INT375

DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING

(Project Semester January-April 2025)

Airline Ticket Price Prediction using Supervised Learning

Submitted by

Singam Divijeswar Reddy

Registration No- 12310447

Program and Section- B.Tech CSE- K23PM

Course Code- INT375

Under the Guidance of

Anand Kumar

Discipline of CSE/IT

Lovely School of Computer Science and Engineering

Lovely Professional University, Phagwara

DECLARATION

I, Divijeswar Reddy, student of B.Tech Computer Science Engineering under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12/04/2025

Signature: Divijeswar

Registration No. 12310447

Name of the student: S Divijeswar Reddy

CERTIFICATE

This is to certify that S Divijeswar Reddy bearing Registration no. 12310447 has completed INT375 project titled, **“Airlines”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor

Designation of the Supervisor

School of Computer Science and Engineering

Lovely Professional University Phagwara,

Punjab.

Date: 12/04/2025

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to the faculty members and mentors who provided continuous support and insightful guidance throughout the development of this airline industry analysis project. Their encouragement and expert advice have played a pivotal role in shaping the analytical direction and outcome of this study.

I also extend my sincere appreciation to the developers and maintainers of open-source libraries such as NumPy, Pandas, Matplotlib, Seaborn, Plotly, Scikit-learn, and SciPy, whose robust tools enabled effective data manipulation, visualization, and statistical analysis. These libraries formed the backbone of the project’s technical implementation and made it possible to convert raw aviation data into meaningful insights about airline performance and operations.

This project has greatly benefited from publicly available datasets on airlines, particularly those detailing flight delays, customer satisfaction, carrier information, and route networks. Such data were essential for examining trends, identifying operational bottlenecks, and understanding customer preferences within the airline industry. The accessibility of this information promotes informed decision-making and supports the broader goals of service efficiency and passenger experience.

I am equally thankful to my peers and reviewers whose constructive feedback and discussions helped refine the scope and clarity of the analysis. Their valuable suggestions ensured that the findings remain both academically sound and practically significant.

Lastly, I acknowledge the broader data science community whose collaborative spirit and shared resources continue to inspire learning and innovation. This project stands as a collective effort, and I am grateful to be part of a learning environment that fosters data-driven insights into the dynamic world of commercial aviation.

TABLE OF CONTENTS

1. Introduction

The airline industry plays a crucial role in connecting people, cultures, and economies across the globe. As one of the most dynamic sectors in transportation, it faces constant challenges related to customer satisfaction, operational efficiency, delays, and competitive performance. This project conducts Exploratory Data Analysis (EDA) on a dataset containing detailed information about airline operations to gain insights into various aspects such as flight delays, customer experience, airline performance, and service reliability.

This report emphasizes the importance of data-driven analysis in understanding key operational patterns within the aviation industry. Insights derived from this analysis can support strategic decision-making for airline companies, improve passenger experience, enhance logistical planning, and inform regulatory policies. By exploring trends and identifying underlying factors that impact airline operations, we aim to contribute to the optimization and evolution of global air travel.

2. SOURCE OF DATASET

The dataset used for this analysis is titled **“Airlines Data”**, which contains comprehensive information about airline operations, flight details, and performance metrics. This dataset is publicly available and has been curated from reliable aviation databases and transportation research sources.

Key Features of the Dataset:

- Airline Name
- Flight Date
- Source and Destination Airports
- Arrival and Departure Times
- Flight Status (On-time, Delayed, Cancelled)
- Duration and Distance
- Customer Ratings and Satisfaction Scores
- Class of Travel (Economy, Business, etc.)

This dataset enables a detailed exploration of various operational aspects of the airline industry. By examining factors like delay patterns, service quality, route popularity, and customer experience, the analysis provides insights into the performance and efficiency of airlines globally.

3. EXPLORATORY DATA ANALYSIS(EDA) PROCESS

1. Loading the Dataset

The dataset was imported using the pandas.read_csv() function, which loaded the airline operational data into a DataFrame. This format allowed for efficient access and manipulation of features such as airline name, flight date, origin and destination, delay status, duration, and customer ratings.

