Expedia Flight Itinerary Analysis

By: Anusha Bhat, Mahima Masetty, Sophie Liu

December 3, 2024

Meet Our Team

Anusha Bhat
Data & Graph Engineer

Mahima Masetty
Data Scientist

Sophie Liu
Machine Learning Engineer

Agenda

- Data Summary 04
- Data Preprocessing 05
- Graph Computing 11
- Predicting Flight Price 18
- Conclusion 34

Understanding the Data: Size, Features, and Missing Values

Data Source: Kaggle
Data Size: 29 GB

Each row is a one-way purchasable itinerary found on Expedia

To/from the following 16 airports:

ATL, DFW, DEN, ORD, LAX, CLT, MIA, JFK, EWR, SFO, DTW, BOS, PHL, LGA, IAD, OAK.

Time Duration:

04/16/2022 - 10/05/2022
~ 6 months

82.1 M rows, 27 Columns

“totalTravelDistance” is missing/null for 7% of the entries

Original Features

- legID
- search date
- flight date
- starting airport
- destination airport
- fare basis code
- travel duration
- number of elapsed travel days
- is basic economy
- is refunable
- is non stop
- base fare price
- total fare price
- seats remaining
- total travel distance
- segments departure time raw & epoch
- segments arrival time raw & epoch
- segments arrival airport codes
- segments departure airport codes
- segments airline names & codes
- segments equipment descriptions
- segments duration in seconds
- segments distance
- segments cabin code

DATA PREPROCESSING

Cleaning and Preparing the Data for Modeling

01 Repartitioned data and dropped columns with non-relevant and/or duplicated information.

02 Feature Engineering

03 Data Transformation

04 Split data into train (75%), validation (15%), and testing (10%) and stored the sets using parquet compression.

Enhancing ‘Segment’ Features: String Processing and Transformation into New Columns

Due to layovers, many columns have entries split into segments:
A || B || C

2022-04-17T07:18:00.000-04:00||2022-04-17T10:16:00.000-04:00	IAD||BOS	ATL||IAD	UA||UA
2022-04-17T15:54:00.000-05:00||2022-04-17T20:24:00.000-04:00	ORD||BOS	ATL||ORD	AA||AA

- Created new columns for each feature containing arrays split on “||”
- Used information in arrays to create the following new features:
  - Layover airports
  - Number of unique cabins, aircrafts, and airlines
  - Number of stops
  - Has first class

segmentsCabinCode	CabinCodes	UniqueCabins	NumUniqueCabins	hasFirstClass
coach||coach||coach	[coach, coach, coach]	[coach]	1	0
coach||coach	[coach, coach]	[coach]	1	0
coach	[coach]	[coach]	1	0
coach||coach	[coach, coach]	[coach]	1	0
coach||coach||coach	[coach, coach, coach]	[coach]	1	0
coach	[coach]	[coach]	1	0
coach||coach	[coach, coach]	[coach]	1	0
coach||coach	[coach, coach]	[coach]	1	0
coach||coach	[coach, coach]	[coach]	1	0
coach||coach	[coach, coach]	[coach]	1	0

Converting Data Types and Transforming Features for Model Readiness

Converting Data Types

- Travel duration from hours to minutes
- Search and flight dates from strings to dates

One-hot Encoding

One-hot encoding of boolean and categorical columns:

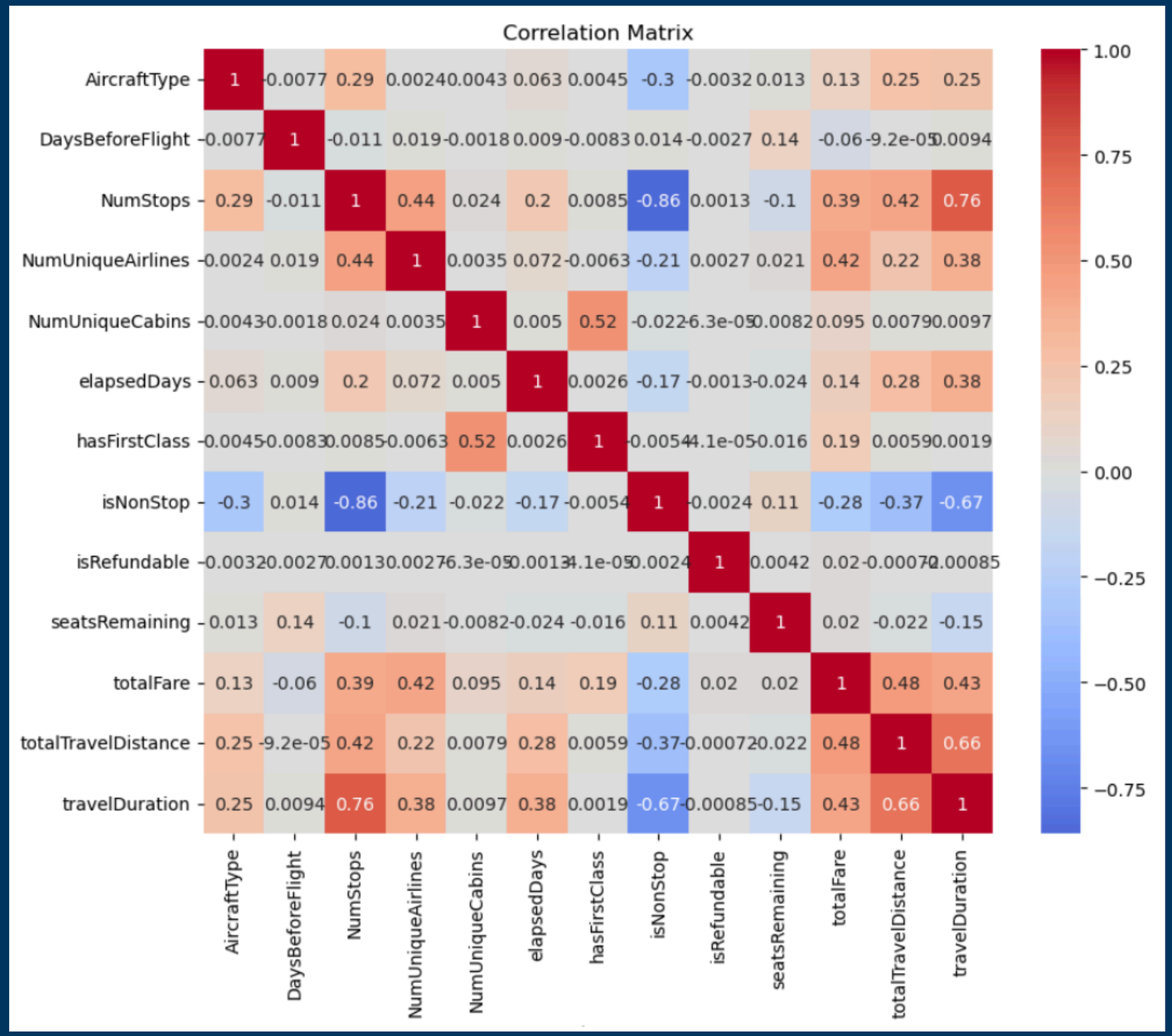
- First class
- Aircraft type
- Is refundable
- Is nonstop

Imputations

Replacing missing values for travel distance using the distribution's median value

Examing Correlations Between Numerical Features

- High correlation with NumStops:
  - IsNonStop, travelDuration
- Moderate correlation with travelDuration:
  - isNonStop, totalTravelDistance.
- Moderate correlation with totalFare:
  - travelDuration, totalTravelDistance, numUniqueAirlines, and numStops.
- Dropped isNonStop since it captures less information than NumStops
- Kept totalTravelDistance and travelDuration for use in graph model



Updated Schema After Data Pre-Processing

Original Schema after dropping initial columns

```
- legId: string (nullable = true)
- searchDate: string (nullable = true)
- flightDate: string (nullable = true)
- startingAirport: string (nullable = true)
- destinationAirport: string (nullable = true)
- travelDuration: string (nullable = true)
- elapsedDays: integer (nullable = true)
- isRefundable: boolean (nullable = true)
- isNonStop: boolean (nullable = true)
- totalFare: double (nullable = true)
- seatsRemaining: integer (nullable = true)
- totalTravelDistance: integer (nullable = true)
- segmentsArrivalTimeRaw: string (nullable = true)
- segmentsDepartureAirportCode: string (nullable = true)
- segmentsAirlineName: string (nullable = true)
- segmentsEquipmentDescription: string (nullable = true)
- segmentsCabinCode: string (nullable = true)
```

New Schema

```
|-- legId: string (nullable = true)
|-- searchDate: date (nullable = true)
|-- flightDate: date (nullable = true)
|-- startingAirport: string (nullable = true)
|-- destinationAirport: string (nullable = true)
|-- travelDuration: integer (nullable = true)
|-- elapsedDays: integer (nullable = true)
|-- isRefundable: integer (nullable = true)
|-- isNonStop: integer (nullable = true)
|-- totalFare: double (nullable = true)
|-- seatsRemaining: integer (nullable = true)
 -- DaysBeforeFlight: integer (nullable = true)
 -- Layovers: array (nullable = true)
     |-- element: string (containsNull = true)
 -- NumStops: integer (nullable = true)
 -- AirlineNames: array (nullable = true)
     |-- element: string (containsNull = true)
 -- NumUniqueAirlines: integer (nullable = true)
 -- AircraftType: integer (nullable = true)
 -- NumUniqueCabins: integer (nullable = true)
 -- hasFirstClass: integer (nullable = true)
 -- FlightArrivalDate: date (nullable = true)
|-- totalTravelDistance: integer (nullable = true)
```

