**SKY INSIGHTS:**

**COMPREHENSIVE U.S AIRLINE DATA ANALYSIS**

**Project Description**

1. **Aim of the Project:**

The primary objective of this project is to conduct a comprehensive analysis of airline data to uncover meaningful trends and patterns related to passenger counts, fare structures, and market dynamics across different year ranges. By examining key variables such as fare, passenger distribution, and market share for both large and low-fare carriers, the project aims to identify significant trends, correlations, and outliers within the dataset.

Through detailed statistical analysis, including correlation assessments and hypothesis testing, the project seeks to reveal how fare changes impact passenger counts and how different airlines compare in terms of market share over time. Additionally, the project will explore the relationships between various metrics, such as the connection between fare and passenger numbers, to provide actionable insights for optimizing pricing strategies and improving route planning.

The ultimate goal is to leverage these insights to support strategic decision-making processes within the airline industry, enhancing operational efficiency and customer satisfaction through data-driven recommendations. The analysis will provide a deeper understanding of market trends and help the airline adapt to evolving customer demands and competitive pressures.

1. **Business Problem or Problem Statement**

**Fare Pricing Optimization**: Airlines often struggle with setting optimal fare prices that maximize revenue while remaining competitive. Without accurate insights into how fare changes impact passenger behaviour, airlines risk either pricing themselves out of the market or missing out on potential revenue. Effective fare pricing requires a deep understanding of demand elasticity and fare sensitivity across different routes and seasons. This problem necessitates the use of historical fare and passenger data to develop data-driven pricing strategies that balance profitability with customer demand.

**Identifying High-Demand Routes**: Determining which routes have the highest passenger demand is crucial for airlines to optimize flight schedules and capacity. Misallocating resources—either by over-serving less popular routes or under-serving high-demand routes—can lead to inefficiencies and lost revenue. By analyzing passenger distribution across different routes and identifying trends, airlines can make informed decisions about route planning and frequency adjustments. This analysis helps in aligning operational strategies with market demand to enhance route profitability.

**Understanding Market Share Dynamics**: Gaining insights into market share and competitive positioning is essential for strategic planning in the airline industry. Airlines need to understand their market share across various routes and fare segments to evaluate their competitive stance and identify growth opportunities. Analyzing market share data helps in assessing the performance relative to competitors, guiding decisions on route expansions, and crafting targeted marketing strategies. This understanding is critical for maintaining a competitive edge and capturing a larger share of the market.

**Our Approach to Solve the Problem**

**Fare Pricing Optimization**: We analyzed fare and passenger data to uncover trends and understand the impact of fare changes on passenger volumes. Through line plots and scatter plots, we assessed fare trends and their relationship with passenger counts. Statistical tests, including hypothesis testing and ANOVA, were conducted to determine the significance of fare adjustments, guiding optimal pricing strategies.

**Identifying High-Demand Routes**: We aggregated and ranked passenger counts by route to highlight high-demand areas. Visualization techniques such as bar plots and heatmaps were used to display passenger distribution and demand intensity, aiding in route optimization and capacity planning.

**Understanding Market Share Dynamics**: By calculating and comparing market share percentages, we evaluated the airline’s competitive positioning. Pie charts and bar plots illustrated market share distribution, helping to identify strengths and market opportunities for strategic decision-making.

1. **Project Description**

**Data Features**

1. tbl: Identifier for the table from which the record is sourced.
2. Year: The year the data record was collected.
3. quarter: The quarter of the year, ranging from 1 to 4.
4. citymarketid\_1: Market ID for the origin city.
5. citymarketid\_2: Market ID for the destination city.
6. city1: Name of the origin city.
7. city2: Name of the destination city.
8. airportid\_1: ID for the origin airport.
9. airportid\_2: ID for the destination airport.
10. airport\_1: Code for the origin airport.
11. airport\_2: Code for the destination airport.
12. nsmiles: Distance between the origin and destination airports, measured in miles.
13. passengers: Total number of passengers on the route.
14. fare: Average fare for the route.
15. carrier\_lg: Code for the largest carrier (by passenger count) on the route.
16. large\_ms: Market share percentage of the largest carrier.
17. fare\_lg: Average fare charged by the largest carrier.
18. carrier\_low: Code for the carrier offering the lowest fare on the route.
19. lf\_ms: Market share percentage of the lowest fare carrier.
20. fare\_low: Lowest fare available on the route.
21. Geocoded\_City1: Geocoded coordinates for the origin city.
22. Geocoded\_City2: Geocoded coordinates for the destination city.
23. tbl1apk: Unique identifier for each route.

