King Saud University
College of Computer and Information Sciences
Department of Information Technology

IT362 – First semester 1446

Phase 2 Report

Section	NAME	ID
56594	Najd Alsabi	443200578
	Leena Almusharraf	443200563
	Hissah Alhano	443200617
	Joud Almutairi	443200544
	Lujain Alharbi	443200811
	Asma Alshilash	443200439

Exploratory Data Analysis (EDA):

This exploratory data analysis (EDA) focuses on understanding patterns in flight-related data, specifically exploring the relationships between flight duration, price, and airline companies. The goal of this analysis is to identify correlations that may exist between these variables and derive insights that could help us understand how various factors influence flight prices and durations. This section provides a detailed description of the EDA process, including data preprocessing, statistical analysis, and visualization techniques.

1. Primary Data

Number of Flights by Airline:

A histogram was created to analyze the distribution of flights across different airlines. The results showed that "Saudia" is the most frequently used airline by a significant margin, followed by "Flynas". (Figure 1)

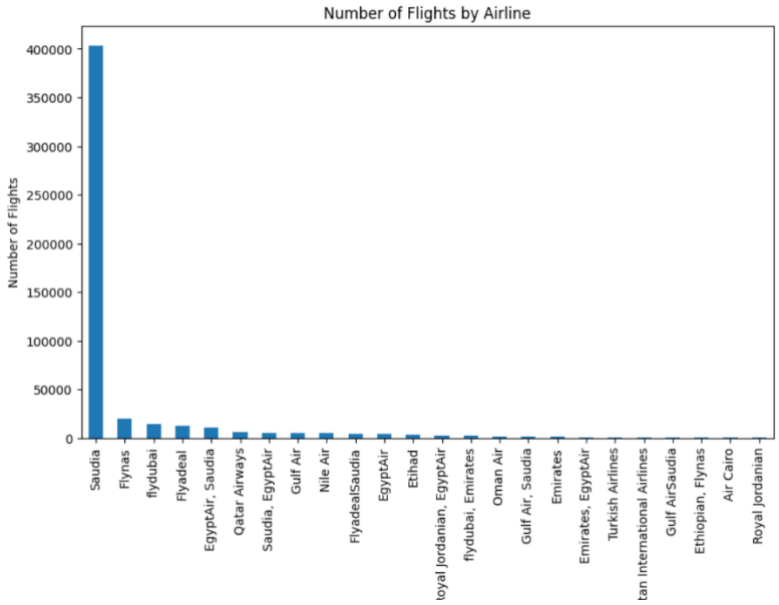

Figure 1

Flight Duration Distribution:

A histogram of the flight durations was generated to examine the distribution. The plot revealed that most flights have a duration between 50 and 250 minutes, and the most common is 1 hr 50 min, with a smaller number of outlier flights having longer durations. The data appeared slightly right-skewed, indicating the presence of a few long-haul flights that last significantly longer than the majority. (Figure 2)

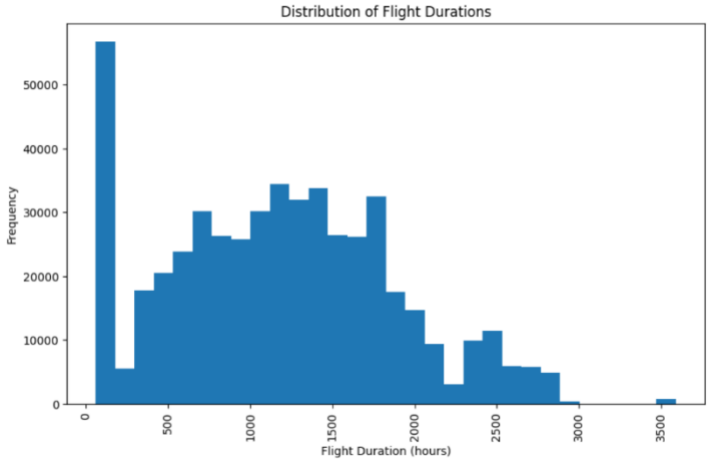

Figure 2

Price Distribution:

Similarly, a histogram of flight prices was created, showing that the majority of flights are priced between 500 SAR and 1500 SAR with the most frequent being 1018 SAR. The price distribution was also right-skewed, with a few flights priced significantly higher. This distribution suggests that while most flights are moderately priced, there are premium flights with much higher costs, likely reflecting additional factors such as time of booking, or demand. (Figure 3)

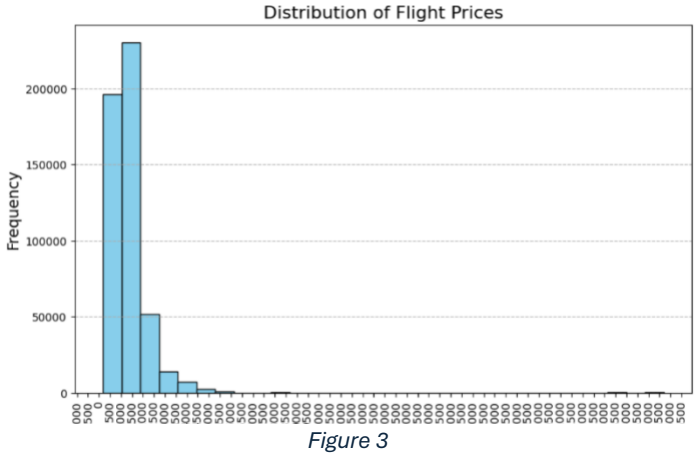

Figure 3

Number of Flights by Departure Date:

A bar chart was created to visualize the number of flights based on the departure date. The chart revealed a clear pattern in flight scheduling over time with the highest frequency being between October and December 2024. This suggests that airlines schedule more flights during periods of high demand, which aligns with the assumption that holiday periods see greater travel volumes. (Figure 4)

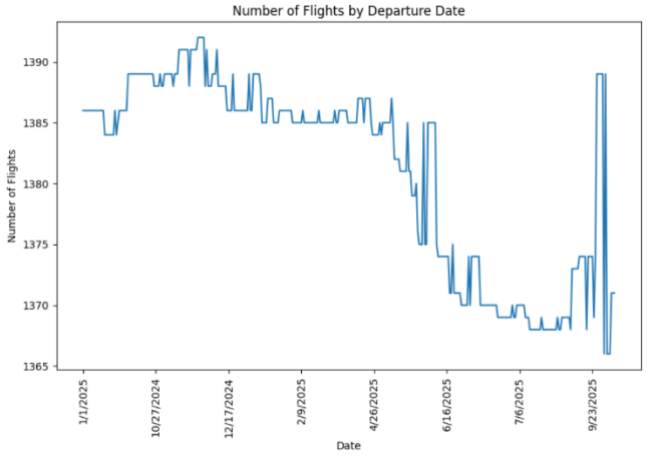

Figure 4

Most Common Departure Times:

An analysis of the most common departure times revealed distinct patterns, with peak times occurring during the early morning (9 AM) and early evening (4 PM - 9 PM). These times correspond to the periods when travelers are likely heading to work or returning home, or when business travelers are catching flights for early meetings. (Figure 5)

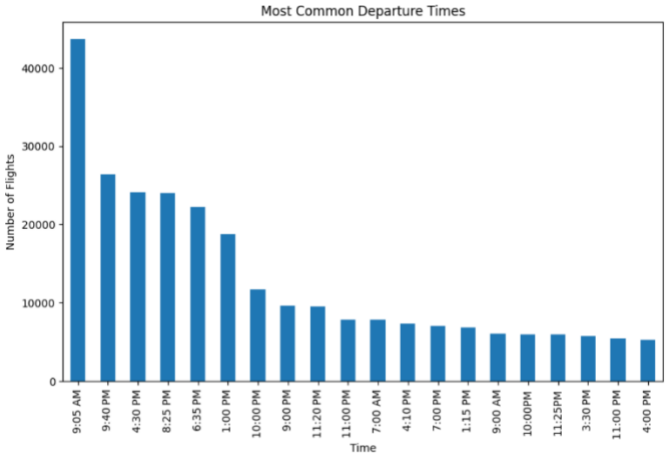

Figure 5

Most Common Arrival Times:

Similar to departure times, the most common arrival times tended to cluster around early morning and evening hours. The distribution of arrival times reflects the common practice of flights being scheduled for early morning or evening landings, allowing travelers to maximize their days. (Figure 6)

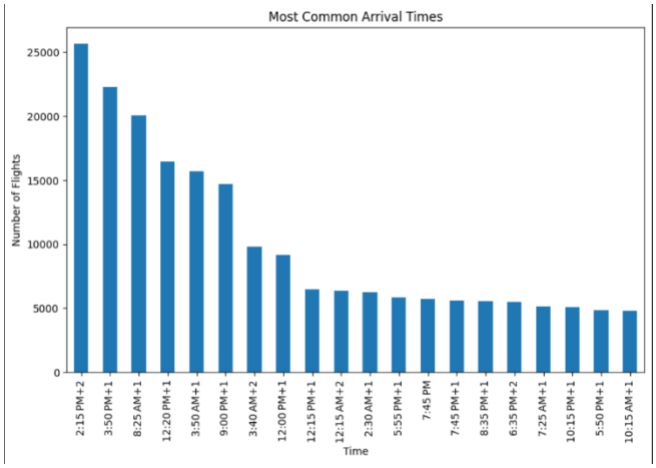

Figure 6

Number of Flights by Stops:

Flights were categorized based on the number of stops (non-stop, one-stop, two-stop flights, etc.). The majority of flights in the dataset were 2 stops. (Figure 7)

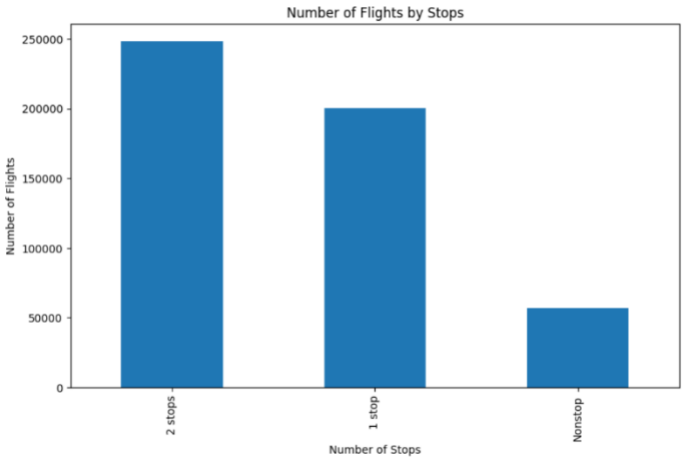

Figure 7

Flights by Departure City:

The distribution of flights across departure cities was visualized using a bar chart. Larger cities such as RUH, EAM, and JED naturally had the highest number of departing flights. (Figure 8)

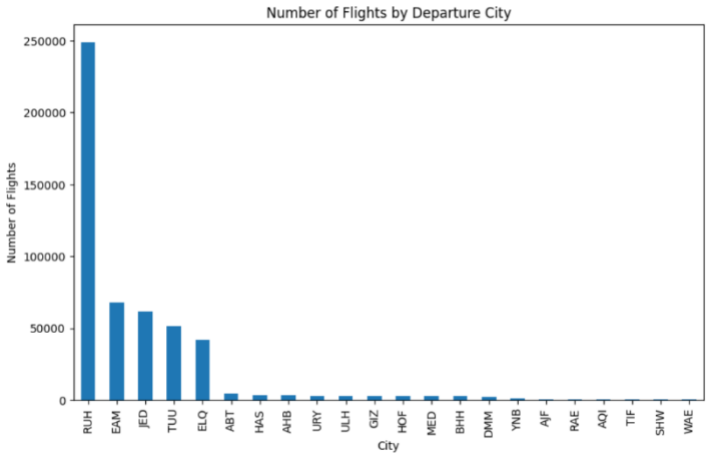

Figure 8

Flights by Arrival City:

Similarly, the number of flights by arrival city followed a similar pattern, with RUH receiving the bulk of flights. This is consistent with the notion that large areas attract more traffic due to both business and leisure travel. (Figure 9)

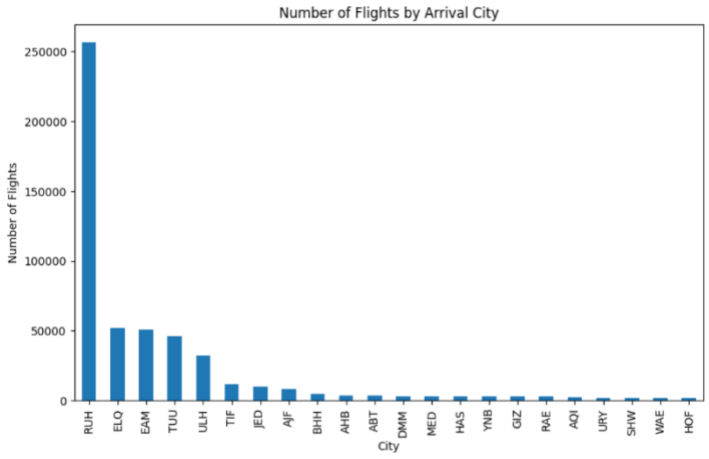

Figure 9

Cross-tabulation between Departure City and Arrival City:

A cross-tabulation was performed to examine the relationship between departure cities and arrival cities. The result highlighted specific patterns in city pairs, indicating the frequency with which flights connect particular cities. This type of analysis helps identify major flight routes between the destinations. The heatmap shows a dark red cell at the intersection of RUH and EAM, indicating a substantial number of flights on this route. (Figure 10)

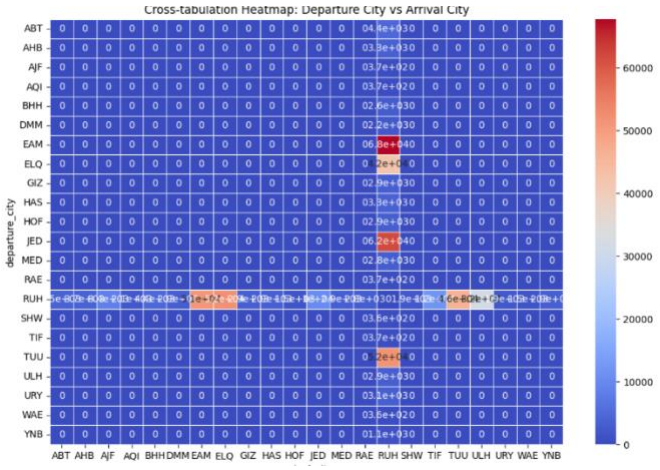

Figure 10

Correlation Analysis:

To further explore the relationships between variables, we conducted a correlation analysis.

Before conducting correlation analysis, the categorical variables (such as airline companies and city names) were encoded numerically. This process, known as factorization, ensures that these categorical data points can be used in numerical calculations. Each company or city was assigned a unique numeric code to facilitate the correlation analysis.

Correlation Matrix and Heatmap:

A correlation matrix was computed to examine the relationships between flight duration, price, and airline company. The following key observations were made:

Flight Duration and Price:

A weak positive correlation (0.073) was observed between flight duration and price. This suggests that, on average, longer flights tend to be slightly more expensive, though the relationship is weak. Other factors, such as airline pricing strategies or market competition, may play a more significant role in determining flight prices. (Figure 11)

Price and Airline Company:

A moderate positive correlation (0.38) was identified between flight price and airline company. This indicates that certain airlines tend to charge more for their flights, which may reflect differences in brand value, or operating costs. The analysis suggests that the airline is a more important determinant of price than flight duration. (Figure 11)

Flight Duration and Airline Company:

A weak negative correlation (-0.11) was found between flight duration and airline company. This implies that the duration of a flight is not strongly tied to the specific airline, indicating that airlines operate flights of varying durations without any specific pattern.

