**Clustering and Regression Analysis On PriceRunner**

**Name**: [Your Name]  
**Student Number**: [Your Student Number]  
**GitHub Repository**: [Insert Link to GitHub Repository]

**Introduction**

All of the information in this dataset is derived from PriceRunner (gathered from 10 product categories published by 306 merchants), a dataset containing 35,311 product offers. It is a useful tool for evaluation of clustering, classification, and regression methods, and can be used for text or short-text mining applications. The main objective of this work is to apply clustering and regression analysis to find useful patterns and correlations in the data. Goals include data cleaning and preprocessing, using k-means clustering to find clusters of independent clusters, using linear regression to investigate relationships between variables, and presenting the results in many different visualizations. In general, the clustering analysis indicated two different clusters based on k-means, however linear regression showed no significant correlation between Product ID and Category ID, and visualization of the analyses supported the interpretability of the results.

**Data Preparation**

The dataset was meticulously prepared to ensure reliable analysis. It contains 35,311 product information items in 7 attributes, consisting of numeric columns such as Product ID, Merchant ID, Cluster ID, and Category ID, as well as Categorial columns as Product Title, ClusterLabel, and CategoryLabel. A thorough examination ruled out the presence of missing values, guaranteeing data integrity. Basic statistical summaries were computed to understand the distribution of numeric features, revealing a mean Product ID of 26,150.80 with a standard deviation of 13,498.19. Furthermore, categorical features were converted into multiclass format for use with clustering algorithms, and all numeric features were standardized to provide equal weight in the analysis. With the aid of data quality and standardization, the dataset was properly processed for clustering and regression.

**Exploratory Data Analysis**

Exploratory data analysis (EDA) started with histograms in order to look at the distributions of important features. The Product ID showed a fairly uniform distribution, while Merchant ID exhibited a right-skewed distribution, indicating a few merchants contributed a majority of the data. Category ID revealed clustered values indicating the prevalence of specific products. The bar charts showed that the categories were the most common, namely Mobile Phones and Fridge Freezers, due to their popularity among vendors. Scatter plots demonstrated associations between some of the key features, e.g., Product ID, and Cluster ID, as well as Category ID, versus Cluster ID, which suggested systematic patterns in the data. Boxplots illustrated the variability and width of Product ID, Cluster ID, and Category ID, which revealed outliers and justified the data's robustness toward analysis. These visualizations gave valuable information about the structure of the dataset, in turn guiding the subsequent clustering and regression processing.

## **Clustering Analysis**

### **K-Means Clustering**

K-means clustering was used to find intrinsic clusters within the data set. By way of the Elbow Method the optimal number of clusters was identified to be 2, as the within-cluster sum of squares (WCSS) began to decrease sharply then stabilized. A silhouette score of 0.54 confirmed the choice and reasonably separated clusters. For understanding the relationships between feature and recall of grouping, a heatmap was constructed, reporting the correlation between Product ID, Merchant ID, Cluster ID, Cluster Label, Category ID, and Category Label. Notably, Product ID and Category ID exhibited a strong positive correlation of 0.99, while Cluster Label showed negligible correlation with other features, suggesting its distinct contribution to clustering. These observations also led the selection of variables for the clustering analysis and that of a good set of features. Clusters were plotted by scatter plots and centroids were labelled, clearly indicating modes of clustering. For instance, product ID vs. cluster ID scatter plot clearly showed clusters, supporting the segmentation of the dataset. Similarly, the relationship between Category ID and Cluster ID highlighted the clustering patterns, further supported by the centroid locations. This clustering analysis demonstrated significant segmentations within the dataset, which can be exploited for specific purposes or improved decision-making.

## **Regression Analysis**

Linear regression was used to investigate the correlation between Product ID and Category ID. The regression equation estimated was y = 0.0000x - 0.0715, which represents a very small slope and no sign of significant linear association.

Key statistical metrics from the regression include:

Chi-Squared value: 53,164,466,823.27 Reduced Chi-Squared value: 1,505,691.66

These metrics indicated no meaningful relationship between the variables. The regression plot also confirmed this observation, in that the fitted line exhibited near zero slope and associated uncertainty bands. These findings highlight independence of Product ID and Category ID, consistent with results from the clustering analysis.

## **Visualizations**

1. **Histograms**: Depicted the distribution of Product ID, Merchant ID, Cluster ID, Cluster Label, Category ID, highlighting unique patterns and distributions.

