Clustering Analysis of USD/INR Exchange Rate Data Using K-means

Unit 5

Avinash Bunga

Master of Science in Information Systems and Business Analytics

Park University

CIS607HOF1P2024 Applied Business Forecasting

Professor: Dr. Abdelmonaem Jornaz

September 15, 2024

Clustering Analysis of USD/INR Exchange Rate Data Using K-means

Introduction

This analysis aimed to use K-means clustering to identify patterns in the USD/INR exchange rate data from 2003 to 2021 and demonstrate predictive capabilities using new data. Clustering is a powerful technique in predictive analytics that helps group data points based on their similarities. By identifying clusters within the dataset, this analysis aims to uncover underlying patterns that could assist in understanding market behavior and aid in financial decision making (Kaggle, n.d.).

Dataset Selection and Input Variables

Dataset Selection: The dataset chosen for this analysis consists of daily USD/INR exchange rate data from 2003 to 2021, sourced from Yahoo Finance. This dataset was selected because it is relevant to the financial domain and aligns well with the aim of identifying patterns in currency exchange rates. The data includes key exchange rate values such as Open, High, Low, and Close for each trading day.

Input Variables: The input variables used for the clustering analysis were:

- **Open**: The exchange rate at the beginning of the trading session.
- **High**: The highest exchange rate during the trading session.
- **Low**: The lowest exchange rate during the trading session.
- **Close**: The exchange rate at the end of the trading session.

Reasoning for Selection: These variables were chosen because they provide a comprehensive overview of daily trading activity and capture the key points of exchange rate fluctuations. By including Open, High, Low, and Close, the analysis captures both the volatility and the overall trend of the exchange rates, which are critical for clustering the data meaningfully (Kaggle, n.d.; Yuan et al., 2022).

Data Preparation

Data Cleaning: Initial data cleaning involved handling missing values, dropping unnecessary columns, and standardizing the data to ensure all variables contribute equally to the clustering process. Standardization was performed using the StandardScaler to bring all variables to a similar scale, making them comparable for distance calculations in the clustering algorithm.

Data Split: The dataset was randomly divided into two chunks: 80% data for training the model and 20% data for testing. This split was essential for evaluating the model's ability to generalize and predict the cluster membership of new, unseen data points (Agarwal, n.d.).

Methodology

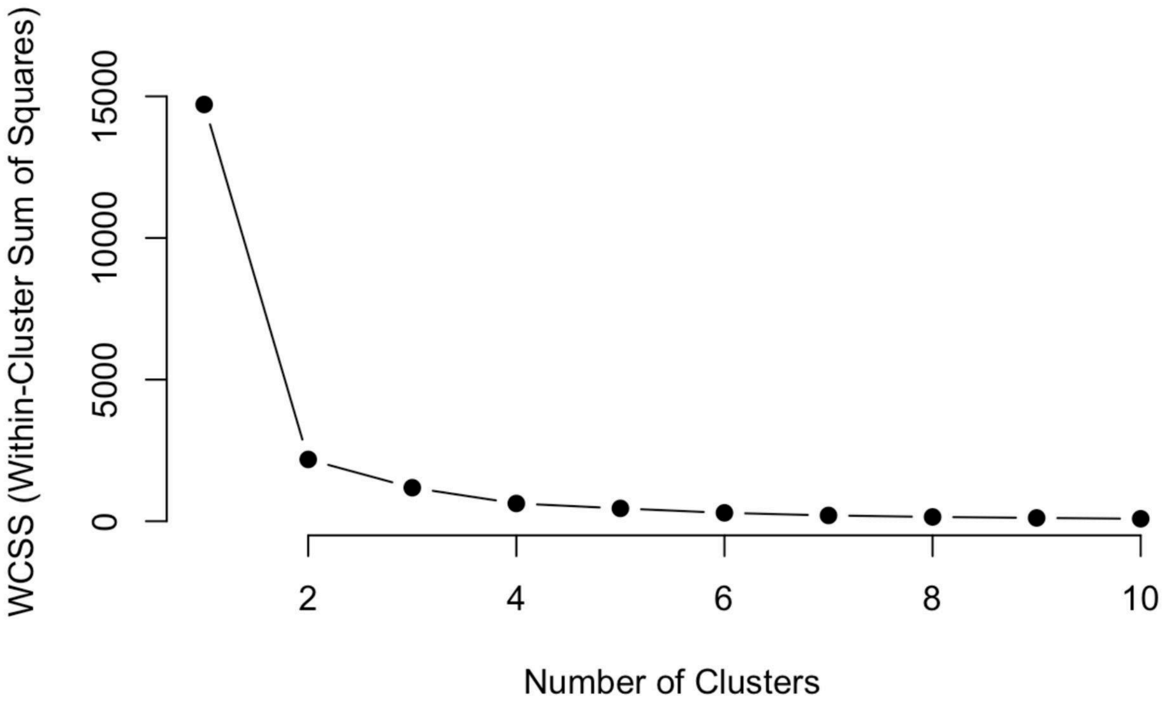
Scaling: To ensure the features contribute equally to the clustering analysis, the data was scaled using the StandardScaler. This technique standardizes the data by centering each feature around the mean and scaling it according to the standard deviation (Mulani, 2022).

