Identifying Customer Spending and Satisfaction Patterns Through Supermarket Data Analysis

Minh Trang Le and Nguyen To To Nguyen

2024-05-12

1.0 Introduction

1.1 Project Description

In this project, our objective is to identify the key factors that drive sales performance and customer satisfaction as reflected in customer ratings. To achieve this, we analyzed the historical sales data of a supermarket. Specifically, we examined records from three different branches of a supermarket over a three-month period, spanning from December 31, 2018, to March 29, 2019. Our approach involved segmenting customer data based on Gender (Male and Female) and Customer Type (Member and Non-members), allowing us to gain insights into their spending patterns and overall satisfaction.

The first guiding question is: "Is there a relationship between total price spent and Customer Type or Gender?" We used hypothesis testing to determine whether there are statistically significant differences in total spending between Male and Female customers, as well as between Members and Normal (Non-members). Moreover, we utilized clustering techniques to identify distinct customer segments based on spending patterns and demographic characteristics. Our goal is to explore how total spending is linked to customer type and gender.

The second guiding question we address is: "What factors drive customer satisfaction based on the given ratings?" We used a linear regression model to associate input variables such as gender, customer type, quantity purchased and total spending to quantify the impact of each factor on rating. By doing so, we gained insights into which genders and customer types contribute most significantly to overall rating. Additionally, we employed K-nearest neighbor (KNN) method to assess whether there exists statistically significant differences in customer satisfaction across different genders or customer segments. This analysis helps us identify which specific customer groups are associated with higher levels of satisfaction.

The domain of interest is customer segmentation within the retail sector, specifically focusing on supermarket environments. Our primary objective is to extract valuable insights that can inform decision-making processes for retail businesses. By gaining a deeper understanding of customers, these businesses can upgrade their services to meet individual customer needs which can improve sales and service quality. Given the critical role customers play in the retail setting, gaining a deeper understanding of their preferences and behaviors can significantly enhance profit margins and brand image.

1.2 Background

Customer Segmentation in Retail

Customer segmentation is a marketing strategy that involves dividing the customer base into distinct groups based on shared characteristics. These characteristics can include demographics (age, gender, income level), purchasing behavior (frequency, amount spent), and product preferences. By segmenting customers, supermarkets can tailor marketing campaigns, product offerings, and promotional strategies to better resonate

with each group's specific needs and preferences. In our project, we segmented customers based on their gender and membership status.

Customer Satisfaction in the Supermarket Industry

Supermarkets face a highly competitive environment with multiple chains striving to secure customer loyalty. Understanding what drives customer satisfaction is crucial for supermarkets to retain customers and boost sales. In this project, we analyzed how various factors, including customer membership, gender, quantity purchased, and total spending, influence customer satisfaction as reflected in their ratings on a scale of 1 to 10.

2.0 Data Description

The data was downloaded as a csv file from Kaggle.

We used 5 variables for our analysis which are Customer.type, Gender, Quantity, Total, Rating.

The variables that we did not use are Invoice.ID, Branch, City, Product.line, Unit.price, Tax.5., Date, Time, Payment, cogs, gross.margin.percentage, gross.income as these variable were not directly related to our objectives of analyzing customer spending and satisfaction patterns.

The data fields:

- **Customer.type** - categorical: Type of customers
  - Members - customers using member card.
  - Normal - customers without member card.
- **Gender** - categorical: Gender type of customer (Male and Female)
- **Quantity** - numeric: Total number of products purchased by customer
- **Total** - numeric: Total price including tax
- **Rating** - numeric: Customer satisfaction rating on their overall shopping experience (On a scale of 1 to 10)

2.1 Data Exploration

We selected a subset of the dataset, focusing on 5 variables using the subset function. The data consisted of 1000 observations related to these 5 variables of interest.

There are no duplicates, missing values, or extreme outliers in this data set, so no modifications are necessary.

As Customer.Type and Gender are categorical variables, we changed the data types of these variables to factor.

2.2 Data Visualization

2.2.1 Visualization of the "Total" variable

We used a histogram to represent the distribution of the total spending by customers of the supermarket.

- The descriptive statistics of the "Total" variable:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.68  124.42  253.85  322.97  471.35 1042.65
```

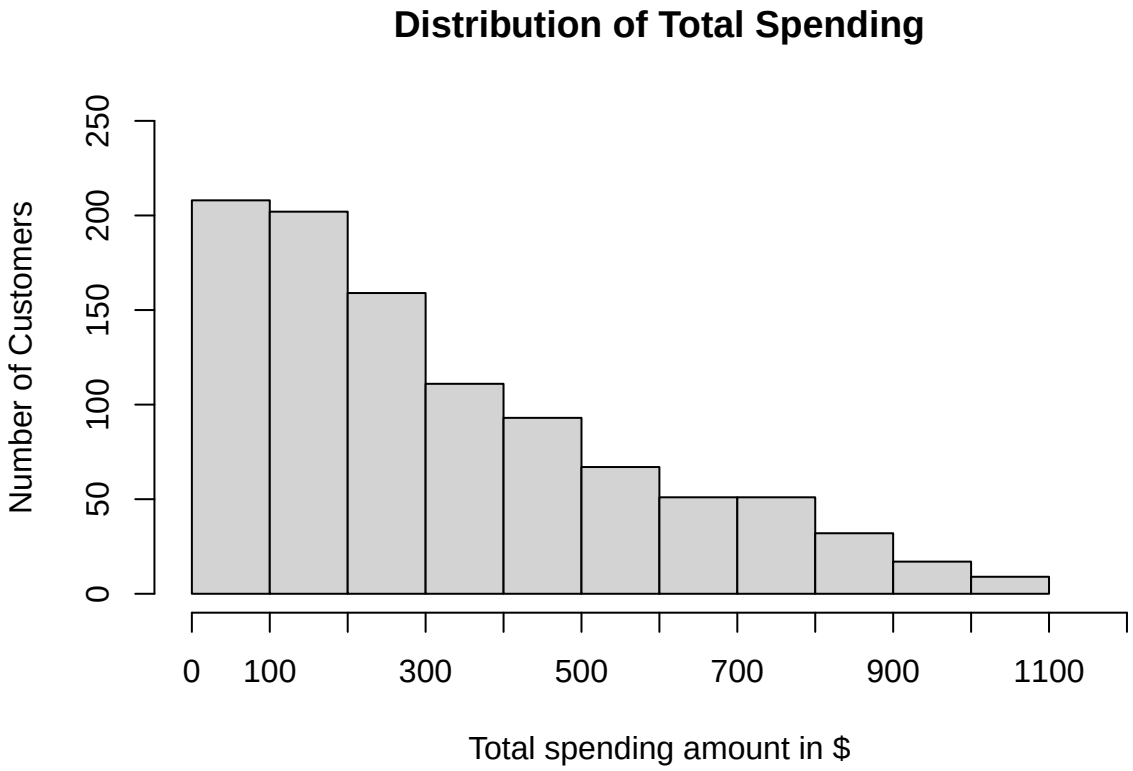

Figure 1: The Distribution of Total Spending

The x-axis represents the total spending amount of customers, binned into ranges. The bin cut points occur every $100, resulting in specific ranges such as: Less than $100, $100 to $200, $200 to $300, and so on, following increments of $100.

On the y-axis, we represented the number of customers who had a total spending amount within a particular bin. For instance, a label "150" might indicate that 150 customers had a total spending amount between $200 and $300

The graph is right-skewed, as most spending occurred between $100 and $200, with only a small number of customers spending more than $1000. The leftmost column, representing customers who spent less than $100, has the highest frequency of around 200 customers. The histogram suggests that a higher proportion of customers fall into lower spending bins.

2.2.2 Visualization of the "Rating" variable

We used a histogram to represent the distribution of the Rating given by the customers of the supermarket.

- The descriptive statistics of the "Rating" variable:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.000   5.500   7.000   6.973   8.500  10.000
```