Flight Route Analysis for Expedia

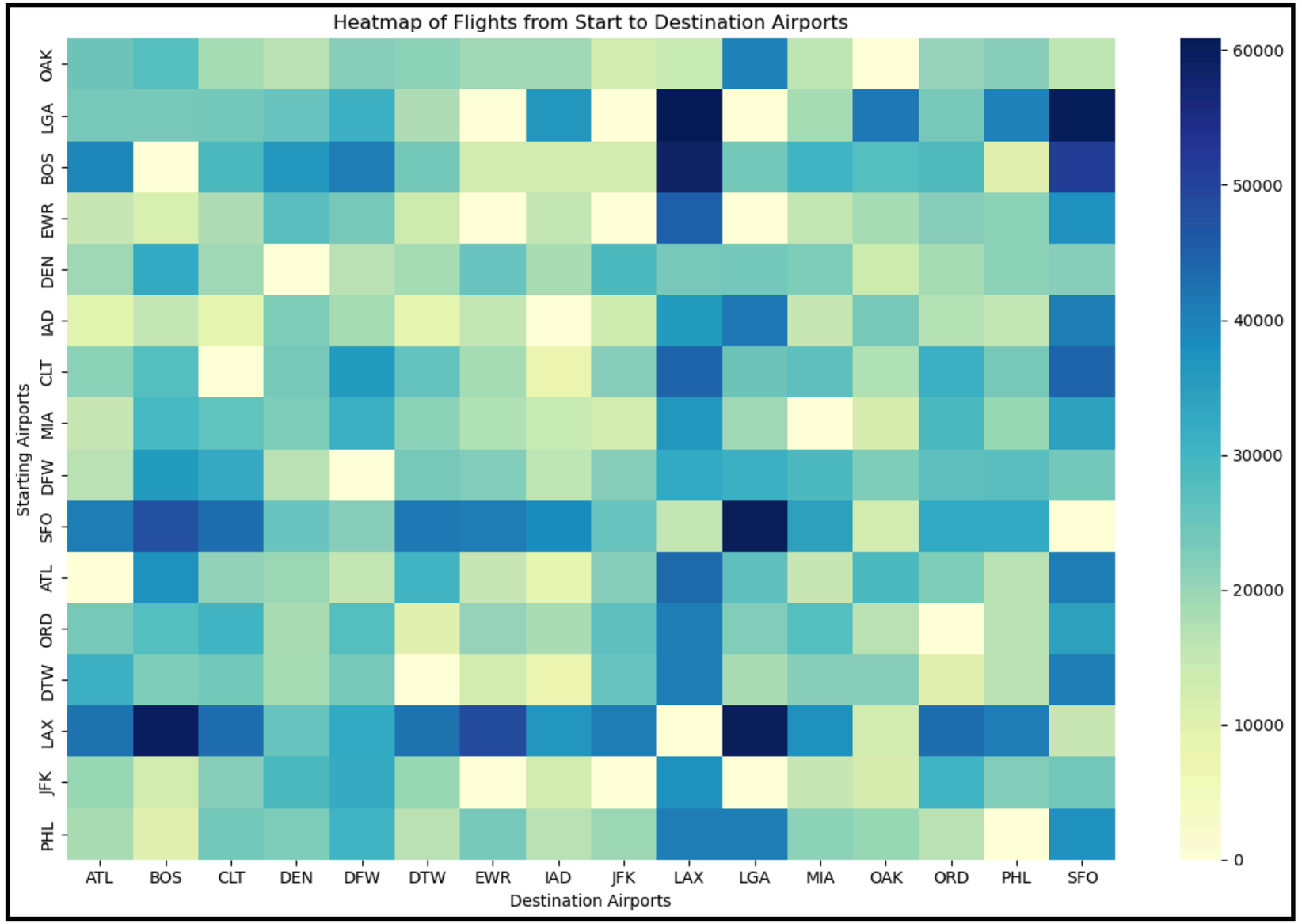
Constructing a graph based model to analyze flight paths between various airports.

Offer the best flight routes

Demand Forecasting for Different Routes

Identify Strategic Hubs for Operational Efficiency

Airport Combinations with the Highest and Lowest Flight Traffic



- Top Combinations:
  - LGA and LAX
  - LGA and SFO
  - BOS and LAX
  - BOS and SFO
- IAD and DEN have the lowest counts for combinations with all airports

Constructing a Graph-Based Model for Flight Route Analysis

- Airports represented as the vertices - 16 nodes
- Nonstop flight routes represented as edges - 644,592 edges
- Departure Airports are the source nodes
- Arrival Airports are the destination nodes
- Graph has one connected component
- All airports have triangles
  - ATL, DTW, DEN, ORD, and LAX have the most triangles (92 each)

Airport Name	Degree
BOS	120403
ORD	117895
LAX	115600
LGA	104591
ATL	104264

Airport Name	Page Rank
LAX	1.419
BOS	1.405
ORD	1.377
ATL	1.254
LGA	1.236

Utilizing the Graph Model to Answer Customer-Specific Questions

Using queries, motif finding, and shortest paths, we can answer the following customer use cases:

- Customer 1: I'm interested in taking a cheap getaway trip. What are the cheapest cities I can fly to from ORD?
- Customer 2: I'm interested in flying to the east coast. What is the shortest route for me to get to an east coast city from ORD?
- Customer 3: I'm interested in a multicity trip flying in and out of ORD. What cities can I visit?

Customer 1

Destination Airport	Avg. Fare ($)
OAK	199.21
LGA	211.42
JFK	216.39
DTW	220.70
ATL	229.59

Customer 2

Destination Airport	Stops	Distance (mi)
CLT	0	592
IAD	0	594
ATL	0	600
PHL	0	672
JFK	0	720

Customer 3

A few itinerary suggestions:

ORD –> LAX –> OAK –> DTW –> ORD

ORD –> BOS –> DEN –> OAK –> ORD

ORD –> IAD –> OAK –> DFW –> ORD

Fare Prediction for Expedia

Building linear and non-linear models to predict flight ticket fares and comparing model performances.

