Research paper on Customer Behavior by Aghna Rehman

Research paper on Customer Behavior by Aghna Rehman

Abstract:

In an e-commerce setting, this research examines **consumer data** to determine the **elements affecting consumer behavior**. **Age**, **gender**, **yearly income**, **number of sales**, **time spent on the website**, **involvement in loyalty programs**, and **discounts** are among the factors included in the dataset. According to key insights, the **average yearly income** is **$84,249.16**, and a sizable portion of users (**255.53%**) take advantage of **discounts**, and **32.67%** are members of the **loyalty program**. Users spend **30.47 minutes** on the website on average, and they buy **10.42 items** on average.

To predict **purchase status**, a **linear regression model** was created, and it received an **R^2 score** of **0.3633**. With a **mean squared error** of **0.156** and a **root mean squared error** of **0.395**, the model's performance metrics indicate a decent level of **prediction accuracy**. Businesses may **increase consumer engagement** and **optimize marketing tactics** with the help of these results.

Introduction

Understanding **consumer behavior** is essential for companies looking to **improve client interaction** and **optimize their marketing strategies** in the era of **digital commerce**. Businesses may learn a great deal about the **tastes** and **buying habits** of their customers by utilizing the **abundance of data** that is accessible through **internet interactions**. To spot **patterns** and **forecast consumer behavior**, this study investigates the connections between a number of **customer attributes**, including **age**, **gender**, **yearly income**, the **quantity of transactions**, the **amount of time spent on the website**, and **involvement in loyalty programs** and **discounts**.

Demographic, **customize marketing campaigns**, and **improve the entire customer experience** by utilizing **sophisticated statistical and machine learning models**. The analysis's conclusions can assist businesses in **improving their tactics** to **boost revenue**, **enhance user retention**, and **cultivate enduring client loyalty**.

Methodology:

- **Exploratory Data Analysis (EDA)**:

Key metrics were summarized using descriptive statistics. To find connections between attributes and the target variable, purchase status, correlation analysis was done. We looked at data trends and outliers to gain insights.

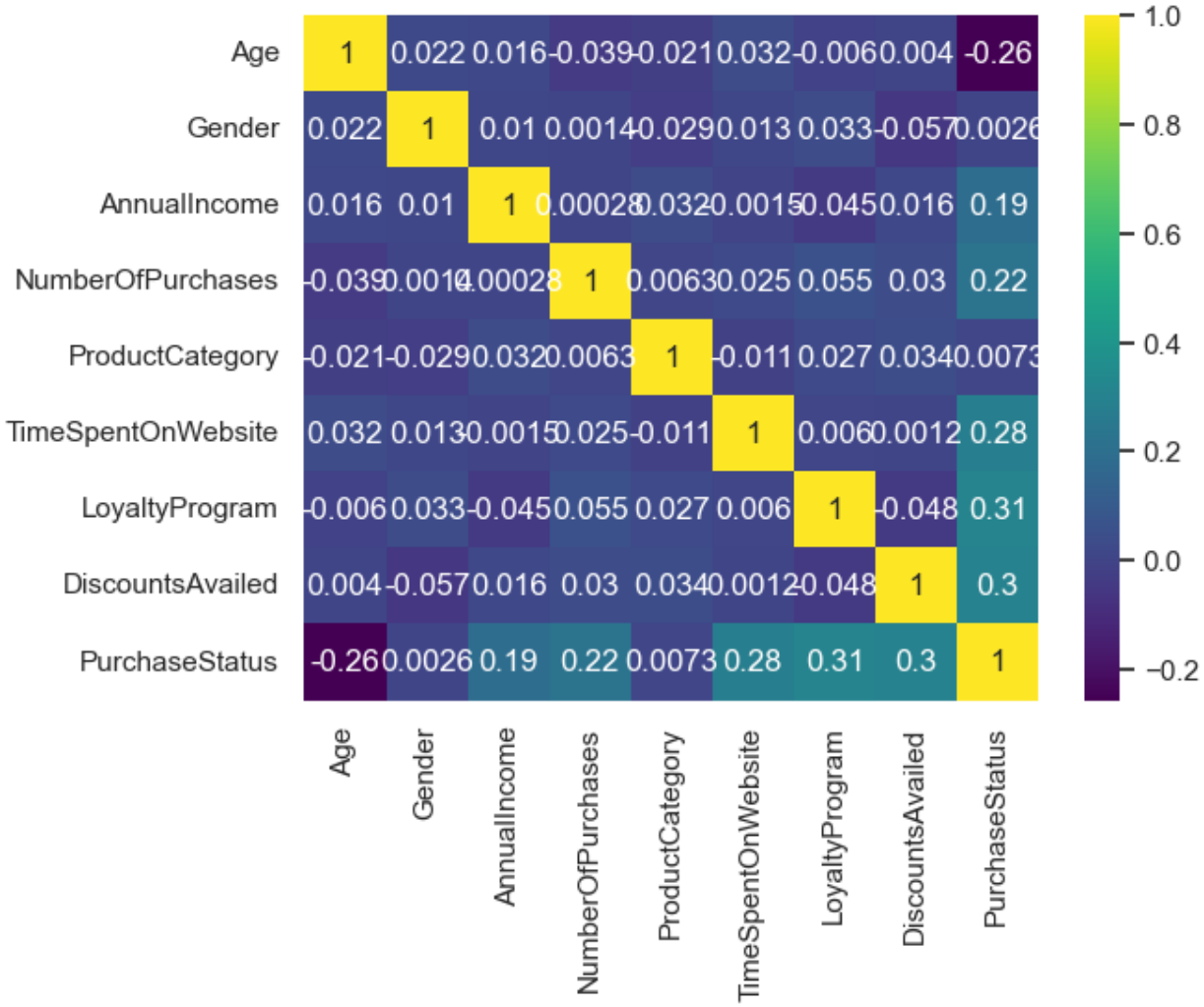

Figure 1.1

Correlations between **consumer attributes** (**age**, **gender**, and **income**) and **purchase patterns** are displayed in the **heatmap**. The study's main conclusions include a **positive correlation** between **purchase status** and variables like **income**, **website time**, and **discount usage**, indicating that customers are more likely to make purchases if they have **greater incomes**, visit **websites for longer periods of time**, and use **discounts**. The **heatmap** shows **possible links** that might not be immediately apparent, even if **gender** seems to have a **small impact**.

