K-Means Clustering of Tax Rate and Average Monthly Income

Name : Dhanyasree kanchrala

Email ID : dk24abc@herts.ac.uk

Student ID : 23070191

**Git hub Link :** https://github.com/dk24abc/Clustering-and-Fitting

Introduction

The financial aspects of different countries' income, cost of living, and various spending categories have been documented in this dataset. These consist of 500 entries that span several years. The important metrics are `Average Monthly Income`, which averages $4,291.25 with a standard deviation of $2,179.22, while the average of `Cost of Living` is $3,716.23. Housing costs take about 34.97% of income, while tax rates come at an average of 22.4%. Other major cost percentages are savings, 14.93%; healthcare, 12.38%; and transportation, 12.48%. The dataset ranges from approximately 2011, and it is close to zero, showing a reasonably even distribution of the data. Kurtosis values, though, indicate a slight peakiness. This data shall prove valuable in understanding the financial condition and cost structures in various regions.

## Visualizations

1. **Heat map Correlation**
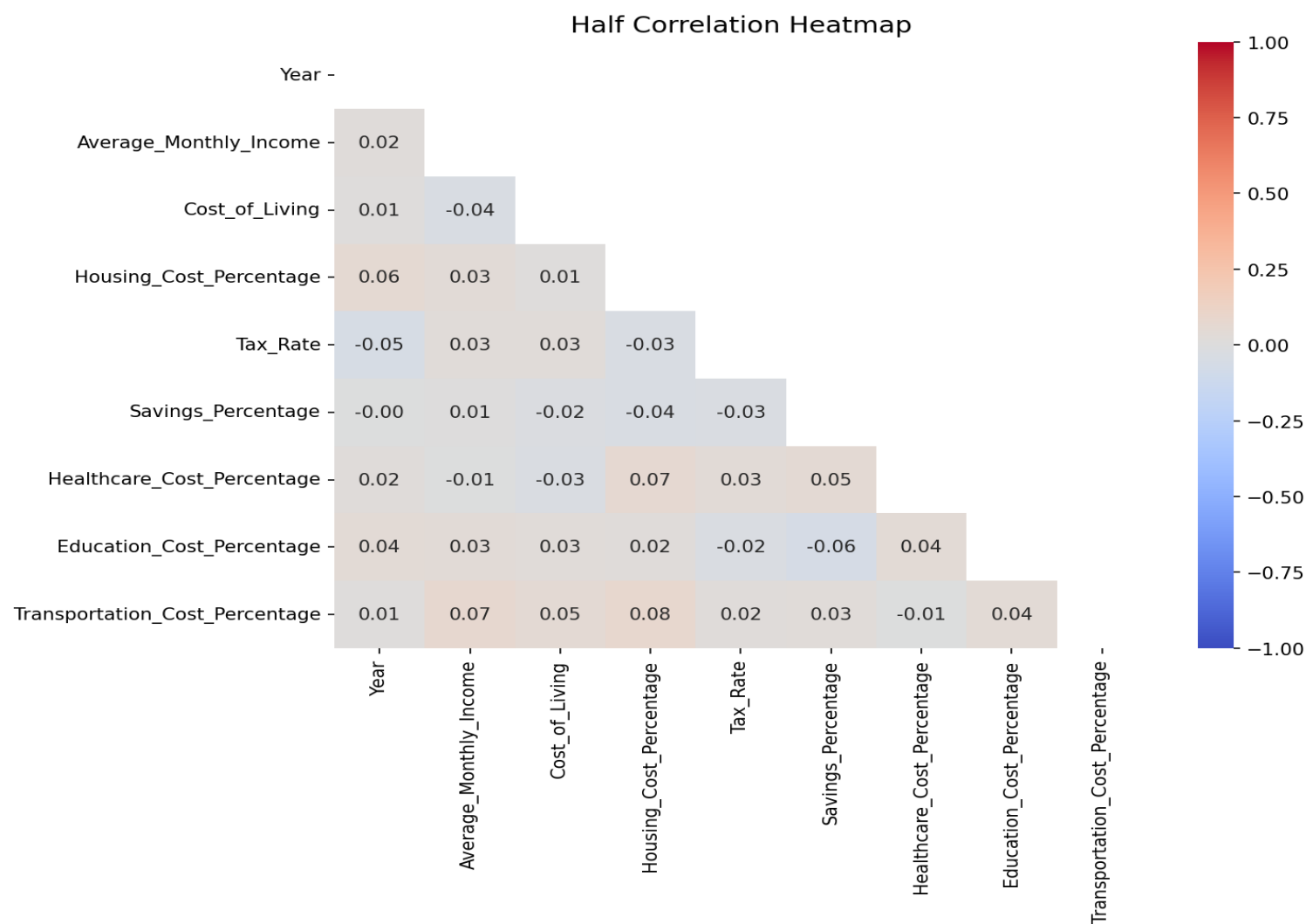
The half correlation heatmap visualizes the relationship of financial variables such as Year, Average Monthly Income, Cost of Living, and various cost percentages. It reveals very weak positive correlations between Year and both Average Monthly Income (0.02) and Cost of Living (0.01), with a weak negative correlation between Average Monthly Income and Cost of Living at -0.04. There is also a 0.08 weak positive correlation between Housing and Transportation Cost Percentages. The following heat map shows the interaction of these financial factors.

