Flight Prices

Asma, Hissah, Joud, Najd, Leena, Lujain

Introduction

Our project analyzes a dataset containing key attributes such as airline, departure time, duration, stops, and ticket price to uncover patterns that affect airfare. By identifying these trends and building a predictive model, this work aims to help travelers make smarter booking decisions and support airlines in optimizing pricing strategies.

Data Collection

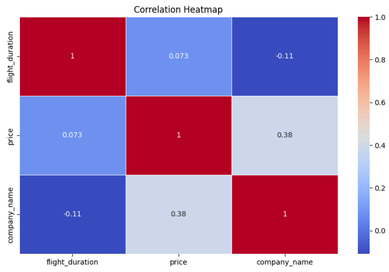
Our dataset was collected on September 24, 2024 through web scraping from Google Flights. It includes 505,504 rows and 11 columns covering flights from September 24, 2024, to September 24, 2025.

Objectives

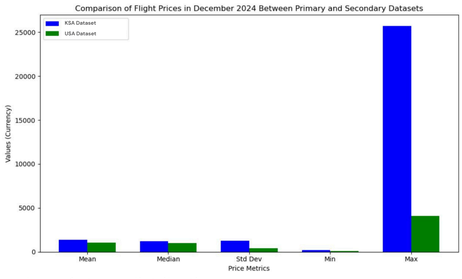
- How does the departure date affect flight prices?
- What is the price difference between flights within Saudi Arabia and those within the USA?
- How does the number of stops in a flight route impact the overall ticket price?
- Does the flight duration affect the ticket price?
- Does the timing of booking influence the ticket price? For instance, is there a price difference when booking a month in advance compared to a week before the travel date?

Data Analysis

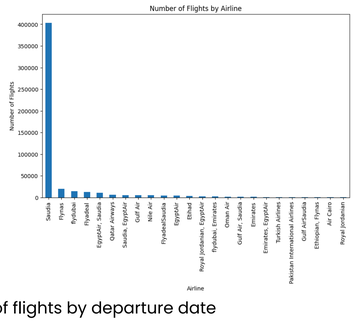
Correlation between price and flight duration and Airline



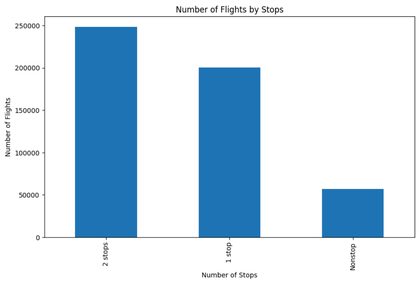
Comparison of Flight Prices in December 2024 Between KSA and USA Datasets



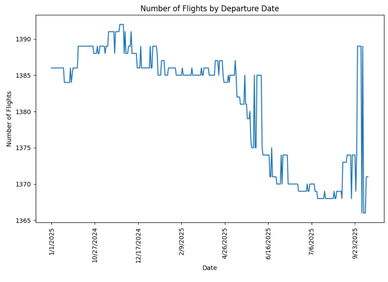
Number of flights by airline



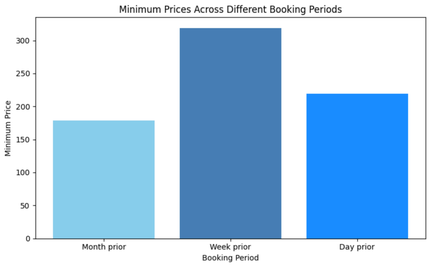
Number of stops distribution



Number of flights by departure date



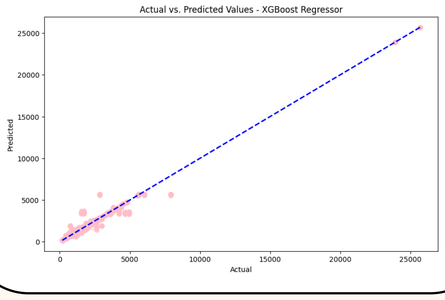
Minimum Prices Across Different Booking Periods



Models And Findings

We conducted a regression analysis using 10 models, with the top performers being XGBoost (R^2 = 0.9764), KNeighbors Regressor (R^2 = 0.9686), and Random Forest (R^2 = 0.9653).

For clustering, K-means with k = 9 achieved the best performance with a silhouette score of 0.215. In comparison, GMM scored 0.173, and DBSCAN performed poorly with a silhouette score of -0.357.



Conclusion

The analysis revealed that flight ticket prices are primarily influenced by factors such as flight duration, number of stops, airline choice, and departure timing, with longer durations, multiple stops, and premium airlines leading to higher prices. XGBoost was identified as the best regression model for predicting prices and KMeans with k=9 as the most effective clustering approach. These insights can help airlines refine pricing strategies and travelers plan cost-effective bookings. Future work could include using real-time data and external factors like weather or economic conditions to improve predictions further.