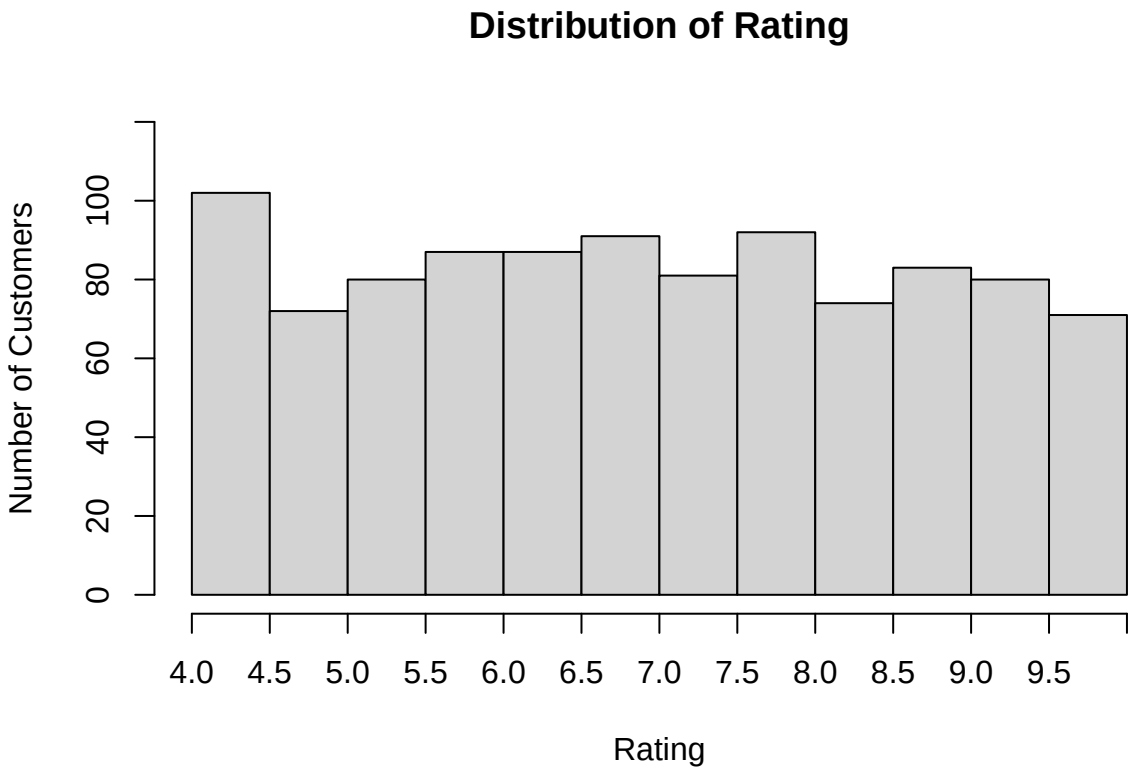

Figure 2: The Distribution of Rating

The x-axis represents the rating given by customers, binned into ranges. The bin cut points occur every 0.5, resulting in specific ranges such as: 4.0 to 4.5, 4.5 to 5.0, 5.0 to 5.5, and so on, following increments of 0.5.

On the y-axis, we represent the number of customers who rate within a particular bin. For instance, a label "80" might indicate that 80 customers rate between 4.5 and 5.0.

The distribution of ratings remains consistent, with minimal fluctuations. However, there was a noticeable concentration of data points on the left side of the histogram around 4.0 to 4.5, with approximately 100 customers. The data ranges from 4.00 to 10.00, indicating that customers tend to avoid giving ratings lower than 4.00.

2.2.3 Visualization of the "Gender" variable

We created a table to visualize the distribution of the levels of the variable "Gender":

Table 1: Gender Levels in the Dataset

Gender	Frequency
Female	501
Male	499

The data set is balanced as there are 501 Female and 499 Male customers.

2.2.4 Visualization of the "Customer.type" variable

We visualized the distribution of the levels of the variable "Customer.type" by creating a table:

Table 2: Customer Type Levels in the Dataset

Customer Type	Frequency
Member	501
Normal	499

The data set is balanced as there are 501 Member and 499 Normal customers.

2.2.5 Visualization of the "Quantity" variable

We used a histogram to represent the distribution of the total quantity purchased by customer

- The descriptive statistics of the "Quantity" variable:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    3.00    5.00    5.51    8.00   10.00
```

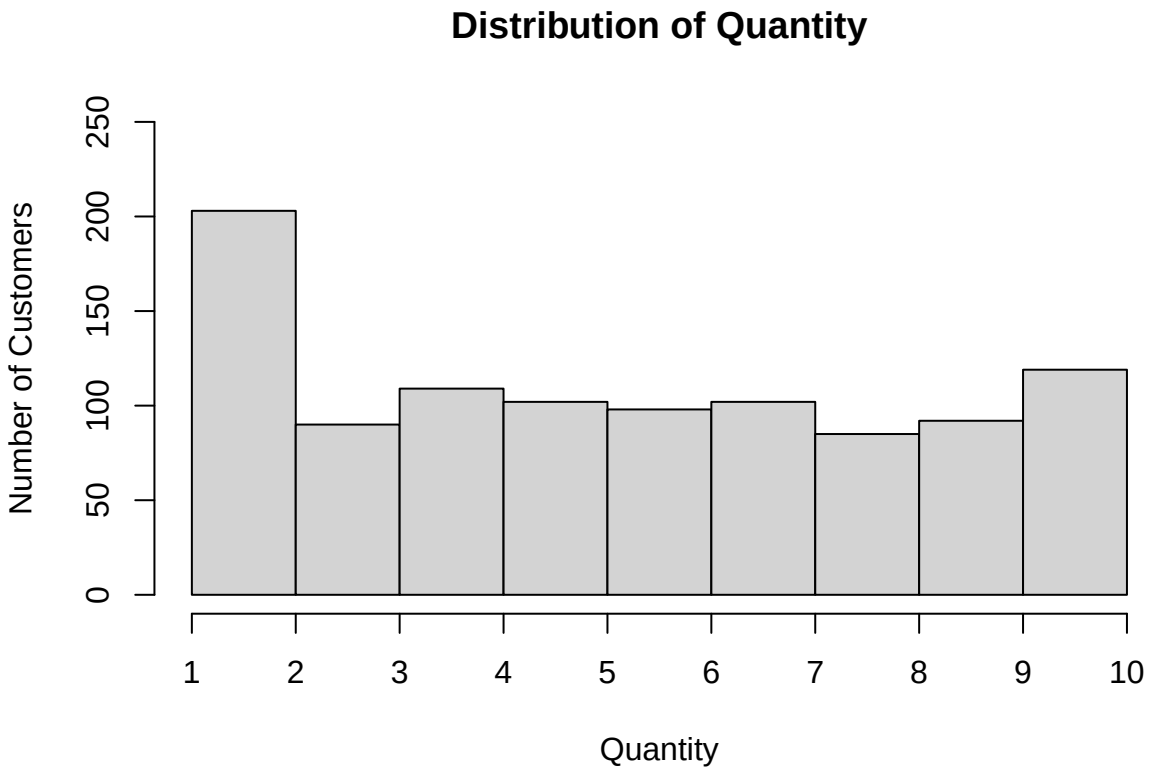

Figure 3: The Distribution of Quantity

The x-axis represents the amount of products customers purchased, binned into ranges. The bin cut points occur every 1 product, resulting in specific ranges such as: 1 to 2, 2 to 3, 3 to 4, and so on, following increments of 1 product.

On the y-axis, the value represents the number of customers who purchased within each bin range. For example, a label "100" might indicate that 100 customers purchased between 2 to 3 products.

The other bar in the histogram exhibits a similar frequency of approximately 100 customers. However, the leftmost column, representing the range of "1 to 2 products," has the highest frequency, with around 200 customers. This suggests that purchasing 1 to 2 products is the most popular quantity.

3.0 Analysis

PART 1: Investigating Spending Patterns by Customer Type and Gender

3.1 Hypothesis Testing for the Relationship between Customer's Gender and Total Spending

We investigated the question: "Is the total spending statistically different between Male and Female customers?"

Box plot of Total Spending per Gender

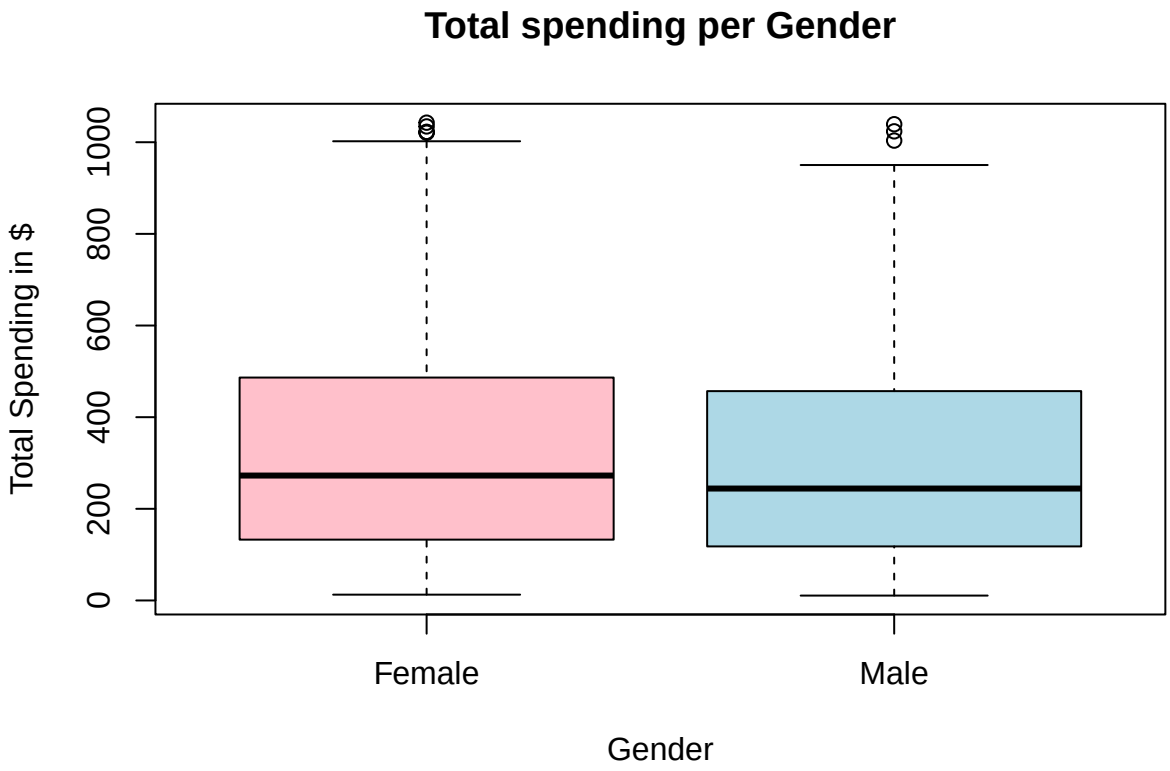

Figure 4: The distribution of Total spending per Gender

The median total spending is higher for Female customers compared to Male customers since the center line for "Female" is positioned higher. The IQR is also larger for Female customers. There are outliers for both genders, represented by the circles above the whiskers. This indicates that there are some Female and Male customers with total spending amounts that fall outside the typical range of the data.

Data Processing

We created two subsets from the data frame: one containing only Male customers and another containing only Female customers. Our objective is to investigate the correlation between gender and total spending. The male subset consists of 499 observations across 5 variables, while the female subset comprises 501 observations related to the same 5 variables of interest.

3.1.1 Method

We performed a hypothesis testing using the t.test function to compare the means of two gender-based

distributions: one for Males and another for Females. In this analysis, the independent variable is Gender and the dependent variable is the Total Spending.

- The null hypothesis: The true difference in means of the average total spending between group Male and group Female is equal to 0.
- The alternative hypothesis: The true difference in means of the average total spending between group Male and group Female is not equal to 0.

We performed a t-test and the p-value = 0.1181

3.1.2 Results

The t-test for the correlation between Gender and Total Spending:

Since the p value of this t-test is $0.1181 > 0.05$, we fail to reject the null hypothesis. There is not a significant difference in means of the average total spending between group Male and group Female.

We proceeded to calculate the descriptive statistics for total spending within the two gender groups to see if they relate to the results of the t-test:

- Descriptive statistics for the Total Spending of Male customers

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.68  117.95  244.23  310.79  456.83 1039.29
```