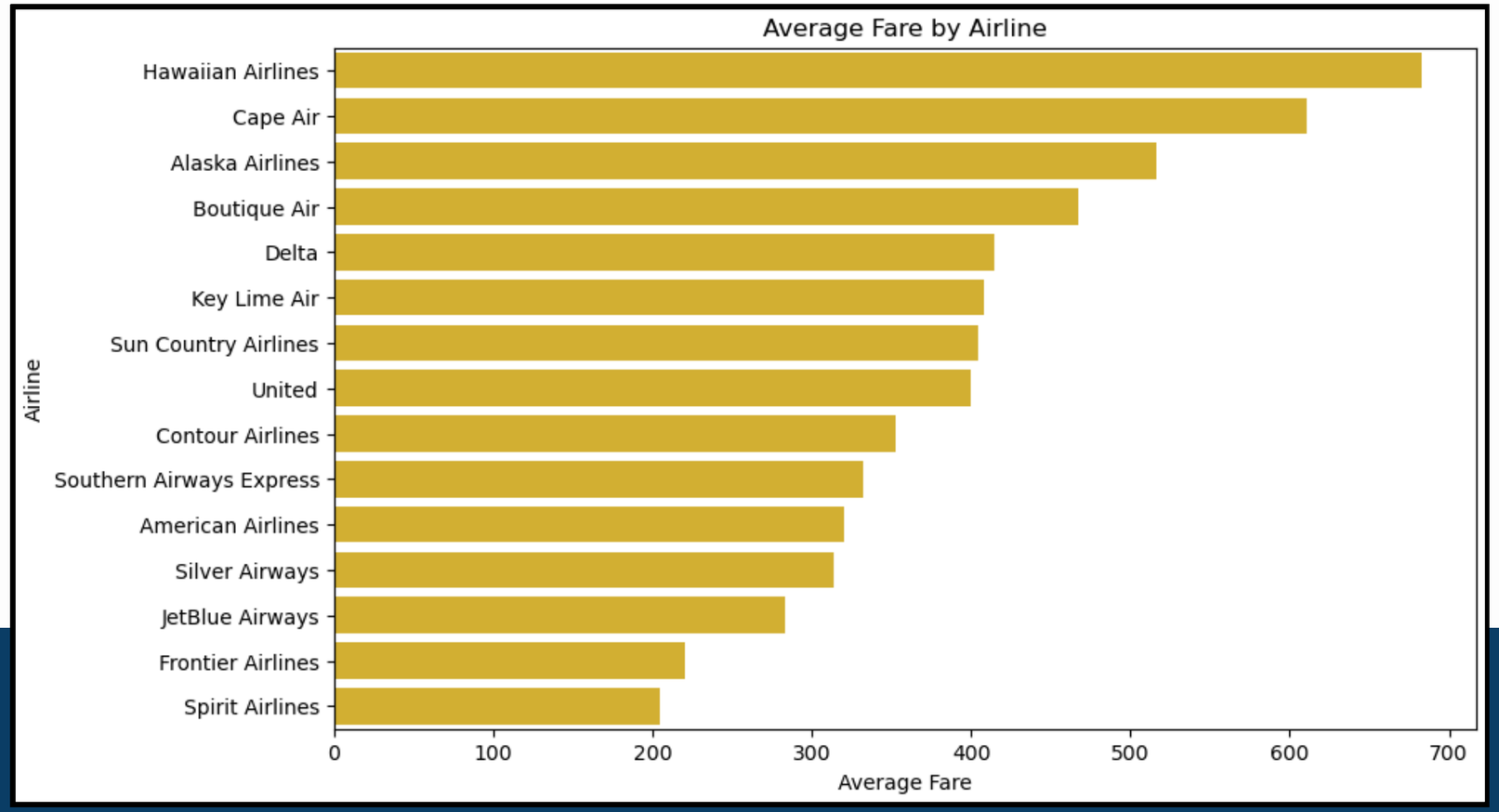
Improved Customer Experience

Optimized Revenue Management

Increased Conversion Rates

Reduced Price Volatility

Average Fare Differs By Airline

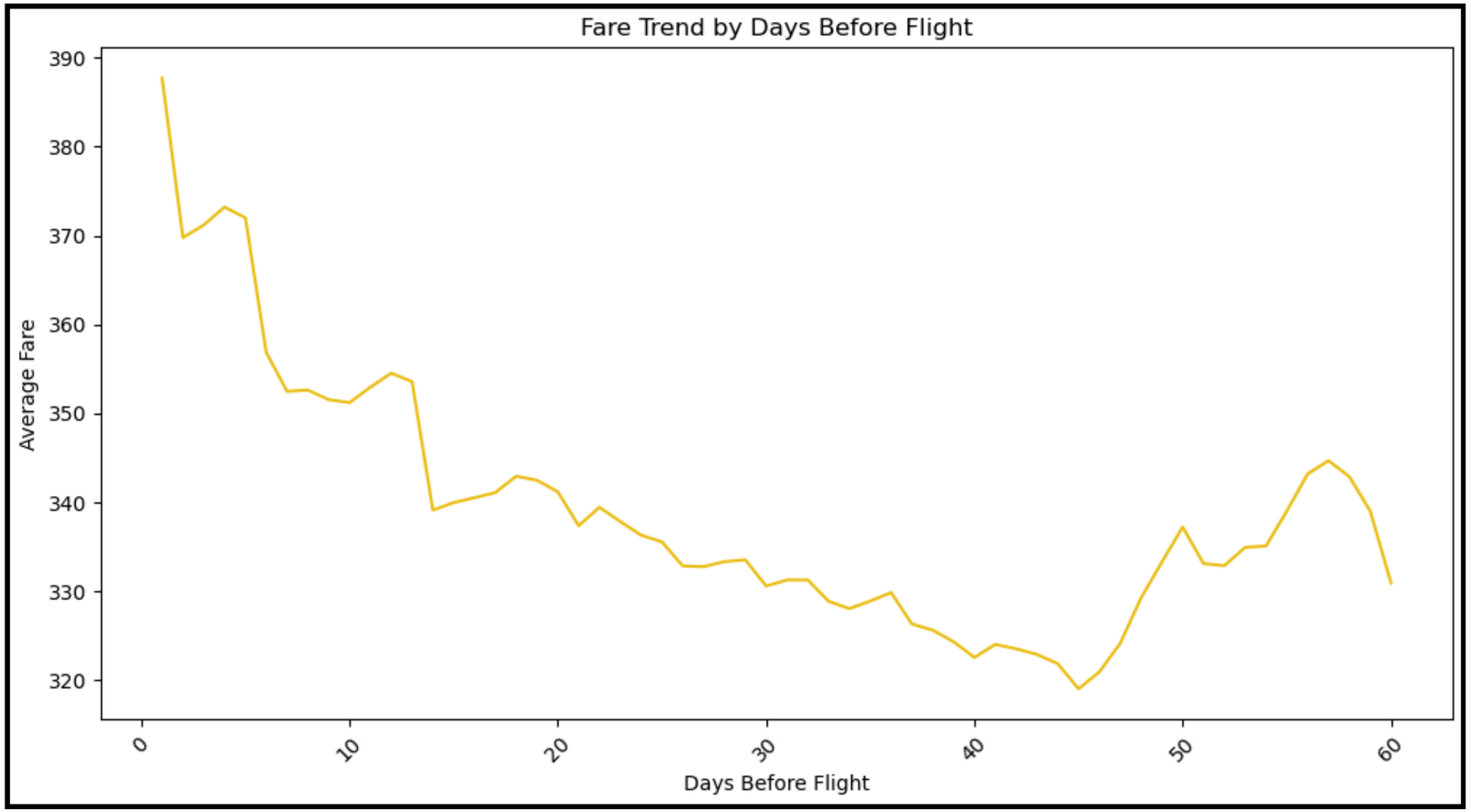


There is significant variability in average flight fare by carrier, with Hawaiian Airline costing the most and Spirit Airlines costing the least on average.

Fares Surge as Flight Date Approaches, With a Notable Pre-45-Day Increase

As days before flight decreases from 45-0, fare increases on average. However, between 45-60 days before the flight, the fare increases with increase in days before flight.

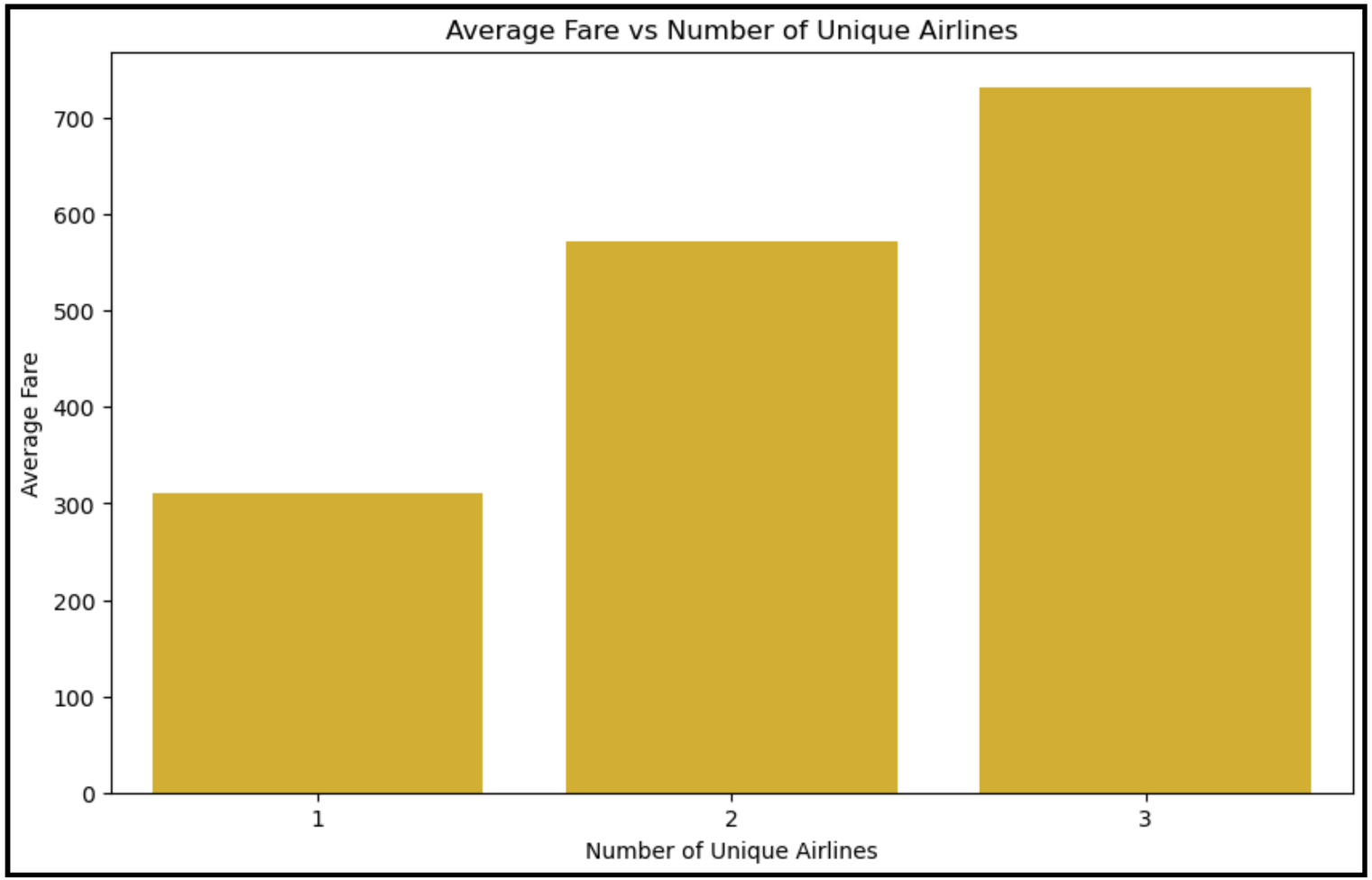
This suggests that early booking (45-60 days before) might not always guarantee the lowest fares, while last-minute bookings (closer to the flight date) could result in significant price hikes.



