Team 1: Sandhya Karki, Qibin Huang, Rakesh Venigalla
DATS 6103: Summary Report
Professor Sushovan Majhi
December 2023

Customer Trends Dataset

Introduction

In today's rapidly evolving digital landscape, seizing opportunities quickly can be a key driver of a company's growth, with data mining playing a pivotal role. Data mining enables companies to analyse extensive datasets, uncover operational insights, and predict future trends. This approach is crucial in staying ahead in the competitive e-commerce sector.

Our project leverages a detailed dataset from Kaggle, encompassing various consumer information such as demographics, purchase history, payment methods, and buying frequency. With 3,900 records, this dataset will help us understand consumer behaviours and preferences, aiding retailers in tailoring personalised shopping experiences.

However, we're taking a step further in our research. Our dataset is AI-generated, specifically using GPT technology. In this study, we aim to not only delve into the nuances of our database's content but also to evaluate how an AI and GPT-generated dataset, which has become an integral part of our lives and a tool we rely on in many ways, compares to real-world scenarios. We seek to understand the accuracy and relevance of AI-generated data in mirroring actual consumer behaviour and market trends.

This project is designed to explore the intricate relationships between various factors, such as customer demographics, product categories, purchase amounts, and their impact on shopping trends. By analysing this AI-generated data, we hope to provide insights that will help retailers improve customer satisfaction, customise marketing strategies, and drive sales growth, while also assessing the reliability and applicability of AI-generated data in real-world business contexts.

SMART Questions

1. Which factors (such as Gender and location) have the most significant impact on the purchase amount?
   Objectives: Identify the key features that influence the purchase amount.

2. What is the relationship between review ratings and repeat purchases?
   Objectives: Using this analysis, we reveal if higher-rated products or services lead to more repeat purchases, indicating customer satisfaction.

3. Correlation between the Purchase Amount & customer's Repurchase Behaviour & the Product Category
   Objectives: How the amount spent on purchases is related to customer loyalty and the type of products they buy.

About the Dataset

Description Of Data:

The dataset used for this project is a large collection of retail customer data obtained from Kaggle. This comprehensive dataset totals over 3,900 rows, each detailing various aspects of retail customer behaviour and customer information.

Key Features of the Dataset:

- Customer Demographics: This includes vital information such as the age, gender, and possibly the occupation and income levels of the customers. These demographic details are instrumental in understanding the diversity of the customer base and in identifying specific trends or preferences within different demographic groups.

- Purchase History: The dataset contains extensive records of past purchases made by each customer. This aspect of the data reveals patterns in purchase frequency, preferred product categories, and average spending amounts, offering insights into customer loyalty and purchasing power.

- Product Categories: Detailed classifications of products purchased allow for an analysis of popular categories and preferences among different demographics. This information is crucial in understanding which products are favoured by certain segments of the customer base.

- Payment Methods: Records of how purchases are made (e.g., cash, credit/debit cards, online payment systems) provide an understanding of the preferred transaction modes for customers. This can be indicative of the customer's purchasing behaviour and financial tendencies.

- Shopping Frequency: The dataset also tracks the frequency of customer visits or purchases, indicating customer engagement and loyalty. This metric is important for assessing repeat business and the effectiveness of marketing strategies.

Preparing The Data

Before delving into the analysis, it was crucial to prepare the dataset to ensure the accuracy and relevance of the findings. The data preparation process involved several key steps to refine and structure the dataset for effective analysis.

- Data Cleaning: The initial step involved thorough cleaning of the dataset. This included identifying and addressing missing values, removing duplicates, and correcting any inconsistencies or errors in the data. For instance, missing demographic information was handled appropriately to ensure a complete and accurate representation of the customer base.

- Data Transformation: Certain variables in the dataset required transformation to make them suitable for analysis. For example, age ranges were converted into numerical values, and categorical data, such as gender and payment methods, were encoded for easy analysis. This step was vital in standardising the data format for subsequent analytical procedures.

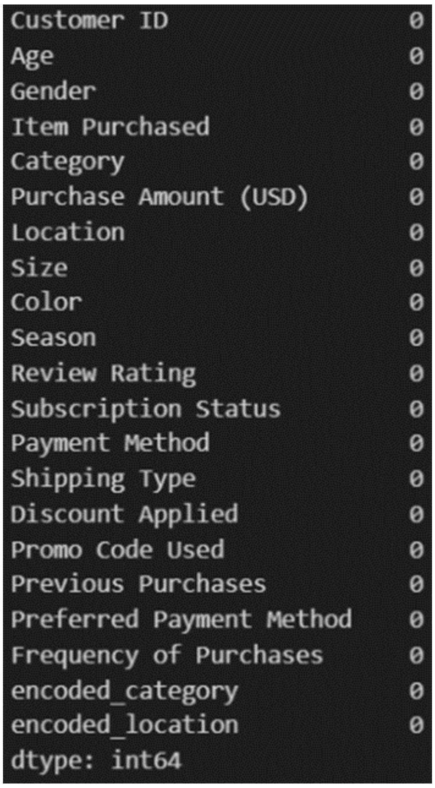
- Feature Selection: We performed feature selection to focus the analysis on the most relevant aspects. This involved identifying and selecting the variables most likely to influence customer behaviour and purchase patterns. Age, gender, product categories, and purchase frequency were prioritised based on their potential impact on shopping trends.