**Scope and Objectives**

**Fare Analysis:** The primary objective was to examine how fare trends evolved over time and their influence on passenger volumes. By analysing fare data from different periods, we aimed to uncover patterns in fare fluctuations and their correlation with passenger counts. This insight helps in identifying optimal pricing strategies that align with customer demand and maximize revenue.

**Identification of Top Airlines:** We focused on identifying the top 5 airlines based on passenger counts. This involved aggregating passenger data for each airline and ranking them to pinpoint the leaders in terms of passenger volume. Analysing trends for these top airlines allowed us to understand their performance over time and how fare changes impacted their market position.

**High-Demand Routes**: Another key objective was to assess route efficiency by identifying the top 5 boarding and destination points. We aggregated passenger counts for various city pairs and ranked them to find the most frequently used routes. This analysis provided insights into route popularity, helping to optimize flight schedules and resource allocation.

**City Pair Analysis:** We determined the top 5 city pairs with the highest passenger counts and analysed their trends across various parameters. This included fare trends, passenger distribution, and market share. By evaluating these city pairs, we aimed to understand which routes were most lucrative and how fare changes affected passenger preferences.

**Trend Analysis of Top City Pairs:** For the identified top city pairs, we conducted a detailed trend analysis to understand fare variations over time. We examined quarterly fare trends, exploring how fare fluctuations influenced passenger volumes for these routes. This analysis provided a comprehensive view of how different parameters affected the top city pairs.

**Visualization and Statistical Testing:** To support our analysis, we utilized various visualization techniques such as line plots, scatter plots, and heatmaps. These visualizations helped in illustrating trends and relationships between variables. Additionally, we applied statistical tests such as ANOVA and hypothesis testing to validate our findings and ensure the robustness of our conclusions.

**Technologies and Methodologies**

**Data Collection and Preparation**:

* **Data Sources**: The analysis was based on airline data, which included various columns such as fare, passenger counts, city names, and year ranges.
* **Data Cleaning**: Initial data preparation involved cleaning the dataset to handle missing values and outliers. Techniques such as replacing outliers with median values and filtering extreme values were employed to ensure data quality.

**Exploratory Data Analysis (EDA)**:

* **Descriptive Statistics**: Basic statistical measures, including means, medians, and standard deviations, were calculated to understand the data distribution.
* **Visualizations**: Various plots were created to visualize trends and relationships:
* **Line Plots**: Used to analyze fare trends over time and by carrier.
* **Bar Graphs**: Displayed top city pairs and their passenger counts across different year ranges.
* **Scatter Plots**: Explored relationships between variables such as large\_ms and fare\_lg, nsmiles and fare, etc.
* **Heatmaps**: Illustrated correlations between different variables to identify significant relationships.

**Trend Analysis**:

* **Top Airlines**: Identified the top 5 airlines by aggregating passenger data and analyzing their performance over time.
* **Boarding and Destination Points**: Determined the top 5 boarding and destination cities by summing passenger counts.
* **City Pairs**: Analyzed top city pairs based on passenger volume and evaluated trends in fare and other parameters.

**Statistical Testing**:

* **ANOVA**: Used to test for significant differences in passenger counts across different year ranges. The ANOVA test helped determine if fare changes had a statistically significant impact on passenger volumes.
* **Chi-Square Test**: Applied to examine the independence between year ranges and city pairs. This test assessed whether the distribution of passengers among city pairs was significantly different across year ranges.

