Clustering Exchange Rates

CS305: Machine Learning
Victoria Lu

Problem and Motivation

Hierarchical clustering and K-means clustering are two common clustering approaches in machine learning to put datapoints into groups based on how similar they are. In this project, I attempt to explore these approaches as ways to assess the similarity between different currencies relative to the dollar.

Many currencies are considered to trade together due to connections in the underlying economies. One way to assess these relationships is to look at the correlation between the two exchange rates (both relative to a different currency, say the dollar) over time based on a linear regression approach. However, this approach assumes a linear and constant relationship between the currencies. However, exchange rate dynamics are influenced by a wide array of factors such as economic policies, geopolitical events, and market sentiment, and it is unlikely that all the underlying relationships are linear. Therefore, I wanted to see if clustering can better capture the nuances of exchange rate dynamics.

Data Description, Cleaning, and Visualization

Raw data

The raw dataset Foreign Exchange Rates from Kaggle contains the daily average exchange rate of 51 currencies where dollar is the quote currency from 01-01-1995 to 11-04-2018 (e.g., the Euro column is how much U.S. Dollar you would have to pay to buy 1 Euro). Many columns do not have any observations in the early years. Given this, as well the fact that economies may undergo significant structural changes over time, I opted to only focus on about 6 years of data, from 2013-01-01 to 11-04-2018.

Data Anomalies, Missingness, and Imputation

Upon closer examination of this subsample, I noticed that a few columns, including 'Bahrain Dinar', 'U.A.E. Dirham', 'Libyan Dinar', 'Qatar Riyal', 'Rial Omani', and 'Saudi Arabian Riyal' have constant values throughout, so I dropped these columns. With the remaining columns, there are 3,446 missing currency-year combinations. I chose to further drop 'Iranian Rial' because almost 30% of its time-series is missing. For the other currencies, I imputed the missing values with linear interpolation using two adjacent temporal datapoints for each missing value.

Scaling and reshaping

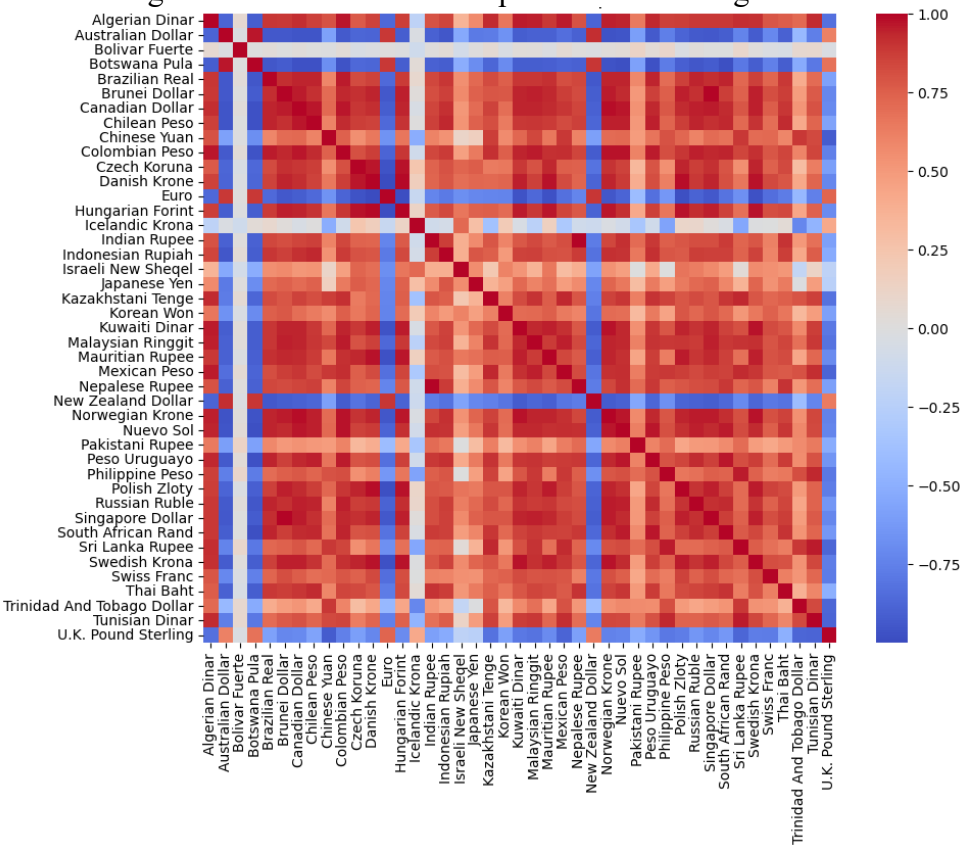
Since different exchange rates have different magnitudes, I scaled all the columns. To perform clustering, I further reshaped the data so that each row corresponds to a currency and the daily exchange rates are treated as features that characterize each currency. The cleaned dataset has dimension 43 rows/currencies x 1,273 columns/days.