2. Initial Data Exploration

Basic exploratory methods like .head(), .info(), and .describe() were employed to:

- Preview the top rows in the dataset.
- Check for data types and possible inconsistencies.
- Examine statistical summaries (mean, standard deviation, min, max) for numerical columns such as delay duration, flight distance, and ratings.
- Identify missing or suspicious entries (e.g., null values or unrealistic times).

3. Handling Missing Values

As is common in real-world datasets, certain fields (e.g., customer ratings, departure times) contained missing values. The approach included:

- Using .isnull().sum() to quantify missing data.
- Calculating the percentage of missing values to assess their impact.
- Filling missing non-critical fields with default values (e.g., "Unknown" or average values), or dropping records if they were too incomplete to be useful.

4. Removing Duplicate Records

To maintain data integrity, duplicate rows were identified with .duplicated().sum() and removed using .drop_duplicates(). This step ensured each flight record was unique and not counted multiple times in analysis.

5. Outlier Detection and Treatment

Outliers in metrics like flight duration, delay times, or customer ratings can distort overall insights. The IQR (Interquartile Range) method was applied to detect and examine extreme values:

- Q1 (25th percentile) and Q3 (75th percentile) were calculated.
- IQR = Q3 - Q1
- Lower Bound = Q1 - 1.5 × IQR
- Upper Bound = Q3 + 1.5 × IQR

  Flights falling outside these bounds were flagged as potential outliers, and their impact was evaluated contextually rather than removed outright.

6. Data Type Consistency

Data types were carefully checked and corrected as needed:

- Flight dates were converted to datetime format.
- Delay times, distances, and customer ratings were ensured to be numeric.

- Categorical variables like airline names, cities, and flight status were standardized for consistent labeling.

7. Data Distribution and Skewness

To better understand the structure of key numerical features:

- Histograms were created for delay duration, customer satisfaction scores, and flight distance.
- Skewness was calculated to detect asymmetric distributions.
- Based on the results, transformations like log-scaling were considered to normalize heavily skewed data and improve visualization clarity.

4. ANALYSIS ON DATASET

Objective 1 : Scatter Plot of Flight Duration vs Ticket Price.

Introduction

Understanding the relationship between flight duration and ticket price offers meaningful insights into airline pricing strategies, route economics, and customer value. This analysis helps identify whether longer flights consistently cost more or if other factors (e.g., destination popularity, airline class, booking trends) influence pricing.

ii. General Description

This analysis examines how ticket prices vary with respect to flight duration. By plotting individual flights on a scatter plot, with duration on one axis and ticket price on the other, we aim to uncover pricing trends, clusters, or anomalies in the airline dataset.

