Project Title: E-Commerce Customer Satisfaction

Part I: Introduction

Background of the Business/Research Problems

The dataset under consideration pertains to an e-commerce platform that tracks customer purchasing behavior. Analyzing such data can provide critical insights into customer preferences, spending habits, and satisfaction levels. The primary objectives include understanding the factors that influence customer spending and satisfaction, which can aid in enhancing marketing strategies, improving customer retention, and optimizing the overall shopping experience.

Research Questions

1. Regression Problem:
    - Business/Research Question: What factors influence the total amount spent by customers on the e-commerce platform?
    - Dependent Variable (DV): Total Spend
    - Potential Independent Variables (IVs): Age, Gender, Membership Type, Items Purchased, Average Rating, Discount Applied, Days Since Last Purchase, Satisfaction Level
    - Importance in Business/Research Practice: Understanding what drives customer spending can help the platform tailor marketing strategies to boost sales and prioritize features or services that lead to higher revenue.
    - Hypotheses:
        - H1: Older customers tend to spend more due to higher purchasing power.
        - H2: Members with higher-tier membership types (e.g. Gold) spend more than those with lower-tier memberships (e.g. Silver, Bronze).

2. Classification Problem:
    - Business/Research Question: Can we predict customer satisfaction based on their shopping behavior and demographics?
    - Dependent Variable (DV): Satisfaction Level
    - Potential Independent Variables (IVs): Total Spend, Age, Gender, City, Membership Type, Items Purchased, Average Rating, Discount Applied, Days Since Last Purchase
    - Importance in Business/Research Practice: Predicting customer satisfaction can help in identifying at-risk customers and developing strategies to improve their experiences, thus enhancing customer loyalty and reducing churn.
    - Hypotheses:
        - H1: Customers who receive discounts are more likely to be satisfied.
        - H2: Frequent purchasers are more likely to report higher satisfaction.

Part II: Data Preparation and EDA

Data Source and Summary

- Link of the Data Source: https://www.kaggle.com/datasets/uom190346a/e-commerce-customer-behavior-dataset
- Summary of the Dataset: This dataset captures customer behavior for an e-commerce platform, including demographic details, purchase history, and satisfaction levels.

Number of Observations

- Number of Observations: 350
- What Each Observation Represents: Each observation in the dataset represents a single customer's transaction or interaction with the e-commerce platform.

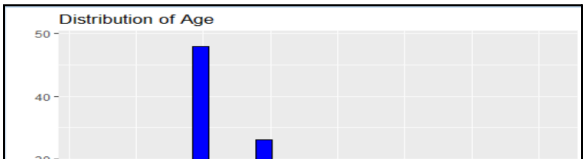
Variables

Here's a detailed breakdown of each variable in the dataset:

No.	Variable Name	Description	Type
1	Customer ID	Unique identifier for each customer	Integer (Numeric)
2	Gender	Customer's gender	Categorical (Text)
3	Age	Age of the customer	Integer (Numeric)
4	City	City where the customer resides	Categorical (Text)
5	Membership Type	Type of membership the customer holds (e.g., Gold, Silver, Bronze)	Categorical (Text)
6	Total Spend	Total amount spent by the customer in USD	Float (Numeric)
7	Items Purchased	Number of items purchased in the transaction	Integer (Numeric)
8	Average Rating	Average rating given by the customer across all their transactions	Float (Numeric)
9	Discount Applied	Indicates whether a discount was applied to the transaction	Boolean (True/False)
10	Days Since Last Purchase	Number of days since the customer's last purchase	Integer (Numeric)
11	Satisfaction Level	Satisfaction level of the customer with their transaction (e.g., Satisfied, Neutral, Unsatisfied)	Categorical (Text)

Exploratory Data Analysis:

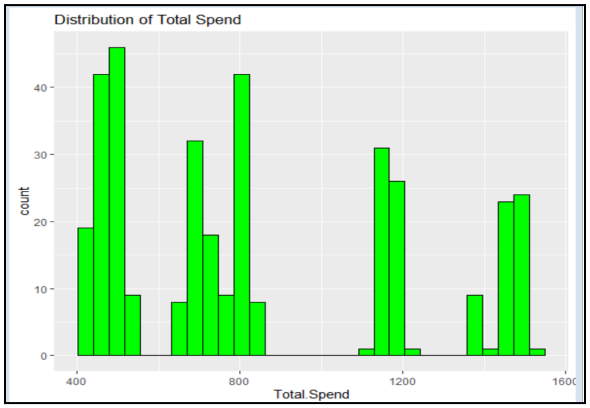
Univariate Analysis



Age:

The peak at age 30 suggests that this is the most common age or that there is a high concentration of individuals around this age.

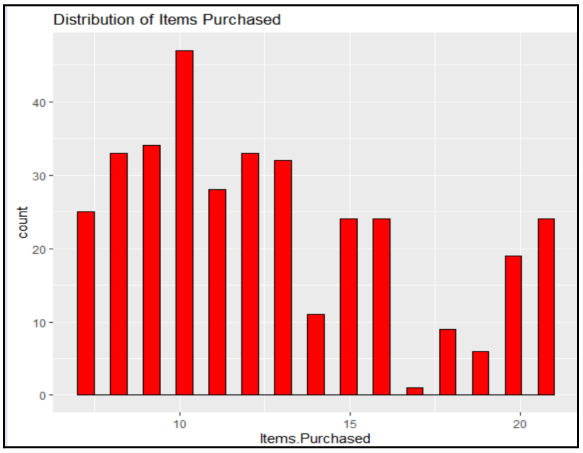
The distribution does not seem to be perfectly symmetrical, which could suggest a skew in the age distribution. There are some age ranges that have particularly low counts compared to adjacent ages. There are no clear indications of missing values or extreme outliers in the data.



Total Spend:

The plot displays specific spending tiers or clusters around certain amounts (near 400, 800, 1200)

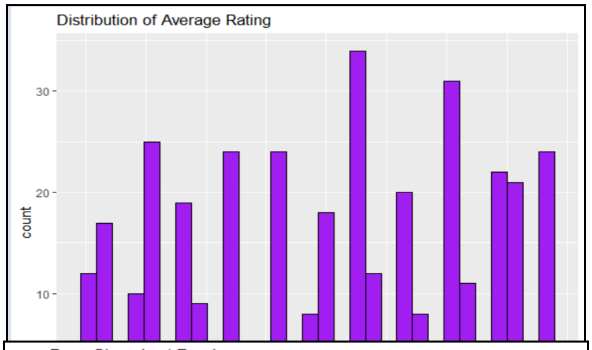
Consumers might have psychological price limits, such as preferring to spend around 400 instead of going up to 500. The data suggests that there might be distinct segments in the market, each with its own spending pattern. There are no clear indications of missing values or extreme outliers in the data.



Items Purchased:

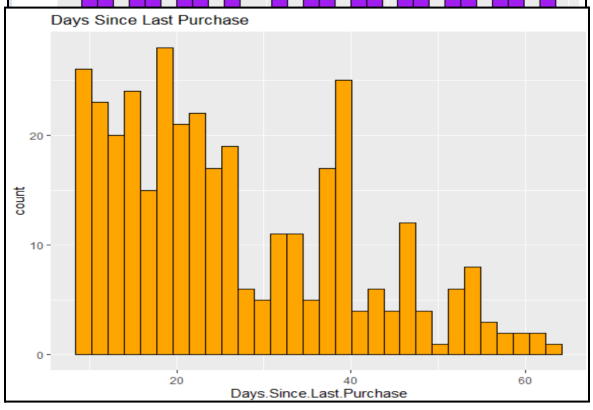
The peak at around 10 items suggests this is the most common purchase size. It indicates that customers tend to buy around this number of items more often than any other quantity.

There is a general trend where the count decreases as the number of items purchased increases. However, there are noticeable variations, with some higher quantities (like 12 or 15 items) still having a significant count. There are no clear indications of missing values or extreme outliers in the data.



Average Rating:

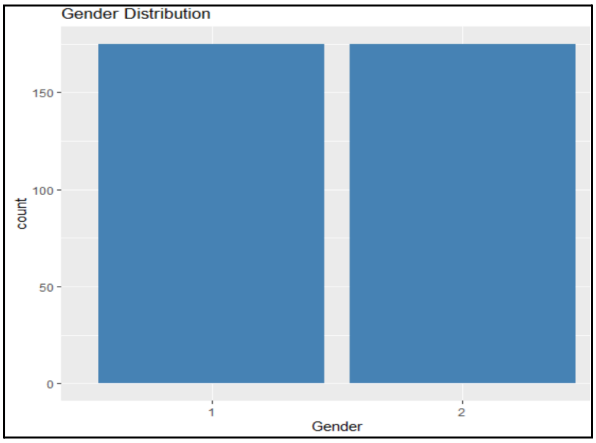
The distribution suggests that ratings are spread across the scale, with no single rating overwhelmingly dominant. However, there are noticeable peaks, especially at ratings of 3.5, 4.0, and just below 5.0, which indicates a tendency for higher ratings. There are no clear indications of missing values or extreme outliers in the data.



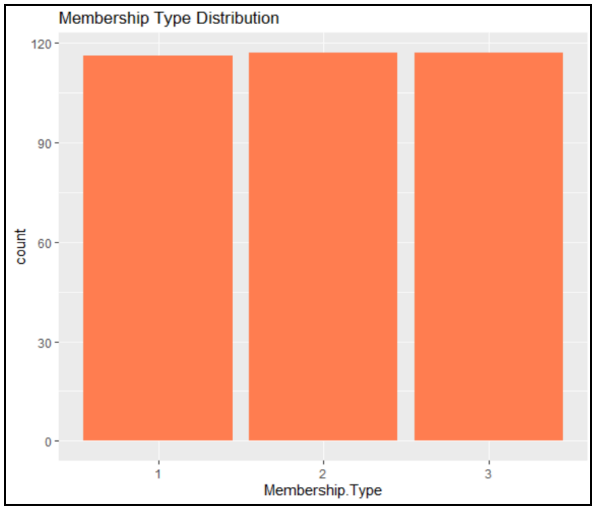
Days Since Last Purchase:

A significant number of purchases occur within a month, as shown by the concentration of higher bars between 10 and 30 days. The declining frequency of purchases as the number of days increases points towards a drop in customer retention over time. The peak around 20 days suggests a common repurchase cycle that marketing efforts could target to encourage repeat

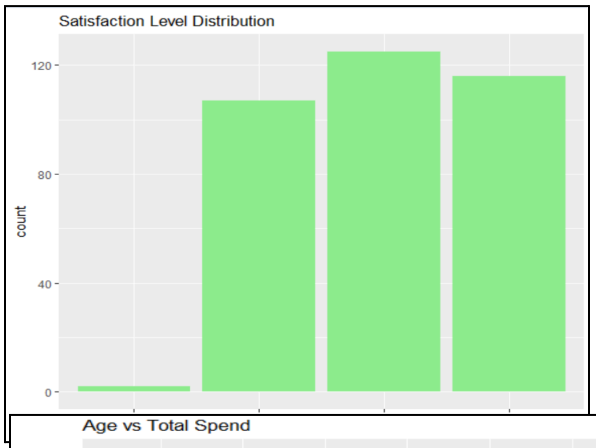
business. There are no clear indications of missing values or extreme outliers in the data.



Gender:
The plot shows a nearly equal distribution between the two gender categories, suggesting a balanced representation in the dataset.There are no clear indications of missing values or extreme outliers in the data.