- Descriptive statistics for the Total Spending of Female customers

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.69  132.76  272.58  335.10  486.44 1042.65
```

The descriptive statistics show that on average, Female customers spend $335.10, slightly more than their Male counterparts who spend $310.79. Additionally, the distribution of spending varies between the two groups. For Male customers, the median spending is $244.23, with a range spanning from $10.68 to $1039.29. In contrast, Female customers have a median spending of $272.58, and their spending range extends from $12.69 to $1042.65. These differences highlight a subtle difference in total spending between the genders.

Histogram Plots

We created histogram plots to visualize the total spending distributions between the two Gender subsets.

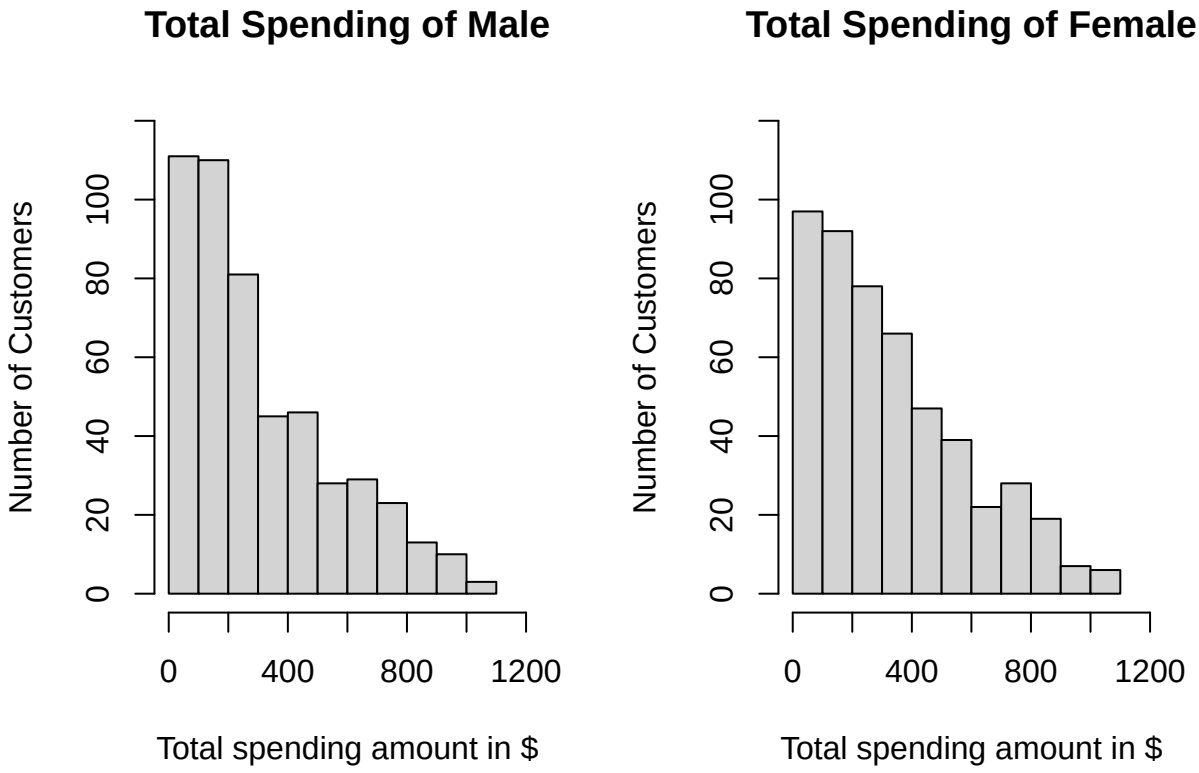

Figure 5: Comparison of Total Spending Distributions by Gender

From Figure 7, both histograms display a similar distribution pattern with a right skew and a peak in lower spending. From the graph, we can understand that although the descriptive statistics reveal slight differences between the two groups, these differences are not substantial enough for the hypothesis test to establish statistical significance.

3.2 Hypothesis Testing for the Relationship between between Customer's types and Total Spending

We investigated the question: "Is the total spending statistically different between Member and Non-member customers?"

Box plot of Total Spending per Customer Type

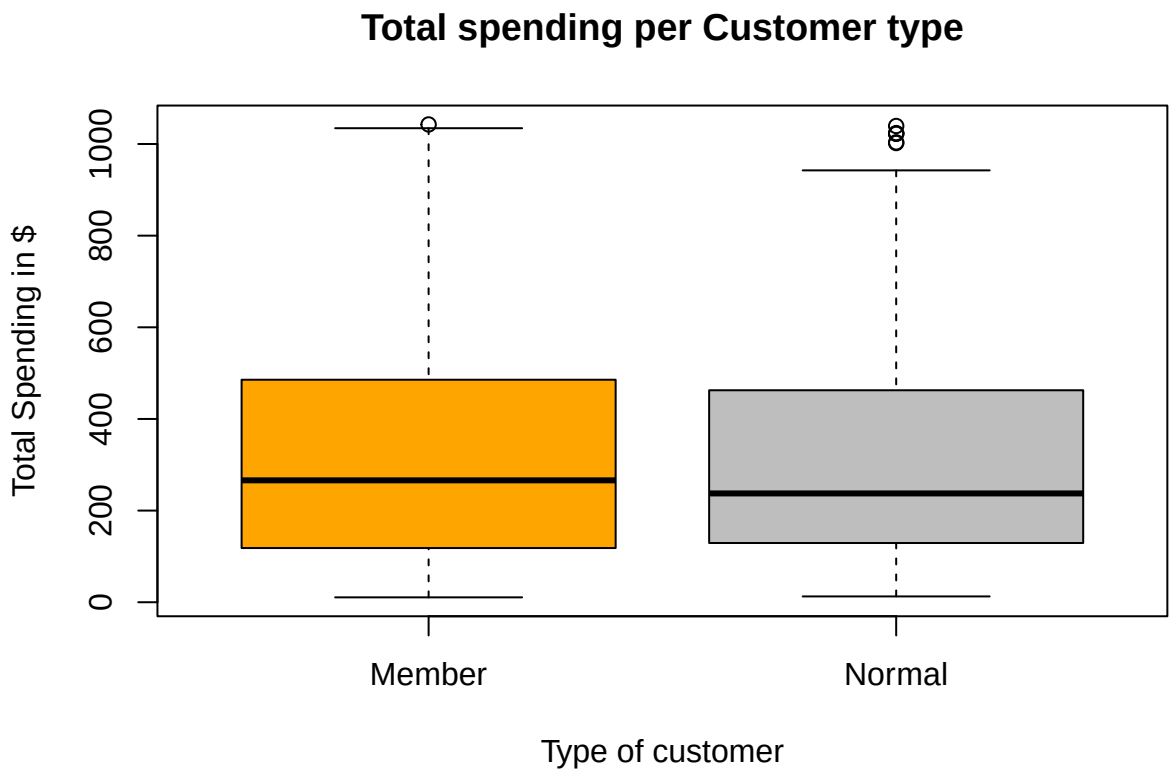

Figure 6: The distribution of Total spending per Customer Type

The median total spending appears to be higher for "Member" customers compared to "Normal" customers because the center line for "Member" is positioned higher on the plot. The IQR also appears to be larger for "Member" customers compared to "Normal" customers as the box for "Member" customers has a greater height.

Overall, the box plot suggests that "Member" customers tend to have higher total spending than "Normal" customers. The spread of data (IQR) is also larger for "Member" customers, indicating greater variability in their spending habits.

Data Processing

From the data frame, we generated two distinct subsets: one exclusively containing Member customers and another comprising Normal (non-member) customers. Our objective is to explore the correlation between each customer type and their total spending. Specifically, the Member subset consists of 501 observations across 5 variables. The Normal subset comprises 499 observations of the same 5 variables of interest.

3.2.1 Method

We conducted a hypothesis testing using the t.test function to compare the means of two customer type

distributions: one for Members and another for Normal (non-member) customers. The independent variable is Customer Type and the dependent variable is the Total Spending.

- The null hypothesis: The true difference in means of the average total spending between group Member and group Normal is equal to 0.
- The alternative hypothesis: The true difference in means of the average total spending between group Member and group Normal is not equal to 0.

We performed a t-test and the p-value = 0.5344

3.2.2 Result

The t-test for the correlation between Customer Type and Total Spending:

Since the p value of this t-test is $0.1181 > 0.05$, we fail to reject the null hypothesis. There is not a significant difference in means of the average total spending between group Member and group Normal

We then analyzed the descriptive statistics of the total spending for the two customer groups to see their correlation with the t-test results:

- Descriptive statistics for the Total Spending of Member customers:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.68  118.25  266.03  327.79  485.57 1042.65
```

