Week 3.1
Using Python Tools for Unsupervised Learning and Feature Engineering
(K-Means Clustering)

Produced By: Abdullah Abdul Majid*(linkedin)*

Completion Date: July 19, 2024

Email: abdullahpisk@gmail.com

Jupyter Notebook Links:

GitHub

Table of Contents:

Problem Statement:

Learn to apply K-means clustering to segment customer data, analyze patterns without predefined labels and plot the result.
Dataset Used: Mall Customer Segmentation Data from Kaggle.

Purpose:

The purpose of this assignment is to inspect and preprocess the dataset. Then use the tools from the scikit-learn library to apply K-means clustering on the data and use Elbow plots to determine the optimal amount of clusters. Analyze the resulting clusters and produce a scatter plot with appropriate feature pairs.

Process:

Dataset Loading and Initial Inspection

- Imported all the necessary libraries, i.e pandas, seaborn, matplotlib and sklearn.
- Downloaded the Mall Customer Segmentation Data from Kaggle.
- Loaded the dataset into a pandas dataframe (loan) using *read_csv* on the *train.csv* file.
- Verified the dataset loading using print.

Inspecting and Preprocessing the Dataset

- Inspected the dataset using *info* and *describe*.

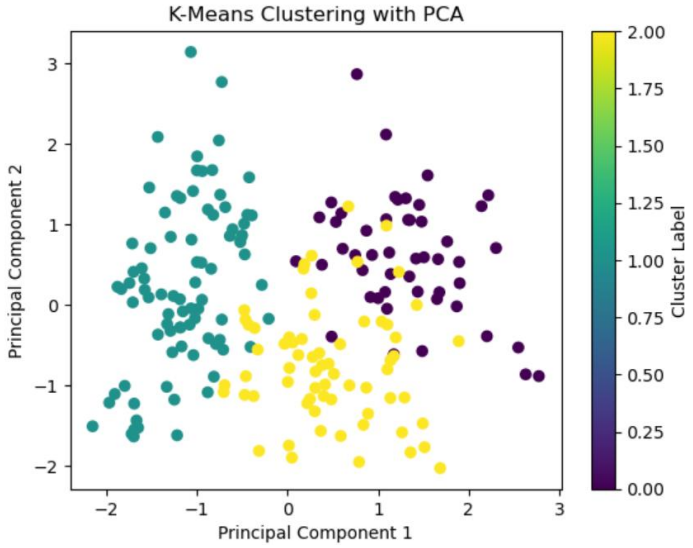
- No duplicates were found when *isnull()* was used.
- No null values were found.
- Used Label Encoding to numericize the Gender column.
- Dropped the useless CustmerID column.
- Used boxplots to check for potential outliers, none were found.

Building and Fitting the K-Means Clustering Model

- Selected the columns for clustering
- Standardized the dataframe using StandardScaler, this is important since the values are on different scales.
- Initialize K-Means with arbitrary initial K=3.
- Next up, we train the model by using *fit.*

