LAB 5: CLUSTERING KNN

OCTOBER 22

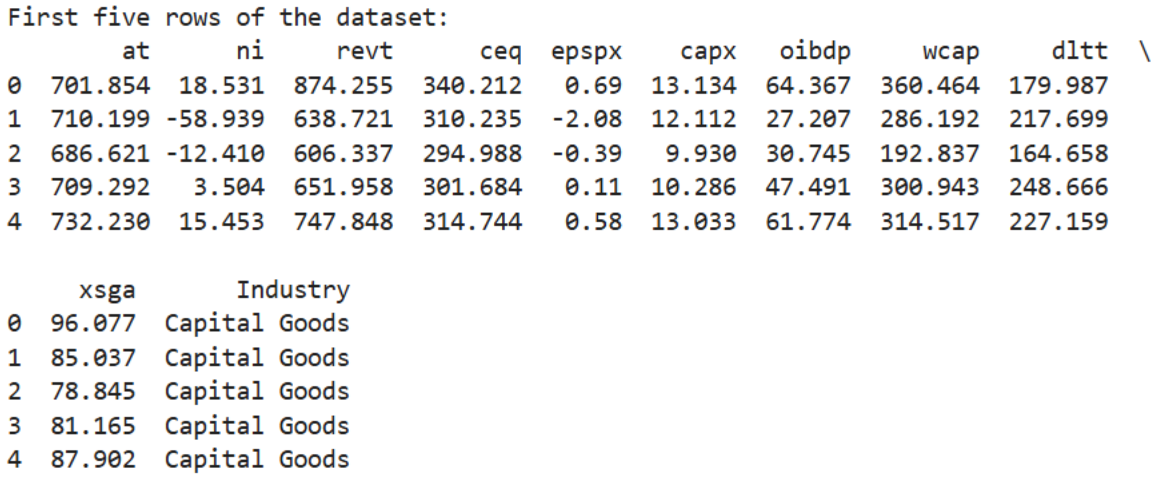
BSAN6070 INTRODUCTION TO MACHINE LEARNING

Authored by: Andres Sebastian Gaibor Heredia

1. Clustering KNN

We will continue to use our financial dataset which comprises 26,154 companies from fiscal year 2000-2024. We are dropping the categorical variables that have missing values as we reviewed, they are not crucial for the analysis. The resulting dataframe contains 235,996 records and 301 fields. The target variable will be “ggroup” which is the Global Industry Classification Group for the companies. The explanatory variables will be “at” Total Assets, “ni” Net Income, “revt” Total Revenue, “ceq” Total Common/Ordinary Equity, “epspx” Earnings Per Share (Basic) Excluding Extraordinary Items, “capx” Capital Expenditure, “oibdp” Operating Income Before Depreciation, “wcap” Working Capital, “dltt” Long-Term Debt and “xsga” Selling General & Administrative Expenses.

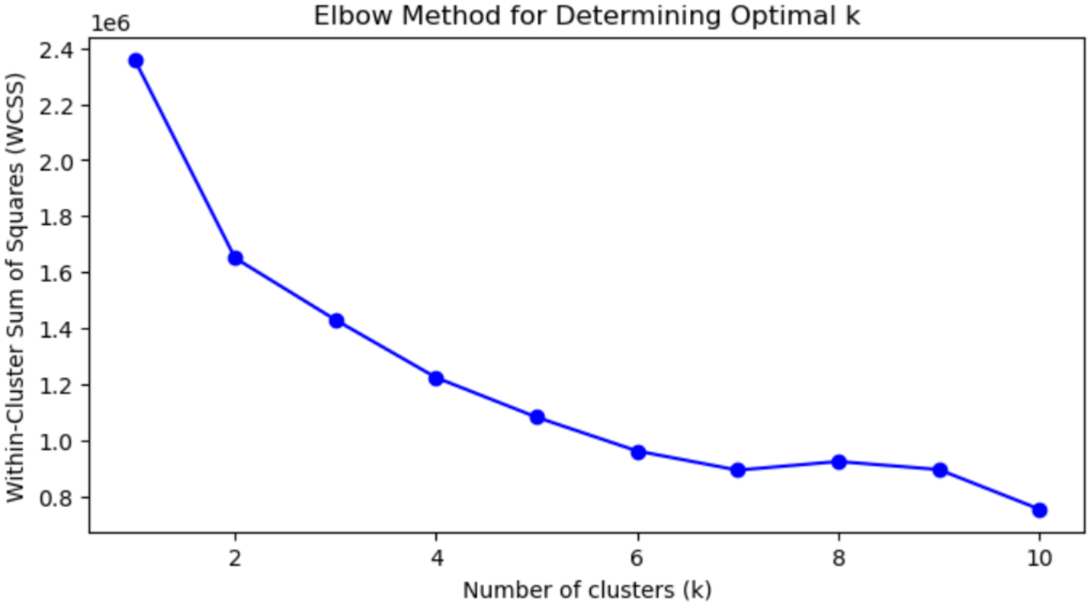
We define X as a matrix with the chosen explanatory variables and y as the field “ggroup”. Then we map the respective industry names and add them into our knn_data dataframe with the rest of the explanatory variables. The resulting dataframe is shown below.



```
First five rows of the dataset:
        at       ni     revt      ceq  epspx    capx   oibdp     wcap     dltt  \
0  701.854   18.531  874.255  340.212   0.69  13.134  64.367  360.464  179.987
1  710.199  -58.939  638.721  310.235  -2.08  12.112  27.207  286.192  217.699
2  686.621  -12.410  606.337  294.988  -0.39   9.930  30.745  192.837  164.658
3  709.292    3.504  651.958  301.684   0.11  10.286  47.491  300.943  248.666
4  732.230   15.453  747.848  314.744   0.58  13.033  61.774  314.517  227.159

     xsga        Industry
0  96.077  Capital Goods
1  85.037  Capital Goods
2  78.845  Capital Goods
3  81.165  Capital Goods
4  87.902  Capital Goods
```

