Stockholm School of Economics
Saga Tran — 25770
Course: BE903 - Current Topics in Data Science for Business

Assignment 1 - Analyzig Market Survey Data

Date: 2025/04/04 5:59 pm

Question 1: K-means Clustering

** Throughout this analyses we have made the assumption that the segment in the target variable and therefore we have chosen to not include it in most of the analyses.*

(i) EDA

From table 1, we note that there are no people under the age of 18, and that the bulk of the people in the survey are between the age of 33 and 48. For the binary variables of "ownHome" and "gender" we can see that they are very evenly split, since their means lie around 50%. Finally, we see that about 10% of the respondents are subscribers, and that the average individual has one kid.

The histogram in figure 1 can be used to make an initial analysis of the distribution of each feature. The **age** distribution is right skewed, with a peak between 20 and 40 years. There is also a decline in frequency after age 60, but no real outliers. The **income** distribution looks similar, with most individuals earning between 25000 and 50000. The few who earns an income of 100 000 or more, could be treated as outliers. The **Kids** distribution is heavily skewed toward 0 kids, with a sharp drop-off as the number of kids increases. This suggests that the (very few) people who have 7 kids should be treated as outliers.

The correlation heatmap additionally displays the dependence between the features in the dataset. Figure 2 highlights a moderate positive relationship between age and income (60%), which suggests that older individuals have higher incomes (reasonable due to career advancement). There is also a weaker negative correlation between age and kids (-29%), indicating that younger people are more likely to have kids. This is not the most straightforward. However, if kids only refer to children between ages $0-18$, it becomes more reasonable since you tend to be younger when first starting a family. Other variables suggests little to no relationships.

(ii) K-means Clustering Algorithm

Before applying the K-Means algorithms, we use python *sklearns* standard scaler to standardize numerical features, that is, age, income, and kids. We do this since the K-Means algorithms sometimes can be sensitive to huge differences in scale (since it measures the euclidean distances between data points). The size of each cluster and the features means are presented in tables 2 and 3.

First, we note that cluster 1 contains the most observations, and cluster 0 the least. Inspecting the feature means, we directly see that clusters 1 and 2 display quite similar means, except for the number of kids that is notably different. To the contrary, the remaining clusters differs heavily in both age and income. Cluster 0 seems to capture young people with lower incomes, and cluster 3 captures older people with higher incomes, who also have a significantly lower subscription rate than the rest of the groups.

(iii) Visualising the cluster solution

The 4 clusters (numbered 0-3) with their centroids are shown graphically in figure 3. These results align with the comments with regards to the means in the previous point. We see

that cluster 0 and cluster 3 are the most distinct with regards to age and income, whilst cluster 1 and 2 overlap to a large extent.

(iv) Boxplot visualisations

Again, figures 4 and 5 highlights the results we have commented on previously, with clear differences between clusters 0 and 3, and a lot smaller differences between clusters 1 and 2. Cluster 0 seems to contain the young individuals with low income, and cluster 3 the older individuals with higher income. Individuals assigned to cluster 1 has a bit higher income on average, while cluster 2 contains people who are a bit older on average.
Additionally, we can note that there are a few outliers, primarily to cluster 2. This means that a small number of people deviate significantly from the general pattern of the other people in that cluster. The presence of outliers could point to limitations in the clustering approach, such as an unsuitable algorithm or an inappropriate number of clusters, causing some points to be poorly assigned.

(v) ANOVA Hypothesis testing

The ANOVA test for income produces an F-statistic of approximately 176 with p-value equal to zero. This allows us to reject the null hypothesis, which states that *the mean income is the same across all clusters*. Thus, there is significant evidence that the mean income differs across the clusters. The ANOVA test for age produces an even higher F-statistic of 351, with a p-value of zero. Hence, there is also significant evidence that mean age differs across clusters. The large F-statistic produced in both tests indicates a substantial variation in both income and age between the cluster groups, and the extremely small p-value confirms that this difference is not due to randomness but is statistically meaningful. Concluding, this suggests that the clustering process has effectively identified distinct income profiles within the dataset.