The correlation heatmap provided a visual representation of these relationships, highlighting the weak associations between the variables. (Figure 11)

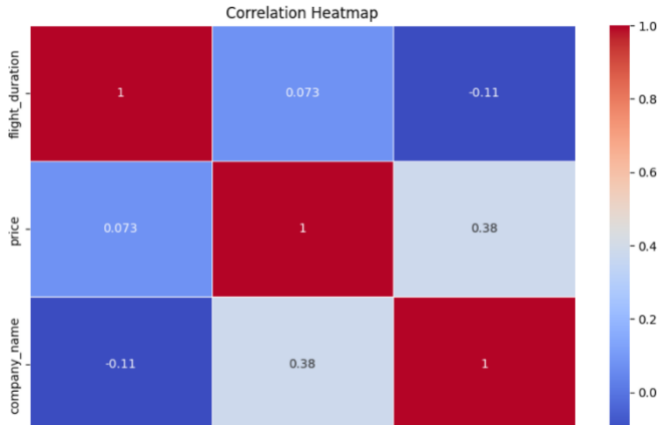

Figure 11

2. Secondary Data

Tools and Libraries Used:

For this analysis, the following Python libraries were employed to ensure efficient data processing, visualization, and statistical computations:

- **pandas**: Utilized for data manipulation, cleaning, and transformation.
- **matplotlib and seaborn:** Employed for creating visualizations that help in understanding the distributions and correlations within the dataset.
- **numpy:** Used for performing numerical operations, including calculations and data transformations.

These libraries are standard tools for conducting data analysis, offering a wide range of functions that support various stages of the analytical process.

Data Cleaning and Preprocessing:

The currency was converted from US dollars $ to SAR In order to facilitate comparisons with our primary dataset.

Dataset shape:

Our dataset contains 317,260 rows, which represent individual records, and 11 columns, which represent different features or attributes of the data.

Exploratory Data Analysis (EDA):

The primary objective of the EDA phase was to explore the distribution of key variables and the relationships between them.

Price distribution:

We plotted the distribution of flight prices using a histogram displaying the frequency of flight prices across specified bins. This visualization further emphasizes the distribution of prices, showing how many flights fall within different price ranges. Both plots serve to highlight the variations in flight prices within the dataset, aiding in understanding pricing patterns. (Figure 12)

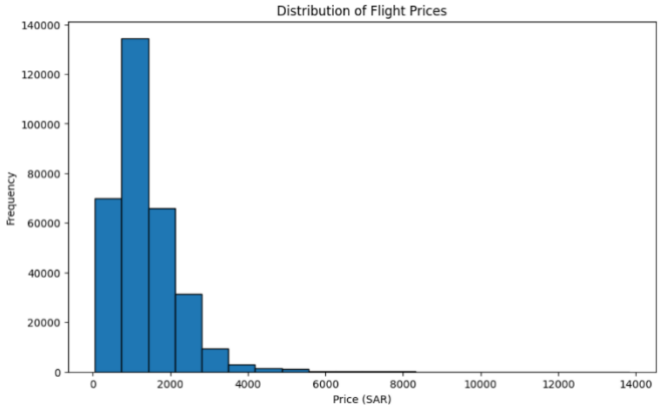

Figure 12

Number of stops distribution:

We visualized the distribution of flights using a bar chart based on the number of stops. It has helped to understand how many flights operate with zero, one, two, or more stops, providing insights into flight options available to travelers. The most frequent is flights with 1 stop. (Figure 13)

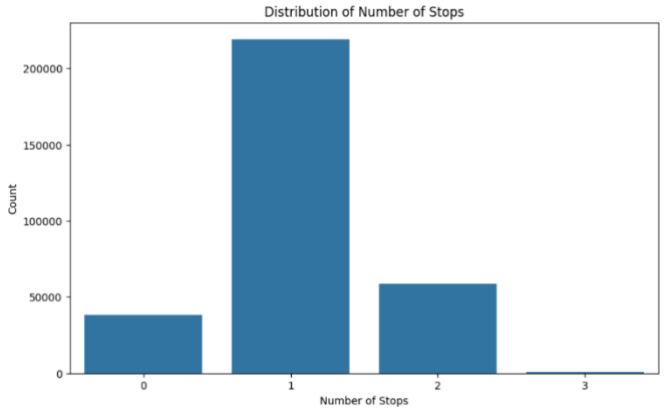

Figure 13

Prices trend over time:

We implemented a line plot depicts the trend of flight prices over time based on the departure date. It helps identify any fluctuations or patterns in pricing, allowing for a better understanding of how prices vary with departure dates. Notably, prices between July and August 2024 were the highest, indicating a peak in demand during that period. (Figure 14)

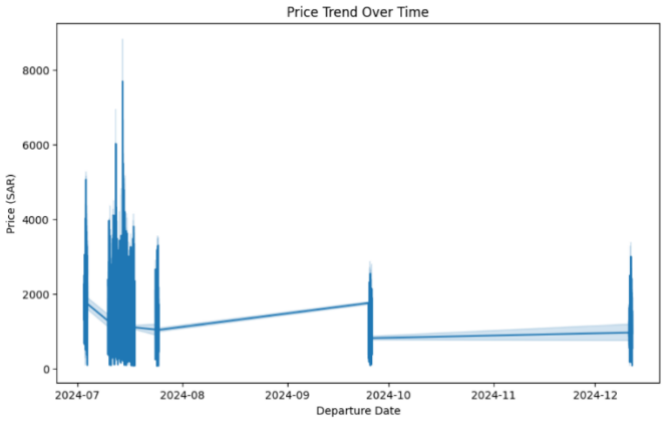

Figure 14

Flight lands next day graph:

We implemented a bar chart to illustrate the frequency of flights landing the next day, with '1' indicating a yes and '0' a no. The data reveals that most flights do not land the next day, highlighting the distribution of overnight travel in the dataset. (Figure 15)

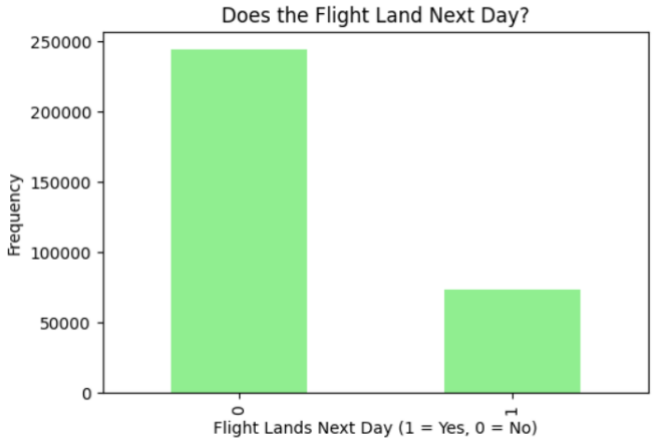

Figure 15

Airline distribution:

We implemented a bar chart illustrates the distribution of airlines in the dataset, The results showed that "United Airlines" is the most frequently used airline by a significant margin, followed by "Alaska Airlines". (Figure 16)

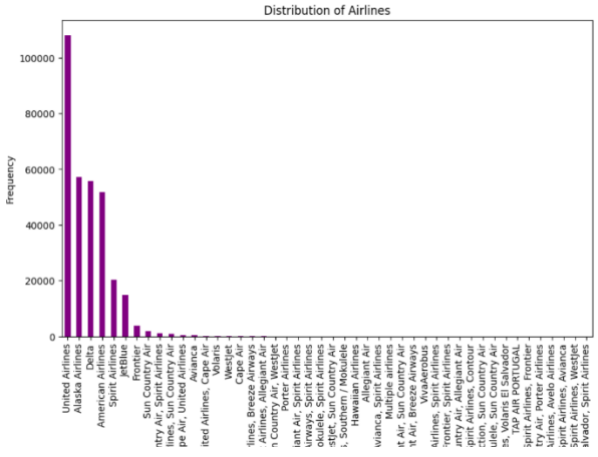

Figure 16

Correlation Analysis:

To further explore the relationships between variables, we conducted a correlation analysis.

Before conducting correlation analysis, the categorical variables (such as airline companies and city names) were encoded numerically. This process, known as factorization, ensures that these categorical data points can be used in numerical calculations. Each company or city was assigned a unique numeric code to facilitate the correlation analysis.

Number of stops and Price:

A moderate positive correlation (0.36) was identified between the number of stops and airline company. indicating that flights with more stops tend to be more expensive. (Figure 17)

Price and Airline Company:

A weak positive correlation (0.23) between airline company and price indicates that prices may vary slightly based on the airline. However, this relationship is weak, implying that other factors, such as service quality and operational costs, likely have a more significant impact on flight pricing. (Figure 17)

Number of stops and Airline Company:

A correlation analysis value of 0.23 between the number of stops and airline company indicates a weak positive correlation. This suggests that there is a slight tendency for the number of stops to vary across different airlines. However, this weak relationship implies that the number of stops is not a strong predictor of which airline is operating a flight, indicating that airlines have diverse routing practices that are influenced by factors other than the airline itself. The correlation heatmap visually illustrates these weak associations between the variables. (Figure 17)

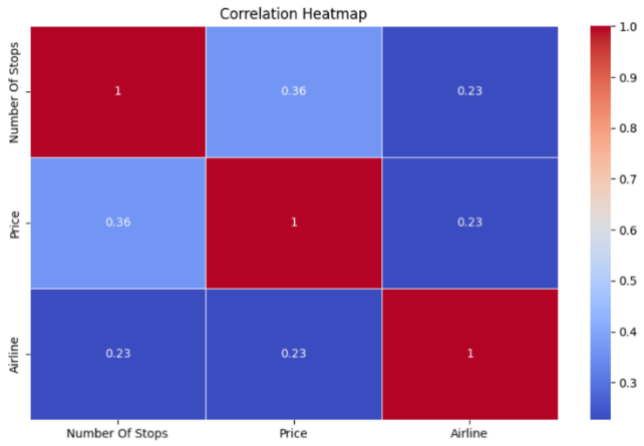

Figure 17

Flight Lands Next Day VS Price:

We implemented a scatter to visualize the relationship between whether a flight lands the next day (binary variable) and its price. Showing if overnight flights tend to be priced differently compared to same-day arrivals. The distribution shows that there are price variations for both scenarios (landing the same day or the next day), but it seems that flights that land the next day may have slightly higher prices concentrated in certain ranges, suggesting a potential impact of overnight flights on pricing. (Figure 18)

