Analysis of Correlation between Airfare Price Dynamics in Relation to Crude Oil Price Fluctuations

1. Introduction

1.1 Project Overview

The primary objective of this project is to develop a robust predictive model that accurately forecasts changes in airfare prices as a response to fluctuations in crude oil prices. This analysis aims to assist the analysis of airline companies in strategizing fare adjustments and provide consumers with insights into potential future airfare trends.

1.2 Research Questions

Final goal is to understand:

1. The impact of oil prices on commercial aviation, operating costs and profitability
2. How do airlines manage oil price volatility and its financial impact?

1.3 Significance of Analysis

The success of the airline industry, particularly in terms of pricing strategy, is closely tied to external economic factors, with crude oil prices being a primary influencer. Through the "Predictive Analysis of Airfare Price Dynamics in Relation to Crude Oil Price Fluctuations," we aim to gain critical insights into how changes in crude oil prices impact airfare costs. This analysis is vital for several key stakeholders in the aviation sector, as it informs strategic decision-making in various aspects of the industry:

Investors and Analysts: Investors in the airline industry and market analysts can leverage the insights from this predictive analysis to gauge the financial health and prospects of airlines, aiding in investment decisions and market predictions.

Airlines and Pricing Managers: Understanding the correlation between crude oil prices and airfare enables airlines and pricing managers to make informed decisions on fare adjustments. This insight helps in optimizing profitability while remaining competitive in the market.

Operations and Financial Planners: For those involved in operations and financial planning within airlines, this analysis provides a framework to anticipate cost changes and budget accordingly. It allows for more effective risk management and financial forecasting.

Policy Makers and Regulatory Bodies: Regulators and policy makers can use the findings from this analysis to understand the economic pressures on the aviation industry. This understanding is crucial for creating fair and effective policies and regulations.

Consumers and Travel Agencies: By understanding the trends in airfare in relation to crude oil prices, consumers and travel agencies can make more informed decisions about travel planning and cost management.

Ultimately, the goal of our analysis is to provide actionable insights that not only guide airlines in pricing strategies under fluctuating crude oil prices but also help in shaping the overall resilience and responsiveness of the aviation industry to external economic factors. This analysis is a step towards enhancing the economic efficiency and market adaptability of the airline sector.

2. Data Sources and Datasets

We used both primary and secondary data for our analysis.

2.1 Primary Data Source:

Airfare trend in relation to crude oil Prices
Source: Google Forms.
Purpose of the Survey: The survey aims to understand how fluctuations in airfare and crude oil prices influence consumer travel decisions. Key objectives include assessing the frequency of air travel, examining sensitivity to airfare changes, understanding consumer awareness of the economic factors affecting airfares, and exploring responses to airfare increases. This information will help airlines and industry stakeholders make informed decisions regarding pricing strategies and customer engagement.
Data Collection Process: The survey was distributed through various social media channels and email bulletins to engage a wide-ranging audience. The response collection spanned four weeks to guarantee a substantial sample size for broad representation.

2.2 Secondary Data Sources:

Ease my trip (Indian website for flight bookings) : Indian Airline Price Prediction

- **Source:** A publicly available on make my trip website
- **Purpose:** The objective of the study is to analyze the flight booking dataset obtained from "Ease My Trip" website and to conduct various statistical hypothesis tests to get meaningful information from it.
- **Variables:** Dataset contains information about flight booking options from the website Easemytrip for flight travel between India's top 6 metro cities. There are 300261 data points and 11 features in the cleaned dataset.

MacroTrade: Oil price worldwide
- **Source**: weekly oil prices in Brent, OPEC basket and WTI futures 2020-2023

- **Purpose:** Used to understand the change in oil prices to understand the effect on airfares.
- **Variables:** Includes detailed text reviews, star ratings, and customer names.

US Energy Administration: American Airline Price Prediction
- **Source:** An aggregated dataset about flight information.
- **Purpose:** Utilize comprehensive airfare transaction data, which includes ticket sales data from a variety of airlines.
- **Variables:** Ticket prices, Airline, dates

3. Information Quality

In the process of doing predictive analysis, we encountered and addressed a lot of data quality concerns that could have potentially compromised our findings.

3.1 Merging and joining data from multiple sources
We synthesized data from diverse origins into a unified dataset with precision, ensuring consistency and alignment across all data points, akin to a well-orchestrated symphony.