Visualizations

Before applying clustering, I visualized the correlations between the currencies using a heatmap (see Figure 1). A few key observations can be noted from Figure 1. First, there is a much larger number of red entries compared to blue entries, which indicate that most currencies are

positively correlated as opposed to negatively correlated with other currencies. In addition, the currencies that correspond to mostly blue entries/are negatively correlated with other currencies are mainly developed market currencies such as the Euro, Australian Dollar, New Zealand Dollar, and U.K. Pound Sterling. Furthermore, the Bolivar Fuerte and Icelandic Krona are two currencies that have minimal correlation with other currencies overall.

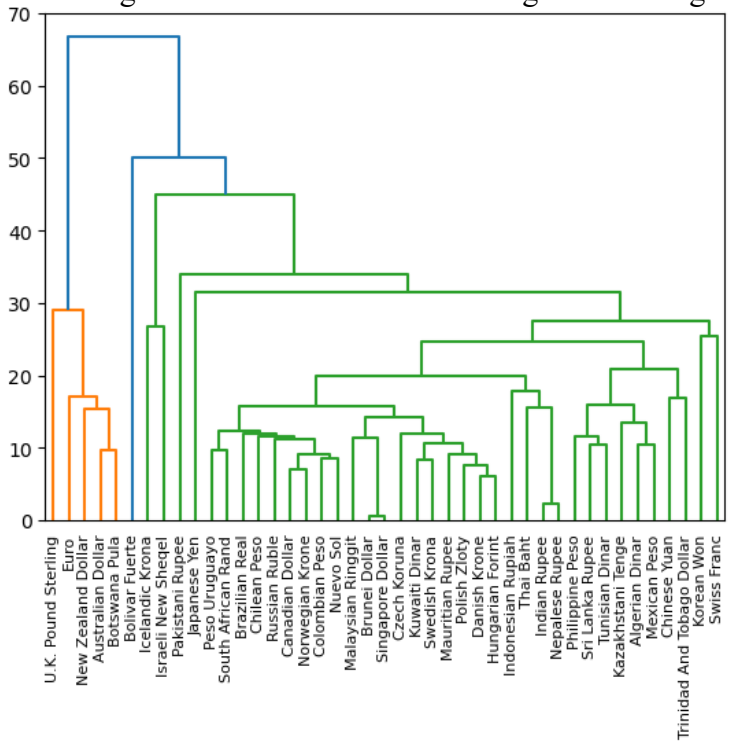
Figure 1: Correlation Heatmap between Exchange Rates



Methodology and Results

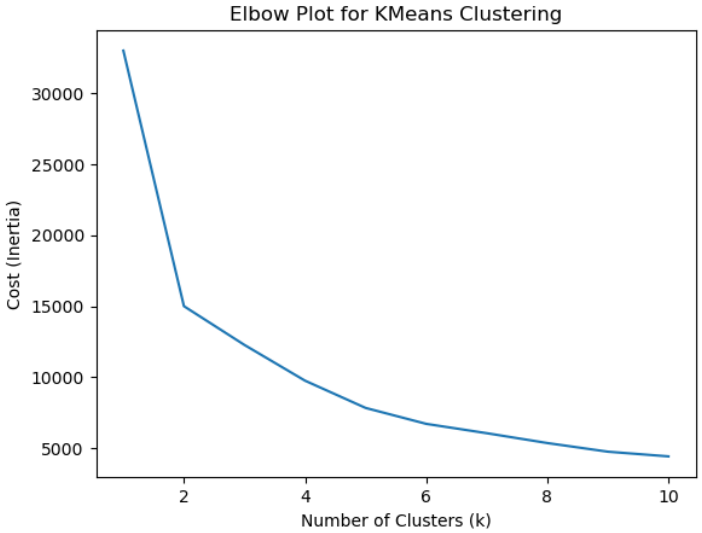
Using the cleaned dataset, I first performed hierarchical clustering. The algorithm works by assigning each point to its own cluster, and then repeatedly merging the two closest clusters based on Euclidean distance until all points are in one cluster. Figure 2 shows the dendrogram produced using the average linkage approach. For the most part, the branches agree with the correlation results, where currencies that are mostly negatively correlated with other currencies belong to one branch and currencies that are mostly positively correlated to each other belong to another branch.

Figure 2: Dendrogram for Hierarchical Clustering with Average Linkage



I also tried out K-means clustering, which works by randomly assigning each point to one of k clusters, calculating the mean of each of the k clusters, assigning each point to the cluster with the closest mean, and repeating the previous three steps until convergence. Based on the elbow plot (see Figure 3), which shows the within-cluster sum of squares (SSD) using different k values, at 2 clusters, adding more clusters no longer significantly reduces the SSD, though there is still some decrease from 4 to 6 clusters. To gain more intuition, I chose k to be 4 or 6. Tables 1 summarizes the results, which mostly agree with the hierarchical clustering dendrogram.