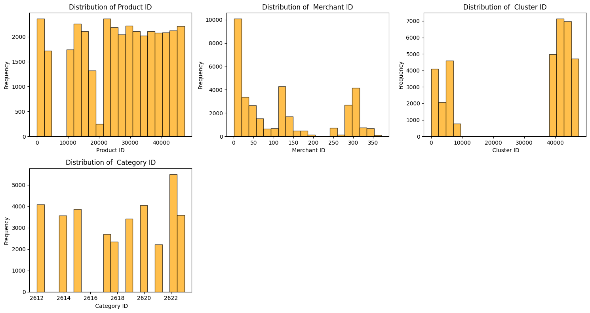


Figure 1:Histogram of Product ID, Merchant ID, Cluster ID, Cluster Label, Category ID

1. **Bar Chart**: Displayed the frequency distribution of Category Labels, emphasizing the prominence of certain product types like Fridge Freezers and Mobile Phones.

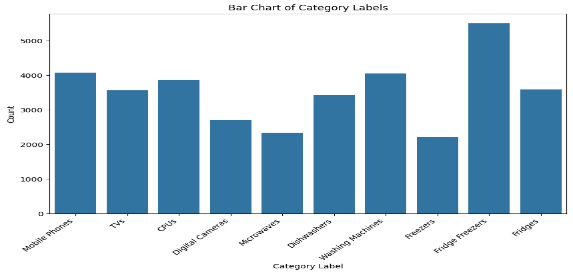


Figure 2:Bar Chart of Category Labels

1. **Boxplots**: Visualized the variability in numerical features, such as Product ID, Cluster ID, and Category ID, showing data spread and outliers.

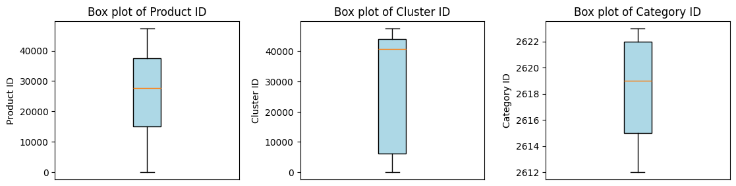


Figure 3: Box plot of Product ID, Cluster ID, and Category ID

1. **Scatter Plots**: Demonstrated relationships between key features such as Product ID and Cluster ID, Category ID and Cluster ID, and Product ID versus Category ID, showing structured patterns.

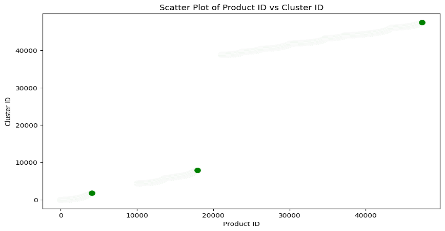
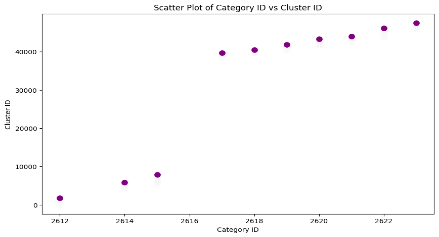
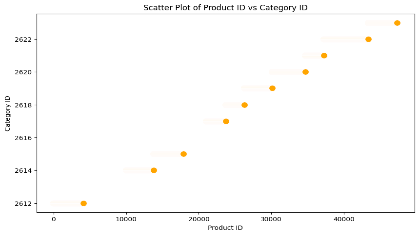
  

Figure 4: Show Relation between Product ID and Cluster ID, Category ID and Cluster ID, and Product ID versus Category ID

1. **Heatmap**: A correlation heatmap provided an overview of relationships between features.