3.2 Dealing with outliers
Outliers were scrutinized and managed effectively, maintaining the analytical robustness of our dataset. We employed statistical methods to either integrate or exclude these anomalies, depending on their relevance to the predictive models.

3.3 Data Standardization
Standardization procedures were implemented to achieve uniformity in our dataset, enabling comparability and coherence across different data formats and scales.

3.4 Handling Missing Values
We employed sophisticated imputation techniques to address the issue of missing values, thereby preserving the dataset's completeness and enhancing the reliability of our analysis.

3.5 Currency Conversion
Recognizing the necessity of a common financial baseline, we meticulously converted all financial data into a single currency. This standardization was critical to eliminate any potential distortion due to currency fluctuation effects on our analysis.

4. Methods and Tools

For our project of Predictive Analysis, we employed a blend of advanced programming tools, data analysis software, and data collection platforms to thoroughly process and scrutinize the datasets at hand.

The selection of methods and tools was strategic to address the specific analytical needs of our research questions.

4.1 Python Jupyter Notebook

Usage: Python was primarily utilized for data cleansing, focusing on the removal of null or missing values and enhancing the overall quality of the dataset.
Scripts and Concepts: We crafted scripts for data cleaning and manipulation, involving text normalization, parsing, and transformation operations using Python libraries such as `pandas` and `numpy`. For statistical modeling and predictions, we implemented packages such as `scikit-learn`.

4.2 Tableau

Usage: We utilized Tableau for our data visualization needs, leveraging its robust and dynamic visualization capabilities to interpret our findings effectively.
Visualizations: Through Tableau, we crafted a series of compelling visual narratives, including line charts for trend analysis and scatter plots for identifying patterns between airfare prices and crude oil price changes.

4.3 Google Forms

Usage: For gathering primary data, we designed and circulated a survey focusing on consumer travel habits and their responses to airfare changes.
Data Collection: We chose google forms because of its user-friendly interface that allowed for straightforward data export into CSV and Jupyter Notebook, streamlining the subsequent data analysis process.

5. Data Wrangling Process

5.1. Data Formatting:

structured and organized raw data into a consistent and usable format. The date formats in all three datasets had to be changed in order into (YYYY-QQ) format to be able to merge it to ensure that data is standardized, clean, and ready for analysis. This step involved reading three datasets: U.S. airlines' airfare data, Indian airlines airfare data and oil prices, then reshaping and cleaning the data. For instance, dropping unnecessary columns like 'Geocoded_City1', 'Geocoded_City2', etc., from the airline dataset. A new column 'year_quarter' is created by combining 'Year' and 'quarter', which is then reordered to the first position. This is crucial for uniformity and ease of analysis later on.

5.3 Data Profiling:

Exploring and summarizing the characteristics of a dataset. It helped in understanding the data's distribution. Data profiling is essential for gaining insights into data quality and potential issues.Data profiling includes understanding the basic structure of the data, such as using **df_us_airlines.info()** to get an overview of the data types and non-null values. Identifying unique values in columns and checking for missing values are also part of this process.

