MALL CUSTOMER SEGMENTATION

CLUSTER ANALYSIS & STATISTICS ANALYSIS

Presented By: Timileyin Anthony

CONTENT

- **Introduction**
- **Data Sources**
- **Data Preprocessing**
- **Exploratory Data Analysis (EDA)**
- **Univariate Clustering Analysis**
- **Bivariate Clustering Analysis**
- **Multivariate Clustering Analysis**
- **Statistics Analysis**
- **Recommendations**

Introduction

Welcome to our Customer Segmentation and Analysis project. In today's competitive market, understanding customer behavior is crucial for businesses to tailor their strategies effectively. This project leverages advanced data analysis techniques to segment customers based on demographic and spending patterns, providing actionable insights to enhance marketing strategies and boost customer loyalty.

Through a methodology approach, we've cleaned and preprocessed the data, explored key relationships, performed clustering to identify distinct customer groups, and conducted statistical tests to validate our findings. Our goal is to offer clear and practical recommendations that can help businesses attract, engage, and retain customers more efficiently.

Data Sources

The data used in this project was sourced from the Kaggle website, a popular platform for data science and machine learning datasets. The dataset contains information on mall customers, including demographic details and spending behavior.

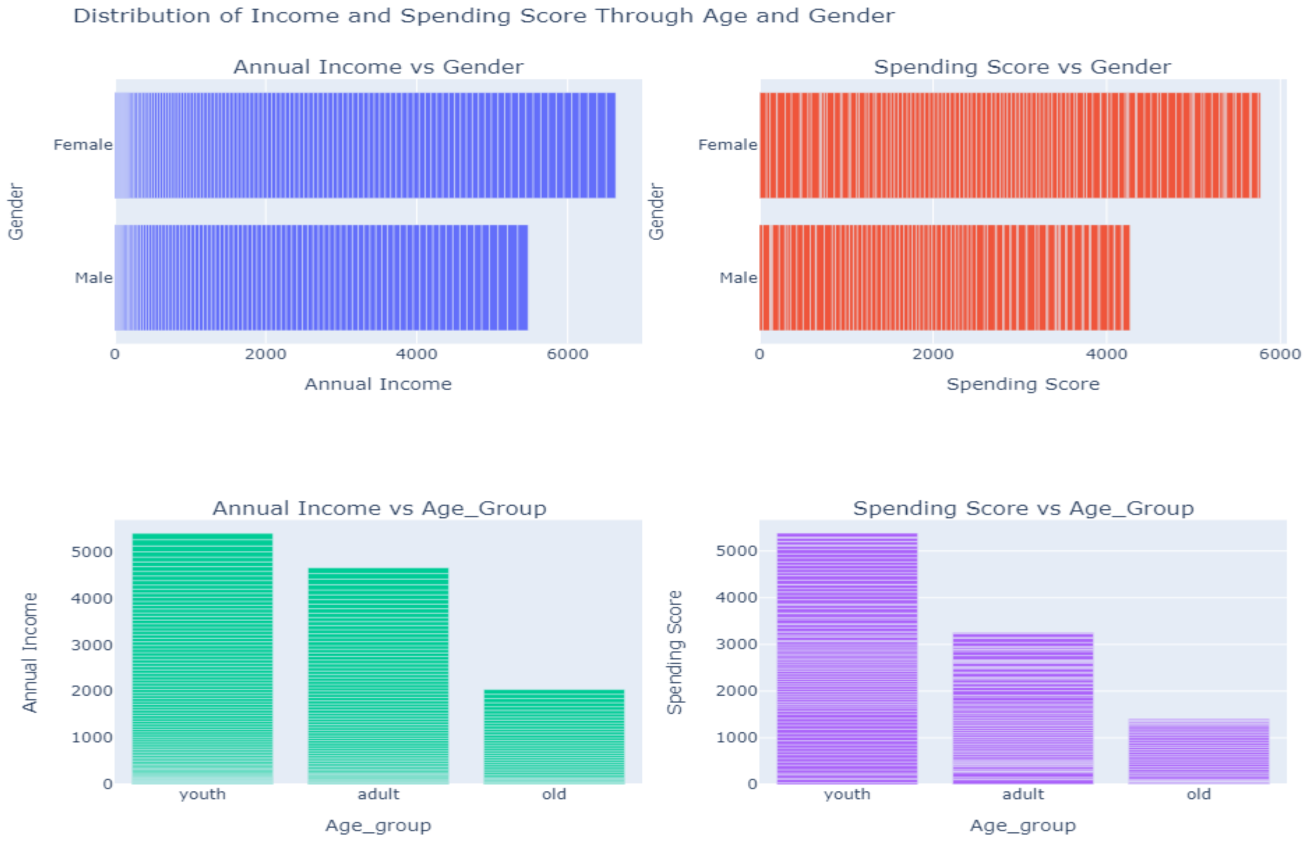
Dataset Details
Source: Kaggle
Link to Dataset: https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python
Description: The dataset includes the following columns:
CustomerID: Unique identifier for each customer.
Gender: Gender of the customer (Male/Female).
Age: Age of the customer.
Annual Income (k$): Annual income of the customer in thousand dollars.
Spending Score (1-100): Score assigned to the customer based on their spending behavior.

Data Preprocessing

In this project, I applied several data preprocessing steps to prepare the dataset for analysis

- Age Grouping: We categorized the age of customers into different groups (Youth, Adult, Old) using list comprehensions.
- Label Encoding: To convert the gender column into a numeric format.
- Standard Scaler: I applied StandardScaler to normalize the data by removing the mean and scaling to unit variance.
- Principal Component Analysis (PCA): PCA was used to reduce the dimensionality of the dataset.
- Eblow: This graph finds the optimal K value in clustering
- K-means: I performed K-means clustering to segment customers based on their spending behavior, age, and income levels.
- Univariate Analysis: I conducted a clustering analysis on individual variables (Annual Income and Spending Score) to understand their distribution and identify distinct segments.
- Bivariate Analysis: Clustering was performed to pair Annual Income and Spending Scores to analyze the relationship between them and identify customer segments.
- Multivariate Clustering: We extended the clustering analysis to multiple variables, and all available features to identify customer segments. This involved using PCA to simplify the feature space before applying Kmeans clustering.
- Statistical Analysis: I conducted a statistical analysis to examine the relationship

EXPLORATORY DATA ANALYSIS (EDA)



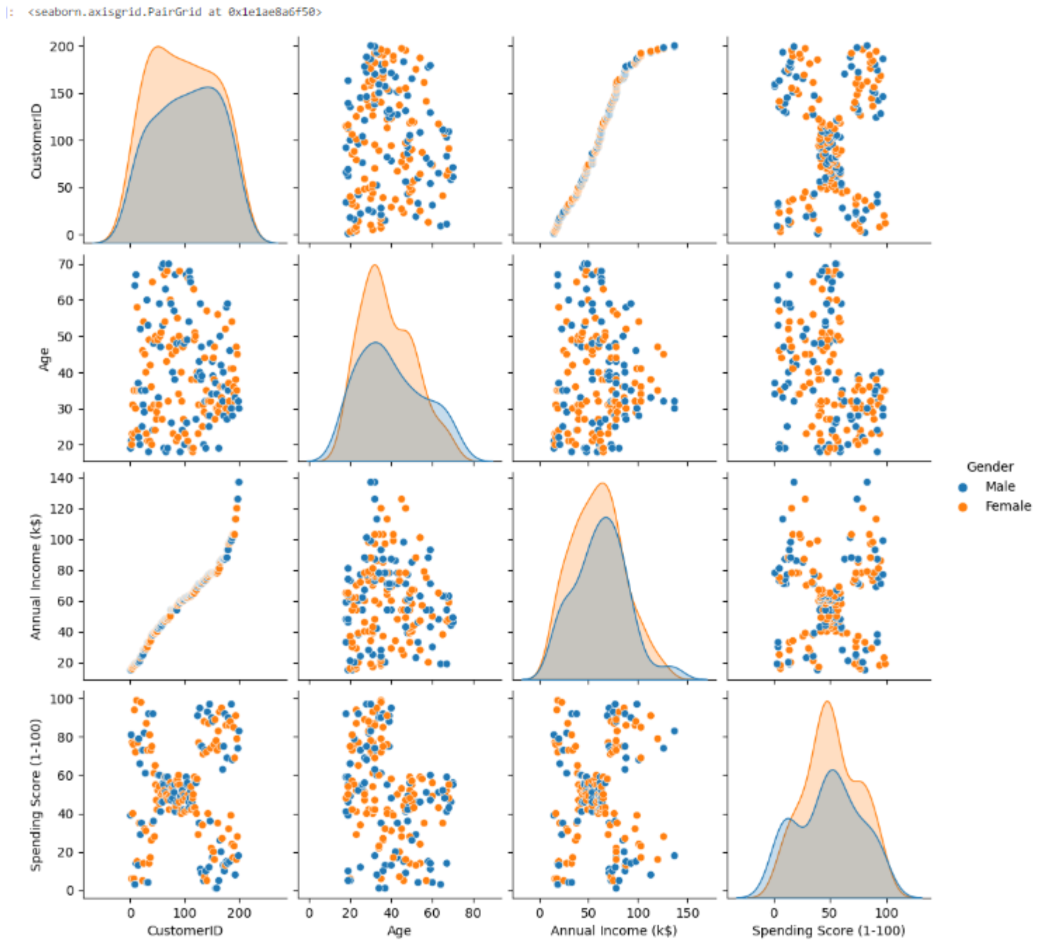
Interpretation

- Women earn and spend more money annually compared to men.
- Young people earn and spend more compared to adults and older people, who have the lowest earnings and spending scores

EXPLORATORY DATA ANALYSIS (EDA)

