**Predicting Customer Purchase Behavior in an e-Commerce Platform**

**Summary**

This project uses machine learning techniques to predict customer purchase behavior on an eCommerce platform. The primary goal is to identify factors that influence purchasing decisions and develop a model that can predict whether a customer will buy. This report outlines the data preprocessing, exploratory data analysis, model selection, training, evaluation, and conclusions drawn from the study.

**Project Background**

ECommerce platforms generate vast amounts of data related to customer behavior, product preferences, and purchase history. Leveraging this data can significantly enhance marketing strategies, inventory management, and customer relationship management.

**Objective**

The objective of this project is to build a predictive model to forecast customer purchase behavior based on historical data. This can help in targeting potential customers, optimizing marketing campaigns, and improving sales conversion rates.

**Data Description**

**Data Collection**

The dataset used in this project consists of customer interaction data collected from an eCommerce platform over a specific period. The data includes attributes such as:

* Customer ID
* Session ID
* Page Views
* Time Spent on Site
* Product Categories Viewed
* Previous Purchase History
* Demographic Information (Age, Gender, Location)
* Device Used (Desktop, Mobile, Tablet)
* Purchase Indicator (Target Variable)

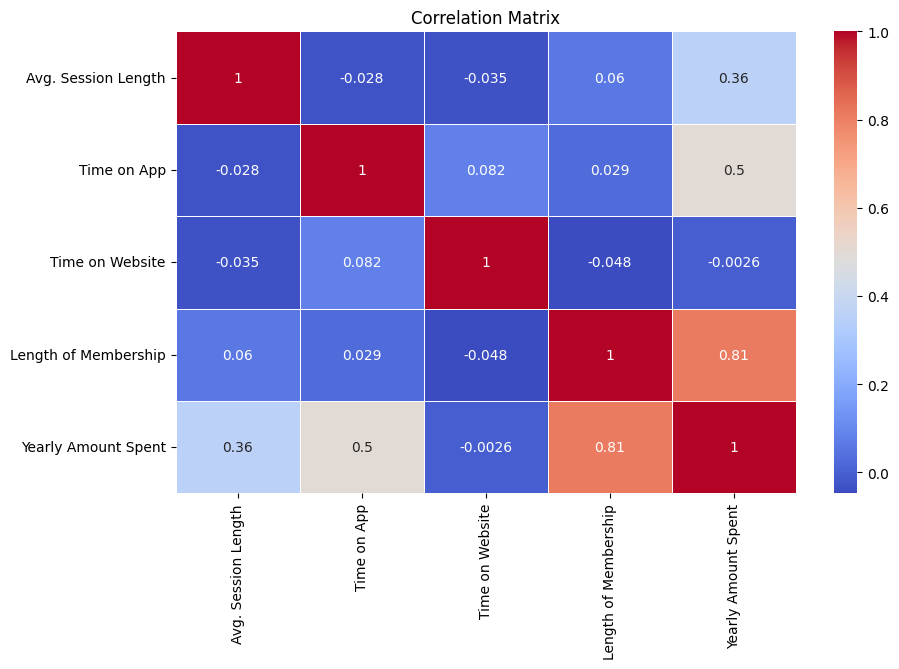
**Data Preprocessing**

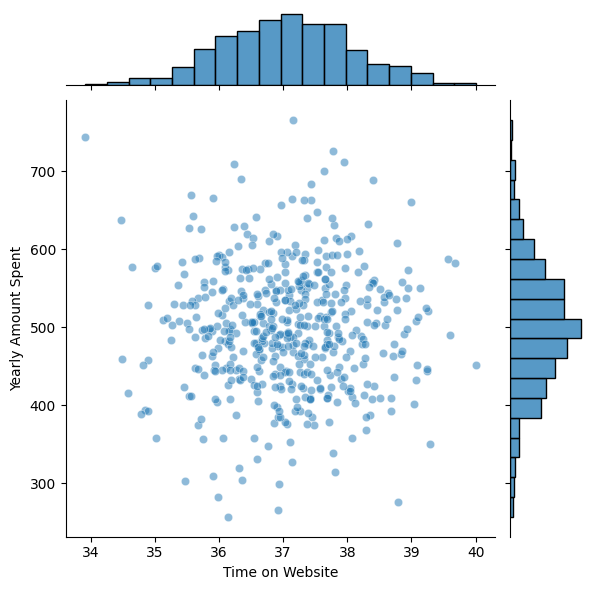
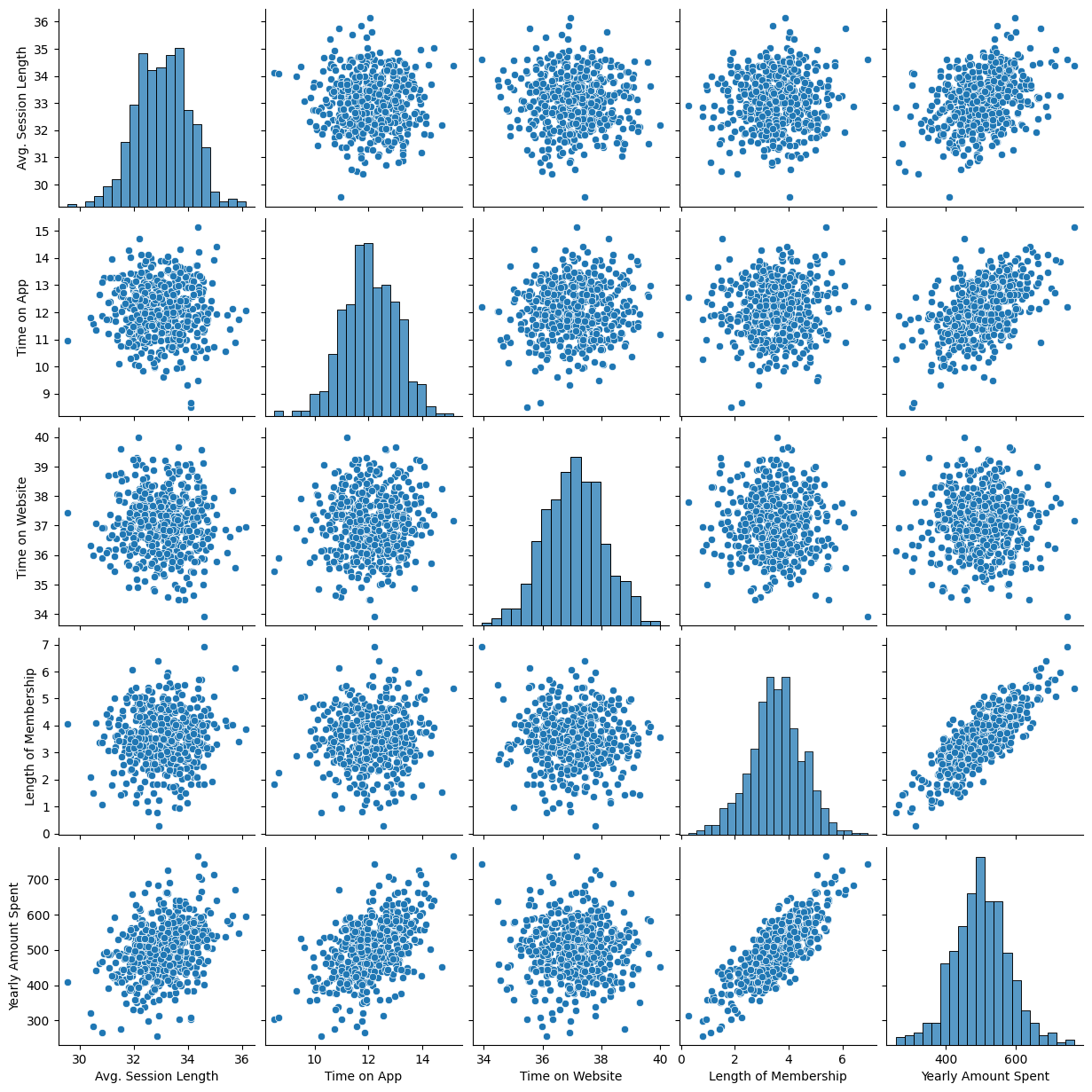
**Data preprocessing steps included:**

