Mall Customers Cluster Analysis

By: Katlyn Thomas, Cole Johnson, and Teja Talluri

Dataset Overview

Mall Customers Dataset: A collection of survey responses from consumers who shopped at a mall

Variables:

- Customer Id: Unique ID
- Age
- Gender
- Income: annual income in $k
- Spending Score: score assigned to customer based on customer behavior and spending nature

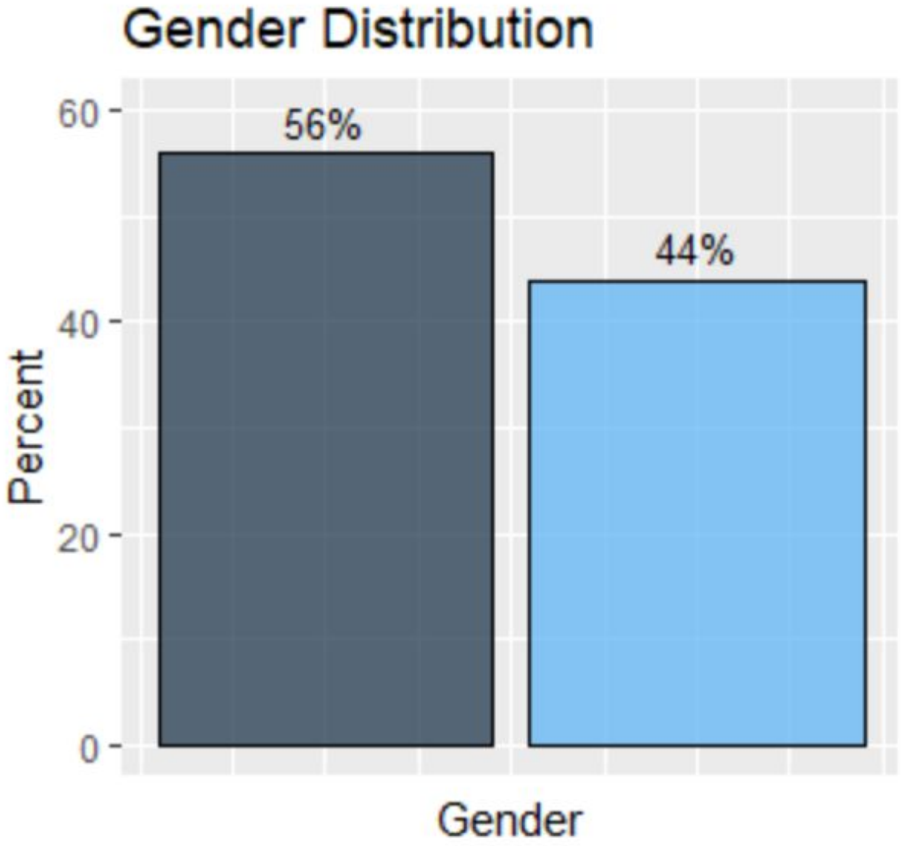
Goal: Use survey responses to identify the optimal number of target customer segments(clusters) for the purpose of improving marketing efforts

Factors to Consider: What are the patterns amongst survey responses? What consumer cluster should we cater our marketing strategy towards?

Exploratory Data Analysis

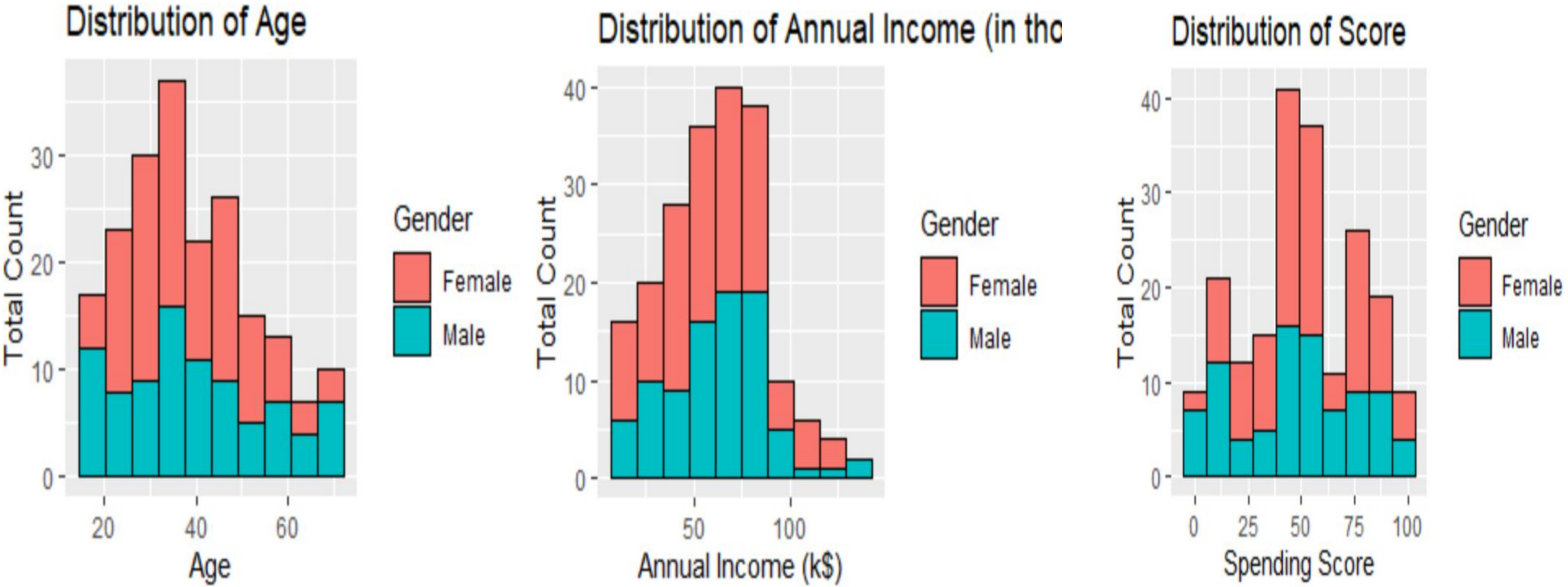
Includes:

- Gender of respondents (Male/Female)
- Distribution of Age, Income, and Score by gender
- Comparison of Age, Income, and Spending Score to Gender via scatterplot



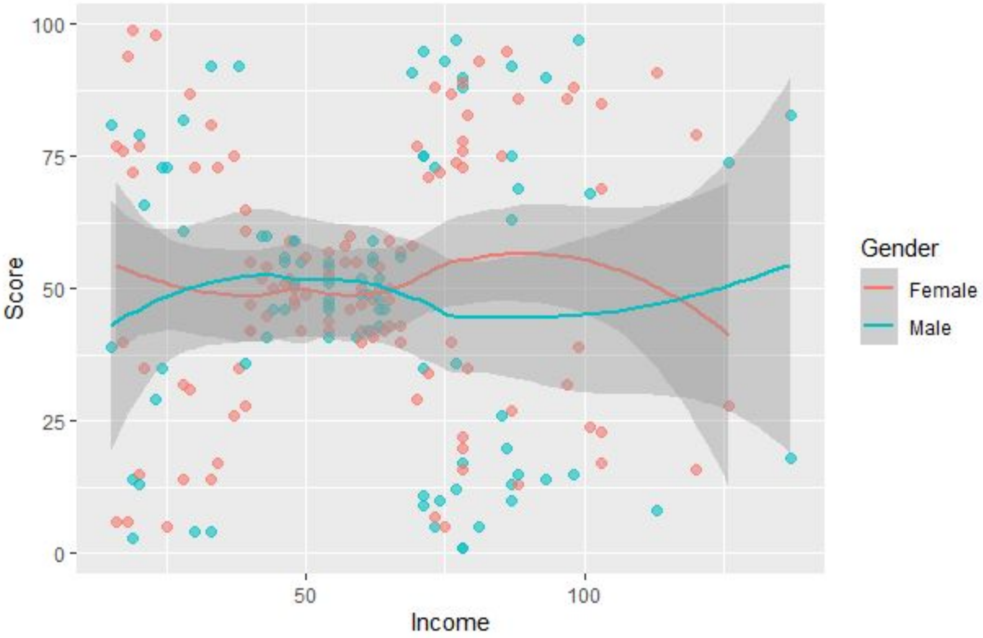
Results:

- 56% of respondents were female
- 44% of respondents were male

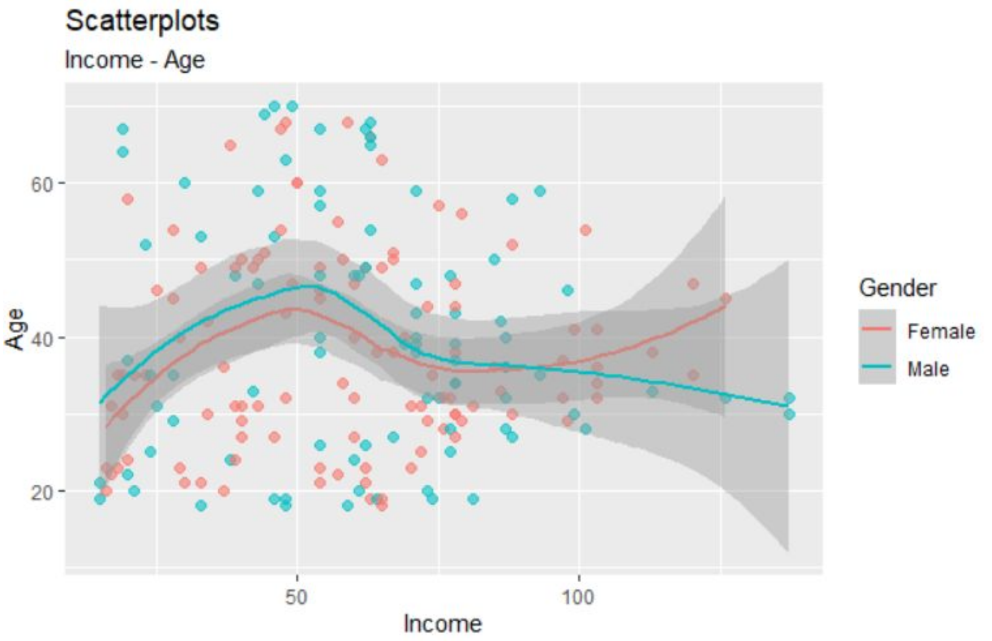


Visualizations of Distributions by Gender for Age, Score, and Salary

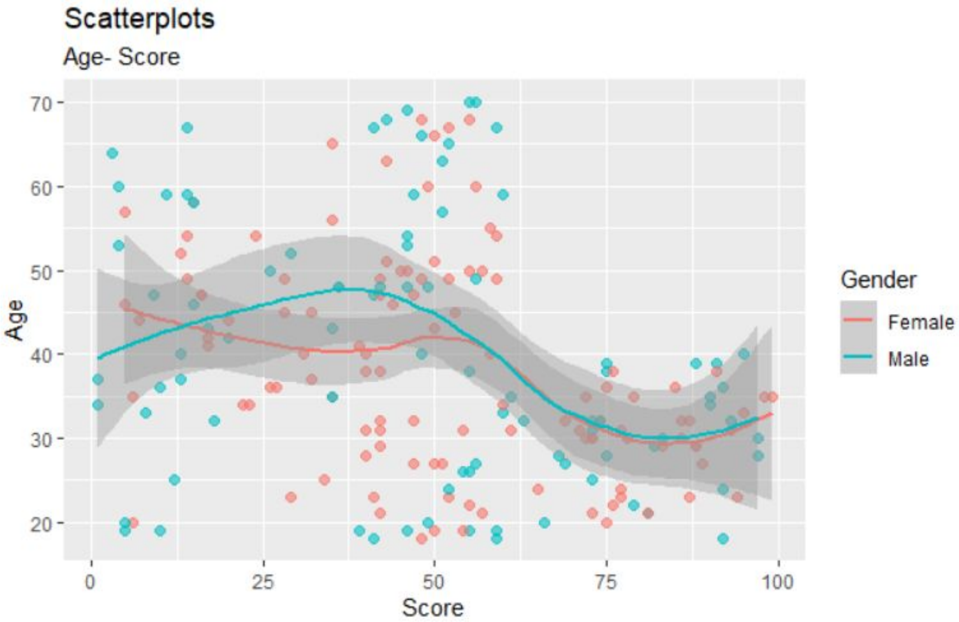
Visualization: Income & Score by Gender



Visualization: Income & Age by Gender



Scatterplot: Age & Score by Gender



Cluster Analysis

Includes:

- K-means Cluster Analysis
    - Analysis #1: Non- Scaled and Non-Standardized Variables
    - Analysis #2: Scaled and Standardized Variables

K-Means Cluster Analysis

Methodology:

1. Check for outliers
2. Convert using transformation
3. Standardize Data
4. Run K-Means
5. Box Plots

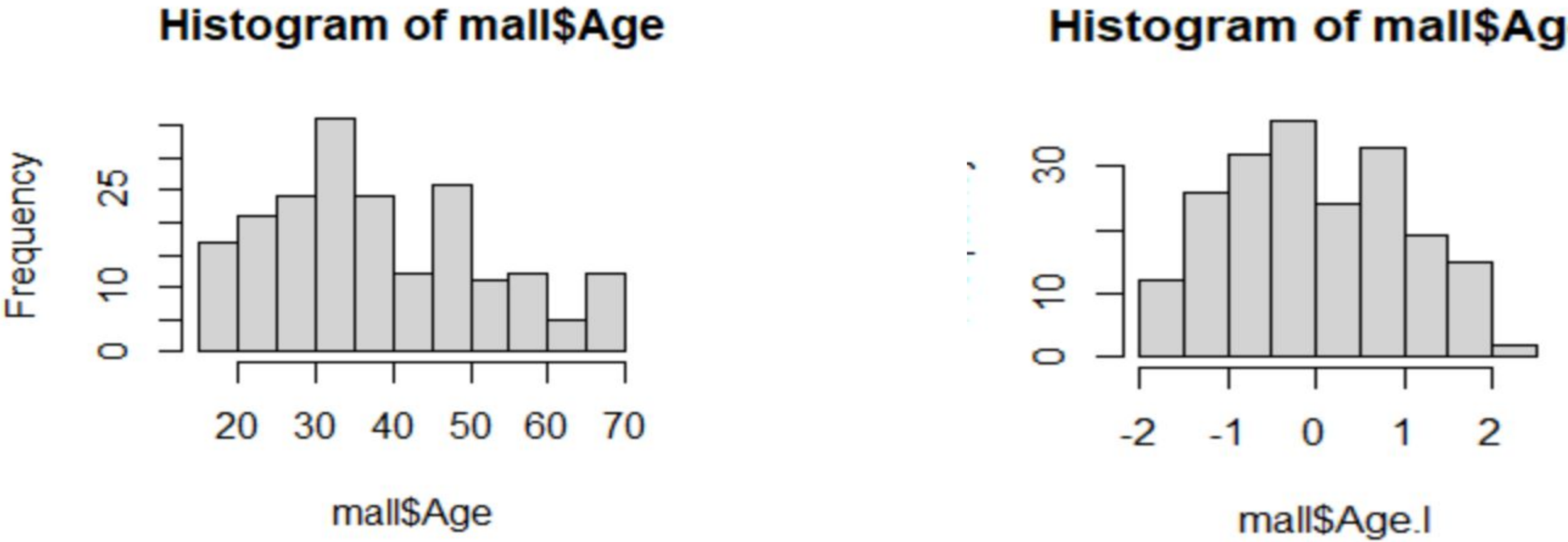
Methodology:

- Used the default Euclidean distance and complete linkage methods in the code
- Renamed columns to simplify data
- Identified outliers and scaling discrepancies via Histograms of the variables: Age, Income, Score
- Performed K-means Analysis for :
  - Scaled and normalized dataset
  - Unscaled and non-normalized dataset

Purpose:

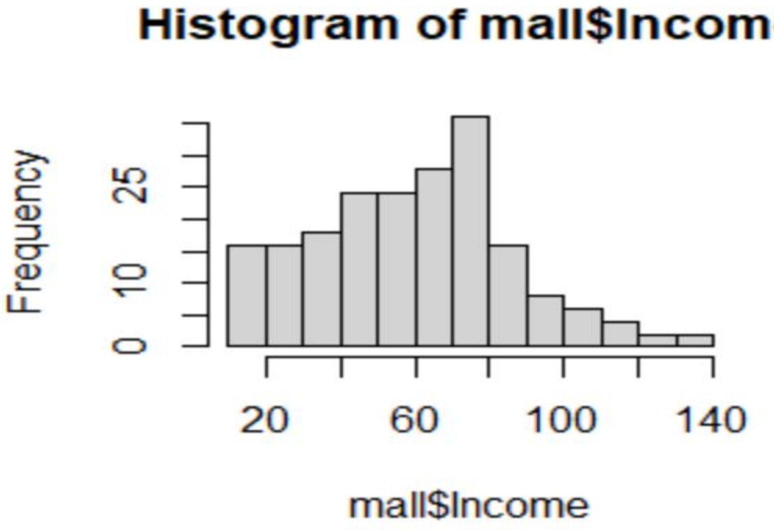
- Identify optimal amount of K clusters
- Identify if optimal K varies based on the type of data used (standardization)
- Identify best cluster to market to

