Flight Price Prediction

Big Project Progress Presentation

Will & Kevin

Introduction & Motivation

Motivation:	Flight prices seem to change rapidly and for no reason • Even refreshing a page can show a change in the price of a flight Predicting flight prices could help consumers save hundreds per flight
Research Question:	What factors influence the price of a flight, and can we predict the price of a flight with machine learning?

Data Source

- Amadeus
  - Provides tech solutions for the travel industry
  - API access for free
- API Details:
  - Provides prices & features for bookable flights given airport codes for route and a date
  - Limited to 2,000 calls per month
  - 1 API call = all flights for 1 day between two specified airports

Data Collection

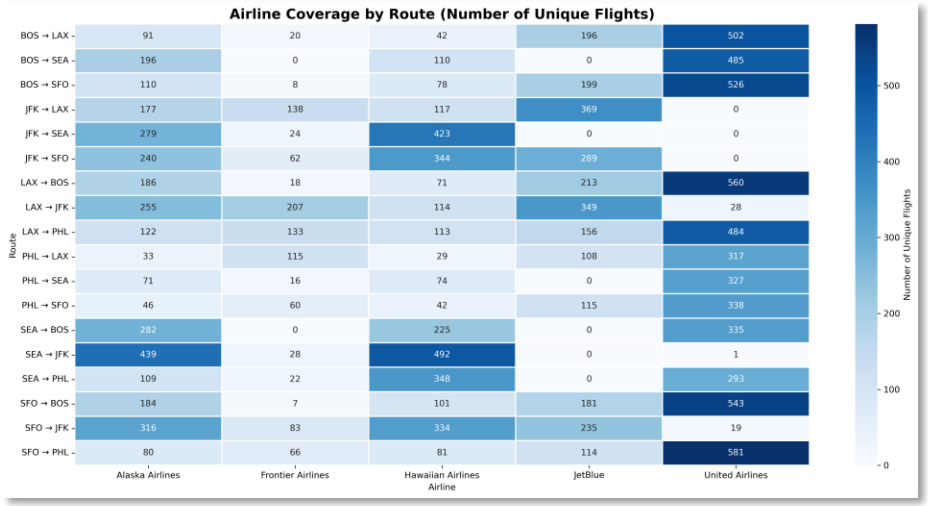
- Scope:
  - Limited to cross-country flights
  - Limited to 6 major airports (3 east coast, 3 west coast)
  - Limited to 14 days in advance
- API Scheduler
  - Runs .py script with API calls 2x per day (12:00 am & pm UTC)
  - 252 API calls per run (18 routes x 14 days)
  - Uses multiple API keys to avoid call limit
  - ~12,000 flight observations per run
  - Each run returns a .csv file and uploads to Dropbox
  - Scheduling done via GitHub workflows (link)

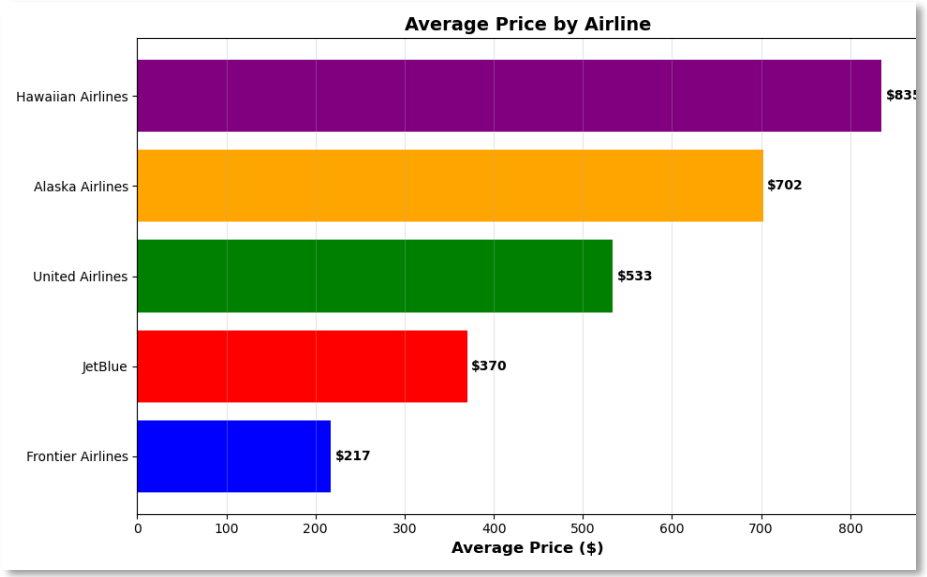
Final Dataset

```
Data Collection Metrics:
  • Total observations: 229,924
  • Unique flights tracked: 15,224
  • Flights with price changes: 13,500
  • Average price change: $154.67
  • Collection period: 2025-11-10 to 2025-11-19
```

