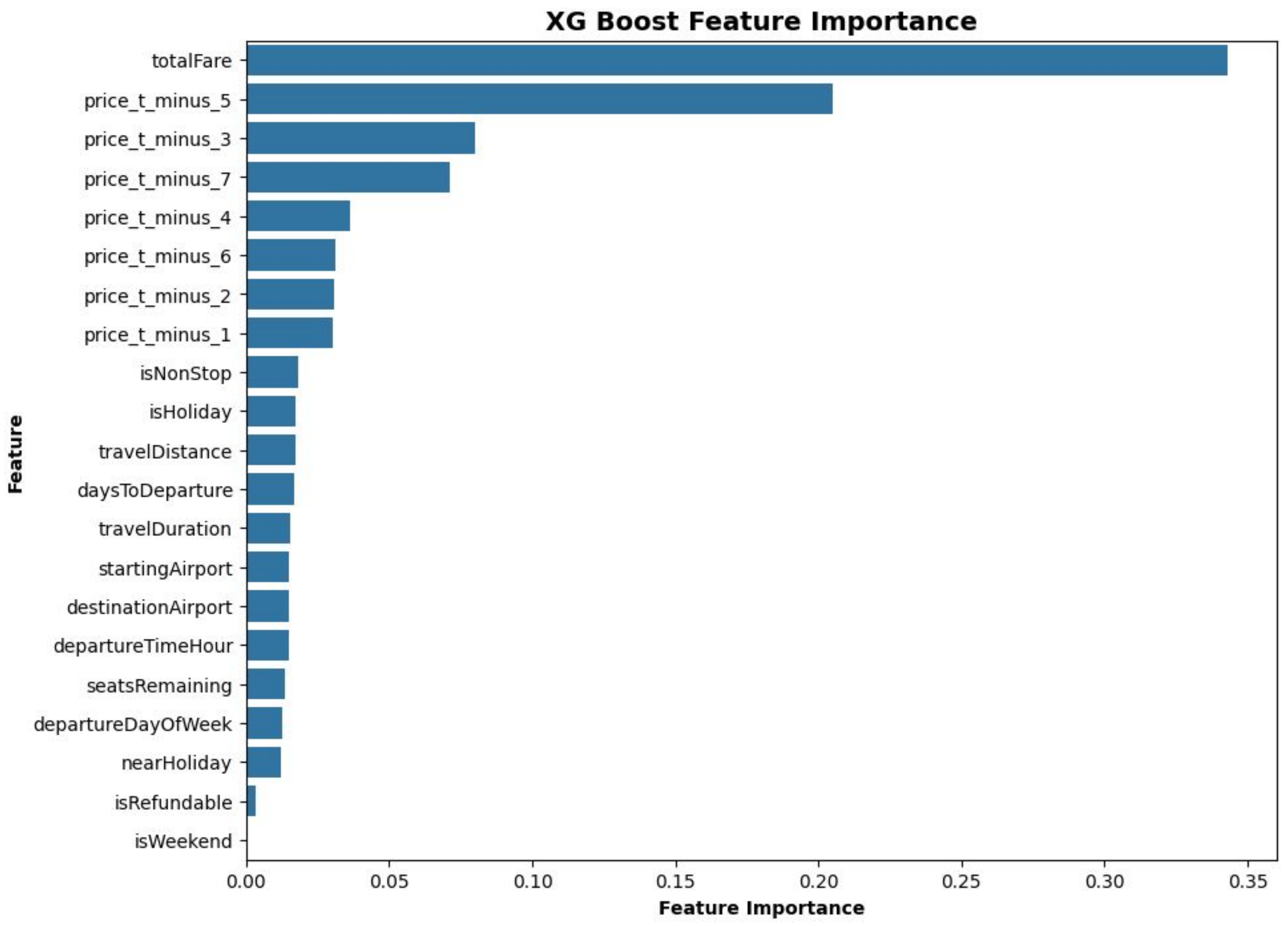
Residual variation suggests exogenous factors are uncaptured by the model



Feature Importances

Total Fare (today's price) is the #1 predictor of tomorrow's price

Lagged prices beat out all other features/predictors



To the demo!

Recommendations & Future Work

Recommendations & Future Work

1. **Real-time Prediction System:** Deploy a model API that continuously updates with new flight data
2. **External Data Sources:** Incorporate weather conditions, real-time seat availability, and airline promotions to improve accuracy
3. **Hyperparameter Tuning:** Further refine model parameters for even lower error rates
4. **Transformer-based Models:** Consider using Time Series Transformers to improve price forecasting
5. **Integration:** Future work should integrate real-time airline promotions & seat availability for better accuracy

Thank you!

Appendix

Data Properties

Data Challenges

1. Total travel distance was imputed using median-based techniques.
2. Categorical variables (e.g., airport codes, airlines) → Encoded using one-hot encoding.
3. Non-stationary trends in ticket prices → Feature transformations were applied.