Clustering Analysis: K-means clustering was applied to the training data to identify groups with similar characteristics. The K-means algorithm assigns data points to clusters by minimizing the within cluster variance, making it effective for grouping data based on feature similarities (Anello, 2023).

Choice of Clusters: The Elbow method was used to determine the ideal number of clusters. The Elbow plot, as shown in **Figure 1**, shows the Within Cluster Sum of Squares (WCSS) against the number of clusters. The point where the curve bends, known as the "elbow," suggests the optimal number of clusters. Based on this method, three clusters were chosen for further analysis (Bobbitt, 2022).

Figure 1 shows the Elbow plot used to determine the optimal number of clusters. The elbow point at three clusters indicates the diminishing returns of adding more clusters beyond this point.

Figure 1: Elbow Method for Optimal Clusters



Clustering Analysis Results

Table 1 below shows the centroids of the three clusters identified by the K-means algorithm. Each centroid represents the average values of the Open, High, Low, and Close columns within each cluster.

Table 1: Centroids of the Clusters Identified by K-means

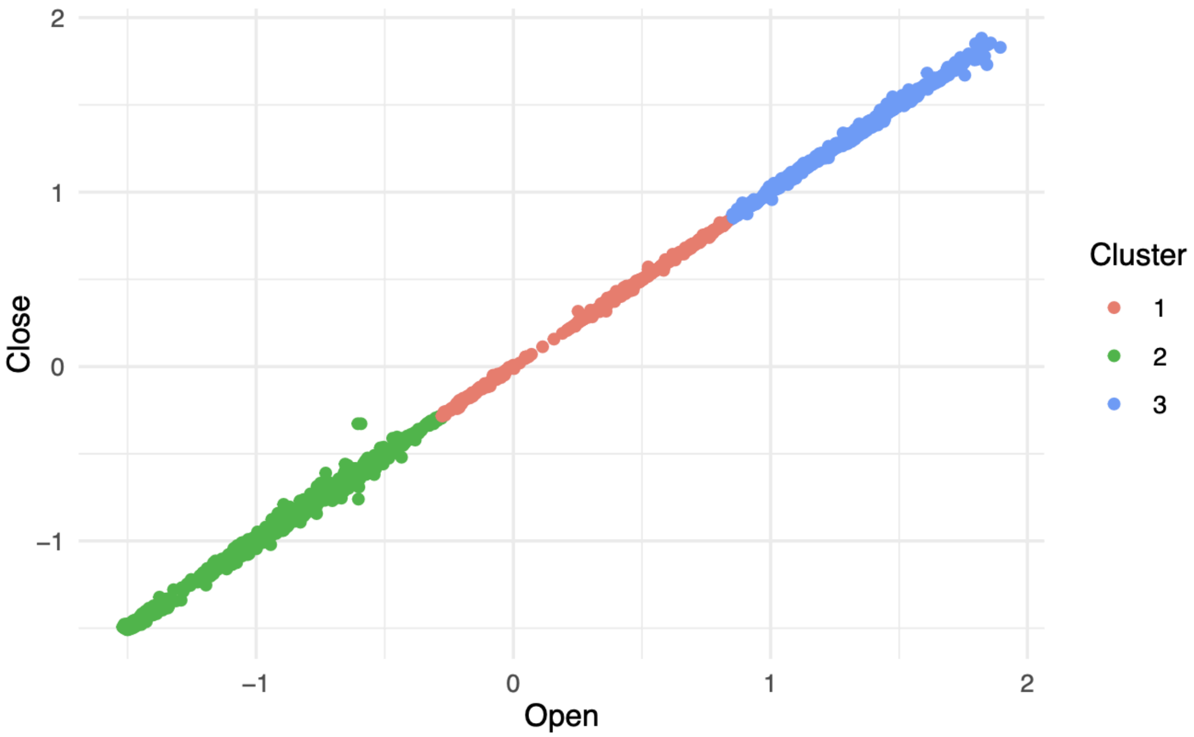
Cluster	Open	High	Low	Close
1	0.4101404	0.4108888	0.4090473	0.4098997
2	-0.9640201	-0.9640081	-0.9639211	-0.9641544
3	1.2654113	1.2646949	1.2662584	1.2658644