- Currently 229,924 observations & growing
- Screenshot below shows the features given by the API

origin	destination	departure_date	days_until_departure	price	currency	airline	number_of_stops	departure_time	arrival_time	total_duration	aircraft_type	cabin_class	bookable_seats
JFK	LAX	2025-11-11	1	144.27	EUR	F9	1	2025-11-11 06:25:00	2025-11-11T12:25:00	PT9H	32Q,32Q	ECONOMY	4
JFK	LAX	2025-11-11	1	149.13	EUR	F9	1	2025-11-11 07:59:00	2025-11-11T19:53:00	PT14H54M	32Q,32Q	ECONOMY	4
JFK	LAX	2025-11-11	1	149.13	EUR	F9	1	2025-11-11 07:59:00	2025-11-11T20:42:00	PT15H43M	32Q,32N	ECONOMY	4
JFK	LAX	2025-11-11	1	210.02	EUR	F9	0	2025-11-11 11:29:00	2025-11-11T14:32:00	PT6H3M	32Q	ECONOMY	4
JFK	LAX	2025-11-11	1	256.07	EUR	AS	1	2025-11-11 07:00:00	2025-11-11T15:17:00	PT11H17M	73H,73J	ECONOMY	7



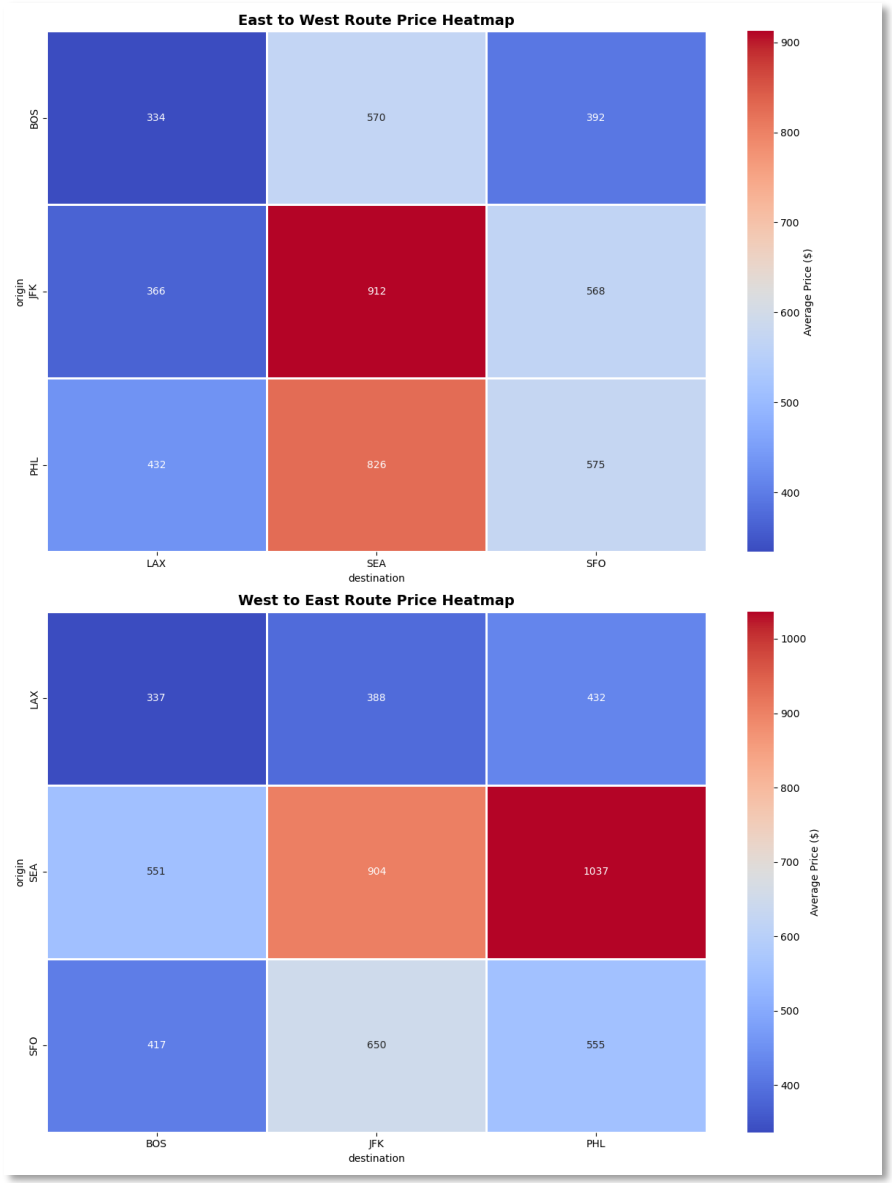


EDA: Airline-Level Analysis

- United: highest route coverage; moderate prices
- Frontier: few flights, very low prices
- Alaska & Hawaiian: specialize in low-competition routes -> charge premium for tickets