- Data Segmentation: The dataset was segmented based on different demographic criteria and shopping behaviours. This segmentation allowed for a more detailed and nuanced analysis, enabling us to examine patterns and trends within specific customer groups.

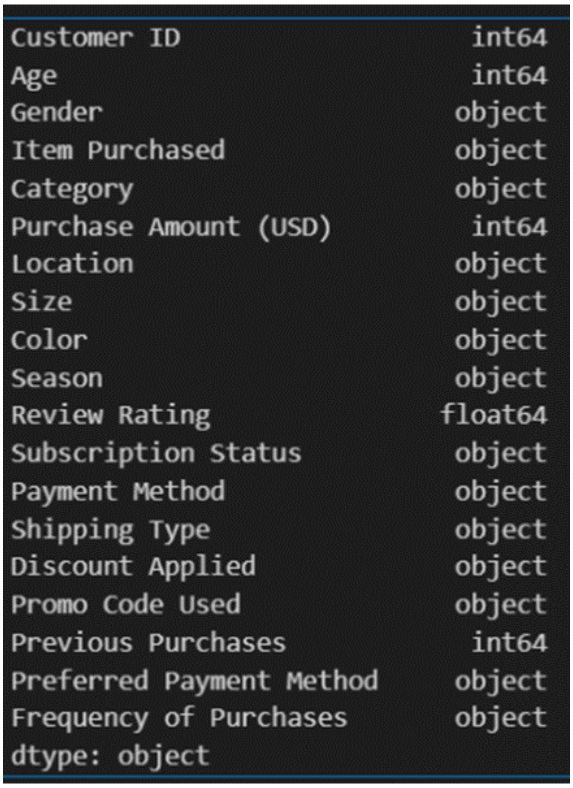
- Handling Outliers: Outliers in the data were identified and addressed. In cases where outliers represented genuine customer behaviours, they were retained to preserve the integrity of the dataset. However, where outliers were due to data entry errors or other anomalies, they were corrected or removed.

Cleaning Data:

Importing the Dataset



```
Customer ID                 0
Age                         0
Gender                      0
Item Purchased              0
Category                    0
Purchase Amount (USD)       0
Location                    0
Size                        0
Color                       0
Season                      0
Review Rating               0
Subscription Status         0
Payment Method              0
Shipping Type               0
Discount Applied            0
Promo Code Used             0
Previous Purchases          0
Preferred Payment Method    0
Frequency of Purchases      0
encoded_category            0
encoded_location            0
dtype: int64
```



```
Customer ID                   int64
Age                           int64
Gender                       object
Item Purchased               object
Category                     object
Purchase Amount (USD)         int64
Location                     object
Size                         object
Color                        object
Season                       object
Review Rating               float64
Subscription Status          object
Payment Method               object
Shipping Type                object
Discount Applied             object
Promo Code Used              object
Previous Purchases            int64
Preferred Payment Method     object
Frequency of Purchases       object
dtype: object
```

FIG. 1 Shows no null values

FIG. 2 Category of data

In our dataset, there aren't any missing values, and we've got 22 columns, each holding different types of information like numbers, words, and decimals. Now, for our analysis, we're narrowing down our focus to just three columns that seem most important. This helps us concentrate on specific data that we believe will give us the most useful insights. By selecting these three columns, we're aiming to dig deeper into the information they hold and uncover meaningful patterns or trends that can guide our analysis and decision-making process.

Exploratory Data Analysis

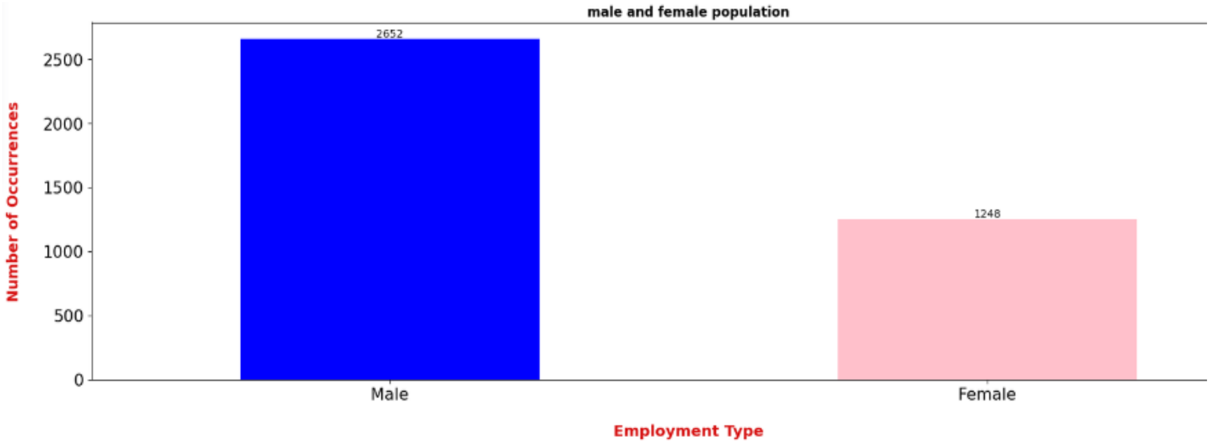


FIG. 3 Male and Female Population

This bar graph shows the number of customers based on their gender. On the horizontal line (x-axis), you'll see the genders - males and females. The vertical line (y-axis) represents how many customers belong to each gender.

