PROJECT

HARINI SARAVANAN

BU ID: U33572460

MET CS:555- Foundations of Machine Learning

TITLE: Optimizing E-Commerce Through Data Insights and Analytics

RESEARCH SCENARIO:

You work as a data analyst for an e-commerce company and have been tasked with identifying key customer behavior patterns and optimizing membership benefits to improve customer satisfaction and revenue.

RESEARCH QUESTIONS:

- What are the distinct customer segments based on spending and behavior?
- How does membership type influence total spend and satisfaction level?
- What factors significantly influence customer satisfaction?

DATASET DESCRIPTION:

Source: https://www.kaggle.com/code/youssefabdelghfar/e-commerce-customer-behavior/input

VARIABLES:

- **Customer_ID**: Unique identifier for each customer.
- **Total_Spend**: Total amount spent by the customer (in dollars).
- **Items_Purchased**: Number of items purchased by the customer.
- **Days_Since_Last_Purchase**: Time since the last purchase (in days).
- **Average_Rating**: Average product rating provided by the customer (1–5 scale).
- **Membership_Type**: Type of membership (Gold, Silver, Bronze).
- **Satisfaction_Level**: Customer satisfaction score (Satisfied, Neutral, Unsatisfied).
- **Age**: Age of the customer.
- **Discount_Applied**: Total discount received by the customer (in dollars).
- **City:** City of the customer
- **Gender:** Gender of the customer (Male, Female)

NOTES

VARIABLES USED IN THE ANALYSIS:

Total Spend, Items Purchased, Days Since Last Purchase, Average Rating and Satisfaction Level

DATA CLEANING PROCEDURES:

- Handled missing values by imputing it with mean.
- Detected the outlier and removed it from the dataset.
- Normalised numeric values for clustering.
- Encoded categorical variables (transforming it to a numeric format for easy analysis).

STATISTICAL METHODS USED:

CUSTOMER SEGMENTATION:
method used: K-Means Clustering

why this method has been chosen?
K-Means is ideal for grouping customers into segments based on similar spending habits and purchase behaviors.

variables used:
Total Spend, Items Purchased, Days Since Last Purchase, and Average Rating

explanation:
This method helps identify distinct customer groups to target them with personalized marketing strategies.

MEMBERSHIP OPTIMIZATION:
method used: ANOVA

why this method has been chosen?
ANOVA is effective for comparing spending patterns and satisfaction levels across different membership types.

variables used:
Total Spend and Satisfaction Analysis

explanation:
This method evaluates if membership type significantly impacts spending or satisfaction to refine membership strategies.

NOTES

SATISFACTION ANALYSIS:
method used: Regression Analysis

why this method has been chosen?
Regression helps uncover the relationship between customer satisfaction and factors like spending, purchases, and ratings.

variables used:
Satisfaction Level, Total Spend, Items Purchased, Days Since Last Purchase, Average Rating

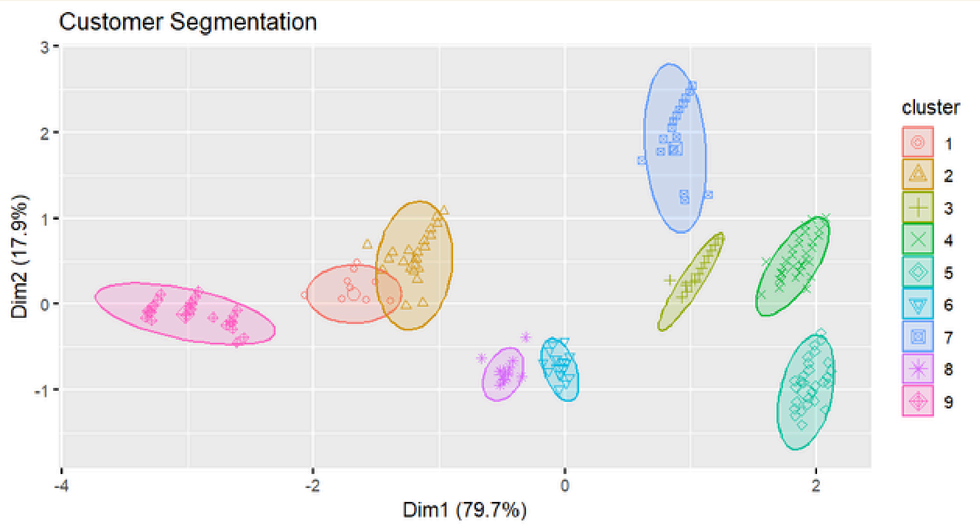
explanation:
This method predicts how various factors influence satisfaction, guiding improvements for better customer experiences.

