# Report on CryptoClustering

## Introduction

This project aims to analyze how cryptocurrencies are affected by 24-hour and 7-day price changes using unsupervised learning techniques. Specifically, the goal is to perform clustering using K-Means and Principal Component Analysis (PCA) on cryptocurrency market data to identify distinct groups and patterns within the data.

## Data Preparation

The first step involved loading the `crypto_market_data.csv` file into a pandas DataFrame. Once loaded, the data was normalized using the `StandardScaler` from scikit-learn to ensure that all features have a mean of 0 and a standard deviation of 1. This step is crucial for ensuring that the clustering algorithm performs optimally.

The normalized data was then used to create a new DataFrame, with the "coin_id" column from the original DataFrame set as the index. This allowed for easier referencing and plotting of the data points.

## Finding the Best Value for k

To determine the optimal number of clusters (k), the elbow method was employed on both the original scaled data and the PCA-transformed data. The elbow method involves the following steps:

1. Creating a list with k values ranging from 1 to 11.
2. Computing the inertia (sum of squared distances of samples to their closest cluster center) for each k value.
3. Plotting the inertia values to visually identify the "elbow point," where the rate of decrease in inertia slows down.

The plots revealed that the best value for k is 4. This means that the data can be effectively grouped into 4 distinct clusters.

## Clustering with K-Means

### Using Original Data

With the optimal k value determined, the K-Means model was initialized and fitted using the original scaled data. The model predicted the clusters for each cryptocurrency, and these predictions were added as a new column to the original DataFrame. A scatter plot was then created using hvPlot, with the x-axis representing the 24-hour price change percentage and the y-axis representing the 7-day price change percentage. The plot was colored based on the predicted clusters, and the "coin_id" column was included in the hover information to identify each cryptocurrency.

### Using PCA Data

Next, PCA was performed to reduce the features to three principal components. The explained variance of the three components was found to be about 89.5%, indicating that these components capture 89.5% of the total variance in the data. This significant reduction in dimensionality simplifies the data while retaining most of the important information.

The K-Means model was then reinitialized and fitted using the PCA data. The predicted clusters were added to the PCA DataFrame, and another scatter plot was created, this time with the x-axis representing the first principal component (PC1) and the y-axis representing the second principal component (PC2). The plot was colored based on the predicted clusters, and the "coin_id" column was included in the hover information.

## Analysis and Comparison

Composite plots were created to compare the clustering results from the original data and the PCA data. The comparison revealed that the PCA-transformed data provided a clearer separation of the clusters. This improved clarity is attributed to the reduction in dimensionality, which helps in mitigating the curse of dimensionality and making the clusters more distinct and easier to interpret.

## Conclusion

The elbow method was used to find that the best value for k is 4 for both the original and PCA-transformed data, indicating that the optimal number of clusters does not differ between the two data sets. The total explained variance of the three principal components was about 89.5%, meaning that these components capture 89.5% of the total variance in the data.

The PCA-transformed data, which uses fewer features, provides a clearer separation of the clusters, making the distinctions between different groups more evident. This project successfully demonstrated the use of unsupervised learning techniques to analyze cryptocurrency market data. By applying PCA, the data's dimensionality was significantly reduced, leading to more distinct and interpretable clusters. The visualizations created using hvPlot provided clear insights into the clustering results, highlighting the impact of using fewer features for clustering.