

Danielle Van Sant

D212 – Data Mining II

Task 1

Dr. Middleton

## **Part I: Research Question**

Is it possible to use K-Means clustering to segment customers into groups based on number of children and income?

The goal of this data analysis is to segment customers according to number of children and income. K-means clustering is a popular unsupervised [machine learning algorithm](https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/) used for partitioning a dataset into a pre-defined number of clusters. The goal is to group similar data points together and discover underlying patterns or structures within the data.

## **Part II: Technique Justification**

The clustering analysis tool I selected is KMeans clustering, KMeans clustering is especially advantageous for seeing patterns in the data that may not be visible to the naked eye. KMeans is good for large datasets where a human may not be able to see groups of data.

All clusters are assumed to have the same variance by K-means. This indicates that for every cluster, the distribution of data points around the cluster center is approximately the same. K-means might not work well if the variances of the clusters differ noticeably.

## Table1- Python Libraries

|  |  |
| --- | --- |
| import pandas as pd | data modeling, data analysis and data manipulation |
| import numpy as np | working with arrays |
| import matplotlib.pyplot as plt | creating high-quality visualizations and graphs |
| from sklearn.cluster import KMeans | used to divide data points into clusters |
| from sklearn import preprocessing | converting raw data into useful information |
| import seaborn as sns | provides a high-level interface for drawing attractive and informative statistical graphics |
| from sklearn.metrics import silhouette\_score | measures the similarity of a data point within its cluster (cohesion) compared to other clusters (separation) |
| from sklearn.preprocessing import StandardScaler | removes the mean and scales each feature/variable to unit variance |

## **Part III: Data Preparation**

One data preprocessing goal was to scale the data using standardscaler. According to an article on LinkedIn, scaling data prior to clustering is necessary because, “Clustering algorithms work by measuring the similarity or distance between data points. If your data has different scales, units, or ranges, the clustering results might be skewed or misleading. For example, if you cluster customers based on their age and income, the income variable might dominate the distance calculation because it has a much larger range than age. To avoid this, you need to preprocess and scale your data so that each variable has a comparable impact on the clustering outcome.” (*What Are Some Best Practices for Preprocessing and Scaling Your Data Before Clustering in Python?*, 2023)

## Table 2 - Data Set Variables

|  |  |
| --- | --- |
| Children | Continuous |
| Income | Continuous |

To begin the data analysis there are several steps that were followed. First the dataset was manipulated so as to only include the two variables of interest.

A screen shot of a computer

Description automatically generated

From the above screenshot it is apparent that the next step is to scale the data as income is a much larger value than children. The data was scaled using standardscaler from preprocessing.

A screenshot of a computer

Description automatically generated

A copy of the cleaned dataset is provided with the PA task 1 submission.  
 Part IV: Analysis

The optimal number of clusters in the data set is three. The optimal number of clusters was determined using the elbow method. The elbow method plots number of clusters vs inertia. The formula to calculate inertia is below.

Inertia = Σ(distance(point, centroid)^2)

A graph with a line

Description automatically generated

From the elbow plot above it is clear that the optimal number of clusters is 2. Below is a screenshot of the centeroids for each cluster.

A screenshot of a computer code

Description automatically generated

## Silhouette Score

The silhouette score provides the quality of the clusters mainly the cohesion and separation. Cohesion shows how close the data points are to other datapoints within its cluster whereas separation tells us how far it is from other clusters. Below is a screenshot of the silhouette score for the model. This number is low which means that there is not good separation between the clusters and the points are not cohesive. Based on this score it is not recommended to use this model.

A screen shot of a computer code

Description automatically generated

A graph with a line

Description automatically generated

A graph with blue dots

Description automatically generated

## **Part V: Data Summary and Implications**

The two clusters that I created are not good quality. I determined the quality of the clusters by running the silhouette score. The silhouette score provides a number between -1 and 1. A score closer to 1 indicates that the data points are close to one another, and the clusters are far away from each other. A silhouette score of 1 also indicates that there is similarity between the data points in the individual clusters. A number closer to zero indicates that the quality is average, data points are a little spread out with some overlapping of clusters. I have included a screenshot below that shows the silhouette scores for three varying amounts of clusters. You can see that all three have a low score. Of the three options for number of clusters the one with the highest score is two clusters which is what was reflected in the elbow plot.

A pink screen with white text

Description automatically generated

A limitation of my cluster analysis technique is the fact that the analysis is only as good as the number of clusters chosen. The variables I selected proved to not make quality clusters. When your clusters are not good quality it can mean that you need to start the analysis over. Starting an analysis from the beginning will cost the company time and money.

The results of my clustering analysis show that additional investigation into the data is required. More variables should be analyzed to see what if any segments prove to be useful in an analysis. The analysis did not provide any actionable insight. My recommendation is that the company should not move forward with this data analysis. Based on the silhouette score I believe the company should not use two clusters to segment their customers. The reason I do not believe this is a good idea is because total variance is below 50% which makes it closer to 0.

# Reference List

*What are some best practices for preprocessing and scaling your data before clustering in Python?* (2023, April 3). www.linkedin.com. https://www.linkedin.com/advice/3/what-some-best-practices-preprocessing-scaling