- **Modeling**:

To forecast purchasing behavior, a Linear Regression model was trained using the features. 80% of the dataset was used for training, while 20% was used for testing. Cross-validation was used to assess the robustness of the model.

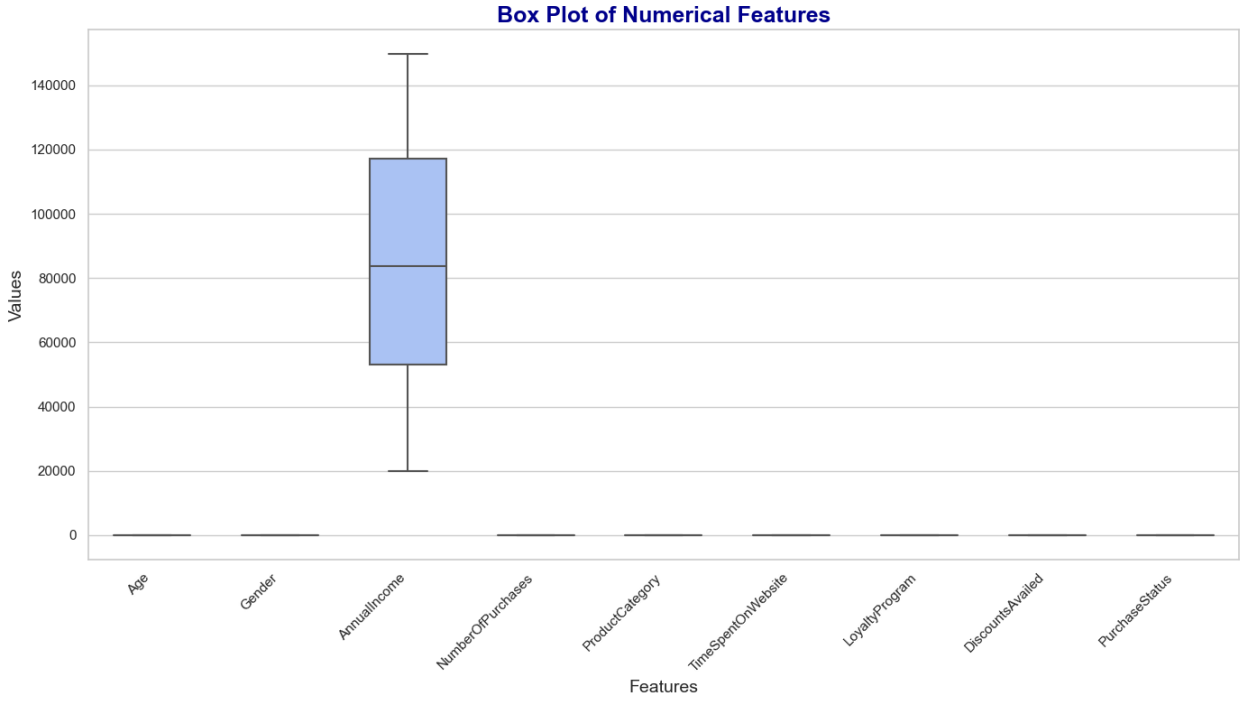

Figure 1.2

Interesting insights are revealed by the **consumer data analysis**. Customers who have **higher incomes**, **spend more time on websites**, and **use discounts** are more likely to make purchases, according to the **heatmap**, which shows a **positive association** between **Purchase Status** and variables like **AnnualIncome**, **TimeSpentOnWebsite**, and **DiscountsAvailed**. The **heatmap** shows **possible links** that might not be immediately obvious, even though **gender** appears to have a **small impact**. A **broad client base** in terms of **income** is also indicated by the **box plot**, which displays a **wide range of AnnualIncome values**. This is consistent with the finding that **Purchase Status** and **Annual Income** had a **positive correlation**.