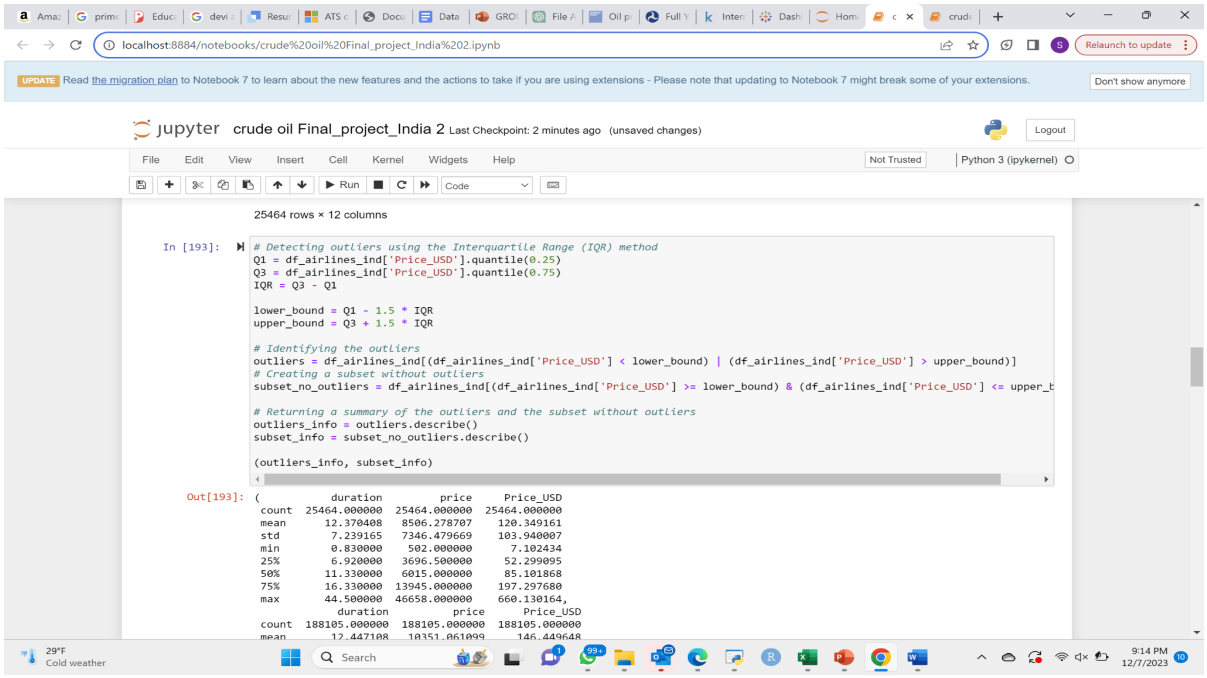
5.4 Data Preprocessing:
Cleaning and preparing the data for analysis. Steps include dropping duplicates, handling missing values, and detecting outliers using the Interquartile Range (IQR) method. For outlier detection, a function **detect_outliers** is defined and applied to the 'fare' column, helping to identify and potentially remove extreme values that might skew the analysis.

5.5 Merging Datasets:
The datasets are merged on the 'year_quarter' column. This is a critical step to bring together relevant information from both the airline and oil price datasets, enabling a comprehensive analysis of how these two variables might interact for both Indian and American airlines.

5.6 Statistical Analysis:
Statistical analysis is performed, especially in the outlier detection and handling stage. Descriptive statistics like quartiles are used to understand the distribution of 'fare' and 'average_price'. Techniques to extract insights and patterns from data. It includes tasks like hypothesis testing, regression analysis, and descriptive statistics. Statistical analysis helps in understanding relationships and making data-driven decisions.

5.7 EDA (Exploratory Data Analysis) and Visualization:
Visualization techniques, such as graphs and charts, are used to present data in a more understandable way. EDA is conducted through plotting graphs. Matplotlib and Seaborn libraries are used to create visualizations like line plots and histograms, allowing for an intuitive understanding of the data trends and distributions.The plots compare average fares and oil prices over time, providing visual insights into the relationship between these variables.

5.8 Modeling:
A Linear Regression model is built to predict airline fares based on oil prices. The dataset is split into training and test sets, and the model is trained on the training set.
Post-training, predictions are made on the test set, and the model's performance is evaluated using metrics like Mean Squared Error (MSE) and R-squared, providing insights into the accuracy and effectiveness of the model.



localhost:8884/notebooks/crude%20oil%20Final_project_India%202.ipynb

Relaunch to update

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.

Don't show anymore

jupyter crude oil Final_project_India 2 Last Checkpoint: 2 minutes ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

