Data Mining and Business Intelligence

Experiment - 3

Name: **Abhishek Vishwakarma**
Class: **D15C**
Roll No: **73**

Aim: To perform exploratory data analysis and visualization on the dataset using python.

Introduction: Exploratory Data Analysis (EDA) is an essential step in the data analysis process that involves summarizing and visualizing the dataset to understand its underlying structure, detect patterns,identify anomalies, and generate insights. It helps in making informed decisions about data preprocessing, feature engineering, and model selection in later stages of analysis. EDA consists of descriptive statistics, data visualization, and correlation analysis, which provide a comprehensive understanding of the dataset.

Descriptive analysis - Central tendency:

Definition: Central tendency refers to statistical measures that identify the center or typical value in a dataset. The most common measures are mean, median, and mode, which help us understand the "average" behavior of numerical variables such as ticket price, duration of flights, and number of days left before departure.

Execution: We can calculate mean, median, and mode for price, duration, and days_left in our dataset.

Inference:

The **mean** price gives the average cost of flights across all airlines.
The **median** price tells us the middle value, which is useful if there are extreme outliers (very expensive tickets).
The **mode** may reveal frequently occurring values, e.g., common prices due to standard fare classes.
This helps us summarize the central behavior of ticket pricing and flight durations.

Descriptive Analysis – Dispersion

Definition: Dispersion measures how spread out the data is. The main metrics are range, variance, and standard deviation. These values show how much variability exists in ticket prices, flight durations, and days left.

Execution: We can calculate **range, variance, and standard deviation** for numerical columns

Inference:

If ticket prices have a **high standard deviation**, it means fares fluctuate a lot (likely due to different airlines, classes, and booking times).

Days left may show smaller **dispersion** if most people book close to travel dates.

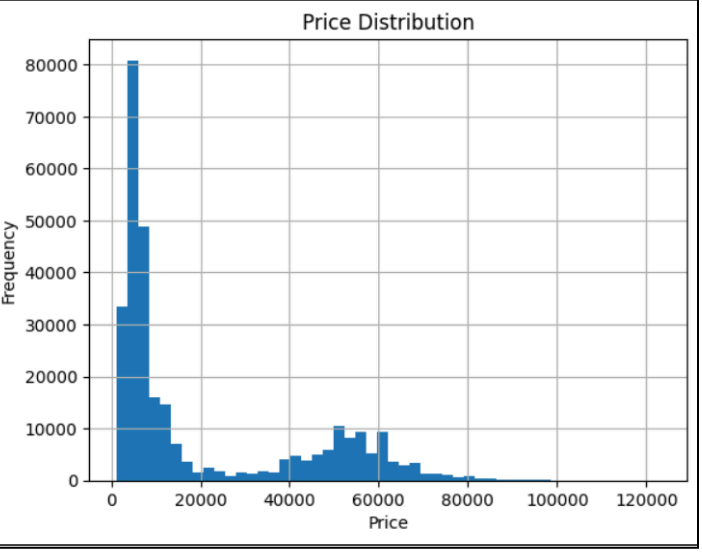
Understanding dispersion helps us see whether the data points are tightly packed around the mean or widely scattered.

Data Visualization:

Definition: Data visualization represents data graphically, making it easier to interpret patterns, trends, and outliers. It uses charts like histograms, box plots, scatter plots, bar charts, and heatmaps.

1) **Histogram** – For price distribution.
   Histogram shows how prices are distributed (skewed toward low/mid prices or spread evenly).

```
# Histogram – For price distribution.
df['price'].hist(bins=50)
plt.title('Price Distribution')
plt.xlabel('Price')
plt.ylabel('Frequency')
plt.show()
```