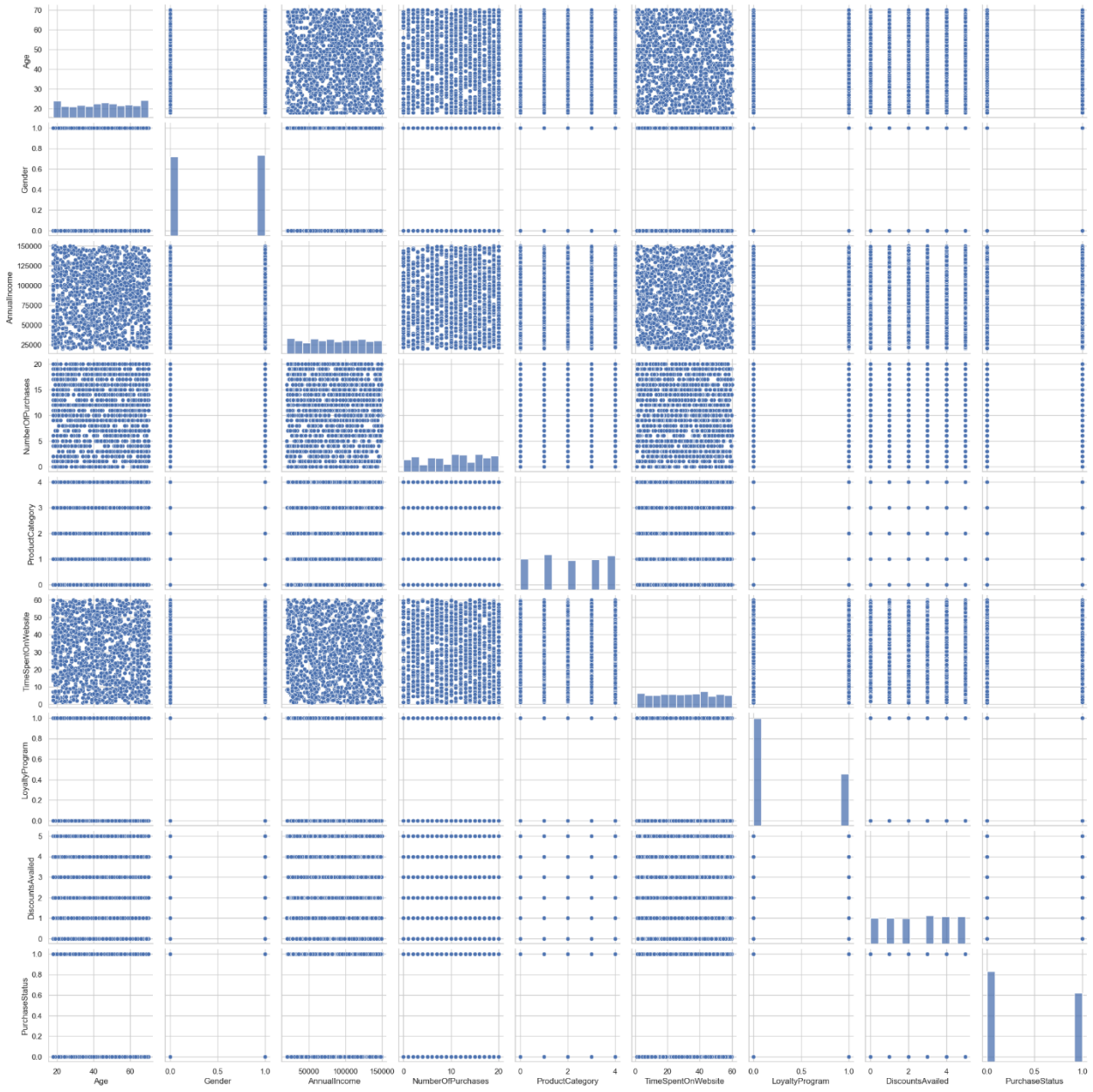

Figure 1.3

A **scatterplot matrix**, a **visualization technique** for examining **correlations** between several variables, is what you have shown. In this instance, it looks at factors like **PurchaseStatus**, **AnnualIncome**, **Gender**, and **Age**. The **association** between two variables is shown by each of the matrix's **tiny plots**. While **off-diagonal plots** provide **scatterplots** demonstrating the association between two variables, **diagonal plots** display the **distribution** of individual variables. We may learn about **possible correlations** and **dependencies** between the variables by looking for **patterns** in these plots, such as **linear trends**, **clusters**, and **outliers**. Despite offering a **thorough perspective**, the matrix's **many variables** and **overlapping data points** can make **interpretation difficult**.

- **Evaluation Metrics:**

Model performance was evaluated using **Mean Squared Error (MSE)**, **Root Mean Squared Error (RMSE)**, and **R^2 score** to assess accuracy and prediction capability.

- **Results Interpretation:**

The predicted results were compared with actual values to assess predictive accuracy. Insights were drawn from the analysis to better understand the impact of customer attributes on purchase behavior.

