

VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

**A4- Cluster Analysis**

**Azim Ziyan A**

**V01102412**

**Date of Submission: 08-07-2024**

# TABLE OF CONTENTS

|  |  |  |  |
| --- | --- | --- | --- |
| 1. Introduction | 3 |  | |
| 2. Objectives | 3 |  |  |
| 3. Business Significance | 4 |  |  |
| 4. Results and Interpretations | 4 |  |  |

1. INTRODUCTION

For this assignment, background factors from the "Survey.csv" dataset will be used to profile respondents using a thorough cluster analysis. The dataset includes a range of socioeconomic and demographic characteristics, such as age, income, work status, and degree of education. The main objective is to use clustering techniques, such KMeans, to separate individuals with similar features and find unique groups within the respondent population in order to discover patterns. The best number of clusters will be determined using techniques like silhouette analysis and the Elbow approach after preparing the data to handle missing values, encode categorical variables, and scale features. This study attempts to offer a deeper knowledge of the respondent profiles by looking at the generated clusters. This insight can help with focused interventions, policy-making, and customized marketing strategies. Through this rigorous cluster analysis, the assignment not only enhances skills in unsupervised learning and exploratory data analysis but also demonstrates practical applications of clustering in understanding and addressing diverse respondent backgrounds.

# OBJECTIVES

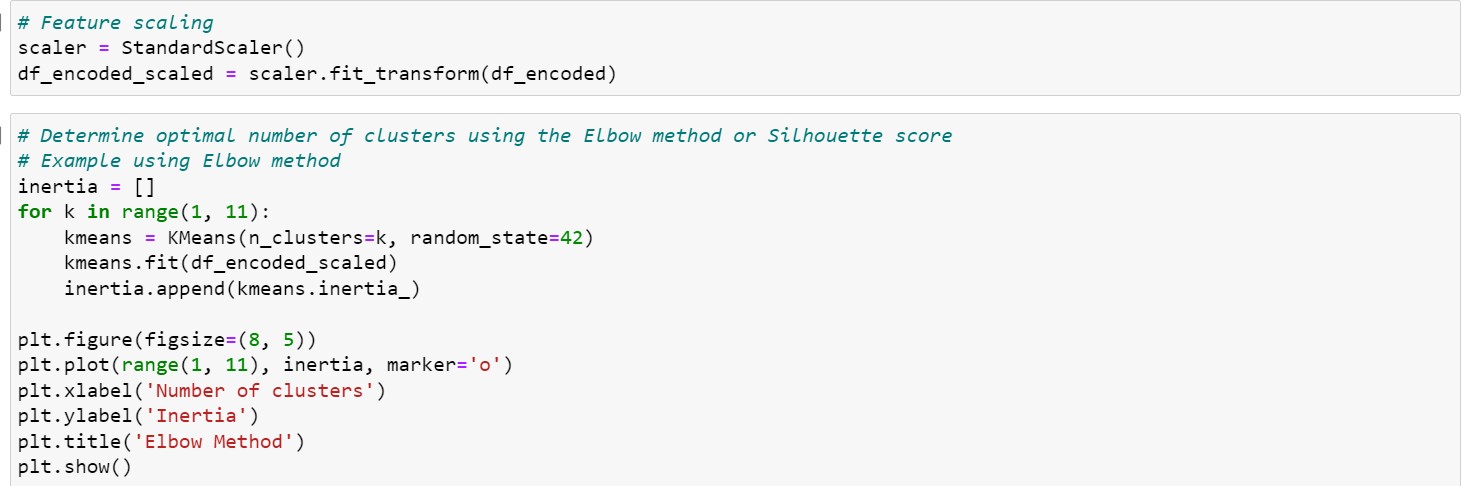
* 1. Cluster Analysis
* Identify Distinct Respondent Groups
* Understand Respondent Profiles
* Data Preprocessing
* Optimal Number of Clusters
* Cluster Analysis Application
* Evaluate Clustering Results