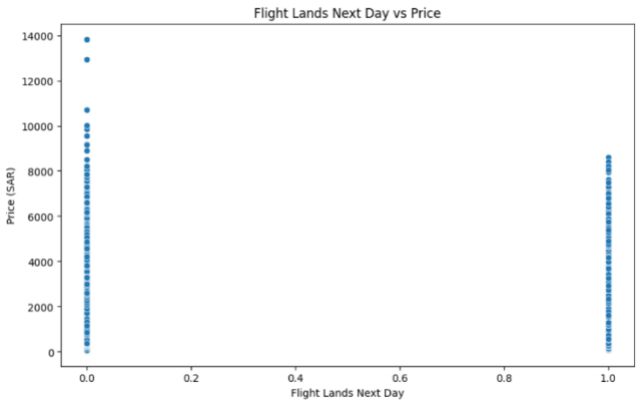

Figure 18

Identifying Outliers:

We implemented a box plot to visually identify outliers in the numerical columns, particularly focusing on the "Price" variable. The boxplot effectively highlights the distribution of flight prices, indicating the presence of several significant outliers above the upper whisker of the plot. Most prices are concentrated around a lower range, with a long tail of high-value prices representing outliers.

The thick box in the plot shows the interquartile range (IQR), which encompasses the middle 50% of the data, while the whiskers extend to 1.5 times the IQR. Any points outside of this range are plotted as individual circles and are considered potential outliers. In this case, flight prices beyond approximately 2,500 units are clearly outliers, with some flights reaching prices as high as 14,000. (Figure 19)

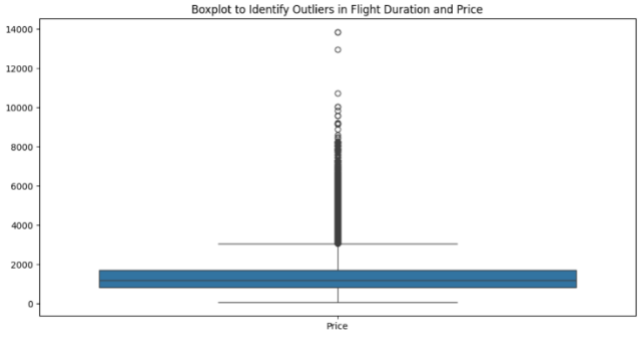

Figure 19

Comparing The Primary And Secondary Datasets

Summary Statistics:

In this analysis, we transformed the Price column in the secondary dataset from Dollars to Riyals (SAR) to align it with the price in the primary dataset. This transformation standardized the data, enabling a direct comparison of key price metrics such as mean, median, and standard deviation across both datasets. This ensures a more accurate evaluation of flight pricing trends between the two datasets.

Price Statistics	Primary dataset	Secondary dataset
Mean	1364.71	1264.23
Median	1191.0	1215.0
Standard deviation	1209.28	574.21
Min	179.0	56.25
Max	25688.0	6937.5

Comparison of Key Frequencies:

Frequency of Airlines:

- Primary Dataset:

In the primary dataset, the airline with the highest frequency of flights is **Saudia** with a total of **331,879 flights**. This indicates that Saudia is the dominant airline in this dataset, likely reflecting its significant presence and network within the region covered by this dataset.

- Secondary Dataset:

in contrast, for the secondary dataset, the airline with the highest frequency is **United Airlines**, operating **76,376** flights. This shows that United Airlines has a major presence in the secondary dataset.

Metrics from each data source side by side:

	Primary Flight Duration	Secondary Flight Duration	Primary Price	Secondary Price	Primary Number Of Stops	Secondary Number Of Stops	Primary Flight Lands Next Day	Secondary Flight Lands Next Day
count	425192.0	59560.0	425192.0	59560.0	425192.0	59560.0	425192.0	59560.0
mean	1099.73..	667.4475..	1364.71...	1272.31..	1.26..	1.28..	0.370..	0.299..
std	697.39..	243.84..	1209.28..	592.14..	0.67..	0.53..	0.48..	0.45..
min	60.0	1.0	179.0	56.25	0.0	0.0	0.0	0.0
25%	555.0	484.0	841.0	855.0	1.0	1.0	0.0	0.0
50%	1060.0	660.0	1191.0	1226.25	1.0	1.0	0.0	0.0
75%	1570.0	836.0	1663.0	1642.5	2.0	2.0	1.0	1.0
max	3590.0	1437.0	25688.0	6937.5	2.0	3.0	1.0	1.0

Frequency of Departure Cities:

- Primary Dataset:

In the primary dataset, the city with the most frequent departures is **Riyadh**, with a total of **207,320 flights**. This emphasizes Riyadh's significance as a major departure point for flights within the region and to international destinations.

- Secondary Dataset:

In the secondary dataset, **Chicago O'Hare (ORD)** takes the lead as the most frequent departure city, with **26,700 flights**. This reflects Chicago O'Hare's central role in the U.S. aviation network as a major departure hub for both domestic and international flights.

Frequency of Departure Dates (by Month):

- Primary Dataset: When looking at the frequency of departures by month in the primary dataset, the month with the most flights is **November 2024**. This suggests a potential peak in travel demand during this period, which could be linked to holidays, business events, or seasonal trends in the region.

