**E-commerce Customer Device Usage Analysis**

Probability and Statistics Project Report

**Submitted By:**

Patel Swapnilkumar Chandubhai

22BAI1308

**Abstract**

This project report documents an analysis of the E-commerce Customer Device Usage dataset. The goal is to understand relationships between various customer behavioural metrics, such as average session length and time spent on different platforms, and their effect on the total yearly spending. The project leverages exploratory data analysis, correlation analysis, and multiple linear regression models to quantify the impact of these variables on customer expenditure. The model is evaluated using RMSE, MAPE, and R2R^2R2 values, which demonstrate the strength of the model in predicting yearly spending based on user behaviour data.

**Introduction**

Understanding customer behaviour is crucial for e-commerce businesses to optimize user experience and increase spending. By analysing factors such as time on website, time on app, and length of membership, this study seeks to identify key determinants of annual customer spending. The insights derived from this model can guide e-commerce platforms in strategizing personalized user engagement approaches.

**Problem Statement**

As e-commerce becomes increasingly competitive, businesses need data-driven insights into what drives customer spending. Identifying patterns in user behaviour, such as how much time users spend on various devices or platforms, could inform better engagement strategies and product recommendations. This project aims to develop a linear regression model that accurately predicts yearly spending based on metrics like session length, platform usage, and membership duration.

**Methodology**

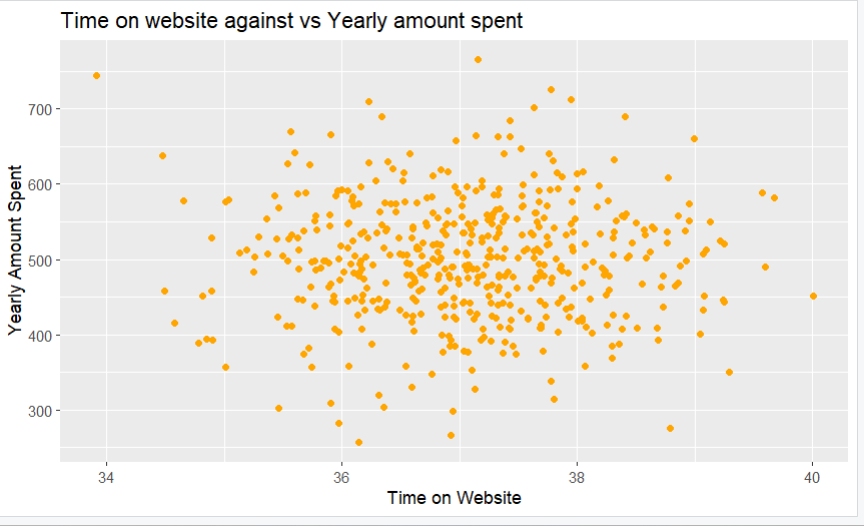
**Dataset**

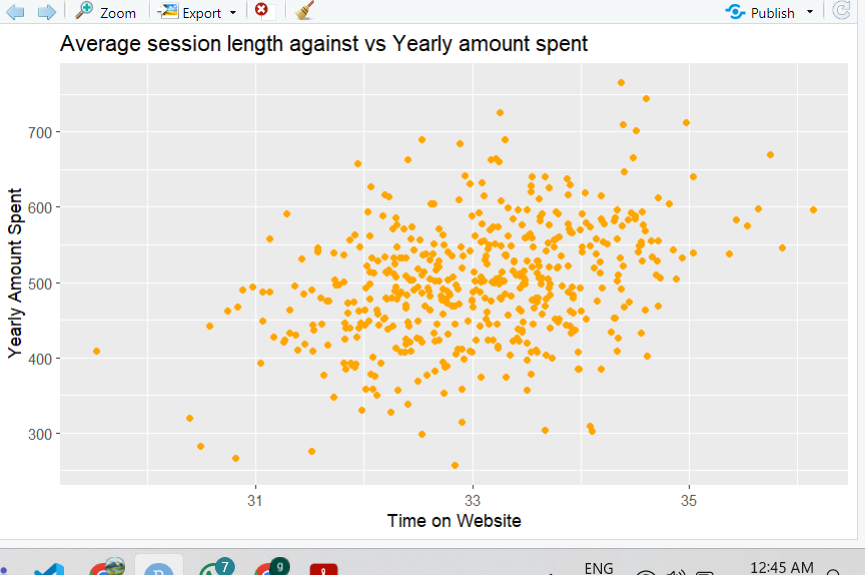
The dataset contains several metrics related to customer behaviour on an e-commerce platform. Key attributes include:

* **Avg. Session Length**: Average time spent per session.
* **Time on App**: Total time spent on the app.
* **Time on Website**: Total time spent on the website.
* **Length of Membership**: How long the user has been a member.
* **Yearly Amount Spent**: The total yearly amount spent by the customer (target variable).