Regression Type	**Logistic Regression**	**Linear Regression**
Why Chosen	Logistic regression is used when satisfaction is *binary* (e.g., Satisfied or Not Satisfied)	Linear regression is used when satisfaction is measured on a *continuous scale* (e.g., ratings).
Explanation	It identifies which factors increase the *likelihood of a customer being satisfied*, helping target areas for improvement.	It predicts how changes in factors like *spending or purchases impact satisfaction scores*, offering insights for optimization.

NOTES

RESULTS AND CONCLUSION:

CHART 1: CUSTOMER SEGMENTATION



Observation:
This chart shows a segmentation of customers into 9 distinct clusters based on their characteristics and behaviors. The clusters are defined by a combination of variables like *Total Spend, Items Purchased, Days Since Last Purchase, Average Rating*.

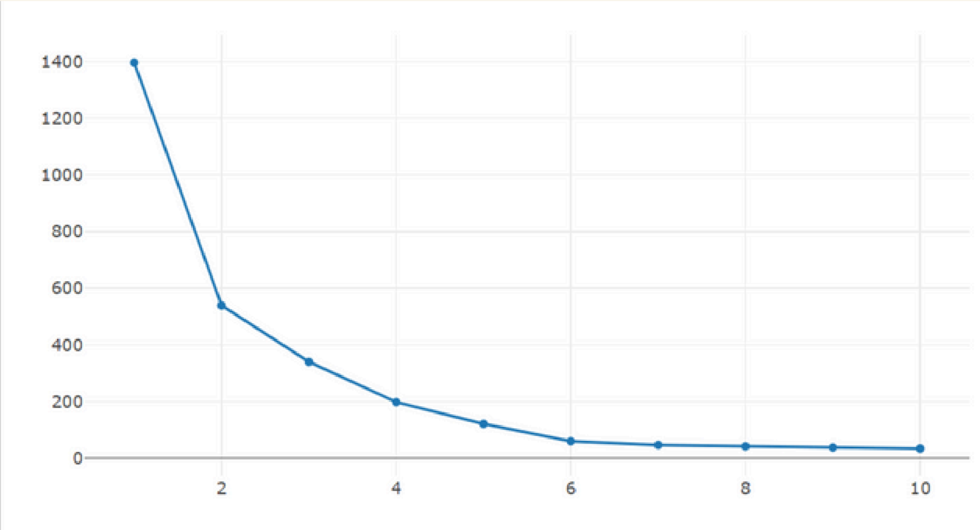
Possible Insights:

- Analyze the specific characteristics of each cluster to identify potential target segments for marketing campaigns.
- Tailor product recommendations and promotions to the preferences of each cluster.
- Use clustering to predict customer behavior and churn risk.

Conclusion:
Customer segmentation can help target marketing efforts. For example, focusing on each segment and approaching them uniquely will result in increased profits and more loyal customers, such as by providing incentives.