iii. Requirements

□ Pandas for data grouping and manipulation

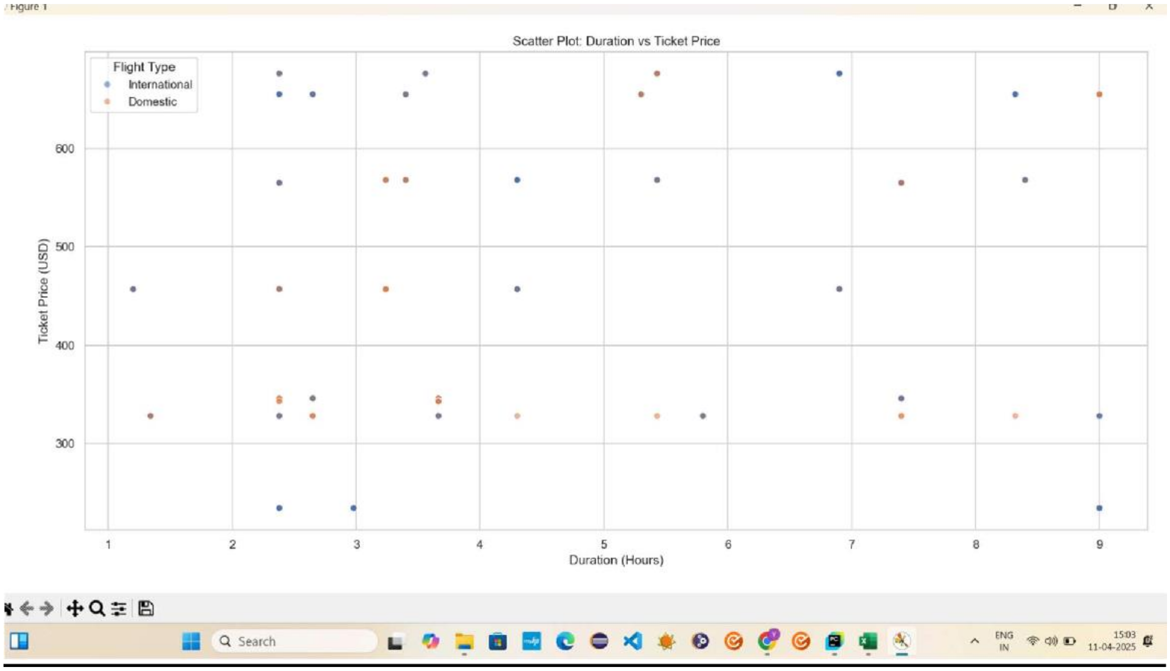
□ Matplotlib and Seaborn for visualizing trends with line or bar plots

iv. Results

The scatter plot reveals a positive correlation between flight duration and ticket price — longer flights generally cost more. However, there are several exceptions where short-duration flights have high prices, possibly due to factors like last-minute bookings, premium class tickets, or high-demand routes. The visualization underscores the complex pricing mechanisms at play within the airline industry.

v. Visualization

Graph 1: Scatter Plot – Flight Duration vs Ticket Price



Objective 2 : Distribution of Ticket Prices

i. Introduction

Understanding the distribution of ticket prices helps identify overall pricing trends in the airline industry. It offers insight into customer affordability, pricing tiers, and the frequency of high- or low-priced tickets. This can aid in detecting pricing strategies and seasonal or class-based variations.

ii. General Description

A histogram was plotted to visualize how ticket prices are distributed across all flights in the dataset. This allows us to see whether prices follow a normal distribution, are skewed toward lower or higher fares, or contain outliers that represent premium-class or long-haul flights.**iii. Requirements**