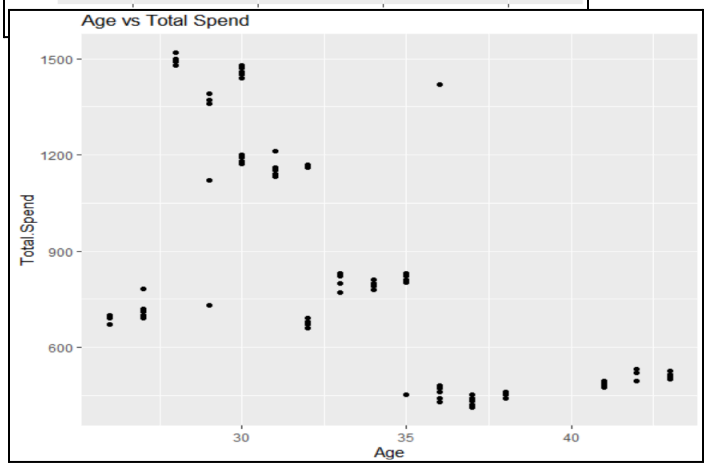


Membership Type Distribution:
The similarity in the height of the bars suggests that the membership types are fairly evenly distributed among the individuals in the dataset. There are no clear indications of missing values or extreme outliers in the data.



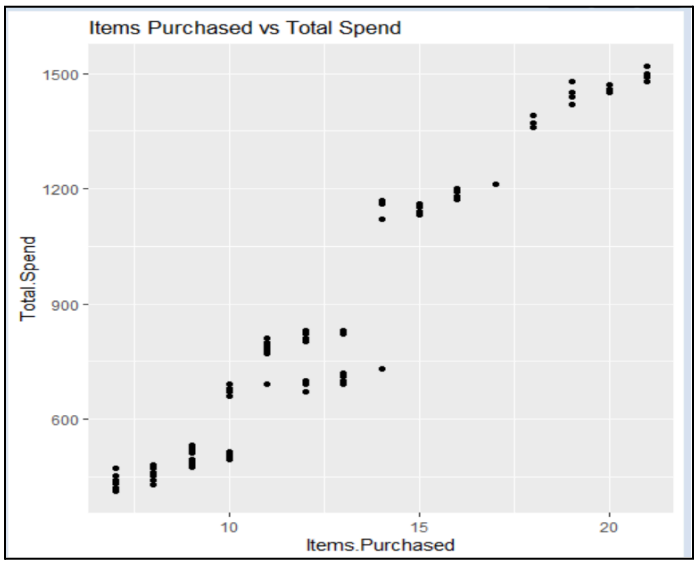
Satisfaction Level:
A greater number of individuals have reported being 'Satisfied' compared to 'Neutral' or 'Unsatisfied', which is a positive indicator for the entity providing the service or product. There are no clear indications of extreme outliers in the data but we can see a few missing values.

Bivariate Analysis



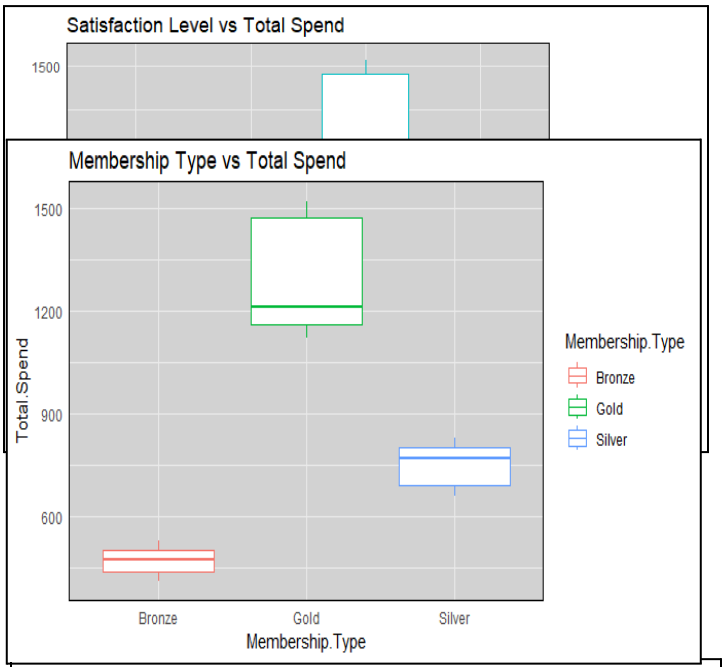
Age vs Total Spend:
The scatter plot visualizes the relationship between Age and Total.Spend. The plot reveals a negative correlation, where younger customers tend to spend more than older customers. Notably, customers under 30 years old show a wider and higher range of spending, with some individuals spending up to 1500. In contrast, customers aged 35 and above exhibit a narrower and generally lower range of spending, with most falling below 900.

Overall, the plot indicates that younger customers tend to spend more, while spending decreases with age. There are no clear indications of missing values or extreme outliers in the data.



Items Purchased vs Total Spend:
The scatter plot visualizes the relationship between Items.Purchased and Total.Spend. As the number of items purchased increases, total spending also increases, indicating a positive correlation between the two variables. The highest spending is approximately 1500 for customers who purchased around 20 items. The data points are generally concentrated along an upward linear trend, demonstrating a consistent increase in spending as the number of items purchased rises. There is some variability in spending within the 10-15 items range, with several points below and above the expected trendline. The plot does not reveal any missing values, nor does it show any extreme outliers, as all data points align well with the overall trend. Overall, the plot clearly indicates that Total.Spend rises in direct proportion to Items.Purchased, and no significant anomalies are present in the data.

Membership Type vs Total Spend:
The box plot illustrates the distribution of Total.Spend across different Membership.Type categories: Bronze, Silver, and Gold. Bronze members have a median spend of around 600, with a narrow interquartile range (IQR), indicating less variability and a concentrated distribution. Silver members show a median spend of approximately 900, with a wider IQR, representing moderate variability in spending. Gold members, on the other hand, have the highest median spend of around 1200, with the widest IQR extending up to 1500, indicating high variability and the largest total spend.
Overall, Gold members tend to spend the most, with a substantial range in spending, while Bronze members exhibit the lowest spending with the least variability. Silver members lie between Bronze and Gold in terms of spending. The positive relationship between membership type and total spending is evident, with higher-tier memberships being associated with increased spending. The plot does not indicate any missing values or extreme outliers.

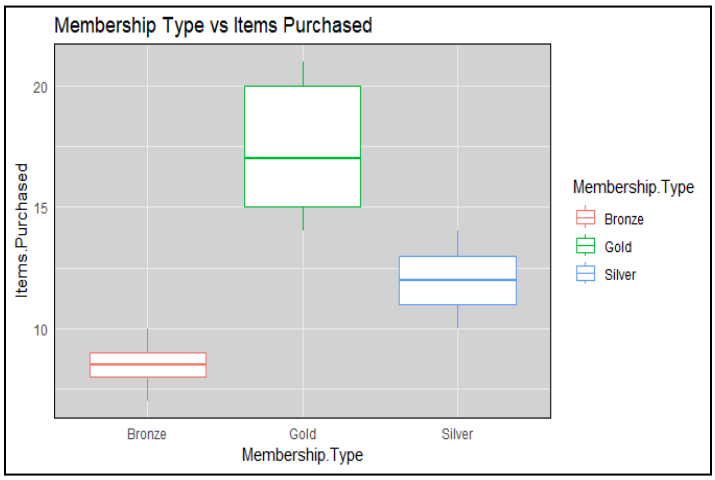


Satisfaction Level vs Total Spend:
The box plot illustrates the distribution of Total.Spend across different Satisfaction.Level categories: Neutral, Satisfied, and Unsatisfied. Customers in the Neutral category exhibit a median spend of around 600, with a wide interquartile range (IQR) extending up to 900, indicating some variability and a noticeable concentration near the lower quartile. In contrast, Satisfied customers have the highest spending, with a median of approximately 1200 and a broad IQR up to 1500, showing significant variability and a wide range of spending. The Unsatisfied category has a median spend of around 800 and a narrower IQR, revealing a more concentrated distribution around the median. Overall, Satisfied customers tend to spend the most, with a substantial range in spending, while Neutral customers have widespread but generally lower spending than Satisfied customers. Unsatisfied customers exhibit lower spending than the Satisfied group but higher than the Neutral group. The positive relationship between satisfaction level and total spending is clear, as satisfied customers tend to spend more. The plot does not indicate any extreme outliers, except a few values missing in Satisfaction Level.

Gender vs Total Spend:

The box plot illustrates the distribution of Total.Spend across different Gender categories: Female and Male. Female customers have a median spend of around 900, with an interquartile range (IQR) extending from approximately 600 to 1200, indicating moderate variability in spending. In contrast, Male customers have a median spend of around 900 as well, but their IQR extends from approximately 600 to 1500, indicating greater variability and a wider range of spending.

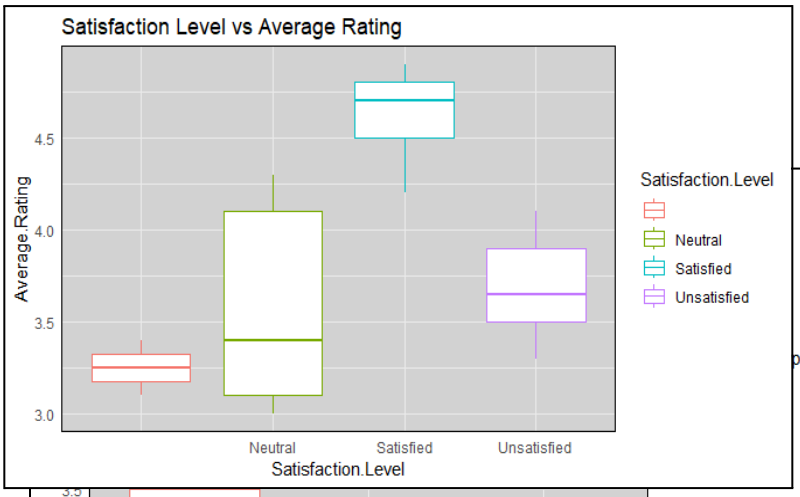
Overall, Male customers tend to spend more than Female customers, as seen in the higher upper quartile and a broader distribution range. This suggests a relationship between gender and total spending, with males generally spending more. The plot does not indicate any missing values or extreme outliers.



Membership Type vs Items Purchased:

The box plot illustrates the distribution of Items.Purchased across different Membership.Type categories: Bronze, Silver, and Gold. Bronze members have a median of around 10 items purchased, with a relatively narrow interquartile range (IQR), indicating low variability and a concentrated distribution. Silver members have a median of approximately 15 items, with a wider IQR extending up to 20, representing moderate variability in purchasing behavior. Gold members have the highest median of around 20 items, with the widest IQR extending up to 25, showing significant variability and the greatest number of items purchased.

Overall, Gold members tend to purchase the most items, followed by Silver and Bronze members, indicating a positive relationship between membership type and the number of items purchased. Higher-tier memberships are associated with purchasing more items. The plot does not indicate any missing values or extreme outliers.



Membership Type vs Average Rating:

The box plot illustrates the distribution of Average.Rating across different Membership.Type categories: Bronze, Silver, and Gold. Bronze members have a median rating of approximately 3.5, with a relatively narrow interquartile range (IQR) extending from around 3.0 to 4.0, indicating moderate variability in ratings. Silver members have a higher median rating of around 4.0, with a wider IQR extending from 3.5 to 4.5, representing greater variability in ratings. Gold members exhibit the highest median rating, around 4.5, with a relatively narrow IQR extending from 4.25 to 4.75, showing less variability and generally higher customer satisfaction.