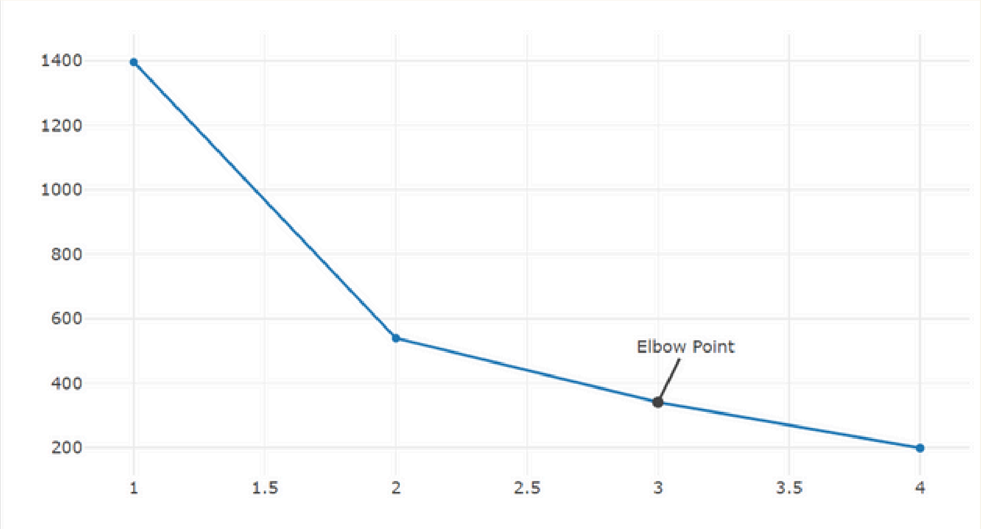
NOTES

CHART 2: ELBOW CURVE



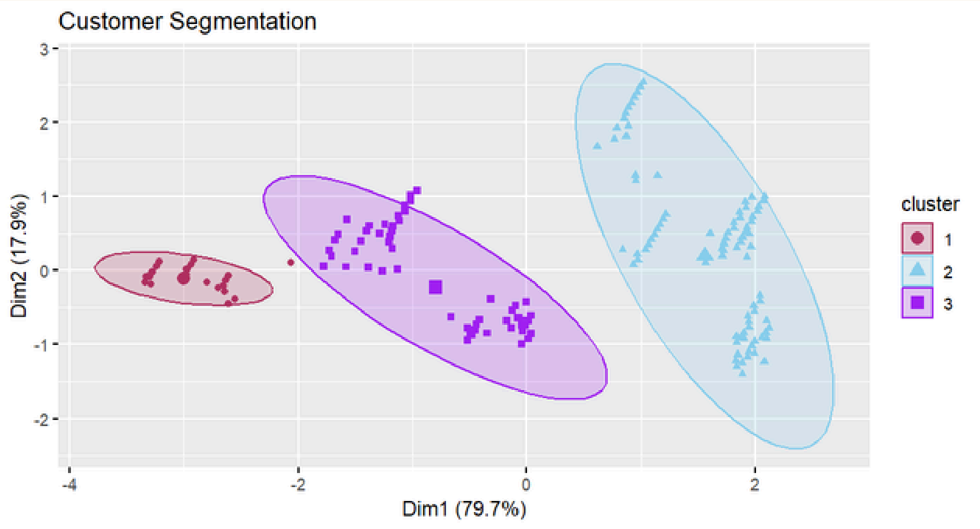
Observation:

- These charts show Elbow Curve, a common technique used to determine the optimal number of clusters in a clustering algorithm, often K-Means.
- In the given chart, the elbow point seems to be around K = 3. This suggests that dividing the data into 3 clusters might be the most optimal choice.
- Therefore, after dividing the clusters into 3, we get the following elbow curve and the cluster chart.