- Descriptive statistics for the Total Spending of Normal customers:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.69  129.23  237.43  318.12  462.61 1039.29
```

The descriptive statistics reveal that the mean spending for Member customers is $327.79, slightly higher than that of Non-Member customers at $318.12. Additionally, the median and range differ between the two groups. For Member customers, the median spending is $266.03, with a range spanning from $10.68 to $1042.65. In contrast, Normal customers have a median spending of $237.43, with a range from $12.69 to $1039.29. These findings indicate a subtle disparity in total spending between the two customer segments.

Histogram Plots

We created histogram plots to visualize the total spending distributions between the two Customer Type subsets.

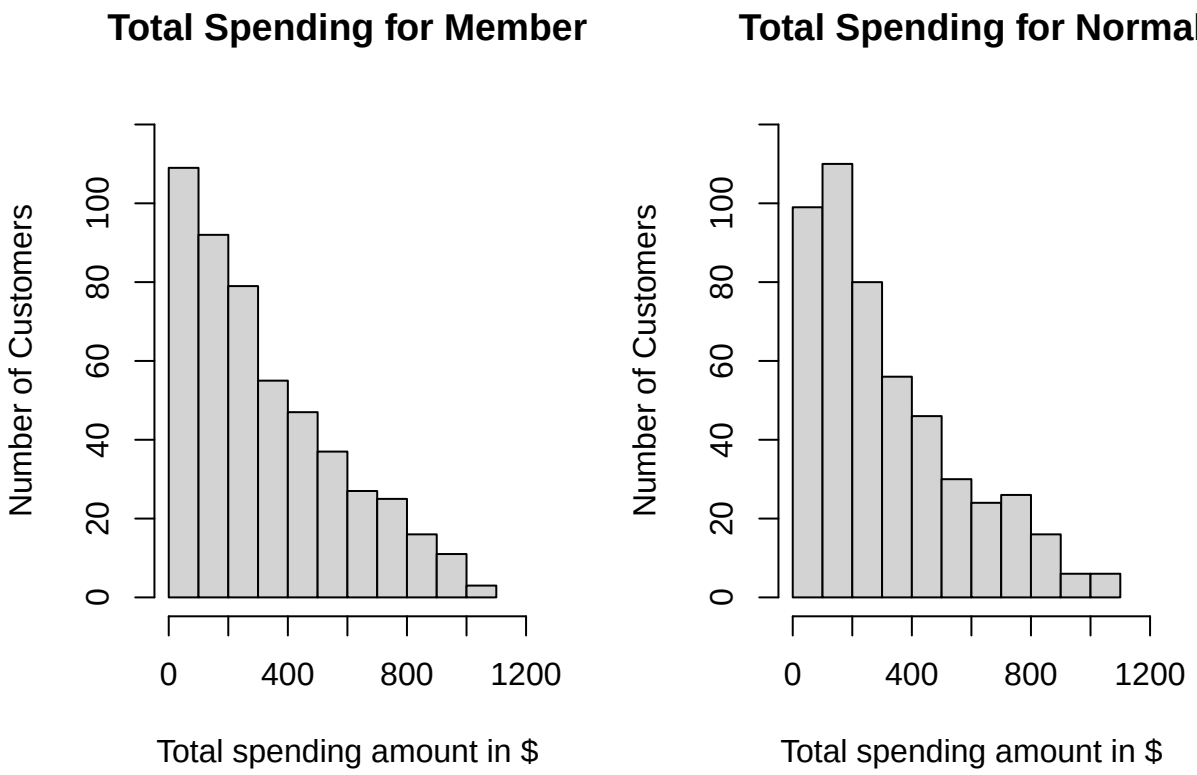

Figure 7: Comparison of Total Spending Distributions by Customer Type

From Figure 9, both histogram shapes have a right-skewed distribution, with peaks at lower spending levels. Similar to the gender groups, we observed that although the descriptive statistics reveal slight differences between the two groups, these differences are not substantial enough for the hypothesis test to conclude a significant result.

Since the hypothesis testing did not yield significant results, our next step is to employ clustering on customer data.

3.3 K-Means Clustering using Total spending and Quantity

In this section, we performed K-Means clustering on customer data based on gender and customer membership to identify distinct customer segments. Our objective is to explore how customer type or gender impact the total amount spent and the quantity of products purchased.

3.3.1 Method

Data Processing

We clustered the data using Total Spending and Quantity, ignoring Gender and Customer type. We then compared cluster membership with the Gender and Customer.type levels to see if there is any correlation.

We created a subset that includes only the 'Total' and 'Quantity' attributes because we are specifically concentrating on those for clustering. The subset comprises 1000 observations of 2 variables of interest. The first 5 rows of data are shown below:

```
##      Total Quantity
## 1 548.9715        7
## 2  80.2200        5
## 3 340.5255        7
## 4 489.0480        8
## 5 634.3785        7
```

Since the range and scale of these two values are very different, we also scaled the data to be used for clustering:

```
##         Total   Quantity
## 1  0.91914693  0.5096752
## 2 -0.98723557 -0.1744526
## 3  0.07141032  0.5096752
## 4  0.67544187  0.8517391
## 5  1.26649176  0.5096752
```

Choosing the Optimal Number of Clusters

We used both the Elbow method and Nbclust function to find the optimal number of clusters.

First, we called the wssplot function passing in the scaled spending and quantity data as the parameter.

The within-cluster sum of squares (WSS) plot was used to generate the Elbow curve. The "elbow" point on the curve indicates the optimal number of clusters to use in the k-means clustering algorithm. In Figure 10 below, the elbow occurs at 2, suggesting that 2 clusters would be the optimal choice.

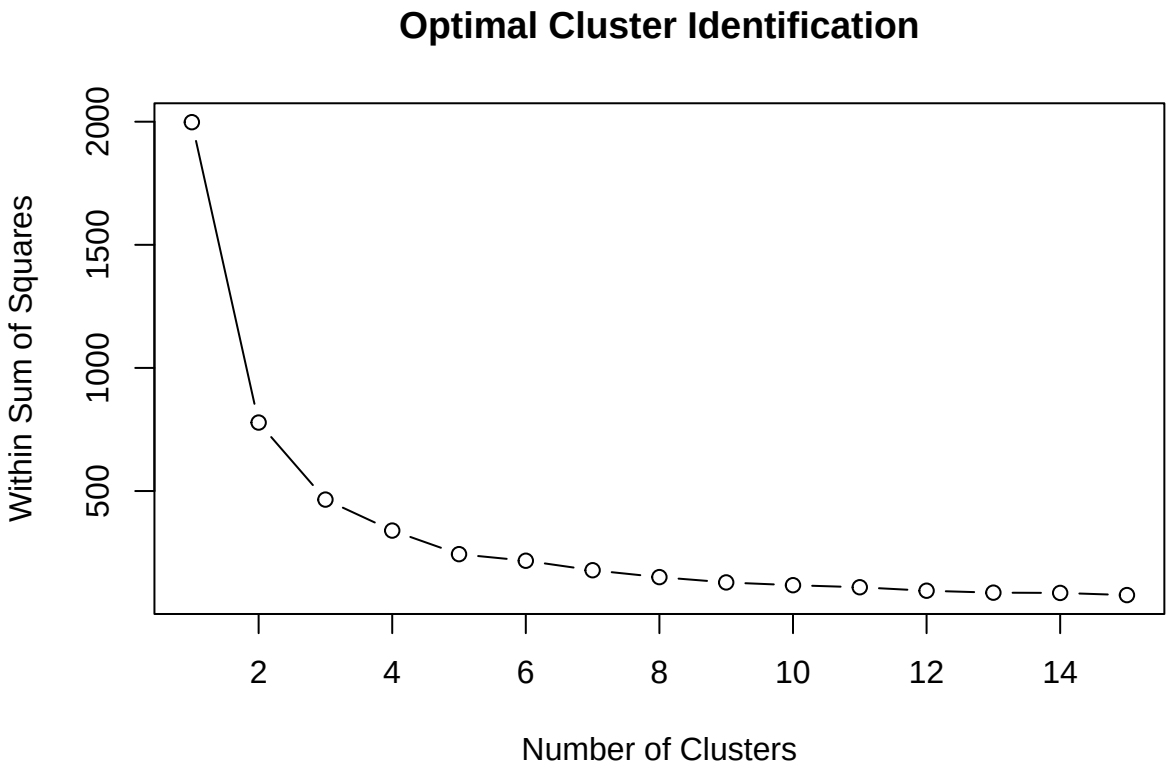

Figure 8: Within Sum of Squares (WSS) Plot

Next, we used the NbClust function to determine the optimal number of clusters. The Table 3 from the result of NbClust shows the number of clusters and the number of indices that reported that clustering as "best".

Table 3: Number of clusters chosen by different indices

Number of clusters	Number of indices
0	2
1	1
2	9
3	6
5	1
9	4
10	3

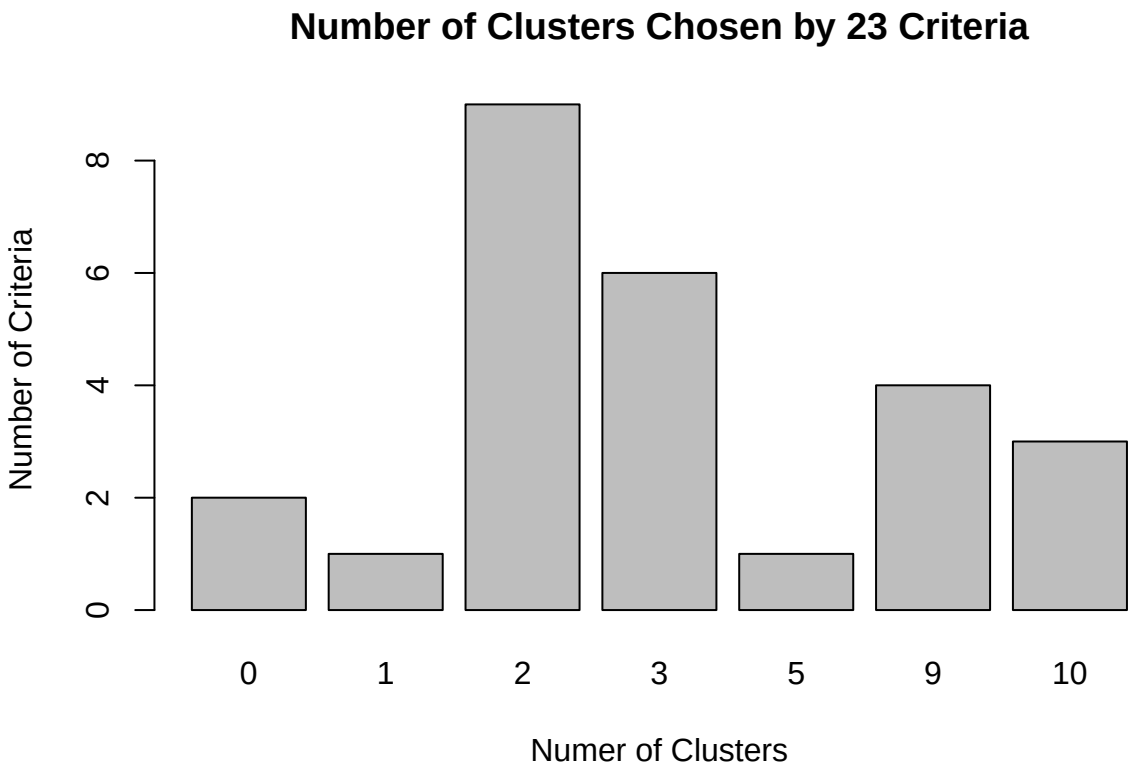

Figure 9: Number of clusters chosen by 23 Criteria

23 indices in total were used by NbClust to determine the optimal number of clusters and 9 indices proposed 2 as the optimal number of clusters. Therefore, according to NbClust, 2 is the optimal number of clusters.

3.3.2 Results

K-means Algorithm

The k-means algorithm was carried out by the function "kmeans". The three parameters we used are: the data to be clustered, the number of clusters, and the number of "runs" of the algorithm, called "nstart".

Table 4: Results of K-means clustering

Size	WithinSS	BetweenSS
440	778.0623	1219.938
560	778.0623	1219.938

From Table 4 that shows the results of K-means clustering, there are 2 clusters, 1 cluster has 560 data points and the other has 440 data points. The sum of squared distances (withinss) inside clusters (778.0623) is smaller than the (betweenss) between clusters (1219.938). We think this is a good result as the total sum of squared distances within clusters is smaller which means that data points within a cluster are similar to each other and there is a high cohesion within clusters. The sum of squared distances between different clusters are large so the data points from different clusters are dissimilar which indicates a good separation between clusters.

We then used the plotcluster() function to visualize clustering of the data. Principal Component Analysis (PCA) is a dimension reduction technique. It condenses the information from multiple variables into a smaller set of "new" variables known as components. These components capture the most significant patterns in the data. According to the PCA, the axes of the plot shows the first two principal components. The plotcluster() function computes the principal component scores for each observation and creates a plot that highlights clusters by color.

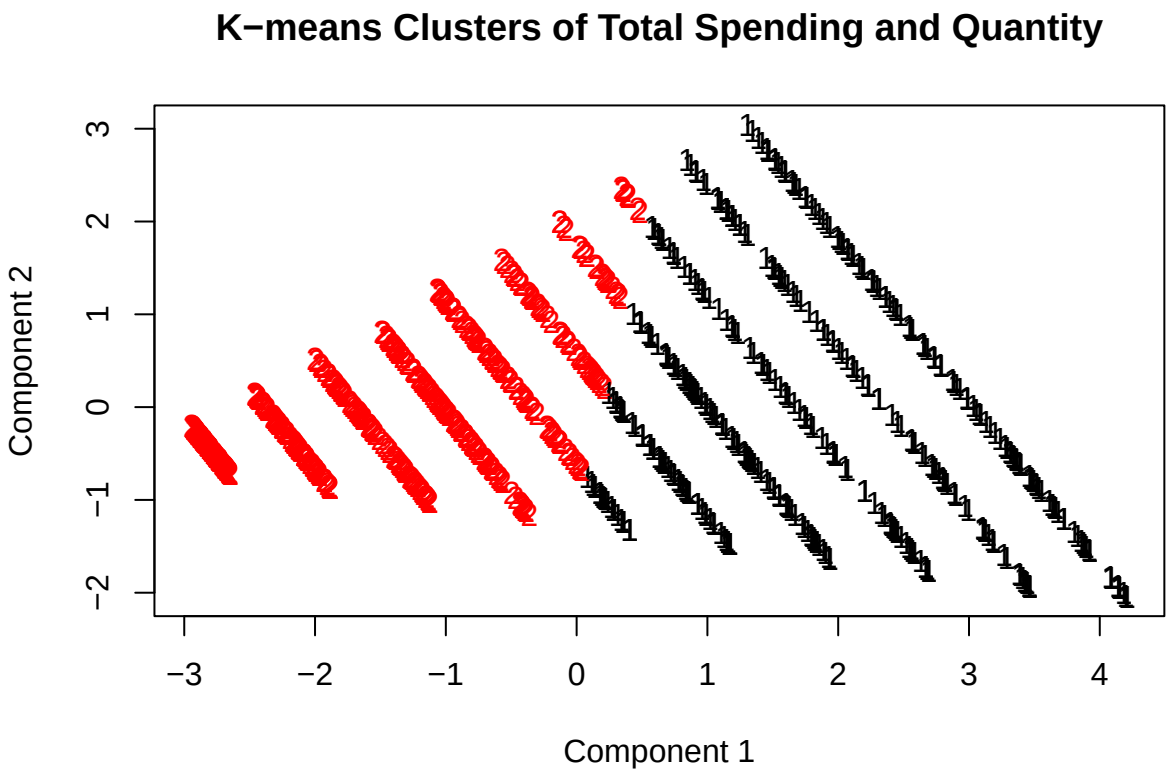

Figure 10: K-means Clusters of Total Spending and Quantity

In Figure 12, the horizontal axis represents one component of the data. It ranges from approximately -3 to +4. The values on this axis indicate different spending and purchase patterns. The vertical axis represents another dimension of the data. It ranges from approximately -2 to +3. Similar to the horizontal axis, the values on this axis also relate to spending and purchase behavior.

The cluster plot visualizes spending and purchasing behavior across two dimensions. The red cluster represents lower spending and purchasing, while the black cluster represents higher spending and purchasing. The plot contains two types of clusters:

- Red Cluster: The red points are concentrated on the left side of the plot. They have negative values along the Component 1 axis, indicating lower spending and purchasing. The red cluster spreads between approximately -3 and 0 on the Component 1 axis.
- Black Cluster: The cluster is on the right side of the plot. The cluster points have positive values along the Component 1 axis, ranging from about 0 to 4 which indicates higher spending and purchasing. The black cluster covers a wide range of values on the Component 2 axis.

Correlate K-means Clustering with Gender

The contingency table for the levels of Gender and their frequencies in each cluster is presented:

```
##
##     Female Male
##   1    235  205
##   2    266  294
```

In the contingency table, we observed a difference in the number of Female and Male customers. Specifically, Cluster 1 has more Female customers (235) compared to Male customers (205), while Cluster 2 has more Male customers (294) compared to Female customers (266).

Correlate K-means Clustering with Customer Type

The contingency table for the levels of Customer Type and their frequencies in each cluster is shown:

```
##
##     Member Normal
##   1    225    215
##   2    276    284
```

In the contingency table, there exists a difference in the number of Member and Normal customers. Specifically, in Cluster 1, there are more Normal customers (284) compared to Member customers (276). In Cluster 2, there are more Member customers (225) compared to Normal customers (215).

3.4 Hierarchical Clustering using Total spending and Quantity

3.4.1 Method

In this section, we performed Hierarchical clustering on customer data using gender and customer membership to identify distinct customer segments.

Data Processing

To achieve this, we created a subset containing only the 'Total' and 'Quantity' attributes, as we are specifically focusing on these variables for clustering. The subset consists of 1000 observations with 2 variables of interest. Next, from the scaled data set mentioned above, we generated a distance matrix for the subset. The default distance metric used is Euclidean.

We then performed a hierarchical clustering using the distance matrix and plotted a dendrogram of the result.

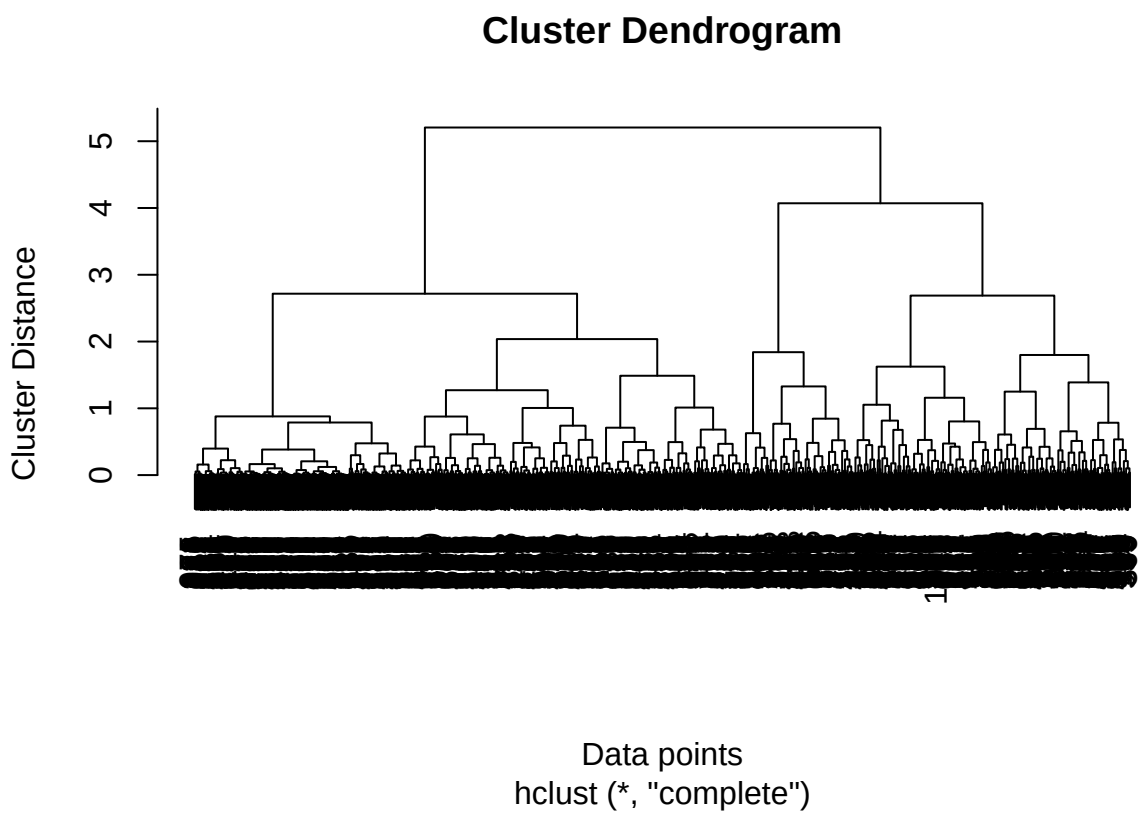

Figure 11: Cluster Dendogram of the 'Total' and 'Quantity' variables

Given that this is an agglomerative algorithm, 2 clusters are created at the end of the algorithm, as shown in Figure 13.

3.4.2 Results

We selected a cut point of 2 clusters on the tree and created a plot showing the data points within each cluster.

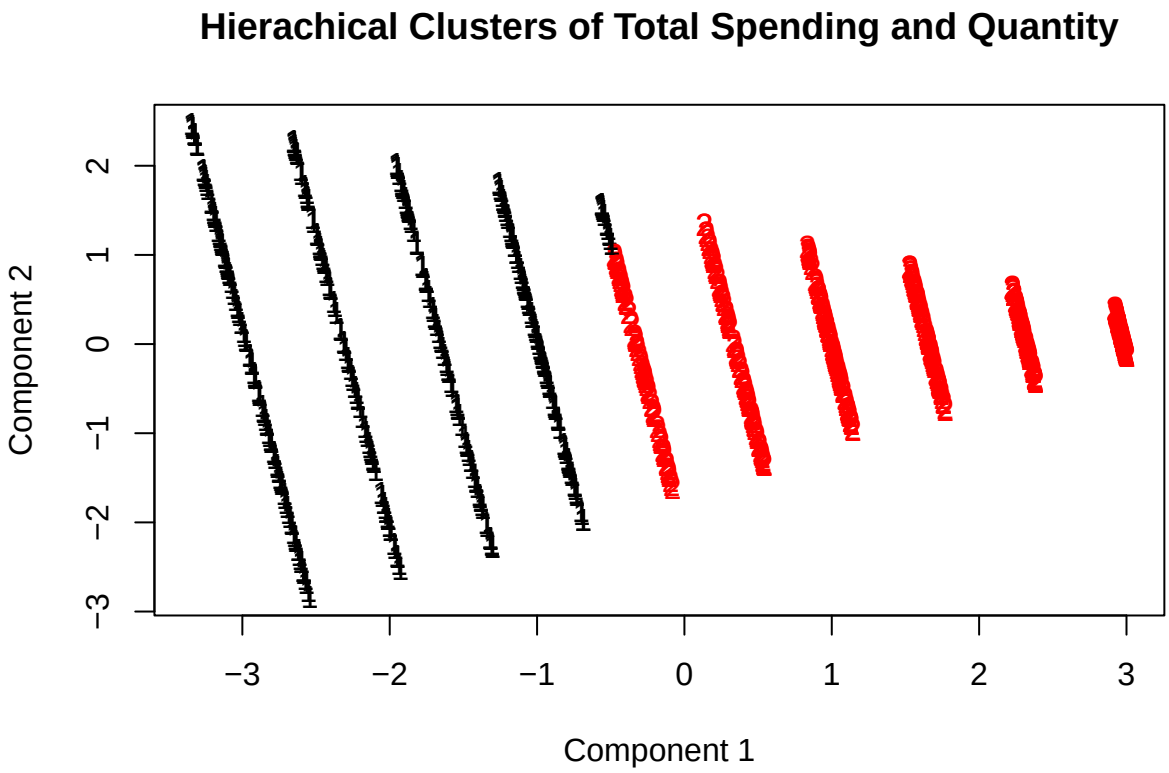

Figure 12: Hierachical Clusters of Total Spending and Quantity

According to Principal Component Analysis (PCA), the cluster plot displays the first two principal components. The plotcluster() function computes the principal component scores for each observation and creates a color-coded plot to highlight clusters.

In Figure 14, the horizontal axis represents one component of the data, ranging from approximately -3 to +3. Values on this axis indicate different spending and purchase patterns. The vertical axis represents another dimension of the data, ranging from approximately -3 to +2. Similar to the horizontal axis, values on this axis relate to spending and purchase behavior.

The cluster plot visualizes spending and purchasing behavior across two dimensions. The black clusters represent lower spending and purchasing, while the red clusters represent higher spending and purchasing. From the plot, we can see two types of clusters:

- Black Cluster: The data is clustered together on the left side of the plot and follow a slanted pattern. They have negative values along the Component 1 axis, indicating lower spending and quantity purchased. The black cluster spreads between approximately -3 and 0 on the Component 1 axis and covers a wide range of values on the Component 2 axis.
- Red Cluster: The cluster is on the right side of the plot and also follows a slanted pattern. It has positive values along the Component 1 axis, ranging from about 0 to 3 which indicates higher spending and quantity purchased.

Correlate Hierarchical Clustering with Gender

Additionally, we presented a table that displays the frequencies of Gender and Customer Type levels in each cluster.