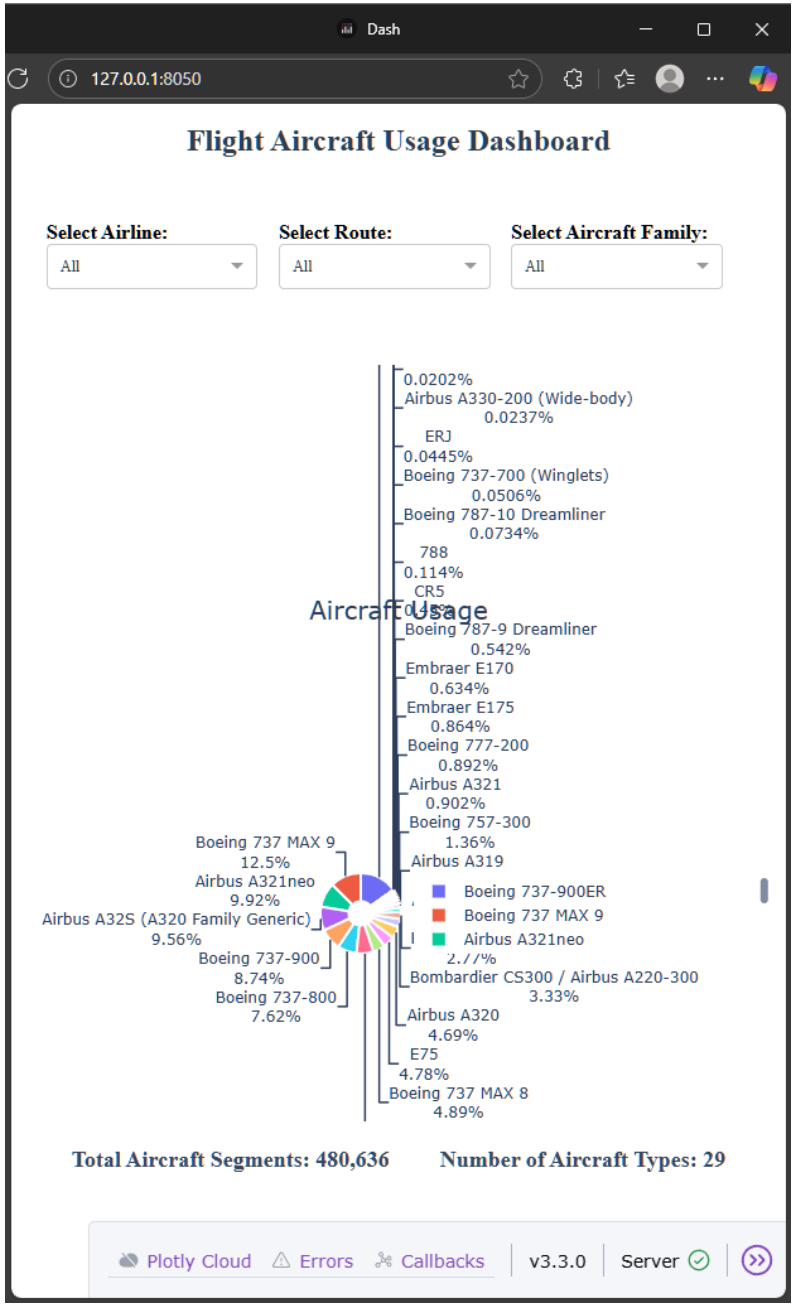
EDA: Route-Level Analysis

- Most Routes: East -> West is cheaper than West -> East
- Flights to & from Seattle are most expensive
  - Dominated by Hawaiian Airlines & Alaska Airlines (less low-cost offerings)
- Cheaper routes have more competition between major low- and mid-cost airlines

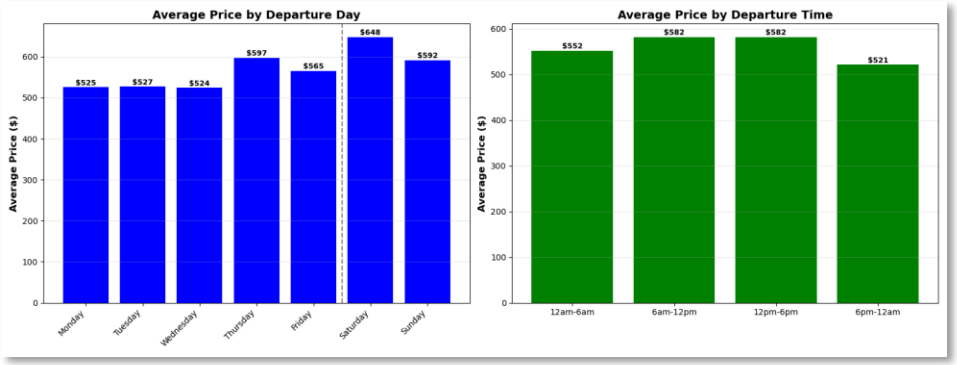


EDA: Aircraft Usage Analysis

- Most popular type: Boeing 737-900ER
- Most popular family: 737 NG & MAX, then A320ceo & A320neo
  - Full-cost airlines (United, Hawaiian, Alaska) go for 737
    - Still have A320 & A321s
  - Low-cost (Frontier, JetBlue) go for Airbus (A320/A321) only



