# 

# 

BUSINESS INTELLIGENCE

CST3340

COURSEWORK 2

STUDENT’S NAME: KABIR SAKA OLAOLUWA

STUDENT’S NUMBER: M01015304

# **Table of Content**

SECTION 1.1 - INTRODUCTION

SECTION 1.2 - DISCUSSION OF DATA CLEANING

SECTION 2.1 - DATA ANALYSIS AND VISUALIZATION

SECTION 3.1 - DATA MINING ALGORITHM

SECTION 3.2 - ANOMALIES

SECTION 3.3 - DATA PRE-PROCESSING

SECTION 3.4 - FURTHER ANALYSIS USING EXPONENTIAL SMOOTHING

SECTION 3.5 - SUMMARY OF RESULTS

SECTION 4.1 - DATA ETHICS

SECTION 5 - CONCLUSION

# **Introduction**

The dataset used in this report contains detailed air traffic passenger statistics, which provides insights into numerous elements of aircraft operations. It includes important information such as departure and arrival locations, airline operations, terminals, boarding areas, fare categories, and passenger counts. The dataset, obtained from [Kaggle,](https://www.kaggle.com/datasets/thedevastator/airlines-traffic-passenger-statistics/discussion?datasetId=2573278&sortBy=voteCount) contains a mix of categorical, numerical, and time-series data, making it versatile for performing various analyses. The inclusion of dates for every event adds to its efficiency by allowing extensive time-based trend analysis to detect seasonal patterns and growth in trends in flights across time.

This dataset provides an in-depth view of airline operations and Passenger behaviors, making it ideal for trend detection and performance analysis. The dataset consists of **16 columns**, each providing specific details about the air traffic data:

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| **Activity Period** | Date when an event or activity occurred. | Numeric |
| **Operating Airline** | Name of the airline managing the flight itinerary. | String |
| **Operating Airline IATA Code** | IATA code of the airline indicating the airport location of operation. | String |
| **Published Airline** | Name of the airline publishing the fare for the flight. | String |
| **Published Airline IATA Code** | IATA code of the airline publishing the fare. | String |
| **GEO Summary** | Summary of the geographic region for the activity/event. | String |
| **GEO Region** | Detailed geographic region where the activity/event occurred. | String |
| **Activity Type Code** | Type of activity recorded. | String |
| **Price Category Code** | Fare category for the flight. | String |
| **Terminal** | Terminal associated with the flight. | String |
| **Boarding Area** | Boarding area or gate for the flight. | String |
| **Passenger Count** | Total number of passengers on the flight. | Integer |
| **Adjusted Activity Type Code** | Adjusted type of activity to account for missing data. | String |
| **Adjusted Passenger Count** | Corrected passenger count for missing or incomplete data. | Integer |
| **Year** | Year of the recorded activity. | Integer |
| **Month** | Month of the recorded activity. | Integer |

## **Discussion of data cleaning undertaken on the dataset**

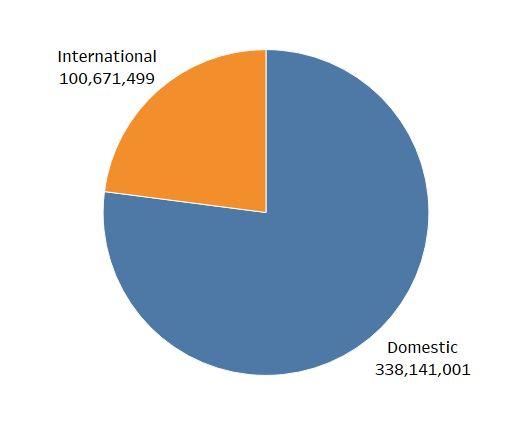
Certain columns in the dataset, like **Operating Airline IATA Code** and **Published Airline IATA Code**, had missing values. Considering the IATA code was a standardized identifier for airlines and their operating locations, addressing these blanks was critical for data accuracy.

To resolve this issue, the corresponding airline names were used as references to look up their respective IATA codes online. This information was widely available and easily accessible due to its standardized use in the aviation and logistics industries. The missing IATA codes were accurately filled in by cross-referencing airline names with credible web sources, assuring the dataset’s consistency.

# **Data Analysis and Visualization**

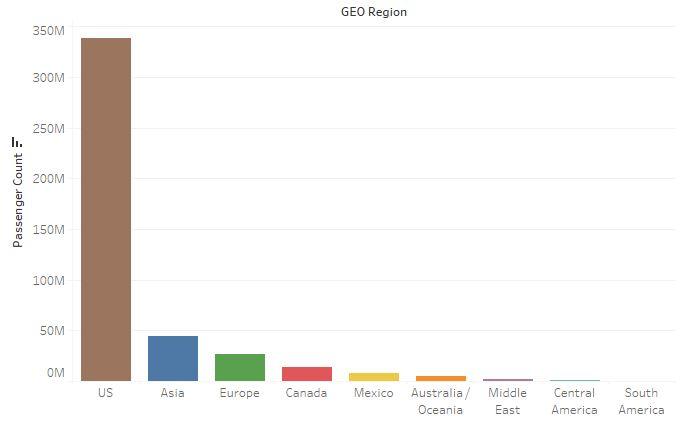
Tableau was implemented to compare the relationship between several columns to gain deeper insights and a better understanding of the data since the dataset focuses on passenger statistics.

Here are several graphs analyzing the relationship between various columns in the dataset:



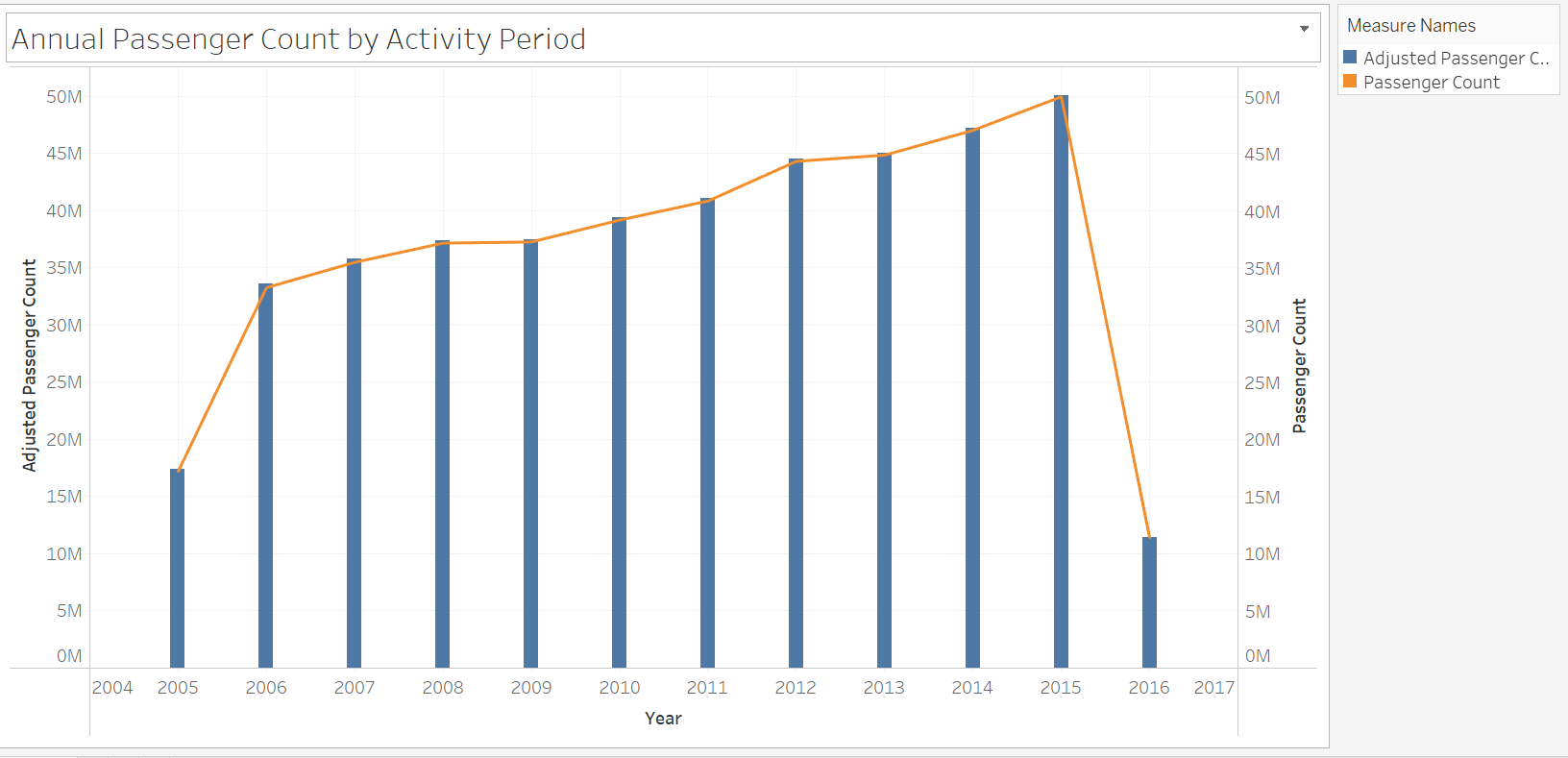
*Figure 1: GEO Summary vs. Passenger Count*

The chart illustrates the number of passengers for both domestic and international flights. The figures given show that domestic flights attract a significantly higher volume of travelers than global flights. This trend implies a preference for in-country travel, possibly due to convenience, flight cost, and frequency of flights. The difference in passenger counts highlights the dynamics of the travel sector for domestic and international routes.



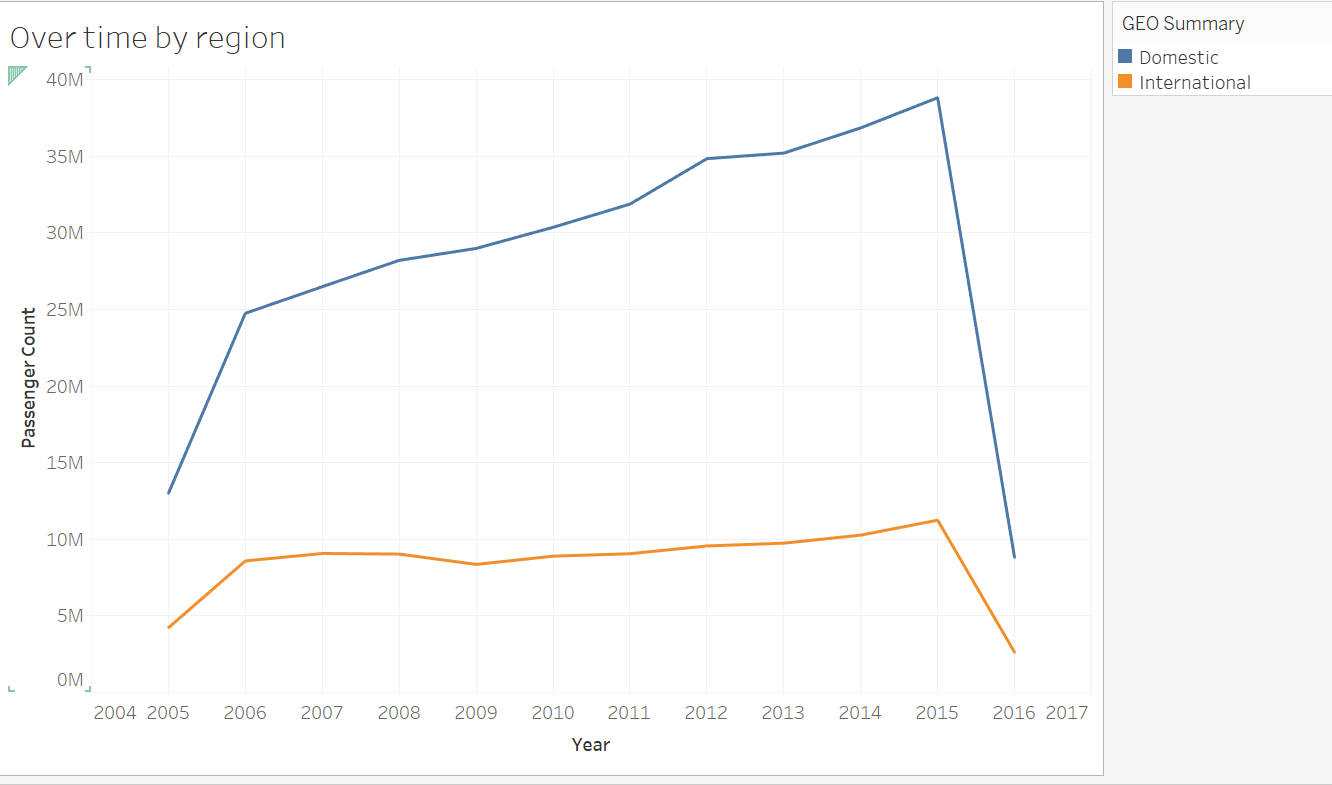
*Figure 2: Passenger Count by GEO Region*

The bar chart depicts how passenger inflows differ among regions. The United States stands out with a massive passenger count, but understanding the underlying reasons for this dominance is more complex. Interestingly, the difference is not limited to just Asia; even Canada, which is geographically close to the US, has a lower passenger count, defying the notion that distance alone does not determine the variations in travel volume. Overall, the data suggests that factors beyond geographical proximity influence passenger volume between these regions.