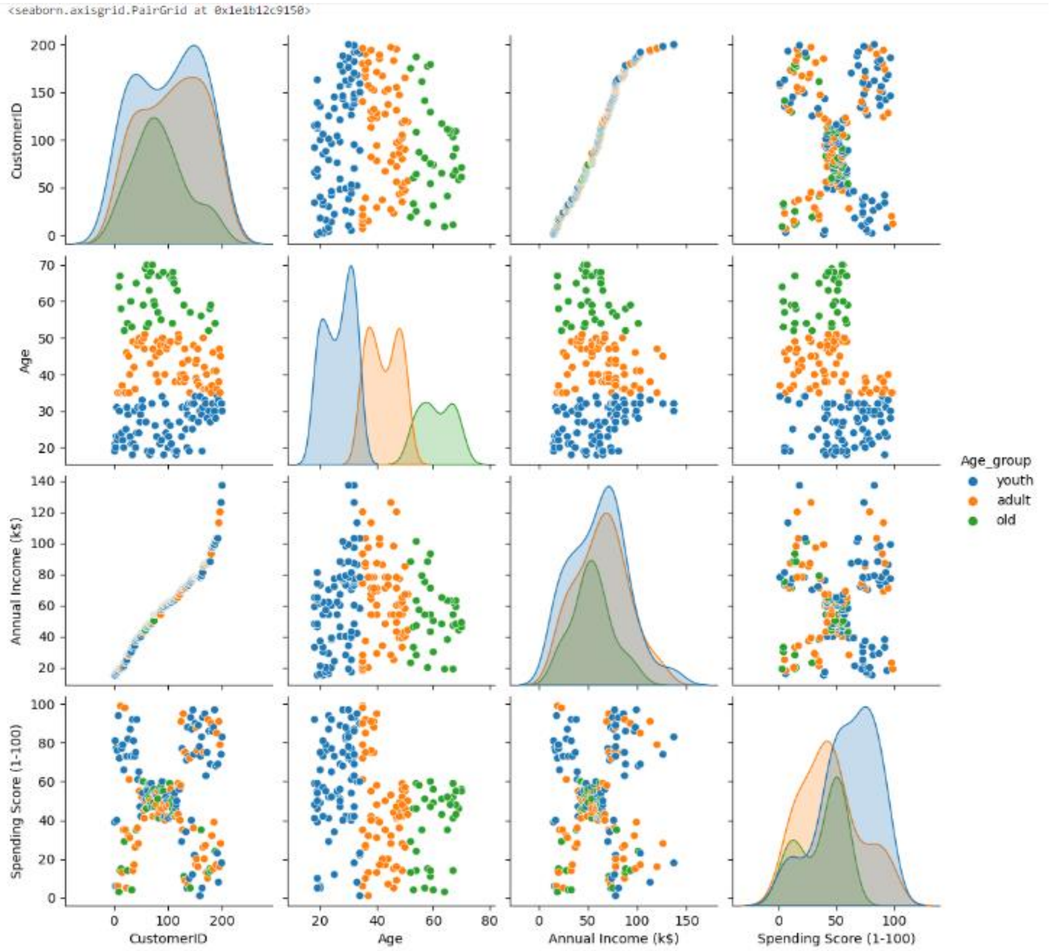
The Different Relationships of Gender and Age_Group in the Data

Gender



```
: <seaborn.axisgrid.PairGrid at 0x1e1ae8a6f50>
```

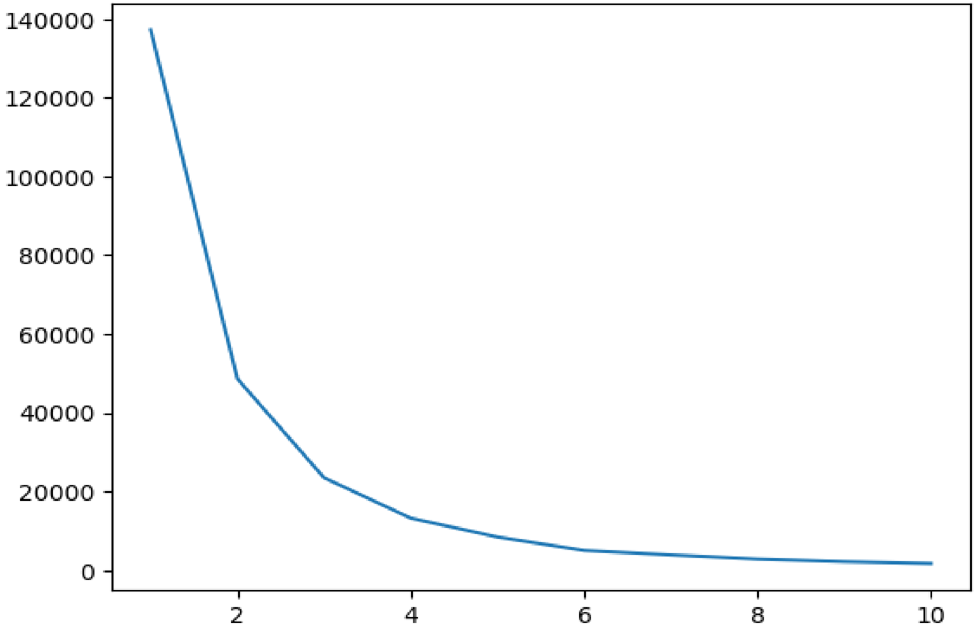
Age_Group



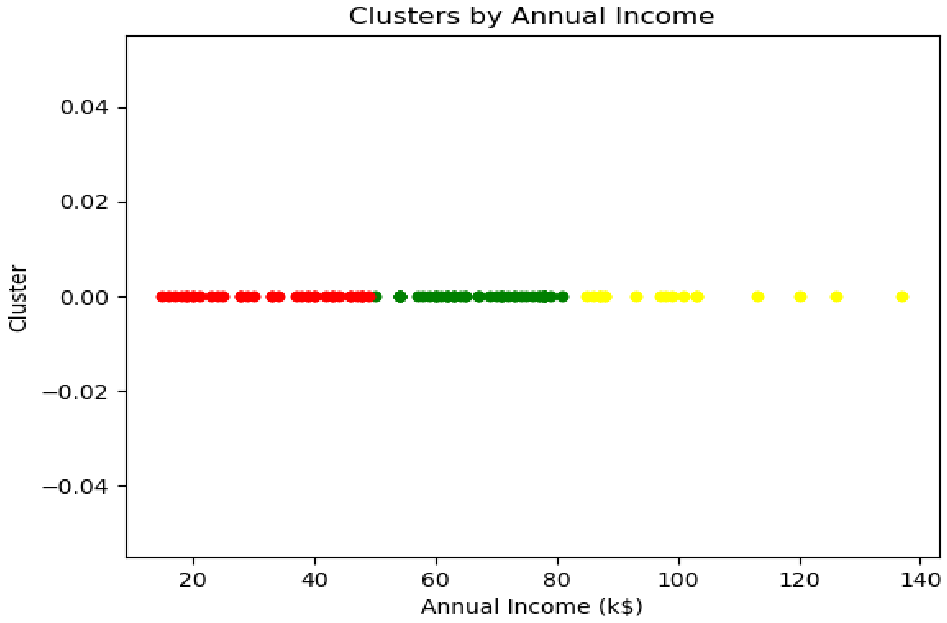
```
<seaborn.axisgrid.PairGrid at 0x1e1b12c9150>
```

Univariate Clustering Analysis

Annual Income



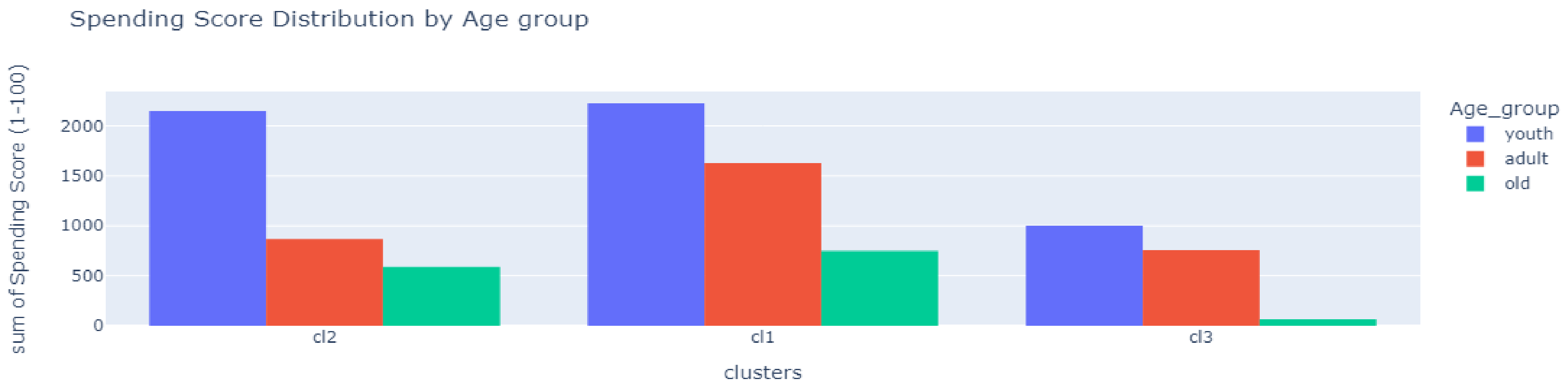
```
Text(0.5, 1.0, 'Clusters by Annual Income')
```



Interpretation of the eblow graph and findings

- The elbow point on the graph occurred at 3 clusters
- This suggests that dividing the data into 3 clusters
- There are three distinct groups of customers based on their annual income levels

Annual Income Clusters on Spending Score by Age Group



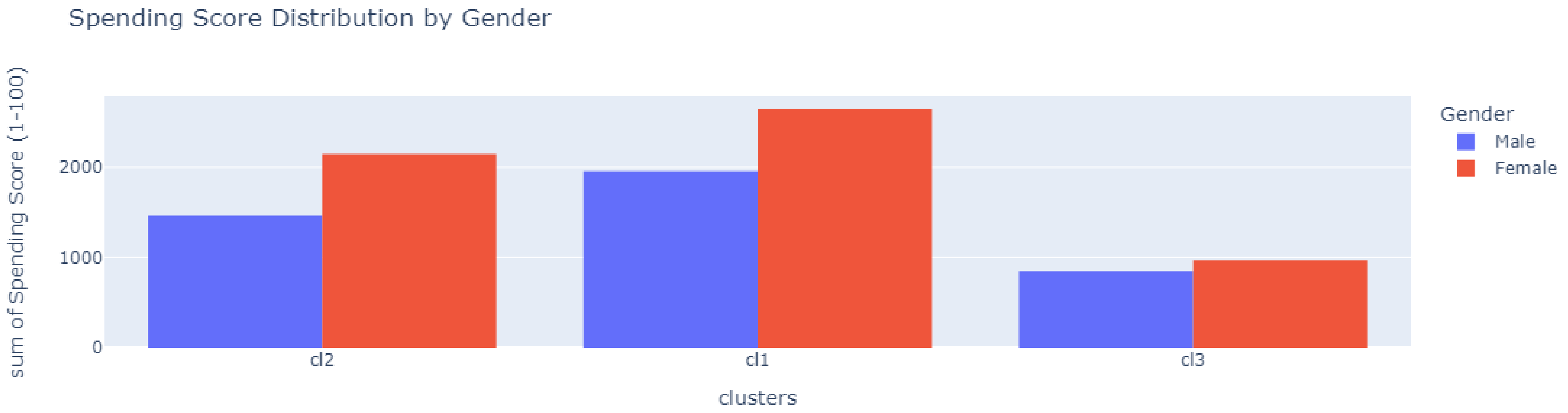
Cluster Interpretation

- In cl2 youth age group has the highest number of customers, followed by the adult age group and then the old age group
- cl1 also has a large number of customers in the youth age group, followed by the adult age group and then the old age group
- In cl3 has the lowest number of customers overall

Recommendations

- For Cluster 2 and Cluster 1, Tailor marketing strategies and promotions that appeal to each age group's spending because they have a meaningful number of customers across all age groups