2 . Bar Chart

This bar chart, shows the average of the monthly income in a country like Australia, India, Russia, South Africa, Brazil, China, Japan, Canada, Germany, Mexico, United States, and France. The x-axis represents the countries, while the y-axis represents the average monthly income, ranging from 0 to 5000. Bars colored in shades of blue to red are used, and black error bars show the variability or uncertainty in the data. This chart compares clearly the average incomes across different countries, highlighting great differences and variability in the level of income.

Linear Fit Analysis
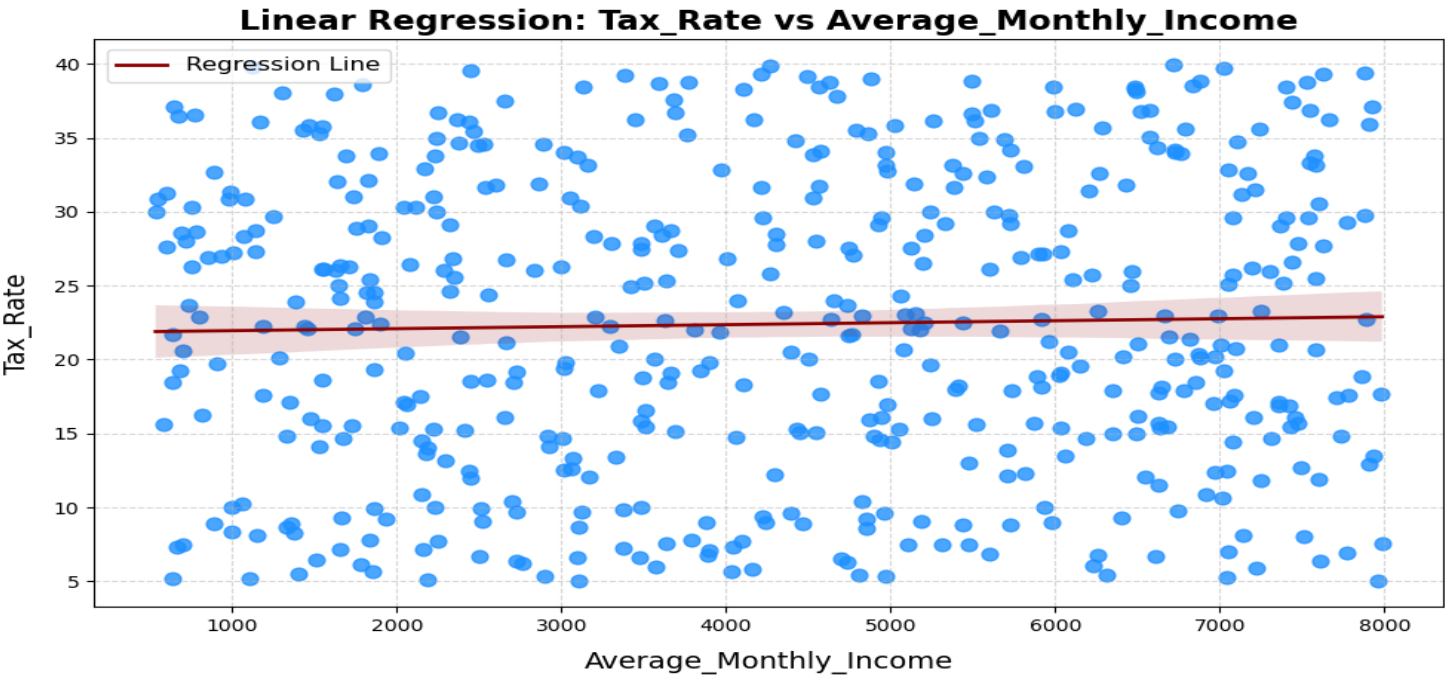
This scatter plot, "Linear Regression: Tax Rate vs Average Monthly Income," illustrates the relationship between average monthly income (x-axis) and tax rate (y-axis). The blue dots reflect the individual data values for each observation. The resulting red regression line shows that there is a slight positive correlation in these data; as average monthly income increases, the tax rate tends to increase. The confidence interval is the shaded area around the regression line and represents the uncertainty in the regression estimate. This plot gives a good insight into the relationship between income and tax rate.