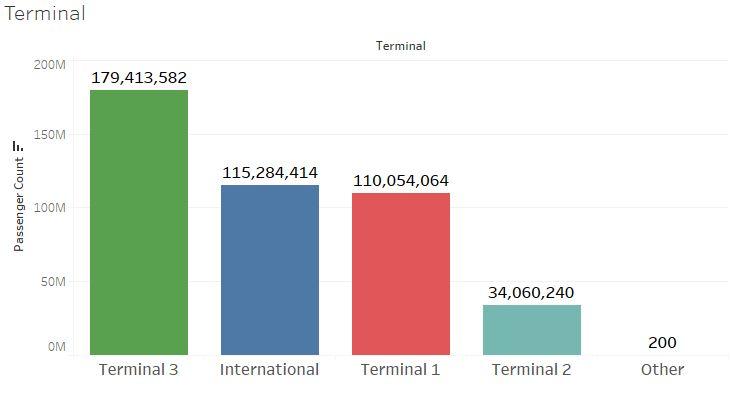
*Figure 3: Annual Passenger Count by Activity Period*

Figure 3 provides a detailed overview of passenger counts from 2005 to 2016. This chart illustrates the annual changes in passenger count, clearly indicating a consistent upward trend from 2005 to 2015. However, a notable decline in passenger counts was evident in 2016. It is important to note that this decline was due to an incomplete data set, as the figures for that year only cover three months.



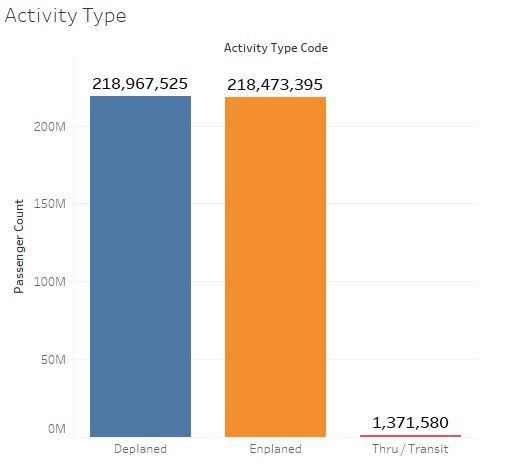
*Figure 4: Comparison of Passenger Count by Activity Period for Domestic and International Flights*

Figure 4 illustrates a line chart that provides a more detailed analysis than Figure 3. It distinguishes between passenger counts for local and international flights over the years. Notably, the chart reveals that the number of passengers traveling on domestic flights consistently outnumbers those on international flights, indicating a significant trend in travel preferences. This separation of data provides valuable insights into the shifting dynamics of flights, showing a higher demand for local flights compared to international flights.



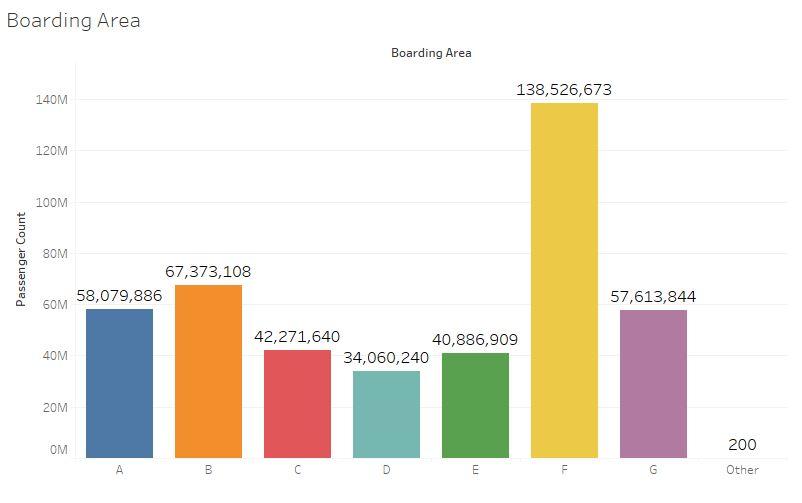
*Figure 5: Passenger Count by Airport Terminal*

This bar chart provides a detailed breakdown of passenger counts across various terminals, highlighting the International terminal, terminals 1, 2, and 3, and additional terminals categorized as "others." The data represented in the chart offers insights into passenger flow, providing a comprehensive understanding of traffic distribution among the different terminals.



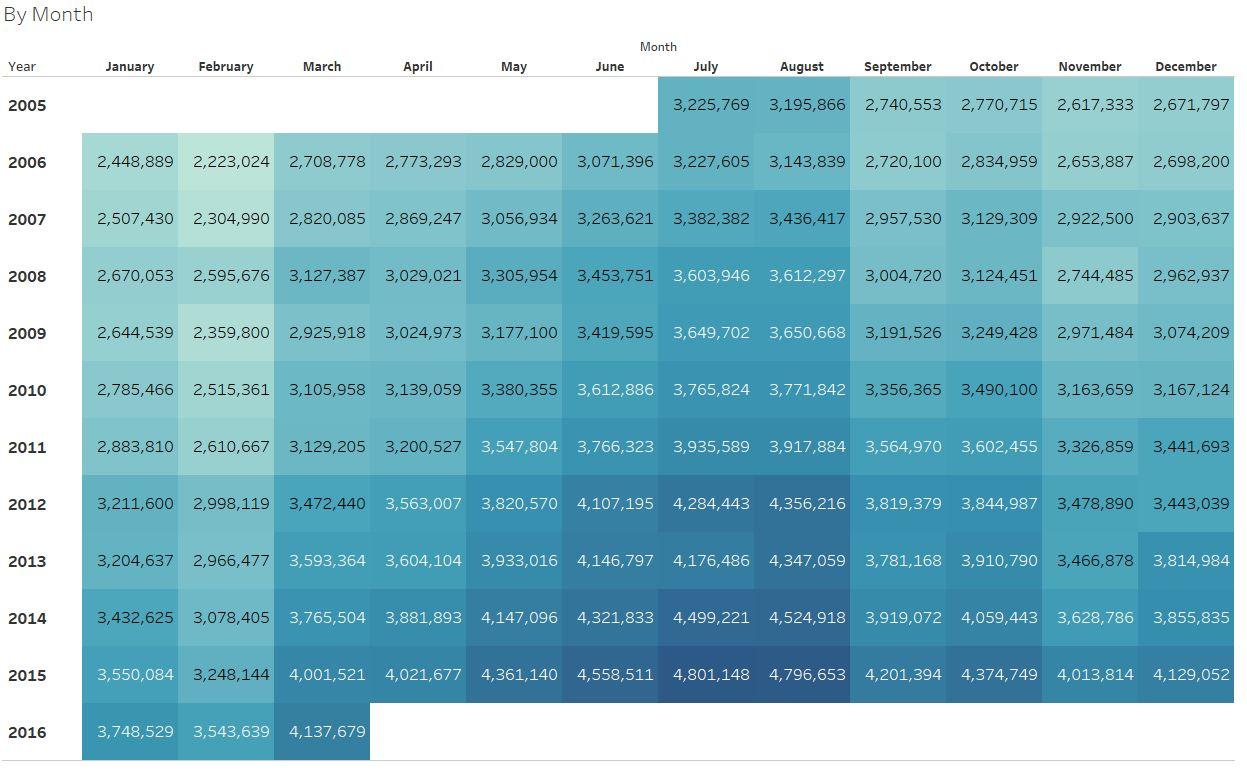
*Figure 6: Passenger Count by Activity Type*