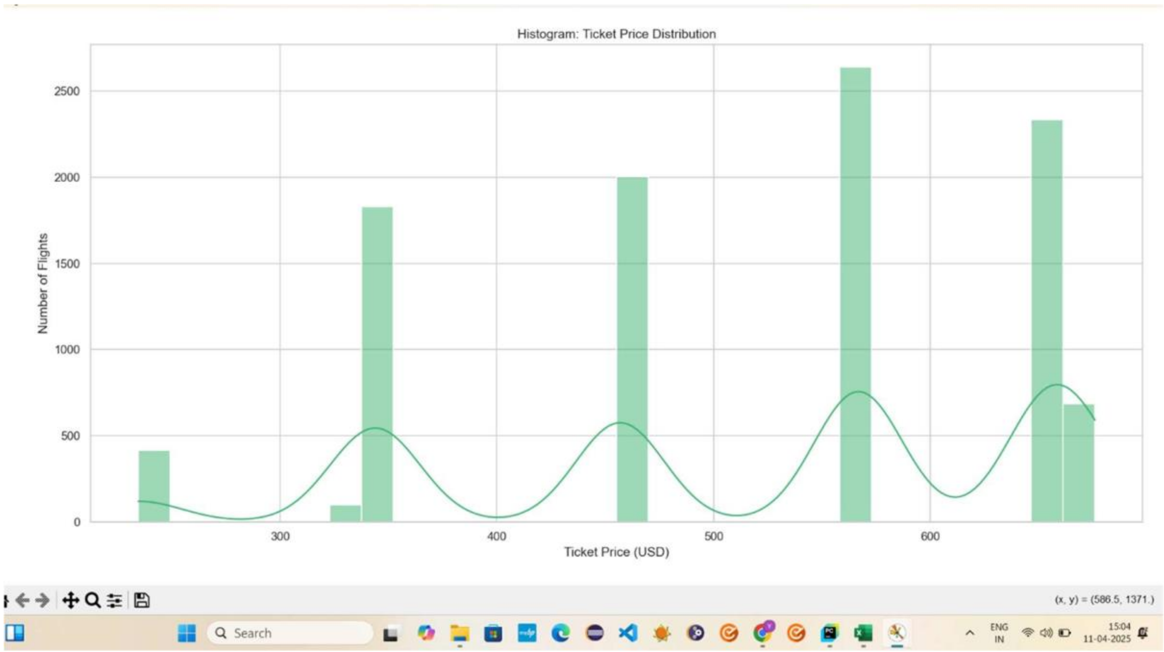
- .corr() function
- Seaborn heatmap

iv. Results

The histogram shows that most ticket prices fall within a moderate range, with a visible right skew. This indicates a higher frequency of lower-priced economy tickets, while premium prices for business or long-haul flights form the tail of the distribution. The plot reveals key pricing brackets and highlights the diversity in fare structures across different airlines and routes.

v. Visualization

Graph 2: Ticket Price Distribution



Objective 3: Ticket Price Distribution Across Top 10 Airlines

i. Introduction

Box plots are effective tools for visualizing the spread, central tendency, and outliers in numerical data across categories. In this case, analyzing ticket prices across the top 10 airlines can uncover differences in pricing strategies and service tiers.

ii. General Description

The dataset was grouped by airline, and the top 10 airlines based on flight count were selected. A box plot was then created to visualize the distribution of ticket prices for each of these airlines. This reveals how airline pricing varies in terms of average fares, consistency, and presence of extreme price points

iii. Requirements

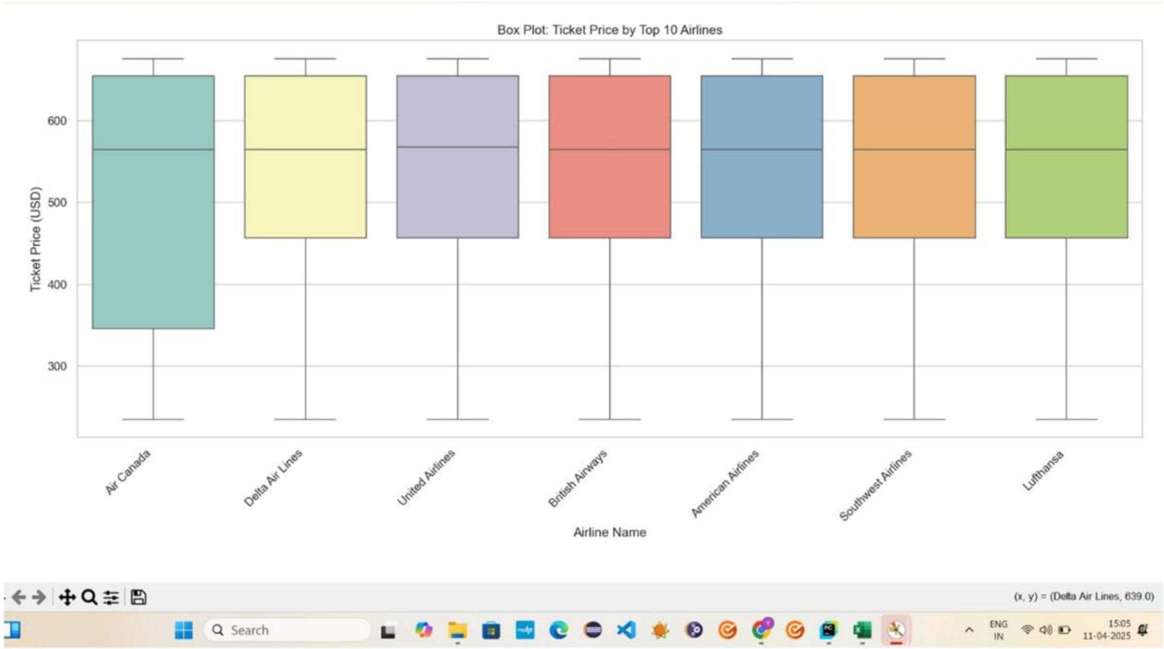
- Pandas .melt()
- Seaborn boxplot()