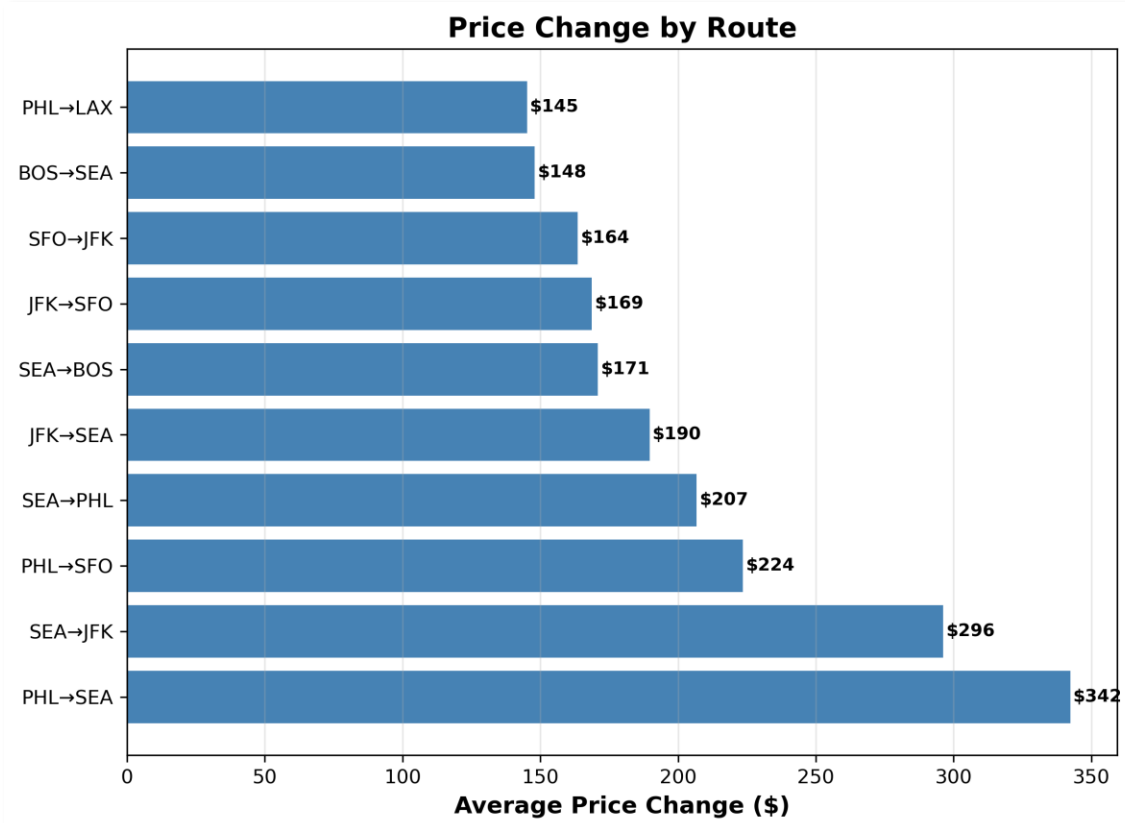
EDA – Temporal Analysis

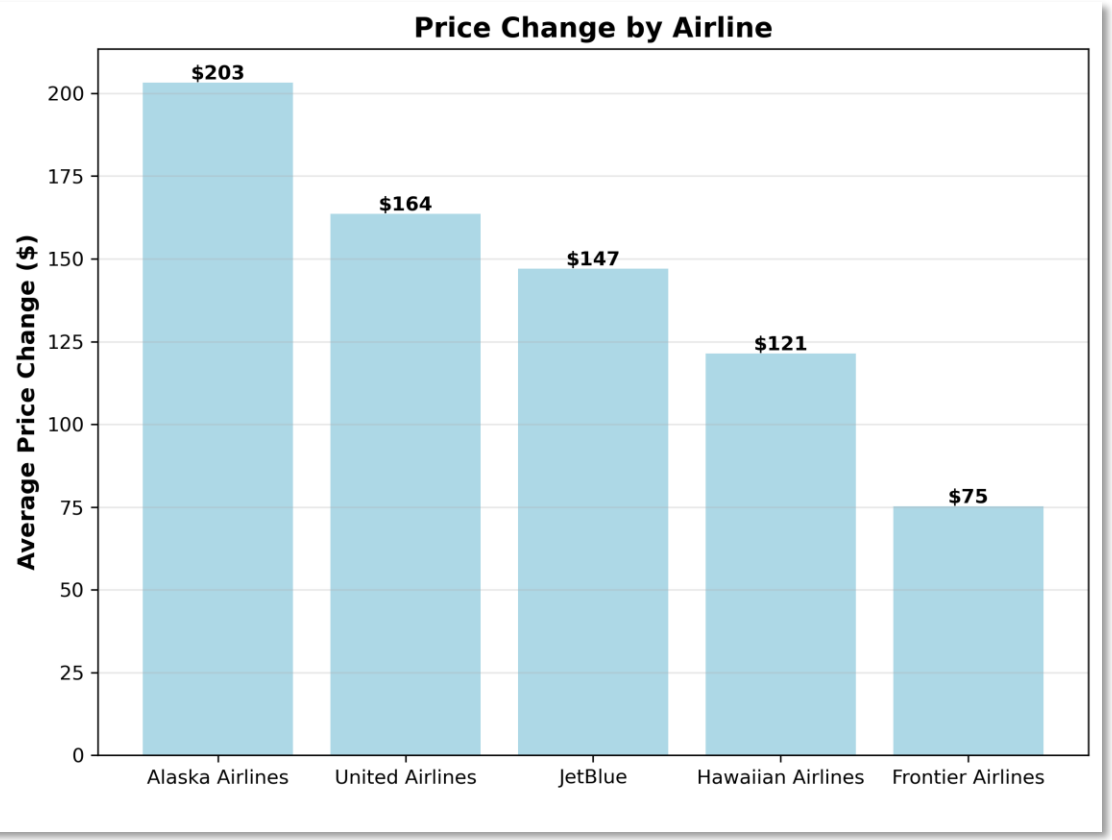


- Prices peak closest to departure
- Significant drop in price 7 days out
- Likely Reason:
  - Airlines price higher for last-minute bookers AND for bookers who plan far in advance
- Cheapest flights depart Monday through Wednesday
- Cheapest time to fly is at night (red-eye flights priced cheaply)

EDA – Price Change Analysis

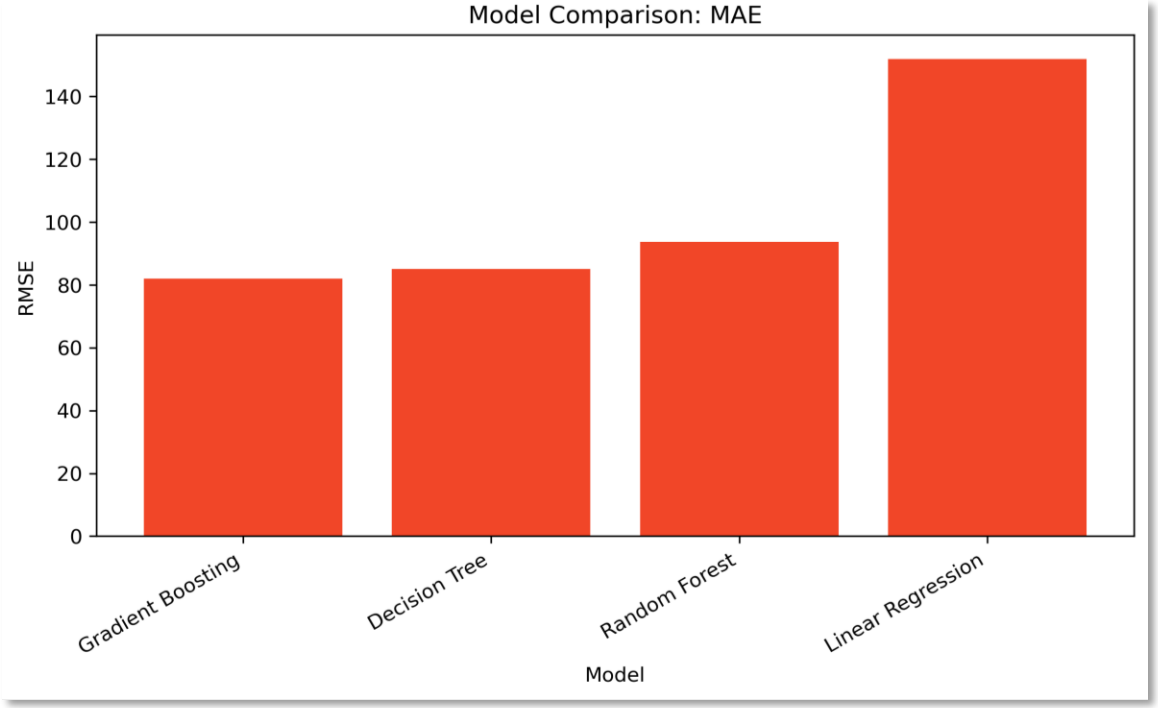
- Because of the scheduled API calls, we can track the same flight's price over time
- Below shows the average change in price for a specific flight (max listed price – min listed price)