This chart gives a comprehensive overview of flight activity, including the various categories of passenger movement. It focuses on the number of passengers who deplaned (exited the aircraft), enplaned (boarded the aircraft), and those in transit (changed flight). This data provides valuable insights into the travel and overall passenger dynamics at the airport.



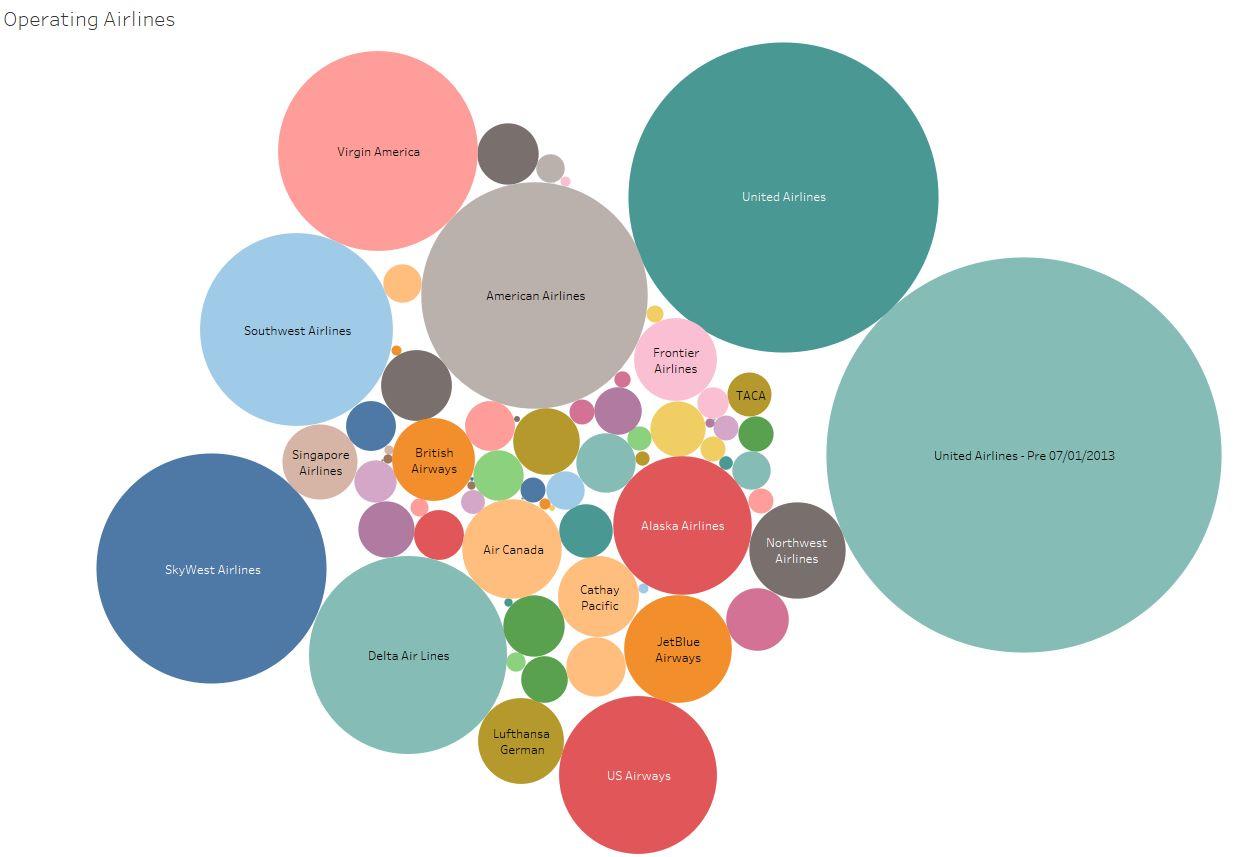
*Figure 7: Passenger Count by Boarding Area*

This chart provides a detailed overview of passenger counts at various boarding areas within the terminal. It illustrates the number of passengers utilizing each boarding zone, allowing for a better understanding of the distribution of foot traffic. By reviewing this data, one can identify areas with a higher amount of travelers and less commonly used boarding areas. This information is crucial for optimizing resource allocation and overall operational efficiency at the terminal.



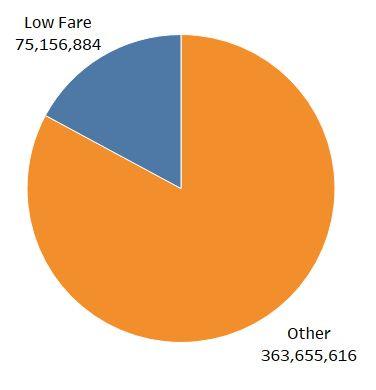
*Figure 8: Monthly Passenger Count by Activity Period*

This chart shows a complete overview of passenger counts from 2005 to 2016, detailing the number of passengers for each month. The data presented allows for a month-by-month comparison, highlighting trends and peaks in passenger numbers over time. This visualization helps in understanding seasonal fluctuations and overall growth or decline in passenger traffic during specific time periods.



*Figure 9: Operating Airlines Distribution*

This chart illustrates the airline with the highest number of passengers for the specified period. The data highlights trends in passenger counts, providing a clear comparison of various airlines' performance in attracting travelers. It also provides insights into the factors influencing these figures, such as route availability, pricing strategy, and general airline reputation.



*Figure 10: Passenger Count Based on Low Fare/Cheap Fare Categories*

This pie chart provides a visual representation of passenger counts for low-fare flights in comparison to other types of flights. This indicates that low-fare flights attract a smaller number of passengers overall. This trend could be attributed to various factors, such as flight timing, the availability of facilities, or customer preferences for comfort or direct travel options. Understanding this passenger distribution can offer valuable insights into market demand and passenger patterns in the airline sector.