(vi) Determining optimal number of clusters

The within-cluster sum of squares measures the total variance within each cluster. As K increases, WCSS decreases because more clusters fit the data more closely. The optimal K is where the WCSS reduction rate slows, forming an "elbow" shape on the plot. Figure 6 shows that the optimal number of clusters likely is 3 or 4.

(vii) Computing accuracy

Since K-means is unsupervised and does not inherently know the original labels, "correct classification" requires that we map each K-means cluster to the most similar original segment and then count the matches. To do this, we created a contingency table to align clusters with segments based on maximum overlap. Each cluster was mapped to the original segment with the highest count in that cluster, as displayed in table 4. This assumes the "correct" label for a cluster is the dominant segment within it. Therefore, "Correctly classified" points are those where the K-means cluster maps to the original segment they belong to, and the accuracy becomes:

$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{total points}} \times 100\%$$

We achieved the results presented in table 5. One thing that we must note here is that two clusters are mapped to the same segment. This is because members of suburb mix and moving up are split between clusters 1 and 2, and in both cases more members of the suburb mix are assigned to that particular cluster. However, comparing the K-means solution in figure 3 with the original segmentation in figure 7, we believe that this representation is quite accurate. Therefore, the conclusion is that the original segmentation may not be perfect.

Question 2: PCA

(i) Principal Component extraction

We start by applying the python *sklearns* standard scaler to all (both numeric and binary) features. We then use the build-in sklearn.decomposition PCA tool to extract 2 principal component. Results are presented in table 6.

For PC1, age, income, and ownHome is dominant, which captures a socioeconomic progression. Higher loadings indicate older, wealthier individuals who own homes. For our analysis in **question 1**, we see that this is more common in the Travelers segment. PC2, on the other hand, is dominated by gender, kids, and subscribe. The positive kids loading suggests families align with this axis. Since gender and subscription is high, PC2 may distinguish female subscribers from male non-subscribers (male = 0, female = 1), orthogonal to wealth/age. Lower loadings on income, ownHome, and age indicates that this component focuses on other traits.

(ii) Variance proportion

The proportion of variance explained by each PC is precisely what the $explained\_variance_ratio\_in scikit-learn provides. This value represents the fraction of the total variance in the dataset that each PC accounts for, an$

(iii) PC scores with cluster membership

Figure 8 displays the principal component scores with the cluster membership as determined in Question 1.

(iv) Individual PC scores

Figure 9 identifies the individual with the highest score for the first principal component, and the individual with the highest score on the second principal component.

The individual with highest PC1 score is a female of age 63, who owns her home and has no kids nor subscriptions, with an income of 105984. The individual with the highest PC2 score is also a woman, but 37 years old with 5 kids, and a subscription, who has an income of 69852. We can conclude that high PC1 probably aligns with an older and richer profile. This is consistent with the loadings in table 6. Contrary, high PC2 suggests a different dimension (female with many kids and moderate income level), tied to loadings like gender or kids. This contrast aligns with PC1 as a wealth/age axis and PC2 as a family/gender axis, which reveals distinct life stages or roles in the data.

Appendix

	age	gender	income	kids	ownHome	subscribe
count	300	300	300	300	300	300
mean	40.557	0.497	52071	1.283	0.460	0.117
std	12.683	0.501	19724	1.362	0.499	0.322
min	18	0	10557	0	0	0
25%	33	0	41275	0	0	0
50%	39	0	53186	1	0	0
75%	48	1	64236	2	1	0
max	73	1	114615	7	1	1