Elbow Plot and Silhouette

Elbow Plot:

This is the elbow curve for the best k in k-means clustering, which helps in deciding on the best number of clusters for any dataset. The x-axis shows the number of clusters, ranging from 2 to 10, while the y-axis reflects the Within-Cluster Sum of Squares (WCSS). The plot comprises a line with data points showing the WCSS values for each k. A red 'X' identifies the position at k=4, which points to the best number of clusters. The "elbow" in k=4 is such a position where the rate of decrease in WCSS gets slowed down and therefore, k=4 is the best number of clusters for this data set.

.**Silhouette Score**: Best Cluster Configuration : The highest **4 clusters silhouette score = 0.42**, which indicates excellent cluster separation

Elbow Curve for Optimal k

K-Means Clustering
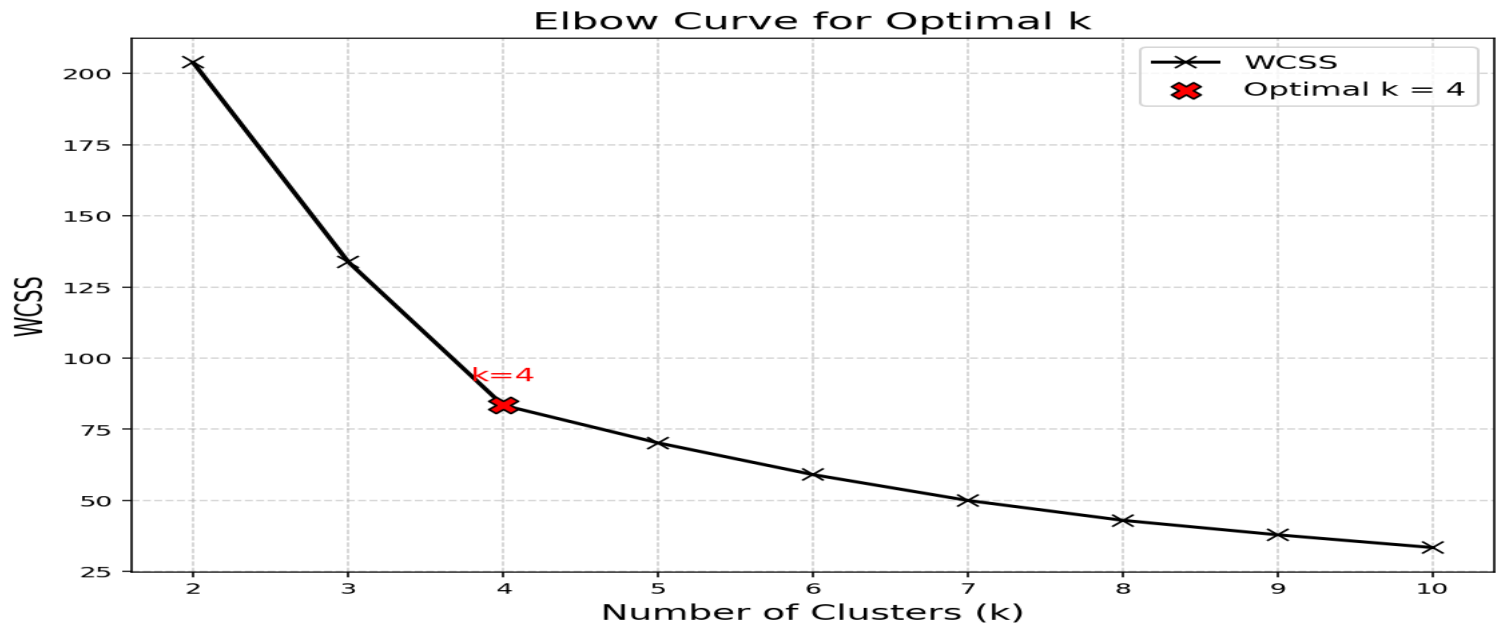
K-Means Clustering with k=4
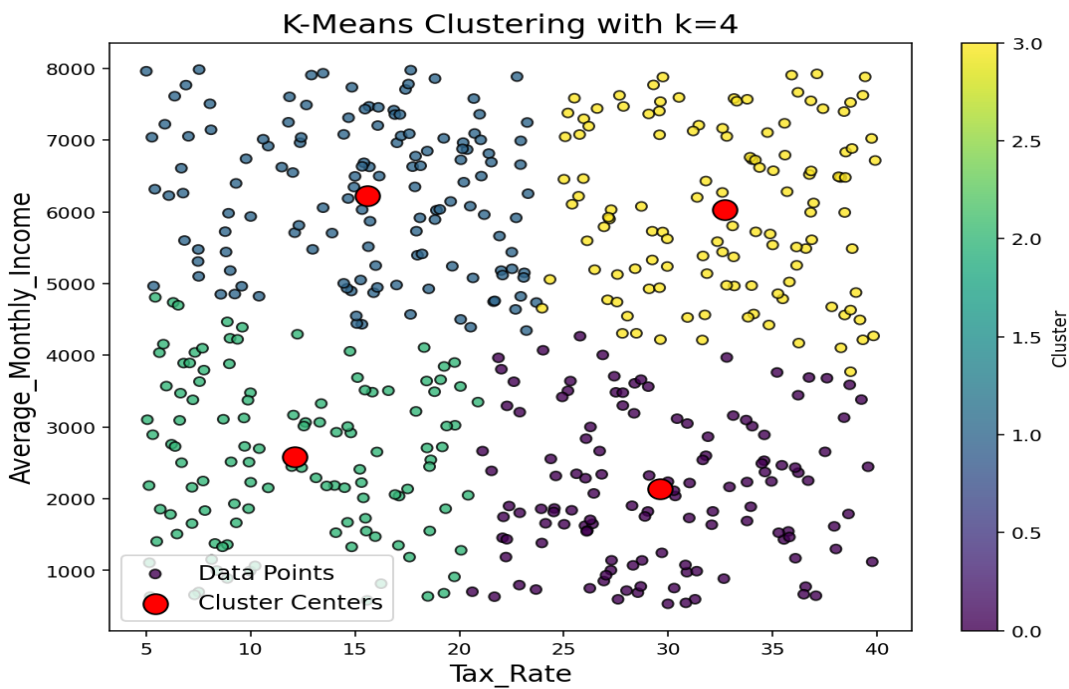
This scatter plot, "K-Means Clustering with k=4,". shows the relationship between Tax Rate (x-axis) and Average Monthly Income (y-axis). Data points are color-coded into four clusters: blue, green, yellow, and purple. Red circles indicate the cluster centers, which mark the average values for tax rate and income for each cluster. The plot shows how the K-Means algorithm has grouped this data into four clear clusters based on these variables. This helps in understanding the distribution of tax rate and income, further helping in identifying central trends in the dataset.

Conclusion:

The dataset depicts a broad perspective of the financial conditions in various countries, which brings into view the underlying facts about income, cost of living, and tax rates. The statistical analysis gives evidence of the relationship between income and tax rates: with an increase in income, the rate of tax also goes up. The clustering based on income and tax rates identified four different groups, hence giving an elaborative understanding of the financial trends across the regions. Such a comparison of average monthly income across countries encompasses large global differences, hence offering a meaningful background to financial planning and policy making