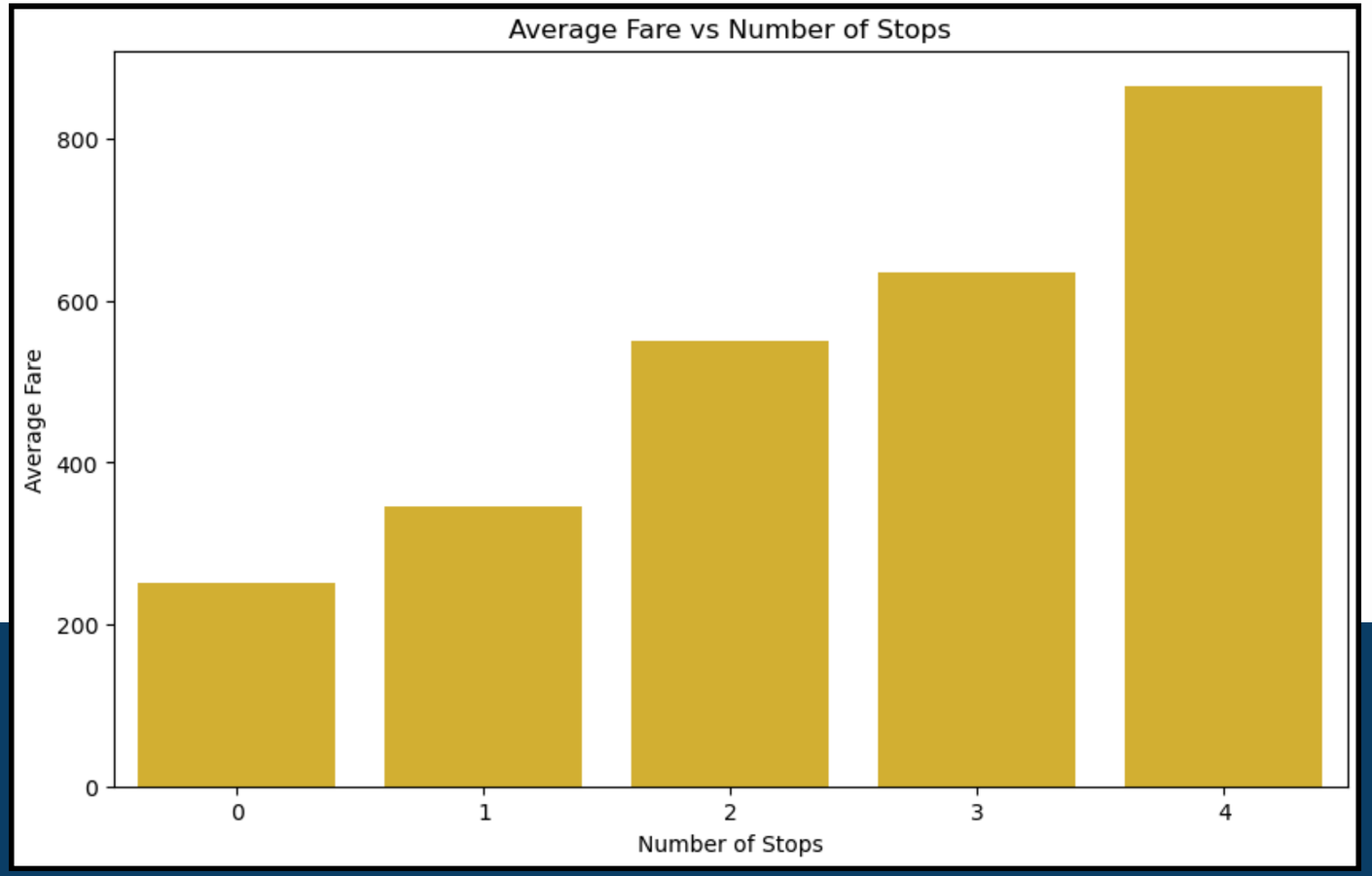
Direct Relationship Between Average Fare and Number of Unique Airlines

As there is an increase in the number of unique airlines an itinerary uses, its average fare also increases.

This is likely because the higher the number of unique airlines in an itinerary, the more layovers it has and the longer distance it travels.



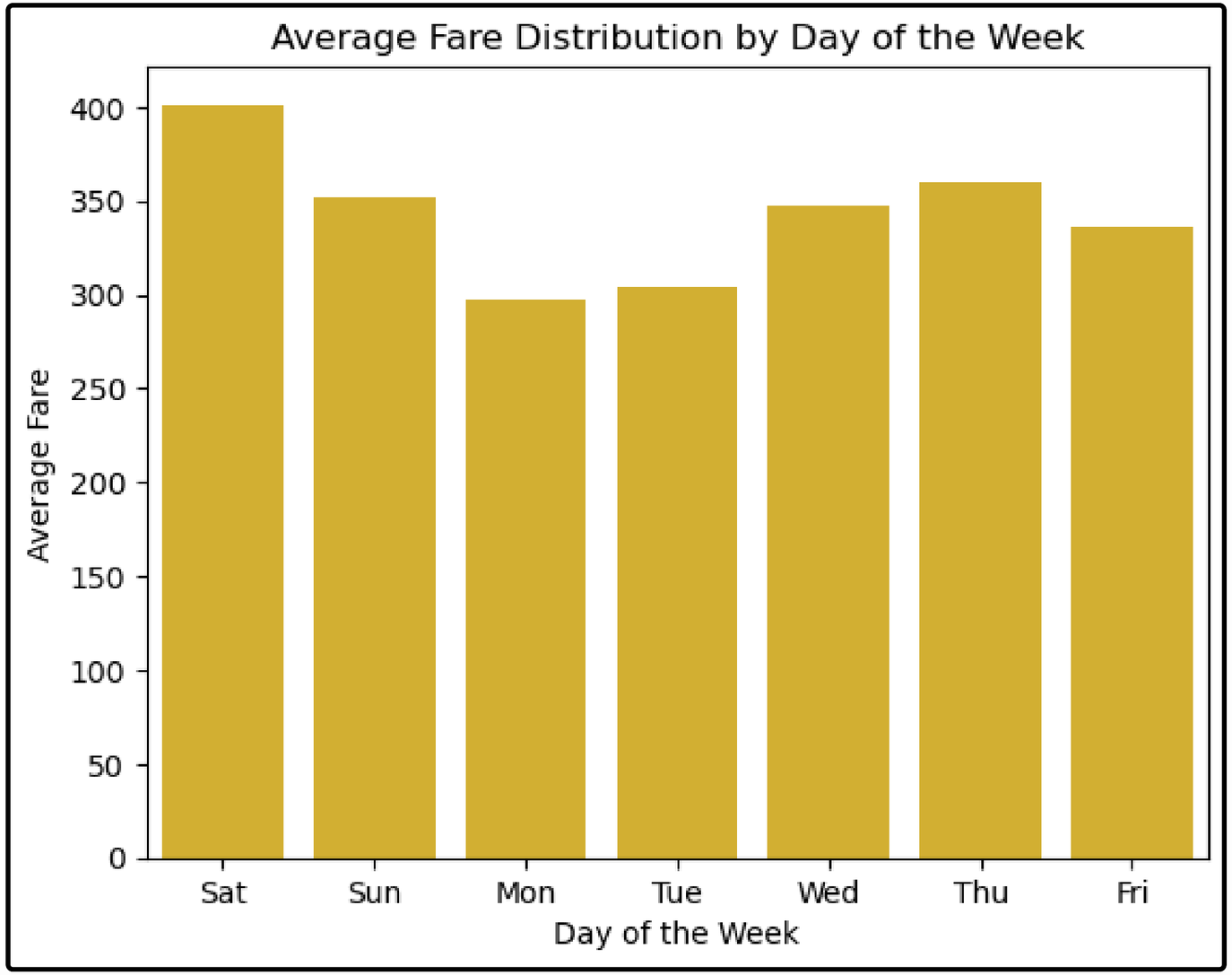
Direct Relationship Between Average Fare and Number of Stops



As the number of stops in a flight's journey increases, the average fare of the flight also increases. Similarly, as the number of stops in a flight's journey decreases, it's average fare also decreases.

Flight Price Distributions Across Days of the Week

Monday and Tuesday tend to have the lowest average flight prices while Saturday tends to have the highest average flight prices.



MACHINE LEARNING MODELS

Transforming Categorical Data for Model Success

One-hot Encoding
Starting and Destination Airport columns

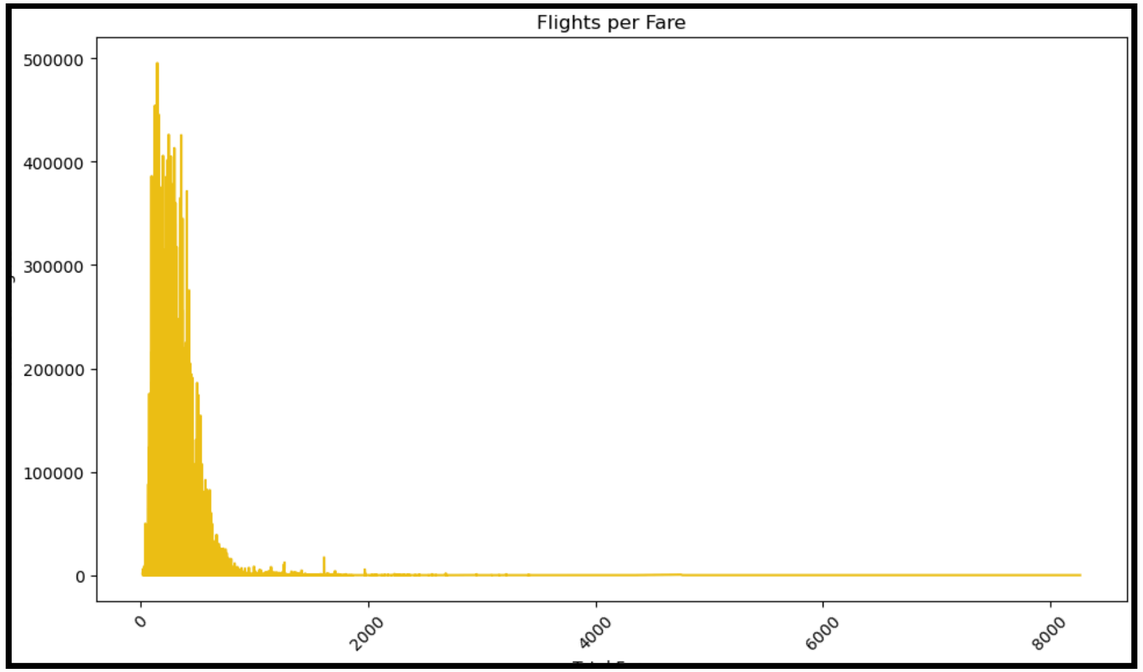
startingAirport	destinationAirport	startingAirport_onehot	destinationAirport_onehot
DTW	ATL	(15,[11],[1.0])	(15,[7],[1.0])
DEN	MIA	(15,[10],[1.0])	(15,[8],[1.0])
DEN	MIA	(15,[10],[1.0])	(15,[8],[1.0])
LAX	PHL	(15,[0],[1.0])	(15,[9],[1.0])
ATL	DFW	(15,[7],[1.0])	(15,[2],[1.0])