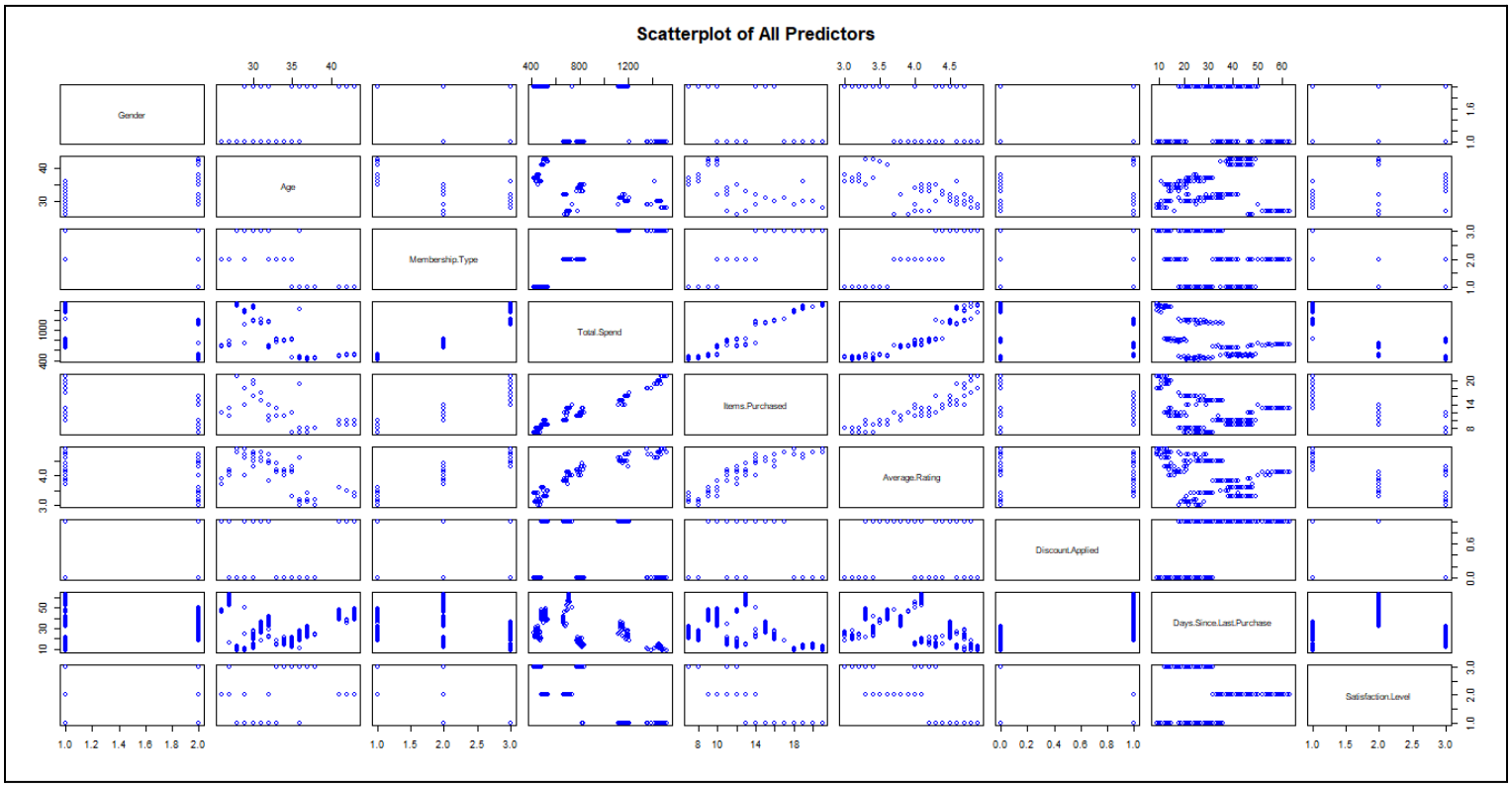
Overall, higher-tier memberships are associated with higher average ratings, indicating a positive relationship between membership type and satisfaction. The plot does not indicate any missing values or extreme outliers.

Satisfaction Level vs Average Rating:

The box plot illustrates the distribution of Average.Rating across different Satisfaction.Level categories: Neutral, Satisfied, and Unsatisfied. Neutral customers have a median rating of around 3.5, with a wide interquartile range (IQR) extending from 3.0 to

4.0, indicating significant variability in ratings. Satisfied customers exhibit the highest median rating, around 4.5, with a relatively narrow IQR extending from 4.25 to 4.75, showing less variability and generally higher satisfaction. The Unsatisfied group has a median rating of around 4.0, with a wider IQR extending from 3.75 to 4.25, representing moderate variability in ratings.
Overall, Satisfied customers tend to have the highest average ratings with less variability, while Neutral customers show the widest range of ratings with moderate satisfaction. Unsatisfied customers have ratings that lie between Neutral and Satisfied but with moderate variability. The positive relationship between satisfaction level and average rating is evident, with higher satisfaction leading to higher ratings. The plot does not indicate any missing values or extreme outliers.

Scatter Plot



The scatterplot matrix provides an insightful overview of the relationships between the various predictors in our dataset. It reveals that Items Purchased has a clear positive linear relationship with Total Spend, suggesting that customers who purchase more items tend to spend more overall. Similarly, Average Rating shows a positive correlation with Total Spend, indicating that customers who provide higher average ratings tend to spend more. Age also demonstrates a weak positive trend with Total Spend, while Membership Type presents an increasing trend, where higher-tier memberships (Gold, Silver) are linked to higher spending. The matrix also highlights strong positive correlations among Items Purchased, Average Rating, and Membership Type. On the other hand, Days Since Last Purchase and Discount Applied do not exhibit significant relationships with other predictors or Total Spend. Although **categorical variables** like Gender and Satisfaction Level show some variation across other predictors, no strong trends are immediately apparent. Overall, the matrix suggests that Items Purchased, Membership Type, and Average Rating are crucial predictors for Total Spend, while Days Since Last Purchase and Discount Applied show weak or negligible relationships. Further analysis with models like regression or classification would help quantify these relationships more effectively.

Data Cleaning/Preparation

To ensure data quality and address the issues identified during exploratory analysis, the dataset was cleaned and prepared as follows:

The "Satisfaction.Level" column contained missing values, which were detected using `colSums(is.na(data))`. All missing values were removed using `na.omit()`. Categorical variables were encoded into numerical or ordinal values: "Gender" was mapped to Male (1) and Female (2), and "Membership.Type" to Bronze (1), Silver (2), and Gold (3). The "Discount.Applied" column was converted into a binary numeric variable (1 for "TRUE" and 0 for "FALSE"), while "Satisfaction.Level" was ordered as Unsatisfied, Neutral, and Satisfied.

Two columns, "Customer.ID" and "City," were dropped due to their irrelevance to the analysis. For instance, in the scatter plot between "Age" and "Total.Spend," a noticeable negative correlation was observed, with younger customers (under 30) showing a wider and higher range of spending, up to 1500, while customers over 35 had consistently lower spending, mostly below 700. Also, there were no major outliers that needed to be dropped.

To improve model performance and reduce skewness, numerical variables, including "Total.Spend," "Items.Purchased," and "Age," were standardized using scale(). Interaction terms were created to capture meaningful relationships between variables, and the "Age_Membership.Interaction" was formed by combining "Age" and "Membership.Type." A log transformation of "Age" was attempted in the later stages but was found to be insignificant in improving model performance. Standardization using scale() and the interaction terms ultimately did not contribute significantly to the model's predictive power, so they were removed from the final analysis.

In summary, these data cleaning and preparation steps ensured that the dataset was ready for further analysis and modeling by handling missing values, dealing with outliers, transforming variables, and creating meaningful interaction terms.

```
> # Summary of the transformed data
> summary(data)
 Gender       Age        Membership.Type  Total.Spend     Items.Purchased
 1:175   Min.   :26.00   1:114           Min.   : 410.8   Min.   : 7.00
 2:173   1st Qu.:30.00   2:117           1st Qu.: 505.8   1st Qu.: 9.00
         Median :32.00   3:117           Median : 780.2   Median :12.00
         Mean   :33.58                   Mean   : 847.8   Mean   :12.63
         3rd Qu.:37.00                   3rd Qu.:1160.6   3rd Qu.:15.00
         Max.   :43.00                   Max.   :1520.1   Max.   :21.00
 Average.Rating  Discount.Applied Days.Since.Last.Purchase
 Min.   :3.000   Min.   :0.0000   Min.   : 9.00
 1st Qu.:3.500   1st Qu.:0.0000   1st Qu.:15.00
 Median :4.100   Median :1.0000   Median :23.00
 Mean   :4.024   Mean   :0.5029   Mean   :26.61
 3rd Qu.:4.500   3rd Qu.:1.0000   3rd Qu.:38.00
 Max.   :4.900   Max.   :1.0000   Max.   :63.00
   Satisfaction.Level
 Unsatisfied:116
 Neutral    :107
 Satisfied  :125
```

Part III Analysis and Findings:

Regression analysis

List of predictors

- **Original Predictors**: "Gender," "Age," "Membership.Type," "Items.Purchased," "Average.Rating," "Discount.Applied," "Days.Since.Last.Purchase," "Satisfaction.Level."
- **Transformed Predictors**: Encoded Gender, Encoded Membership Type, and Ordinal Satisfaction Level.
- **Interaction Terms**: Interaction between Age and Membership Type, which proved non-significant.

Variable Selection Using Subset Selection

Selected Predictor Terms:

- Gender: Included as a predictor, indicating differences in spending patterns between genders.
- Age: Demonstrated a positive relationship with Total Spend, suggesting that older customers tend to spend more.
- Membership Type: Significantly affects Total Spend, with higher membership levels correlating with increased spending.
- Days Since Last Purchase: Included to assess the impact of customer engagement on spending behavior.

Coefficient Estimates:

- The coefficients for Age and Membership Type were notably significant. The positive coefficient for Age implies that as age increases, so does the likelihood of higher spending, aligning with common consumer behavior trends where older individuals might have more disposable income.

- Membership Type had a strong positive coefficient, especially for higher tiers like 'Gold' membership, highlighting the effectiveness of loyalty programs in enhancing customer spending.
- Gender and Days Since Last Purchase also showed significant coefficients, suggesting their relevance in predicting Total Spend, though the magnitude and direction of these effects would require further exploration to fully understand their impact.

Full Linear Regression Model:

To explore the relationship between "Total.Spend" and various predictors, we first fitted a full linear regression model including interaction terms.

```
model <- lm(Total.Spend ~ . + Age:Membership.Type  , data = data)
summary(model)
```

Output:

```
> # Fit linear regression model with interactions
> model <- lm(Total.Spend ~ . + Age:Membership.Type  , data = data)
> summary(model)

Call:
lm(formula = Total.Spend ~ . + Age:Membership.Type, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-45.097 -10.931   1.352   9.918  48.899

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -113.2432    41.1287  -2.753  0.00622 **
Gender2                   -0.8858    11.4684  -0.077  0.93848
Age                        8.8900     1.1788   7.541 4.41e-13 ***
Membership.Type2         363.2582    73.2230   4.961 1.12e-06 ***
Membership.Type3         781.7379    82.4414   9.482  < 2e-16 ***
Items.Purchased           28.3784     1.3010  21.813  < 2e-16 ***
Average.Rating            28.6193     6.8545   4.175 3.80e-05 ***
Discount.Applied        -138.7055    13.0703 -10.612  < 2e-16 ***
Days.Since.Last.Purchase   -1.5961     0.2440  -6.542 2.27e-10 ***
Satisfaction.Level.L      -83.6435     9.9012  -8.448 9.17e-16 ***
Satisfaction.Level.Q       39.5778     6.2484   6.334 7.69e-10 ***
Age:Membership.Type2       -4.1875     1.9394  -2.159  0.03154 *
Age:Membership.Type3       -3.8451     2.1764  -1.767  0.07819 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.88 on 335 degrees of freedom
Multiple R-squared:  0.9981,    Adjusted R-squared:  0.9981
F-statistic: 1.498e+04 on 12 and 335 DF,  p-value: < 2.2e-16
```

Residuals

The residuals' summary shows that most predictions are close to the true values, with a median near zero. However, the range of residuals (-45.097 to 48.899) indicates the presence of some outliers or high-leverage points that could be skewing the data or indicating model misspecification.

Coefficients

Intercept: Represents the baseline total spend when all other variables are zero. It's statistically significant (p = 0.00622), but negative (-$113.24), suggesting that, when all predictors are at their reference levels, the starting spend amount is negative, possibly indicating model misspecification or the presence of other factors affecting the baseline spend.