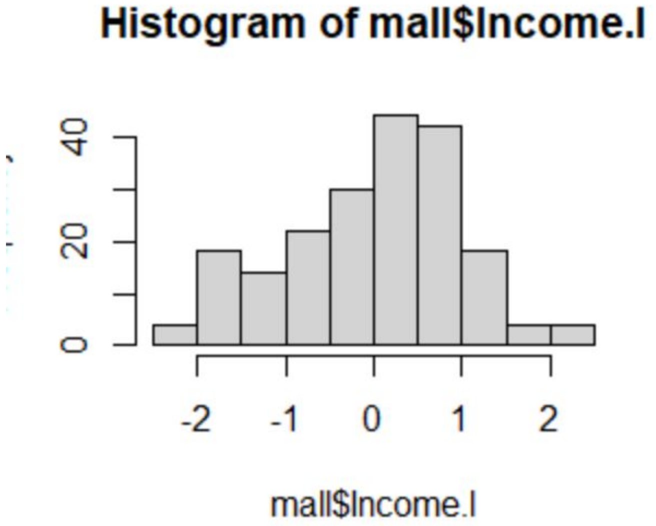
Histogram for Age



Left: Distribution of Age without rescaling, **Right:** Distribution of Age with rescaling (sqrt) and standardization

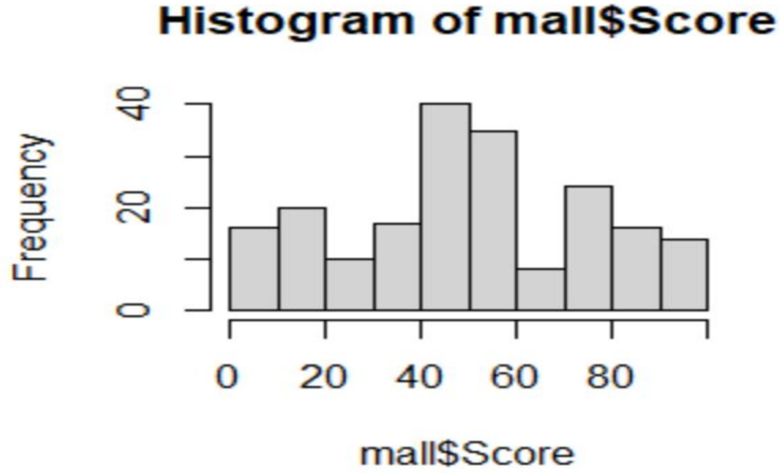
Histogram for Income

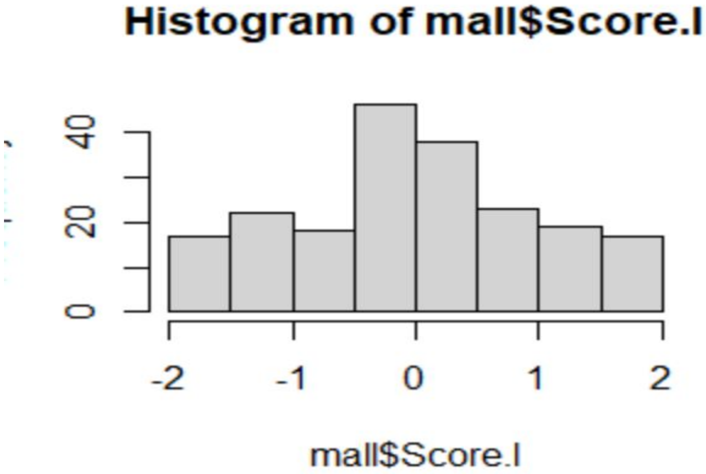




Left: Distribution of Income without rescaling, **Right:** Distribution of Income with rescaling (sqrt) and standardization