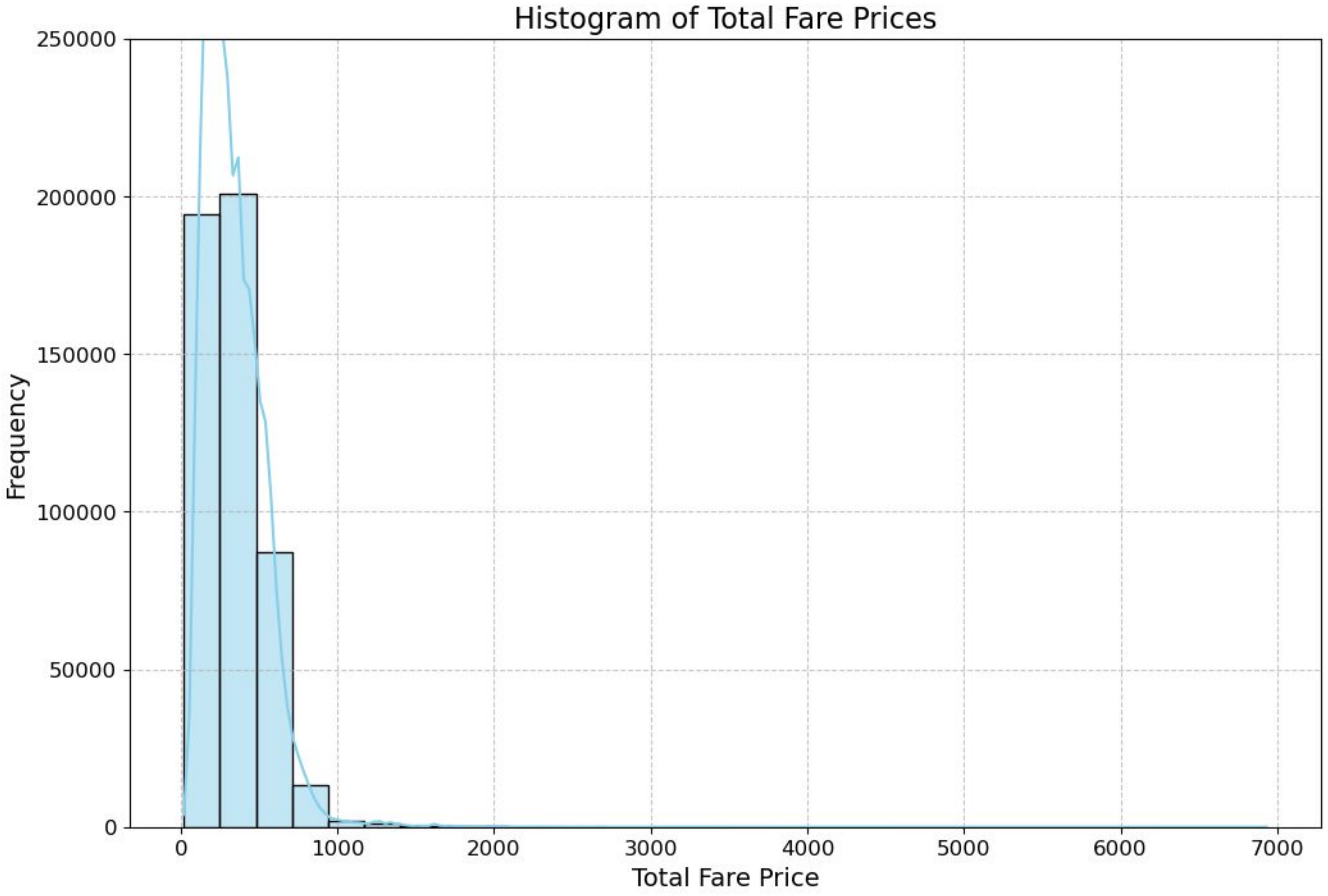
Key Data Insights

1. Flight prices increase closer to departure, but the rate varies based on seasonality.
2. Some airlines exhibit more volatile pricing trends than others.
3. The number of available seats influences price, lower availability leads to higher fares.