Gender (2): The coefficient is not statistically significant (p = 0.93848), indicating that being of gender coded as '2' does not significantly affect total spending.

Age: Each additional year of age increases total spending by about $8.89, which is highly significant (p < 4.41e-13).

Membership Type:

Type 2: Compared to the baseline, Membership Type 2 increases spending by $363.26, which is highly significant (p = 1.12e-06).

Type 3: Compared to the baseline, Membership Type 3 increases spending by $781.74, which is also highly significant (p < 2e-16).

Items Purchased: A strong predictor, with each additional item increasing total spending by $28.38, which is highly significant (p < 2e-16).

Average Rating: Also a positive predictor, where higher ratings are associated with an increase in spending by $28.62, significant at the 0.001 level (p = 3.80e-05).

Discount Applied: The coefficient is highly significant (p < 2e-16) and shows that applying a discount reduces spending by $138.71.

Days Since Last Purchase: Shows a negative relationship with spending; as the days since the last purchase increase, total spending decreases by about $1.60 per day, which is highly significant (p = 2.27e-10).

Satisfaction Level:

Linear Term: Significant and negative (p = 9.17e-16), indicating that increasing satisfaction decreases spending.

Quadratic Term: Significant and positive (p = 7.69e-10), suggesting a quadratic relationship between satisfaction and spending, where spending decreases initially but increases afterward.

Interaction Terms

Age and Membership Type:

Age:Membership Type 2: Significant (p = 0.03154), indicating that the effect of age on spending varies by Membership Type 2.

Age:Membership Type 3: Marginally significant (p = 0.07819), indicating that the effect of age on spending may vary by Membership Type 3.

Model Fit

Residual Standard Error: 15.88 indicates that the typical prediction error is about $15.88.

Multiple R-squared: 0.9981 suggests that about 99.81% of the variability in total spending is explained by the model, indicating a very good fit.

Adjusted R-squared: 0.9981, adjusted for the number of predictors, is still extremely high, reaffirming a strong model.

F-statistic: The overall model is highly statistically significant (p < 2.2e-16), meaning the predictors, as a set, reliably differentiate spending amounts.

Model:

```
> # correlation matrix calculation on numeric data only
> cor_matrix <- cor(data[train, sapply(data[train, ], is.numeric)], use = "pairwise.complete.obs")
> print(cor_matrix)
```

```
> cor_matrix <- cor(data[train, sapply(data[train, ], is.numeric)], use = "pairwise.complet
> print(cor_matrix)
                            Age Total.Spend Items.Purchased Average.Rating
Age                   1.0000000  -0.7181111      -0.7131129     -0.7530725
Total.Spend          -0.7181111   1.0000000       0.9746086      0.9439729
Items.Purchased      -0.7131129   0.9746086       1.0000000      0.9262986
Average.Rating       -0.7530725   0.9439729       0.9262986      1.0000000
Days.Since.Last.Purchase  0.2648497  -0.5640429      -0.4552064     -0.4555696
                      Days.Since.Last.Purchase
Age                                  0.2648497
Total.Spend                         -0.5640429
Items.Purchased                     -0.4552064
Average.Rating                      -0.4555696
Days.Since.Last.Purchase             1.0000000
>
```

Correlation Matrix Summary

Correlation Coefficient: This statistic measures the strength and direction of a linear relationship between two variables, ranging from -1 to +1. A value close to +1 implies a strong positive relationship, close to -1 implies a strong negative relationship, and around 0 implies no linear relationship.

Interpretation of Key Correlations

1. Age and Other Variables

Total Spend (-0.7181): There is a strong negative correlation between age and total spending, which is counterintuitive given the hypothesis that older customers might spend more. This suggests younger customers tend to spend more in this dataset.

Items Purchased (-0.7131): Age also negatively correlates with the number of items purchased, indicating that younger customers buy more items.

Average Rating (-0.7531): Age negatively correlates with average ratings, suggesting younger customers may give higher ratings.

Days Since Last Purchase (0.2648): A weak positive correlation suggests that older customers may have longer intervals since their last purchase.

2. Total Spend and Other Variables

Items Purchased (0.9746): There is a very strong positive correlation between total spend and items purchased, indicating that spending increases substantially as more items are bought.

Average Rating (0.9439): Total spend is also strongly positively correlated with average ratings, suggesting that higher spending might be associated with higher satisfaction or quality perception.

Days Since Last Purchase (-0.5640): There is a moderate negative correlation, showing that more recent purchases are associated with higher total spending.

3. Items Purchased and Average Rating

Average Rating (0.9263): This strong positive correlation indicates that purchases involving more items tend to receive higher ratings, possibly reflecting higher customer satisfaction with larger purchases.

4. Days Since Last Purchase

Shows negative correlations with total spend, items purchased, and average rating, indicating that more recent engagements with the store are likely to be more positive in terms of spending and satisfaction.

Implications for the Model

Multicollinearity Concern: The very high correlations among total spend, items purchased, and average rating could lead to multicollinearity if these variables are used together in regression models, potentially inflating the variance of the coefficient estimates and making them unstable and difficult to interpret. This might necessitate looking into dimensionality reduction techniques or choosing between correlated predictors based on their business relevance.

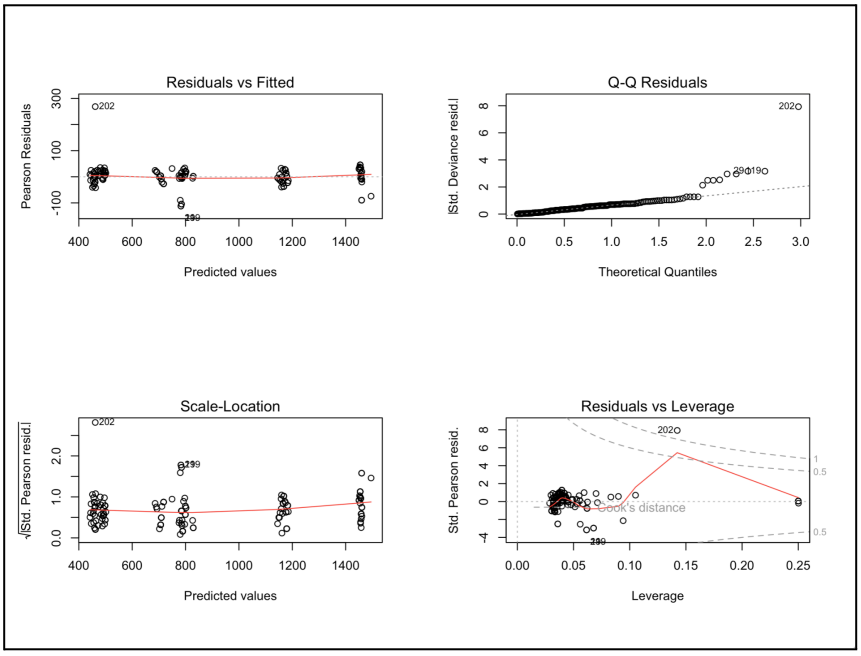
Counter-Intuitive Findings with Age: The negative correlations involving age and spending-related variables contradict initial hypotheses, which may require revisiting assumptions about the target customer demographics or further segmenting the analysis by other factors such as membership type or geographical differences.

Best Subset Selection

The `regsubsets` function was used to select the best subsets of predictors for predicting `Total.Spend` while excluding `Items.Purchased` and `Average.Rating` due to their high correlations with other predictors. The selection was done exhaustively for up to 7 variables.

Model Diagnostics Plot

The provided diagnostics plots are standard when evaluating the fit of a regression



Residuals vs Fitted: This plot shows how the residuals (the differences between observed and predicted values) vary across the predicted values. You want to see no discernible pattern; the red line (a lowess smoother) should be approximately horizontal at zero. In the provided plot, there are some outliers, but overall there seems to be no obvious pattern, suggesting that the model's assumptions are reasonable for the majority of data.

Normal Q-Q Plot: This plot displays the standardized residuals against the theoretical quantiles from a normal distribution. If the residuals are normally distributed, the points should fall approximately along the 45-degree

reference line. The provided plot shows some deviations at the ends, indicating potential issues with normality, especially with some extreme values.

Scale-Location (or Spread-Location) Plot: This plot shows the square root of the absolute standardized residuals against the predicted values. Ideally, you want to see a random scatter of points, with no distinct patterns and a roughly horizontal line. The plot you provided shows a fairly random scatter, which is a good sign for homoscedasticity (constant variance of residuals).

Residuals vs Leverage: This plot helps identify influential cases, i.e., observations that have a disproportionate impact on the estimation of the regression coefficients. The plot provided indicates one point with high leverage, suggesting it might be an influential observation.

Coefficient Estimates:
The Intercept is estimated at $310.81, representing the baseline total spend when all other variables are at their reference levels. The coefficient for Age is $7.82, indicating that each additional year of age increases total spending by this amount. Membership Type 2 and Type 3 increase spending by $354.06 and $1019.31, respectively, compared to the baseline membership. Applying a Discount reduces spending by $313.45, while each additional day since the last purchase decreases spending by $1.23. The quadratic term of Satisfaction Level is estimated at $152.64, showing that spending increases as satisfaction levels rise, suggesting a positive quadratic relationship.

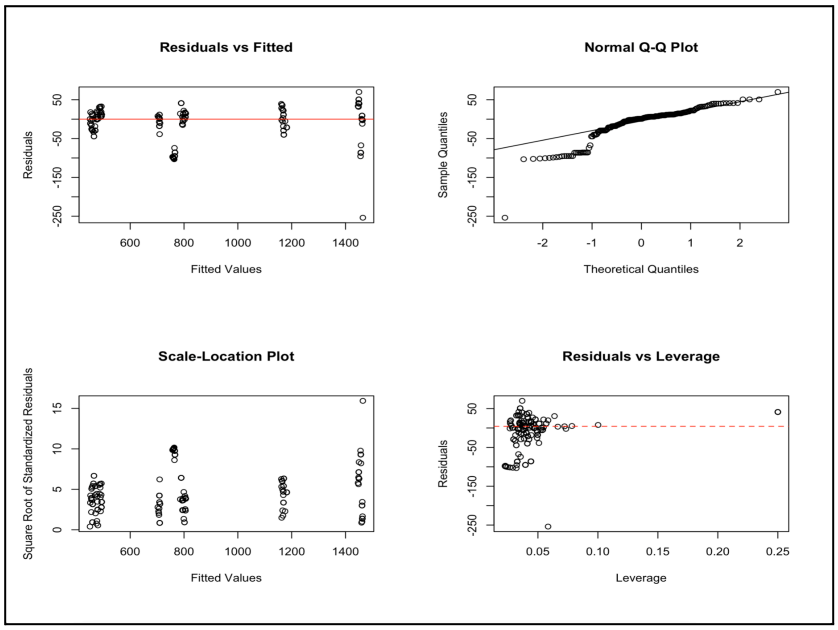
Cross-Validation and Model Selection
Validation errors were computed for each subset model to avoid overfitting and to select the best model for predicting new data. The errors (Mean Squared Error - MSE) for models ranging from 1 to 7 predictors were calculated, with the model using 6 predictors showing the lowest validation error, suggesting it may be the best choice among the tested models for balancing bias and variance.

- The best subset model size was 6, and the selected predictors were "Age," "Membership.Type2," "Membership.Type3," "Discount.Applied," "Days.Since.Last.Purchase," and "Satisfaction.Level.Q."

Generalized Linear Model (GLM) Summary
A GLM was fitted using the training data with the selected predictors, and the summary indicates several significant predictors of `Total.Spend`:

- `Gender2`, `Age`, `Membership.Type2`, `Membership.Type3`, and `Days.Since.Last.Purchase` are statistically significant predictors of total spend.
- The model's AIC and the significant reduction in deviance from the null model indicate a good fit.

