Mini Project 1: Structured Data

IST652 - Scripting for Data Analysis

Khushi Shetty
SUID: 855125581

Exploratory Data Analysis of an Airline dataset using python.

Data and its Source

The dataset used in this analysis is the "Airline Dataset.csv." This dataset contains comprehensive information about passengers, flights, and various attributes relevant to air travel. Understanding the dataset's structure and contents is crucial for the analysis.
The dataset serves as a valuable source for understanding flight and passenger characteristics, making it a valuable resource for airlines to assess their operations and customer demographics.

Data Source: Kaggle
Airline Dataset (kaggle.com)

Data Exploration and Data Cleaning

The data exploration and cleaning process consisted of several crucial steps to prepare the dataset for analysis:

1. Initial Data Load: I began by loading the dataset into a Pandas dataframe, allowing me to easily manipulate and explore the data.
2. Column Removal: Unnecessary columns, including "First Name," "Last Name," "Airport Country Code," and "Airport Continent," were dropped. This action simplifies the dataset, focusing on key attributes.
3. Date Conversion: The "Departure Date" column was converted to a datetime data type. This conversion is essential for conducting date and time-related analysis.
4. Date Format Standardization: I identified inconsistencies in date formatting, with some dates using hyphens ('-') instead of forward slashes ('/'). Standardizing this formatting ensures consistent date representation and avoids potential issues in date parsing.
5. Duplicate Entry Check: To maintain data integrity, I checked for duplicate entries. The analysis confirmed that there were no duplicate records, assuring the uniqueness of each entry.
6. Missing Value Check: To ensure data completeness, I checked for missing values. The dataset was found to be free of missing values, making it clean and ready for analysis.

Comparison Questions:

I addressed three key comparison questions in this analysis:

Analysis 1: Flight Status by Airport Continent and Gender

Unit of Analysis: Airport continent
Comparison: To understand the distribution of passengers by gender across different flight statuses for each airport continent.
Computation:

- Group the data by "Airport Continent," "Flight Status," and "Gender."
- Calculate the count of flights for each combination.
- Display the results, providing the counts of male and female passengers for each combination.

Summary of Findings:
The analysis revealed the distribution of flight statuses across various airport continents and genders.

1. Continent with the Highest Total Flights: North America has the highest total number of flights, with a total of 32,033 flights. This indicates that North America is a major hub for air travel.
2. Flight Status Distribution by Continent:
    a. Africa: Africa has a balanced distribution of flight statuses, with roughly similar numbers of flights that were Cancelled, Delayed, and On Time.
    b. Asia: Asia also shows a balanced distribution, with slightly more flights reported as On Time compared to Cancelled and Delayed.
    c. Europe: Europe has a similar distribution pattern, with an even distribution among the three flight statuses.
    d. North America: North America has a significant number of flights in all categories, indicating a robust aviation industry. Delayed flights are slightly more common than Cancelled or On Time flights.
    e. Oceania: Oceania also has a balanced distribution of flight statuses, with all three categories having a similar number of flights.
    f. South America: South America shows a distribution pattern similar to Africa and Oceania, with no single category dominating.
3. Total Flights by Continent:
    a. North America dominates in terms of the total number of flights, followed by Asia and Europe.
    b. Africa, South America, and Oceania have significantly fewer total flights compared to the top three continents.

Analysis 2: Distribution of Passengers in Various Age Groups

Unit of Analysis: Age groups
Comparison: Understanding the distribution of passengers within different age groups, segmented by gender.
Computation:

- Define bin edges and labels for age groups.
- Create a new column "Age Group" by binning the "Age" column.
- Calculate the count of passengers in each age group by gender.
- Calculate the total count and percentage of males and females in each age group.

Summary of Findings:
The analysis presents the distribution of passengers across age groups and their gender.

1. Age Group "0-17":
This age group has a large number of passengers as compared to most of the age groups, indicating a significant portion of travelers are children and teenagers.

2. Age Group "18-29":
Young adults aged 18-29 make up a substantial portion of passengers.

3. Age Group "30-39":
This age group represents individuals in their thirties, and while the count is lower than the previous group, it's still a significant passenger segment.

4. Age Group "40-49":
Passengers in their forties also constitute a considerable segment.

5. Age Group "50-59":
This age group has a substantial passenger count, indicating the importance of catering to the preferences and requirements of travelers in their fifties.

6. Age Group "60+":
The "60+" age group stands out with the highest total count, suggesting that older travelers are a significant portion of the airline's customer base. Airlines should focus on services that meet the unique needs of this demographic.

In terms of gender, the analysis shows a balanced distribution of male and female passengers in all age groups.

Analysis 3: Distribution of Flight Statuses Across Different Months

Unit of Analysis: Months
Comparison: Understanding how flight statuses vary across different months.
Computation:

- Extract the month from the 'Departure Date' column.
- Analyze flight activity by month, calculating the count of flights for each status.
- Display the results.

Summary of Findings:
The analysis reveals how flight statuses change throughout the year.
January and March have the highest counts of Delayed flights, while August and July have the highest counts of On Time flights. These variations could be due to a range of factors, including weather, holidays, and travel patterns.

Program Description

The analysis program was implemented using Python, leveraging essential libraries such as NumPy, Pandas, and Matplotlib. The analysis followed a structured approach, encompassing data loading, cleaning, analysis, and visualization. Each step was meticulously executed to answer the comparison questions.

The Python code facilitated efficient data manipulation, analysis, and visualization, ensuring the delivery of valuable insights for the airlines.

Output Files

Three primary output files were generated as part of this analysis:

1. "flight_status_gender_per_continent.csv":
This CSV file provides flight status information segmented by airport continent and gender. It shows how many male and female passengers are in various flight status groups, helping us understand passenger demographics and flight performance by region. The file contains 18 rows of data, where each row represents a count for female and male passengers for the combination of flight status and a specific airport continent.

2. "age_group_counts.csv":
This CSV file presents the distribution of passengers in various age groups. It includes counts of male and female passengers in different age groups, along with the percentage of each gender in each age category, offering insights into the demographics of passengers based on age.
The file contains 6 rows of data, where each row represents an "Age Group."

3. "flight_status_by_month.csv":
This CSV file contains the analysis of flight status by month, providing a month-by-month breakdown of the number of flights that were Cancelled, Delayed, or On Time. This data offers valuable insights into seasonal variations in flight performance.
The file contains 12 rows of data, where each row represents a month.

Source Data File

The source data file for this analysis is "Airline Dataset.csv." This dataset serves as the foundation for the analysis, offering a comprehensive view of passenger and flight data, making it a valuable asset for airlines and aviation professionals.