```
##
## clusterCut2 Female Male
##           1    220  194
##           2    281  305
```

In the contingency table, there exists differences between the number of Female and Male customers. Cluster 1 has more Female customers (220) compared to Male customers (194), while in Cluster 2, there are more Male customers (305) compared to Female customers (281)

Correlate Hierarchical Clustering with Customer Type

```
##
## clusterCut2 Member Normal
##           1    215    199
##           2    286    300
```

In the contingency table, we observed differences between the number of Member and Normal customers across the two clusters. Specifically, Cluster 1 has more Member customers (215) compared to Normal customers (199), while Cluster 2 has more Normal customers (300) compared to Member customers (286).

In conclusion, both K-means Clustering and Hierarchical Clustering consistently identified 2 clusters as the optimal choice. Interestingly, this number of clusters aligned with the levels of Gender and Customer Type. This suggested a potential relationship between Total spending and Quantity with the Gender and Customer Type classification.

Model Validation using Statistics

To further validate the clustering results, we calculated the average total spending and average quantity purchased for both Male and Female customers. The results are as follows:

Table 5: Average Spending By Gender

Gender	Average Spending ($)
Female	335.0957
Male	310.7892

Table 6: Average Quantity Purchased By Gender

Gender	Average Quantity Purchased
Female	5.726547
Male	5.292585

The results reveal that, on average, Female customers spend $335 and buy 5.73 products while Male customers spend $311 and buy 5.29 products on average. These findings suggest that Female spending and quantity purchased tend to be slightly higher than those of Male customers.

We also found the average spending for Normal and Member customers:

Table 7: Average Spending By Customer Type

Customer Type	Average Spending ($)
Member	327.7913
Normal	318.1229

Table 8: Average Quantity Purchased By Customer Type

Customer Type	Average Quantity Purchased
Member	5.558882
Normal	5.460922

The analysis reveals that Member customers spend an average of $328 and buy 5.56 products, while Normal customers spend an average of $318 and buy 5.46 products. Although there is not that much of a difference in the quantity bought, Members tend to have slightly higher spending and quantity purchased compared to Normal customers.

PART 2: Interpreting Customer Satisfaction Drivers Through Rating Analysis

Visualization of Rating per Different Factors

We used box plots to visualize the distribution of Rating across different factors:

Box plot of Rating per Gender

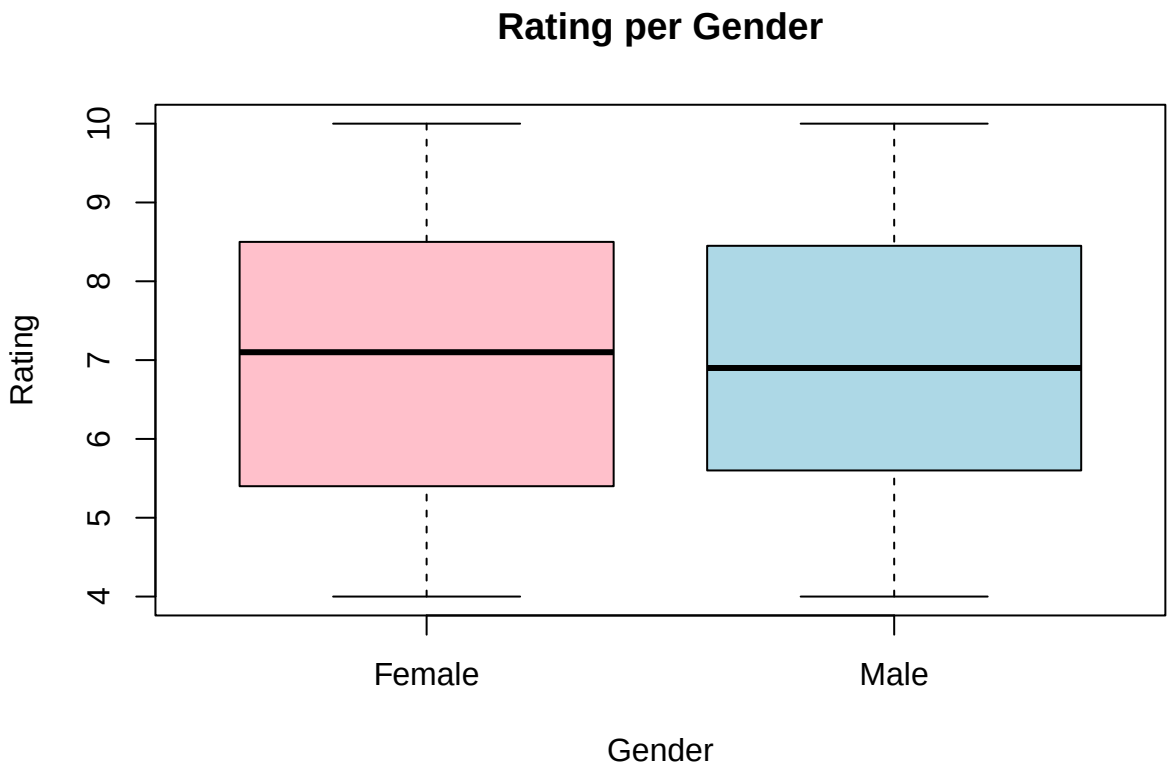

Figure 13: The distrbituion of Rating per Gender

The median rating is higher for Female customers compared to Male. The IQR for Female customers is also larger than their Male counterparts. There are no outlier for both genders.

The box plot suggests that Female customers tend to rate higher than Male customers. However, the spread of data is also larger for Female customers, indicating greater variability in their ratings.

Box plot of Rating per Customer Type

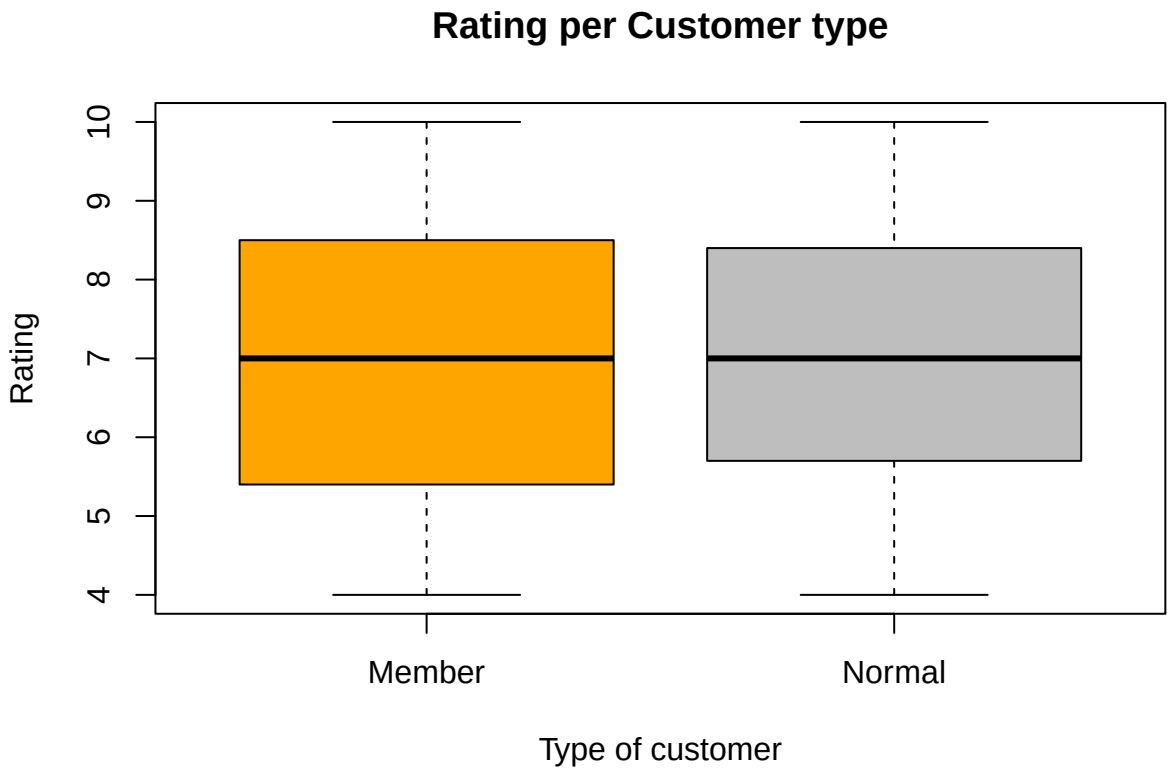

Figure 14: The distrbituion of Rating per Customer Type

The median rating appears to be similar for both “Member” and “Normal” customers because the center lines for “Member” and “Normal” are positioned at roughly the same level on the plot. The IQR appears to be larger for “Member” customers compared to “Normal” customers. There are no outlier for both customer types.

The box plot suggests that there is no significant difference in median rating between member and normal customers. However, the IQR is larger for member customers, indicating greater variability in their ratings.

3.5 Linear Regression Model for Continous Rating Variable

We explored whether customer types, genders, total spending, or quantity purchased have an impact on the rating. To quantify the influence of each factor on the rating, we constructed a linear regression model and a KNN model using the sales_data.df data set.

3.5.1 Method

We fitted a linear model using the sales_data.df data set. We used the original modified data set, which consists of 1000 observations across 5 variables without any further subsetting.

- The independent variables are "Customer.type", "Gender", "Quantity", "Gender" and "Total"
- The dependent variable is "Rating"

3.5.2 Results

Table 9: Result of the Linear Regression Model for the Continous Rating variable

Coefficients	p-value
Customer.type (Normal)	0.568
Gender (Male)	0.923
Quantity	0.654
Total purchased	0.263

Surprisingly, all the variables from Table 9 have p-values greater than 0.05, indicating that they do not significantly contribute to explaining the model's outcome. All the predictors in this model are not significant. This outcome is unexpected, as we initially hypothesized that information about customer demographics, membership status, and purchase history would be predictive of supermarket ratings. However, the results from the regression model reveal no correlation between these factors and the rating provided by customers.

3.6 Linear Regression Model For Discrete Rating variable

3.6.1 Method

As the linear model for the continuous variable did not yield any significant factors, we opted to transform the 'Rating' variable into a discrete variable using the 'cut' function for binning

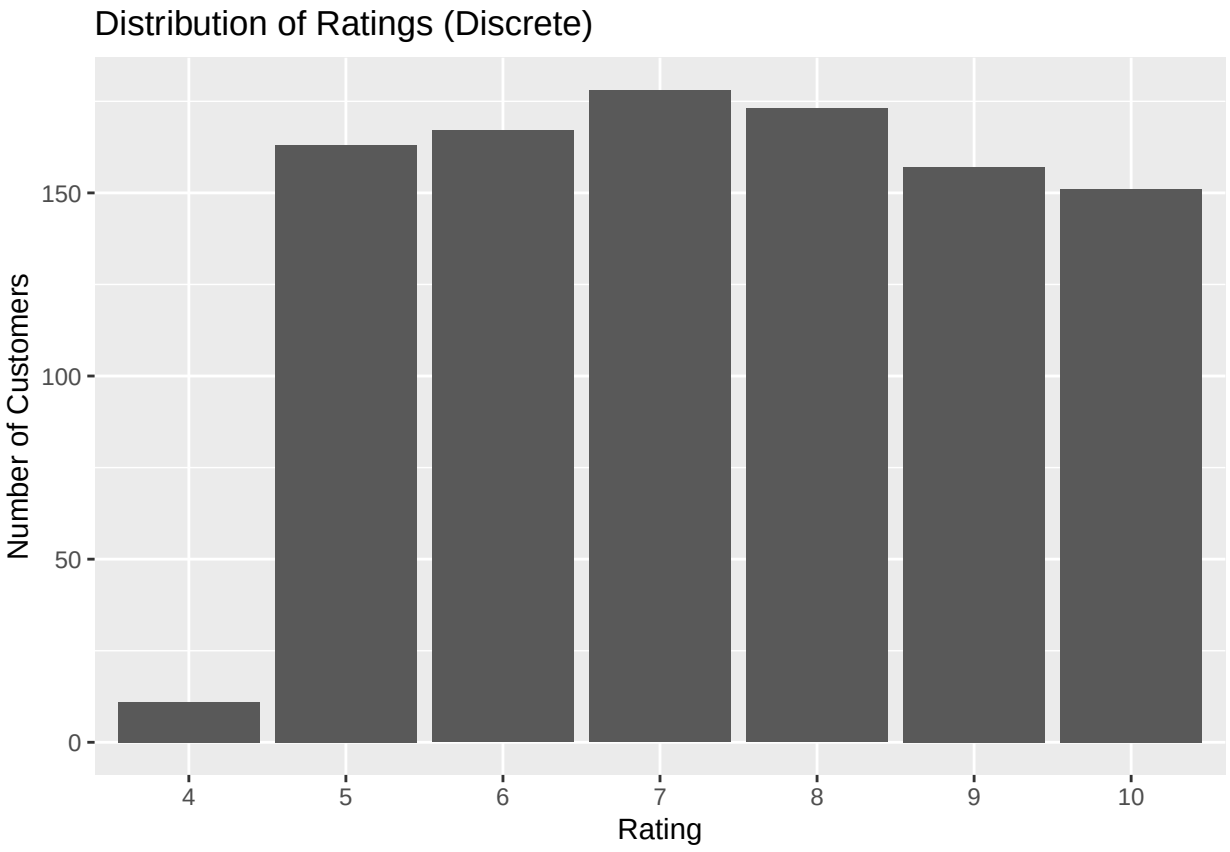

Figure 15: The distrbituion of Discrete Rating Variable

The bar chart displays the distribution of ratings. The rating ranges from 4 to 10, with no ratings below 4. The least common rating is 4, while the majority of people tend to rate around 7.

We fit a linear model again using the sales_data.df data set with the modified Rating variable.

- The independent variables are "Customer.type", "Gender", "Quantity", "Gender" and "Total".
- The dependent variable is the discrete "Rating" variable.

3.6.2 Results

Table 10: Result of the Linear Regression Model for the Discrete Rating variable

Coefficients	p-value
Customer.type (Normal)	0.392
Gender (Male)	0.812
Quantity	0.773
Total purchased	0.352

We attempted to categorize the 'Rating' variable into discrete numerical values but the outcome remained unchanged. The regression model results indicate no significant correlation between these factors and the

ratings provided by customers as all the factors from Table 10 have p-value > 0.05. This suggests that the issue lies not in whether 'Rating' is treated as a continuous or discrete variable but rather that there are no significant factors within the model to explain the variation in ratings.

3.7 K-nearest neighbor (KNN) Regression Model

3.7.1 Method

We predicted Rating, given Customer.type, Gender, Quantity and Total as the predictors. The original data of 1000 observations and 5 variables will be used.

We splitted the data into training and test sets, allocating 75% of the original data set to the training set and 25% to the test set. The KNN training set comprises 752 observations across 5 variables, while the test set contains 248 observations with the same 5 variables.

To select the optimal model, we performed cross-validation on the training data. Using root mean square error (RMSE) as the criterion, we identified the best value for the hyperparameter 'k.' The final chosen value for 'k' is 43, resulting in an RMSE of 1.744070.

3.7.2 Results

We used the varImp function to display the "importance" of the predictors:

```
## loess r-squared variable importance
##
##               Overall
## Total         100.000
## Customer.type   4.720
## Quantity        3.211
## Gender          0.000
```

The most critical variable is Total, which has an importance score of 100.00. However, the variables Customer.type (4.720), Quantity (3.211), and Gender (0.000) show minimal importance. This aligns with the findings from the linear regression model, which indicates that neither Customer.type nor Gender significantly impacts the ratings provided.

Given this lack of significant factors in both Regression and KNN Model to explore the impacts of Rating, it is possible that other unaccounted variables influence the Rating within this data set. There remains room for improvement, and researchers or analysts interested could explore additional factors to enhance the model.

4.0 Conclusions

4.1 Conclusions about the Spending Patterns by Customer Type and Gender

For the first guiding question, "Is there a relationship between total price spent and customer type or gender?", both K-means Clustering and Hierarchical Clustering consistently identify two clusters as the optimal choice which aligns with the levels of Gender and Customer Type. After analyzing the clustering results and statistical calculations, we found differences in spending patterns and the amount of products purchased between Female and Male customers, as well as between Member and Non-member customers. Specifically, Female and Member customers tend to spend and purchase more than Male and Normal customers. These results aligned with our expectations. This insight is valuable for business owners, as they can create targeted promotional campaigns to attract Female and Member customers, thereby amplifying their spending. Additionally, they can develop services to encourage more Male customers to make purchases and provide advantages to incentivize more people to become Members.

However, it is important to note that the observed difference is not highly significant. The hypothesis test does not find any significant difference in total spending between Female and Male customers, as well as between Member and Normal customers. The data set only contains information on approximately 1000 customers over a three-month period, which could potentially undermine our conclusions. With a larger dataset and a longer observation period, we believe a stronger relationship can be established.

4.2 Conclusions about the Customer Satisfaction Drivers

Both Regression and KNN Model could not find any significant factors to identify which factors impact Rating. This outcome is surprising, as we initially anticipated that certain features would influence the overall rating. However, it is essential to consider the limitations of our data set and modeling techniques. With the availability of more data, further exploration and perhaps more sophisticated models, we may uncover hidden patterns or interactions that contribute to the rating.

5.0 References

[Kaggle Data Set] (https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales)

[Principal Component Analysis] (https://www.geeksforgeeks.org/principal-component-analysis-pca)

[Plots & Figures in RMarkdown] (https://uoepsy.github.io/scs/rmd-bootcamp/06-figs.html)