**Advanced Analytics**:

* **Correlation Analysis**: Calculated correlation coefficients to understand the relationships between variables like nsmiles, fare, and passengers.
* **Pair Plots**: Visualized the relationships between multiple variables to identify patterns and trends.

**Data Visualization Tools**:

* **Matplotlib and Seaborn**: Utilized for creating various plots, including line plots, bar graphs, scatter plots, and heatmaps. These libraries provided a comprehensive suite of visualization tools to effectively communicate insights.
* **Jupyter Notebook**: Employed for conducting the analysis, executing code, and documenting results interactively.

1. **Functionalities**

**Data Cleaning and Preparation**:

* **Handling Missing Values**: Missing values were imputed using median values for numerical columns and mode for categorical columns. This ensures that data completeness does not skew analysis.
* **Outlier Detection and Treatment**: Outliers were detected using statistical methods (e.g., IQR) and replaced with median values to prevent distortion in analyses like fare trends and passenger counts.
* **Data Transformation**: Data was transformed as needed, including scaling and normalization, to prepare it for various types of analysis and visualization.

**Trend Analysis of Fares**:

* **Aggregation**: Fare data was aggregated by year and quarter to analyze trends over time. For example, average fares for each year range (e.g., 2001-2005) were calculated.
* **Visualization**: Line plots were created to visualize fare trends by quarter and year, providing insights into how fare levels have changed over different time periods.

**Identification of Top Airlines**:

* **Passenger Count Aggregation**: Passenger counts were summed by airline to identify the top 5 airlines with the highest total passenger volumes.
* **Ranking**: The top airlines were ranked based on their passenger counts, allowing for a focus on major players in the industry.

**Analysis of Boarding and Destination Cities**:

* **Aggregation by City**: Boarding and destination city data were aggregated to determine the total number of passengers departing from and arriving at each city.
* **Top Cities Identification**: The top 5 cities for both boarding and destination were identified based on passenger volumes, providing insights into major hubs and popular destinations.

**City Pair Analysis**:

* **City Pair Creation**: Combined city names into city pairs to represent routes (e.g., “City A to City B”).
* **Aggregation and Ranking**: Passenger counts were aggregated by city pair to identify the top 5 routes with the highest traffic. This helps in understanding which routes are most heavily traveled.

**Correlation and Relationship Analysis**:

* **Correlation Matrix**: A correlation matrix was computed to explore relationships between variables such as fare, passenger counts, and market share.
* **Heatmap Visualization**: A heatmap was generated to visually represent the strength and direction of correlations between different variables.

**Statistical Testing**:

* **T-Tests**: Conducted t-tests to compare fare averages and passenger counts across different year ranges to assess statistical significance of observed differences.
* **ANOVA**: Performed ANOVA to test for differences in market share across multiple time periods, helping to determine if trends are statistically significant.

**Visualization of Data Trends**:

* **Line Plots**: Used to show trends in fares and passenger counts over time, revealing patterns and fluctuations.
* **Scatter Plots**: Created to explore relationships between variables such as large\_ms and fare\_lg, and nsmiles and passenger counts.
* **Bar Charts**: Displayed top 5 city pairs, boarding cities, and destination cities to easily compare and contrast different metrics.

**Interactive Analysis with Jupyter Notebook**:

* **Notebook Integration**: Employed Jupyter Notebooks to perform interactive data analysis. Code execution, visualization, and results were managed within an iterative and flexible environment.
* **Documentation**: Provided detailed explanations and comments within the notebook to ensure clarity and reproducibility of the analysis.

**Error Handling and Exception Management**:

* **Validation Checks**: Implemented validation checks for input data and processing steps to catch and handle errors gracefully.
* **Exception Handling**: Included exception handling in code to manage issues such as data format inconsistencies or calculation errors, ensuring robustness in analysis.

1. **Code Implementation**

The project is implemented using Python, leveraging key libraries such as pandas, numpy, seaborn, and matplotlib for data manipulation and visualization. The code is organized into distinct sections, each focusing on specific aspects of the analysis.