- Secondary Dataset: In the secondary dataset, the month with the highest frequency of flights is **July 2024**. This aligns with the summer travel season, particularly in the Northern Hemisphere, when vacation travel and family trips tend to peak.

Comparison of Flight Price Statistics Between Primary and Secondary Datasets:

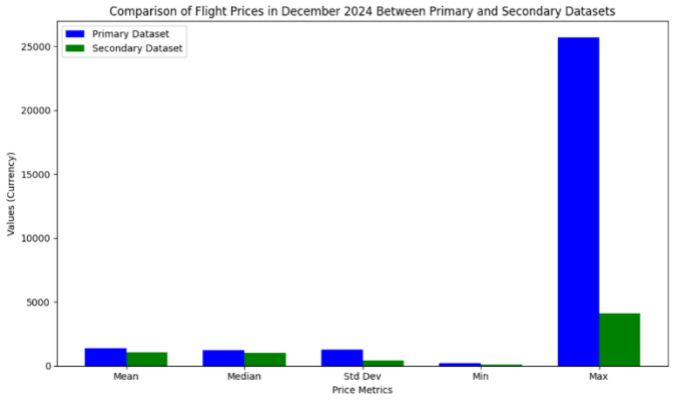

Figure 20

The bar chart compares flight prices from primary and secondary datasets in December 2024 across various price metrics: mean, median, standard deviation, minimum, and maximum. The mean, median, and standard deviation values are fairly similar between the two datasets, with the primary dataset being slightly higher. The minimum prices are almost identical, but the maximum price shows a huge difference—over 25,000 in the primary dataset versus around 5,000 in the secondary. This could be due to seasonal variations, or last-minute flights.

Flight Comparison On 15 September 2024 Between Primary and Secondary Datasets:

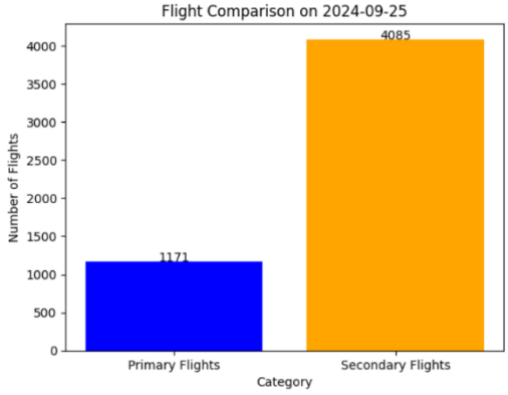

Figure 21

This bar chart compares the number of flights between the primary and secondary datasets on September 25, 2024. The primary dataset shows significantly fewer flights (1,171) compared to the secondary dataset (4,085). One reason for the lower number of flights in the primary dataset could be the start of the school season in Saudi Arabia, where the primary data is sourced. As September marks the beginning of the academic year, there may be reduced demand for flights, leading to fewer scheduled flights during this period. In contrast, the secondary dataset, which is from the U.S., shows a much higher number of flights. This difference could be due to varying seasonal travelling patterns in the U.S.

Comparison of Daily Flight Count in December 2024 Between Primary and Secondary Datasets:

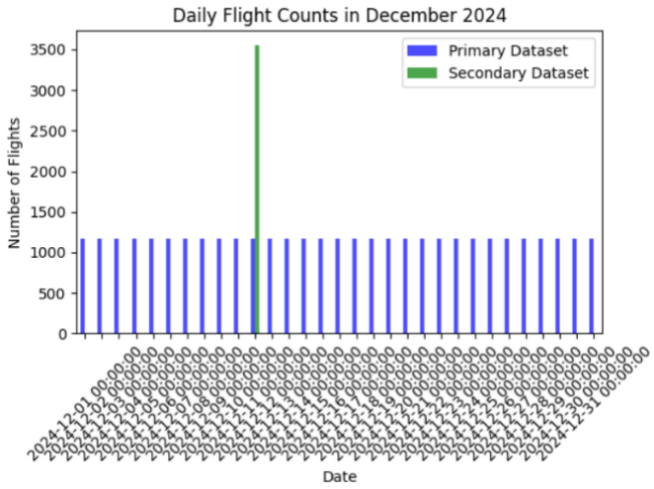

Figure 22

This bar chart displays the daily flight counts in December 2024 for both the primary (blue) and secondary (green) datasets. The primary dataset shows consistent daily flight numbers, hovering around 1,200 flights per day throughout the month. In contrast, the secondary dataset shows a notable spike on a single day, reaching over 3,500 flights, while maintaining a lower count or no activity on other days.

The spike in the secondary dataset is likely due to the Christmas season in the U.S., when travel demand increases sharply for the holidays. Meanwhile, in Saudi Arabia, December marks the final exam season, which likely contributes to steady, lower flight counts as students focus on academics. Since Saudi Arabia does not celebrate Christmas, the cultural and academic factors result in more uniform travel patterns throughout the month.