Plot

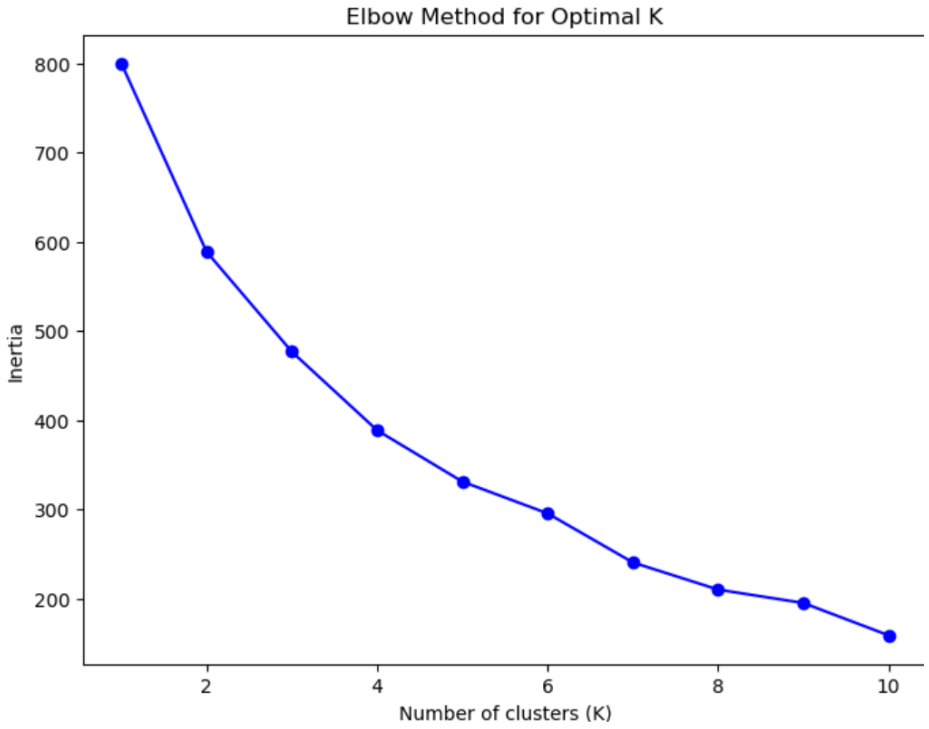
- Used Seaborn ScatterPlot:
-



Not enough clusters.

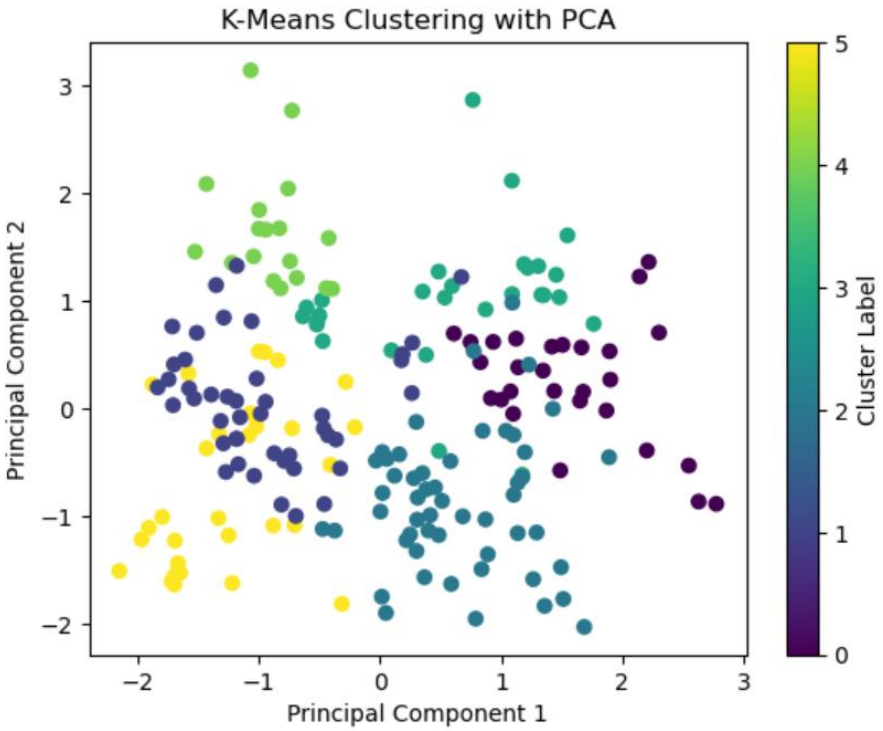
Using Elbow Plot to get K

- Selected K range 1-11.
- Computed K-Means for each value of K.
- Produced a plot:



K-Means Again

- We saw a major jump at K=6, so we selected that.
- Fit the K-Means model again with the new K value.
- New Scatter Plot:


K-Means Clustering with PCA
Principal Component 2
Principal Component 1
Cluster Label
3
2
1
0
−1
−2
−2
−1
0
1
2
3
5
4
3
2
1
0