25464 rows × 12 columns

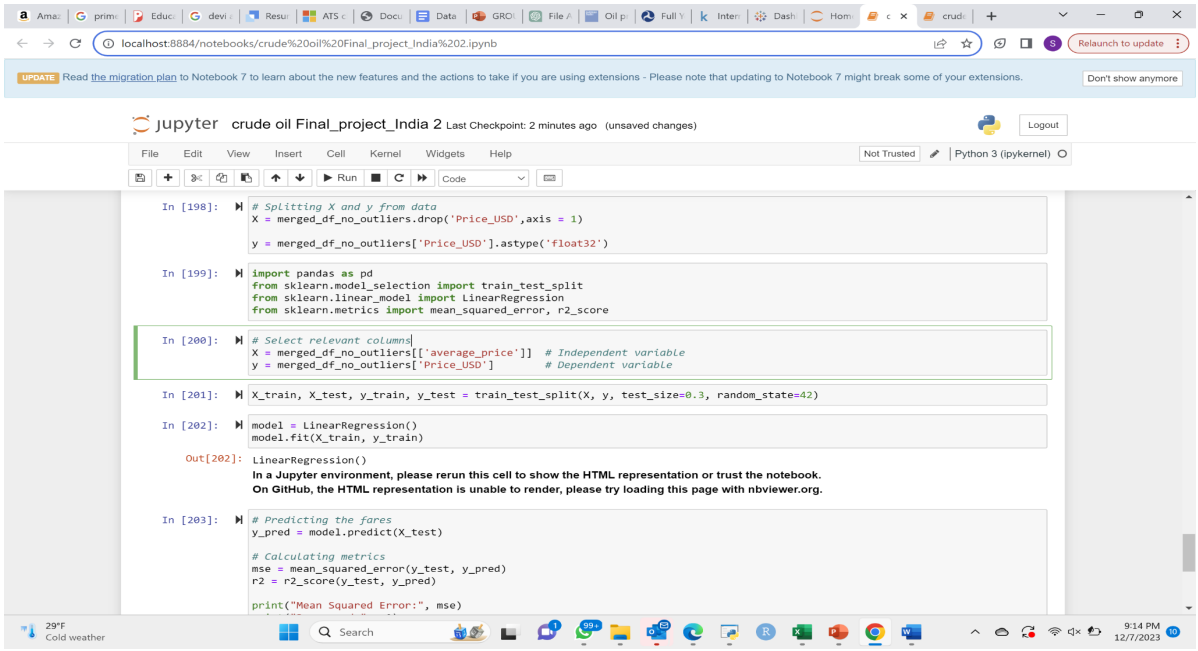
In [193]:
```
# Detecting outliers using the Interquartile Range (IQR) method
Q1 = df_airlines_ind['Price_USD'].quantile(0.25)
Q3 = df_airlines_ind['Price_USD'].quantile(0.75)
IQR = Q3 - Q1

lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Identifying the outliers
outliers = df_airlines_ind[(df_airlines_ind['Price_USD'] < lower_bound) | (df_airlines_ind['Price_USD'] > upper_bound)]
# Creating a subset without outliers
subset_no_outliers = df_airlines_ind[(df_airlines_ind['Price_USD'] >= lower_bound) & (df_airlines_ind['Price_USD'] <= upper_b

# Returning a summary of the outliers and the subset without outliers
outliers_info = outliers.describe()
subset_info = subset_no_outliers.describe()

(outliers_info, subset_info)
```

Out[193]:
```
(           duration         price     Price_USD
 count  25464.000000  25464.000000  25464.000000
 mean      12.370408   8506.278707    120.349161
 std        7.239165   7346.479669    103.940007
 min        0.830000    502.000000      7.102434
 25%        6.920000   3696.500000     52.299095
 50%       11.330000   6015.000000     85.101868
 75%       16.330000  13945.000000    197.297680
 max       44.500000  46658.000000    660.130164,
              duration          price      Price_USD
 count  188105.000000  188105.000000  188105.000000
 mean       12.447108   10351.061099     146.449648
```

29°F Cold weather

Search

9:14 PM 12/7/2023



localhost:8884/notebooks/crude%20oil%20Final_project_India%202.ipynb

Relaunch to update

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.

Don't show anymore

jupyter crude oil Final_project_India 2 Last Checkpoint: 2 minutes ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

In [198]:
```
# Splitting X and y from data
X = merged_df_no_outliers.drop('Price_USD',axis = 1)

y = merged_df_no_outliers['Price_USD'].astype('float32')
```

In [199]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
```

In [200]:
```
# Select relevant columns
X = merged_df_no_outliers[['average_price']]  # Independent variable
y = merged_df_no_outliers['Price_USD']        # Dependent variable
```

In [201]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

In [202]:
```
model = LinearRegression()
model.fit(X_train, y_train)
```

Out[202]:
```
LinearRegression()
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