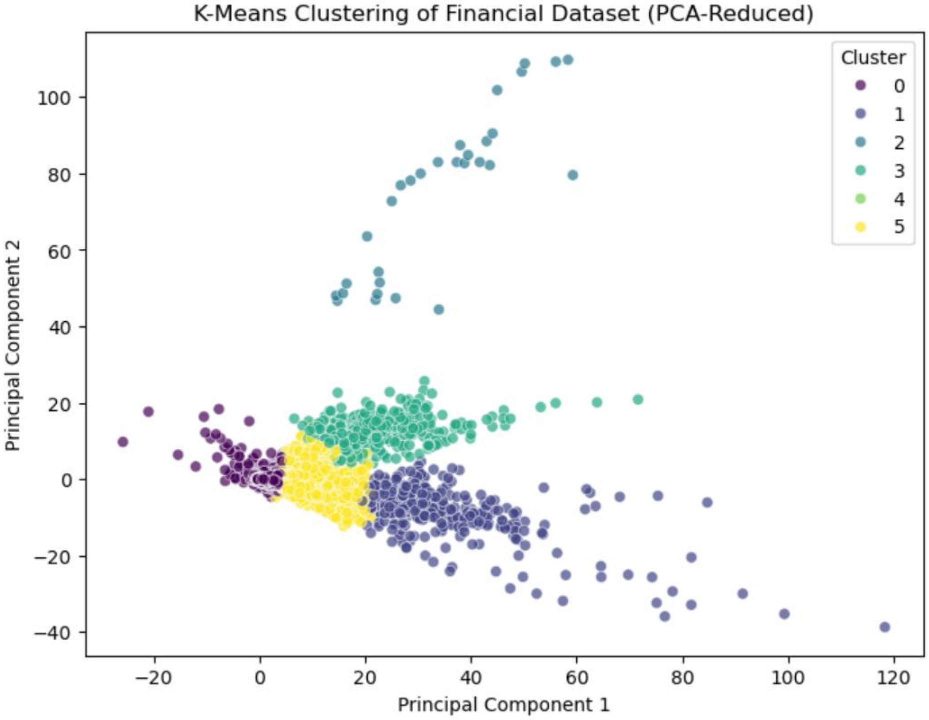
After defining knn_data we standardized the features of the matrix X using StandardScaler() for a better clustering performance. Afterwards we determine the optimal number of clusters using the Elbow Method and a random state = 42.



The Elbow Method is used to determine the optimal number of clusters (k) for K-Means Clustering. It does so by plotting the Within-Cluster Sum of Squares (WCSS) against different values of k. The goal is to find the "elbow point", where increasing k no longer significantly reduces WCSS.

The Y-axis represents Within-Cluster Sum of Squares (WCSS), which measures how close the data points are to their cluster centroids. Lower values indicate better clustering.
The X-axis represents the number of clusters (k).
The curve shows a sharp decrease in WCSS initially, then starts to level off as k increases. As we can see on the plot the elbow point appears to be on clusters k = 7 so we choose the previous number of clusters to optimize the clustering analysis. The number of clusters is then k = 6.

We fit the scaled data using kmeans.fit and predict cluster assignments with kmeans.predicted. Then we plot the visualization of the clustering using PCA reduction to the first two principal components.



This scatter plot represents the results of K-Means clustering applied to the financial dataset, reduced to two principal components using PCA (Principal Component Analysis). Each point represents a data instance, and different colors indicate clusters.

Key Observations

Cluster Distribution: The dataset is divided into 6 clusters (0 to 5). Some clusters (e.g., yellow and purple) are densely packed, indicating well-defined groups. Other clusters (e.g., light blue at the top) have widely spread points, suggesting possible outliers or high variance.
Separation of Clusters: The clusters show good separation, but there are some overlaps.
Cluster 2 (Blue) and Cluster 3 (Teal) Overlap. These two clusters blend into each other, indicating that the separation might not be optimal.

Outliers in the Upper Region: A small group of points (top-right) seems isolated from the main clusters. These could be misclassified points, high-variance data, or true anomalies.

Key Takeaways

- PCA Effectively Reduces Dimensionality: The dataset, originally in higher dimensions, is now visualized in two principal components, capturing most of the variance.
- K-Means Captures Distinct Groups: The model successfully identifies clusters, with clear separation for most.
- Potential for k=5 Instead of k=6: Some clusters are too close, suggesting that k=6 may not be ideal.
- Investigate Outliers: The points far from the main cluster centers should be examined separately. These could be true anomalies or data artifacts