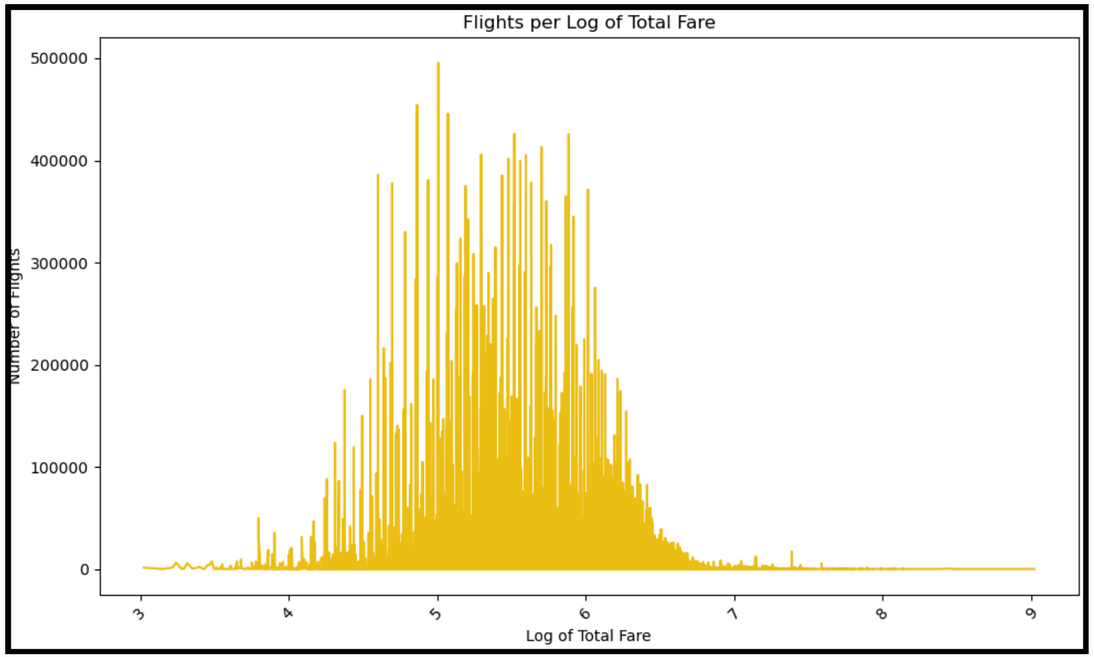
Exploding AirlineNames Array Into Columns
and One-hot Encoding the Columns

AirlineNames	AirlineName_0	AirlineName_1	AirlineName_2	AirlineName_3	AirlineName_4
[Delta, Delta]	Delta	Delta	null	null	null
[American Airline...	American Airlines	American Airlines	null	null	null
[American Airline...	American Airlines	American Airlines	null	null	null
[United, United]	United	United	null	null	null
[American Airlines]	American Airlines	null	null	null	null

AirlineName_0	AirlineName_0_onehot
Delta	(13,[1],[1.0])
American Airlines	(13,[0],[1.0])
American Airlines	(13,[0],[1.0])
United	(13,[2],[1.0])
American Airlines	(13,[0],[1.0])

Transforming Dependent Variable for Model Success: From Skewed to Symmetric





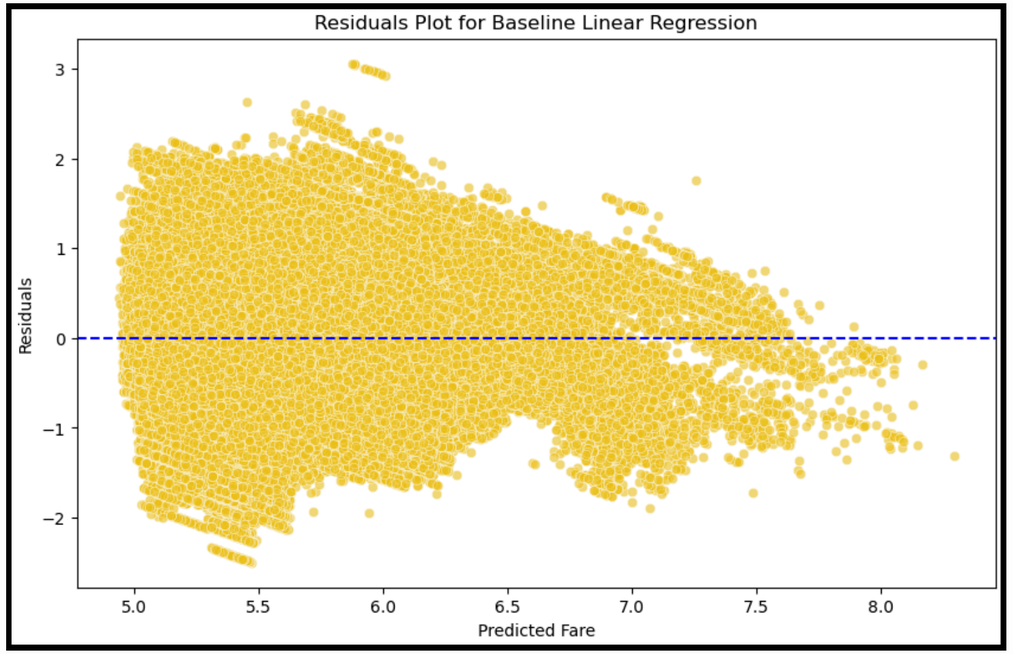
Log-transforming the right-skewed **totalFare** variable helps in achieving a more normal distribution, improves the R^2 of regression models, and improves model alignment with the assumptions of linear regression.

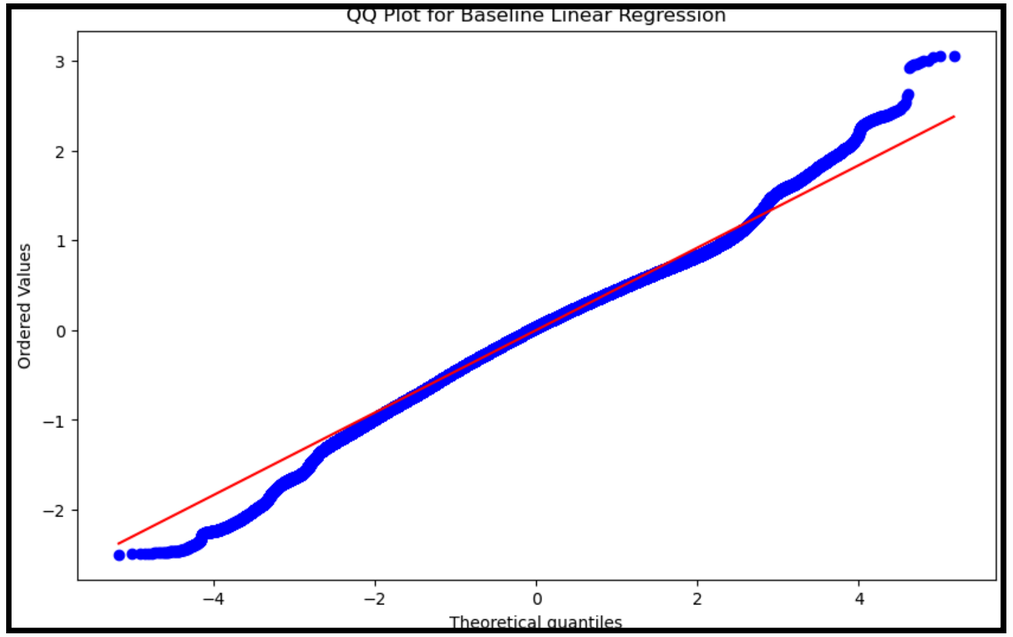
Establishing Baseline Model for Performance Comparison

Developed a linear regression model using features we found to be either moderately or highly correlated with the dependent variable.