2) **Box Plot** – To detect outliers in price

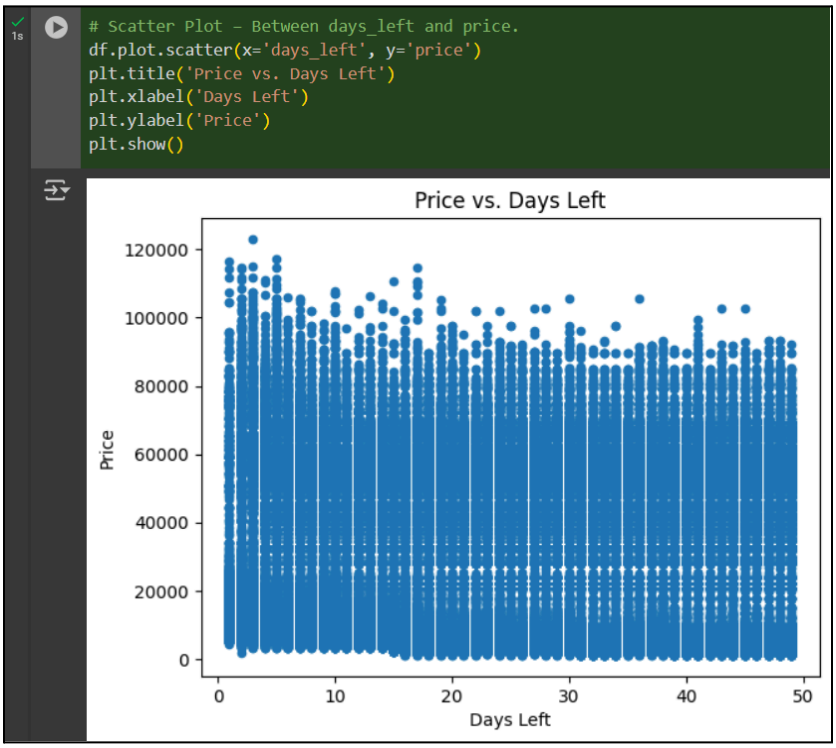
Box Plot highlights price differences between Economy and Business classes.

```
# Box Plot – To detect outliers in price.
df.boxplot(column='price', by='class')
plt.title('Price Distribution by Class')
plt.xlabel('Class')
plt.ylabel('Price')
plt.show()
```

3) **Scatter Plot**: Between days_left and price

Scatter Plot will show the relationship between booking time and price (last-minute tickets may be more expensive).

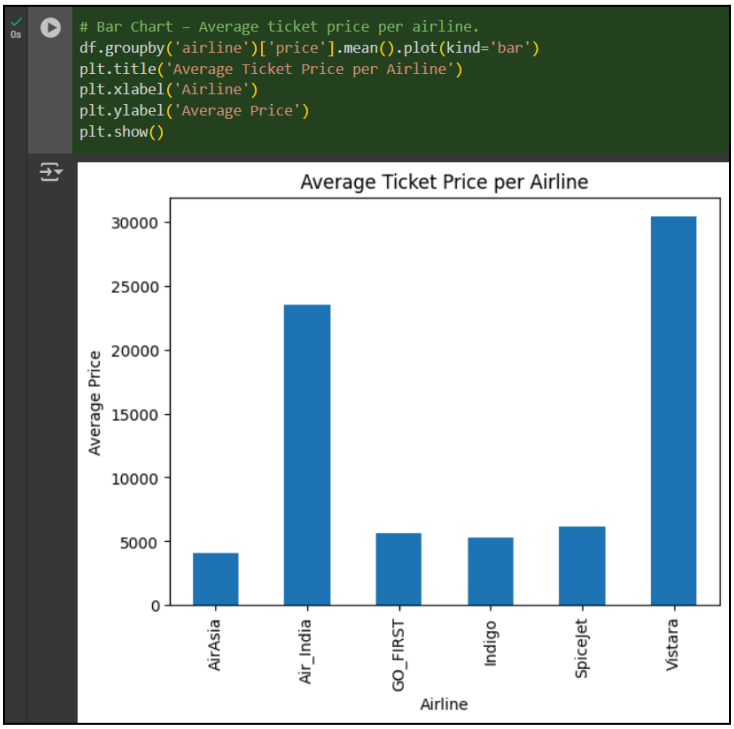
```
# Scatter Plot – Between days_left and price.
df.plot.scatter(x='days_left', y='price')
plt.title('Price vs. Days Left')
plt.xlabel('Days Left')
plt.ylabel('Price')
plt.show()
```



4) **Bar Chart** – Average ticket price per airline.

Bar Chart compares average pricing strategies of different airlines.