**Correlation Analysis**

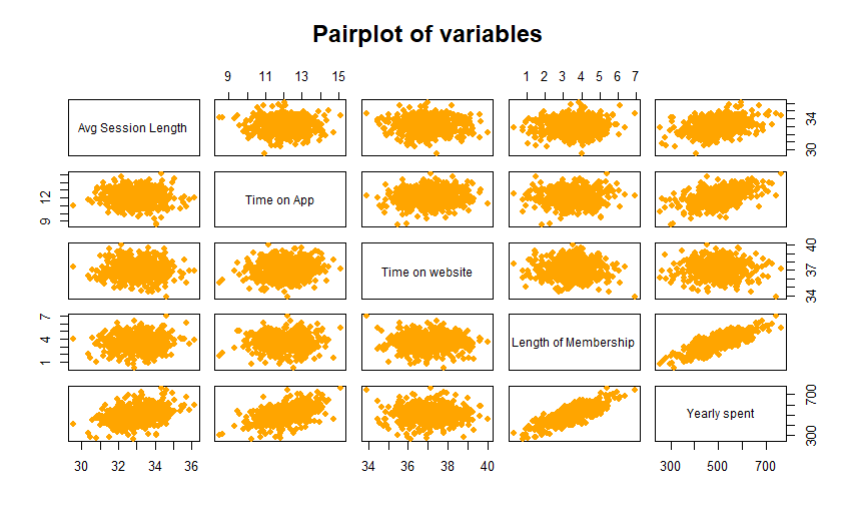
A preliminary correlation analysis was conducted to identify relationships between the features and the target variable, Yearly Amount Spent. Scatter plots revealed that **Length of Membership** shows a strong positive correlation with yearly spending, while **Time on Website** has a relatively low correlation. This insight guided the selection of features for the regression model.



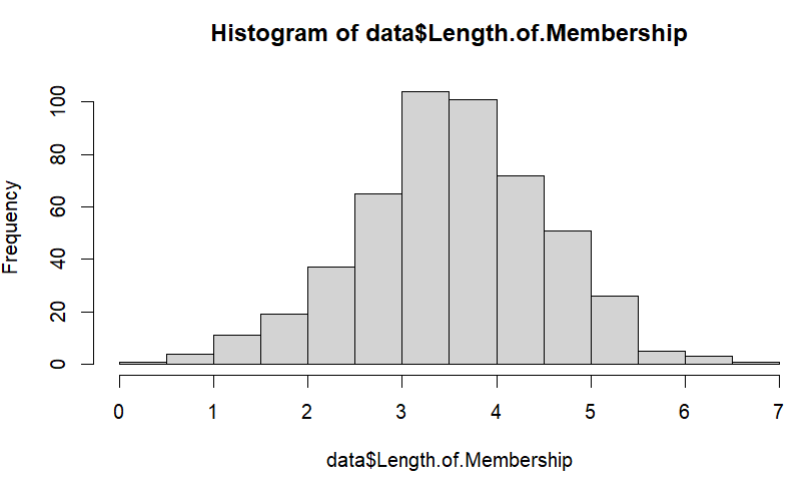


**Exploratory Data Analysis (EDA)**

Visualizations such as scatter plots, histograms, and boxplots were created to examine the distribution of the data and assess normality. These plots helped to confirm that **Length of Membership** is normally distributed and suitable for linear modelling.



*Fig 3:*The pairplot reveals that **Length of Membership is the most significant predic**tor of Yearly Amount Spent. Other factors like Time on App and Time on Website have a weaker influence.



*Fig 4:* The histogram shows the distribution of the "Length of Membership" variable. Here's what we can conclude from it:

**Shape:**

* The distribution appears to be **right-skewed**. This means that there are more customers with shorter membership lengths compared to longer ones.