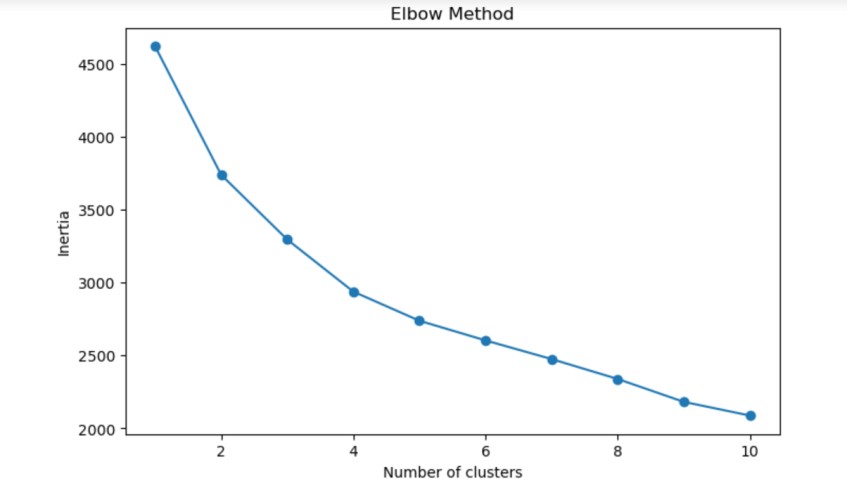
# BUSINESS SIGNIFICANCE

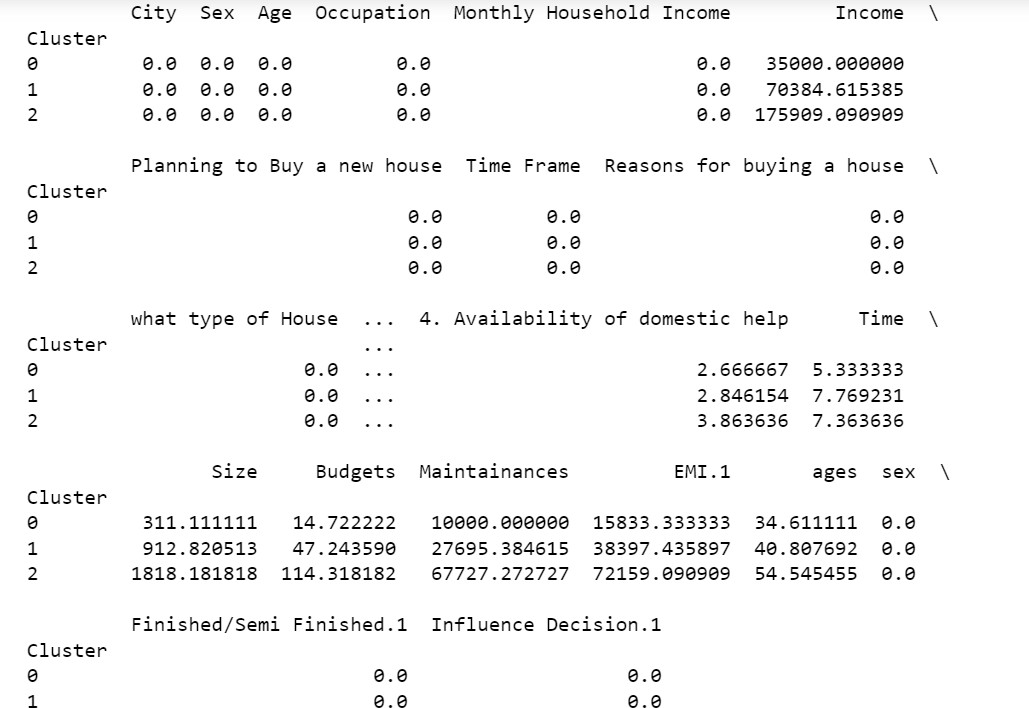
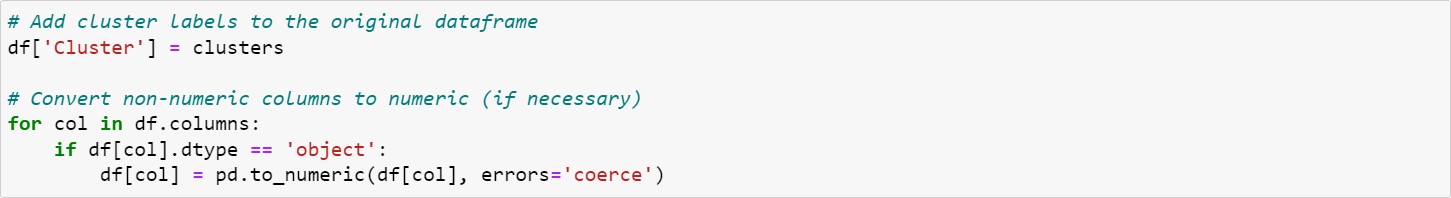
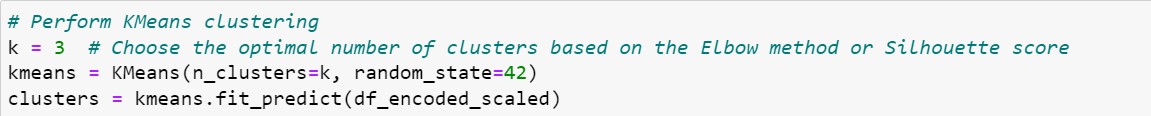
Applying cluster analysis to respondents' background variable-based characterization has important business ramifications in a number of sectors. Businesses can more successfully adjust their products and services to fit the varied demands and preferences of various client segments by identifying separate respondent groups. By using clustering to understand responder profiles, personalized marketing techniques that appeal to particular demographic or behavioral trends can be implemented, improving consumer happiness and engagement. In addition, cluster analysis makes it easier to find high-value markets with more room for revenue expansion and client retention. Businesses can enhance their competitive edge in a constantly changing market by optimizing resource allocation, streamlining operational efficiency, and creating customized strategies through the utilization of insights obtained by cluster analysis. In the end, this analytical technique promotes creativity in addition to bettering decision-making procedures.