Interpretation:

- **Cluster 1**: Represents days with exchange rates moderately above the mean, indicating relatively stable periods.
- **Cluster 2**: Represents days with lower-than-average exchange rates, capturing periods of lower market activity.
- **Cluster 3**: Represents days with the highest exchange rates relative to the mean, showing periods of significant market movement.

Figure 2: visualizes the clusters formed on the training data using K-means clustering. The scatter plot displays data points categorized into three distinct clusters based on their Open and Close values. The colors represent different clusters: Cluster 1 (red), Cluster 2 (green), and Cluster 3 (blue). This visualization illustrates how the K-means algorithm effectively grouped the scaled data points, highlighting the distinct patterns identified in the training data (Sarah, n.d.).

Figure 2: K-means Clustering of Scaled Training Data



Predictive Analytics Results

Table 2 below shows the first five records from the testing dataset, with their predicted cluster assignments. These records were not used in cluster identification, demonstrating the predictive capability of the K-means model.

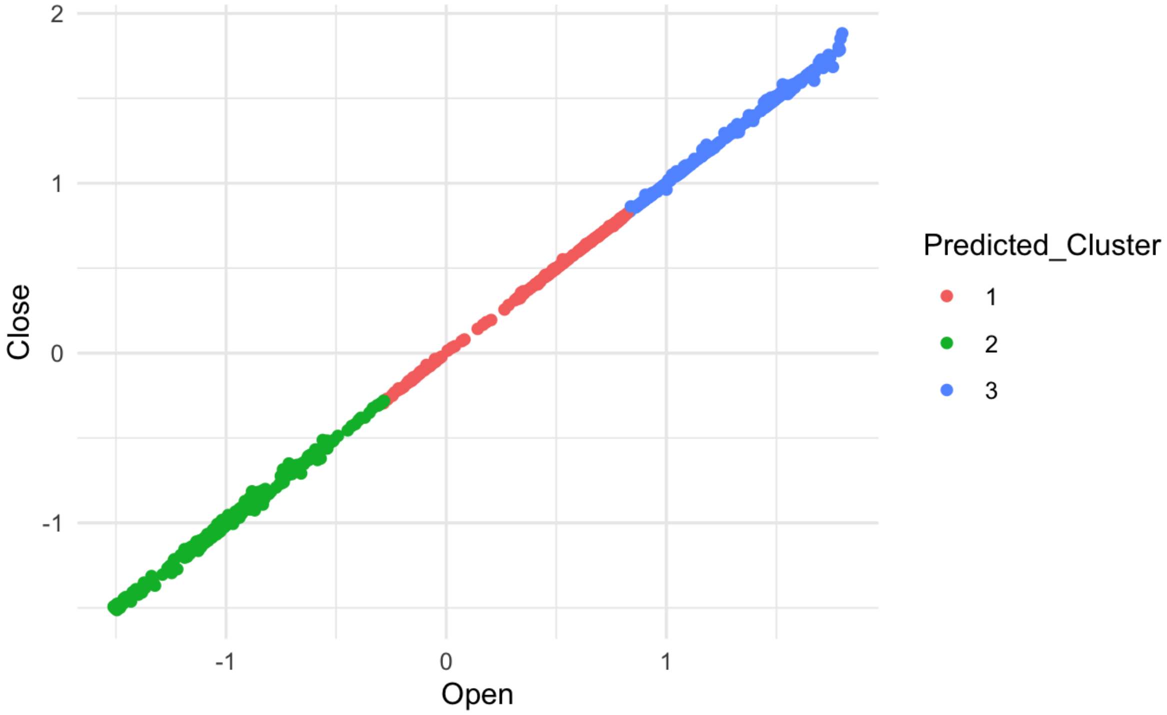
Table 2: Predicted Cluster Assignments for Testing Data