Annual Income Clusters on Spending Score by Gender



Cluster Interpretation

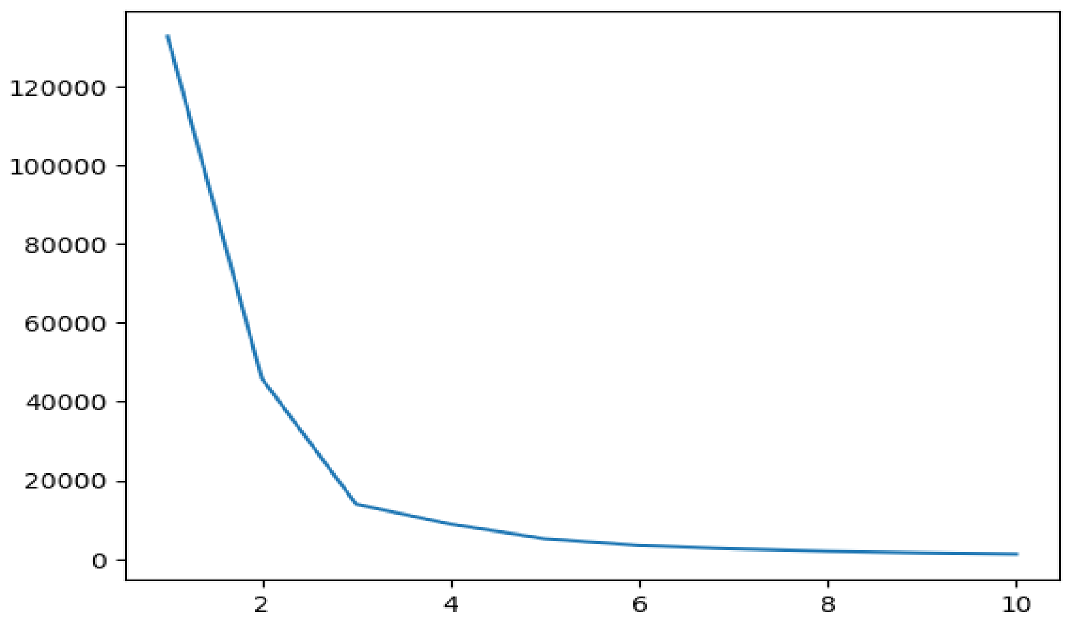
- In all clusters, there are more female customers than male customers
- Spending scores vary, but the trend of females being more general remains consistent

Recommendations

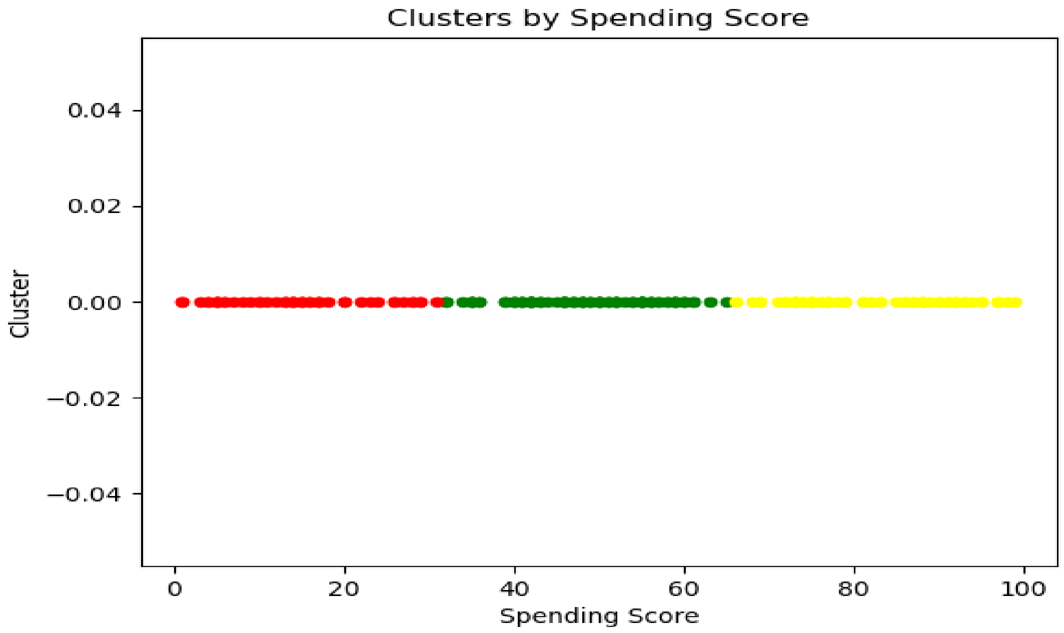
- Develop marketing campaigns that are particularly appealing to female customers
- Create promotions and offers that are tailored to both genders

Spending Score

```
[<matplotlib.lines.Line2D at 0x1e1b2584f10>]
```



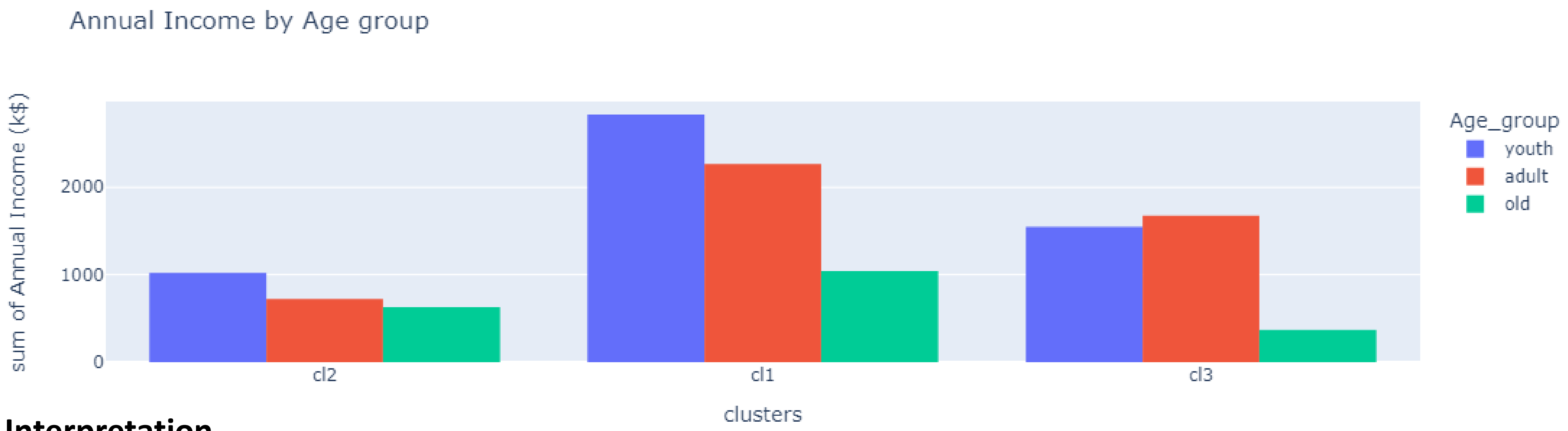
```
Text(0.5, 1.0, 'Clusters by Spending Score')
```



Interpretation of the eblow graph and findings

- The elbow point on the graph occurred at 3 clusters
- This suggests that dividing the data into 3 clusters
- There are three distinct groups of customers based on their Spending Score

Spending Scores Clusters on Annual Income by Age Group



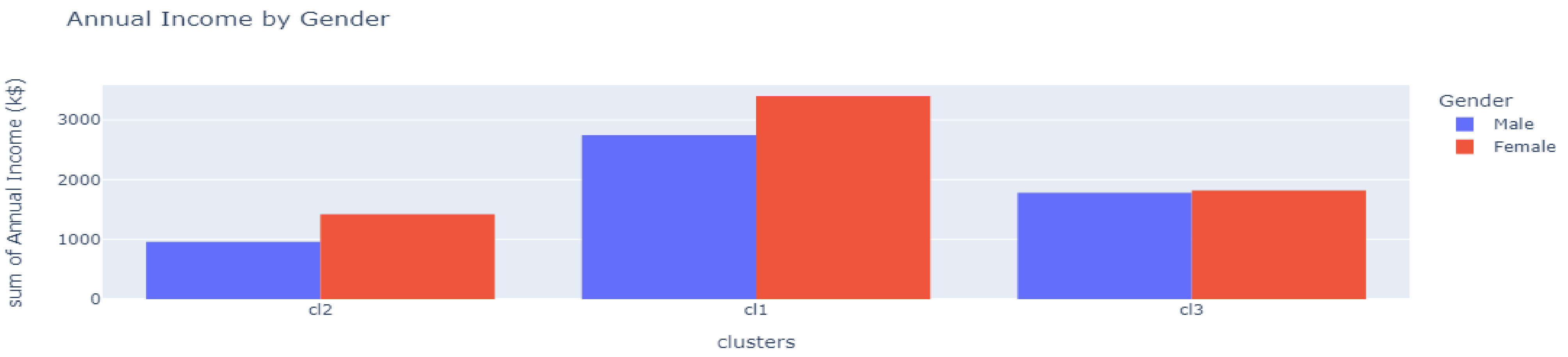
Interpretation

- Cluster 2 has the least number of customers in each age group
- Cluster 1 has the largest number of customers across all age groups
- Cluster 3 has a higher number of adults compared to youths and a smaller number of older customers.

Recommendations

- Focus on niche marketing strategies that appeal to each age group in cluster 2
- In cluster 1, Implement loyalty programs and incentives for young customers to keep them engaged and increase their lifetime value.
- In cluster 3 consider promoting products and services that cater to mid_life needs and Target older customers with specialized offers to increase their engagement and spending

Spending Scores Clusters on Annual Income by Age Group



Interpretation

- Cluster 1 Indicates a higher income group.
- Cluster 2 Has the fewest number of customers, with more females than males.
- Cluster 3 Has a nearly equal number of males and females

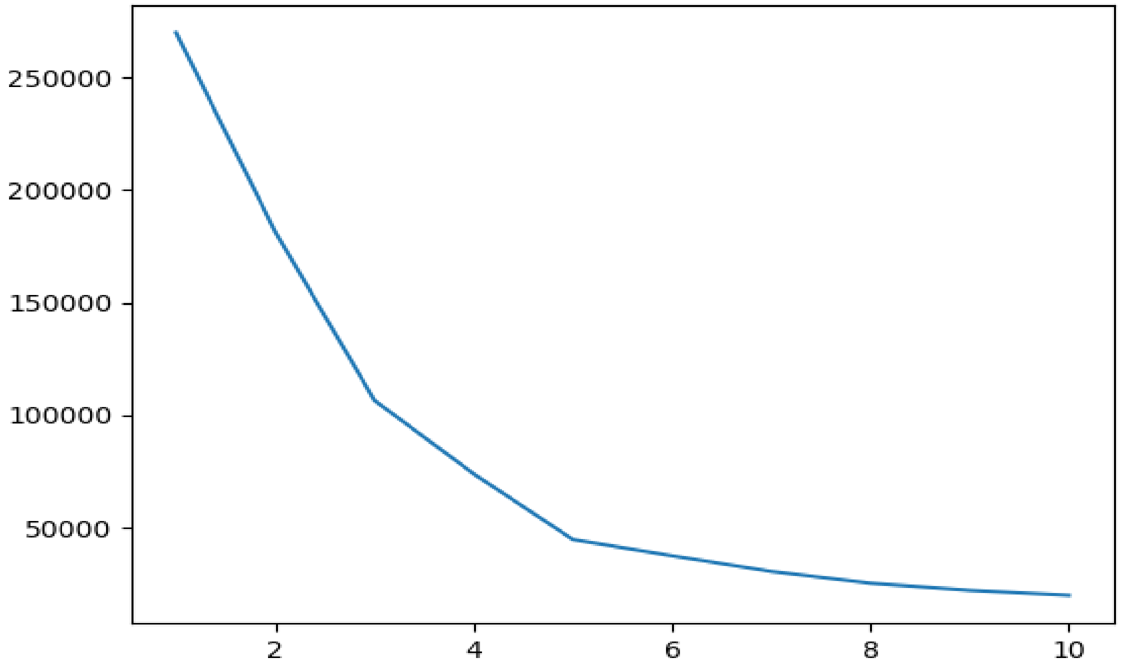
Recommendations

- Cluster 1, Focus on high-income products and premium services and Create gender-specific marketing campaigns to appeal to both males and females.
- Cluster 2, Develop targeted offers to attract more male customers and Promote mid-range products that might appeal to a broader audience.
- Cluster 3, Design inclusive marketing strategies that appeal equally to both males and females. and Offer a variety of products catering to different preferences