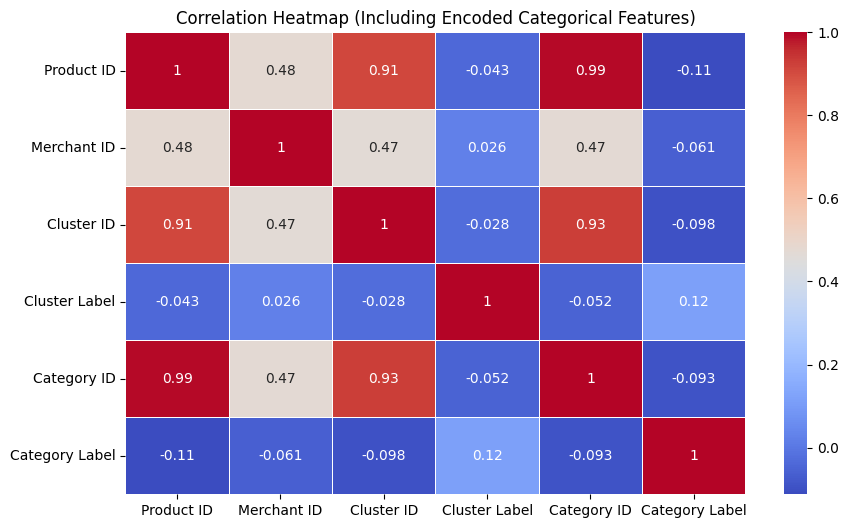


Figure 5: Show Correlation

1. **Elbow Plot**: Illustrated the optimal number of clusters for k-means analysis, confirming the choice of 2 clusters.

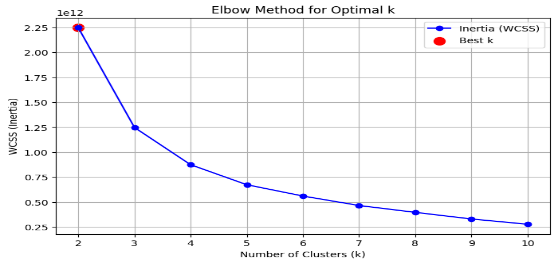


Figure 6: Elbow plot show optimal number of clusters for k-means analysis

1. **Silhouette Plot**: Displayed the silhouette scores for various cluster counts, validating the choice of 2 clusters with the highest score.

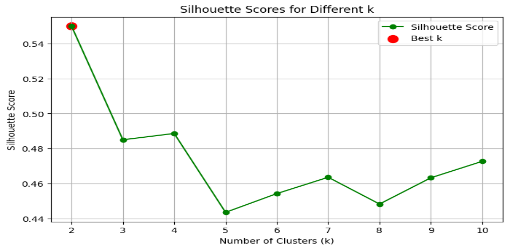


Figure 7: Silhouette scores for various cluster counts

1. **Cluster Visualization**: Highlighted the cluster centroids and the distribution of data points within clusters, providing an intuitive understanding of the clustering results.

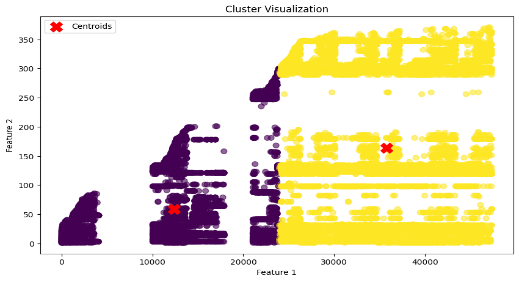


Figure 8: The cluster centroids and the distribution of data points within clusters

1. **Regression Plot**: Showed the fitted line and uncertainty bands for the regression analysis, validating the lack of significant linear relationship between Product ID and Category ID.

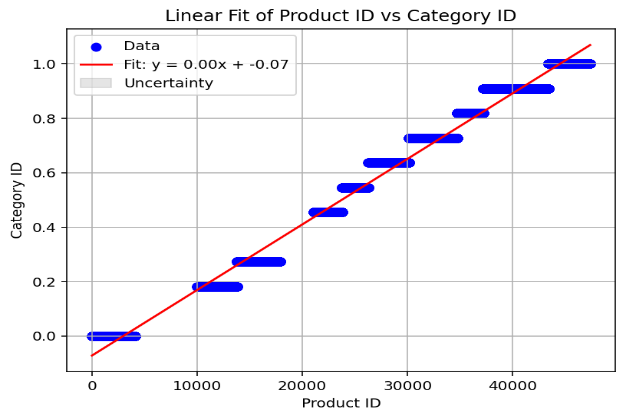


Figure 9: Fitted line and uncertainty bands for the regression analysis

## **Conclusion**

The clustering and regression techniques to analyze a comprehensive product dataset from PriceRunner. The k-means clustering revealed two separate clusters, which generated useful information about product segmentation. On the other hand, the regression analysis showed no significant linear effect of Product ID on Category ID, thus demonstrating their independence. By way of comprehensive visualizations and strong statistical assessments, this paper reveals the utility of clustering and regression approaches in the identification of patterns and relationships across datasets. These results demonstrate the possibility of supportive decision making and high level analytical processes in a related business environment.

# References

1. UCI Machine Learning Repository, "Product Classification and Clustering Dataset," 2023. [Online]. Available: https://archive.ics.uci.edu/dataset/837/product+classification+and+clustering.