Table 1: Descriptive statistics of the dataset

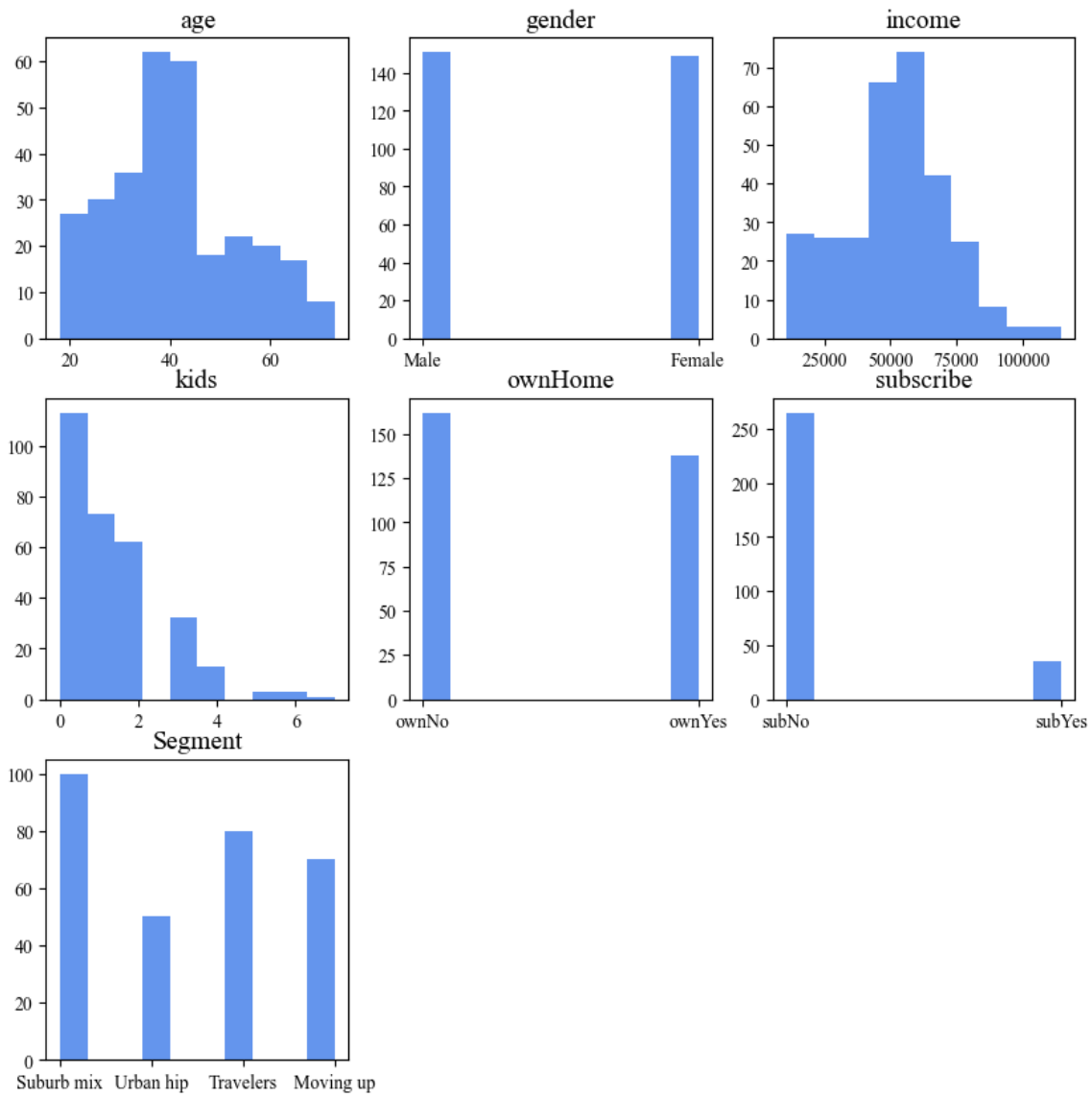

Figure 1: Histogram of each variable

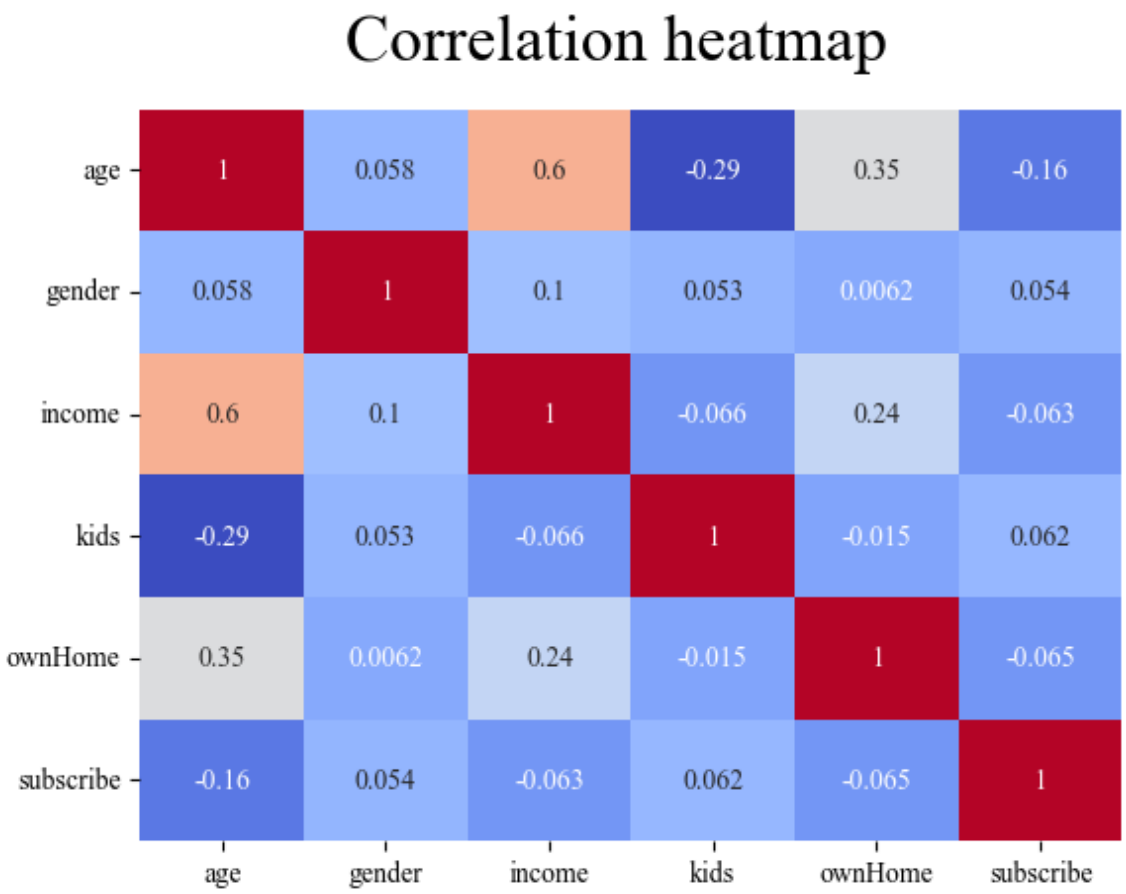

Figure 2: Correlation matrix of relevant variables

	Count
Cluster 0	51
Cluster 1	97
Cluster 2	89
Cluster 3	63

Table 2: Sizes of clusters

	Age	Gender	Income	Kids	OwnHome	Subscribe
Cluster 0	23	0	21655	1	0	0
Cluster 1	38	1	54593	3	0	0
Cluster 2	41	0	52161	0	0	0
Cluster 3	59	0	72685	0	1	0