Open	High	Low	Close	Date	Predicted Cluster
1.7	1.72	1.72	1.73	2020-05-27	3
0.532	0.569	0.524	0.532	2013-11-25	1
-0.592	-0.586	-0.59	-0.592	2012-02-10	2
-1.12	-1.11	-1.11	-1.1	2005-01-20	2
1.47	1.46	1.44	1.5	2018-11-13	3

Interpretation: The results show that each new data point was successfully assigned to the nearest cluster, demonstrating the model's ability to generalize and predict cluster membership for unseen data.

Figure 3 presents the visualization of the predicted clusters for the testing data. This scatter plot shows how each test data point is assigned to its nearest cluster based on the Open and Close values. The colors represent different clusters: Cluster 1 (red), Cluster 2 (green), and Cluster 3 (blue), indicating that the K-means model effectively groups new data points based on their similarities (Sarah, n.d.).

Figure 3: Predicted Clusters for Testing Data



Conclusion

This analysis successfully demonstrated using K-means clustering to identify distinct patterns in the USD/INR exchange rate data. The model effectively grouped data into meaningful clusters based on exchange rate behavior and showed predictive power when applied to new, unseen data points. This approach can be valuable for financial analysis, helping to identify market trends and support decision making in currency exchange.

References

Agarwal, M. (n.d.). *Pythonic Data Cleaning With pandas and NumPy*. RealPython. Retrieved September 15, 2024, from https://realpython.com/python-data-cleaning-numpy-pandas/

Anello, E. (2023). *K-Means Clustering in R Tutorial*. Datacamp. Retrieved September 15, 2024, from https://www.datacamp.com/tutorial/k-means-clustering-r

Bobbitt, Z. (2022, September 8). *How to Use the Elbow Method in R to Find Optimal Clusters*. Statology. https://www.statology.org/elbow-method-in-r/

kaggle (n.d.). *US Dollar / INR Rupee Dataset(2003-2021)*. Kaggle. Retrieved September 15, 2024, from https://www.kaggle.com/datasets/meetnagadia/us-dollar-inr-rupee-dataset20032021

Mulani, S. (2022, August 3). *Using StandardScaler() Function to Standardize Python Data*. Digitalocean. https://www.digitalocean.com/community/tutorials/standardscaler-function-in-python

Sarah, M. (n.d.). *A Comprehensive Guide to Cluster Analysis: Applications, Best Practices and Resources*. DisplayR. Retrieved September 15, 2024, from https://www.displayr.com/understanding-cluster-analysis-a-comprehensive-guide/

Yuan, S., Roover, K. D., & Deun, K. V. (2022, September 9). *Simultaneous clustering and variable selection: A novel algorithm and model selection procedure*. SpringerLink. https://link.springer.com/article/10.3758/s13428-022-01795-7