Figure 3: Elbow Plot for K-Means Clustering



Tables 1: K-Means Clustering Results

4-clusters	6-clusters
Cluster 1: Algerian Dinar, Brazilian Real, Brunei Dollar, Canadian Dollar, Chilean Peso, Colombian Peso, Czech Koruna, Danish Krone, Hungarian Forint, Indian Rupee, Indonesian Rupiah, Japanese Yen, Korean Won, Kuwaiti Dinar, Malaysian Ringgit, Mauritian Rupee, Nepalese Rupee, Norwegian Krone, Nuevo Sol, Peso Uruguayo, Polish Zloty, Russian Ruble, Singapore Dollar, South African Rand, Swedish Krona, Swiss Franc, Thai Baht Cluster 2: Australian Dollar, Botswana Pula, Euro, New Zealand Dollar, U.K. Pound Sterling Cluster 3: Icelandic Krona, Israeli New Sheqel Cluster 4: Bolivar Fuerte, Chinese Yuan, Kazakhstani Tenge, Mexican Peso, Pakistani Rupee, Philippine Peso, Sri Lanka Rupee, Trinidad And Tobago Dollar, Tunisian Dinar	Cluster 1: Australian Dollar, Botswana Pula, Euro, New Zealand Dollar, U.K. Pound Sterling Cluster 2: Algerian Dinar, Chinese Yuan, Kazakhstani Tenge, Malaysian Ringgit, Mexican Peso, Pakistani Rupee, Philippine Peso, Sri Lanka Rupee, Trinidad And Tobago Dollar, Tunisian Dinar Cluster 3: Bolivar Fuerte Cluster 4: Brunei Dollar, Czech Koruna, Danish Krone, Hungarian Forint, Korean Won, Kuwaiti Dinar, Mauritian Rupee, Polish Zloty, Singapore Dollar, Swedish Krona, Swiss Franc Cluster 5: Brazilian Real, Canadian Dollar, Chilean Peso, Colombian Peso, Indian Rupee, Indonesian Rupiah, Japanese Yen, Nepalese Rupee, Norwegian Krone, Nuevo Sol, Peso Uruguayo, Russian Ruble, South African Rand, Thai Baht Cluster 6: Icelandic Krona, Israeli New Sheqel

A few interesting observations can be noted from the hierarchical clustering and K-means clustering results:

1) Overall, the developed market currencies Australian Dollar, Euro, New Zealand Dollar, and U.K. Pound Sterling are very similar to each other. However, it is not very clear why the Botswana Pula is more similar these currencies than the South African Rand (the currency is pegged to a basket of currencies with the South African Rand having the highest weight).
2) Even though I clustered using k=4 and 6, the two main branches in Figure 1 and the elbow plot of K-Means clustering both suggest that dividing the currencies into two clusters with the currencies in the previous point in one cluster and the rest of the currencies in another cluster is sufficient.
3) The Icelandic Krona and Israeli New Sheqel were clustered together when using both 4 and 6 clusters. Because the underlying economies share few similarities, they may have been clustered together simply because they are very different from all the other economies.
4) Cluster 4 with 4-Means clustering and Cluster 2 with 6-Means clustering are mainly comprised of currencies of Asian emerging economies. It's surprising that the Mexican Peso is more similar to these currencies compared to other LATAM currencies like the Brazilian Real and Colombian Peso.
5) Most of the currencies of countries in Central Europe are fairly similar to each other (see Cluster 4 when using 6 clusters). More interestingly, developed market currencies like the

Singapore Dollar, Swiss Franc, and Swedish Krona are more alike these currencies compared to other developed market currencies.

6) Cluster 5 when using 6 clusters is perhaps the hardest to interpret as it is comprised of a range of different currencies. It contains most of the LATAM currencies (e.g., Brazilian Real, Chilean Peso, and Colombian Peso), three emerging market currencies and developed market currencies respectively, as well as a few other currencies. A few similarities shared by the underlying economies include they are generally quite export-oriented and that many are heavily dependent on commodities.

Limitations and Extensions

Since clustering is a form of unsupervised machine learning technique, there is no definitive way to assess how accurate the clusters are, so these techniques can only serve as a way to offer more intuition. With K-Means clustering specifically, there is also some randomness in the results given how the initial two clusters were assigned. Additionally, existing research has shown that hierarchical and K-Means clustering algorithms are biased towards forming spherical clusters as opposed to clusters with irregular shapes. Therefore, if certain currencies are very similar to each other but are distributed somewhat irregularly, they may not be put in the same cluster. It is also hard to assess to what extent this was an issue as the feature space is not 2D or 3D and so one cannot visualize the distribution of the datapoints.

Most of the limitations outlined above are inherent to clustering algorithms and are hard to improve upon. As an extension, one may try to see if using data from a shorter timeframe yields very different results. Restricting to a shorter timeframe means that there are less likely to be major structural changes in the underlying economies, which may reduce the degree of irregularity in the distribution of the datapoints.

Works Cited

Currency exchange rates. (n.d.). Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/datasets/thebasss/currency-exchange-rates

Jin, C., & Malthouse, E. (2016). On the bias and inconsistency of K-means clustering. https://www.scholars.northwestern.edu/en/publications/on-the-bias-and-inconsistency-of-k-means-clustering