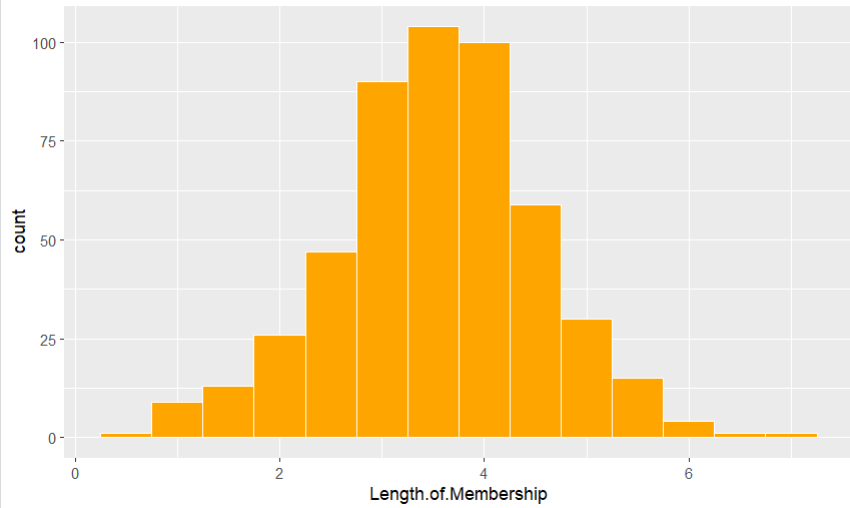
**Central Tendency:**

* The majority of customers seem to have a membership length between 3 and 4 years.

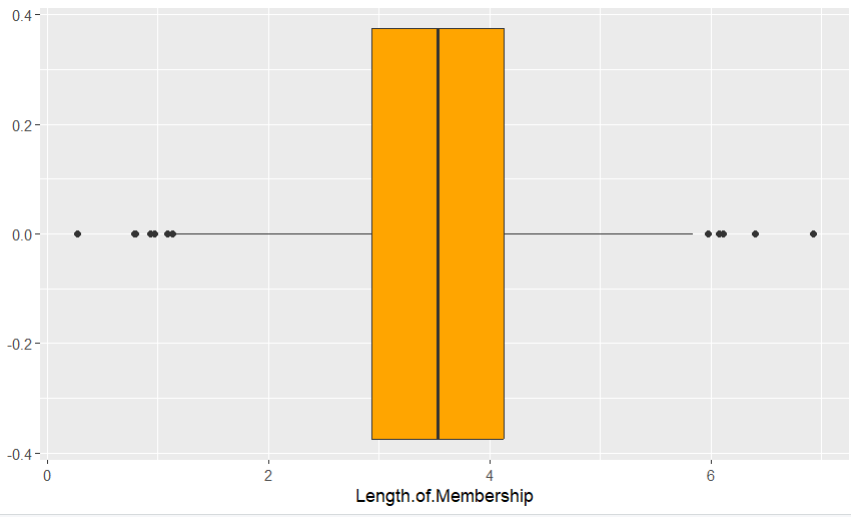
**Spread:**

* The distribution is relatively wide, indicating that there is a significant variation in membership lengths among customers.

*Fig 5:* Histogram Plot for Length of Membership Vs Count of Customers



*Fig 6:* The boxplot below provides a visual summary of the distribution of the "Length of Membership" variable. Here's what we can conclude from it:



**Central Tendency:**

* The median membership length is around 4 years, as indicated by the line within the box.

**Spread:**

* The box shows the interquartile range (IQR), which represents the middle 50% of the data. In this case, the IQR appears to be relatively narrow, suggesting that the majority of customers have membership lengths close to the median.

**Outliers:**

* The dots outside the box represent potential outliers. These are data points that are significantly different from the rest of the data. In this case, there seem to be some customers with much longer membership lengths.

**Overall, the boxplot indicates that most customers have membership lengths around 4 years, with a few outliers having significantly longer memberships.**

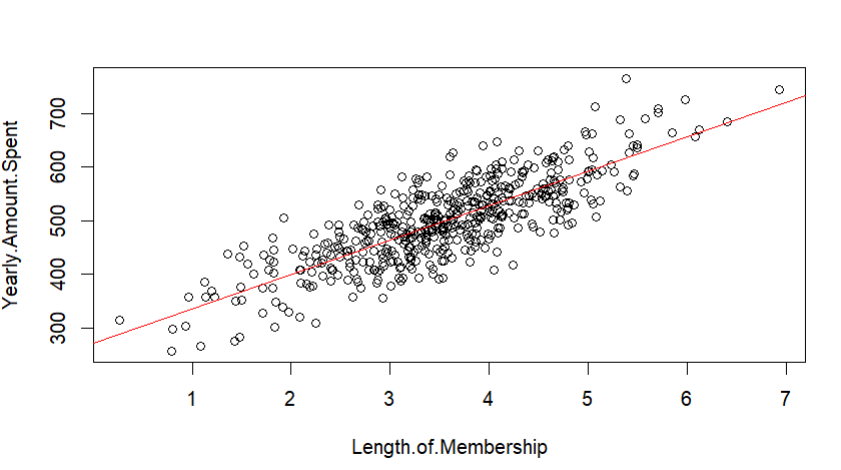
**Prediction Model**

**Linear Regression with Single Predictor**

A simple linear regression model was initially fitted using Length of Membership as the sole predictor. The model's performance metrics indicate that membership duration is a significant predictor of yearly spending, with a p-value below the significance level. The slope (β₁) of approximately 64.219 implies that each additional year of membership increases spending by around $64.21.