Model Performance:
RMSE = 0.4602
R^2 = 0.3664

Features Included:
DaysBeforeFlight, NumStops, NumUniqueAirlines, NumUniqueCabins, travelDuration, isRefundable, hasFirstClass, totalTravelDistance, AircraftType, seatsRemaining, elapsedDays





Forward Selection vs. Backward Elimination Selected Features

Forward Selection

Features Selected:

totalTravelDistance, AirlineName_0_onehot , NumUniqueAirlines, startingAirport_onehot

Backwards Elimination

Features Selected:

travelDuration, elapsedDays, isRefundable, seatsRemaining, DaysBeforeFlight, NumStops, NumUniqueAirlines, AircraftType, NumUniqueCabins, hasFirstClass, startingAirport_onehot, destinationAirport_onehot, AirlineName_0_onehot

Interaction term created for NumUniqueCabins * hasFirstClass as they have a 0.52 correlation

Gradient Boosting Trees

- Captures Non-Linear Relationships
- Handles Feature Importance, enhancing interpretability.
- Robust to Noise and Outliers
- Better Predictive Performance

Feature Selection

totalTravelDistance	0.1973288
travelDuration	0.099217504
seatsRemaining	0.06987792
NumUniqueAirlines	0.0690371
DaysBeforeFlight	0.04918615
hasFirstClass	0.023707151
AirlineName_0_onehot	0.019429704
destinationAirport_onehot	0.014678557
AirlineName_3_onehot	0.0099891145
startingAirport_onehot	0.009672513
AirlineName_2_onehot	0.008560828
AircraftType	0.00854629
AirlineName_4_onehot	0.0066447747
NumUniqueCabins	0.006232057
NumStops	0.006142347
elapsedDays	0.0055798185
AirlineName_1_onehot	0.0036835736
isRefundable	0.0

Parameter Tuning

Used Random Search to optimize model hyperparameters

Model performance comparison

USING FORWARD SELECTION FEATURES

MODEL NAME	Train R^2	Train RMSE
LINEAR REGRESSION	0.4165	0.4417
L1 LASSO	0.2981	0.4844
L2 RIDGE	0.4087	0.4446
GLM	0.4579	0.4579

USING BACKWARDS ELIMINATION FEATURES

MODEL NAME	Train R^2	Train RMSE
LINEAR REGRESSION	0.4252	0.4384
L1 LASSO	0.2515	0.5002
L2 RIDGE	0.4150	0.4422
GLM	0.5767	0.3762

VS

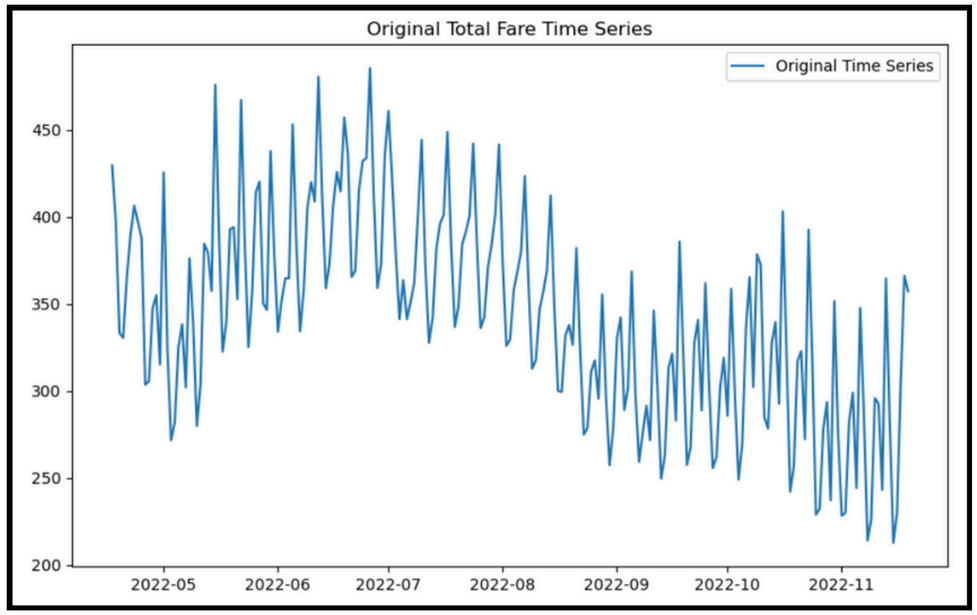
Test RMSE = 0.382

GBT Model

Train R^2	Train RMSE
0.612	0.3693

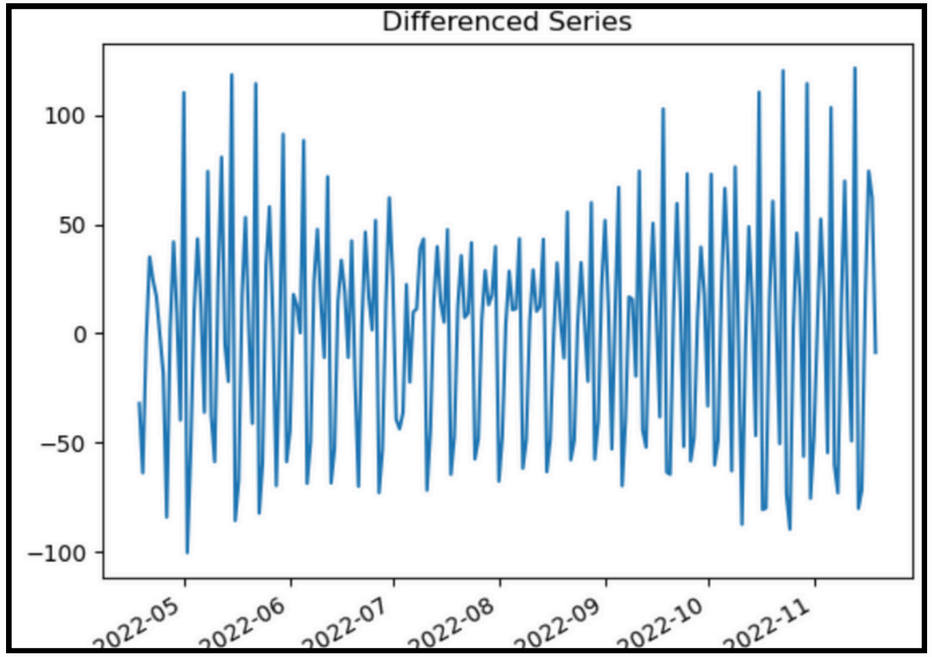
Ensuring Stationary Data for Use in the ARIMA Model

Analyze sequential flight price data collected over time to identify patterns, trends, and seasonality, enabling accurate forecasting of future ticket prices and supporting better decision-making for pricing strategies or purchase timing.



Downward trend/Cyclic Fluctuations
ADF test p-value: 0.769 (> 0.05)
Data is non-stationary

First-order Differencing

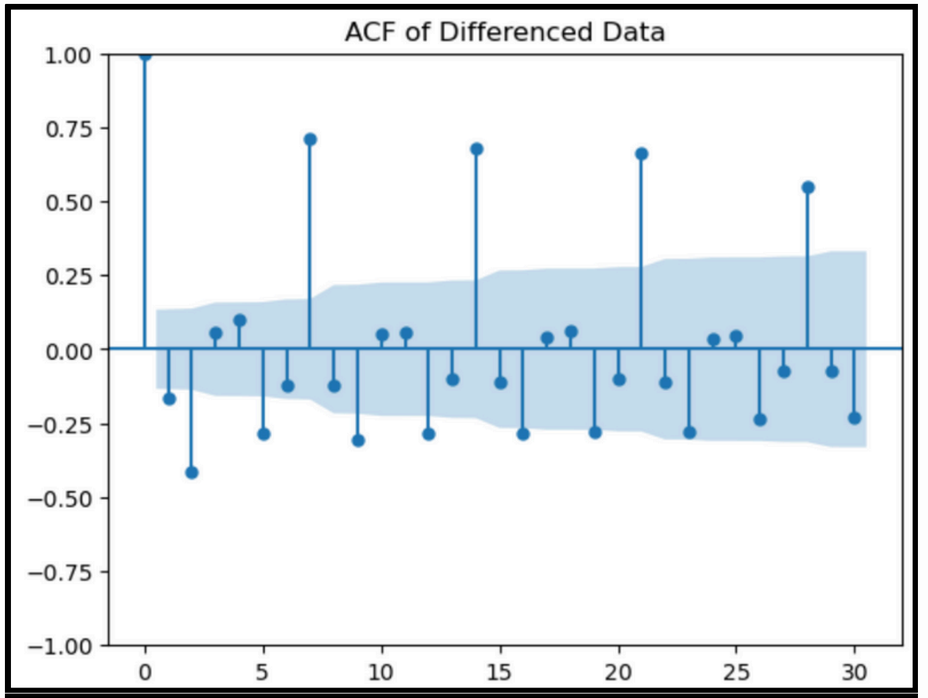


p-value: 1.79e-05 (< 0.05)
Data is stationary after differencing

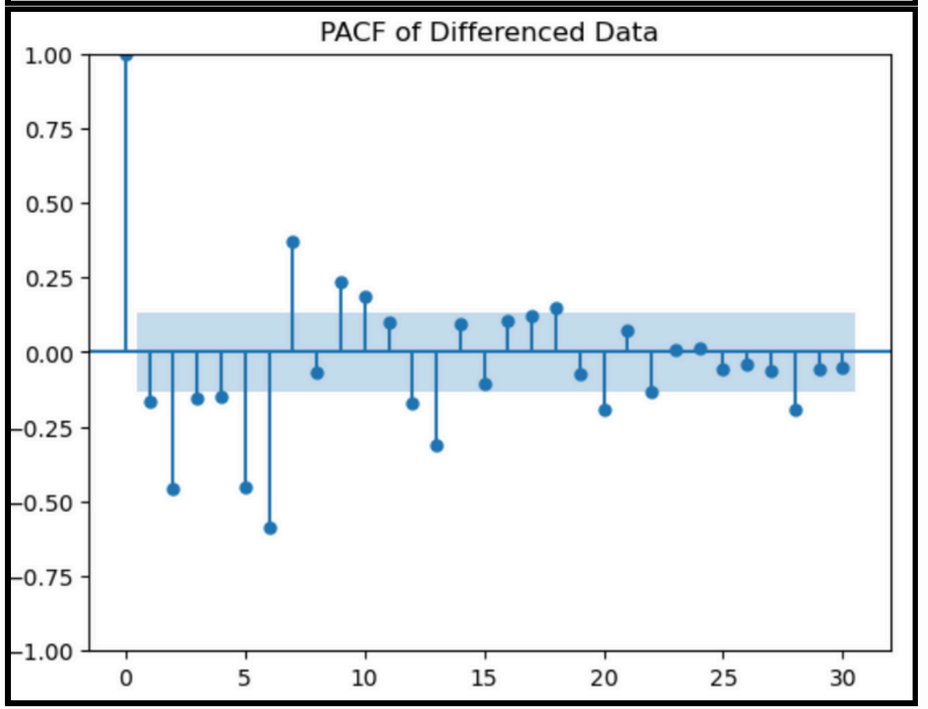
Creating the SARIMA Model for Seasonality of Fares