Comparison of Flight Duration in December 2024 Between Primary and Secondary Datasets:

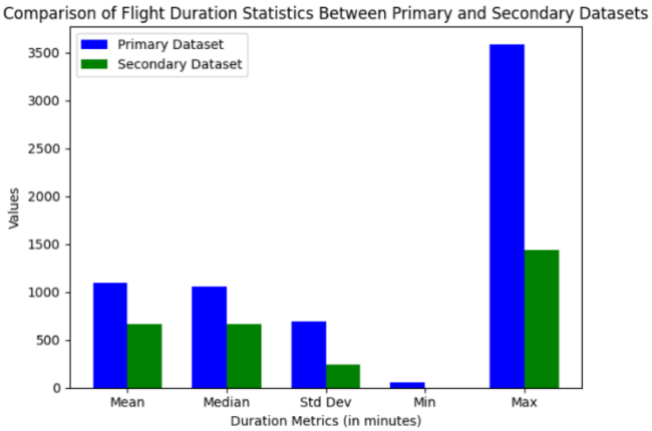

Figure 24

The bar chart presents a comparative analysis of flight duration statistics between the primary and the secondary dataset. We can see that the primary dataset has significantly longer flight durations across various metrics, including a higher mean (1000+ minutes vs. 500~ minutes), median (1000~ minutes vs. 600~ minutes), standard deviation (600~ minutes vs. 300~ minutes), minimum (100~ minutes vs. 0 minutes), and maximum (3,500 minutes vs. 1,500~ minutes) compared to the secondary dataset. These differences suggest that the primary dataset contains flights with generally longer and more variable durations compared to the secondary dataset.

Contextualize Findings :

-Price:

The mean and median prices in both datasets are quite similar, with the primary dataset being slightly higher. This indicates a close alignment in average flight prices, providing a consistent view of price trends across both sources. Despite minor differences, the overall pricing patterns in each dataset reinforce a similar understanding of the market.a consistent understanding of price trends across the two sources.

-**Number of stops:**

Price increases with the number of stops in both datasets, but the primary data shows higher prices overall.

The secondary dataset includes flights with more than 2 stops, which shows a decrease in average price for 3-stop flights, possibly due to longer travel times or less desirable routes.

- Airline:

Both datasets feature a dominant airline (Saudia in the primary and United Airlines in the secondary) with a substantial lead over other airlines. This suggests a similar pattern of concentration where a few airlines account for a large portion of flights.

In both cases, there is a steep drop-off in frequency after the leading airline, indicating that most other airlines operate far fewer flights. This distribution is typical in markets where a few major carriers dominate.

Contradictions and Explanations:

Regional Focus Differences: The primary dataset seems more focused on airlines prominent in the Middle East and regional markets, while the secondary dataset is centered on airlines operating in the U.S. and neighboring regions. This difference in regional focus could explain the different leading airlines in each dataset.

-Time Frame:

- **Seasonality and Price Variability:**

The primary dataset, with its broader and more evenly distributed time frame, is likely to include both peak and off-peak pricing, leading to higher price variability (as seen with a larger standard deviation).

The secondary dataset, emphasizing dates during the summer peak (July) and winter holidays (December), may have fewer low-price records since these periods typically feature higher demand, which could lead to elevated average prices during those times.

- **Impact on Price Ranges:**

The emphasis on high-demand periods in the secondary dataset could explain why its average prices align with the primary dataset but have lower variability. By focusing on peak travel

times, the secondary dataset might lack the low-cost fares often found in off-peak seasons, which are more likely present in the primary dataset.

Contradictions Explained:

Differences in price variability and flight frequencies can be attributed to these time frame differences. The secondary dataset's focus on peak periods explains its lower price variability and higher concentration of flights on specific days, while the primary dataset’s broader scope leads to a wider price range and more evenly distributed flight counts.

Summary of New Insights and Hypotheses

The Exploratory Data Analysis (EDA) of flight data has uncovered significant insights into pricing, flight durations, and airline dominance, revealing complex market dynamics. The primary dataset exhibits a strong prevalence of "Saudia," whereas "United Airlines" emerges as the leader in the secondary dataset, indicating regional variations in airline dominance that reflect geographic market trends. This disparity leads to the hypothesis that the leading airlines may possess more established routes or lower operational costs in their respective areas, contributing to their frequency of flights.

Additionally, while there is a weak positive correlation between flight duration and pricing in both datasets, it suggests that flight prices are more significantly influenced by factors such as demand, competition, and airline pricing strategies rather than duration alone. Notably, flights with a higher number of stops tend to be more expensive, particularly in the secondary dataset, which shows a moderate correlation (0.36) between the number of stops and price. This observation supports the hypothesis that increased stops might reflect higher operational costs or cater to travelers with fewer direct options.

Price trends reveal distinct peaks during high-demand periods, such as July 2024 for the secondary dataset and November 2024 for the primary dataset, suggesting that seasonal travel demand or specific events may drive these fluctuations. Consequently, it can be hypothesized that travelers booking flights in advance or during off-peak times could realize significant cost savings.

Moreover, both datasets display peak departure and arrival times during early mornings and evenings, aligning with typical business and leisure travel patterns. This scheduling may be a strategic choice by airlines to accommodate both business travelers and those aiming to maximize their travel utility.

The analysis also identified notable outliers in flight prices, particularly in the primary dataset, with prices reaching as high as 25,688 SAR. These outliers likely represent last-minute bookings, premium services, or routes with elevated operational costs.

Lastly, despite similar price distributions between the primary and secondary datasets, the primary dataset shows significantly higher maximum prices, indicating different regional market dynamics that may lead to greater variability in pricing, particularly when comparing the Saudi Arabia to the U.S. These insights and hypotheses set the stage for further exploration of the factors influencing flight prices, durations, and operational strategies, paving the way for more targeted analyses or machine learning applications.