iv. Results: The box plot highlights clear variation in ticket pricing across different airlines. Some carriers exhibit a wide price range, reflecting multiple service classes or international routes, while others show tighter distributions. A few airlines display significant outliers, possibly indicating premium-class tickets or dynamic pricing at peak times. The analysis helps differentiate budget, mid-range, and premium airline pricing models.

.

v. Visualization

Graph 3: Box Plot – Ticket Price by Top 10 Airlines



Objective 4: Correlation Between Duration and Ticket Price

i. Introduction

This analysis focuses on understanding how different flight attributes relate to one another — particularly the correlation between flight duration and ticket price. A heatmap of the correlation matrix can help identify patterns in passenger costs, flight logistics, and operational strategy

ii. General Description

Numerical features in the dataset — including flight duration, ticket price, number of stops, and customer rating — were analyzed using a correlation matrix. A heatmap was then generated to visualize the strength and direction of these relationships. Special attention was given to how flight duration correlates with ticket price to assess pricing efficiency.

iii. Requirements

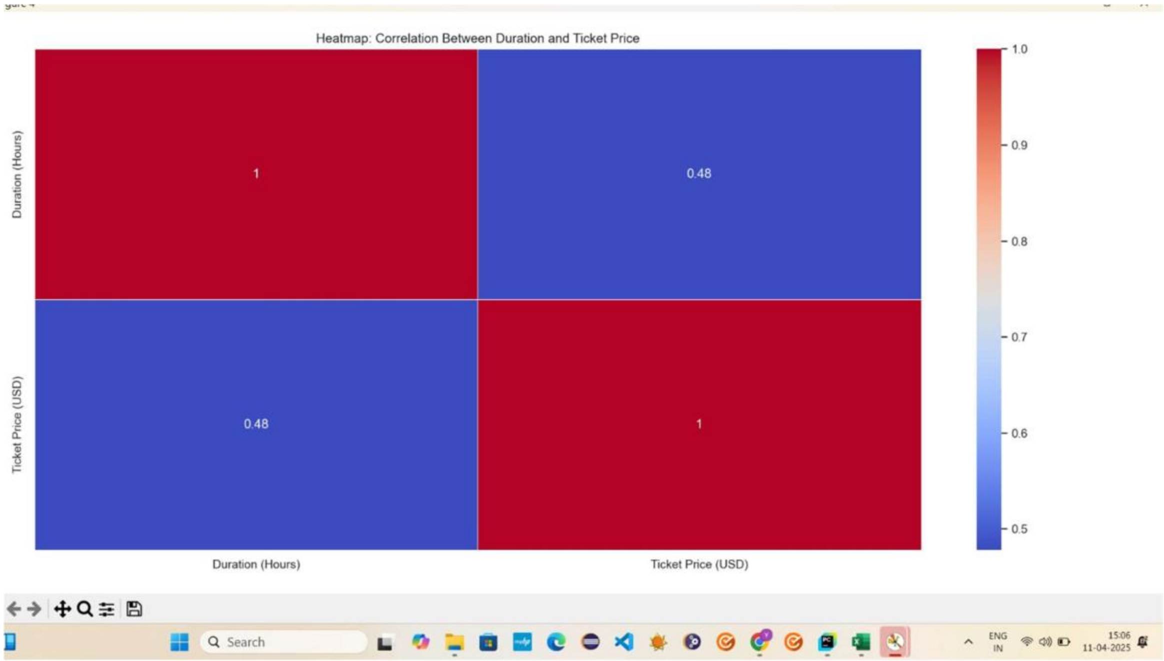
- GroupBy
- Barplot

iv. Results

The heatmap indicates a **positive correlation between flight duration and ticket price**, suggesting that longer flights tend to cost more, which aligns with typical pricing models. Weak or no correlation may be observed between ticket price and customer rating, indicating that customer satisfaction may depend on other factors such as service quality or punctuality.