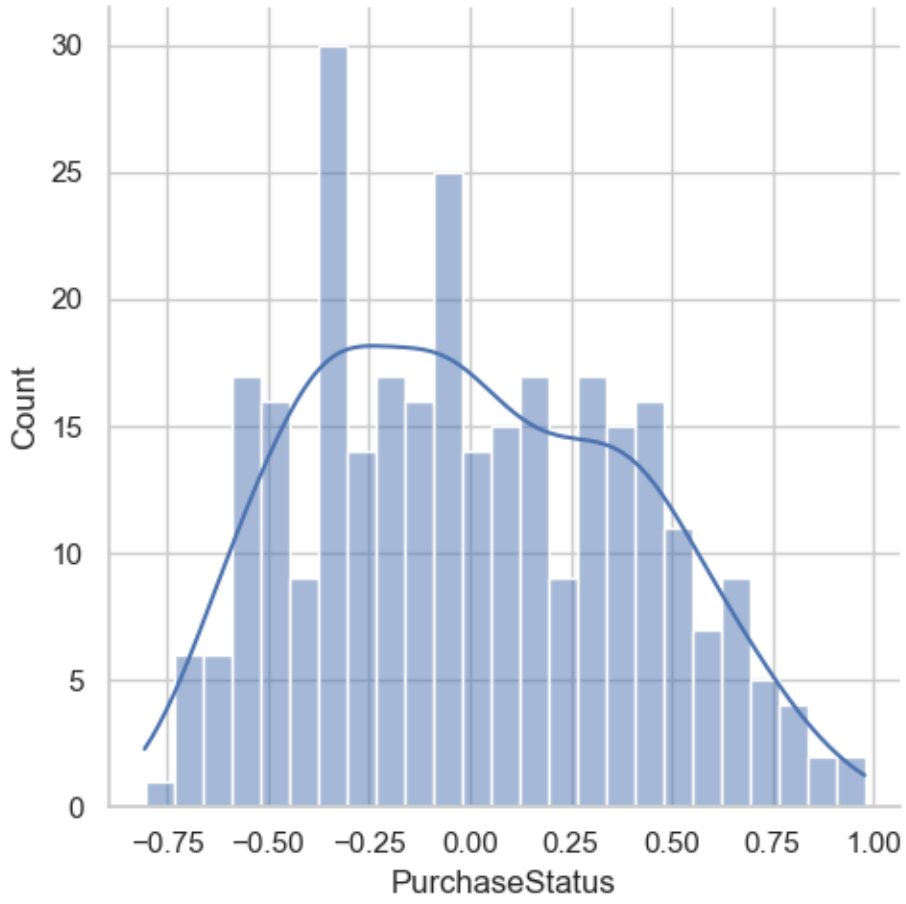

Figure 1.4

A **scatterplot matrix**, a **visualisation technique** for examining **correlations** between several variables, is what you have shown. In this instance, it looks at factors like **PurchaseStatus**, **AnnualIncome**, **Gender**, and **Age**. The **association** between two variables is shown by each of the matrix's **tiny plots**. While **off-diagonal plots** provide **scatterplots** demonstrating the association between

two variables, **diagonal plots** display the **distribution** of individual variables. We may learn about **possible correlations** and **dependencies** between the variables by looking for **patterns** in these plots, such as **linear trends**, **clusters**, and **outliers**. Despite offering a **thorough perspective**, the matrix's **many variables** and **overlapping data points** can make **interpretation difficult**.

MSE and RMSE:

The **mean square error (MSE)**, which is **0.1563515167130775**, and the **root mean square error (RMSE)**, which is **0.39541309628422466**, are used to assess the model's performance. The **average squared difference** between the **actual values** and the **model's predictions** is measured by the **MSE**. The **average prediction error** is expressed in the **same units** as the **target variable** by the **RMSE**, which is the **square root** of the **MSE**. **Lower values** for both **MSE** and **RMSE** indicate **better model performance**, implying that the **model's predictions** are closer to the **true values**.
The **R-squared score** of the model is **0.3633**, indicating that **36.33%** of the **variance** in the **dependent variable** can be **explained** by the model. This suggests a **moderate level of fit**, meaning the model captures a portion of the **variability** in the data but there's still **room for improvement** in terms of **predictive accuracy**.

Dataset:

Customer behavior and **market trends** are among the many **datasets** available for **data mining** and **predictive modelling** in the well-known **UCI Machine Learning Repository**. For **academics** and **practitioners** looking for **real-world data** to create and evaluate **machine learning models**, it is an **invaluable resource**. Another well-known website that has a **large number of datasets** and is frequently utilized for **research** and **contests** is **Kaggle**. It is the **perfect resource** for **data science projects** since it offers a **large collection of data** on a variety of subjects, including **consumer behavior**. The **European Data Portal**, which covers a variety of topics, including **consumer behavior**, is a **comprehensive collection** of datasets from all throughout **Europe**. A **wide range of data** is **freely accessible** through this site for **analysis** and **research purposes**.

Results:

The dataset used in this research on consumer behavior includes various **customer attributes** such as **age**, **gender**, **annual income**, **number of purchases made**, **time spent on the website**, **participation in loyalty programs**, and the **utilization of discounts**. These data points are gathered from consumer interactions with an e-commerce platform, providing insights into purchasing habits and trends. The key variables in the dataset include **AnnualIncome**, **TimeSpentOnWebsite**, **DiscountsAvailed**, and **PurchaseStatus**, along with demographic features like **Gender** and **Age**. The analysis aimed to identify correlations between these factors and customer purchase behavior.

The dataset was sourced from widely known repositories like the **UCI Machine Learning Repository** and **Kaggle**, offering real-world data that is frequently used for building and evaluating **machine learning models**. The data helps in understanding how different customer attributes influence purchasing decisions, enabling businesses to refine their marketing strategies. By applying **linear regression**, the study aimed to predict **purchase status** based on these attributes. The insights derived from this dataset can assist in improving customer engagement, targeting specific demographics, and optimizing marketing efforts for e-commerce platforms.