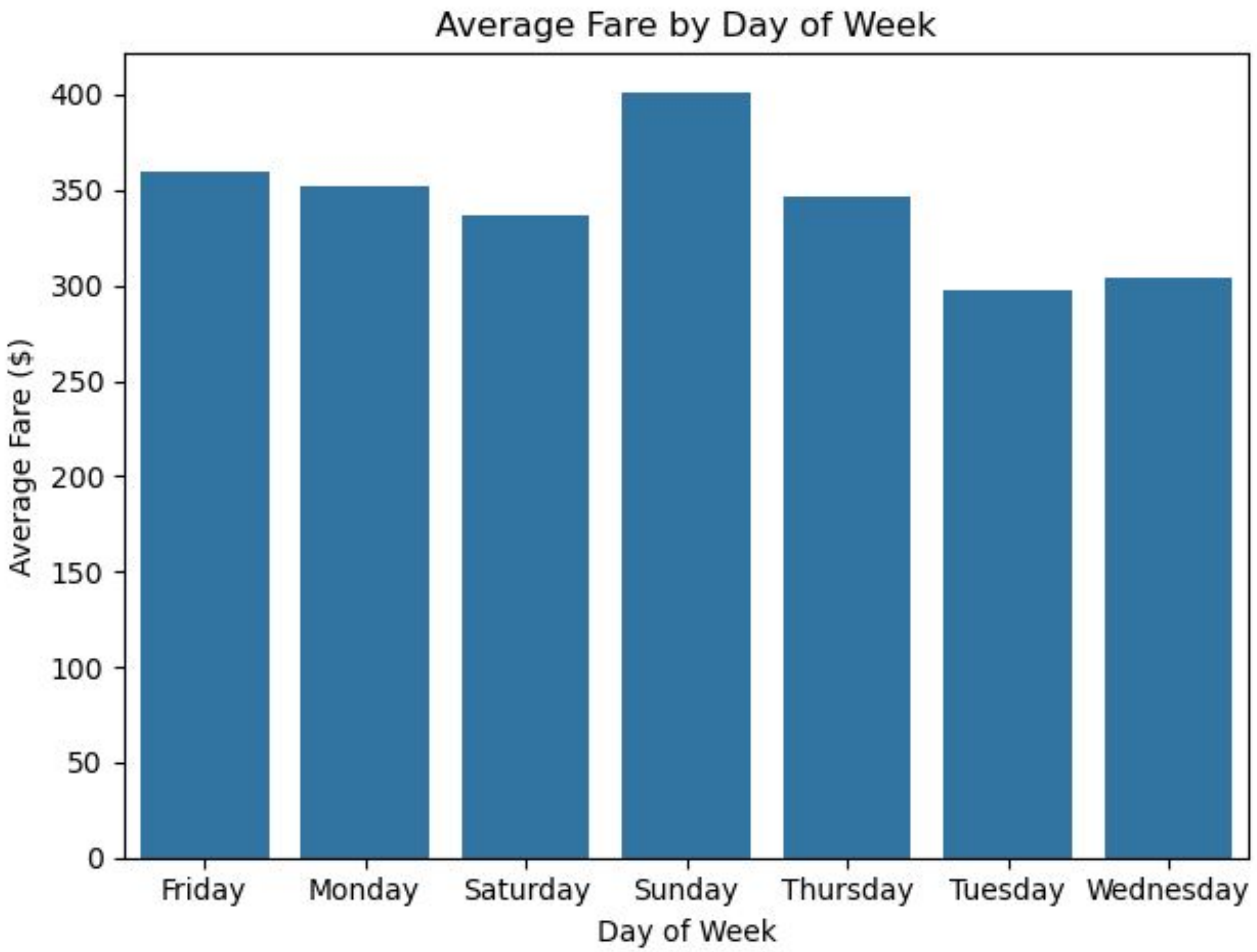
EDA

Histogram of Total Fare Prices

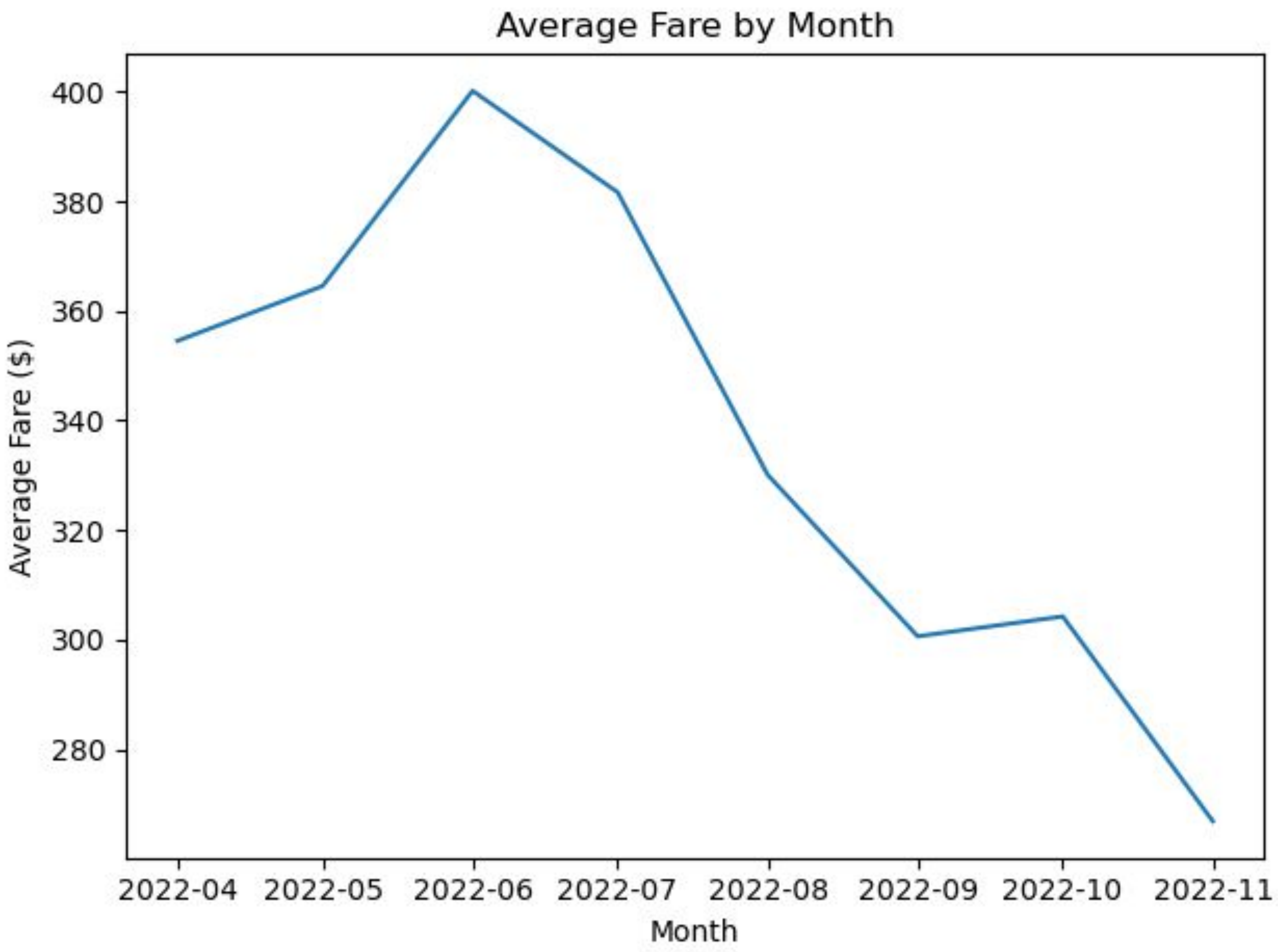
- Most fares are concentrated below $1,000, with a steep drop-off afterward
- There are a few outliers with significantly higher prices (premium/business class tickets)
- This highlights the skewed nature of flight pricing, suggesting the need for transformations in modeling