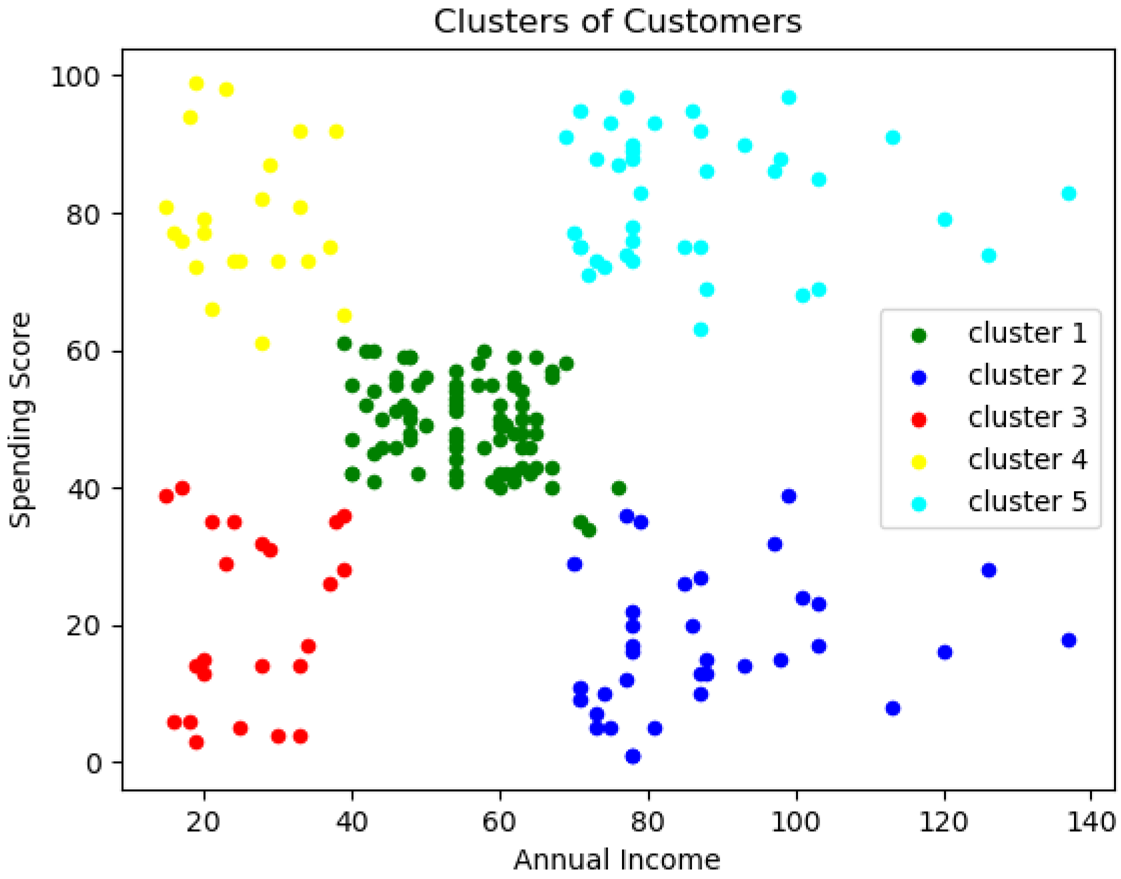
Bivariate Clustering Analysis

Annual Income (k$) vs Spending Score (1-100)

```
[<matplotlib.lines.Line2D at 0x1e1b45c0a10>]
```



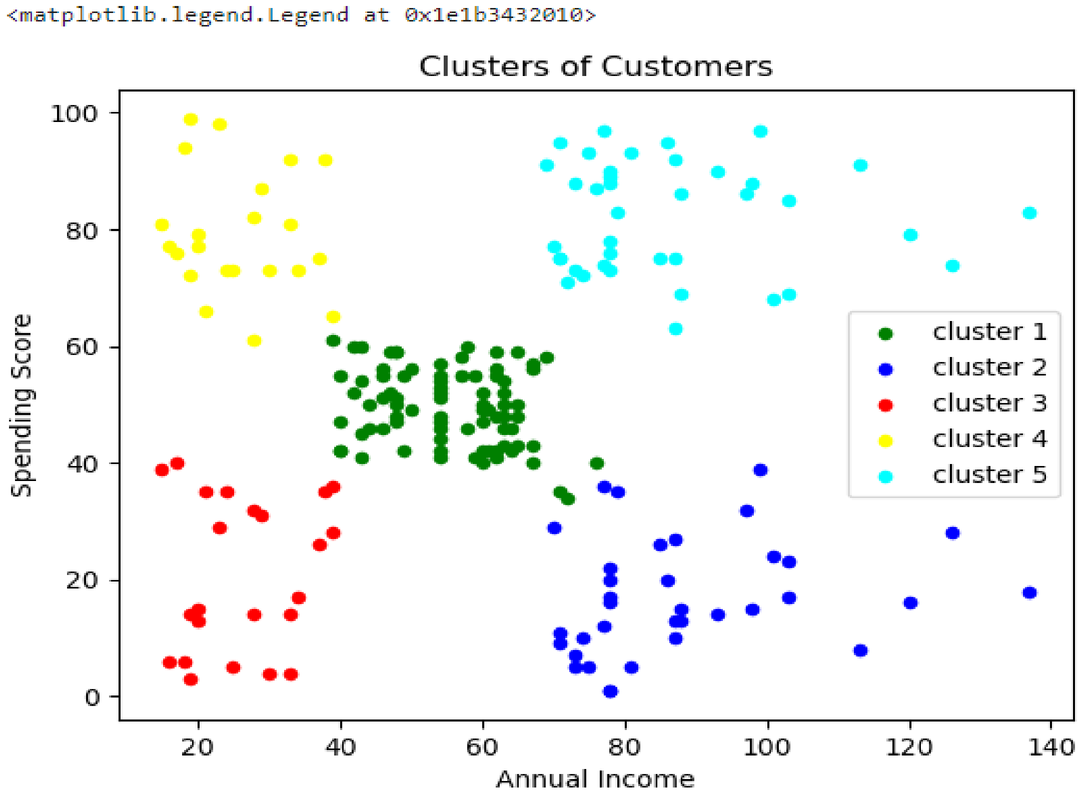
```
<matplotlib.legend.Legend at 0x1e1b3432010>
```



Interpretation of the eblow graph and findings

- This elbow shows 5 clusters.
- 5 clusters is the optimal number for segmenting the data
- Divide the customers into 5 distinct groups

Annual Income (k$) vs Spending Score (1-100)



```
<matplotlib.legend.Legend at 0x1e1b3432010>
```

Recommendations

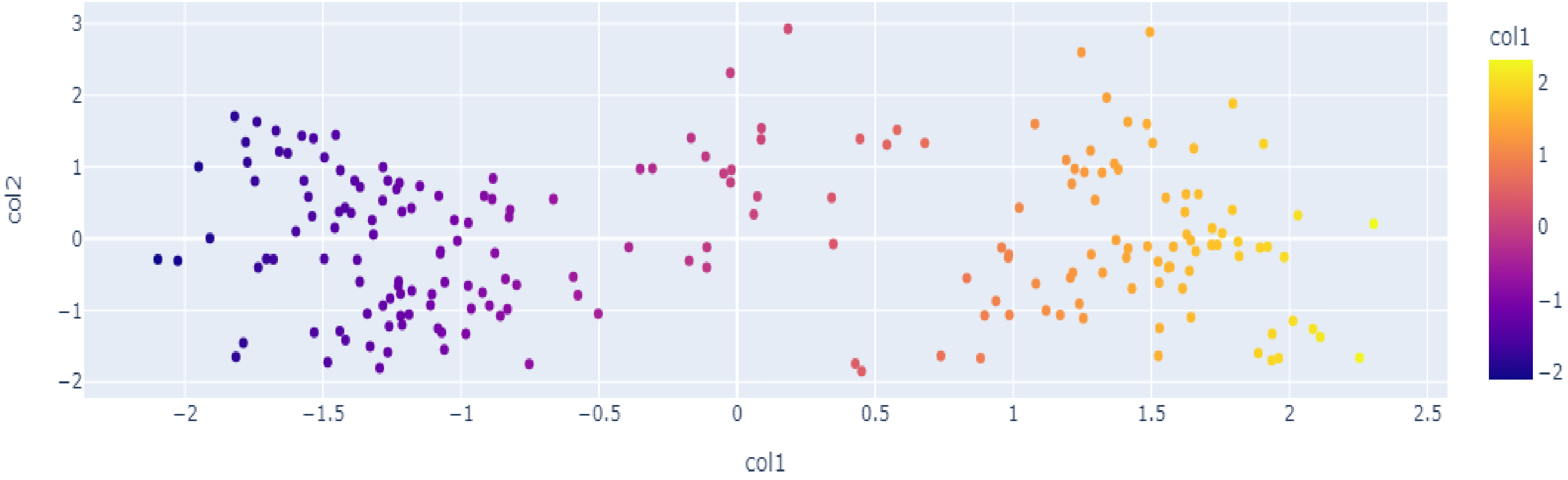
- Cluster 1 = Offer mid-range products and regular promotions to encourage consistent spending.
- Cluster 2 = Implement campaigns that highlight the value and benefits of higher-priced products
- Cluster 3 = Focus on budget-friendly options and frequent discounts
- Cluster 4 = Provide incentives for frequent purchases to keep them engaged.
- Cluster 5 = Offer premium and exclusive products, personalized services, and VIP experiences.