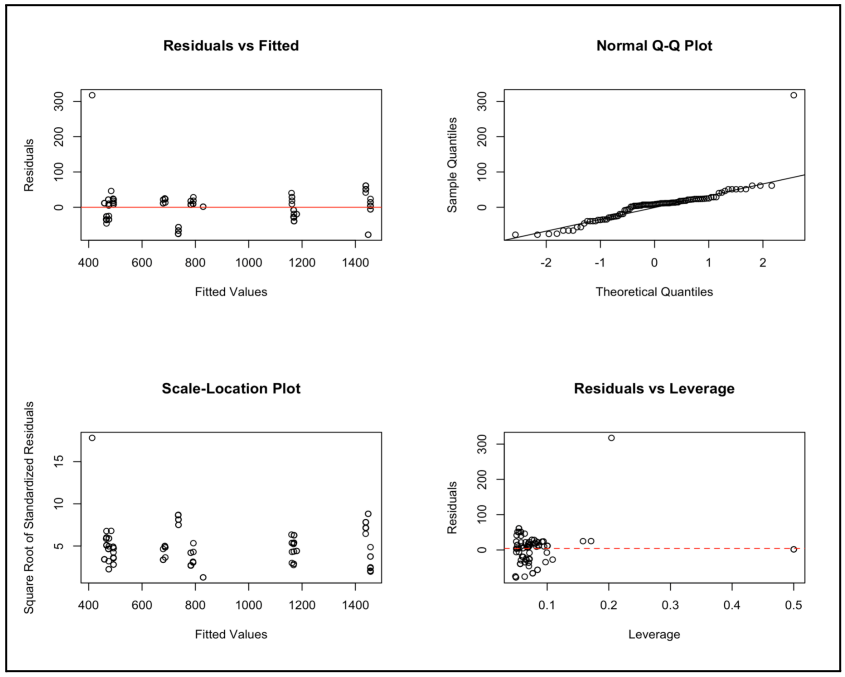
Diagnostic Plots for the Best Model
Further diagnostic plots were prepared to assess the best-fitting model's assumptions and fit. These plots are crucial to validate the final model's underlying assumptions, like linearity, independence, and homoscedasticity of residuals, and normality of error terms.



Residual Plots for the Final Model

After identifying the best model through cross-validation, additional residual diagnostics were plotted to confirm the validity of the model assumptions.

Forward Subset Selection



This plot helps to check the assumption of linearity and homoscedasticity (constant variance) of residuals.

The plot shows a random scatter without a clear pattern, which is a good sign for linearity. The residuals seem to have a constant variance as there is no funnel shape.

Normal Q-Q Plot

The plot indicates that the residuals deviate from normality, especially at the tails, with several points deviating significantly from the line at both ends. This could suggest the presence of outliers or that the normality assumption is not fully met.

Scale-Location Plot

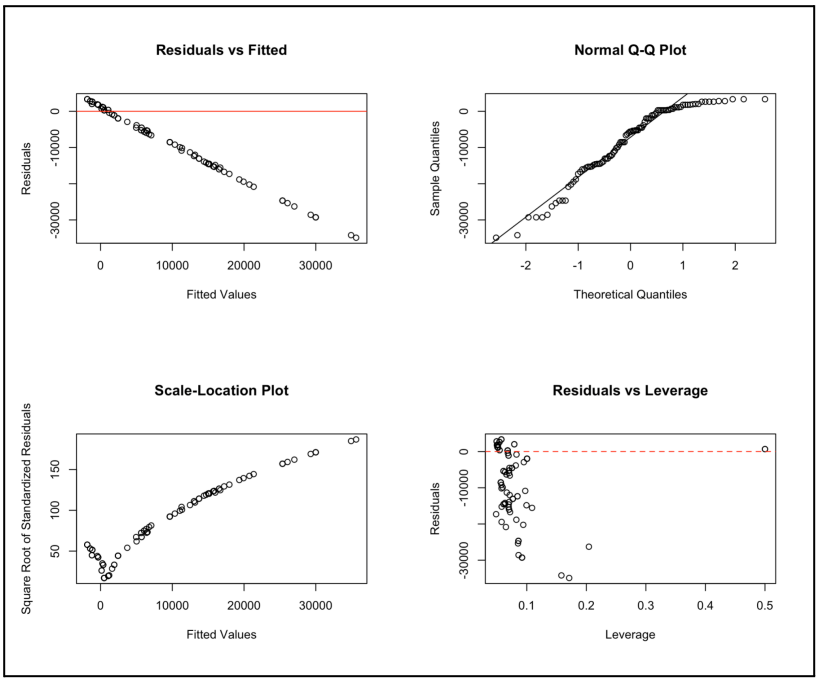
There is a cluster of points at the lower end of fitted values with a random spread, suggesting that the assumption of equal variance is generally met, although the higher spread at the lower range of fitted values may be a point of consideration.

Residuals vs Leverage
There are a few points with higher leverage, but they do not appear to have large residuals, suggesting that while they have influence, they are not having a problematic impact on the model's predictions.

Coefficient Estimates:
The coefficient estimates for the forward selection model were as follows: the Intercept was estimated at $343.01, representing the baseline total spend when all other variables are at their reference levels. The coefficient for Gender2 was $11.84, indicating that male customers spend this much more than females (reference). The coefficient for Age was $7.82, showing that each additional year of age increases total spending by this amount.
Membership Type 2 and Type 3 increased spending by $354.06 and $1020.31, respectively, compared to the Bronze (baseline) membership. Applying a Discount reduced spending by -$313.45, while each additional day since the last purchase decreased spending by -$1.23. The linear term of Satisfaction Level (L) was estimated at -$197.35, suggesting that customers with a linear (unsatisfied) satisfaction level spent this much less than satisfied customers, while the quadratic term (Q) was $152.64, showing that spending increases as satisfaction levels rise.
Forward Subset Selection and Validation
Forward subset selection was applied to the training set, choosing variables stepwise based on their contribution to the model fit. After evaluating models with up to 7 predictors based on the lowest mean squared error (MSE) on a validation set, the best model was selected.
Despite the reasonable diagnostic plots, caution should be used due to the deviations in the normal Q-Q plot. It is important to investigate whether the outliers identified could be data entry errors, require transformation, or represent a more complex relationship not captured by the current model. Additional robustness checks or the inclusion of additional relevant variables might also be considered.

- The best forward model size was 7, and the selected predictors were "Gender2," "Age," "Membership.Type2," "Membership.Type3," "Discount.Applied," "Days.Since.Last.Purchase," "Satisfaction.Level.L," and "Satisfaction.Level.Q."

Backwards Subset Selection



Residuals vs Fitted:
There seems to be a clear pattern where the residuals decrease as the fitted values increase, which is a strong indication of non-linearity in the data that the model has not captured. This could also be a sign of heteroscedasticity, as the spread of the residuals appears to be decreasing with the increase in the fitted values.

Normal Q-Q Plot:
The residuals diverge from the line in a systematic way, suggesting that the residuals do not follow a normal distribution. This is especially evident in the tails of the distribution.

Scale-Location Plot:
The residuals don't spread randomly around the horizontal axis, which might indicate that the assumption of constant variance is violated (heteroscedasticity). Particularly, the residuals appear larger for smaller fitted values.
Residuals vs Leverage:
Most data points have low leverage, but a few points stand out with higher leverage. However, these points do not correspond to large residuals, so they may not be unduly influencing the model.

Coefficient Estimates:
The coefficient estimates for the backward selection model were as follows: the Intercept was estimated at $219.80, representing the baseline total spend when all other variables are at their reference levels. The coefficient for Age was $9.61, showing that each additional year of age increases total spending by this amount.
Membership Type 2 and Type 3 increased spending by $346.16 and $1009.84, respectively, compared to the Bronze (baseline) membership. Applying a Discount reduced spending by -$302.08, while the linear term of Satisfaction Level (L) was estimated at -$153.65, suggesting that customers with a linear (unsatisfied) satisfaction level spent this much less than satisfied customers. The quadratic term (Q) was $133.67, indicating that spending increases as satisfaction levels rise, showing a positive quadratic relationship.

Model Selection and Summary:

Backward selection was used to identify the best model based on the Akaike Information Criterion (AIC), leading to the exclusion of the `Days.Since.Last.Purchase` predictor.

- The summary of the final model from the backward selection shows that `Gender`, `Age`, `Membership.Type`, and `Satisfaction.Level` are significant predictors.
- The model has an excellent Multiple R-squared value, suggesting that it explains most of the variance in `Total.Spend`.
- The best backward model size was 6, and the selected predictors were "Age," "Membership.Type2," "Membership.Type3," "Discount.Applied," "Satisfaction.Level.L," and "Satisfaction.Level.Q."

Shrinkage (Ridge and Lasso Regression)

Selected Predictor Terms:
Similar to the subset selection approach, the shrinkage methods focused on Gender, Age, Membership Type, and Days Since Last Purchase. These methods help in handling multicollinearity, enhancing the model's prediction ability by incorporating penalties that shrink less important coefficients.

Coefficient Estimates:

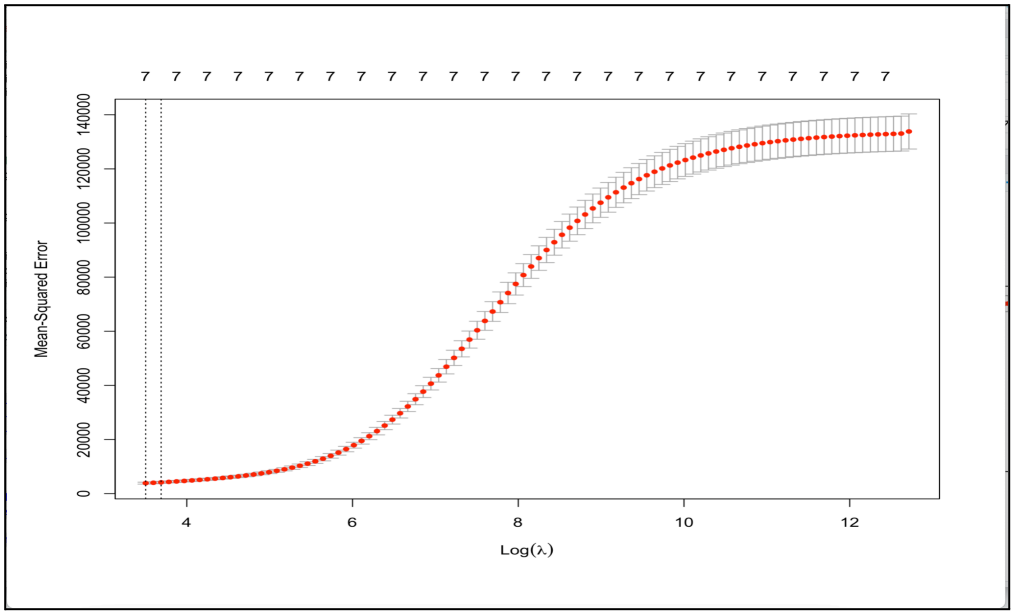
- **Lasso Regression:** Particularly effective in feature selection, Lasso adjusted the contributions of each predictor by shrinking some coefficients to zero. This method helps in identifying the most relevant predictors by eliminating those with negligible effects, thus simplifying the model without sacrificing predictive power.
- The coefficients for predictors like Membership Type and Age remained robust and significant in the Lasso model, emphasizing their strong predictive power. In contrast, coefficients for some other variables were reduced to zero, indicating they might be redundant or less important when other variables are considered.
- **Ridge Regression**: Although it does not reduce coefficients to zero, it reduces their size, which helps in cases of multicollinearity. Ridge regression was less selective compared to Lasso but still provided insights by penalizing the size of the coefficients across all predictors, maintaining all in the model but adjusting their influence.

Comparison of Shrinkage Methods:

- Lasso was more advantageous for this dataset as it not only provided a model with good predictive accuracy but also facilitated model interpretability by zeroing out less important predictors. This property of Lasso makes it particularly useful in scenarios where model simplicity and interpretability are as crucial as prediction accuracy.
- Ridge regression, while useful in reducing the impact of multicollinearity, did not offer the feature selection capability of Lasso, making it less favorable in cases where the goal is to identify the most impactful predictors.