Histogram for Score

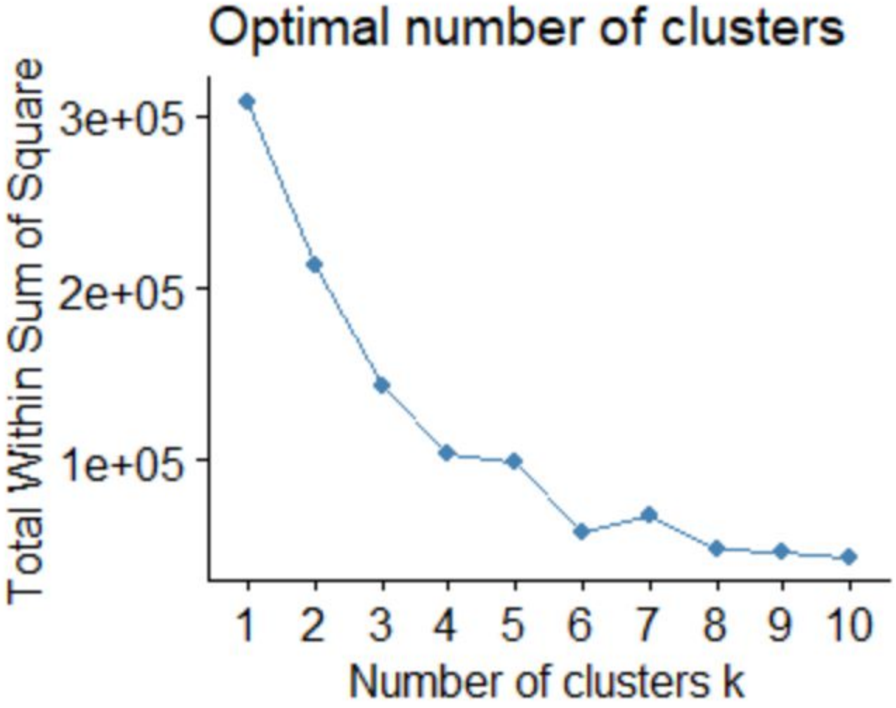




Left: Distribution of Score without rescaling

Right: Distribution of score after standardization

Analysis #1: Original (Non-Scaled/Standardized) Dataset



Interpretation:

- We used the wss method to visualize the optimal number of clusters for our K-means analysis
- Based on the “Elbow Method” we will set K=5 for our K-means analysis

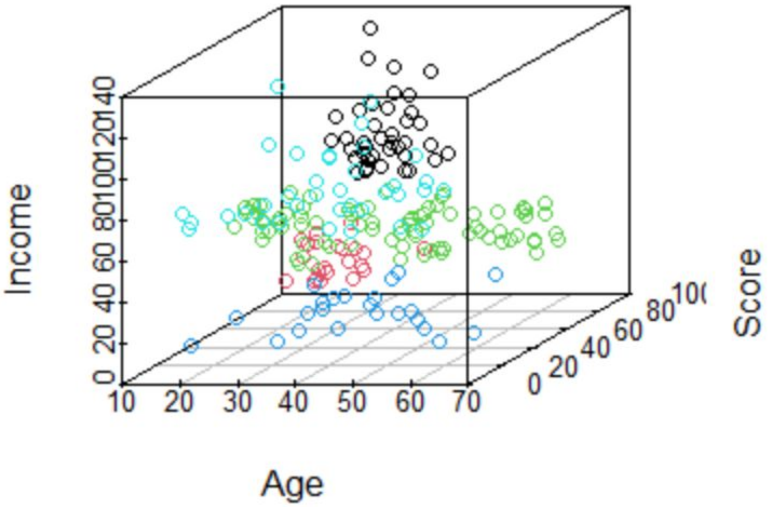
Cluster Analysis #1 Cont.

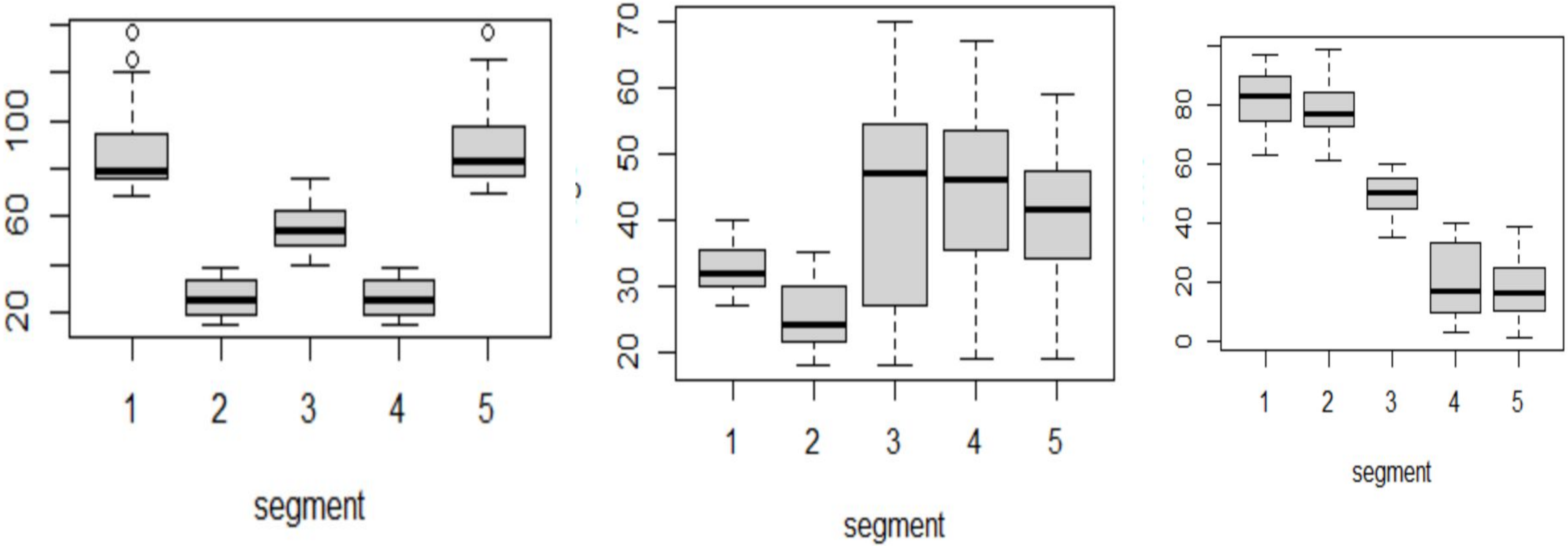
Interpretation:

- 5 distinct segments with minimal overlap
  - Cluster 1: 39 observations
  - Cluster 2: 23 observations
  - Cluster 3: 79 observations
  - Cluster 4: 23 observations
  - Cluster 5: 36 observations

Potential Clusters:

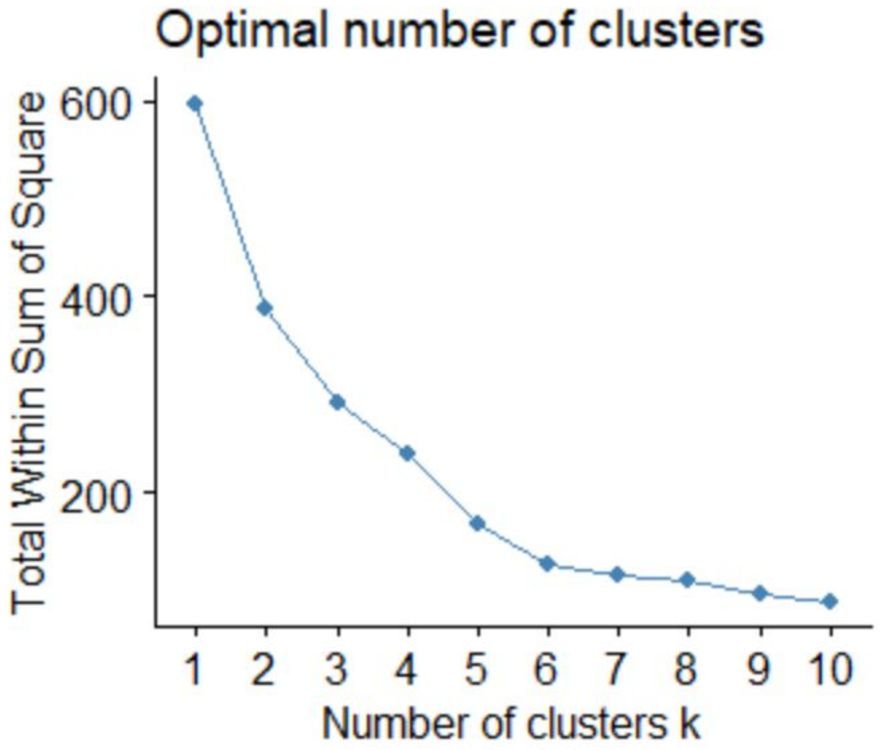
1. High spending score, high income, age < 40
2. High spending score, low income, age > 40
3. Medium spending score, medium income, age covers full range
4. Low spending score, low income, full age coverage
5. Low spending score, high income, age > 40





Box Plots for Cluster Analysis #1 Variables (**from left to right:** Income, Age, Score)

Analysis#2: Scaled & Standardized Dataset



Interpretation:

We used the wss method to visualize the optimal number of clusters for our K-means analysis

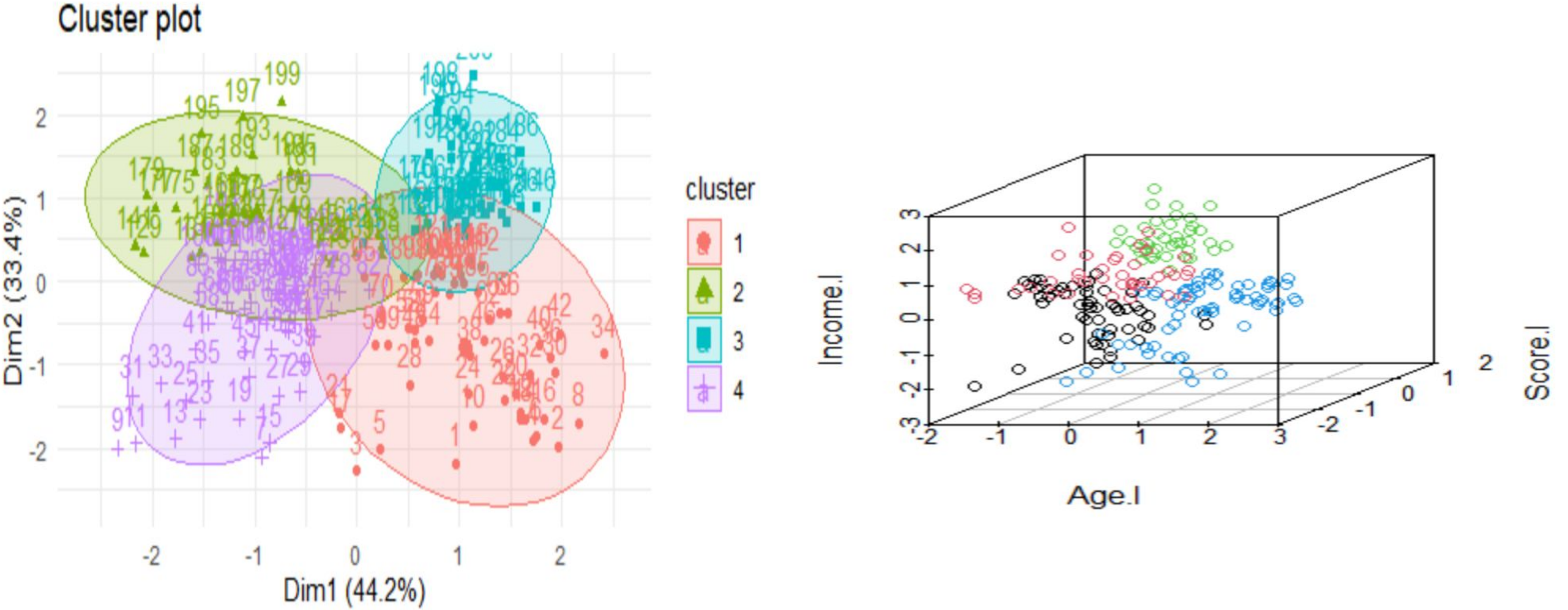
Based on the “Elbow Method” we will set **K=4** for our K-means analysis using scaled & standardized variables