NOTES

CHART 3: UPDATED CUSTOMER SEGMENTATION

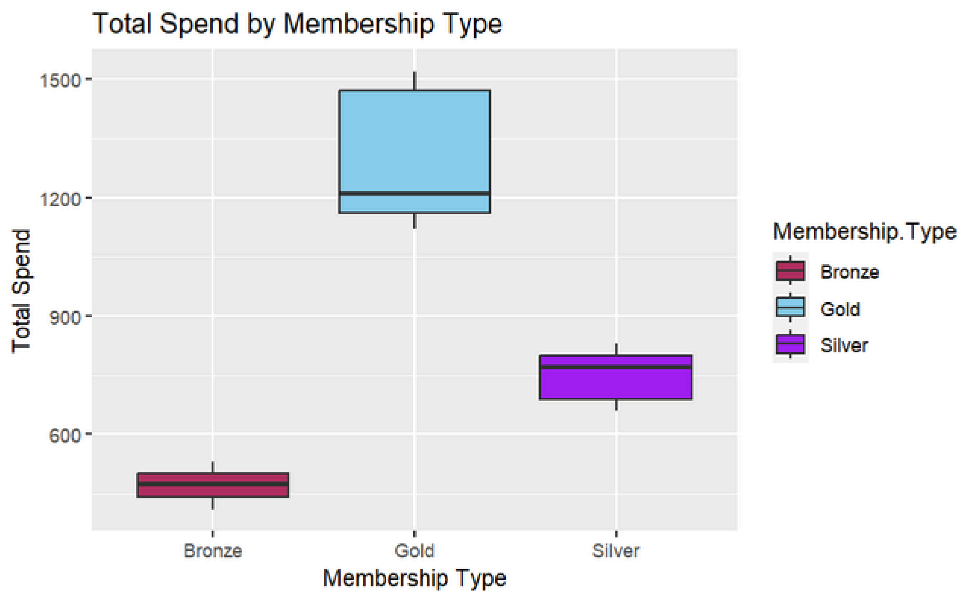


Limitations:

- K-Means is sensitive to outliers, as they can skew the centroid calculation. Outliers can result in the creation of extra clusters or distortion of cluster centroids.
- K-Means requires you to specify the number of clusters (K) in advance. It is difficult to predict the segments before performing any analysis.

NOTES

CHART 4: TOTAL SPEND BY MEMBERSHIP TYPE



Observation:
This chart reveals clear differences in spending patterns across membership types. Gold members, on average, spend significantly more than Silver and Bronze members.

Possible Insights:

- Target marketing strategies and promotions to Bronze and Silver members to encourage increased spending.
- Explore loyalty programs and exclusive offers to retain high-spending Gold members.

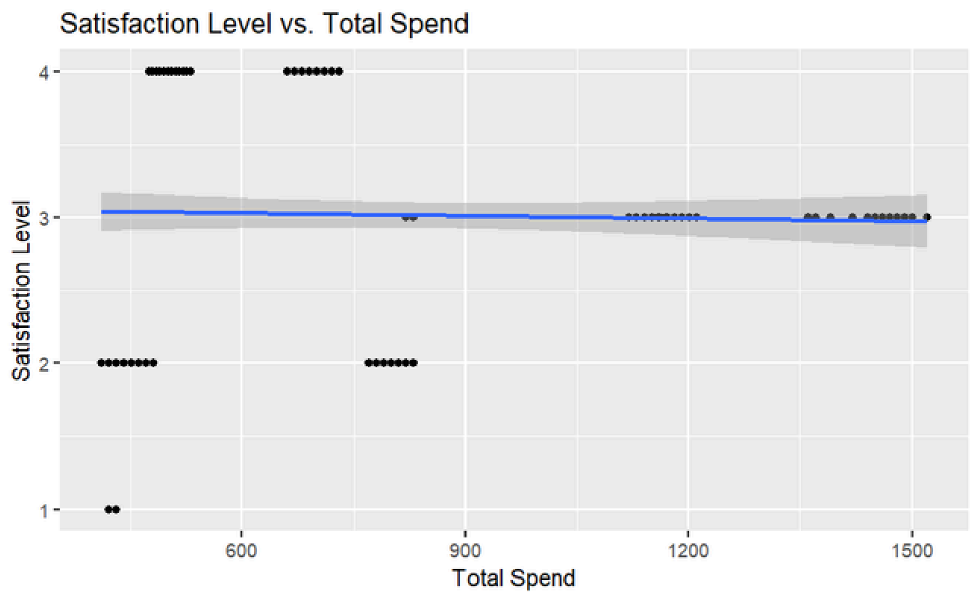
Conclusion:
Optimize VIP memberships to enhance spending further. Focusing on understanding and catering to the distinct needs of each membership type will help create targeted strategies that optimize revenue and customer satisfaction. By identifying spending trends among membership types, businesses can better allocate resources, offering more personalized services to high-value members while fostering growth in lower-spending segments.

Limitations:

- The analysis shows correlations between membership type and spending, but it does not establish a causal relationship, which limits actionable insights.
- Other factors influencing spending, such as seasonality or external economic conditions, may not have been considered, affecting the robustness of the findings.

NOTES

CHART 5: SATISFACTION LEVEL VS. TOTAL SPEND



Observation:
This chart shows a weak positive correlation between *Total Spend* and *Satisfaction Level*. This means that, generally, as customers spend more, their satisfaction tends to increase slightly. However, the relationship is not very strong, indicating that other factors likely play a significant role in determining satisfaction.

