Airline Price Predictor

Summary & Hypothesis:

For our project, we aim to use time series forecasting and machine learning to predict variations in airline ticket prices leading up to departure. The airline industry is a notorious case study of the price elasticity of demand, with nearly all modern airlines utilizing dynamic pricing strategies to adjust ticket prices based on factors such as time to departure, demand, and other market conditions. Our hypothesis is that by modeling how prices change based on time to departure, it is possible to predict when ticket prices are most likely to increase. Specifically, we aim to analyze historical flight price data to identify temporal patterns that will allow us to better understand the fluctuations in price leading up to a flight's departure. This would ultimately help consumers make more informed decisions, increasing confidence in their planning and minimizing unnecessary spending.

Technical Approach:

Our goal is to produce a model capable of predicting the flight price from day T to day $T+k$, where day $T+k$ is the flight departure date, using historical price data from day 0 to day T. To do so, we will preprocess the very likely non-stationary data (since we expect the data to trend upwards with changing means and variance toward the departure day). From there, we will test multiple models including SARIMA (Seasonal AutoRegressive Integrated Moving Average), LSTM (Long Short-Term Memory Networks) and GRUs (Gated Recurrent Units). After testing and evaluating these models, we will benchmark them as a candidate model and fine-tune.

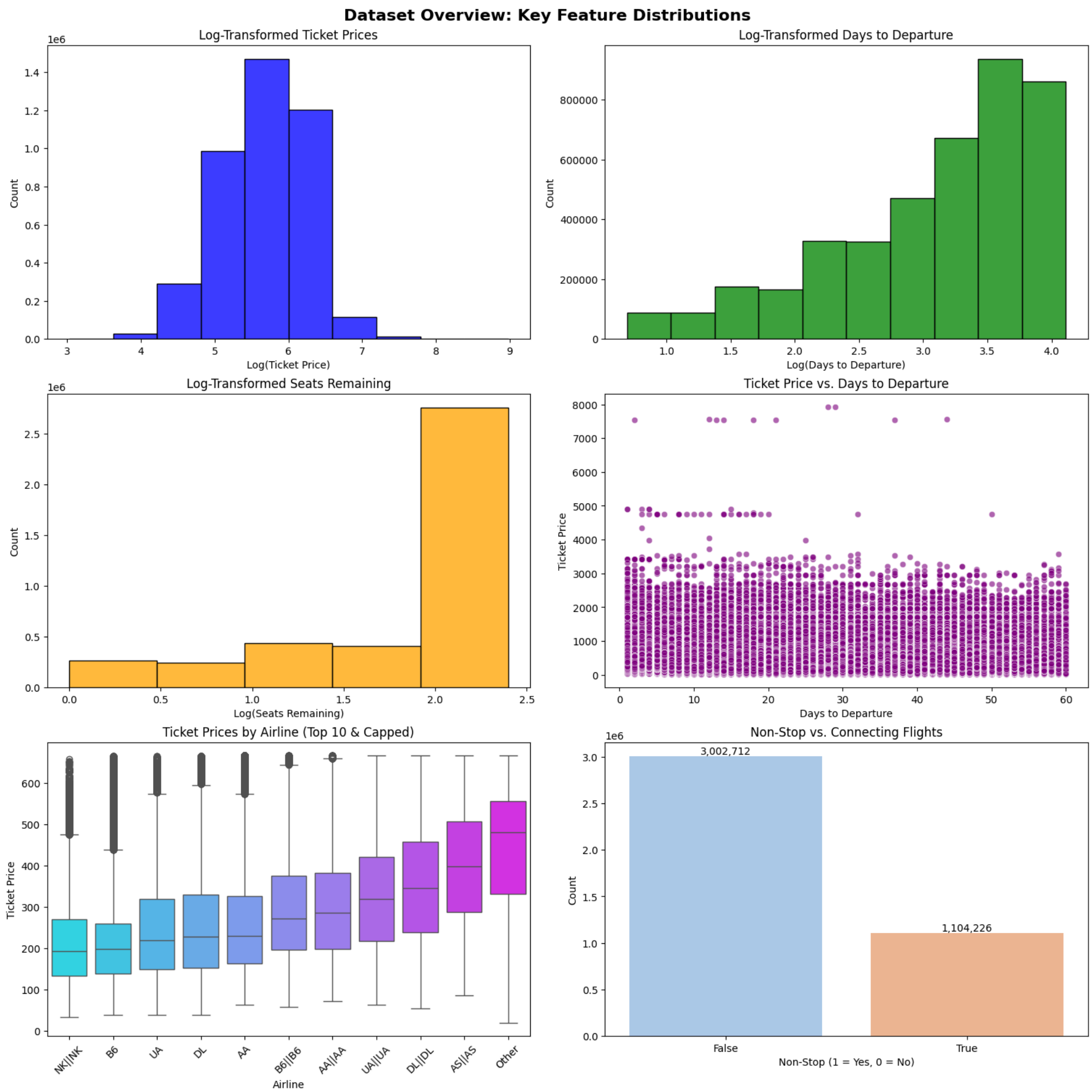
Data Source & Tools:

For our data, we plan to use the Flight Price dataset, which contains 82M+ rows of historical flight price data from one-way flights on Expedia in 2022. The dataset contains features such as ticket prices, routes, airlines, and, most importantly, the date of departure. A preliminary EDA on a stratified sample of 4M rows can be found in the appendix below.

Group Workplan:

Week 1	**Dataset preparation and EDA.** Each team member will do their own EDA and we will combine results.
Week 2–4	**Model development and tuning.** Each team member will develop their own forecast model (6 total models) for the rest of the team to review. Initial model ideas: Seasonal ARIMA, XGBoost, Prophet, LSTM, AR-Net, and State Space Models.
Week 5	**Model evaluation, benchmarking, and interpretation.** Each team member will do this portion for the model they developed.
Week 6	**Final report generation and presentation.** Entire team will review models, compare evaluation statistics, and choose the best forecast model. Work together to develop slides for the final presentation and determine which team members will present (~4 presenters).

Appendix:



Dataset Overview: Key Feature Distributions
Log-Transformed Ticket Prices
Log(Ticket Price)
Count
1e6
Log-Transformed Days to Departure
Log(Days to Departure)
Log-Transformed Seats Remaining
Log(Seats Remaining)
Ticket Price vs. Days to Departure
Days to Departure
Ticket Price
Ticket Prices by Airline (Top 10 & Capped)
NK||NK
B6
UA
DL
AA
B6||B6
AA||AA
UA||UA
DL||DL
AS||AS
Other
Airline
Non-Stop vs. Connecting Flights
3,002,712
1,104,226
False
True
Non-Stop (1 = Yes, 0 = No)