EDA



Average Fare by Day of Week

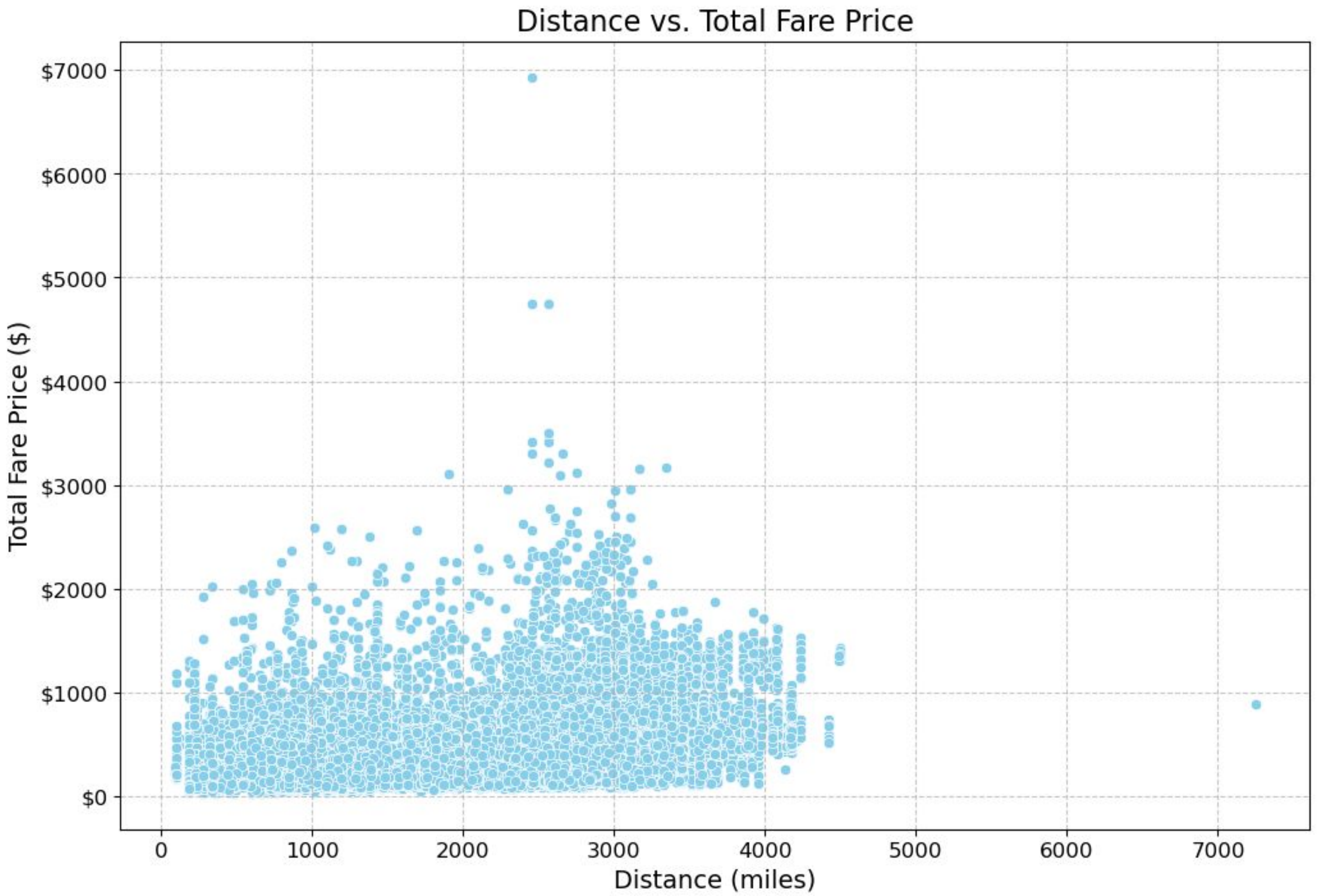


Average Fare by Month

EDA

Scatter Plot of Distance vs. Price

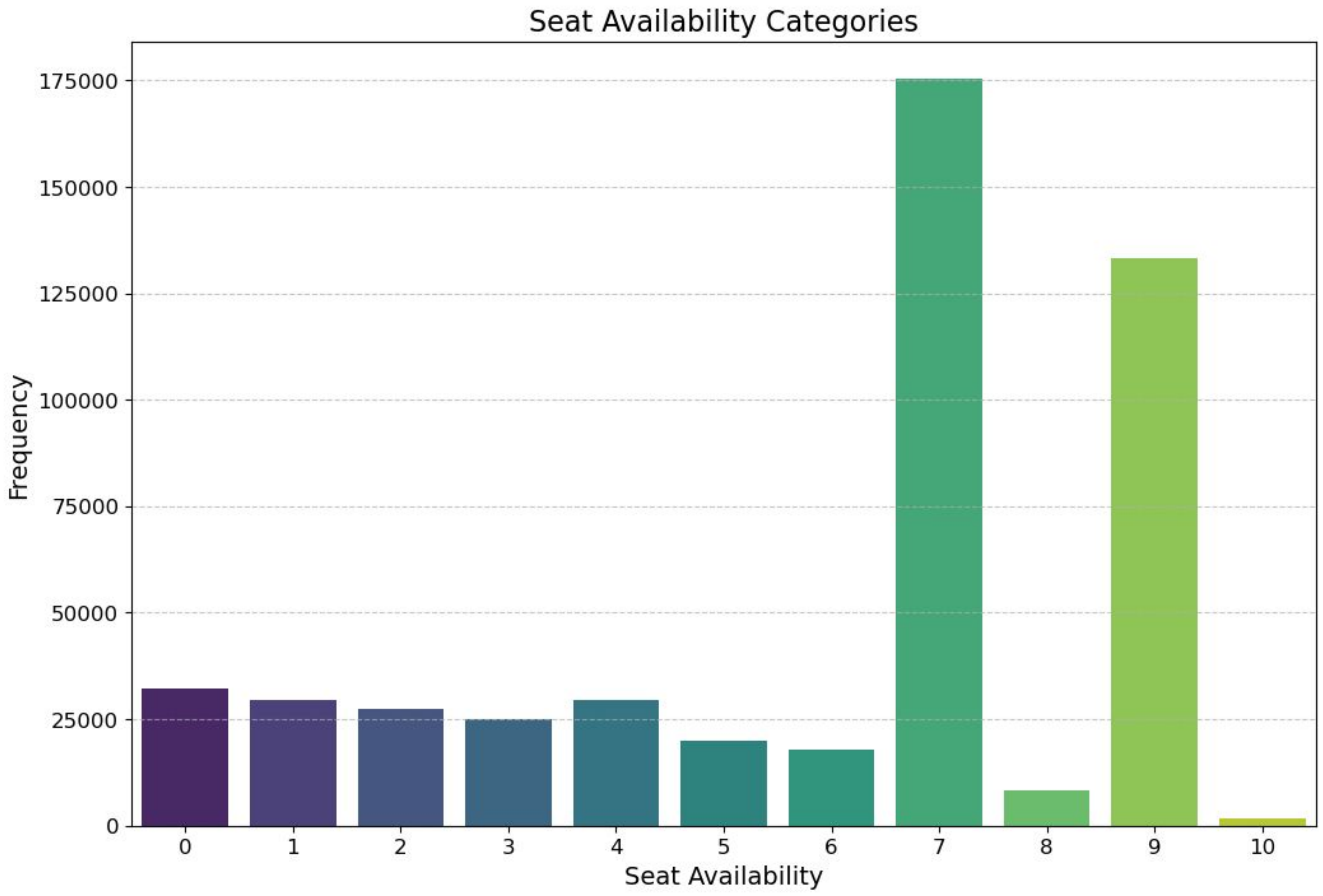
- Longer flights generally cost more, but with high variability
- Some short-distance flights have very high fares, likely due to premium flights or last-minute bookings
- Highlights the non-linear relationship between fare and distance, suggesting other factors influence pricing



