CLUSTERING AND FITTING ANALYSIS OF COUNTRY DATA

Md Atikuzzaman

Student ID: 23082277

GitHub Repository: https://github.com/Atikuzzaman101/Clustering-and-Fitting

Introduction:

This report is the analysis of country-wise data conducted using the techniques of clustering and regression. The aim is to cluster countries on the basis of their economic and geographical characteristics and predict GDP as a function of population and area. This exercise shows how clustering reveals patterns and regression makes forecasts. The dataset includes variables such as population, area, and GDP. K-means clustering has been applied for clustering countries and prediction has been done with a linear regression model. Various plots that convey the valuable information extracted from this analysis have been constructed for explanation purposes.

Data Preparation:

The dataset includes information about countries, such as population, area, and GDP. Before analysis, the following steps were performed:

1. **Data Cleaning:** Missing values in the "GDP ($, millions)" column were replaced with the median value.
2. **Normalization:** Numerical data, including population, area, and GDP, were scaled to a range of [0, 1] to ensure fair clustering.

Clustering Analysis

Objective: Group countries based on similarities in population, area, and GDP.

1. **Methodology**:

   - The k-means algorithm was selected to do the clustering.
   - The **Elbow Method** was used for the determination of the suitable number of clusters, which was found to be 3.

2. **Visualization**:

    - A scatter diagram to present the clusters was created, whereby distinct colors represent each cluster. The diagram highlighted groups of countries with similar traits such as GDP and population.

Regression Analysis

Objective: Predict GDP based on population and area.

1. **Model Design**:

    - A linear regression model was trained using Population and Area as independent variables to predict GDP.
    - The model learned the relationships of these features with GDP and did predictions on unseen data.

2. **Performance Evaluation**:

    - The model performance was at its best with a Mean Squared Error of, hence reflecting its efficiency in predicting GDP.

3. **Visualization**:

    - A scatter plot of actual GDP values against model predictions shows a great coincidence.

Results and Insights

1. **Clustering Results**:

    - Countries were divided into three clusters, each with unique economic and geographical characteristics.
    - The clustering visually separated countries of high GDP from smaller economies, thus making patterns more interpretable.

2. **Regression Results**:

    - The regression model successfully predicted GDP based on population and area.
    - Predictions closely matched actual values, confirming the model's reliability.

Visualizations

1. **Elbow Plot**: Demonstrates the optimal number of clusters determined using the Elbow Method.
2. **Cluster Scatter Plot**: Illustrates groups of countries based on population and GDP.
3. **Regression Scatter Plot**: Shows a comparison between actual GDP values and model predictions.

Conclusion

This analysis has shown the application of clustering and regression techniques in understanding country-level data. The k-means clustering was effective in grouping countries with similar characteristics, while the regression model was able to provide quite accurate predictions of GDP. These methods offer valuable insights for data-driven decision-making in fields like economics and policy planning.