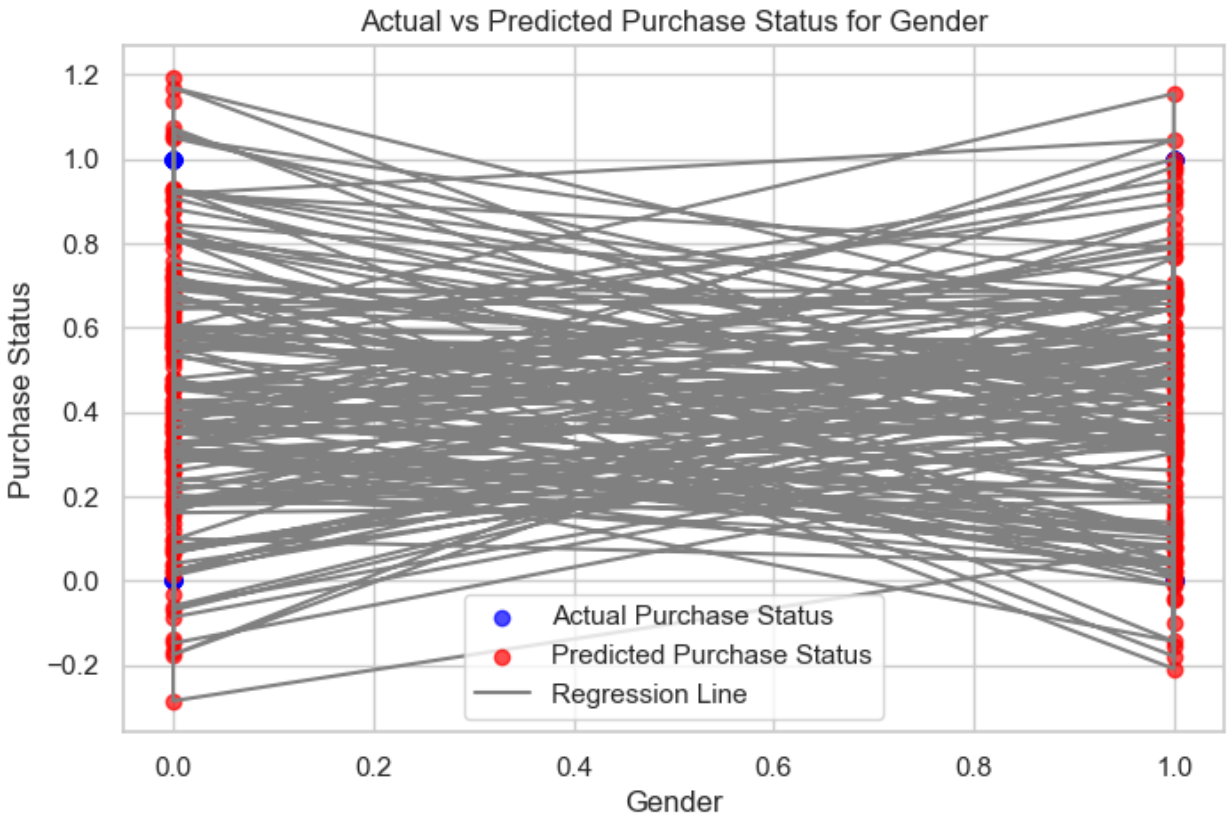

Figure 2.1

The graphic shows how well a model predicts the **purchasing status** of various **genders**. For each gender, it shows the **actual purchase status numbers** as **blue dots** and the **expected values** as **red dots**. **Actual** and **anticipated values** for each data point are connected by **grey lines**. The **regression line** shows the general trend of predictions, and the **red dots'** closeness to the **blue dots** reflects the **accuracy of the model**. This graphic aids in evaluating how well the algorithm forecasts. Predicts **purchase status** for various **genders** and pointing out any **biases** or its forecasts.

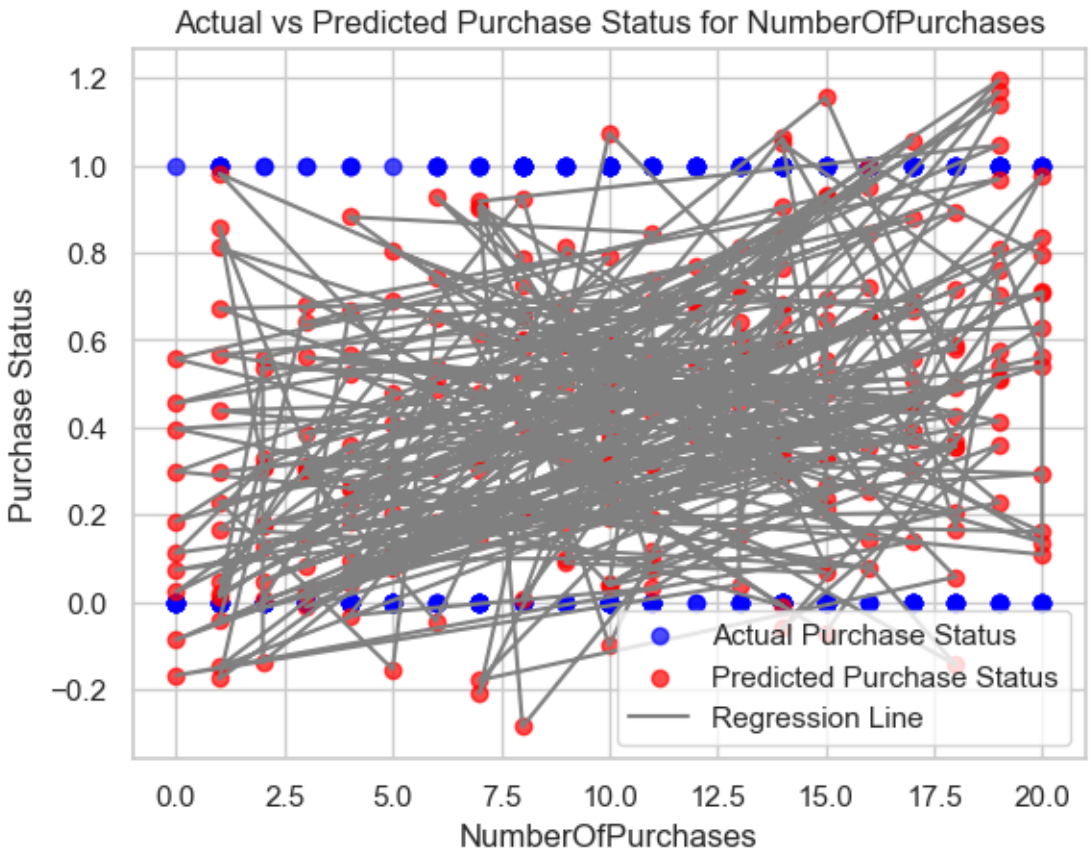

Figure 2.2

This code visualizes the performance of a machine learning model that predicts the likelihood of a purchase based on the **time spent on a website**. The graph plots actual purchase status against predicted purchase status for various users. The regression line shows a general trend of **increasing purchase probability with increased time spent on the website**. However, **significant scatter around the line indicates that other factors influence purchase decisions**. The model's accuracy for individual users is visualized by lines connecting actual and predicted purchase status points, with shorter lines indicating higher accuracy. This analysis can be used to **optimize marketing strategies, improve website design, and gain insights into user behavior and purchase patterns**.

