Clustering Assignment Report

# 1. Introduction

Clustering is an **unsupervised learning technique** used to group similar observations into clusters.  
In this assignment, we applied clustering on the **EastWestAirlines dataset** to identify different segments of airline customers.  
The objective is to help the airline improve **marketing strategies, loyalty programs, and customer targeting**.

# 2. Dataset Description

* **Source**: EastWestAirlines.xlsx
* **Size**: Several hundred customer records with multiple features.
* **Attributes**: Includes frequent flyer information such as:
  + Balance: Miles balance of the customer
  + BonusMiles: Total bonus miles earned
  + FlightMiles: Miles traveled under frequent flyer program
  + BonusTrans: Number of bonus transactions
  + DaysSinceEnroll: Days since enrollment in loyalty program
* The ID# column was removed as it is non-informative.

# 3. Exploratory Data Analysis (EDA)

* **Missing Values**: No significant missing data found.
* **Summary Statistics**: High variation across features, with some skewed distributions.
* **Distribution Plots**: Features like Balance and BonusMiles showed heavy-tailed distributions.
* **Correlation Heatmap**: Strong correlations between BonusMiles and BonusTrans.
* **Boxplots**: Detected several **outliers** in features such as Balance and BonusMiles.

# 4. Outlier Removal

* Outliers were removed using the **Interquartile Range (IQR) method**:
  + Any value outside [Q1 – 1.5 × IQR, Q3 + 1.5 × IQR] was considered an outlier.
* After removal, the dataset became **cleaner**, reducing skewness and ensuring better cluster separation.

# 5. Data Preprocessing

* **Standardization**: Applied StandardScaler to normalize all features.
* **PCA (Principal Component Analysis)**: Reduced dimensions to 2D and 3D for visualization of clusters.

# 6. Clustering Methods Applied

**6.1 KMeans Clustering**

* **Elbow Method** and **Silhouette Score** used for choosing optimal clusters.
* Best result at **k = 3**.
* Interpretation:
  + **Cluster 1**: Low-spending, less active customers.
  + **Cluster 2**: Moderately engaged customers.
  + **Cluster 3**: High-value frequent flyers.

**6.2 Hierarchical Clustering**

* Used **Ward linkage**.
* Dendrogram suggested **3 major clusters**, consistent with KMeans results.

**6.3 DBSCAN**

* Parameters: eps = 1.5, min\_samples = 5.
* Identified **core clusters** and flagged **noise points** (irregular customers).
* Useful for anomaly detection.

# 7. Results & Insights

* Customers naturally fall into **three distinct groups**.
* **Cluster 3** represents the most valuable customers (frequent travelers, loyal program members).
* **Cluster 1** includes occasional travelers with minimal engagement.
* **Cluster 2** contains medium-usage customers who can be converted into loyal customers with targeted offers.

**Business Implication**:  
Airlines should **retain Cluster 3** with premium loyalty rewards,  
**encourage Cluster 2** with promotional offers, and  
**analyze Cluster 1** to understand reasons for low engagement.

# 8. Conclusion

* Outlier removal improved clustering performance.
* Both **KMeans and Hierarchical clustering** agreed on **3 customer segments**.
* **DBSCAN** helped in detecting anomalies.
* PCA visualization confirmed cluster separability.

Final Recommendation: Focus on **high-value customers (Cluster 3)** while increasing loyalty among **moderate users (Cluster 2).**