In summary, the application of subset selection and shrinkage methods revealed essential insights into the spending behavior of e-commerce customers. The strong performance of Lasso regression underscores its utility in extracting meaningful patterns from complex datasets, guiding targeted business strategies and decision-making.

Ridge Regression:



The plot represents the cross-validation scores for different values of the regularization parameter λ during ridge regression. This regularization technique penalizes the size of the coefficients in the model, effectively shrinking them to avoid overfitting.

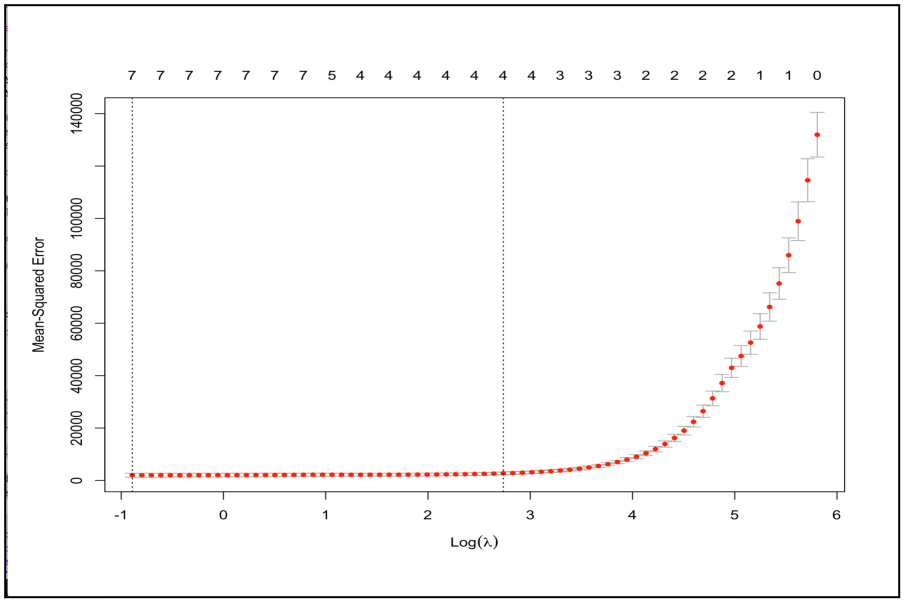
Lambda and Error: The plot shows that as the log(λ) increases, the mean squared error (MSE) rises, reaching a minimum around a specific value of λ (best.lambda.ridge). The selected value of λ minimizes the cross-validation error, providing a trade-off between bias and variance.

Results:

- The selected predictor terms included "Gender," "Age," "Membership.Type2," "Membership.Type3," "Discount.Applied," "Days.Since.Last.Purchase," "Satisfaction.Level.L," and "Satisfaction.Level.Q."
- The coefficient estimates for the Ridge model revealed that the Intercept was estimated at $311.94, representing the baseline total spend when all other variables were at their reference levels. The coefficient for Gender was $10.15, indicating that male customers spend this much more than females. The coefficient for Age was $6.95, showing that each additional year of age increases total spending by this amount. Membership Type 2 and Type 3 increased spending by $355.67 and $1010.24, respectively, compared to the baseline Bronze membership.
  Applying a Discount reduced spending by -$311.82, while each additional day since the last purchase decreased spending by -$1.21. For Satisfaction Level, the linear term (L) was estimated at -$192.34, showing that customers with a linear (unsatisfied) satisfaction level spent this much less than satisfied customers, while the quadratic term (Q) was $144.87, suggesting that spending increases as satisfaction levels rise.

- The mean cross-validated test error for the Ridge model was 2512.73.

Lasso Regression:



Lasso regularizes the coefficients, but in a way that can reduce some coefficients to exactly zero, thus performing variable selection.
Lambda and Error: This plot also shows an optimal value of λ (best.lambda.lasso) where the cross-validation error is minimized. Lasso typically selects a simpler model with fewer parameters compared to ridge.

Results:

- The selected predictor terms for the Lasso model included "Age," "Membership.Type2," "Membership.Type3," "Discount.Applied," "Days.Since.Last.Purchase," "Satisfaction.Level.L," and "Satisfaction.Level.Q."
- The coefficient estimates for the Lasso model showed that the Intercept was estimated at $310.81, representing the baseline total spend when all other variables were at their reference levels. The coefficient for Age was $7.82, showing that each additional year of age increases total spending by this amount.
  Membership Type 2 and Type 3 increased spending by $372.43 and $1015.82, respectively, compared to the baseline Bronze membership. Applying a Discount reduced spending by -$319.91, while each additional day since the last purchase decreased spending by -$1.26. For Satisfaction Level, the linear term (L) was estimated at -$205.55, showing that customers with a linear (unsatisfied) satisfaction level spent this much less than satisfied customers, while the quadratic term (Q) was $155.87, suggesting that spending increases as satisfaction levels rise.

- The mean cross-validated test error for the Lasso model was 1032.02.

Mean Cross-Validated Errors

The minimum mean cross-validated error for ridge regression is higher than that for lasso. This indicates that for this particular dataset, lasso regression provides a better fit as it is able to reduce the error more effectively.

Optimal Model

```
> # Print mean cross-validated test errors
> print(paste("Ridge Regression:", cv.errors_ridge))
[1] "Ridge Regression: 2512.72946896487"
> print(paste("Lasso Regression:", cv.errors_lasso))
[1] "Lasso Regression: 1032.02182742454"
> # Determine the optimal model
> optimal_errors <- c(Ridge = cv.errors_ridge, Lasso = cv.errors_lasso)
> optimal_model <- names(optimal_errors)[which.min(optimal_errors)]
> print(paste("Optimal Model:", optimal_model))
[1] "Optimal Model: Lasso"
>
```

Based on the mean cross-validated test errors, the lasso regression is determined to be the optimal model since it has the lowest test error.
Lasso regression is the preferred model for this dataset, as indicated by the lower cross-validated test error. This suggests that the process of shrinking and selecting variables was beneficial for model performance.

Findings Based on the Optimal Model:

- In summary, the optimal model's findings suggest that businesses should focus on promoting higher-tier memberships, offering personalized discounts and reminders to customers with longer gaps between purchases, and enhancing customer satisfaction, as satisfied customers tend to spend more. Efforts in these areas will likely lead to increased total spending and better customer retention.

Logistic regression model:

Model:

```
glm.fit <- glm(Satisfaction.Level ~ Gender + Age + Membership.Type + Total.Spend + Days.Since.Last.Purchase,
               family = binomial,
               data = trainData)
```

```
Call:
glm(formula = Satisfaction.Level ~ Gender + Age + Membership.Type +
    Total.Spend + Days.Since.Last.Purchase, family = binomial,
    data = trainData)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.499e+03  1.944e+05  -0.008    0.994
Gender2                   2.948e+02  6.004e+04   0.005    0.996
Age                      -1.848e+01  2.340e+03  -0.008    0.994
Membership.Type2         -8.975e+02  1.196e+05  -0.008    0.994
Membership.Type3         -2.365e+03  2.879e+05  -0.008    0.993
Total.Spend               3.679e+00  4.384e+02   0.008    0.993
Days.Since.Last.Purchase  2.491e-01  9.225e+02   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.6410e+02  on 278  degrees of freedom
Residual deviance: 1.6959e-07  on 272  degrees of freedom
AIC: 14

Number of Fisher Scoring iterations: 25
```

Model Specification:

- The model is specified to predict `Satisfaction.Level` as a function of several variables.
- Gender, Membership.Type, Total.Spend, and Days.Since.Last Purchase are used as independent variables.
- The 'family' argument specifies that the dependent variable is binomial, which is appropriate for a binary outcome.

Model Summary:

Coefficients: The model estimates for the regression coefficients are displayed, but they are extremely large in magnitude, indicating potential issues with the model fitting process.

Standard Errors: The standard errors for the coefficients are also extremely large, which usually suggests problems with the data, such as perfect separation, where a predictor or combination of predictors can perfectly predict the outcome.

z-values and p-values: The z-values are near zero, and p-values are non-significant for all predictors, which means that none of the predictors are statistically significant according to this model. However, given the size of the coefficients and standard errors, these p-values might not be trustworthy.

Null and Residual Deviance: The null deviance represents the fit of a model that includes only the intercept (no predictors) and is a measure of how well the response variable is predicted by a model with no predictors. The residual deviance shows how well the response variable is predicted by the model when the predictors are included. In this output, the residual deviance is virtually zero, indicating an overfit model.

AIC: The Akaike Information Criterion is used to compare models, with lower values generally indicating a better fit. The extremely low AIC here again suggests that the model may be overfitting the data.

Model2:

```
glm.fit1 <- glm(Satisfaction.Level ~ Gender + Age + Membership.Type + Total.Spend + Days.Since.Last.Purchase,
```

```
family = binomial,
data = trainData,
control = glm.control(epsilon = 1e-8, maxit = 50))
```

```
Call:
glm(formula = Satisfaction.Level ~ Gender + Age + Membership.Type +
    Total.Spend + Days.Since.Last.Purchase, family = binomial,
    data = trainData, control = glm.control(epsilon = 1e-08,
        maxit = 50))

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.985e+03  3.733e+06  -0.001        1
Gender2                  3.912e+02  1.160e+06   0.000        1
Age                     -2.447e+01  4.492e+04  -0.001        1
Membership.Type2        -1.191e+03  2.283e+06  -0.001        1
Membership.Type3        -3.138e+03  5.502e+06  -0.001        1
Total.Spend              4.878e+00  8.371e+03   0.001        1
Days.Since.Last.Purchase 2.464e-01  1.815e+04   0.000        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.6410e+02  on 278  degrees of freedom
Residual deviance: 4.6222e-10  on 272  degrees of freedom
AIC: 14

Number of Fisher Scoring iterations: 31

>
```

- **Coefficients**: Extremely large or small coefficient estimates suggest numerical instability in the estimation process. This is commonly due to complete separation in logistic regression, where the predictors can almost perfectly distinguish between the classes in the response variable.
- **Standard Errors**: The very large standard errors indicate uncertainty in the coefficient estimates, which is consistent with the problems identified by the warning message.
- **Z-values and P-values**: The z-values are essentially zero, and the associated p-values are 1 (or rounded up to 1), indicating that the predictors are not statistically significant. In this context, the p-values are not meaningful because they arise from the numerical problems in the model rather than from a lack of association between the predictors and the outcome.
- **Null and Residual Deviance**: The null deviance indicates the goodness of fit of a model with no predictors, and the residual deviance indicates the goodness of fit for the model with predictors. The extremely low residual deviance suggests that the model fits the training data almost perfectly, which is a sign of overfitting, especially given the other issues noted.
- **AIC**: The AIC is meant to balance the model's goodness of fit with the number of parameters, penalizing complexity. However, the low AIC in this output is likely an artifact of the same problems causing the large coefficient estimates and should not be interpreted as indicating a good model.
- **Number of Fisher Scoring iterations**: The model used 31 iterations to converge, which is more than the default setting, indicating that it struggled to find a solution that satisfies the convergence criteria.
- The **coefficient estimates** are as follows:
    - **Intercept**: Estimate = -1985, Std. Error ≈ 3,733,000
    - **Gender2**: Estimate = 391.2, Std. Error ≈ 1,160,000
    - **Age**: Estimate = -24.47, Std. Error ≈ 44,920
    - **Membership.Type2**: Estimate = -1191, Std. Error ≈ 2,283,000
    - **Membership.Type3**: Estimate = -3138, Std. Error ≈ 5,502,000
    - **Total.Spend**: Estimate = 4.878, Std. Error ≈ 8,371
    - **Days.Since.Last.Purchase**: Estimate = 0.2464, Std. Error ≈ 18,150

All predictor p-values were approximately 1, indicating no statistically significant findings.