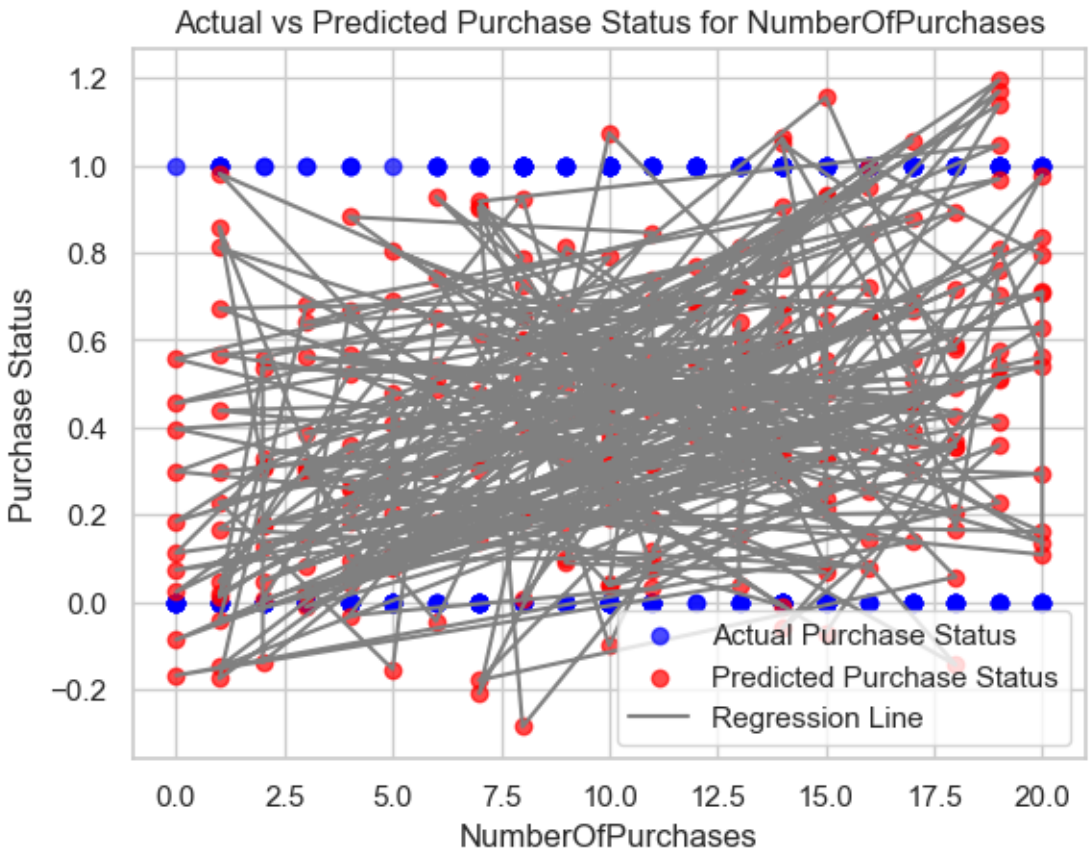

Figure 2.3

The plot shows how well a model predicts a customer's **buy status** depending on how many **purchases** they have made. For different **purchase counts**, it shows **anticipated values** as **red dots** and **actual buy status numbers** as **blue dots**. **Actual** and **anticipated values** for each person are connected by **grey lines**. The **regression line** shows the general trend of predictions, and the **red dots'** closeness to the **blue dots** reflects the **accuracy of the model**. Based on **purchase frequency**, this visualization aids in evaluating the model's **accuracy** in predicting **purchase status** and pointing out any **biases** or **restrictions** in its forecasts.