Possible Insights:

- Consider improving customer experience for lower-spending customers to boost their satisfaction.
- Analyze the distribution of discounts and their impact on satisfaction across different spending levels.

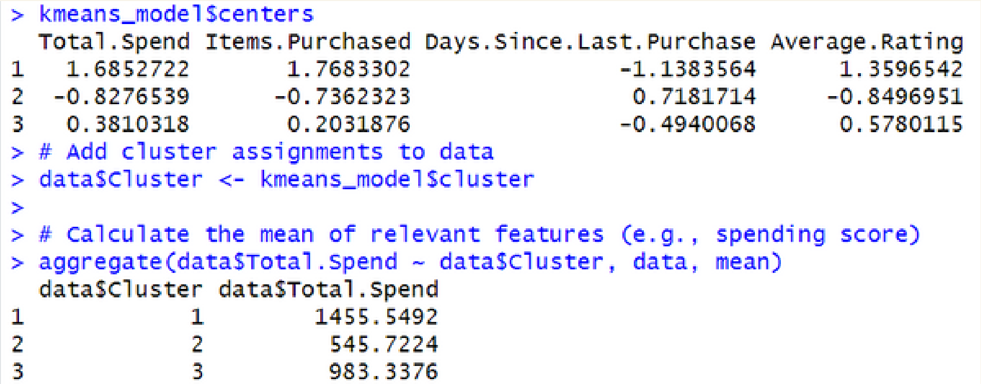
Conclusion:
To improve customer satisfaction, it's important to focus on key factors like offering targeted discounts and personalizing experiences based on age and spending habits. While there is a slight positive correlation between spending and satisfaction, addressing other influential factors can help enhance overall customer satisfaction and loyalty.

Limitations:

- Omitted variables can result in incorrect p-values, making it difficult to determine whether the relationships between predictors and the outcome are statistically significant.

OUTPUT:

What are the distinct customer segments based on spending and behavior?



```
> kmeans_model$centers
  Total.Spend Items.Purchased Days.Since.Last.Purchase Average.Rating
1   1.6852722       1.7683302               -1.1383564      1.3596542
2  -0.8276539      -0.7362323                0.7181714     -0.8496951
3   0.3810318       0.2031876               -0.4940068      0.5780115
> # Add cluster assignments to data
> data$Cluster <- kmeans_model$cluster
>
> # Calculate the mean of relevant features (e.g., spending score)
> aggregate(data$Total.Spend ~ data$Cluster, data, mean)
  data$Cluster data$Total.Spend
1            1        1455.5492
2            2         545.7224
3            3         983.3376
```

HO (Null Hypothesis): There are no distinct customer segments in the dataset based on spending and behavior.
Ha (Alternative Hypothesis): There are distinct customer segments in the dataset based on spending and behavior.

DECISION RULE:
If the clusters show distinct separations with minimal overlaps, reject H_0. Otherwise, fail to reject H_0.

RESULTS:

1. **Optimal Number of Clusters:**

- Using the elbow method, the Within-Cluster Sum of Squares (WSS) decreases significantly at 3 clusters, indicating that 3 is the optimal number of clusters.
- The elbow point was identified with the second derivative approach to minimize WSS.