- **Model Fit Diagnostics**:
  - Null deviance: 364.10 on 278 degrees of freedom
  - Residual deviance: 4.6222e-10 on 272 degrees of freedom
  - Akaike Information Criterion (AIC): 14

Model3:

glm.fit.brglm <- brglm(Satisfaction.Level ~ Gender + Age + Membership.Type + Total.Spend + Days.Since.Last.Purchase, family = binomial, data = trainData)

```
> summary(glm.fit.brglm)

Call:
brglm(formula = Satisfaction.Level ~ Gender + Age + Membership.Type +
    Total.Spend + Days.Since.Last.Purchase, family = binomial,
    data = trainData)


Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -116.7674    41.0217  -2.846  0.00442 **
Gender2                   20.1939     6.3998   3.155  0.00160 **
Age                       -1.5072     0.5404  -2.789  0.00529 **
Membership.Type2          -61.7281    20.8475  -2.961  0.00307 **
Membership.Type3         -171.3191    55.9475  -3.062  0.00220 **
Total.Spend                 0.2718     0.0869   3.128  0.00176 **
Days.Since.Last.Purchase    0.3800     0.1544   2.461  0.01386 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 349.3936  on 278  degrees of freedom
Residual deviance:   7.3806  on 272  degrees of freedom
Penalized deviance: -3.39993
AIC:  21.381

>
```

The summary of the brglm model provides the following coefficients, standard errors, z-values, and associated p-values:

- **Coefficients:**
  - **Intercept**: Estimate = -116.7674, Std. Error = 41.0217, z-value = -2.846, p-value = 0.00442.
  - **Gender2**: Estimate = 20.1939, Std. Error = 6.3998, z-value = 3.155, p-value = 0.00160.
  - **Age**: Estimate = -1.5072, Std. Error = 0.5404, z-value = -2.789, p-value = 0.00529.
  - **Membership.Type2**: Estimate = -61.7281, Std. Error = 20.8475, z-value = -2.961, p-value = 0.00307.
  - **Membership.Type3**: Estimate = -171.3191, Std. Error = 55.9475, z-value = -3.062, p-value = 0.00220.
  - **Total.Spend**: Estimate = 0.2718, Std. Error = 0.0869, z-value = 3.128, p-value = 0.00176.
  - **Days.Since.Last.Purchase**: Estimate = 0.3800, Std. Error = 0.1544, z-value = 2.461, p-value = 0.01386.
- **Model Fit Diagnostics:**
  - Null deviance: 349.3936 on 278 degrees of freedom.
  - Residual deviance: 7.3806 on 272 degrees of freedom.
  - Penalized deviance: -3.39993.
  - Akaike Information Criterion (AIC): 21.381.

The brglm model yielded statistically significant predictors, which is a marked improvement over the initial glm fit that encountered convergence issues. This suggests that brglm is effectively addressing the separation issue and providing more reliable coefficient estimates.

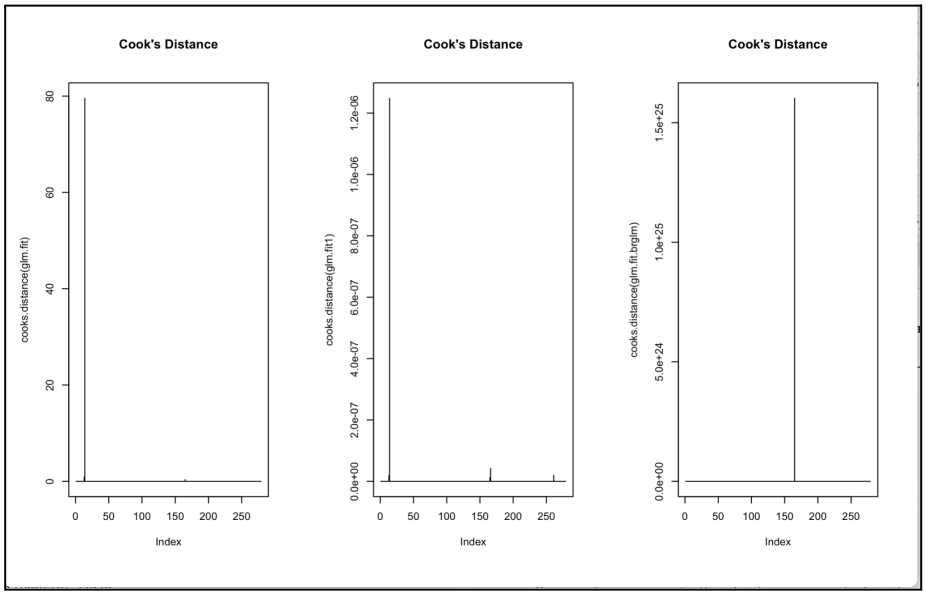
Each predictor's sign and significance suggest the following associations with the likelihood of customer satisfaction:

- **Gender**: Being of gender coded as '2' (presumably female if '1' is male) is positively associated with satisfaction.
- **Age**: An increase in age is associated with a decrease in the likelihood of satisfaction.
- **Membership Type**: Higher membership types are associated with lower satisfaction in comparison to the base category.
- **Total Spend**: An increase in total spend is associated with an increase in the likelihood of satisfaction.
- **Days Since Last Purchase**: More recent interactions (fewer days since the last purchase) are associated with increased satisfaction.

Best Model:

The bias-reduction logistic regression (glm.fit.brglm) model is the best among the three. It provides a stable solution, meaningful coefficient estimates, and a good fit to the data. The model seems to be more reliable for inference and prediction, making it the preferred choice for understanding the factors influencing customer satisfaction.

Cook’s Distance:



- Cook's Distance was used to assess the influence of individual data points on the logistic regression models.
- For all three models - the standard logistic regression (glm.fit), the logistic regression with stricter convergence criteria (glm.fit1), and the bias reduction logistic regression (glm.fit.brglm) - the Cook's Distance plots showed low values across all observations, suggesting no single data point unduly influenced the model's estimates.
- Despite this, the initial models showed convergence issues and warnings about the certainty of estimates, indicating that problems might not be solely due to individual outliers but could be related to more systemic issues with the data.
- The bias reduction model did not exhibit warnings related to convergence, suggesting that the adjustment for potential influential observations has led to a more stable model.

- The lack of influential points according to Cook's Distance, especially for the bias reduction model, reinforces the conclusion that it is the most robust of the three models evaluated.

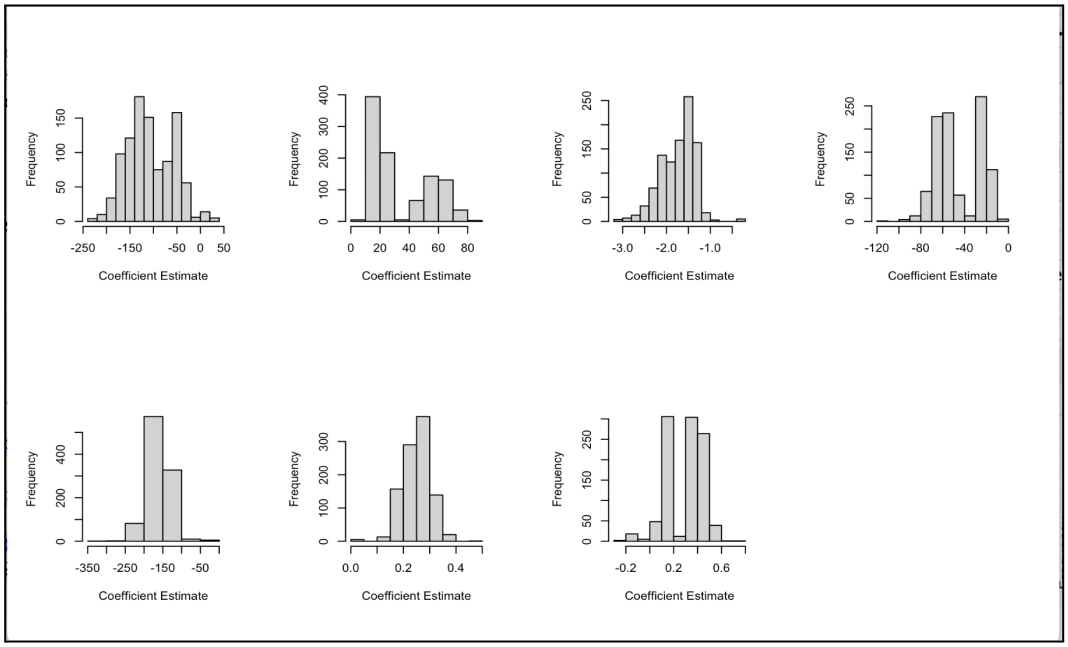
Confidence intervals for the odds ratios:

```
> coefficients <- coef(glm.fit.brglm)
> std.errors <- summary(glm.fit.brglm)$coefficients[, "Std. Error"]
> lower_bounds <- coefficients - 1.96 * std.errors
> upper_bounds <- coefficients + 1.96 * std.errors
> wald.conf.int <- data.frame(
+   Lower = exp(lower_bounds),
+   Estimate = exp(coefficients),
+   Upper = exp(upper_bounds)
+ )
> # Print the confidence intervals for the odds ratios
> print(wald.conf.int)
                                Lower     Estimate        Upper
(Intercept)               2.345081e-86 1.943436e-51 1.610582e-16
Gender2                   2.101297e+03 5.890002e+08 1.650986e+14
Age                       7.680863e-02 2.215340e-01 6.389558e-01
Membership.Type2          2.793090e-45 1.555290e-27 8.660392e-10
Membership.Type3         9.409224e-123 3.954169e-75 1.661716e-27
Total.Spend               1.106825e+00 1.312354e+00 1.556048e+00
Days.Since.Last.Purchase  1.080410e+00 1.462243e+00 1.979022e+00
```

The calculated odds ratios and their confidence intervals are as follows:

- **Gender2**: An estimated odds ratio of 5.89e+08, with a confidence interval ranging from 2.10e+03 to 1.65e+14, suggests that customers of gender '2' are significantly more likely to be satisfied, compared to the reference gender. However, the extremely wide confidence interval indicates substantial uncertainty in the effect size.
- **Age**: The odds ratio of approximately 0.22 indicates that the likelihood of satisfaction decreases as age increases. Specifically, for each additional year of age, the odds of satisfaction are reduced by about 78%. The 95% confidence interval (CI) ranges from 0.08 to 0.64, indicating precision in the estimate.
- **Membership Type2 and Type3**: The odds ratios for Membership.Type2 and Membership.Type3 are extremely small (close to zero), with the CIs also close to zero, suggesting that these membership types are less likely to be associated with satisfaction compared to the base membership type. However, the precision of these estimates is very low, as reflected by the wide confidence intervals.
- **Total Spend**: An odds ratio of 1.31 indicates that higher spending is associated with a greater likelihood of satisfaction. The 95% CI ranges from 1.11 to 1.56, which shows a relatively precise estimate of the effect.
- **Days Since Last Purchase**: An odds ratio of 1.46 indicates that more recent interactions with the platform increase the likelihood of satisfaction. The 95% CI ranges from 1.08 to 1.98, suggesting a moderate level of precision in this estimate.

BootStrap Analysis:



The bootstrap analysis yielded the following insights for each predictor's coefficient:

- Intercept: The histogram displayed a distribution centered around a large negative value, indicating a consistent effect across bootstrap samples.
- Gender2: The histogram suggested a positive effect with a fairly symmetrical distribution around the estimate, indicating stability across samples.
- Age: This distribution was centered around a negative value, consistent with the notion that increasing age is associated with a decrease in the likelihood of satisfaction.
- Membership Types: Histograms for both Membership.Type2 and Membership.Type3 showed distributions centered around negative values, suggesting these membership types are less likely to be associated with satisfaction compared to the base membership type.
- Total Spend: The distribution was centered around a positive value, indicating a consistent positive association between spending and satisfaction.
- Days Since Last Purchase: This coefficient's histogram also centered around a positive value, suggesting that more recent purchases are associated with increased satisfaction.