* **Residual Analysis**: A QQ plot of residuals showed a normal distribution, satisfying a key assumption for linear regression.

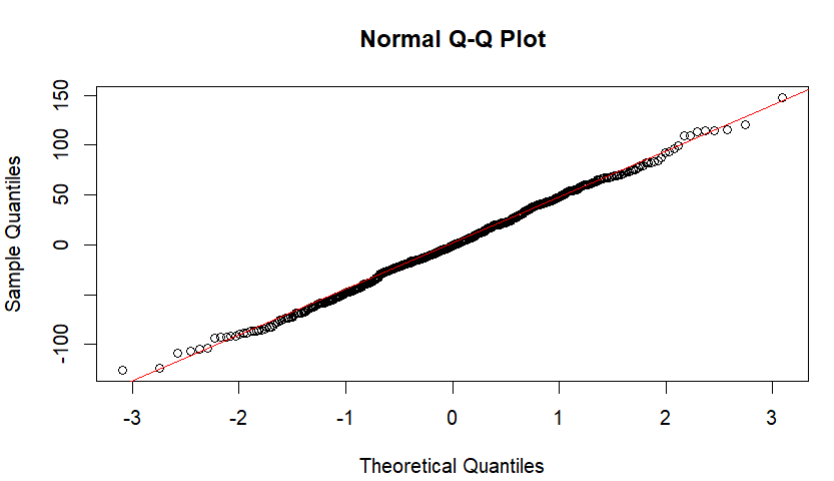
*Fig 7:* The scatter plot with the regression line shows a positive linear relationship between "Length of Membership" and "Yearly Amount Spent." This means that as the length of membership increases, the yearly amount spent also tends to increase.



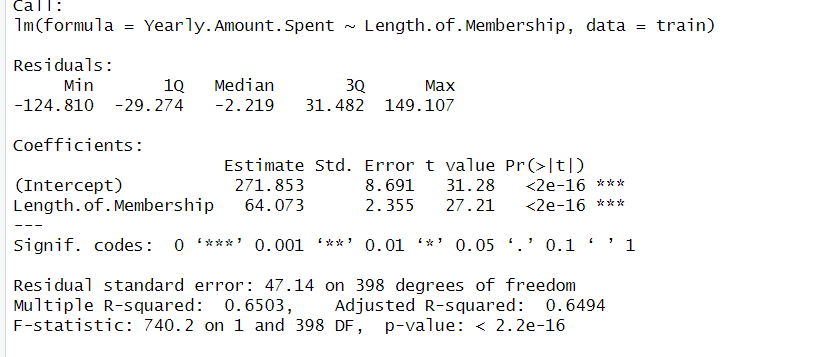
The **red line represents the best-fit line** that minimizes the distance between the data points and the line. This line can be used to predict the yearly amount spent for a given length of membership.

However, it's important to note that there is some variability in the data points around the line, indicating that length of membership is not the only factor influencing yearly spending. Other factors might also play a role.

*Fig 8:* The Q-Q plot suggests that the residuals are approximately normally distributed. This is an important assumption for linear regression as it ensures the validity of hypothesis testing and confidence interval estimation.