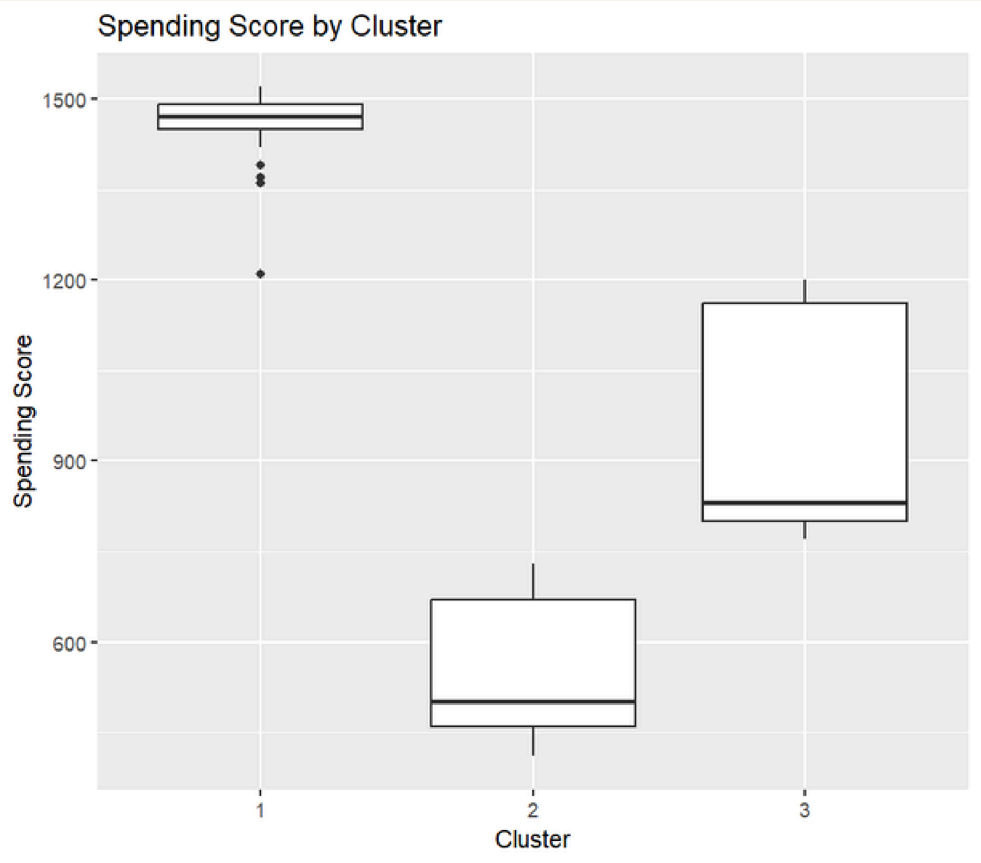
2. **Cluster Visualization:**

- The first visualization (9 clusters) depicts fine-grained segmentation.
- The second visualization (3 clusters) shows broader segmentation of customer behavior.
- Ellipses indicate distinct groups with minimal overlap, suggesting meaningful segmentation.

3. **Cluster Characteristics:**

- Cluster 1 (Red): Customers with low spending and distinct behavior.
- Cluster 2 (Blue): Moderate spenders with average behavior.
- Cluster 3 (Purple): High spenders with premium spending patterns.

NOTES



Observations of the clusters and the WSS strongly indicate segmentation. Hence, **reject H_0**.

INTERPRETATION:

- The data reveals three distinct customer segments: *low spenders, moderate spenders, and high spenders*, characterized by spending patterns and behaviors.
- This segmentation is unlikely to be due to random chance.
- These insights can guide targeted marketing strategies:

-->Focus on Cluster 3 (High Spenders) for premium services and loyalty programs.

-->Design campaigns for Cluster 2 (Moderate Spenders) to boost engagement.

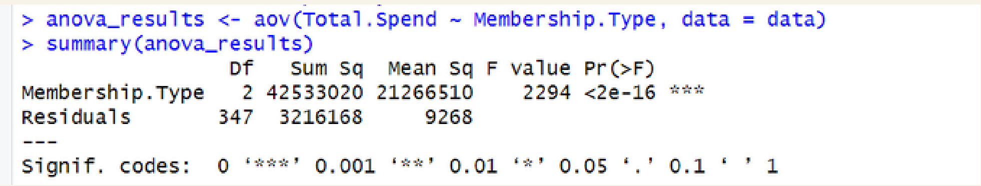
-->Offer entry-level incentives to Cluster 1 (Low Spenders) to enhance participation.

NOTES

How does membership type influence total spend and satisfaction level?

H0 (Null Hypothesis): Membership does not have a significant effect on Total Spending.
Ha (Alternative Hypothesis): Membership does have a significant effect on Total Spending.