Methodology:

- Rescaled [using sqrt()] :Age(Age.l), Income(Income.l).
- Standardized: Score(Score.l) , Age.l, and Income.l

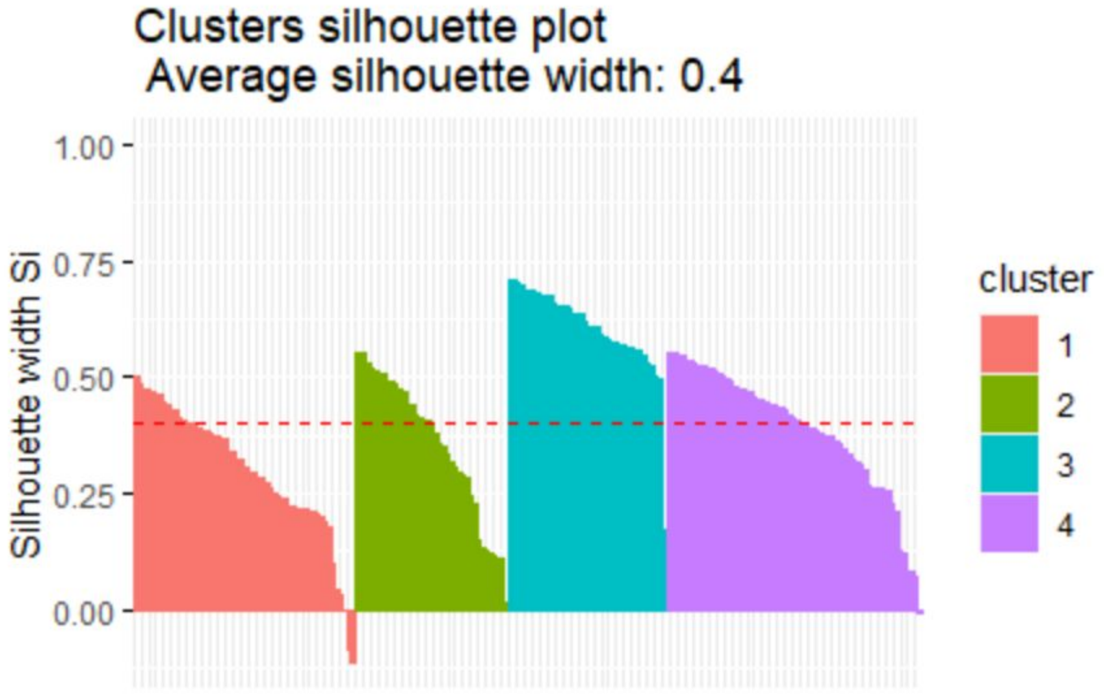
For visualization:

- 3D & 2D Methods
- Silhouette
- Box Plot



Standardized and Scaled Cluster Analysis Visuals (**left**: 2D Model, **right:** 3D Model)

Silhouette Visualization



Clusters silhouette plot
Average silhouette width: 0.4

Output:

```
cluster size ave.sil.width
1        56       0.30
2        39       0.36
3        40       0.61
4        65       0.39
```

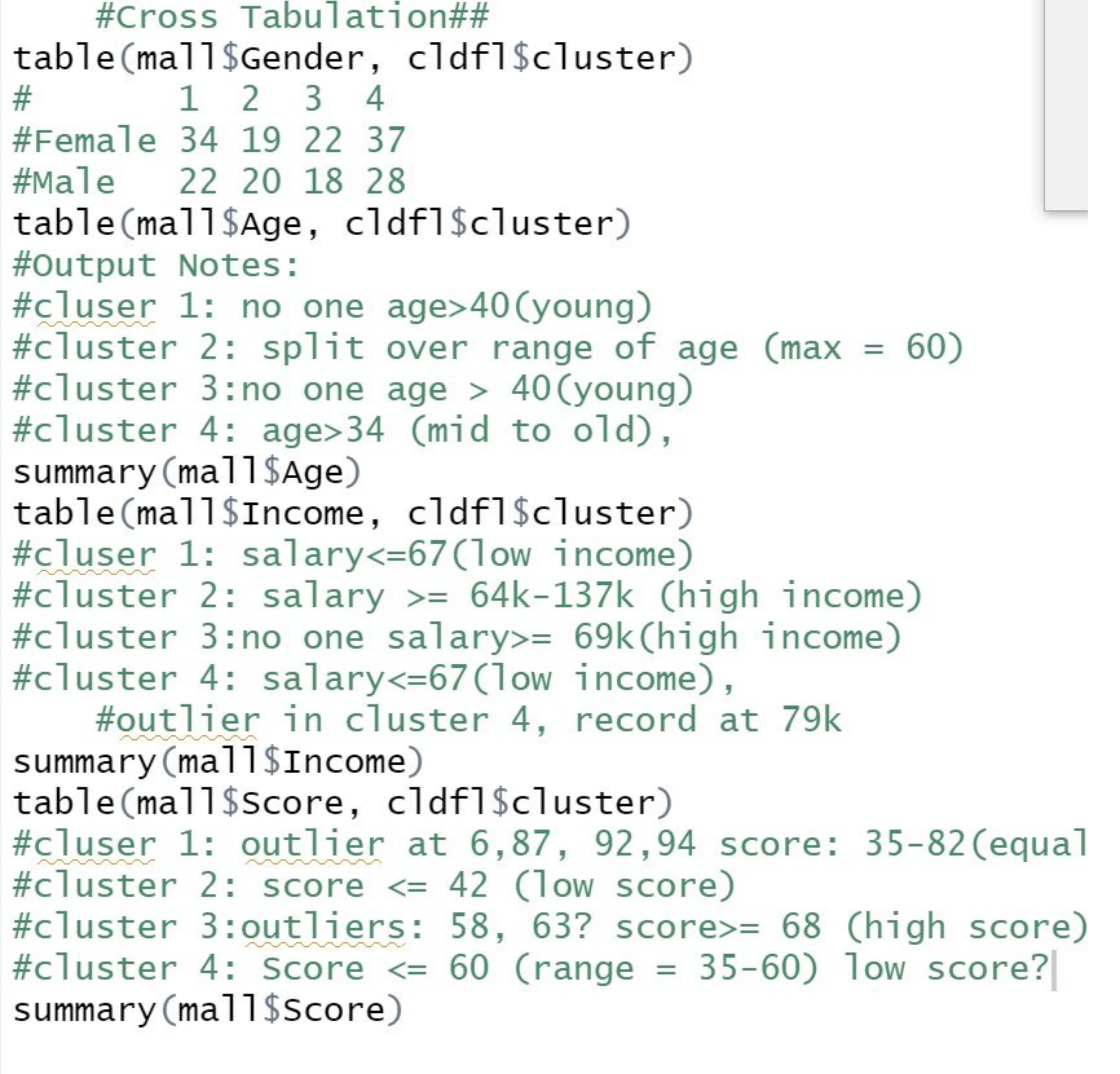
Interpretation:

- Ave.sil.width closer to 1 = better fit
  - Thus cluster 3 = optimal cluster

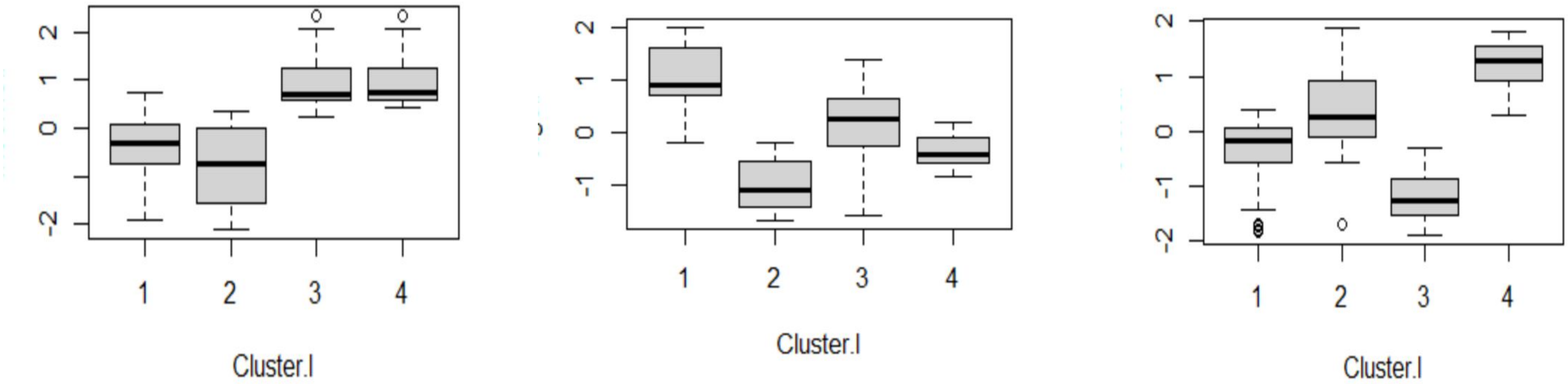
Cross Tabulation to Find Best Cluster

Cluster Types:

1. Cluster 1: young, low income, spending score spread across range
2. Cluster 2: middle age, high income, low spending score
3. Cluster 3: young, high income, high spending score
4. Cluster 4: old), low income, low score



```
    #Cross Tabulation##
table(mall$Gender, cldfl$cluster)
#       1  2  3  4
#Female 34 19 22 37
#Male   22 20 18 28
table(mall$Age, cldfl$cluster)
#Output Notes:
#cluser 1: no one age>40(young)
#cluster 2: split over range of age (max = 60)
#cluster 3:no one age > 40(young)
#cluster 4: age>34 (mid to old),
summary(mall$Age)
table(mall$Income, cldfl$cluster)
#cluser 1: salary<=67(low income)
#cluster 2: salary >= 64k-137k (high income)
#cluster 3:no one salary>= 69k(high income)
#cluster 4: salary<=67(low income),
    #outlier in cluster 4, record at 79k
summary(mall$Income)
table(mall$Score, cldfl$cluster)
#cluser 1: outlier at 6,87, 92,94 score: 35-82(equal
#cluster 2: score <= 42 (low score)
#cluster 3:outliers: 58, 63? score>= 68 (high score)
#cluster 4: Score <= 60 (range = 35-60) low score?
summary(mall$Score)
```



Box Plots for Log Cluster from left to right: Income.l, Age.l, Score.l

Final Recommendations

Based on the optimal **K = 4** identified in the Standardized and Scaled Clustering Analysis:

- We believe the two most important variables are **age** and **income**.

Marketing focus: **cluster 3**

- **Characteristics**: Age = young(27-40), Income = high (>= $69k), Spending Score = high
- Ave.sil.width = .61
- Gender: 22 Female , 18 Male

Cluster 3 Assumptions:

1. Those who are younger and have more disposable income are more likely to spend money
2. Standardized & Scaled **income** variable is skewed, if scaling is fixed this may add more objects to our cluster
3. Removal of overlap amongst clusters can further assist the marketing team in identifying more unfulfilled sub niches or best marketing strategies amongst this cluster