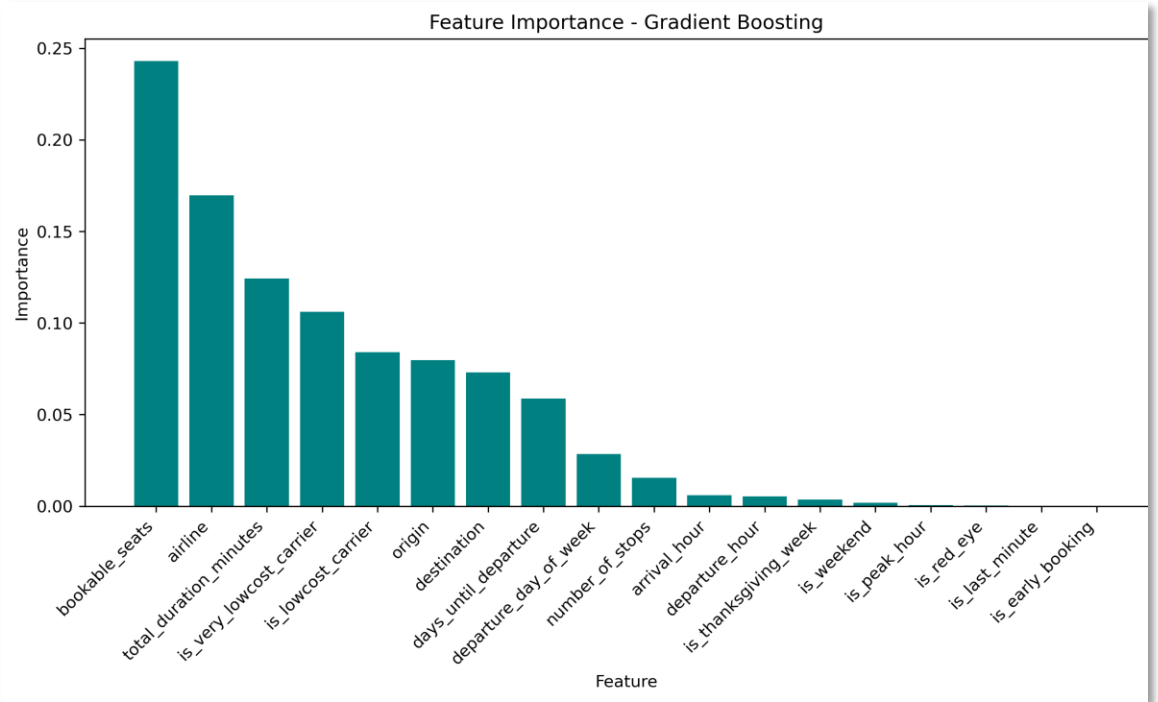




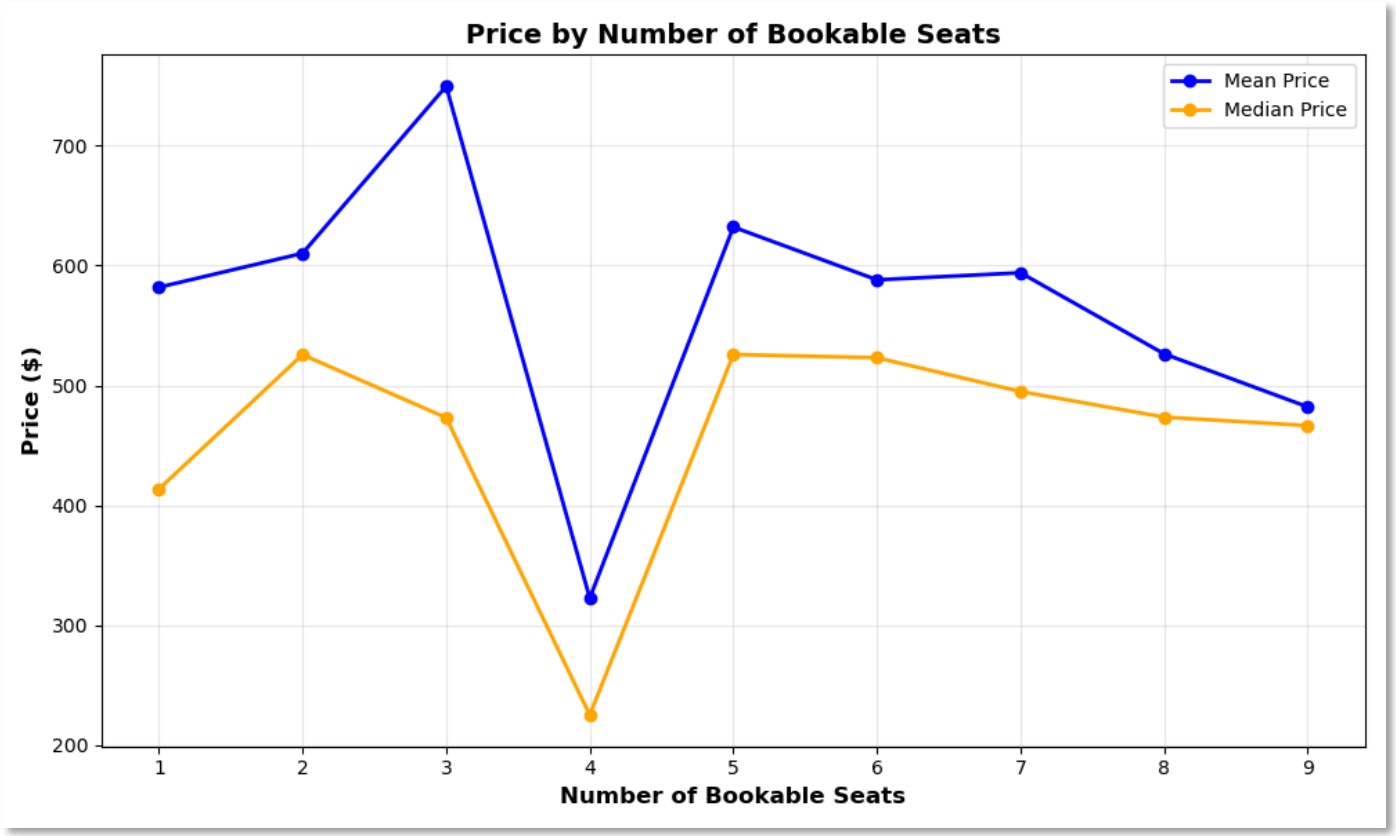
Preliminary Machine Learning Results

- Tested 4 initial models:
  - Linear Regression
  - 3 Tree Models (Decision Tree, Random Forest, & Gradient Boosting)
- Tree models outperformed linear regression
- Lowest MAE (Gradient Boosting) ~ 82 (~16% mean error assuming $500 average price)
- Number of bookable seats, airline, and flight duration most important features
- Temporal features have less predictive power





Note On Bookable Seats



- Non-linear relationship
- Drop in mean & median price at 4 bookable seats
  - Data issue or intentional pricing strategy?

Future Work

- Enhanced Machine Learning Models
  - Neural Networks, XGBoost, etc.
  - Adding further hyperparameter tuning
  - Adding more engineered features
- Added Data Sources
  - Features about airlines
  - Features about weather/climate

Implications & Limitations

- Limitations:
  - Data availability (6 airports, select airlines)
  - Limited date range
  - Limited features from API (had to create *flight_id* separately, etc)
- Ethical, Economic, & Social Implications:
  - Predicting prices could help consumers save money and protect consumers against confusing or exploitative pricing strategies
  - Pricing strategies are kept secret by airlines, secrets could be discovered