DECISION RULE:
Reject H0: if $p<0.05$ else There is no significant evidence to reject H0.



```
> anova_results <- aov(Total.Spend ~ Membership.Type, data = data)
> summary(anova_results)
                 Df   Sum Sq  Mean Sq F value Pr(>F)    
Membership.Type   2 42533020 21266510    2294 <2e-16 ***
Residuals       347  3216168     9268                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

RESULTS:
As we can see, **F value= 2294** which is very high suggesting a strong effect of membership type on total spending.

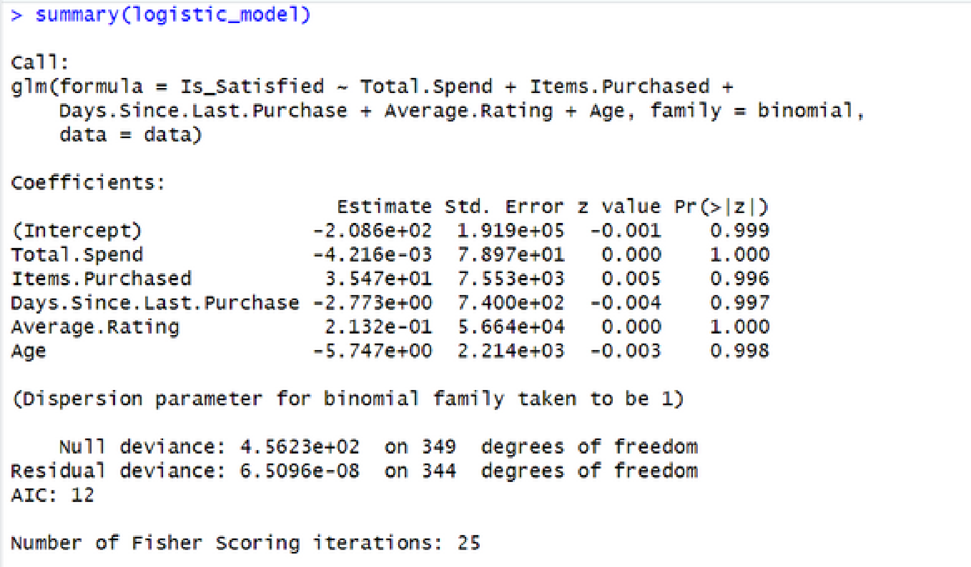
p<2e−16 means the result is extremely significant. Membership type has a statistically significant effect on Total Spend.

INTERPRETATION:
Membership type significantly influences total spending ($p < 0.001$). Gold, Silver, and Bronze members spend differently, and this difference is unlikely to be due to random chance.

NOTES

What factors significantly influence customer satisfaction?

LOGISTIC MODEL



```
> summary(logistic_model)

Call:
glm(formula = Is_Satisfied ~ Total.Spend + Items.Purchased +
    Days.Since.Last.Purchase + Average.Rating + Age, family = binomial,
    data = data)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.086e+02  1.919e+05  -0.001    0.999
Total.Spend              -4.216e-03  7.897e+01   0.000    1.000
Items.Purchased           3.547e+01  7.553e+03   0.005    0.996
Days.Since.Last.Purchase -2.773e+00  7.400e+02  -0.004    0.997
Average.Rating            2.132e-01  5.664e+04   0.000    1.000
Age                      -5.747e+00  2.214e+03  -0.003    0.998

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4.5623e+02  on 349  degrees of freedom
Residual deviance: 6.5096e-08  on 344  degrees of freedom
AIC: 12