- **AR (Auto-Regressive): Uses past values to predict the current value.**
- **I (Integrated): Differencing to remove trends and make data stationary.**
- **MA (Moving Average): Uses past errors to improve predictions.**
- **Seasonal Components: Adds terms to account for periodic patterns.**

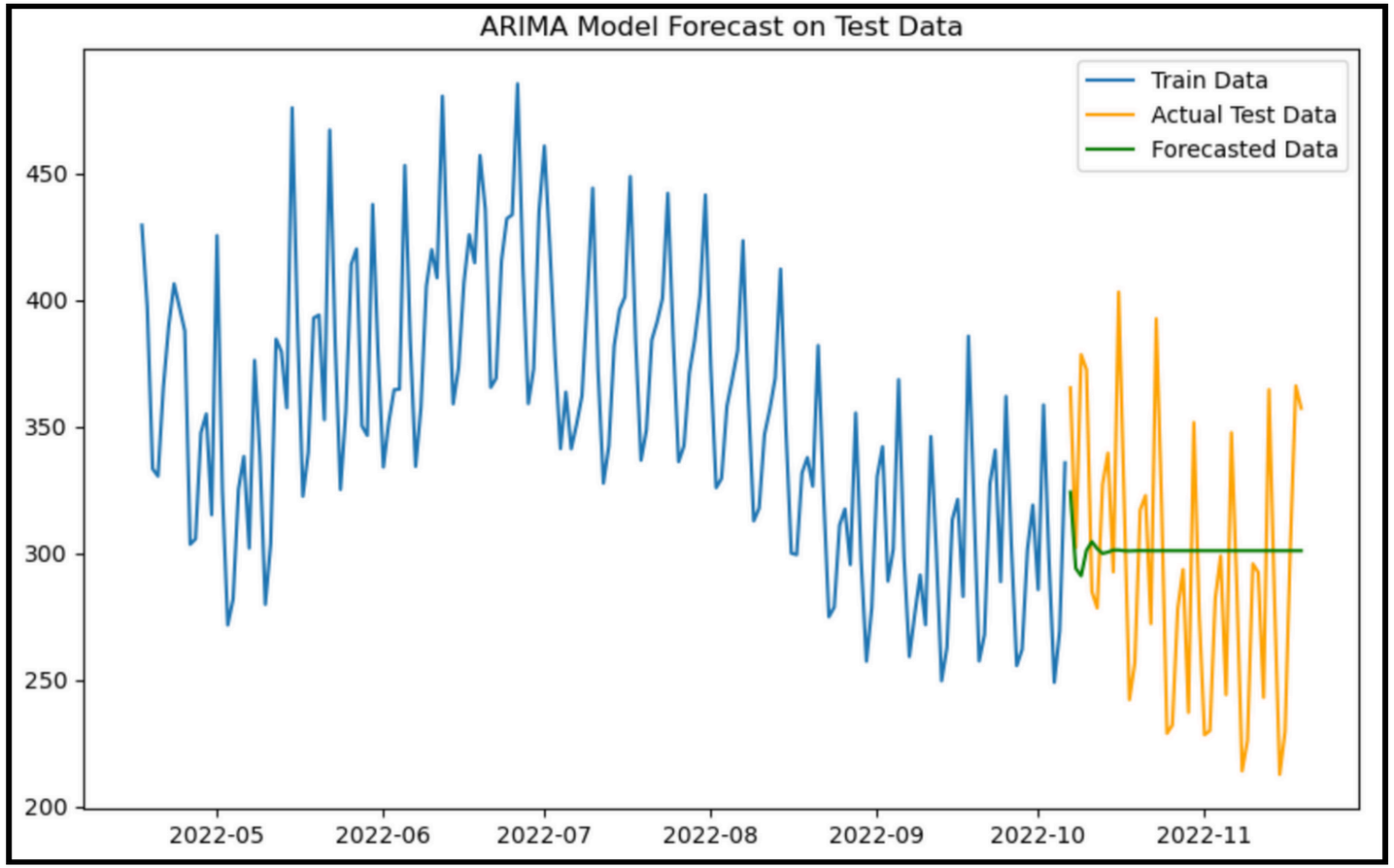
q = 2



p = 2



Evaluating Performance of SARIMA on Test Data



RMSE: 52.1423735511614
MAE: 43.613875720277136

While the SARIMA model captures some general trends, the forecasted values show less variability compared to the actual test data, suggesting that the model struggles with accurately predicting the full range of seasonal fluctuations.

Summary and Next Steps for Optimizing Flight Routes and Pricing

- Utilized data on flight itineraries from Expedia that contains information on flight routes, flight and search dates, aircraft and cabin types, and travel time.
- Implemented a graph-based model to use queries, motify finding, shortest paths algorithm, and page rank to answer customer questions.
  - Can be further used to optimize flight routes and airport hubs.
  - Utilize link-prediction and demand forecasting
- Predicted flight prices using linear regression, GBT, SARIMA model, and compared the model performance.
- Model improvement:
  - More columns to capture price patterns, for example convert flight date to Weekday and/or Month.
  - Continue tuning hyperparameters for GBT

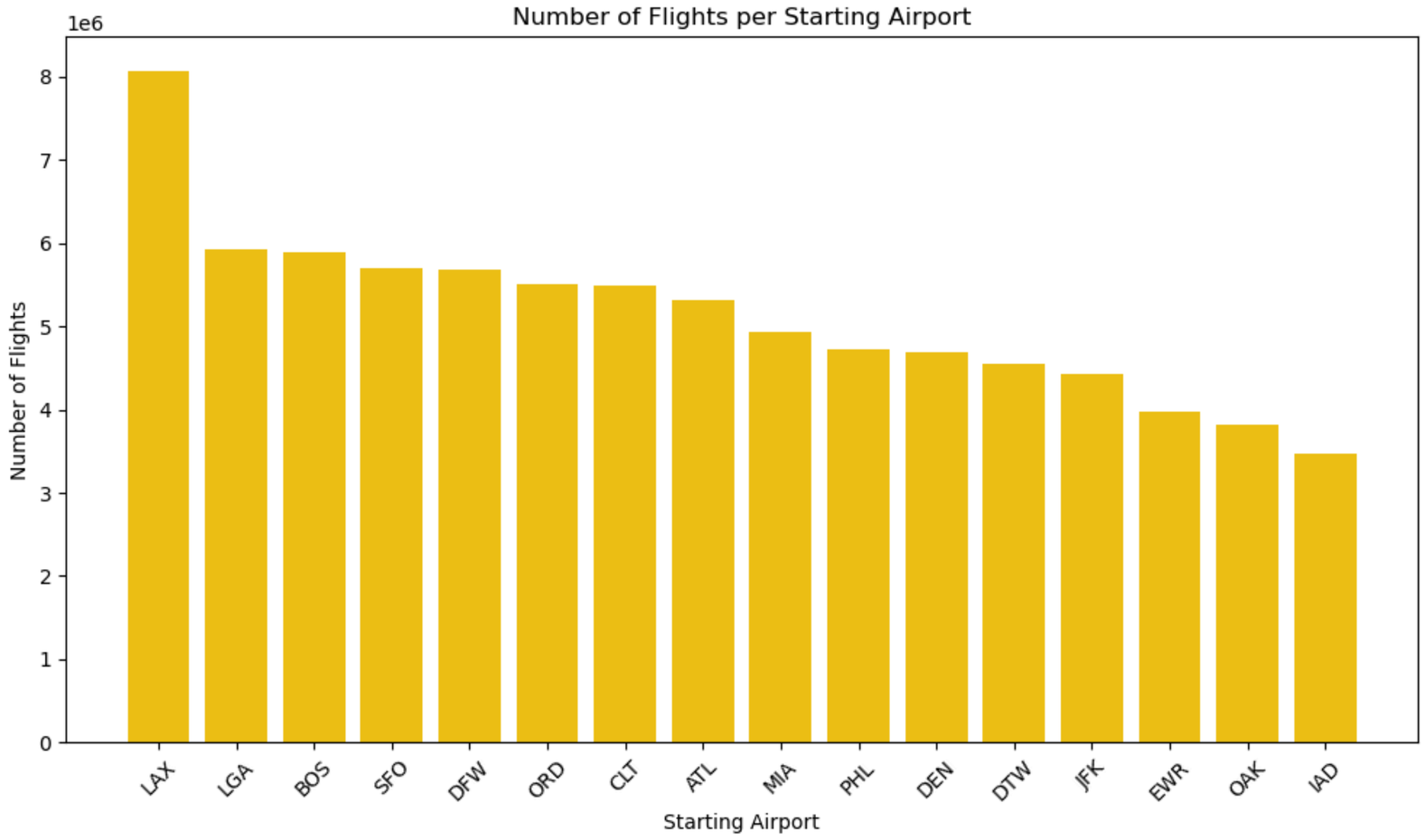
THANK YOU!

ANY QUESTIONS?