EDA

Bar Chart of Seat Availability Categories

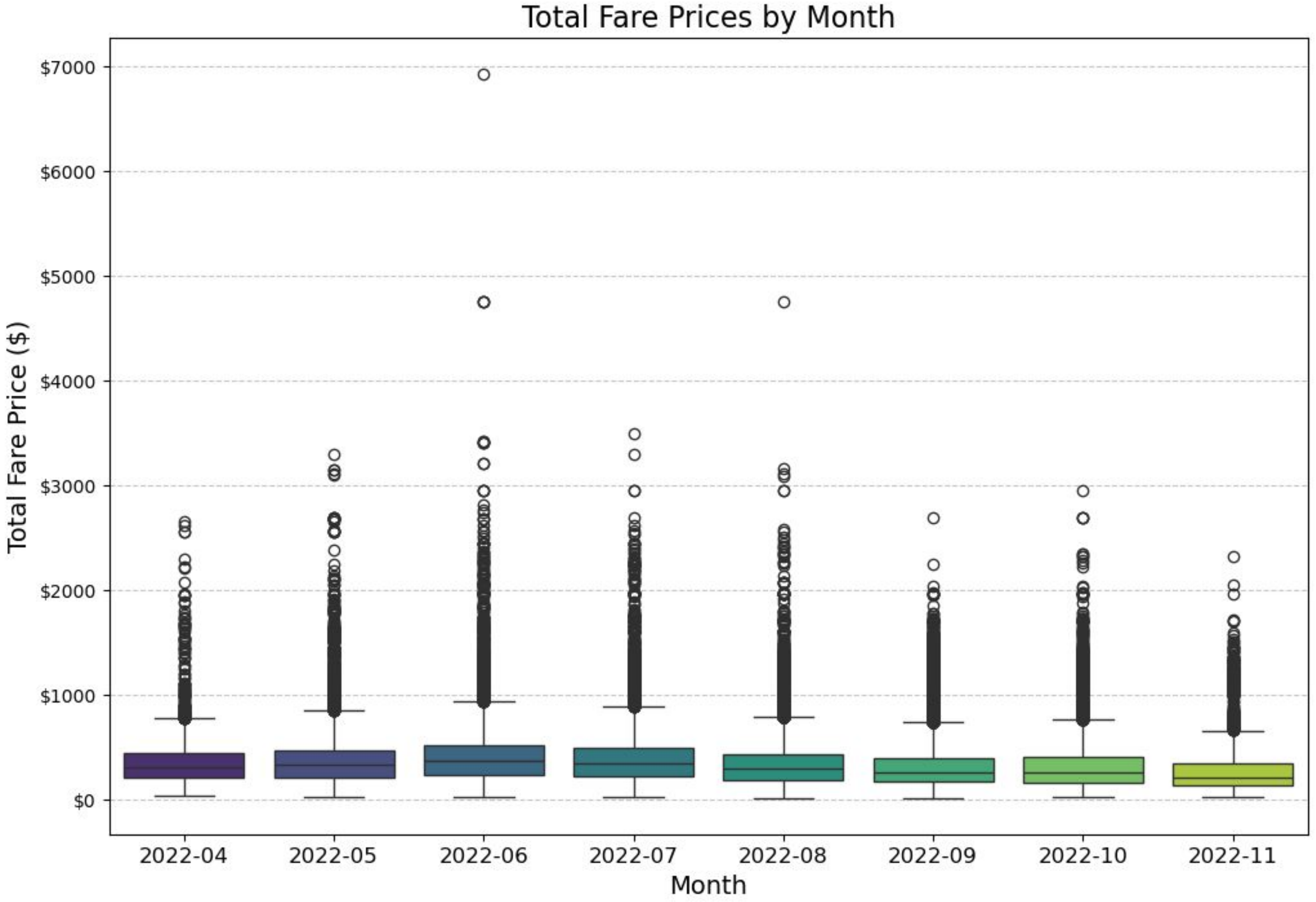
- The dataset shows that most flights have either low availability (0-5 seats) or high availability (7 and 9 seats)
- Understanding seat availability helps in predicting pricing trends, as fewer seats often lead to higher fares



Feature Engineering

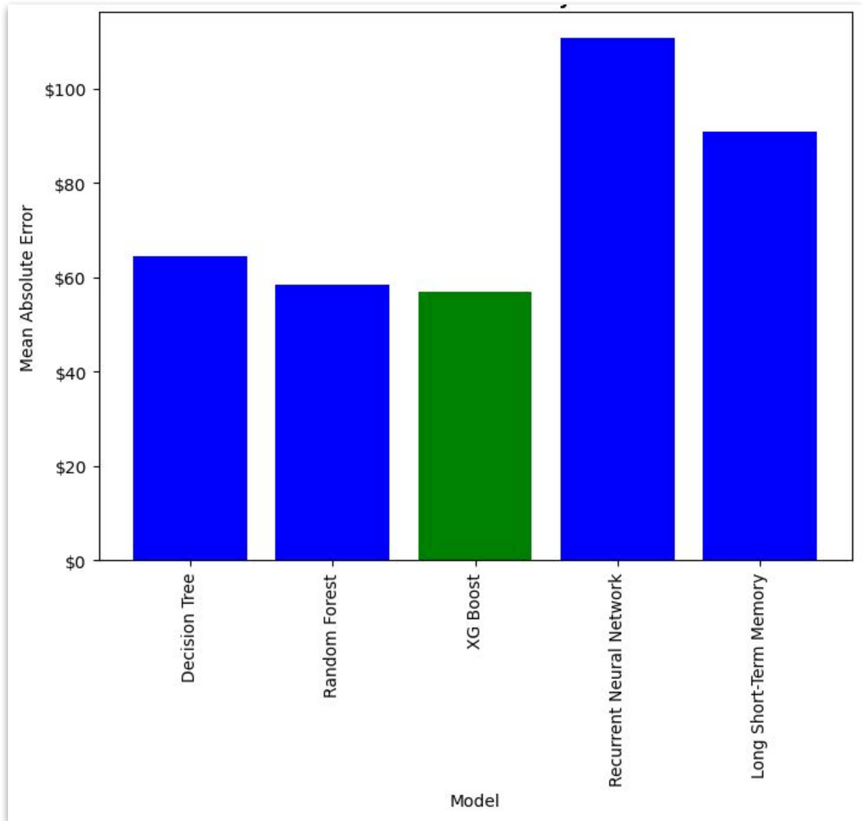
Boxplot of Prices by Month

- Prices show similar distributions across all months with significant outliers (more in summer months)
- Most likely associated with the high travel season during summer that raises average fares