* Handling missing values.
* Encoding categorical variables.
* Normalizing numerical features.
* Splitting the data into training and test sets

**Data Visualization**

Visualizations were used to understand the distribution of data and relationships between features.





**Insights**

* A higher average session length and time spent on the app were positively correlated with the yearly amount spent.
* Length of membership showed a strong positive correlation with yearly amount spent.

**Model Selection**

Several machine learning models were considered, including linear regression, polynomial regression, decision trees, and random forests.

**Model Training**

Models were trained and evaluated using cross-validation.

**Linear Regression**

To establish a simple relationship between the features and the target variable. Linear regression aims to predict the yearly amount spent based on a linear combination of input features.

**Polynomial Regression**

To capture non-linear relationships between the features and the target variable. Polynomial regression extends linear regression by considering polynomial combinations of features.

**Decision Tree**

To model decision-making processes. Decision trees split the data into branches to make predictions, allowing for complex decision boundaries and interactions between features.

**Random Forest**

To improve the accuracy and robustness of predictions by averaging multiple decision trees. Random forests reduce overfitting and capture a broader range of patterns in the data.

**Results and Discussion**

**Model Performance**

**Linear Regression:**

* **MAE**: 8.426
* **MSE**: 103.916
* **RMSE**: 10.194
* **R2**: 0.981

**Polynomial Regression:**

* **MAE**: 8.594
* **MSE**: 109.347
* **RMSE**: 10.457
* **R2**: 0.980

**Decision Tree:**

* **MAE**: 19.697
* **MSE**: 655.237
* **RMSE**: 25.598
* **R2**: 0.879

**Random Forest:**

* **MAE**: 13.307
* **MSE**: 286.342
* **RMSE**: 16.922
* **R2**: 0.947

**Discussion**

**Model Performance Analysis:**

* **Linear Regression**: This model performed the best with an R2 score of 0.981, indicating it explains approximately 98.1% of the variance in the yearly spending. The low MAE and RMSE values further support its high accuracy and precision.
* **Polynomial Regression**: While slightly less accurate than Linear Regression, Polynomial Regression also showed strong performance with an R2 score of 0.980. However, it didn't significantly outperform the simpler Linear Regression model, suggesting that the added complexity might not be necessary.
* **Decision Tree**: This model showed the lowest performance among the models with an R2 score of 0.879. The higher MAE and RMSE values indicate less accurate predictions and larger errors, which might be due to overfitting the training data.
* **Random Forest**: This model performed better than the Decision Tree with an R2 score of 0.947. The lower MAE and RMSE values compared to the Decision Tree indicate that Random Forest was able to generalize better and reduce errors. However, it still didn't surpass the performance of Linear Regression.

**Insights and Recommendations:**

* The strong performance of Linear Regression suggests that the relationship between the features and the target variable is largely linear. This insight can simplify future predictive modeling efforts, focusing on linear techniques.
* The findings emphasize the importance of 'Length of Membership' and 'Time on App' in driving customer spending. Strategies to enhance customer experience on the app and encourage long-term memberships could be effective in increasing yearly spending.
* Future work should explore more advanced models and additional features to potentially capture more complex patterns and improve prediction accuracy.

**Conclusion**

The project successfully demonstrated the effectiveness of various regression techniques in predicting yearly customer spending. Linear Regression, despite its simplicity, proved to be the most accurate and reliable model. These insights can guide eCommerce platforms in strategic decision-making to boost customer engagement and revenue. Future directions include further exploration of Neural networks to enhance predictive performance.