# **Data Mining Algorithm**

For this analysis, Exponential Smoothing has been selected as the data mining algorithm. It is a time series forecasting method that is particularly useful for datasets exhibiting trends and seasonal patterns. Unlike static algorithms like linear regression, Exponential Smoothing adapts to long-term trends and fluctuations in data, making it suitable for forecasting future values in a time series.

This model gives more weight to recent observations while still applying historical data, allowing it to capture trends and seasonality effectively. The algorithm involves three components: Level, Trends, and Seasonality, which are updated as new data arrives. This approach allows the model to adjust its forecasts as the data evolves, making it dynamic and responsive to changes in the underlying trend. The choice of Exponential Smoothing is justified by the nature of the dataset, which represents passenger counts over time. As displayed in the visualizations, there’s a clear upward trend in the passenger volume over the years, along with potential seasonal variations.

The key variables for this analysis are the Activity Period—an independent variable representing the time series, and Passenger Count—a dependent variable, for forecasting.

**Anomalies**

The dataset was clean for key variables with no missing values. There were missing values in irrelevant columns, such as Operating Airline IATA Code, which was manually inputted using publicly available information. The dataset was also checked for outliers and inconsistencies, but no issues were found.

## **Data Pre-processing**

Data pre-processing and transformation were essential for preparing the dataset for analysis and mining. The goal was to construct a model to predict Passenger Counts, so the dataset was simplified by focusing on the most relevant variables. The reduction highlighted two key features: **Activity Period (Date)** and **Passenger Count**, which helped to highlight the relationship between the date and passenger count.

The data reduction steps employed were:

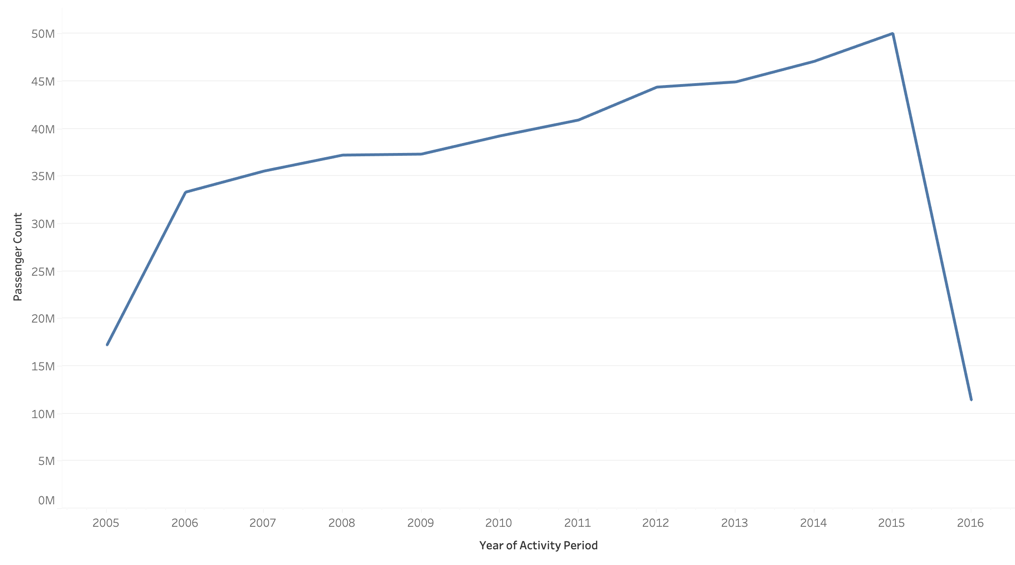
* **Key Variable Identification**: This step involved identifying the most important variables for analysis while discarding less relevant ones. In this case, the **Activity Period (Date)** was selected as the most significant predictor for **Passenger Count**.
* **Dataset Simplification**: This technique reduced the number of variables without losing critical information. The original dataset had sixteen columns, many of which were unrelated to the target variable, **Passenger Count**. The dataset was reduced to two columns—**Activity Period (Date)** and **Passenger Count**.

These pre-processing steps enabled a focused dataset ideal for creating an efficient forecasting model, making it suitable for further analysis.

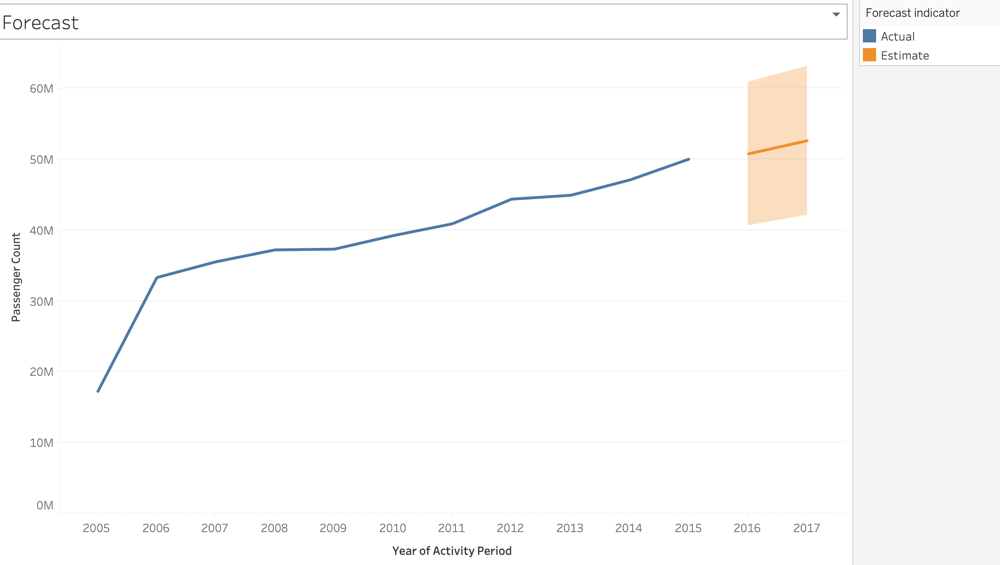
## **Further Analysis Using Exponential Smoothing**

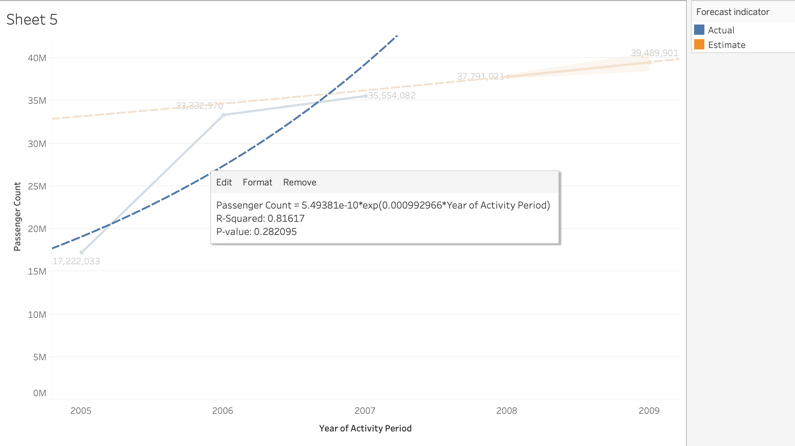
To gain deeper insights into the dataset using the time-series analysis, Exponential Smoothing was chosen as the forecasting method. For this analysis, the key variable has been identified to build a predictive model on Tableau, which focuses on forecasting passenger counts based on the activity period. The following steps outline the application of the exponential smoothing model:

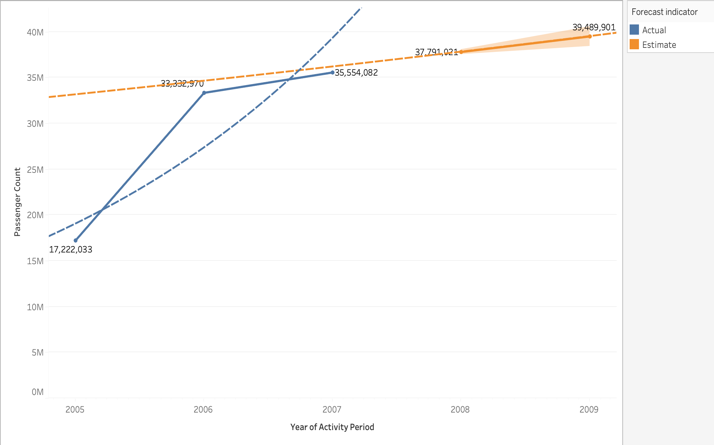
* **Data Import**: The dataset was imported into Tableau for analysis, upon importing, the data was reviewed to ensure that Activity Period was recognized as a date field and Passenger Count was treated as a numeric variable
* **Visualization in Tableau**: A time series plot was drawn in Tableau with Activity Period on the X-axis and Passenger Count on the Y-axis. This allowed for visual analysis of trends and potential seasonal variations. The plot revealed an upward trend in passenger counts during specific periods, along with possible seasonal fluctuations.



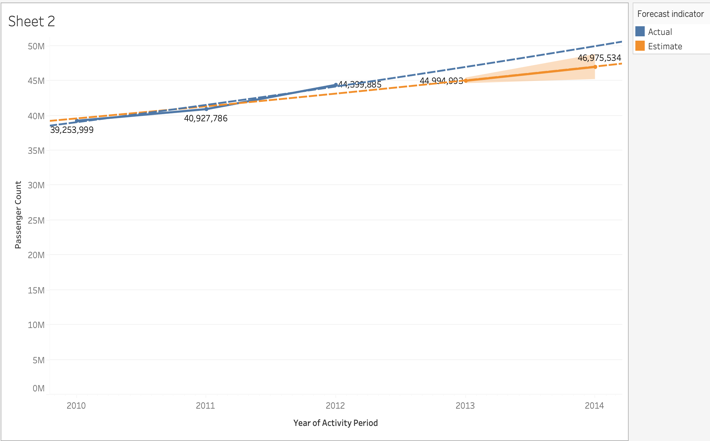
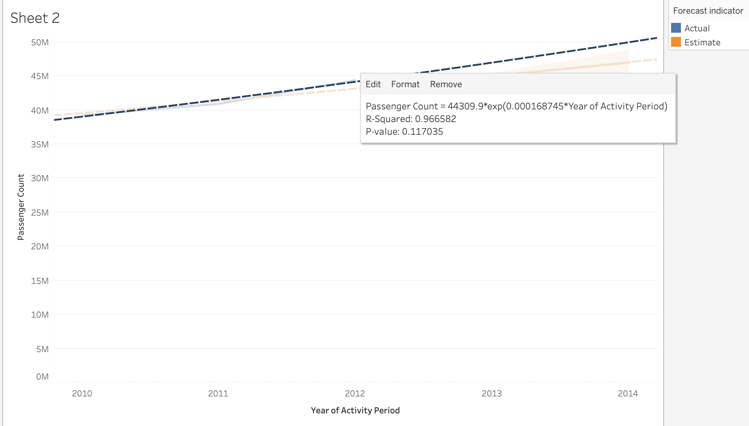
* **Model Selection**: In Tableau, the Exponential Smoothing model was selected from the Analytics pane. Since the dataset showed trend and seasonality, the Seasonal model was implemented. It was chosen due to its ability to capture both the trend and seasonal pattern in the dataset.
* **Forecast Generation**: These forecasts were plotted alongside the actual data, creating a clear visualization of the historical values and predicted future counts. The forecasted values were presented as a shaded area to indicate the uncertainty in the predictions.

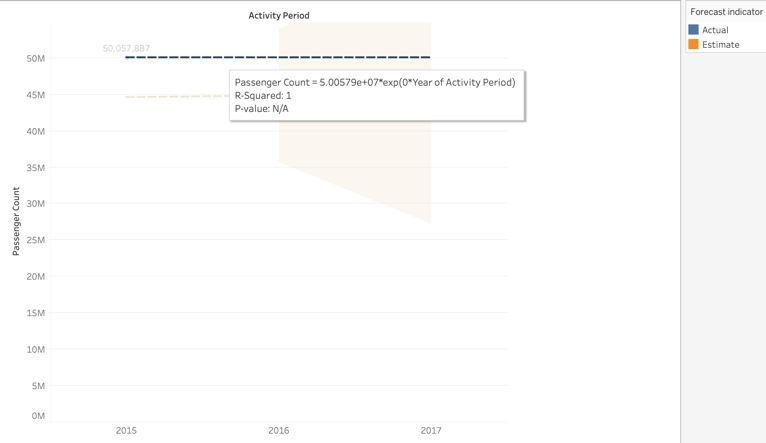


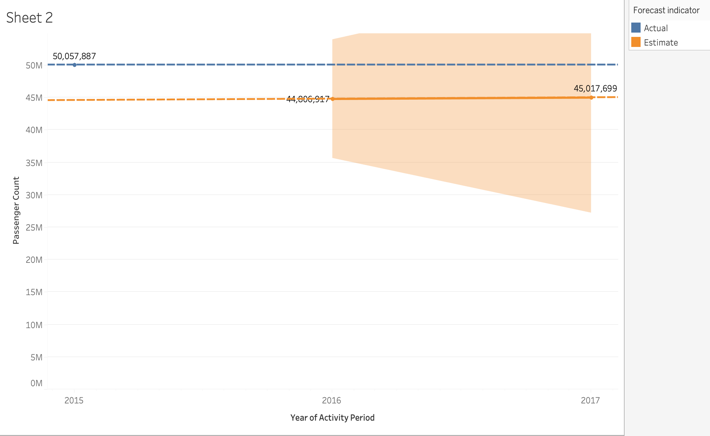
Here are a few time series graphs using the Exponential Trend Line:



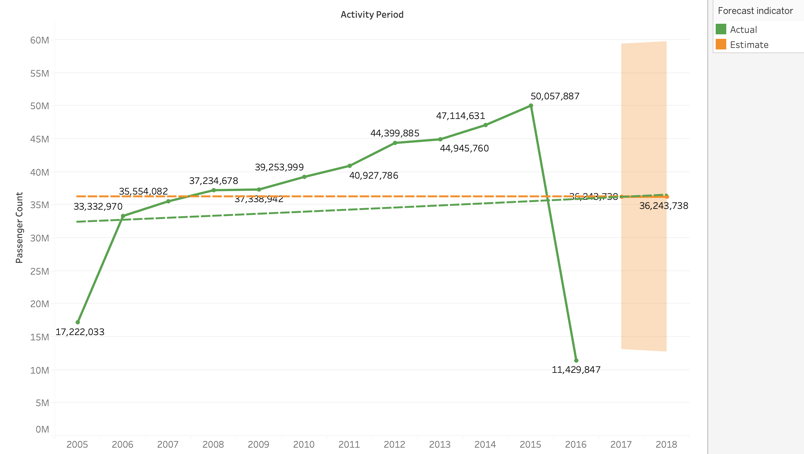
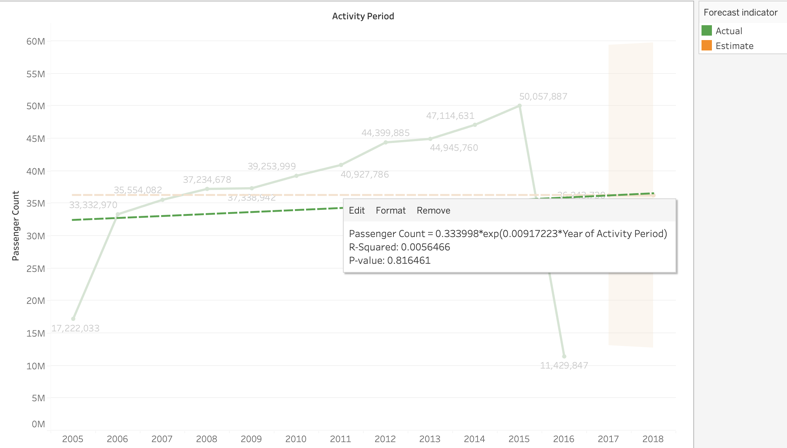
This graph highlights a trend from 2005 to 2007 using an exponential trend line and estimated passenger counts from 2005 to 2009. The blue line shows actual counts, while the orange line shows the estimation of passenger counts, as both demonstrated a close match, indicating the estimation was accurate. The exponential trend line reveals an accelerating passenger growth with an R-squared value of 0.81617 converted to an R-value of 0.90342, showcasing a strong fit.



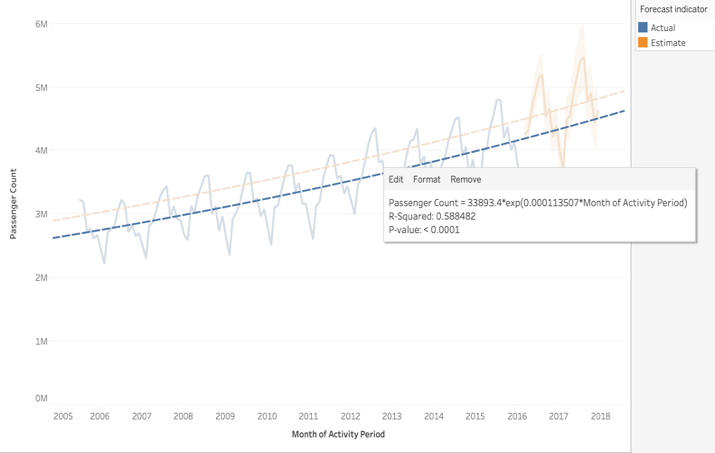
This graph shows the actual and estimated counts from 2010 to 2014. The actual passenger count (blue line) was 39,253,99 in 2010 and rising to 46,975,534 in 2014. The estimated count was 40,927,786 in 2011 and increased to 44,994,993 in 2013. The trend line shows growth in passenger volume with an r-squared value of 1, showing a perfect fit.

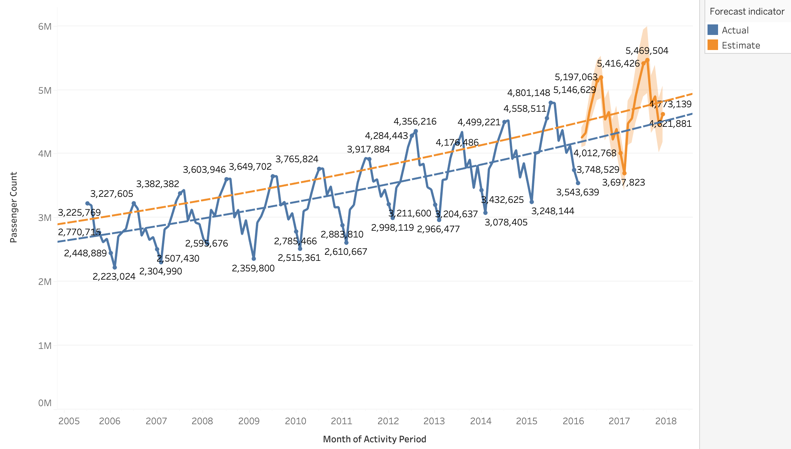


This graph shows the passenger count increased from 2015 to 2016 but decreased from 2016. The actual count exceeded the estimated passenger count in 2015 and 2016. However, in 2017, the actual count was lower than the estimated.