Interpretation

- Cluster 1 = Customers with moderate income and moderate spending scores
- Cluster 2 = Customers with high income but low spending scores
- Cluster 3 = Customers with low income and low spending scores
- Cluster 4 = Customers with low income but high spending scores
- Cluster 5 = Customers with high income and high spending scores

Multivariate Clustering Analysis

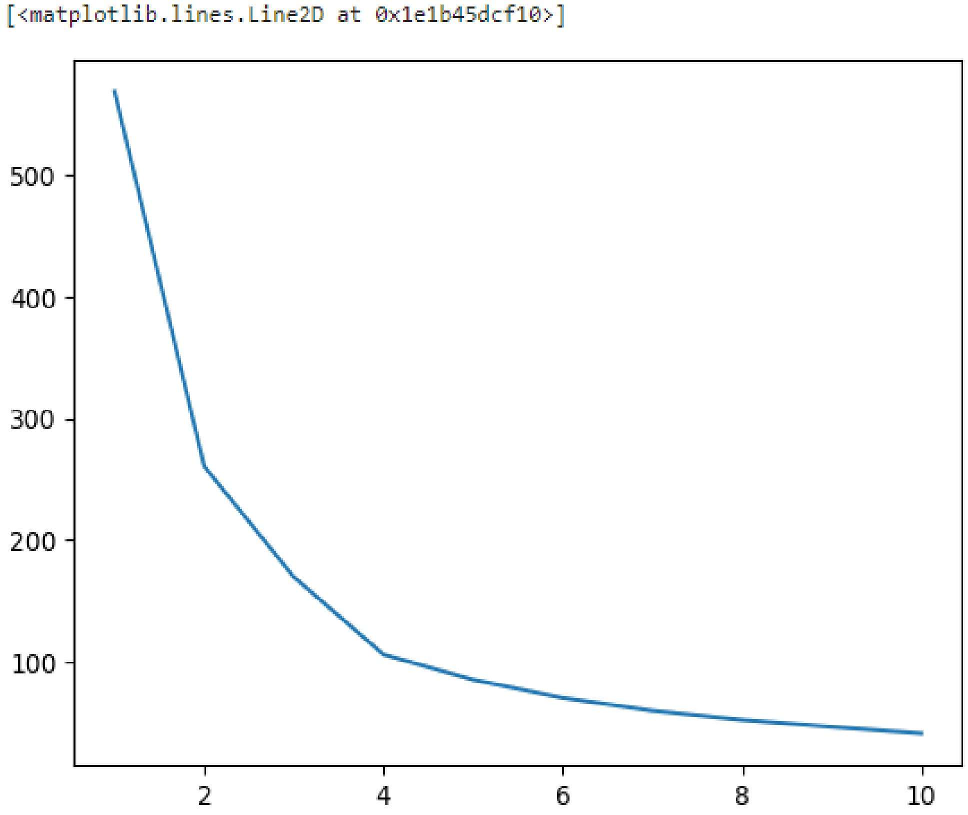
Data Modeling



After Standardize

Reduce the column to two using PCA

Multivariate Eblow Analysis



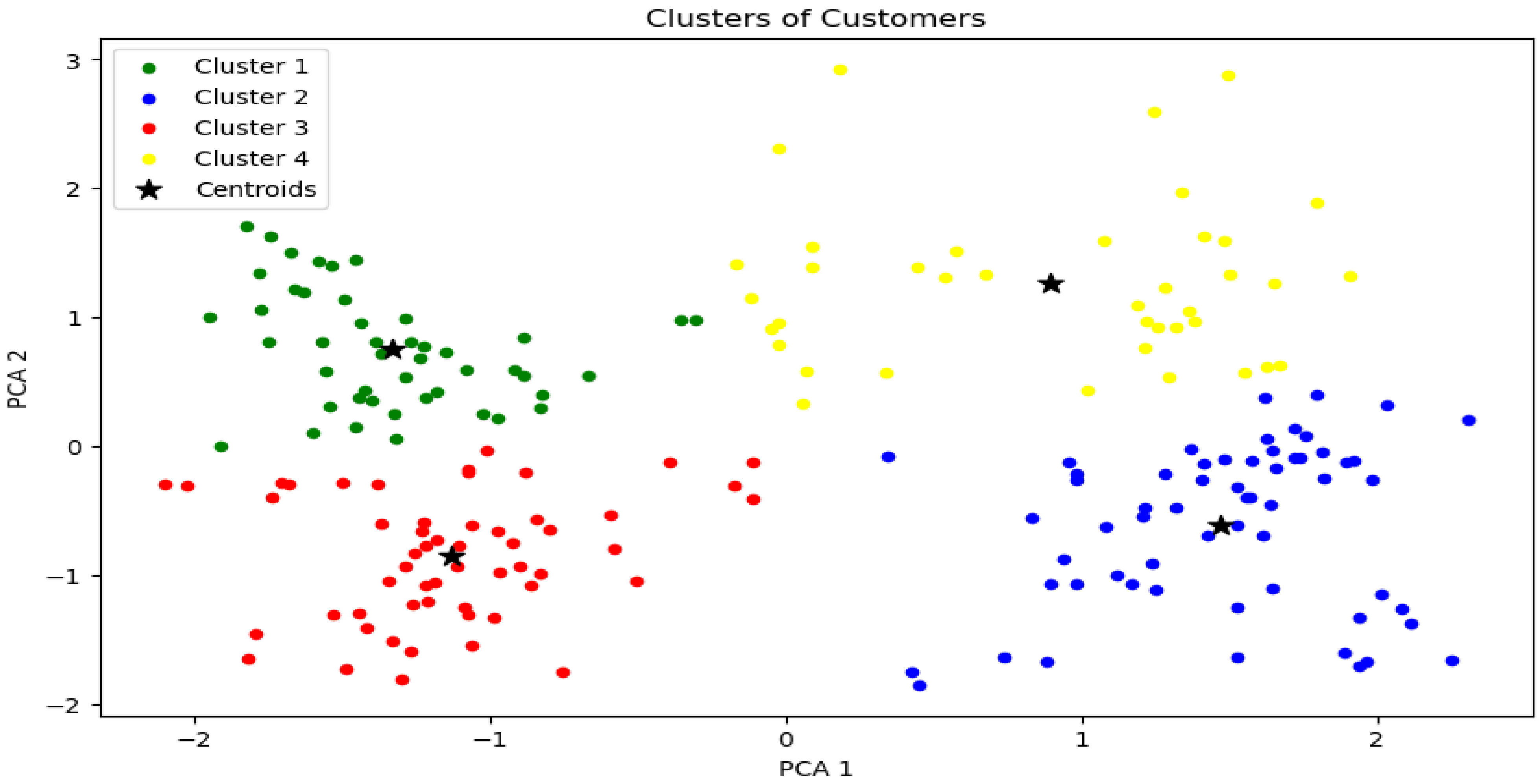
```
[<matplotlib.lines.Line2D at 0x1e1b45dcf10>]
```

Interpretation of the eblow graph and findings

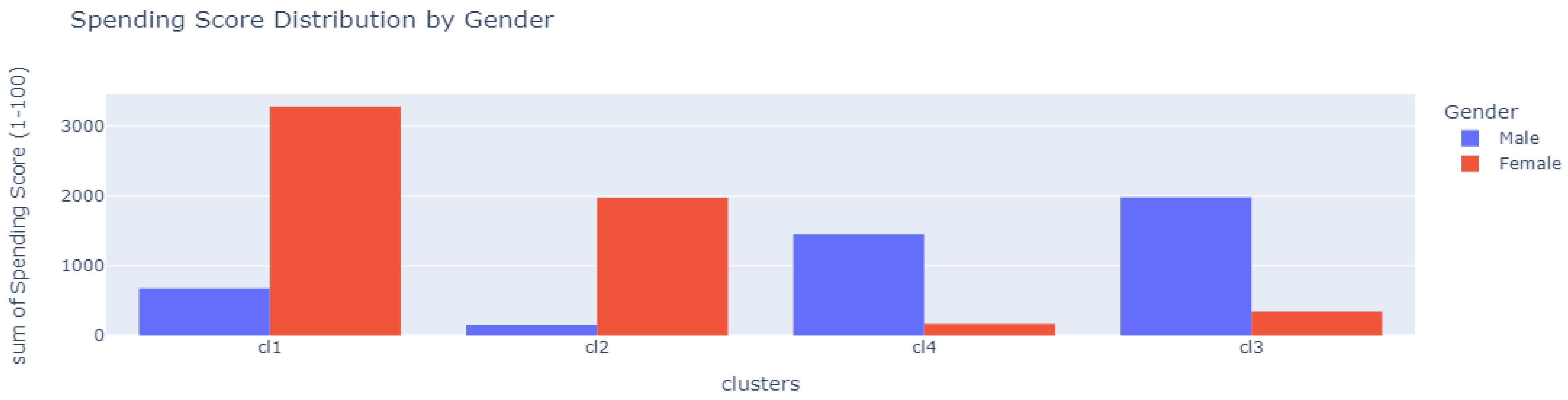
- The number of clusters shows an elbow at 4 clusters
- Indicates that 4 clusters is the optimal number for segmenting the data

Multivariate Cluster Chart

```
<matplotlib.legend.Legend at 0x1e1b4c4c550>
```



Multivariate Cluster Analysis on Spending Score by Gender



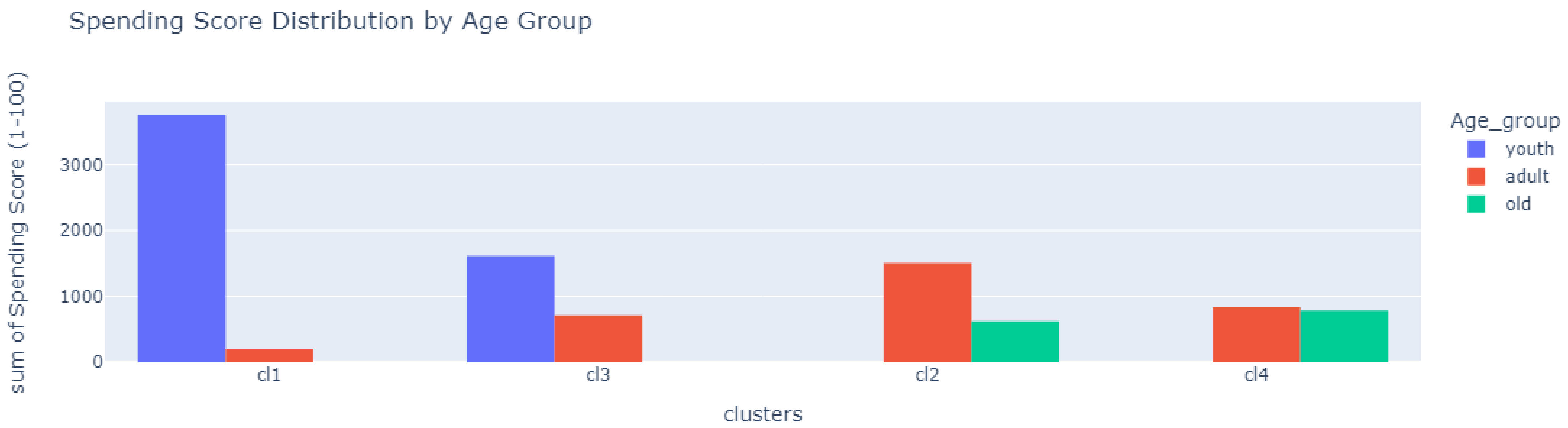
Interpretation

- Cluster 1 and Cluster 2, Have a significantly higher number of female customers compared to males.
- Cluster 3 and Cluster 4, Have a significantly higher number of male customers compared to females.

Recommendations

- In Cluster 1 and Cluster 2, Focus marketing campaigns and promotions on products and services that appeal to female customers.
- For Cluster 3 and Cluster 4, Develop marketing strategies that target male customers, offering products and services that cater to their interests.