Table 3: Mean values for each feature in the clusters

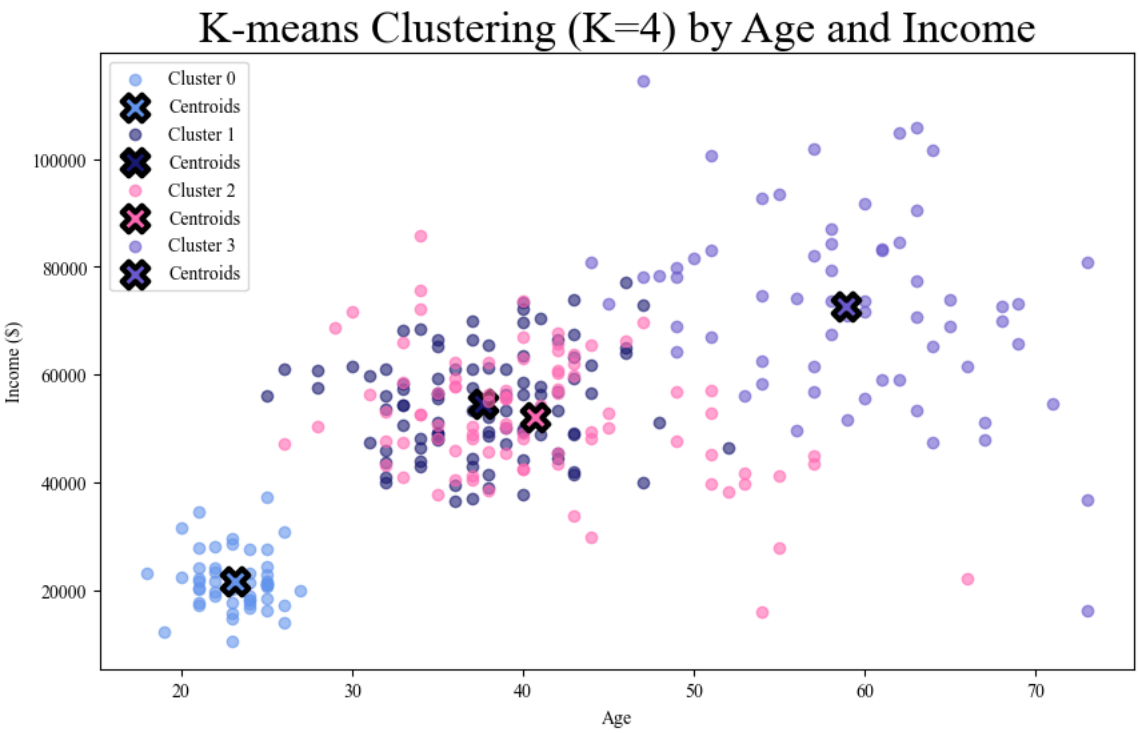

Figure 3: Visualising cluster solution

Figure 4: Boxplot of income across clusters

Figure 5: Boxplot of age across clusters

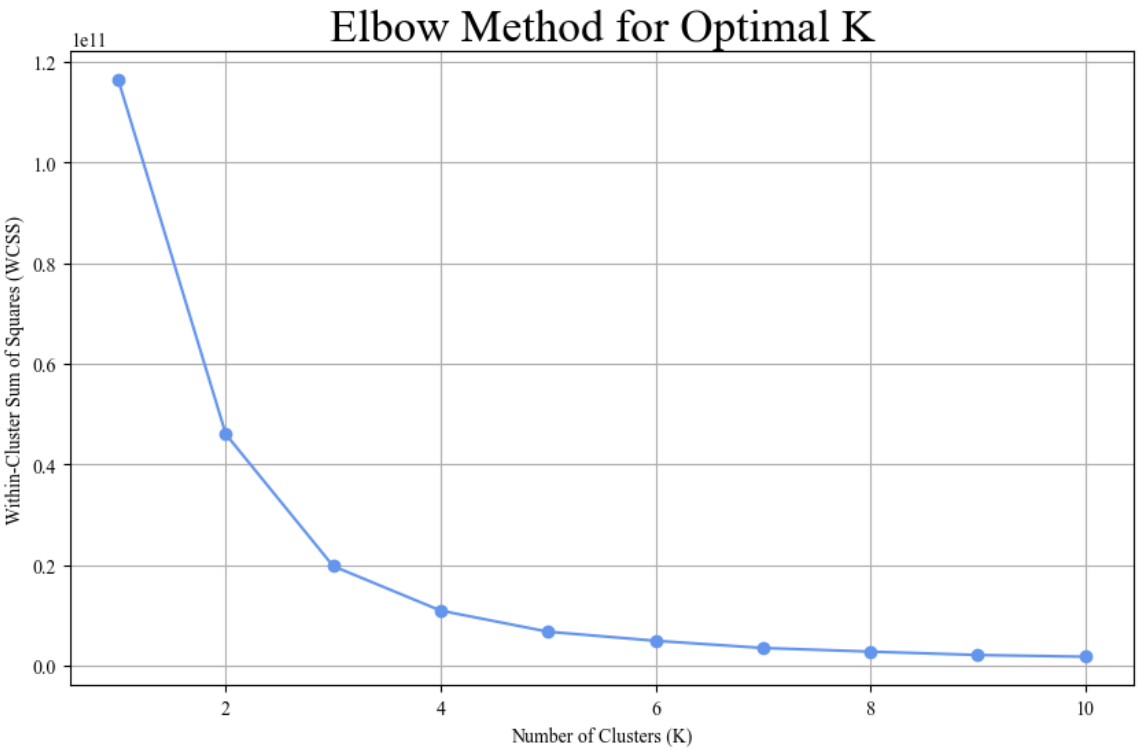

Figure 6: Elbow method for determining optimal no. clusters

Cluster 0 mapped to Urban hip
Cluster 1 mapped to Suburb mix
Cluster 2 mapped to Suburb mix
Cluster 3 mapped to Travelers