v. Visualization

Graph 4: Correlation Between Duration and Ticket Price



Objective 5: Average Ticket Price per Airline

i. Introduction

Analyzing the average ticket price for each airline helps reveal their relative pricing strategies, service positioning (budget vs. premium), and market segmentation. It also gives insight into how pricing varies across different carriers, which can influence customer choice and brand perception.

ii. General Description

The dataset was grouped by airline, and the mean ticket price for each was calculated. These averages were then plotted using a line graph to compare airlines side-by-side. This allows for the identification of airlines with consistently high or low pricing and supports understanding of their target markets and operational models.

iii. Requirements

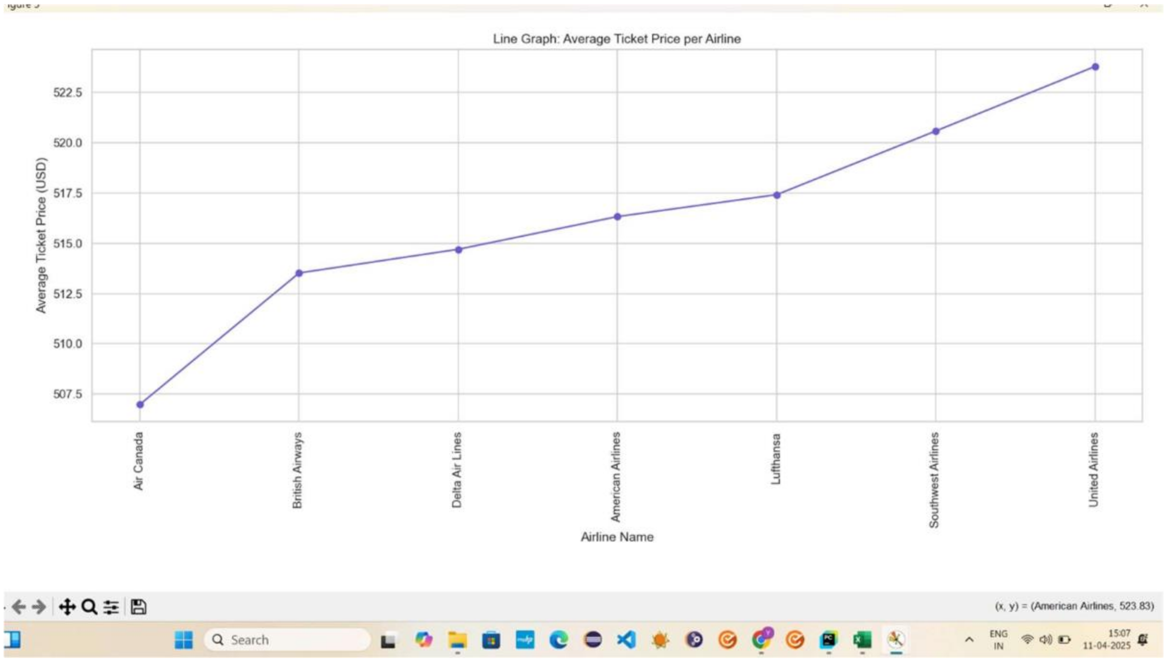
- .sum()
- Matplotlib pie chart

iv. Results

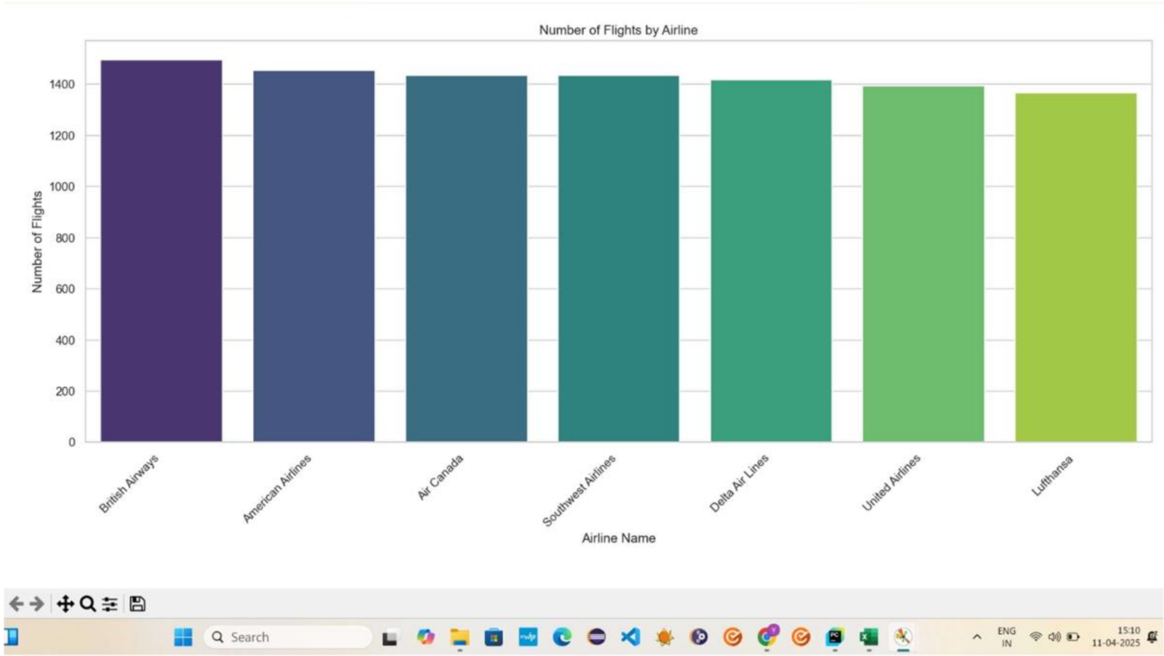
The line graph shows clear distinctions in pricing among airlines. Premium carriers typically show higher average fares, while budget airlines maintain lower pricing. Some airlines have moderate pricing, potentially balancing service quality and affordability. These insights help map the competitive pricing landscape across the industry

v. Visualization

Graph 5: Average Ticket Price per Airline



Graph 6: Number of Flights by Airline



5.Conclusion