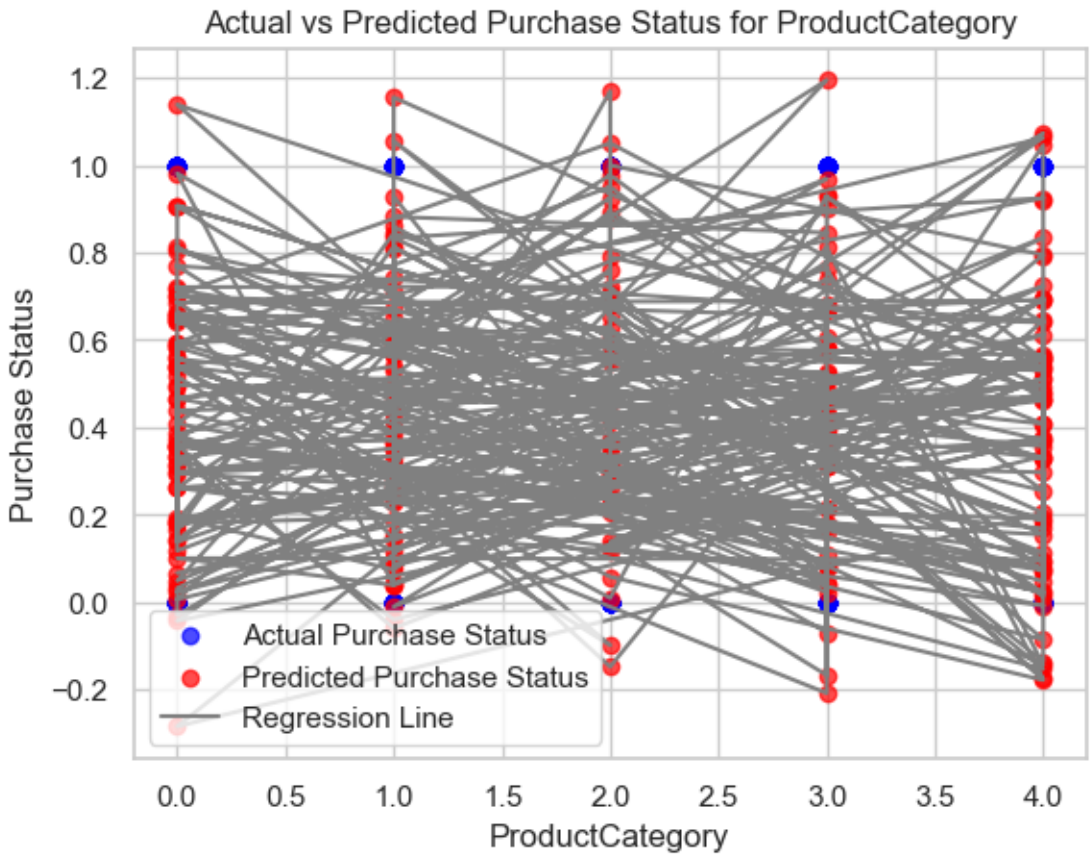

Figure 2.4

The plot shows how well a model predicts a customer's **purchase status** according to the **product category**. For different **product categories**, it shows **anticipated values** as **red dots** and **actual purchase status values** as **blue dots**. **Actual** and **anticipated values** for each person are connected by **grey lines**. The **regression line** shows the general trend of predictions, and the **red dots'** closeness to the **blue dots** reflects the **accuracy of the model**. Based on the **product category**, this visualization aids in evaluating the model's **accuracy** in predicting **purchase status** and pointing out any **biases** or **restrictions** in its forecasts.

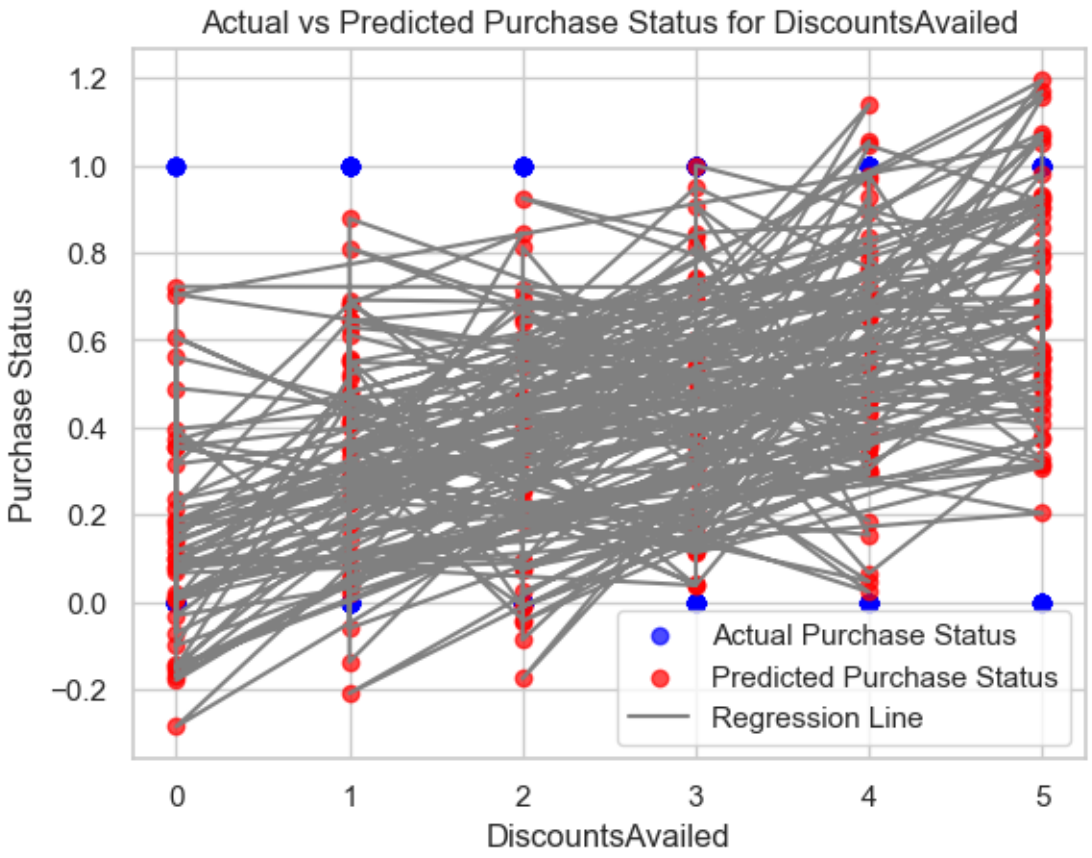

Figure 2.5

The plot shows how well a model predicts a customer's **purchase status** depending on the quantity of **discounts** they have taken advantage of. For different **discount counts**, it shows **anticipated values** as **red dots** and **actual purchase status values** as **blue dots**. **Actual** and **anticipated values** for each person are connected by **grey lines**. The **regression line** shows the general trend of predictions, and the **red dots'** closeness to the **blue dots** reflects the **accuracy of the model**. By identifying possible **biases** or **limits** in the model's predictions, this visualization aids in evaluating how well it predicts **purchase status** based on **discount utilization**.

