AirFly Insights: Data Visualization and Analysis of Airline Operations

INTRODUCTION

The objective of this project is to analyse large-scale airline flight data to uncover operational trends, delay patterns, and cancellation reasons using data visualization techniques. The goal is to help understand airline and airport-level performance and contribute to actionable insights using visual analysis.

Week 1: Project Initialization and Dataset Setup

Loaded Flight_delay.csv into a Pandas DataFrame for analysis. Viewed the first few rows using df.head() to understand the dataset structure.

Examined the last 10 rows of the dataset using df.tail(10) to verify data consistency and completeness at the end of the file.

Checked the dataset size, which contains **14,051,979 elements**, giving an idea of the data volume and helping plan memory and performance optimizations.

Explored the dataset shape using df.shape, revealing **484,551 rows and 29 columns**, which provides an overview of the dataset's dimensions.

Examined the **data types** (df.dtypes) and **column names** (df.columns) to understand the nature of each feature and get an overview of the dataset structure, which guided further cleaning and analysis.

Used df.describe() and df.info() to get statistical summaries, data types, non-null counts, and memory usage, providing a quick overview of dataset quality and structure.

Checked for **missing values** (df.isnull().sum()) and **duplicate rows** (df.duplicated().sum()), ensuring data completeness and identifying potential issues before analysis.

Calculated the **minimum, maximum, and average values** of the Distance column to understand the range and central tendency of flight distances in the dataset. Computed the **average arrival and departure delays** to assess overall flight punctuality.

Used groupby to find the **maximum distance per airline** and the **total number of cancellations per airline**, providing airline-level operational insights.

WEEK 2: Preprocessing and Feature Engineering

Reloaded the dataset Flight_delay.csv into a Pandas DataFrame to begin further analysis.

This ensured a fresh working environment and consistency before applying **data preprocessing tasks**.

Checked for **duplicate rows** using df.duplicated().sum() and removed them with df.drop_duplicates(keep='first'), ensuring data consistency and preventing redundancy in analysis.

Verified duplicates again using df.duplicated().sum() to confirm that **0 duplicates remained** after cleaning, ensuring the dataset was free from redundancy.

Checked for **missing values** in each column using df.isnull().sum() to identify data quality issues for preprocessing.

Handled missing values by replacing nulls in the **Org_Airport** and **Dest_Airport** columns with 'unknown', ensuring completeness of categorical data and avoiding errors in analysis.

Converted the **Date** column to datetime format using pd.to_datetime(), enabling time-based analysis and easier handling of temporal trends.

Extracted new time-based features from the Date column, including **Month**, **DayOfWeek**, and **Hour**, to support trend analysis and visualization of delays across different time periods

Created a new feature **Route** by combining the Origin and Dest columns, allowing analysis of flight delays on specific routes.

Saved the cleaned and preprocessed dataset as **Flight_delay_cleaned.csv** for future analysis, ensuring data consistency and reusability.