Number of Fisher Scoring iterations: 25
```

H0 (Null Hypothesis): None of the predictors (*Total Spend, Items Purchased, Days Since Last Purchase, Average Rating, Age*) have a significant effect on the likelihood of a customer being satisfied.

Ha (Alternative Hypothesis): At least one predictor has a significant effect on the likelihood of a customer being satisfied.

DECISION RULE:
Reject H0: if the p-value of a predictor is < 0.05. Otherwise, there is no significant evidence to reject H0.

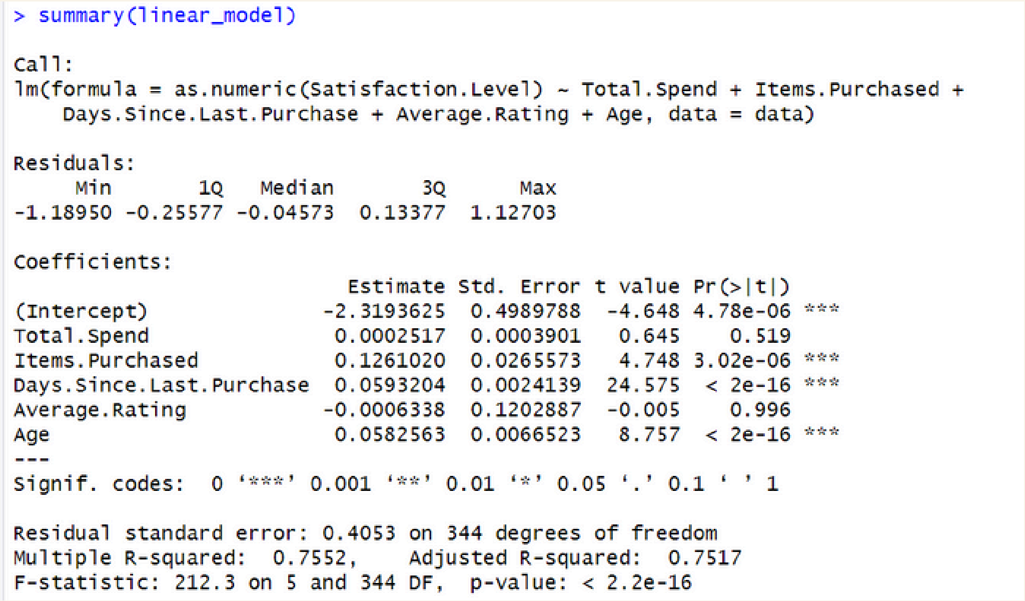
RESULTS:

- None of the predictors in the model are significant (p>0.05 for all predictors).
- The logistic regression model fails to provide meaningful results, as the coefficients do not have a significant impact on the likelihood of customer satisfaction.

Interpretation:
The logistic regression model does not explain customer satisfaction well. Revisit data preprocessing (e.g., scaling, removing multicollinearity) or consider alternative models or predictors to improve the fit.

NOTES

LINEAR MODEL



```
> summary(linear_model)

Call:
lm(formula = as.numeric(Satisfaction.Level) ~ Total.Spend + Items.Purchased +
    Days.Since.Last.Purchase + Average.Rating + Age, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.18950 -0.25577 -0.04573  0.13377  1.12703

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -2.3193625  0.4989788  -4.648 4.78e-06 ***
Total.Spend               0.0002517  0.0003901   0.645    0.519
Items.Purchased           0.1261020  0.0265573   4.748 3.02e-06 ***
Days.Since.Last.Purchase  0.0593204  0.0024139  24.575  < 2e-16 ***
Average.Rating           -0.0006338  0.1202887  -0.005    0.996
Age                       0.0582563  0.0066523   8.757  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4053 on 344 degrees of freedom
Multiple R-squared:  0.7552,    Adjusted R-squared:  0.7517
F-statistic: 212.3 on 5 and 344 DF,  p-value: < 2.2e-16
```

H0 (Null Hypothesis): None of the predictors (*Total Spend, Items Purchased, Days Since Last Purchase, Average Rating, Age*) significantly affect the Satisfaction Level.

Ha (Alternative Hypothesis): At least one predictor has a significant effect on the Satisfaction Level.

DECISION RULE:
Reject H0: if the p-value of the overall model (F-statistic) or a specific predictor is < 0.05. Otherwise, there is no significant evidence to reject H0.

RESULTS:
As we can see, **F value= 212.3** which is very high indicating that the predictors collectively explain satisfaction well.

p<2.2e−16 means the result is extremely significant. At least one predictor has a significant effect on the Satisfaction Level.

Interpretation:
The linear regression model explains 75.5% of the variation in satisfaction levels (R^2=0.7552).

Significant predictors like *Items Purchased, Days Since Last Purchase, and Age* provide actionable insights, while *Total Spend and Average Rating* do not seem to influence satisfaction directly.

NOTES

SIGNIFICANT vs NON-SIGNIFICANT PREDICTORS

Significant Predictors (p<0.05)	**p value**	**Result**
Items Purchased	3.02e−06	Satisfaction increases significantly with the number of items purchased.
Days Since Last Purchase	<2e−16	Satisfaction increases with more time since the last purchase.
Age	<2e−16	Older customers are more satisfied.

Non-Significant Predictors (p>0.05)	**p value**	**Result**
Total Spend	0.519	Spending does not significantly impact satisfaction.
Average Rating	0.996	Ratings have no measurable effect on satisfaction.