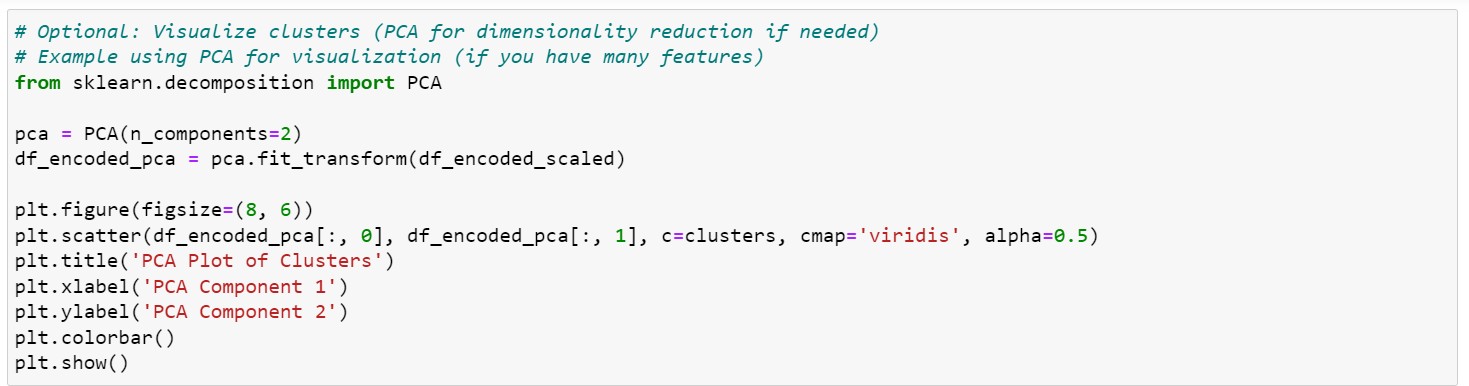
# RESULTS AND INTERPRETATIONS

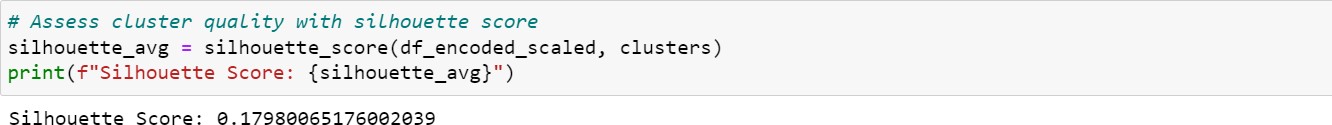
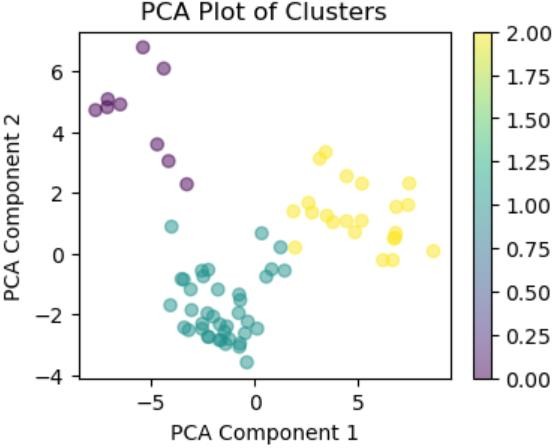
* + - Python