Table 4: Cluster mappings

Number of correctly classified points:	207 out of 300
Accuracy:	69.00%

Table 5: K-Means Accuracy

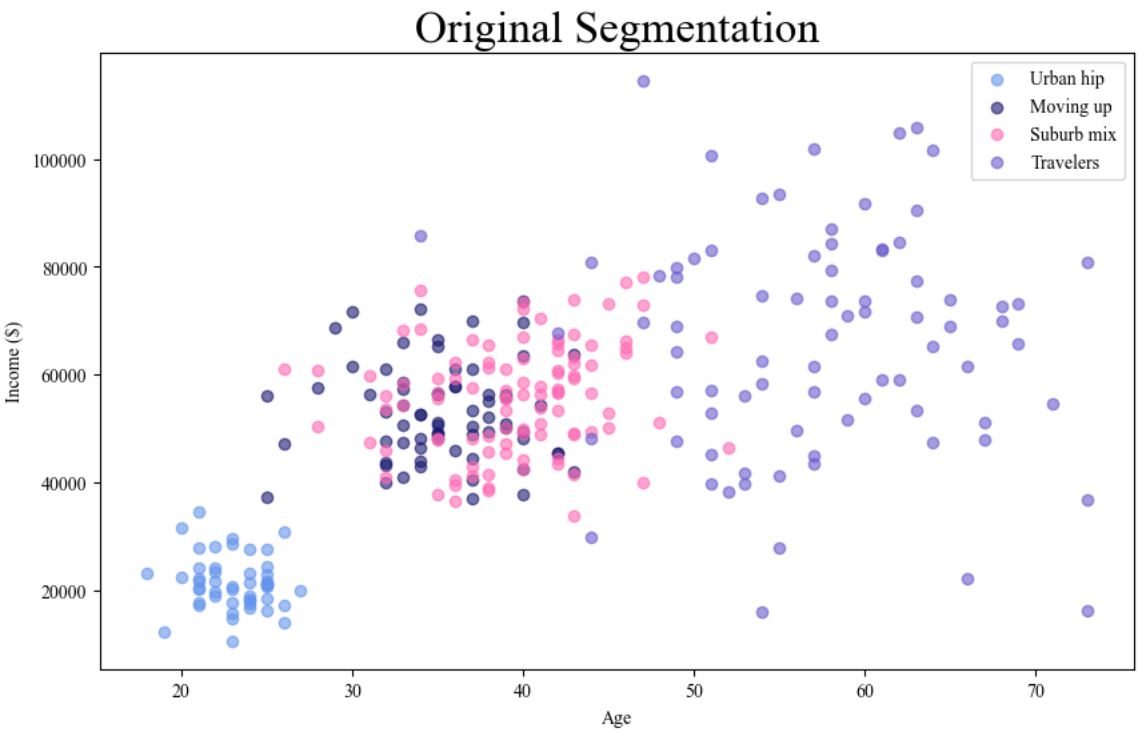

Figure 7: Scatter plot of original segmentation according to segments

	PC1	PC2
Age	0.640	-0.020
Gender	0.078	0.687
Income	0.562	0.225
Kids	-0.260	0.501
OwnHome	0.405	0.125
Subscribe	-0.192	0.459

Table 6: Extracted Principal Components (Loadings)