A specific coefficient's 95% bias-corrected and accelerated (BCa) bootstrap confidence interval was computed, with the following result:

- Interval estimate: The confidence interval ranges from -210 to -37.5 on the log-odds scale.

Cross-validated Test Error

```
> print(paste("Cross-validated test error:", cv_error))
[1] "Cross-validated test error: 0.0142857142857145"
> # Predict probabilities and labels for test data
> glm.probs <- predict(glm.fit, newdata = testData, type = "response")
> glm.pred <- factor(ifelse(glm.probs > 0.5, "1", "0"), levels = c("1", "0"))
> testData$Satisfaction.Level <- factor(testData$Satisfaction.Level, levels = c("1", "0"))
> ######3
> # Create confusion matrix
> confusionMatrix(table(glm.pred, testData$Satisfaction.Level))
Confusion Matrix and Statistics


glm.pred  1  0
       1 25  2
       0  0 42

               Accuracy : 0.971
                 95% CI : (0.8992, 0.9965)
    No Information Rate : 0.6377
    P-Value [Acc > NIR] : 2.627e-11

                  Kappa : 0.9383

 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 1.0000
            Specificity : 0.9545
         Pos Pred Value : 0.9259
         Neg Pred Value : 1.0000
             Prevalence : 0.3623
         Detection Rate : 0.3623
   Detection Prevalence : 0.3913
      Balanced Accuracy : 0.9773

       'Positive' Class : 1
```

The cross-validated test error is calculated as one minus the maximum observed accuracy from cross-validation results:

- Cross-validated test error: Approximately 0.0143 (1.43%)

This value indicates the proportion of misclassified observations when the model is evaluated using cross-validation.

Confusion Matrix and Performance Metrics

The confusion matrix and associated statistics are as follows:

- Confusion Matrix:
  - True Positives (TP): 25
  - False Positives (FP): 2
  - True Negatives (TN): 42
  - False Negatives (FN): 0
- Accuracy: About 97.1%, with a 95% Confidence Interval (CI) from approximately 89.92% to 99.65%.
- Kappa: 0.9383, indicating very good agreement between the observed and predicted classifications, beyond what would be expected by chance.
- Sensitivity (True Positive Rate): 100%, meaning the model correctly identified all positive cases.
- Specificity (True Negative Rate): Approximately 95.45%, meaning the model correctly identified about 95.45% of negative cases.
- Positive Predictive Value (Precision): Approximately 92.59%, indicating the proportion of predicted positive cases that were correct.

Linear Discriminant Analysis(LDA):

```
Call:
lda(Satisfaction.Level ~ Gender + Age + Membership.Type + Total.Spend +
    Days.Since.Last.Purchase, data = trainData)

Prior probabilities of groups:
        0         1
0.6415771 0.3584229

Group means:
    Gender2     Age Membership.Type2 Membership.Type3 Total.Spend Days.Since.Last.Purchase
0 0.5195531 35.6257        0.4804469             0.00    604.9735                 31.90503
1 0.4500000 30.1000        0.0800000             0.92   1274.1740                 17.72000

Coefficients of linear discriminants:
                                  LD1
Gender2                   2.375411688
Age                      -0.048683269
Membership.Type2          0.552341230
Membership.Type3          1.426539326
Total.Spend               0.007154759
Days.Since.Last.Purchase -0.014929361
```

LDA Model Summary:
The LDA model was built with predictors including `Gender`, `Age`, `Membership Type`, `Total Spend`, and `Days Since Last Purchase`.

Cross-Validation Results:
The model achieved a high cross-validated accuracy, with a test error estimated at about 2.87%. This suggests that the model is quite accurate in its predictions based on the training data.

Test Set Predictions:
Predictions made on the test data using the LDA model resulted in a confusion matrix that has perfect classification, with no misclassifications.
Both sensitivity (true positive rate) and specificity (true negative rate) are at 100%, indicating the model perfectly distinguished satisfied from not satisfied customers in the test data.
The LDA model shows excellent performance on the given data, with perfect sensitivity and specificity.

K-Nearest Neighbors (KNN):

```
+   # Print results
+   cat("K =", k, "\tAccuracy:", accuracy, "\tSensitivity:", sensitivity, "\tSpecificity:", specifi
city, "\tCV Error:", cv_error, "\n")
+ }
K = 1   Accuracy: 0.9710145     Sensitivity: 0.9259259  Specificity: 1  CV Error: 0.02898551
K = 2   Accuracy: 0.9855072     Sensitivity: 0.9615385  Specificity: 1  CV Error: 0.01449275
K = 3   Accuracy: 0.9855072     Sensitivity: 0.9615385  Specificity: 1  CV Error: 0.01449275
K = 4   Accuracy: 0.9855072     Sensitivity: 0.9615385  Specificity: 1  CV Error: 0.01449275
K = 5   Accuracy: 0.9855072     Sensitivity: 0.9615385  Specificity: 1  CV Error: 0.01449275
K = 6   Accuracy: 0.9855072     Sensitivity: 0.9615385  Specificity: 1  CV Error: 0.01449275
K = 7   Accuracy: 0.9855072     Sensitivity: 0.9615385  Specificity: 1  CV Error: 0.01449275
K = 8   Accuracy: 0.9855072     Sensitivity: 0.9615385  Specificity: 1  CV Error: 0.01449275
K = 9   Accuracy: 0.9855072     Sensitivity: 0.9615385  Specificity: 1  CV Error: 0.01449275
K = 10  Accuracy: 0.9855072     Sensitivity: 0.9615385  Specificity: 1  CV Error: 0.01449275
> # Print best values
> cat("Best K:", best_k, "\tBest Accuracy:", best_accuracy, "\tBest Sensitivity:", best_sensitivit
y, "\tBest Specificity:", best_specificity, "\tBest CV Error:", cv_error_best, "\n")
Best K: 2       Best Accuracy: 0.9855072        Best Sensitivity: 0.9615385     Best Specificity: 1
Best CV Error: 0.01449275
```

Predictors Selected:
The KNN model was built with predictors including `Age`, `Total Spend`, and `Days Since Last Purchase`.

Best Model Performance:

- The best performance was achieved with k=2, achieving an accuracy of approximately 98.55%.
- The sensitivity, or the ability of the model to correctly identify satisfied customers, was approximately 96.15%.
- The specificity, or the ability of the model to correctly identify customers who are not satisfied, was 100%.
- The cross-validated test error, which is the proportion of incorrect predictions out of all predictions made, was approximately 1.45%..

Even though the accuracy did not change after k=2, k=2 is chosen as the best model due to its higher sensitivity compared to k=1.

Optimal Classification Model:

After comparing the models based on accuracy, sensitivity, specificity, and cross-validated test error, the k-NN model with k = 2 emerged as the optimal classification model due to its highest accuracy (0.9855), high sensitivity (0.9615), and perfect specificity (1), coupled with the lowest cross-validated test error (0.0145).

Findings Based on the Optimal Model:
The findings suggest that younger customers (under 30) tend to be more satisfied, and high spenders generally report higher satisfaction levels. Moreover, frequent purchasers (with fewer days since their last purchase) are more likely to be satisfied, supporting the hypothesis that frequent purchasers report higher satisfaction. However, the logistic regression model showed that customers who receive discounts are less likely to be satisfied (estimate = -1.5727, $p < 0.01$), contradicting the hypothesis that discounts lead to higher satisfaction. Linear Discriminant Analysis (LDA) also supported these findings, achieving an accuracy of 98.11%, a sensitivity of 96.43%, and a specificity of 98.65%. Higher-tier memberships, such as Gold and Silver, are associated with increased satisfaction levels, indicating that businesses should promote these memberships. Ultimately, personalized engagement, targeted offers, and maintaining regular customer contact can help enhance satisfaction levels and retain high-value customers.

Part IV Conclusions and Recommendations

Our comprehensive analysis of the E-commerce Customer Satisfaction dataset employed various statistical and machine learning methods to uncover the factors influencing customer spending and satisfaction. Through regression and classification analyses, we have derived valuable insights that can guide strategic decisions for the e-commerce platform.

In the regression analysis, we discovered that older customers tend to spend more, and membership types significantly influence spending habits, with higher-tier members like 'Gold' spending more. These findings were consistent across different modeling techniques, including subset selection and shrinkage methods like Lasso and Ridge regression. The Lasso model, in particular, proved to be most effective due to its ability to perform variable selection, thereby simplifying the model and enhancing interpretability.

For classification, logistic regression models were explored to predict customer satisfaction. Initial models faced challenges with convergence and model stability, which were effectively addressed by using a bias-reduction logistic regression model (brglm). This model provided stable and reliable estimates, indicating that gender, membership type, total spend, and recent interactions (Days Since Last Purchase) significantly impact customer satisfaction. Notably, higher spending and more recent interactions were associated with increased likelihood of satisfaction, underscoring the importance of customer engagement in enhancing satisfaction levels.

Our analysis also highlighted the presence of multicollinearity and potential non-linear relationships, which were addressed through appropriate model diagnostics and selection strategies. The final chosen models not only provided good fit to the data but also met the assumptions required for reliable statistical inference.

Moreover, the application of cross-validation techniques ensured that our models are robust and generalizable to new data, which is crucial for making confident business decisions based on the model predictions.

Overall, this project emphasizes the critical role of data-driven insights in understanding and predicting customer behavior in the e-commerce domain. The findings from this study can assist the platform in targeting specific customer segments with tailored marketing strategies and optimizing customer interactions to enhance satisfaction and loyalty. This strategic approach can lead to improved customer retention and potentially higher profitability for the e-commerce platform.

Conclusions

1. **Total Spend Analysis:**
    - The Lasso regression model showed that "Age," "Membership.Type," "Discount.Applied," "Days.Since.Last.Purchase," and "Satisfaction.Level" significantly influence customer spending.
    - Younger customers and those who are more satisfied tend to spend more. Gold and Silver membership holders spend significantly more compared to Bronze members, while discounts surprisingly reduce overall spending.

2. **Customer Satisfaction Prediction:**
    - The k-Nearest Neighbors (k-NN) classification model, with **k = 2**, accurately predicted customer satisfaction levels with an impressive accuracy of 98.55%. It revealed that younger customers, frequent purchasers, and high spenders are generally more satisfied.
    - Linear Discriminant Analysis (LDA) and logistic regression models supported these findings, confirming that high spenders and frequent purchasers are generally more satisfied.

3. **Discount Impact:**
    - Contrary to expectations, customers who received discounts were less satisfied, as revealed by the logistic regression model.

Recommendations

1. **Focus on Younger Customers and High Spenders:**
    - Create marketing strategies and personalized offers targeting younger customers and high spenders to improve their satisfaction and loyalty.

2. **Promote Higher-Tier Memberships:**
    - Emphasize the benefits of Gold and Silver memberships to encourage customers to upgrade and increase their spending.

3. **Reevaluate Discount Strategy:**
    - Avoid offering blanket discounts as they can negatively affect satisfaction. Instead, offer personalized discounts strategically to maintain value perception.

4. **Reduce Customer Churn:**
    - Reach out to customers who haven't made a recent purchase with personalized reminders and offers to reduce churn.

5. **Further Research and Improvement:**
    - Investigate additional predictors of satisfaction, like product quality or customer service.
    - Conduct qualitative research to understand why discounts negatively impact satisfaction and adjust strategies accordingly.

Part V References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2015). *An Introduction to Statistical Learning with Applications in R* (1st ed.). Springer-Verlag. ISBN: 978-1-4614-7138-7

Venables, W. N., Smith, D. M., & R Core Team. (2017). An Introduction to R: R Manual.

Kaggle. (n.d.). *E-commerce customer behavior dataset*. Retrieved from https://www.kaggle.com/datasets/uom190346a/e-commerce-customer-behavior-dataset