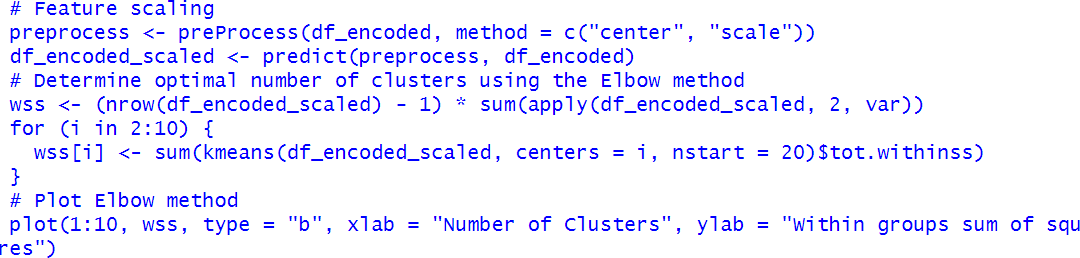


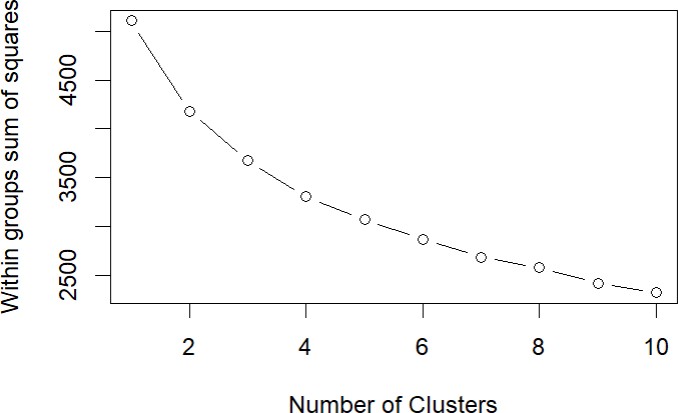


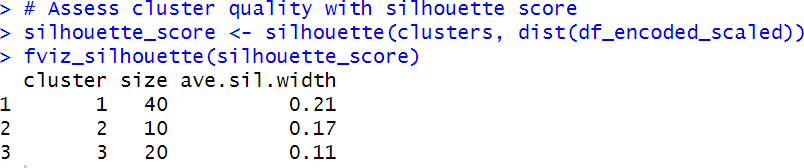
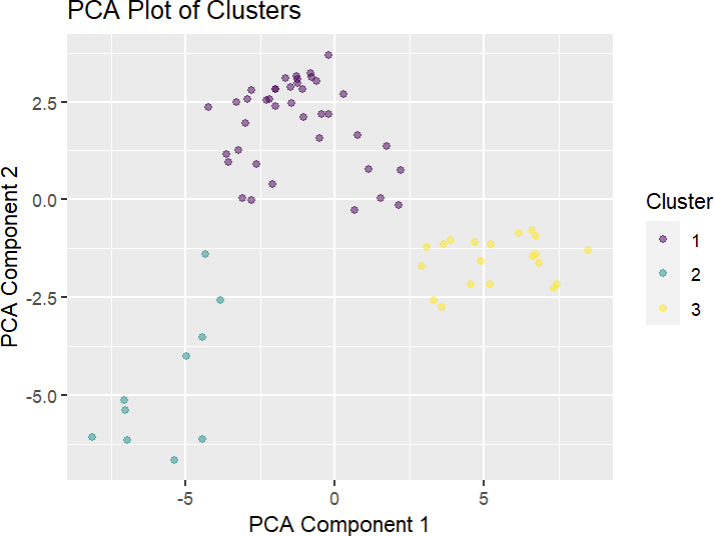
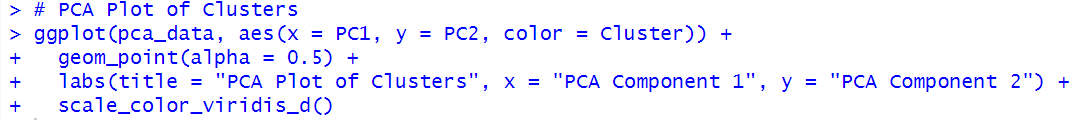
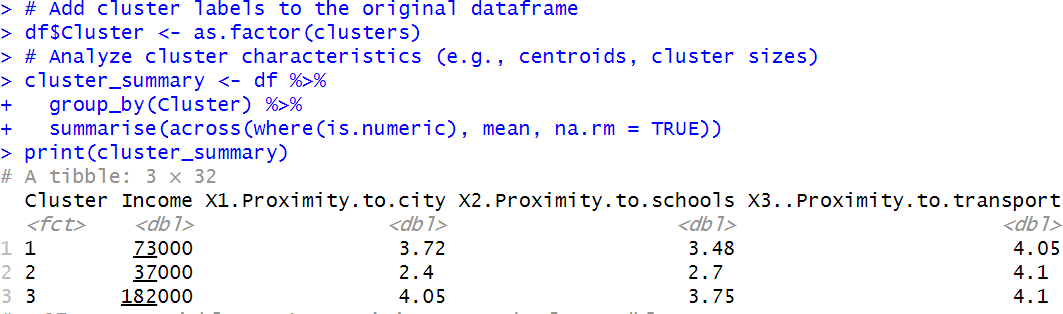


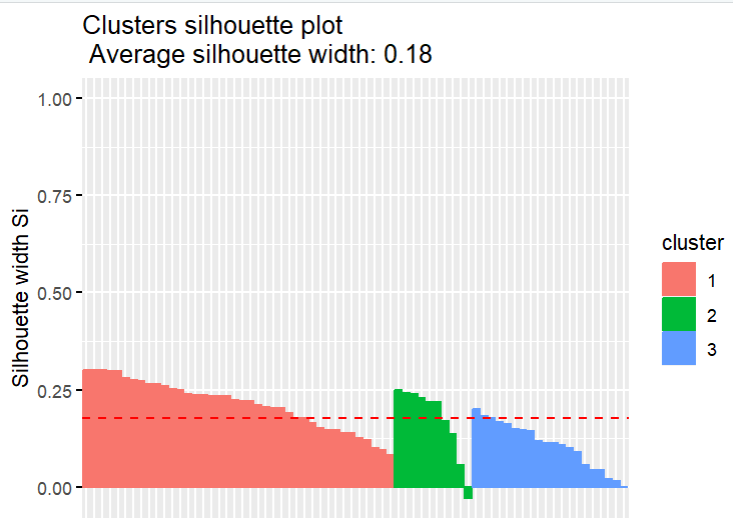


* + - R









Interpretation

Drawing from the encoded and scaled dataset, KMeans clustering studies conducted in R and Python yielded findings that show three different responder clusters, each with different background variables. With Cluster 0 having the lowest average income at roughly $35,000, Cluster 1 at roughly $70384, and Cluster 2 substantially higher at $175,909, the Python clusters demonstrate notable disparities in income levels. Aside from that, characteristics like housing preferences and accessibility to city facilities show different averages among clusters, indicating a range of socioeconomic and lifestyle profiles among the participants. Python's Silhouette Score is 0.1798, which denotes a respectable level of cluster quality and a respectable degree of cluster separation.

The clustering results in R show three groupings in a similar manner, but they also shed light on certain amenities and housing preferences. For example, Cluster 2 exhibits higher income levels and preferences for larger unit sizes and better interior design, while Cluster 1 is linked to moderate income levels and preferences for proximity to city attractions. With average silhouette widths ranging from 0.11 to 0.21, indicating medium to good separation between clusters, the silhouette analysis verifies reasonable cluster quality. All things considered, these results provide companies the ability to efficiently divide up their clientele, customize advertising plans, and create niche goods and services that suit the unique requirements and tastes of every group. This increases client contentment and maximizes the use of available resources.