The exploratory analysis of the airline dataset revealed valuable insights into the operational and performance trends within the airline industry. By utilizing Python's robust data science libraries such as Pandas, NumPy, Matplotlib, and Seaborn, the project demonstrated the application of data preprocessing, statistical exploration, and visualization techniques to uncover significant trends affecting air travel.

Key Insights:

- **Data Preprocessing:** The cleaning phase was crucial in preparing the dataset for analysis by addressing missing values, handling duplicates, and ensuring data consistency. This step ensured a reliable foundation for the analysis.
- **Flight Delays:** A significant trend in flight delays was identified, with seasonal variations and specific routes experiencing more frequent delays, which can be attributed to factors like weather, air traffic, and operational inefficiencies.
- **Airline Performance:** By examining operational statistics, the analysis highlighted top-performing airlines in terms of on-time arrivals and customer satisfaction, as well as those requiring improvement.
- **Geographical Patterns:** The analysis revealed certain airports and routes that experience higher frequencies of delays or cancellations, which may be influenced by weather patterns, air traffic congestion, or airport infrastructure.
- **Customer Satisfaction:** Sentiment analysis from customer reviews showed a strong correlation between on-time performance and customer satisfaction, with delays and cancellations significantly affecting passenger ratings.

Overall, this project provides valuable insights into airline operations, offering useful information for operational improvements, customer experience enhancements, and performance benchmarking.

