**Project Report: Predicting Revenue & Clustering Sales Trends**

**1. Introduction**

The goal of this project is to analyze sales trends and predict revenue using a dataset containing information about Netflix subscribers, revenue, advertisements, and engagement. This project employs both **Supervised Learning (Regression)** to predict revenue and **Unsupervised Learning (Clustering)** to identify sales patterns.

**2. Dataset Description**

The dataset contains the following key features:

* **Subscribers**: Number of active users.
* **Ads Revenue ($)**: Revenue generated from advertisements.
* **Engagement (hrs)**: Total watch time of users.
* **Revenue ($) (Target Variable)**: Total revenue generated.

**3. Methodology**

**3.1 Supervised Learning - Revenue Prediction**

To predict revenue, a **Linear Regression Model** was used:

* **Data Preprocessing**:
  + Missing values were handled by dropping incomplete rows.
  + Data was split into **training (80%)** and **testing (20%)** sets.
  + Features were standardized using **StandardScaler**.
* **Model Training & Evaluation**:
  + A **Linear Regression** model was trained on the dataset.
  + The model was evaluated using **Mean Squared Error (MSE), R-squared Score (R²), and Mean Absolute Error (MAE)**.

**3.2 Unsupervised Learning - Clustering Sales Trends**

To group sales trends, **K-Means Clustering** was applied:

* **Feature Scaling**:
  + The dataset was standardized to ensure equal weightage to all features.
* **Finding Optimal Clusters**:
  + The **Elbow Method** was used to determine the best number of clusters.
* **Applying K-Means**:
  + The dataset was clustered into **three groups (k=3)** based on similarities in sales data.
* **Visualization**:
  + Clusters were plotted to observe patterns in subscribers and revenue.

**4. Results & Analysis**

**4.1 Revenue Prediction Results**

* **Mean Squared Error (MSE):** Indicates the average squared difference between actual and predicted revenue.
* **R-squared Score (R²):** Measures how well the model explains the variance in revenue.
* **Mean Absolute Error (MAE):** Shows the average absolute difference between predictions and actual values.
* **Findings:**
  + A lower MSE and MAE indicate a well-performing model.
  + The R² score provides insights into how much variance in revenue is explained by the features.

**4.2 Clustering Insights**

* **Cluster Analysis:**
  + Sales regions were grouped into three clusters based on revenue, subscribers, and engagement trends.
  + The clustering visualization helped identify high-revenue vs. low-revenue patterns.

**5. Conclusion & Future Work**

* The **Linear Regression Model** successfully predicted revenue with reasonable accuracy.
* **K-Means Clustering** identified meaningful patterns in sales trends.
* **Future Improvements**:
  + Testing other regression models (e.g., Random Forest, Gradient Boosting) for better accuracy.
  + Applying advanced clustering techniques like **DBSCAN or Hierarchical Clustering**.
  + Adding more features, such as user demographics, for deeper analysis.

This project demonstrates the power of **Data Science** in revenue forecasting and business decision-making. The insights from this study can help optimize marketing strategies and maximize profits for businesses like Netflix.

**Prepared by:** Dhinakara