Flight Project Report

Team Members: Kenzy Essam , Razan Emad El-Din, Areej Sofy , Amany Ashraf, Yomna El Ameer

# Brief About The Dataset:

# The dataset contain information related to airline loyalty program members. Each row represents a unique member, with various attributes describing their membership and flight activity. Here's a brief summary of the key columns:

* **Member ID**: Unique member identifier.
* **FFP Date**: Date joined the frequent flyer program.
* **First Flight Date**: Date of the member's first flight.
* **Gender**: Member's gender.
* **FFP Tier**: Loyalty program tier.
* **Work Location**: Member’s work city, province, and country.
* **Age**: Member's age.
* **Flight Count**: Total flights taken.
* **Bonus Points**: Total bonus points accumulated.
* **Yearly Spend**: Spending in two specific years.
* **Total Distance**: Distance flown in kilometers.
* **Last Flight Date**: Date of the last flight.
* **Days Since Last Flight**: Time since the last flight.
* **Flight Interval**: Average and maximum days between flights.
* **Exchange Count**: Number of points exchanges.
* **Discount**: Average discount received.
* **Total Points**: Points earned from flights and non-flight activities.

# Dataset Issues:

* **Nulls**: Missing values in columns, which can lead to incomplete or biased analysis.
* **Duplicates**: Repeated rows or entries that can skew your results if not removed.
* **Errors**: Incorrect or invalid data that need to be filtered out.
* **Abbreviations**: Column names or values that are shortened or unclear, making the data harder to interpret.
* **Unclear Column Names**: Column names that are not descriptive, leading to confusion about their meaning.
* **Inconsistent Data Types**: Incorrect data types that affect calculations and analysis.
* **Extra Hyphens (-)**: Unnecessary hyphens in the data that need to be removed for consistency.

# Data Cleaning and Preprocessing by Kenzy and Razan and Areej:

* Using Python pandas
* Using jupyter notebook launched by anaconda
* File Reading:  
  Loaded the dataset using read\_csv() function in Python.
* Display Configuration:  
  Enabled full visibility of rows and columns using set\_option() to ensure comprehensive review of data.
* Handling Null Values:  
  Identified null values using isnull().any(), and replaced missing values in the age column with the mean.  
  Converted the age column data type to integer.
* Duplicate Handling:  
  Checked for duplicates using duplicated() and reviewed them.  
  Dropped all duplicates using drop\_duplicates() and verified the dataset again, confirming no duplicates remained.
* Column Renaming:  
  Renamed columns using the .rename() function to standardize column names for clarity.
* Exporting Cleaned Data:  
  Saved the cleaned dataset to a new Excel file using the to\_excel() function, ensuring the first column was retained as the index.

# Data Cleaning by Amany and Yomna:

* Using power query
* Column Renaming:  
  Renamed key columns for better readability:  
  - bp sum → Total Bonus Points  
  - Sum yr 1 → First Year Points  
  - Sum yr 2 → Second Year Points  
  - SEG\_KM\_SUM → Total Flight Distance.

- LAST\_TO\_END → Last Flight Date

-AVG\_INTERVAL → AVG Days Between Flights

- MAX\_INTERVAL → MAX Days Between Flights

- Point\_NotFlight→ Non Flight Points

* Data Type Modifications**:**

Corrected data types for specific columns:

* Converted *Max Days Between Flights* to the appropriate type.
* Fixed errors in the *Last Flight Date* column.
* Formatted *Avg Discount* to a fixed decimal format.
* Additional Data Cleaning:
* Removed null values in the *Gender* column.
* Replaced abbreviations, e.g., "CA" → "California" in the *Work City* column.
* Applied the Text.Proper() function to standardize text formatting in the *Work City* column.
* Rounded values in *Avg Discount* to two decimal places.

# Visualizations by the Team:

* Using Power BI
* Binning Age Data: Grouping Age into bins of size 10 allows for a clearer understanding of age demographics within the dataset.
* Line Charts:  
  - Flight Distance vs. Last Flight Date:  
   Question Answered: How does flight distance vary over time?  
   This chart tracks the trends in flight distances over different periods, helping to identify any seasonal or time-based variations.
* Bar Charts:  
  - Age Distribution by Bins:  
   Question Answered: What is the distribution of travelers across age groups?  
   This bar chart visualizes the count of travelers in each age group, offering insights

-Exchange count by Gender:

Question Answered: What is the distribution of exchange count by gender This bar chart visualizes the Exchange count by Gander

* into the most common age ranges.  
  - Total Points by City:  
   Question Answered: Which cities have the highest total points accumulation?  
   This chart highlights the cities with the most significant total points, identifying areas with high traveler activity.
* Map Visualization:  
  - Flight Count by Country:  
   Question Answered: What is the geographic distribution of flight counts across countries?  
   This map shows which countries have the most frequent travelers, providing a global perspective on travel patterns.
* Scatter Plot:  
  - Age vs. Total Points:  
   Question Answered: Is there a correlation between Age and Total Points earned?  
   This scatter plot helps analyze whether more frequent travelers tend to earn more bonus points, identifying potential relationships between the two variables.
* Gauge Charts:  
  - Total Flight Distance:  
   Question Answered: How close are travelers to reaching their maximum flight distance?  
   The gauge chart provides a clear snapshot of total flight distance, showing progress towards a target.  
  - Average Second Year Points:  
   Question Answered: How do second-year points compare to expected targets?  
   The gauge chart helps visualize how travelers' second-year points fare in comparison to a set goal.
* Area Charts:  
  - Average Flight Count vs. Average Days Between Flights (by month):  
   Question Answered: How do average flight counts and days between flights change over months?  
   This area chart helps analyze trends in flight frequency and intervals between flights, providing a seasonal or time-based view.

-- Count of Member ID by Month :

Question Answered: How does the count of Member IDs change over the months of the year?

This area chart represents how the count of Member IDs goes up and down throughout the year,

* Pie Charts:  
  - Gender Distribution:  
   Question Answered: What is the gender breakdown of travelers?  
   This chart offers a visual representation of gender distribution within the dataset.  
  - FFP Tier by Count (Donut Chart):  
   Question Answered: How are travelers distributed across different FFP (Frequent Flyer Program) tiers?  
   The donut chart helps visualize the proportion of travelers in each tier, aiding in loyalty analysis.
* Traveler Trends:  
  - Travelers Count by FFP Date (Line Chart):  
   Question Answered: How has the number of travelers changed over time based on their FFP date?  
   This line chart tracks the number of new travelers joining the Frequent Flyer Program over time, helping to analyze membership trends.
* Dashboard Cards