Multivariate Cluster Analysis on Spending Score by Age Group



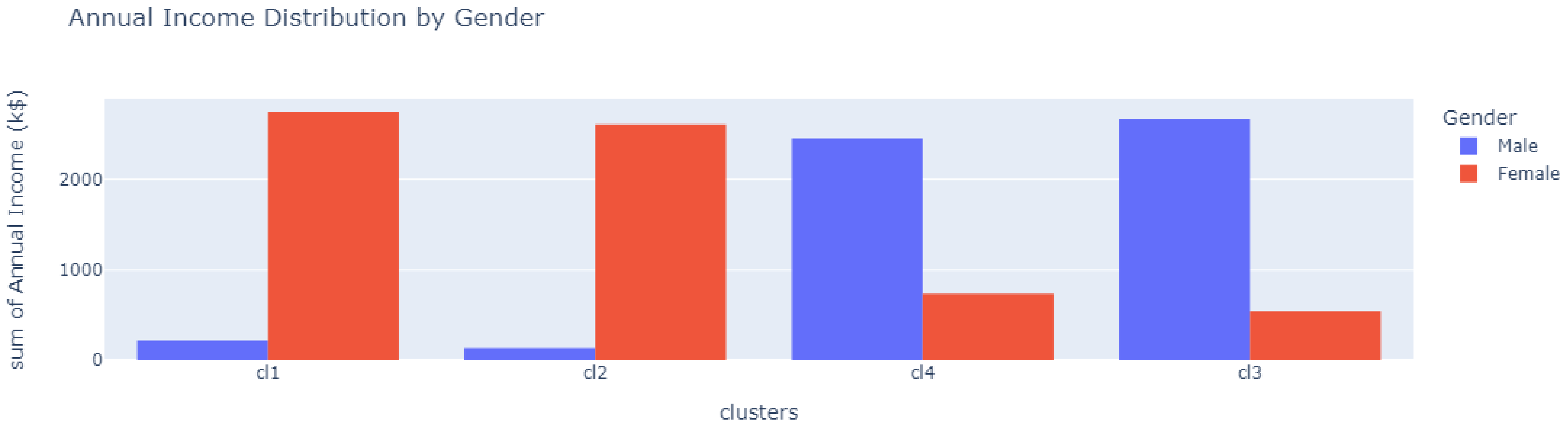
Interpretation

- Cluster 1 is Dominated by youth customers, indicating a younger with high spending scores.
- Cluster 3 Mostly youth but with a notable adult presence.
- Cluster 2 Adults with a significant number of older customers, indicating mature spending habits.
- Cluster 4 Balanced between adults and older customers, suggesting stable and mature spending patterns.

Recommendations

- Cluster 1 Focus marketing efforts on products and services that appeal to young customers.
- Cluster 3 Develop marketing strategies that appeal to both youth and adults.
- Cluster 2 Target marketing towards adults and older customers.
- Cluster 4 Create campaigns that appeal to both adults and older customers.

Multivariate Cluster Analysis on Annual Income by Gender



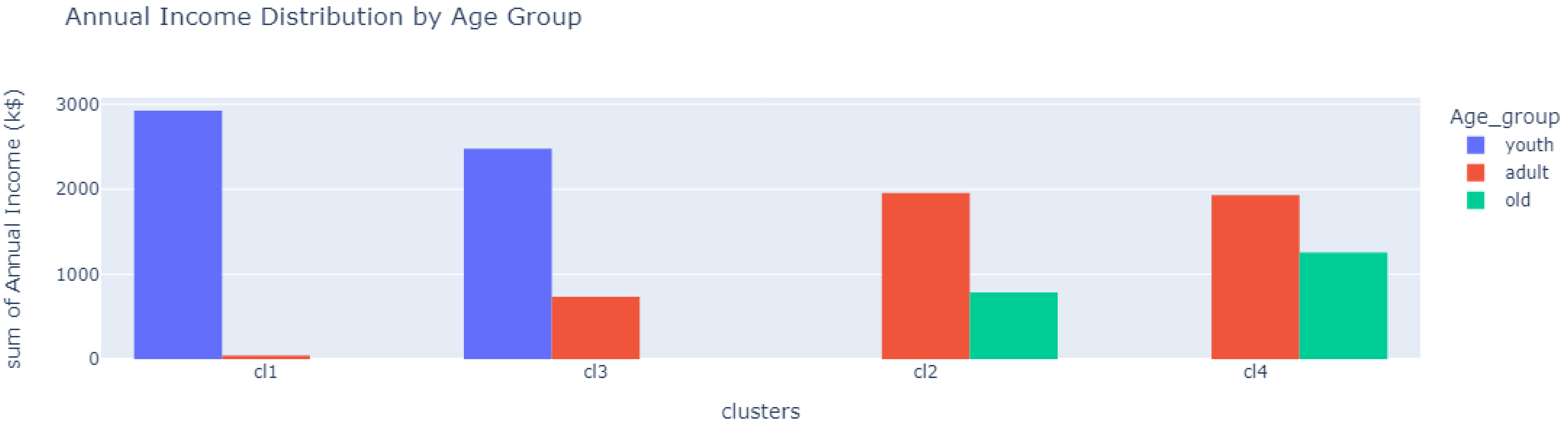
Interpretation

- Cluster 1 and Cluster 2 are dominated by female customers with high annual incomes
- Cluster 3 and Cluster 4 have a higher number of male customers with high annual incomes.

Recommendations

- Cluster 1 and Cluster 2, Focus marketing on high-income female customers. And Promote luxury products, exclusive offers, and premium services that appeal to affluent women.
- Cluster 3 and Cluster 4, Target high-income male customers. And Highlight high-end products and services that cater to their interests.

Multivariate Cluster Analysis on Annual Income by Age Group



Annual Income Distribution by Age Group

Interpretation

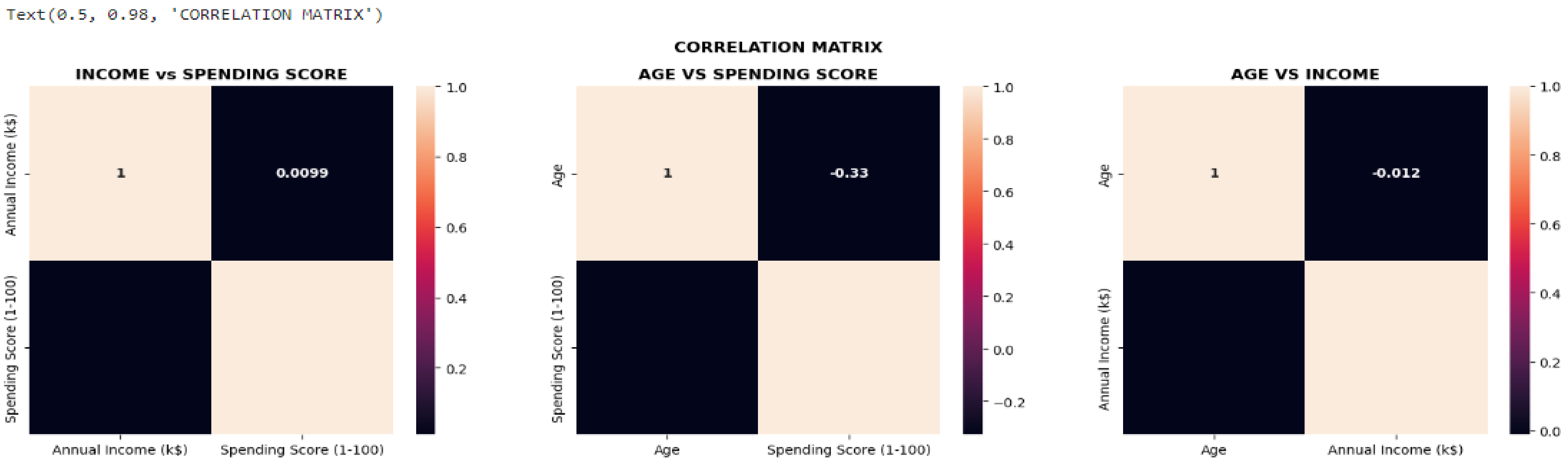
- Cluster 1 and Cluster 3 are dominated by youth customers with higher annual incomes.
- Cluster 2 and Cluster 4 have a mix of adult and older customers with moderate to high annual incomes.

Recommendations

- Cluster 1 and Cluster 3 Target high-income youth customers and Promote trendy products
- Cluster 2 and Cluster 4 Develop marketing strategies for both adult and older customers and Offer a range of products and services that cater to different age preferences and lifestyles.

Statistics Analysis

Correlation Matrix



```
Text(0.5, 0.98, 'CORRELATION MATRIX')
```

Interpretation of Correlation Results

Annual Income vs. Spending Score

•Correlation Coefficient: 0.009903

•The correlation is very close to zero.

•There is no significant linear relationship.

•Customers with different income levels have similar spending scores.

Age vs. Spending Score

•Correlation Coefficient:0.327227

•These indicate a weak negative relationship.

•As age decreases, the spending score decreases weakly.

Age vs. Annual Income:

•Correlation Coefficient: -0.012398

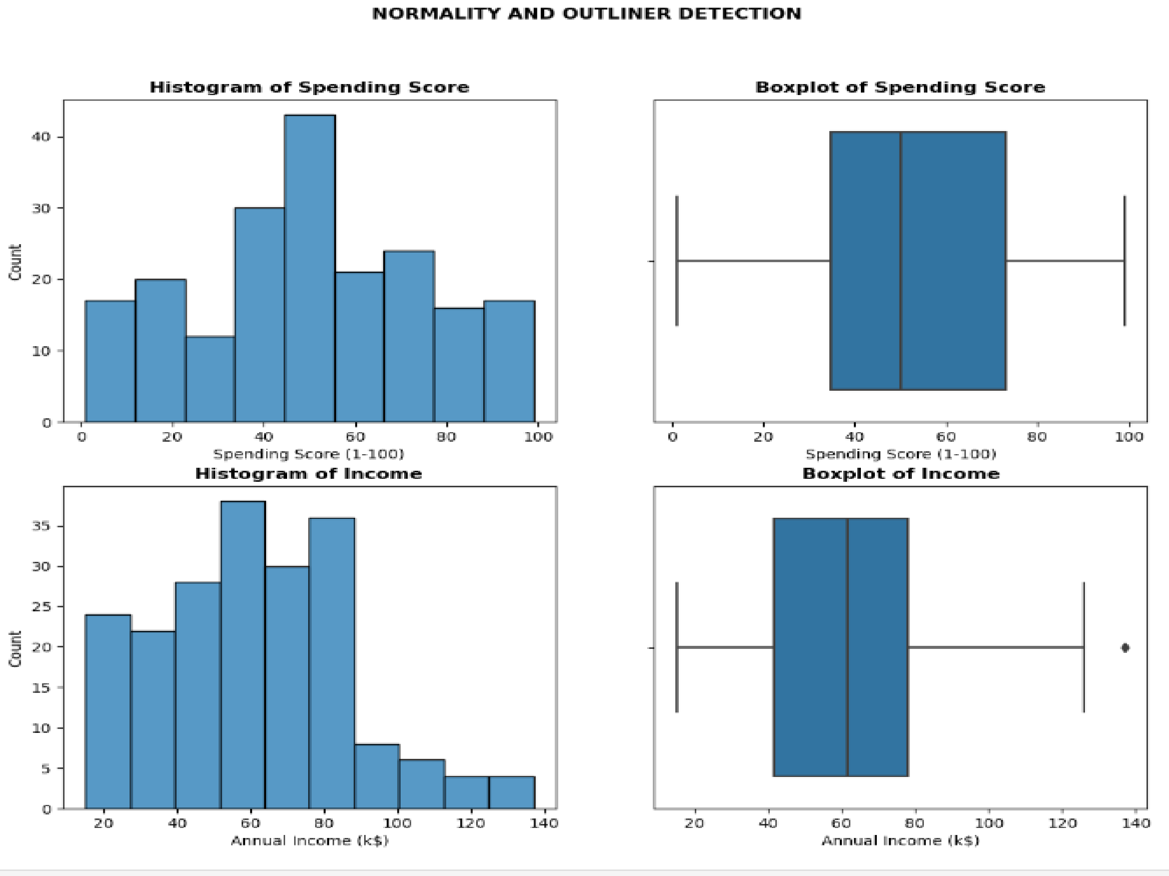
•The correlation is very close to zero but negative.