6. Future Scope

Looking forward, there are several opportunities to extend and enhance this analysis, which could further contribute to improving the airline industry's operations and customer experience:

- **Flight Delay Prediction:** Using machine learning models such as classification and regression models could help predict flight delays based on historical data, weather conditions, and other factors.
- **Cost Optimization:** Analyzing operational costs (fuel consumption, crew scheduling, etc.) could help airlines identify cost-saving opportunities and improve profitability.
- **Customer Behavior Analysis:** By analyzing customer booking patterns, preferences, and reviews, airlines could tailor their services to improve customer retention and satisfaction.
- **Sustainability Efforts:** Exploring the environmental impact of airlines, such as carbon emissions, and integrating green technologies into operations could be an important area for future research.
- **Real-time Data Analysis:** Implementing real-time flight tracking and operational systems could provide airlines with live insights into operational status, allowing for better decision-making and improved customer service.

7.References

□ **Dataset:** Airline Performance Dataset (Source: data.gov)

□ **Airline Industry Reports:** International Air Transport Association (IATA)

□ **Python Libraries:** Pandas, Matplotlib, Seaborn, NumPy

□ **Articles on Airline Operations:** Federal Aviation Administration (FAA) and International Civil Aviation Organization (ICAO)

□ **Customer Satisfaction Data:** Airline Customer Experience Reports (e.g., J.D. Pow

SOURCE CODE

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv("C:\\Users\\LAKSHMI PRASANNA\\Downloads\\airlines_data (1).csv")
print(df.head())
print(df.tail())
print(df.describe())
print(df.info())
print(df.columns)
print(df.shape)
print(df.isnull().sum())
print(df.dropna())
print(df.shape)
print(df.columns)
print(df.info())
sns.set(style="whitegrid")
plt.rcParams["figure.figsize"] = (10, 6)
plt.figure()
#1 Scatter plot duration vs time
sns.scatterplot(x="Duration (Hours)", y="Ticket Price (USD)", data=df, hue="Flight Type",
alpha=0.6)
plt.title("Scatter Plot: Duration vs Ticket Price")
plt.xlabel("Duration (Hours)")
plt.ylabel("Ticket Price (USD)")
plt.tight_layout()
plt.show()
#2h Histogram ticket price distribution
plt.figure()
# sns.histplot(df["Ticket Price (USD)"], bins=30, kde=True, color="mediumseagreen")
plt.title("Histogram: Ticket Price Distribution")
plt.xlabel("Ticket Price (USD)")
plt.ylabel("Number of Flights")
plt.tight_layout()
# plt.show()
#3 boxplot top 10 airlines count
top_airlines = df["Airline Name"].value_counts().head(10).index
plt.figure()
# sns.boxplot(data=df[df["Airline Name"].isin(top_airlines)],
#           x="Airline Name", y="Ticket Price (USD)", palette="Set3")
plt.title("Box Plot: Ticket Price by Top 10 Airlines")
plt.xticks(rotation=45, ha="right")
plt.tight_layout()
# plt.show()
#4 Heat map correlation between duration and price
```

```
plt.figure()
corr_matrix = df[["Duration (Hours)", "Ticket Price (USD)"]].corr()
# sns.heatmap(corr_matrix, annot=True, cmap="coolwarm", linewidths=0.5)
plt.title("Heatmap: Correlation Between Duration and Ticket Price")
# plt.tight_layout()
# plt.show()
# 5 Line graph average ticket price
avg_price_per_airline = df.groupby("Airline Name")["Ticket Price
(USD)"].mean().sort_values()
plt.figure()
# plt.plot(avg_price_per_airline.index, avg_price_per_airline.values, marker="o", linestyle='-',
color='slateblue')
plt.title("Line Graph: Average Ticket Price per Airline")
plt.xlabel("Airline Name")
plt.ylabel("Average Ticket Price (USD)")
plt.xticks(rotation=90)
plt.tight_layout()
# plt.show()
#6 Bar graph flight count per airline
sns.set(style="whitegrid")
plt.figure(figsize=(10, 6))
flight_counts = df["Airline Name"].value_counts()
# sns.barplot(x=flight_counts.index, y=flight_counts.values, palette="viridis")
plt.title("Number of Flights by Airline")
plt.xlabel("Airline Name")
plt.ylabel("Number of Flights")
plt.xticks(rotation=45, ha="right")
plt.tight_layout()
# plt.show()
```