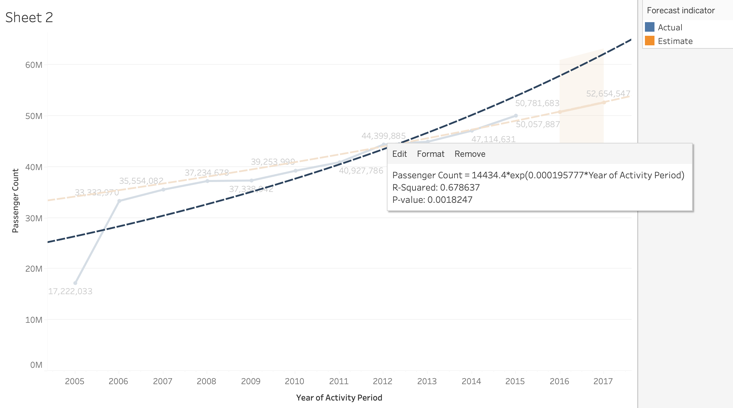


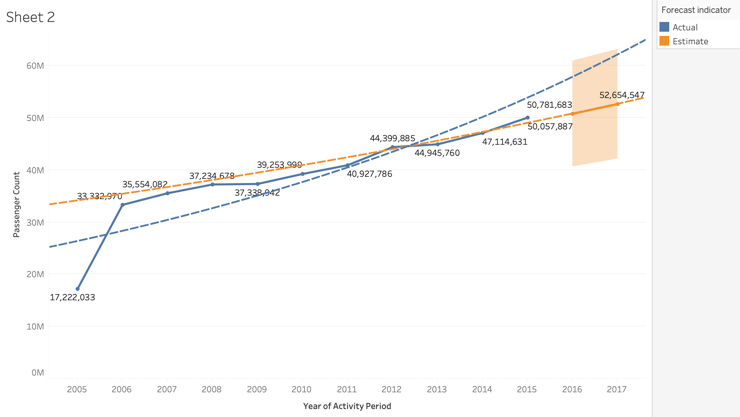
This graph depicts passenger counts over time, peaking at 50,057,887 in 2016 before declining to 11,429,847. From 2005 to 2010, counts were stable with little growth but increased after 2010. The exponential trend line has an R-squared value of 0.0056, corresponding to an R-value of 0.075. This poor correlation indicates that the model doesn't fit the data, implying a lack of consistent exponential growth in passenger trends.





These graphs display the monthly passenger count from 2005 to 2018, highlighting steady growth with fluctuations. The actual passenger count exceeds the estimated values. The exponential trend line has an R-squared value of 0.5885, which translates to 0.767 R-value, indicating a positive relation between passenger count and the activity period. Overall, these figures reflect consistent passenger growth and conservative estimates.





These graphs compare the actual and estimated yearly passenger counts from 2005 to 2017, demonstrating a consistent rise with actual counts exceeding expectations. The slight differences indicate that the estimates were accurate, while the actual growth outpaced expectations. The exponential trend line has an R-squared value of 0.678, which equals 0.823 R-value, showing a solid fit and reflecting constant growth.

**Summary of Results**

The analysis reveals consistent growth in passenger counts over time, with actual counts frequently exceeding estimates. Exponential smoothing and trend lines provided strong fits for the shorter-term trends, especially from 2005 to 2014. The exponential trend model accurately detected trends but failed with anomalies like the decline in 2017 and 2018. Overall, the data highlights consistent increases, seasonal changes, and a tendency for actual passenger counts to outperform conservative forecasts.

# **Data Ethics**

Data ethics is the study of the moral difficulties associated with data collection, processing, sharing, and application. It tackles concerns about algorithms, technology, and data practices, with a focus on protecting individual rights, assuring transparency, and preventing misuse, bias, or discrimination in data processing.

**Ethical Considerations in Data Analysis**

Ethical considerations are the underlying moral standards that guide data throughout its lifecycle, from collection and processing to analysis and distribution. These principles are critical to ensuring trust and integrity in data practices. Key considerations include:

* **Transparency**: Individuals should understand how their data is collected, processed, and used. This involves being transparent about the aim of data collection and how it will be utilized.
* **Bias and Fairness**: Analysts must discover and eliminate biases in the datasets and algorithms. Using skewed data can result in discriminatory consequences, undermining the impartiality and integrity of analysis and decision-making procedures.
* **Privacy**: Protecting individuals' personal information is paramount. Analysts must implement robust measures to prevent unauthorized access, ensuring that sensitive information is kept safe from misuse and breaches.

**Legal Considerations in Data Analysis**

Legal concerns are critical for ensuring that data processes comply with all applicable laws and regulations. This includes:

* **Data Protection Laws**: Analysts must follow strict regulations like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) to protect individuals’ privacy rights and regulate personal data collection, storage, and sharing.
* **Intellectual Property Rights**: Respecting intellectual property regulations is critical to avoiding infringement on data creators’ ownership rights. Analysts must ensure their data usage complies with these standards.
* **Liability**: Understanding the legal ramifications of errors or data misuse is critical in any analytical task. Analysts must be mindful of the potential effects of their conclusions, which may include legal liability for data-related errors or ethical violations.

**Professional Considerations in Data Analysis**

Professional considerations highlight data analysts’ need to maintain a high level of honesty, expertise, and professionalism in what they do. Key features include:

* **Accuracy**: Analysts must prioritize accuracy in data analysis to ensure accurate interpretations and avoid misleading stakeholders.
* **Confidentiality**: Protecting sensitive data must be a top priority. Analysts should place a strict policy to keep confidential information safe from leaks and breaches.
* **Accountability**: Analysts must take full responsibility for the ethical and legal implications of their job. This involves disclosing inaccuracies and working to correct any harmful consequences of their analysis.

# **Conclusion**

### **Overall Visualization Results**

The analysis of the *Airlines Traffic Passenger Statistics* Dataset comprises 15,006 time-series entries across nine columns. Visual analysis of the dataset revealed several key insights into air traffic patterns, including domestic flights consistently recorded higher passenger counts than international flights, indicating the dominance of domestic travel in the dataset.

In addition, the geographical region analysis revealed that flights within the United States carried the most passengers by a significant margin. The second-highest region, Asia, demonstrates a significant gap between these regions. These findings emphasize the significance of domestic travel and the central role of US-based flights in overall passenger traffic.

### **Data Mining Results**

Exponential Smoothing effectively spotted the increased trend in the passenger counts while considering seasonal fluctuations. The model fared particularly well in analyzing shorter terms with steady growth. However, it struggled to account for anomalies, such as the decline observed in 2016, due to incomplete data.

The Exponential Trend Line displays various levels of fit across data segments. R-squared values range from 0.58, indicating a good fit, to 1, showing a perfect fit for specific periods. Strong correlations between these periods demonstrated the model’s ability to accurately track passenger count trends.

### **Business Intelligence**

The findings provide vital business intelligence for airlines, airport operators, and policymakers to help make strategic decisions. Airlines can use the insights to optimize flight schedules, focusing on high-demand routes and seasons. Airports can use these findings to improve resource allocation by altering staff and managing terminal usage depending on passenger patterns during peak periods, thereby enhancing operational efficiency and customer satisfaction. In practice, airlines can present promotional offers for underutilized routes to improve passenger numbers, while airports can capitalize on low-traffic periods to conduct maintenance facilities, reducing passenger disruption.

These data-driven strategies can assist in enhancing operational efficiency, customer satisfaction, and long-term planning, allowing for better forecasting and decision-making across the industry.