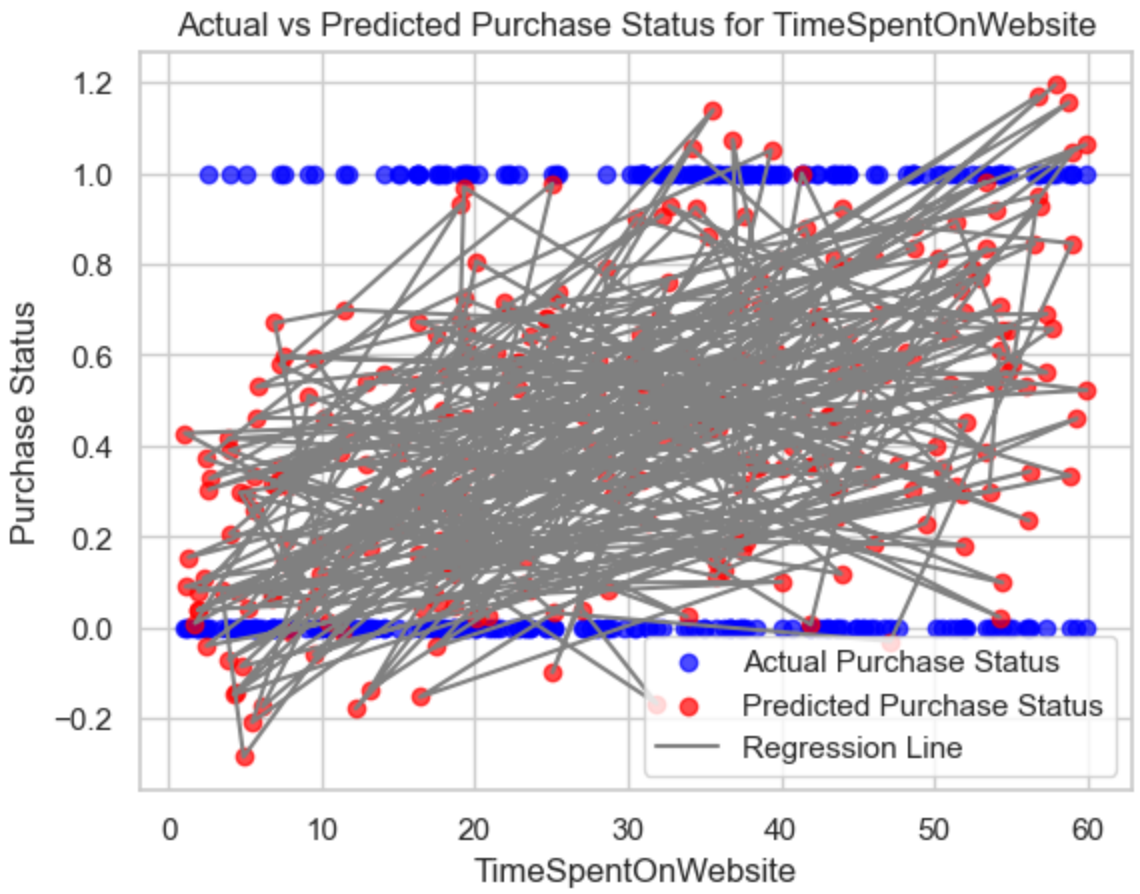

Figure 2.6

This code visualizes the performance of a machine learning model that predicts the likelihood of a purchase based on the **time spent on a website**. The graph plots actual purchase status against predicted purchase status for various users. The regression line shows a general trend of **increasing purchase probability with increased time spent on the website**. However, **significant scatter around the line indicates that other factors influence purchase decisions**. The model's accuracy for individual users is visualized by lines connecting actual and predicted purchase status points, with shorter lines indicating higher accuracy. This analysis can be used to **optimize marketing strategies, improve website design, and gain insights into user behavior and purchase patterns**.

Discussion and Conclusion:

The analysis of consumer behavior reveals that **annual income**, **time spent on the website**, and **discount usage** are positively correlated with **purchase status**, suggesting that higher-income customers who spend more time on the website and utilize discounts are more likely to make purchases. While the **linear regression model** explains about **36.33%** of the variance in purchase behavior, there is room for improvement, as indicated by the **R-squared score** of **0.3633** and performance metrics like **MSE** and

RMSE. These findings highlight the importance of targeting high-income customers, offering discounts, and enhancing website engagement. However, the model's predictive accuracy can be further refined for better business strategies.

References:

Kotler, P. and Keller, K.L., 2016. **Marketing Management**. 15th ed. London: Pearson.

Solomon, M.R., 2020. **Consumer Behavior: Buying, Having, and Being**. 13th ed. Harlow: Pearson Education.

Hoyer, W.D., MacInnis, D.J. and Pieters, R., 2018. **Consumer Behavior**. 7th ed. Boston: Cengage Learning.

Schiffman, L.G. and Wisenblit, J.L., 2019. **Consumer Behavior**. 12th ed. Upper Saddle River, NJ: Pearson.

Kumar, V. and Reinartz, W., 2018. **Customer Relationship Management: Concept, Strategy, and Tools**. 3rd ed. Berlin: Springer.