•There is no significant linear relationship between age and annual income.

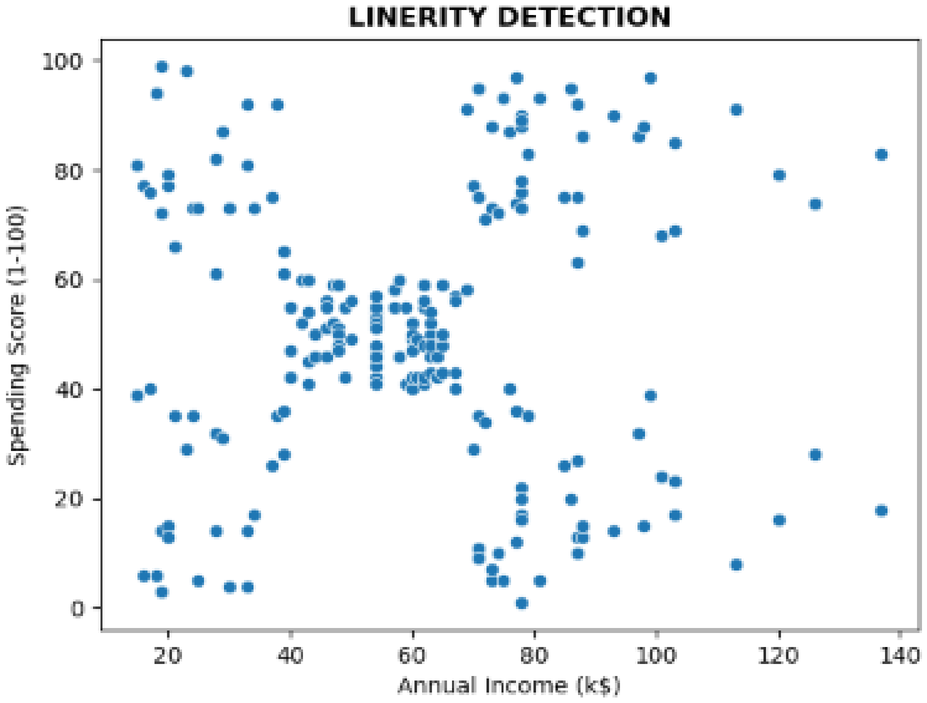
•Customers of different ages have similar annual incomes

Check assumptions of correlation before performing the Hypothesis Test

```
Text(0.5, 0.98, 'NORMALITY AND OUTLINER DETECTION')
```



```
Text(0.5, 1.0, 'LINERITY DETECTION')
```

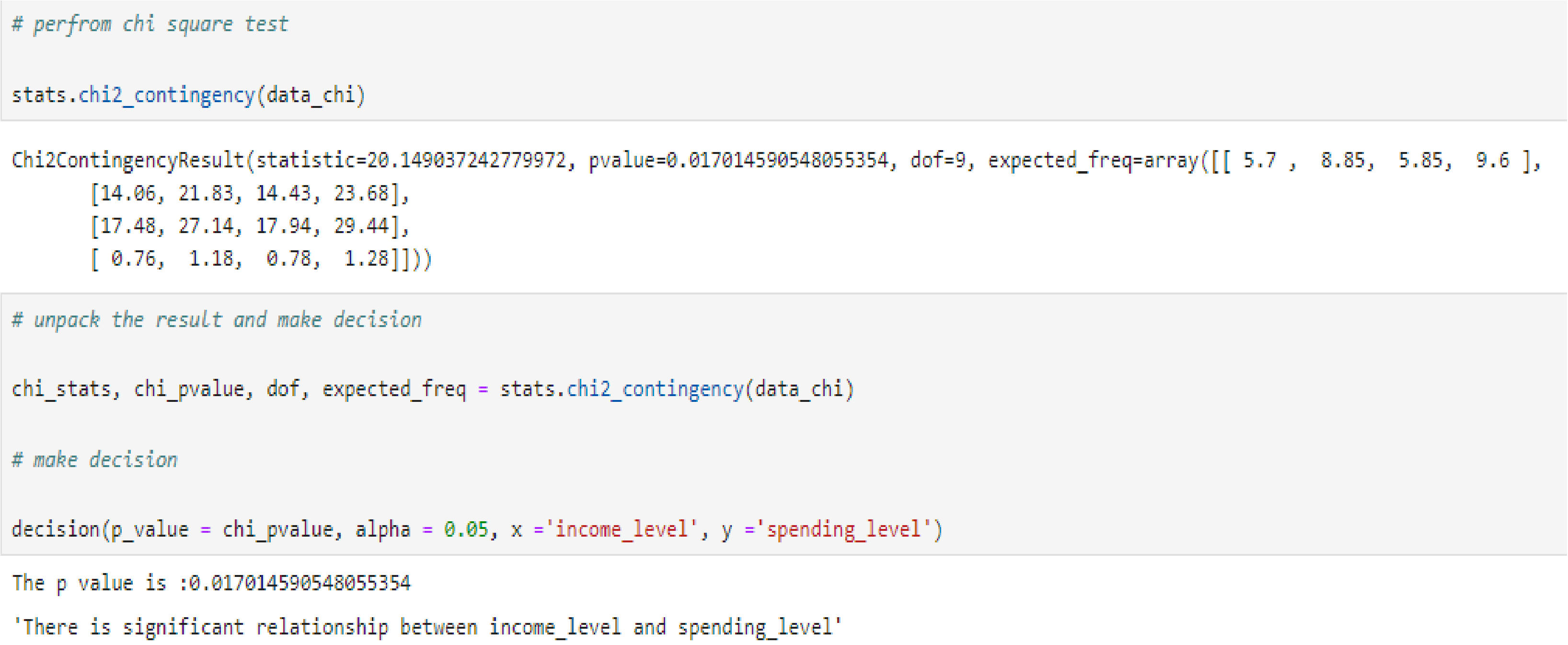


Interpretation of Assumption Checking

- After checking the assumptions for correlation, we found that only the continuous variable assumption is met

CHI SQUARE TEST OF ASSOCIATION

Test of Annual Income and Spending Scores



```
# perfrom chi square test

stats.chi2_contingency(data_chi)
```

```
Chi2ContingencyResult(statistic=20.149037242779972, pvalue=0.017014590548055354, dof=9, expected_freq=array([[ 5.7 ,  8.85,  5.85,  9.6 ],
       [14.06, 21.83, 14.43, 23.68],
       [17.48, 27.14, 17.94, 29.44],
       [ 0.76,  1.18,  0.78,  1.28]]))
```

```
# unpack the result and make decision

chi_stats, chi_pvalue, dof, expected_freq = stats.chi2_contingency(data_chi)

# make decision

decision(p_value = chi_pvalue, alpha = 0.05, x ='income_level', y ='spending_level')
```

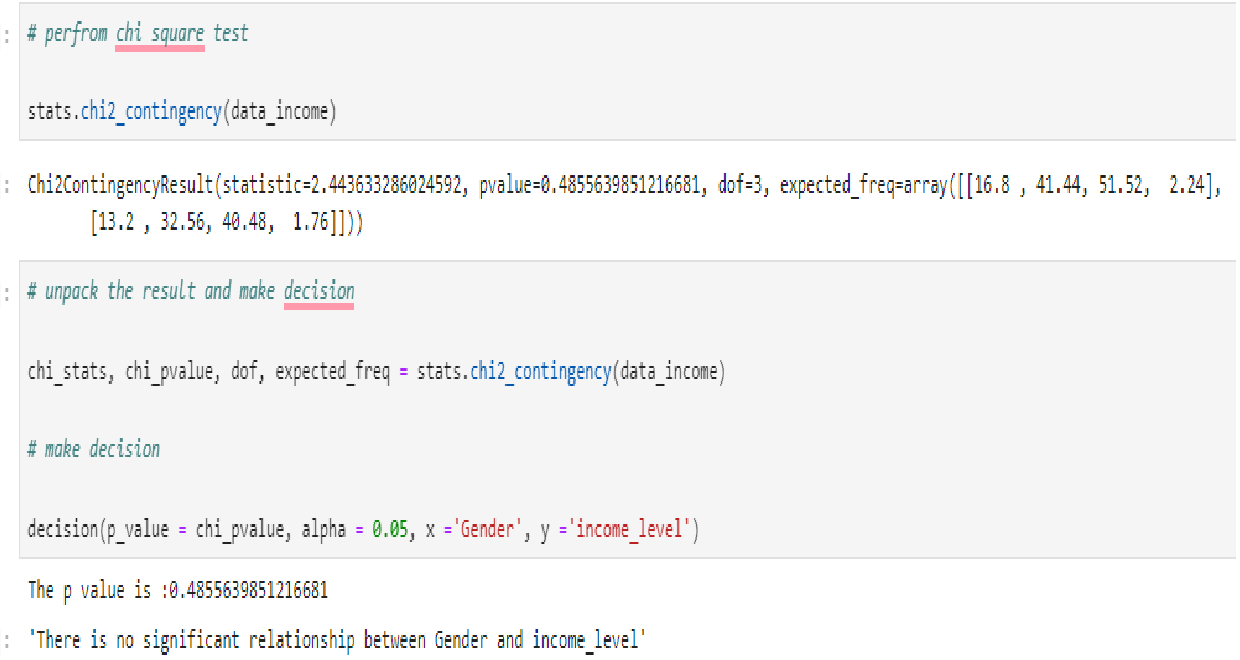
```
The p value is :0.017014590548055354
'There is significant relationship between income_level and spending_level'
```

Chi-Square Test Finding

- The p-value of 0.017 means there is a significant relationship between income levels and spending levels
- This finding is statistically significant
- This indicates that how much someone earns is associated with how much they spend

Gender contributes to income level and spending level

Income level



```
# perfrom chi square test

stats.chi2_contingency(data_income)
```

```
Chi2ContingencyResult(statistic=2.443633286024592, pvalue=0.4855639851216681, dof=3, expected_freq=array([[16.8 , 41.44, 51.52,  2.24],
       [13.2 , 32.56, 40.48,  1.76]]))
```