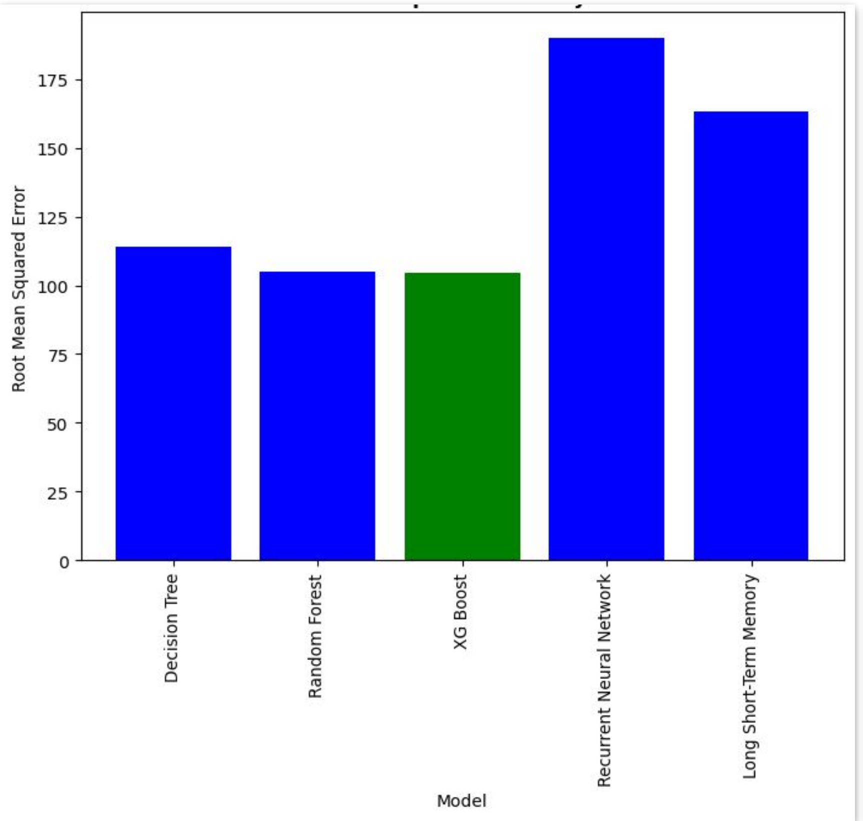


Models Tradeoff Discussion

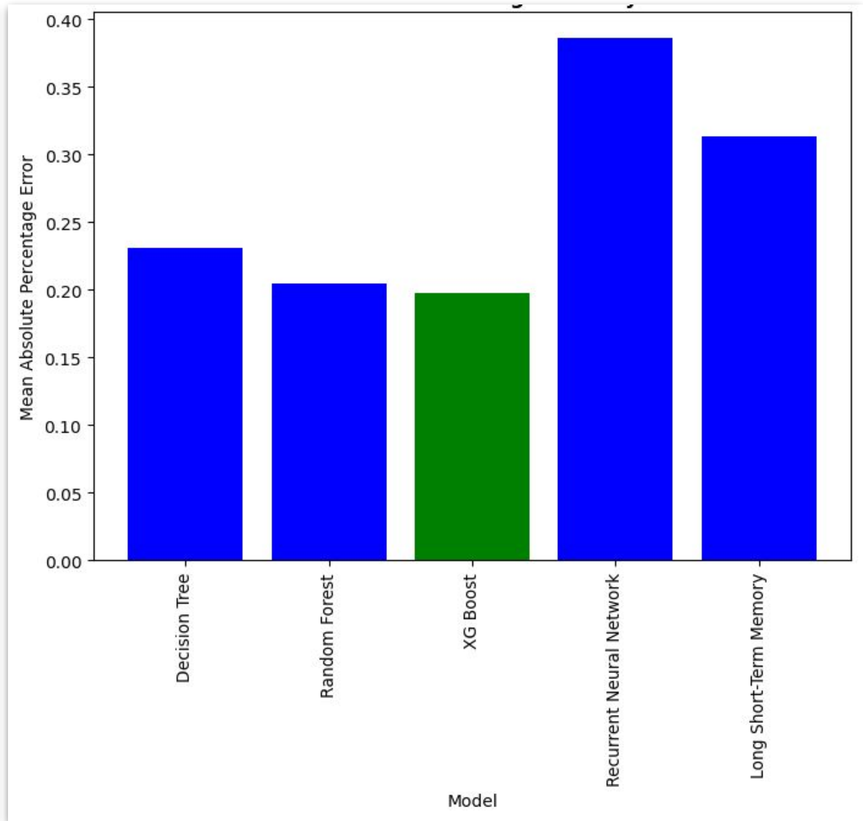
Mean Absolute Error



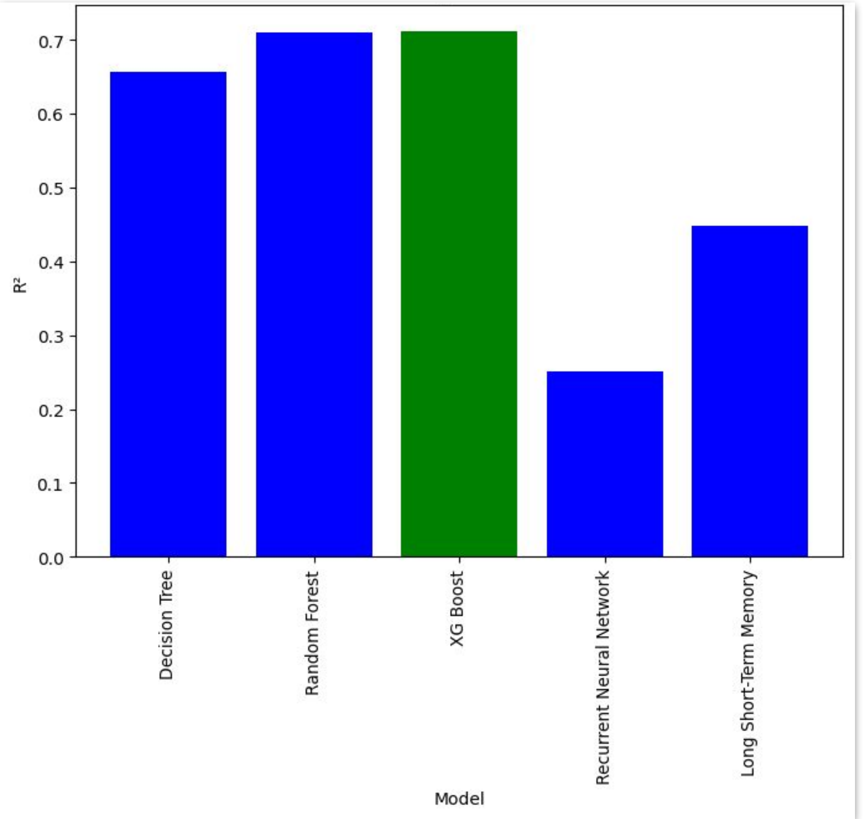
Root Mean Squared Error



Mean Absolute Percent Error



R^2

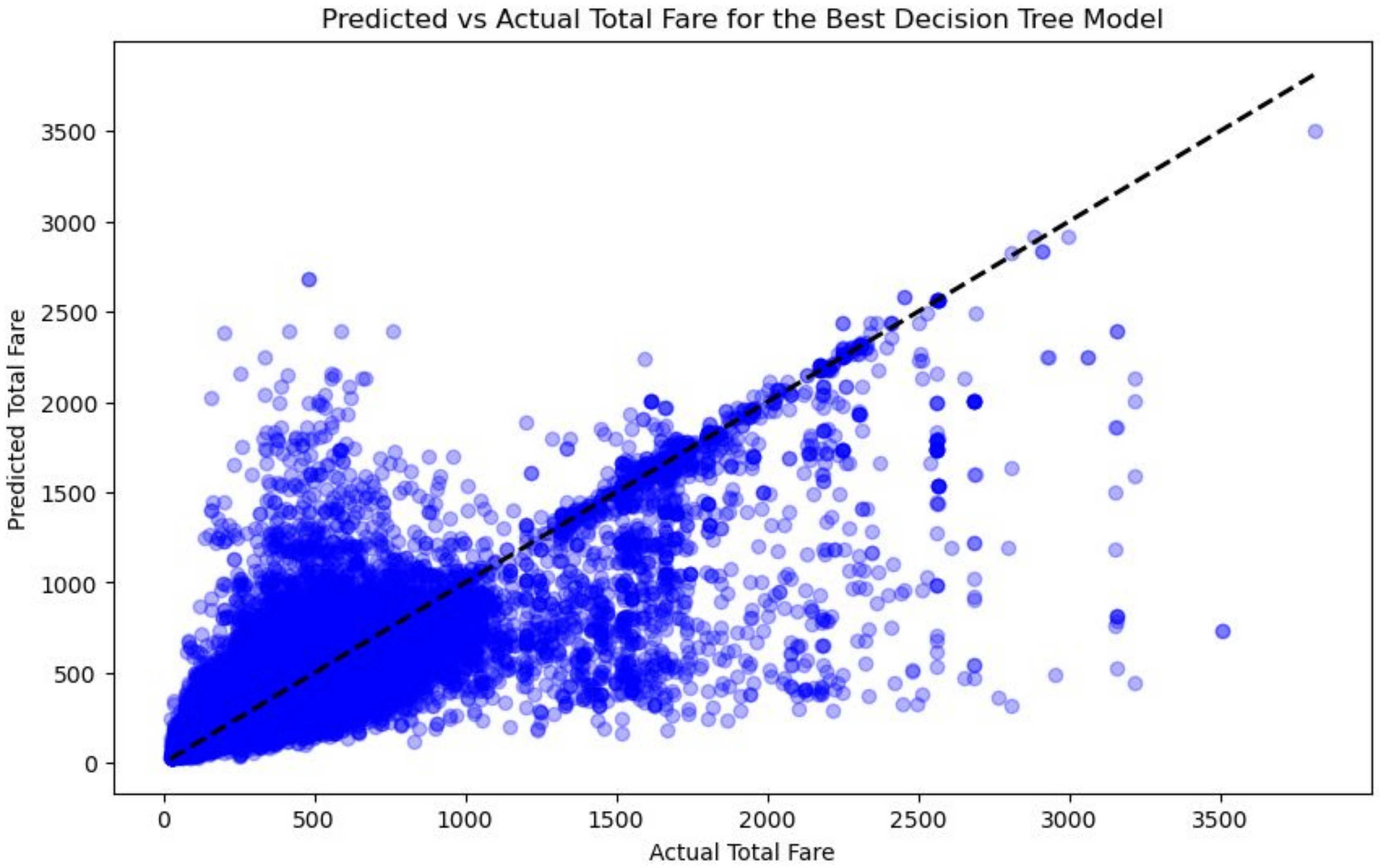


Key Insights

- Tree-based models (Random Forest, XGBoost) outperform deep learning models for flight price prediction

- Feature engineering significantly improved model accuracy, particularly holiday effects and seat availability binning

- Flight prices are highly non-stationary, requiring advanced modeling techniques



Future Research Directions

1. **Expanding the dataset**
   a. include international flights and budget airline fares.

2. **Exploring ensemble methods**
   a. The ones that combine Random Forest and XGBoost for improved predictions.

3. **Investigating reinforcement learning**
   a. In order to reach dynamic pricing recommendations.