	Variance Explained
PC1	31.99%
PC2	18.45%
Total	50.44%

Table 7: Proportion and total variance explained by PC component

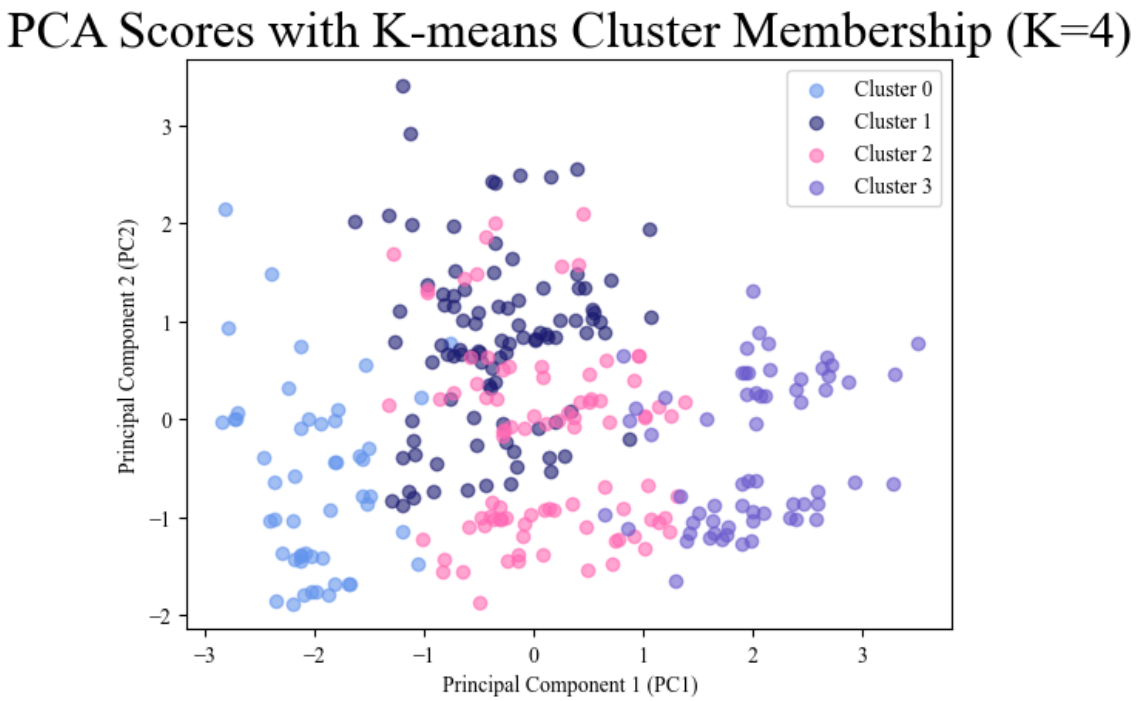

Figure 8: Principal component scores with cluster membership

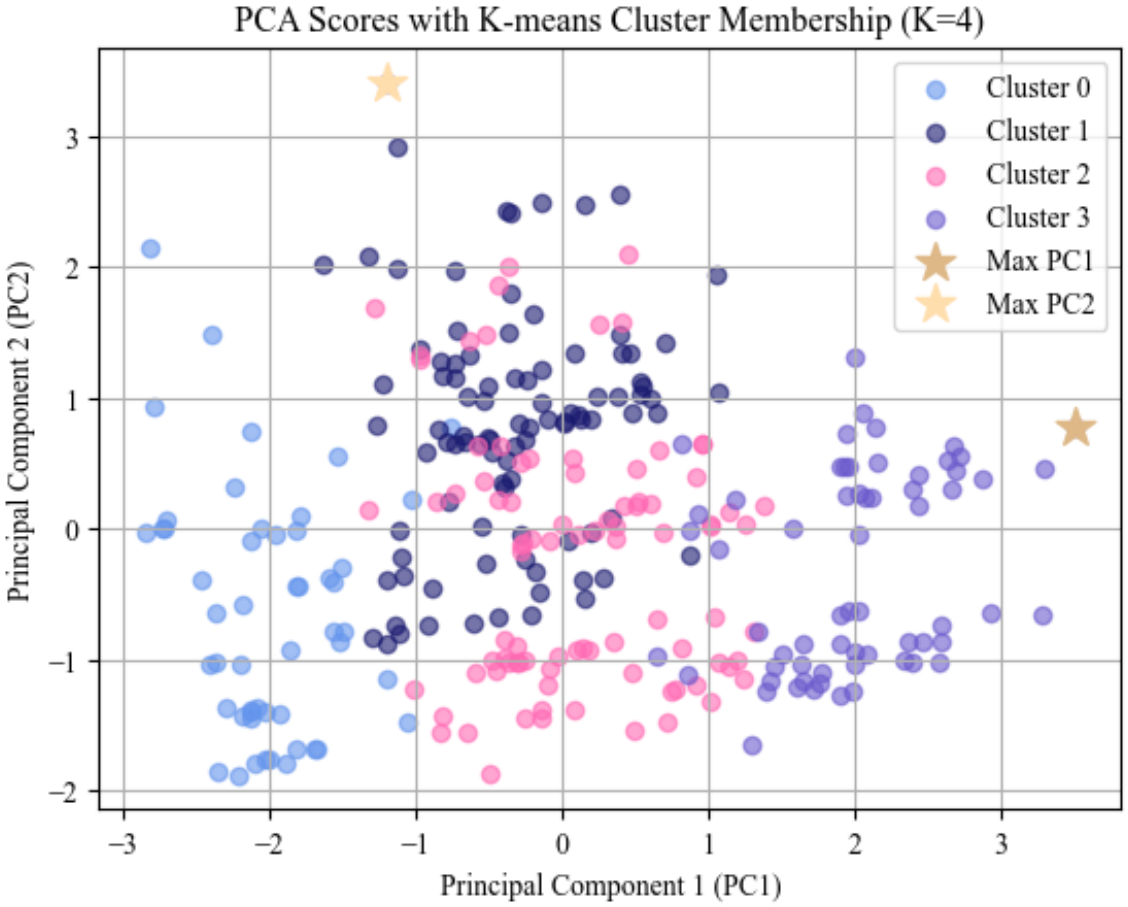

Figure 9: Identification of individuals with the highest & lowest PCs

	Highest PC1 score
Age	63
Gender	1
Income	105984
Kids	0
OwnHome	1
Subscribe	0

Table 8: Individual with highest PC1 score

	Highest PC1 score
Age	37
Gender	1
Income	69852
Kids	5
OwnHome	0
Subscribe	1

Table 9: Individual with highest PC2 score