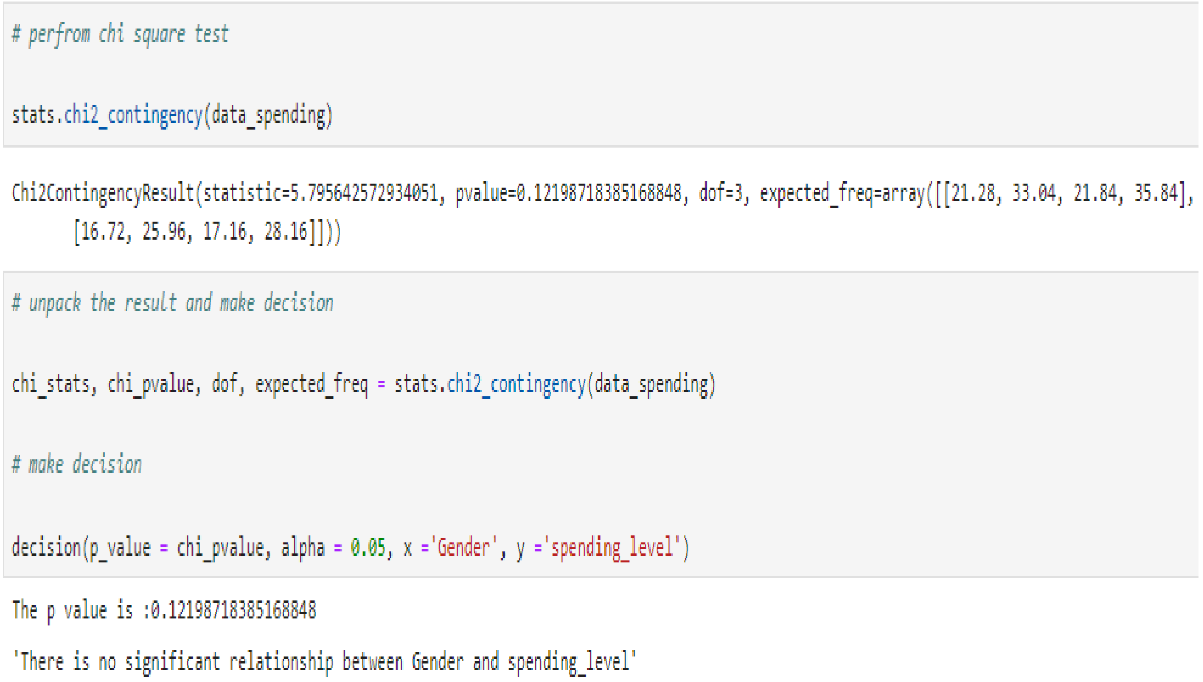
```
# unpack the result and make decision

chi_stats, chi_pvalue, dof, expected_freq = stats.chi2_contingency(data_income)

# make decision

decision(p_value = chi_pvalue, alpha = 0.05, x ='Gender', y ='income_level')
```

```
The p value is :0.4855639851216681
'There is no significant relationship between Gender and income_level'
```

Finding Interpretation

- The p-value of 0.485 means there is no significant relationship between gender and income level.
- This indicates that gender alone is not a strong predictor of income level

spending level



```
# perfrom chi square test

stats.chi2_contingency(data_spending)
```

```
Chi2ContingencyResult(statistic=5.795642572934051, pvalue=0.12198718385168848, dof=3, expected_freq=array([[21.28, 33.04, 21.84, 35.84],
       [16.72, 25.96, 17.16, 28.16]]))
```

```
# unpack the result and make decision

chi_stats, chi_pvalue, dof, expected_freq = stats.chi2_contingency(data_spending)

# make decision

decision(p_value = chi_pvalue, alpha = 0.05, x ='Gender', y ='spending_level')
```

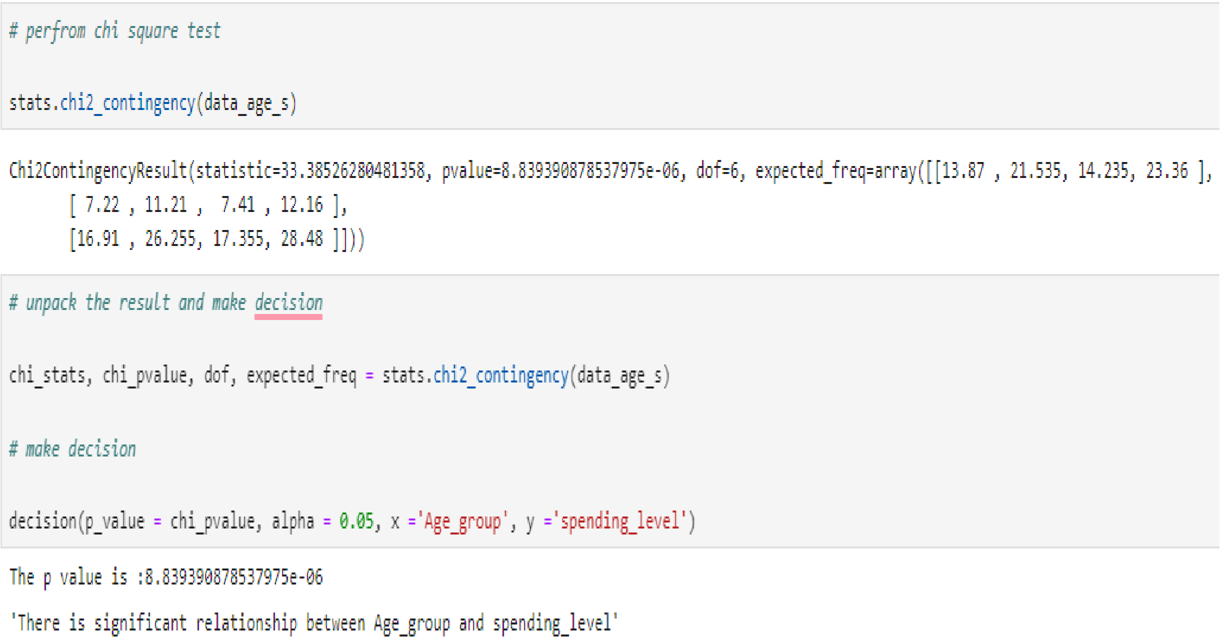
```
The p value is :0.12198718385168848
'There is no significant relationship between Gender and spending_level'
```

Finding Interpretation

- The p-value of 0.122 indicates there is no significant relationship between gender and spending level.
- This means gender is not a decisive factor in determining how much someone spends.
- These findings indicate that spending habits are not strongly influenced by gender alone

Age Group contributes to income level and spending level

Spending level



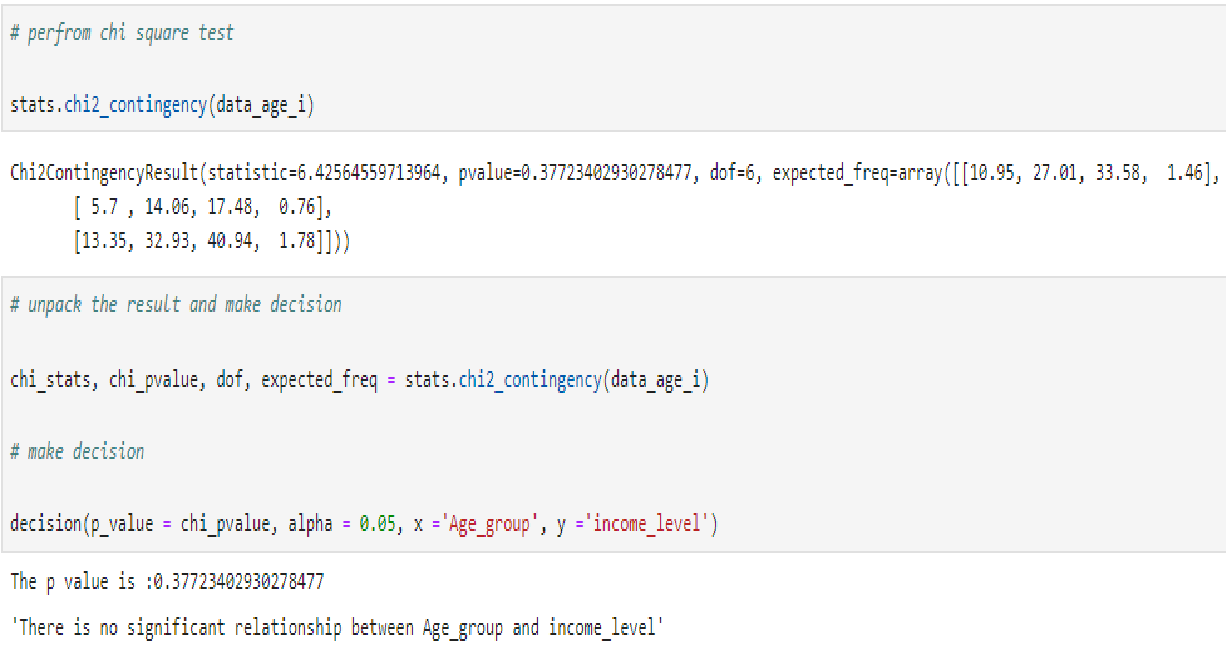
```
# perfrom chi square test

stats.chi2_contingency(data_age_s)
```

```
Chi2ContingencyResult(statistic=33.38526280481358, pvalue=8.839390878537975e-06, dof=6, expected_freq=array([[13.87 , 21.535, 14.235, 23.36 ],
       [ 7.22 , 11.21 ,  7.41 , 12.16 ],
       [16.91 , 26.255, 17.355, 28.48 ]]))
```

```
# unpack the result and make decision

chi_stats, chi_pvalue, dof, expected_freq = stats.chi2_contingency(data_age_s)

# make decision

decision(p_value = chi_pvalue, alpha = 0.05, x ='Age_group', y ='spending_level')
```

```
The p value is :8.839390878537975e-06
'There is significant relationship between Age_group and spending_level'
```

Income level



```
# perfrom chi square test

stats.chi2_contingency(data_age_i)
```

```
Chi2ContingencyResult(statistic=6.42564559713964, pvalue=0.37723402930278477, dof=6, expected_freq=array([[10.95, 27.01, 33.58,  1.46],
       [ 5.7 , 14.06, 17.48,  0.76],
       [13.35, 32.93, 40.94,  1.78]]))
```

```
# unpack the result and make decision

chi_stats, chi_pvalue, dof, expected_freq = stats.chi2_contingency(data_age_i)

# make decision

decision(p_value = chi_pvalue, alpha = 0.05, x ='Age_group', y ='income_level')
```

```
The p value is :0.37723402930278477
'There is no significant relationship between Age_group and income_level'
```

Finding interpretation

- The p-value (8.84e-06) indicates a significant relationship between age group and spending level.
- These data strongly mean that a person's age group is closely linked to how much they spend.
- This finding is statistically significant, That different age groups exhibit distinct spending behaviors.

Finding interpretation

- The p-value of 0.377 suggests that there is no significant relationship between age group and income level.
- These data do not show a meaningful connection between someone's age group and their income level.
- This finding indicates that age group alone is not a strong predictor of income level

Recommendations

Targeted Marketing

- Develop targeted marketing strategies based on age groups rather than gender or income alone.
- Age group has a significant impact on spending behavior
- Tailoring campaigns to different age demographics can lead to more effective results.

Customer Segmentation

- Use clustering techniques to group customers based on their spending habits and preferences.
- Create personalized experiences and promotions that resonate with each customer segment.

Product Placement

- Place products strategically based on spending patterns.
- For example, high-spending age groups could be targeted with premium products.
- while moderate spenders might respond well to value propositions.

Promotional Campaigns

- Design promotions and offers that appeal to specific age groups.
- Consider offering discounts or incentives that align with the spending behavior of each age demographic.

Continuous Monitoring

- Continuously monitor customer behavior and adjust strategies accordingly.

THANK YOU