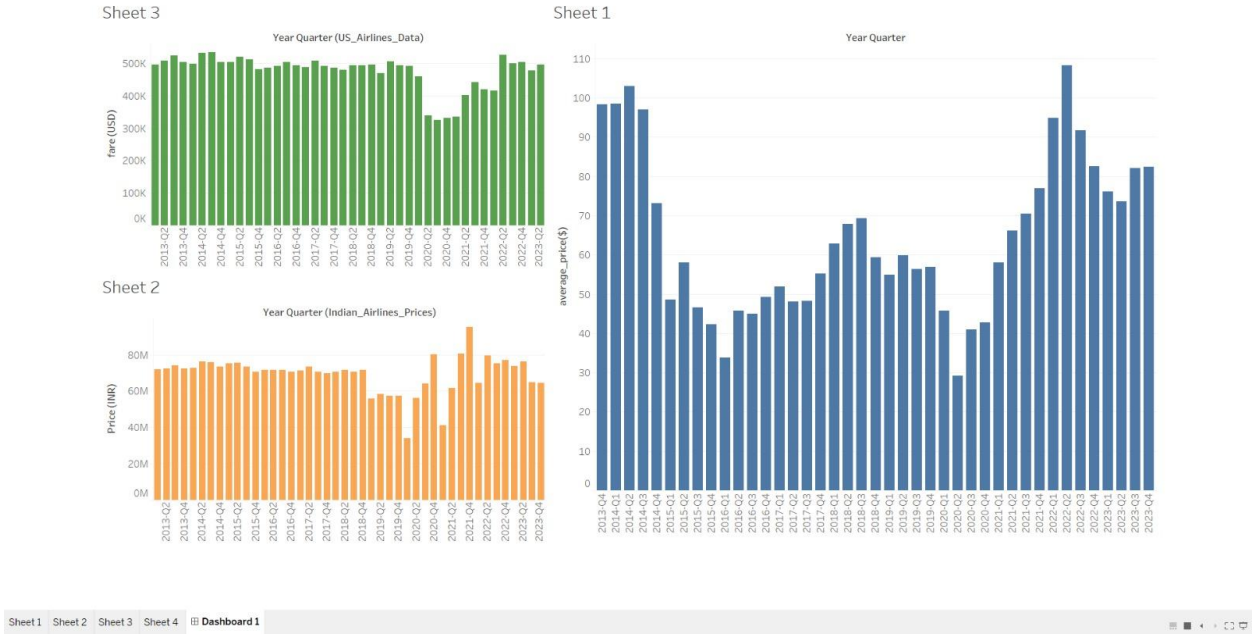
In [203]:
```
# Predicting the fares
y_pred = model.predict(X_test)

# Calculating metrics
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("Mean Squared Error:", mse)
```