APPENDIX

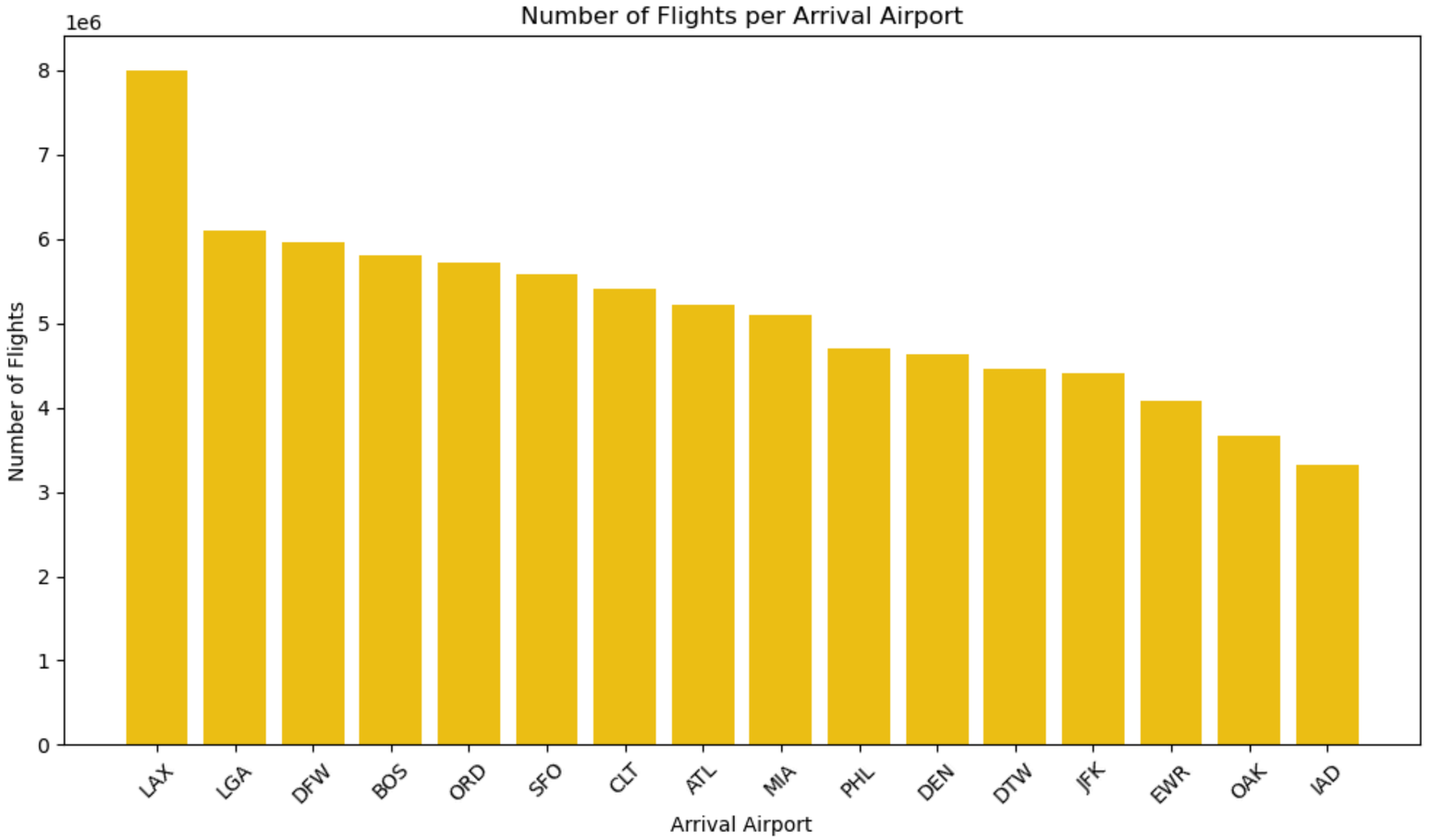
Analyzing Airport Traffic with Departing Flights

- LAX, LGA, BOS are the top 3 airports with the most departing flights.
- OAK and IAD have the least amount of departing flights across all airports.



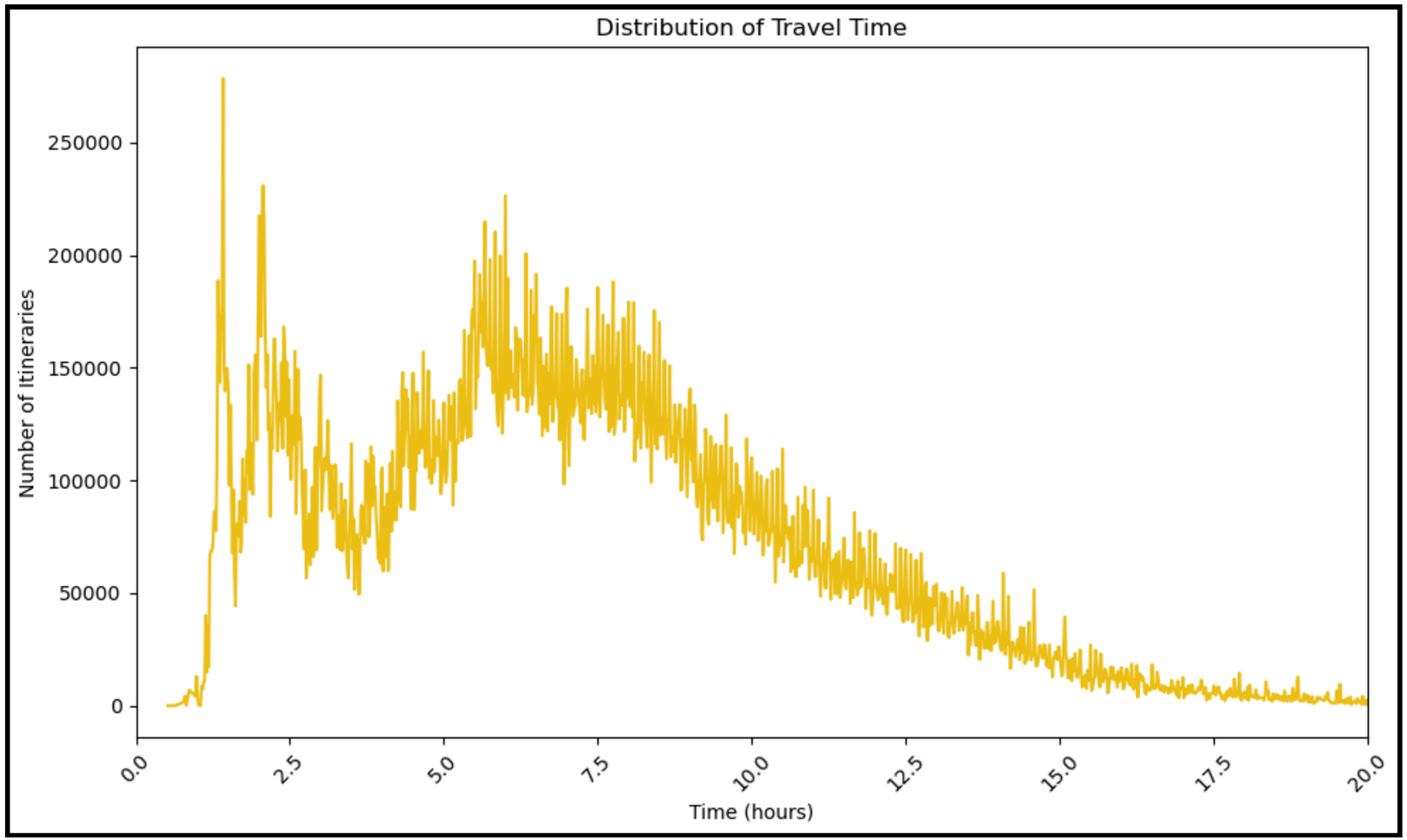
Analyzing Airport Traffic with Arriving Flights

- LAX, LGA, DFW are the top 3 airports with the most arriving flights.
- OAK and IAD again have the least amount of arriving flights across all airports.



Decline in Itineraries with Increasing Travel Time

- Most trips have less than 7.5 hours of travel time
- As travel time exceeds 7.5 hours, there is a sharp decrease in number of itineraries available
- Travel time peaks between 1-3 hours



Link Prediction for Direct Routes Between Airports

- Link Prediction on the graph using ORD as the source node
- Used Jaccard Index score to calculate node similarity
- All airports except for OAK have an index > 0.5 indicating that these airports should have direct routes from ORD in the future
- In fact, all airports have nonstop routes from ORD so this model has good prediction.

Airport Name	Jaccard Index
OAK	0.3125
LGA	0.625
BOS	0.8125
EWR	0.75
DEN	0.875
IAD	0.75
CLT	0.8125

Airport Name	Jaccard Index
MIA	0.8125
DFW	0.875
SFO	0.75
ATL	0.8125
DTW	0.875
LAX	0.875
JFK	0.625
PHL	0.75

SARIMA Model Summary

SARIMAX Results

Dep. Variable:	totalFare	No. Observations:	172
Model:	ARIMA(2, 1, 2)	Log Likelihood	-860.859
Date:	Mon, 02 Dec 2024	AIC	1731.717
Time:	03:18:36	BIC	1747.425
Sample:	0	HQIC	1738.091
	- 172		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2369	0.276	0.858	0.391	-0.304	0.778
ar.L2	-0.3579	0.148	-2.415	0.016	-0.648	-0.067
ma.L1	-0.6964	0.269	-2.587	0.010	-1.224	-0.169
ma.L2	-0.1138	0.243	-0.468	0.639	-0.590	0.362
sigma2	1369.6994	166.032	8.250	0.000	1044.283	1695.116

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	59.82
Prob(Q):	0.91	Prob(JB):	0.00
Heteroskedasticity (H):	0.65	Skew:	1.19
Prob(H) (two-sided):	0.11	Kurtosis:	4.64