*Fig 9:* Showing the Evaluation Matrix for Linear Regression



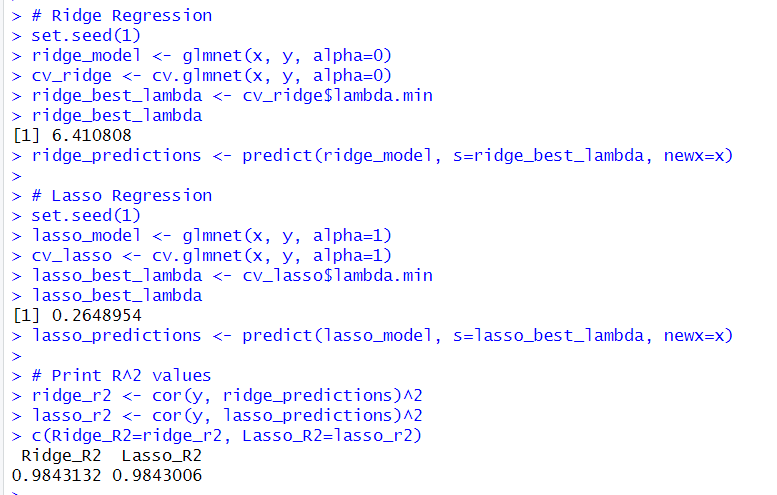
**Regularization Techniques**

**Ridge and Lasso Regression**

Given potential multicollinearity among predictors, Ridge and Lasso regression were employed to enhance model robustness.

* **Ridge Regression**: Applies L2 regularization, shrinking coefficients to minimize the impact of multicollinearity.
* **Lasso Regression**: Utilizes L1 regularization, which can lead to feature selection by shrinking some coefficients to zero.

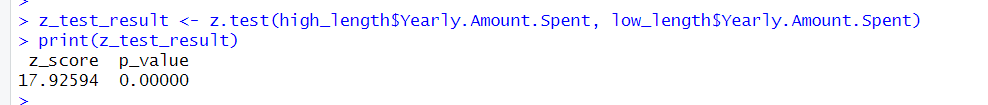
Cross-validation was performed to identify the optimal regularization parameters (lambda), significantly reducing prediction error and improving the model's generalization ability.



**Statistical Analysis**

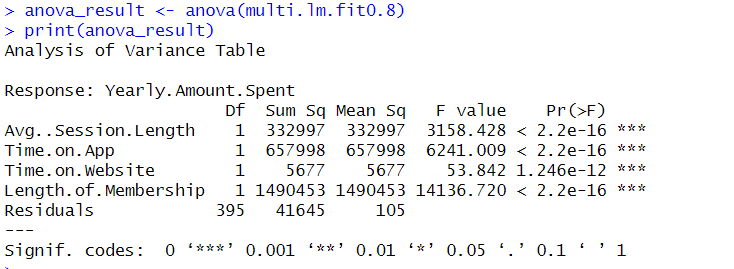
**1. Z-Test**

A Z-test was conducted to compare the spending behavior of customers with high versus low membership lengths. The p-value obtained was below 0.05, indicating a statistically significant difference between the two groups.



**2. ANOVA**

ANOVA was used to evaluate the overall significance of the regression model. The results confirmed that the predictors jointly contribute significantly to explaining the variance in yearly spending.



**Model Evaluation**

**1. Train-Test Split**

The dataset was split into training (80%) and testing (20%) sets. Model evaluation metrics included:

* **Root Mean Squared Error (RMSE)**: Measures the standard deviation of prediction errors.
* **Mean Absolute Percentage Error (MAPE)**: Provides a percentage-based evaluation of prediction accuracy.
* **R-Squared (R²)**: Indicates the proportion of variance explained by the model.

The multiple regression model achieved an R² value of 0.98, significantly outperforming the simple regression model (R² = 0.65).

**Cross-Validation**

A 10-fold cross-validation approach was used to assess the model's performance and ensure robustness against overfitting. The results confirmed the model's stability across different data subsets.

**Discussion**

The analysis revealed that increasing **Avg. Session Length**, **Time on App**, and **Length of Membership** positively impacts yearly spending. Surprisingly, **Time on Website** had a minimal effect, suggesting that efforts to enhance mobile app engagement and customer loyalty may yield better returns than focusing solely on the website experience.

Regularization methods, especially Lasso regression, proved effective in handling multicollinearity and improving model accuracy. By incorporating feature selection, the model achieved a better fit and more interpretable results.

**Conclusion**

This study demonstrates the effectiveness of linear regression and regularization techniques in predicting customer spending. The key findings are:

1. **Length of Membership** is the most significant predictor of yearly spending.
2. Enhancing user engagement on mobile applications is likely to increase annual customer expenditure.
3. Regularization methods such as Ridge and Lasso regression improve model accuracy and generalizability.

Future work could explore non-linear models or advanced machine learning techniques like decision trees and neural networks to capture complex patterns in customer behavior.

**References**

* Kaggle Dataset Link : <https://www.kaggle.com/datasets/iyadavvaibhav/ecommerce-customer-device-usage>
* FreeCodeCamp R tutorial Youtube : <https://youtu.be/_V8eKsto3Ug?feature=shared>
* GeeksForGeeks : <https://www.geeksforgeeks.org/r-tutorial/>