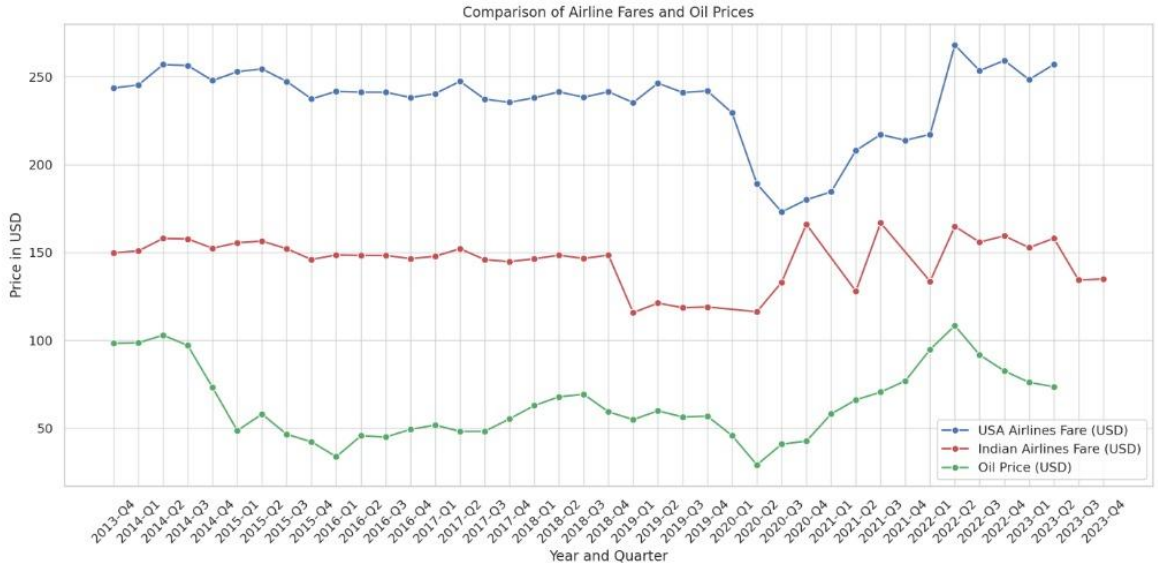
29°F Cold weather

Search

9:14 PM 12/7/2023

6.Visual Analysis





The chart illustrates the fluctuations in prices over time, allowing for a comparison of trends between airline fares in the US and India against the backdrop of oil price changes. The overall trend for each line can provide insights into how sensitive airline fares in both countries are to the changes in oil prices.

7. Final Analysis

Correlation Analysis:

The chart suggests there may be a relationship between oil prices and airline fares, as seen by the general tendency for airline fares to fluctuate in a manner that corresponds to changes in oil prices. However, this relationship is not perfectly direct, indicating that other factors also influence airline fares.

Airline Fare Trends:
The trends show that US airline fares have remained relatively stable with slight fluctuations compared to Indian airline fares, which display more volatility. This could be due to a variety of factors including different market dynamics, fuel hedging practices, and operational efficiencies.

Impact of Oil Prices:
There are periods where the oil price changes seem to have a visible impact on the airline fares, particularly if there is a sharp increase or decrease in oil prices. For instance, a significant drop in oil prices does not always correspond to an immediate decrease in airline fares, which may be attributed to the airlines' pricing strategies and fuel hedging.

Regional Differences:
The graph indicates that the airline industry's response to changes in oil prices may vary by region. For example, Indian airlines show a more pronounced response to oil price changes than US airlines. This difference may be due to the scale of the airlines, regulatory environments, or the elasticity of demand in each country.

Concluding Remarks:
While there is a general trend that airline fares follow the movement of oil prices, the correlation is not absolute. Airline fares are influenced by a complex mix of factors including but not limited to oil prices, such as demand, competition, seasonal trends, and airline-specific cost structures and strategies.
It is important for businesses and analysts to consider these multifaceted influences when making decisions or predictions about the airline industry's pricing patterns.
The analysis underscores the importance of strategic planning in the airline industry to manage the risks associated with volatile oil prices.

8. Contribution

Yalla Surya - Data Analysis and Predictive Modeling

- Worked on Python Jupyter Notebook for advanced data wrangling and preprocessing.

- Conducted statistical analysis and developed predictive models using `scikit-learn.
- Interpreted the modeling results to draw actionable insights.

Shreya Rao - Data Collection and Management

- Designed the survey and managed the distribution using an online survey platform.
- Collected primary data and compiled it into a master dataset.
- Performed initial data cleaning and organization in Microsoft Excel.

Akshita Krishna Gajengi - Data Visualization and Reporting

- Created a series of dynamic visualizations in Tableau to depict trends and patterns in the data.
- Developed an interactive dashboard in Tableau for stakeholders to explore the data findings.
- Conducted a thorough quality check of the final report, visualizations and presentation.

Stuti Saxena- Technical Support and Presentation

- Provided ongoing technical support for Python Jupyter Notebook and Tableau.
- Ensured the accuracy of the data analysis and the reliability of the predictive models.
- Authored the final report, summarizing the methodology, findings, and recommendations.

9. References

1. Dashboards
   https://public.tableau.com/app/profile/akshita.gajengi/viz/DashboardforAirlines_17020012139600/Dashboard1?publish=yes
2. Oil Prices Forecast
   https://www.goldmansachs.com/intelligence/pages/oil-prices-are-forecast-to-trade-between-70-and-100-a-barrel-in-2024.html
3. 2022 US Airline data
   https://www.bts.gov/newsroom/full-year-2022-us-airline-traffic-data
4. Kaggle Data
   https://www.kaggle.com/datasets/rajanand/international-air-traffic-from-and-to-india
5. Data Sources
   https://www.bts.gov/content/annual-us-domestic-average-itinerary-fare-current-and-constant-dollars
6. Dataset Source
   https://gabors-data-analysis.com/datasets/airline-tickets-usa/