**Data Preparation and Cleaning:**

* Data Loading: Data is loaded into pandas DataFrames using pd.read\_csv(). This allows for efficient handling and manipulation of large datasets.
* Handling Missing Values: Missing values in numerical columns are imputed using median values with df.fillna(df.median()), while categorical columns use mode values with df.fillna(df.mode()[0]).
* Outlier Detection: Outliers are detected using the Interquartile Range (IQR) method. Values outside the IQR range are replaced with the median to ensure robust analysis.
* Data Transformation: Numerical features are scaled using StandardScaler or MinMaxScaler to normalize the data and facilitate comparison across different scales.

**Trend Analysis and Visualization:**

* Aggregation: Data is aggregated by year and quarter using groupby() and agg() methods. For example, fare trends are calculated with df.groupby(['Year\_Range', 'quarter'])['fare'].mean().
* Visualization: Line plots are created using seaborn’s sns.lineplot() to visualize trends. plt.figure() and plt.subplot() are used to organize multiple plots in a single figure for comparative analysis.

**Identification of Top Airlines and Routes:**

* Aggregation and Ranking: Top airlines are identified by summing passenger counts with df.groupby('airline')['passengers'].sum().sort\_values(ascending=False). Similarly, top boarding and destination cities, as well as city pairs, are determined using groupby() and sorting functions.
* Data Selection: The top 5 airlines, city pairs, and cities are selected based on their aggregated passenger counts to focus on the most significant routes and destinations.

**Correlation and Statistical Testing:**

* Correlation Matrix: A correlation matrix is computed using df.corr() to explore relationships between variables. A heatmap is generated with sns.heatmap() to visually represent these correlations.
* Statistical Testing: T-tests and ANOVA are performed using scipy.stats.ttest\_ind() and scipy.stats.f\_oneway() to assess statistical significance. These tests help determine if observed differences in fares and market share across time periods are significant.

1. **Results and Outcomes**

**Trend Analysis:**

* Fare Trends: The trend analysis revealed how average fares fluctuated over different quarters and year ranges. Notable trends include seasonal variations in fare prices, with some quarters exhibiting higher average fares due to increased travel demand.
* Passenger Trends: Passenger count trends highlighted peak periods and shifts in travel preferences over the years, identifying periods of high travel volume and low demand.

**Top Airlines and Routes:**

* Top Airlines: The top 5 airlines were identified based on total passenger volumes. This information helps in understanding market share distribution and competitive positioning within the industry.
* Top Boarding and Destination Cities: Analysis of boarding and destination cities provided insights into major hubs and popular travel destinations. This is crucial for airlines to optimize route planning and marketing strategies.
* Top City Pairs: The top 5 city pairs with the highest passenger traffic were identified, revealing the most popular routes and enabling airlines to focus on these high-demand connections.

**Correlation Insights:**

* Fare and Passenger Count: Correlation analysis showed relationships between fare levels and passenger counts, helping to understand how fare changes impact passenger volumes.
* Market Share Dynamics: Analysis of market share across different time periods highlighted shifts in market dominance among airlines, providing insights into competitive dynamics and market trends.

**Statistical Testing:**

* Significant Differences: T-tests and ANOVA results indicated whether observed differences in fares, passenger counts, City pairs and market share across time periods were statistically significant, supporting data-driven decision-making for fare pricing and market strategy.

1. **Conclusion**

The project delivered a detailed analysis of airline data, focusing on fare trends, passenger distribution, and route efficiency. Key findings include the identification of top airlines, major boarding and destination cities, and the most heavily trafficked city pairs. The analysis revealed significant trends in fare fluctuations and passenger volumes, offering actionable insights for optimizing pricing strategies and improving route planning.

The project also provided valuable correlation insights and statistical tests that helped in understanding the relationships between different variables and the significance of observed trends. These insights are crucial for making informed decisions and enhancing operational efficiency within the airline industry.

Moving forward, future developments could include integrating more granular data, such as customer demographics and booking patterns, to refine predictions and strategies. Additionally, incorporating advanced machine learning models could further enhance trend analysis and forecasting accuracy.