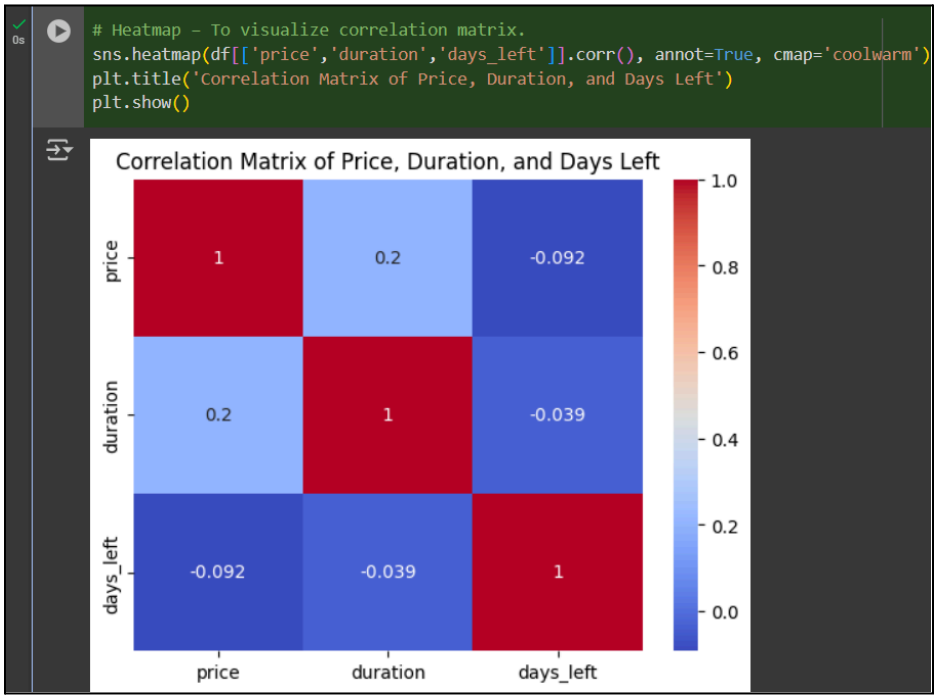
```
# Bar Chart – Average ticket price per airline.
df.groupby('airline')['price'].mean().plot(kind='bar')
plt.title('Average Ticket Price per Airline')
plt.xlabel('Airline')
plt.ylabel('Average Price')
plt.show()
```



5) **Heatmap** – To visualize correlation matrix.

Heatmap provides an overview of relationships among numerical variables.

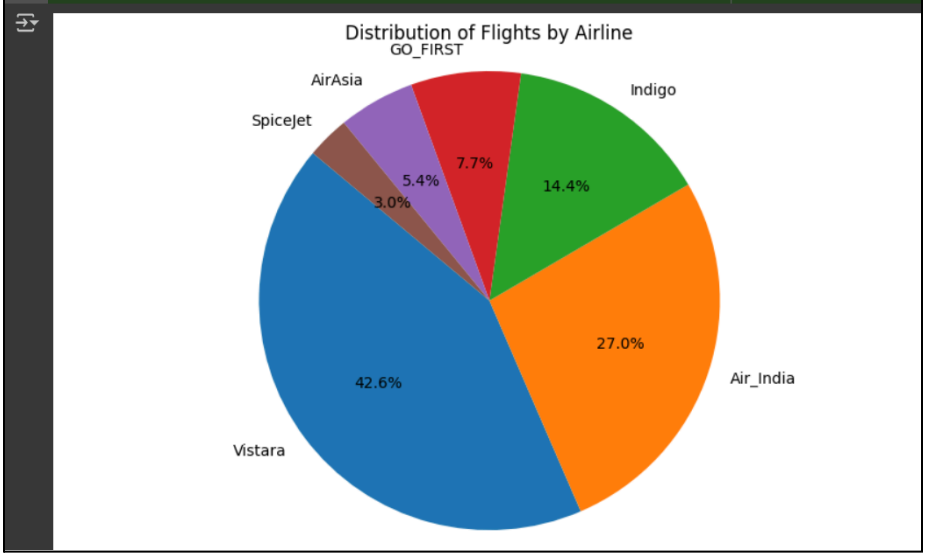
```
# Heatmap – To visualize correlation matrix.
sns.heatmap(df[['price','duration','days_left']].corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Matrix of Price, Duration, and Days Left')
plt.show()
```



6) **Pie-Chart:** The pie chart represents the proportion of flights operated by each airline in the dataset.
From the chart, we can identify which airlines dominate the flight market share and compare their relative contribution to the total number of flights.

```
# Pie Chart – Distribution of flights by airline.
import matplotlib.pyplot as plt

airline_counts = df['airline'].value_counts()
plt.figure(figsize=(10, 6))
plt.pie(airline_counts, labels=airline_counts.index, autopct='%1.1f%%', startangle=140)
plt.title('Distribution of Flights by Airline')
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.
plt.show()
```



Conclusion: Through this experiment, we successfully performed exploratory data analysis (EDA) and visualization on the airline flights dataset. By applying descriptive statistical methods such as central tendency (mean, median, mode) and dispersion (range, variance, standard deviation), we were able to summarize the overall patterns in ticket prices, flight durations, and booking timelines. Correlation analysis helped us understand relationships between variables, such as the effect of days left before departure on flight prices. Furthermore, visualizations like histograms, box plots, scatter plots, bar charts, pie charts, and heatmaps provided clear and intuitive insights into the dataset.