Looking at the graph, it's clear that there are more males than females in this customer dataset.

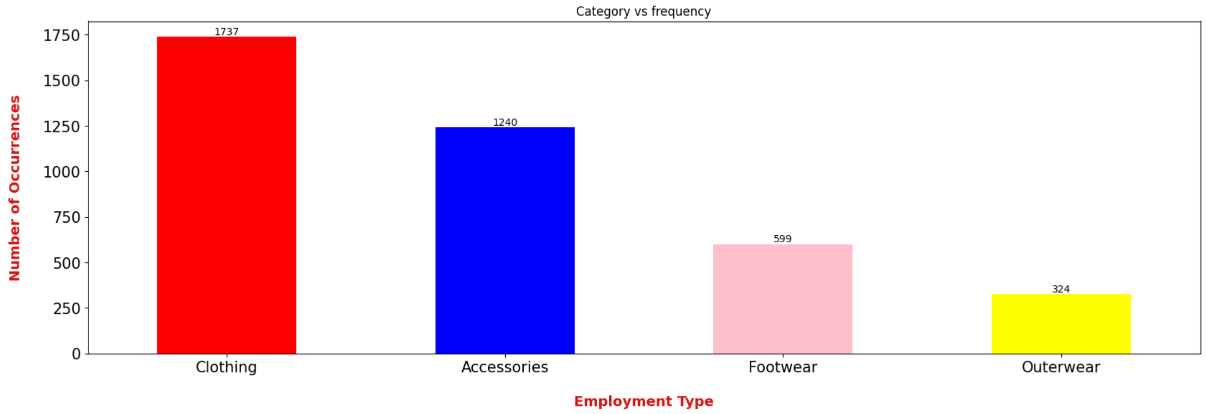


FIG. 4 Category vs Frequency

In the graph, we see a bar plot depicting how often customers buy products in different categories. The x-axis shows the product categories, and the y-axis represents the buying frequency. Notably, clothing emerges as the most popular category, with the highest number of purchases. Similarly, outerwear stands out as the least frequently bought item. This simple visualization gives a quick overview of customer preferences, highlighting the strong demand for clothing and the comparatively lower interest in outerwear among the surveyed customers.

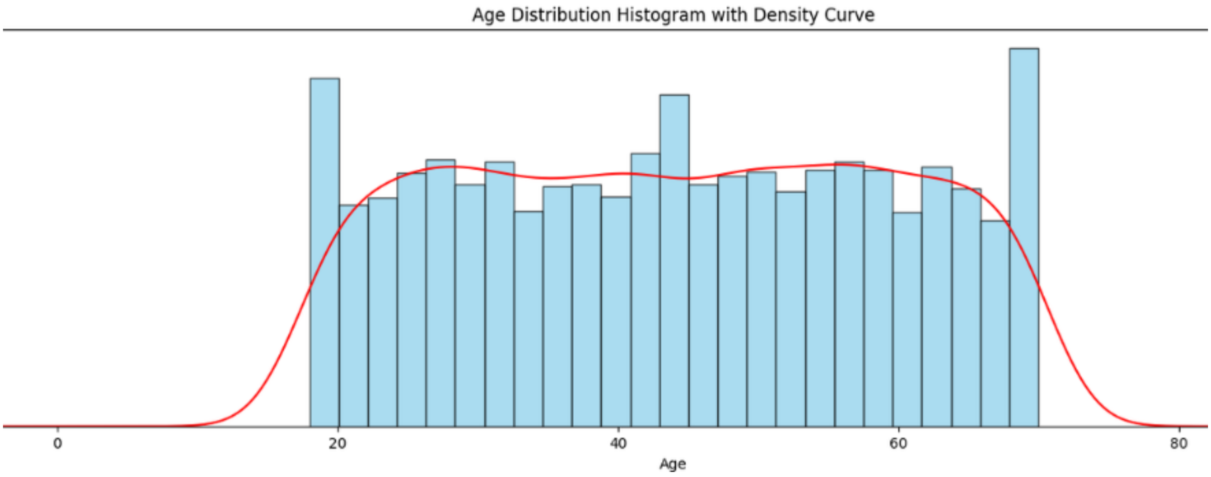


FIG. 5 Age distribution curve

The chart displays the age distribution of customers in a dataset. It reveals that most customers fall between the ages of 20 and 70. Looking at the density curve, we observe fairly even frequencies across various age groups. Notably, ages around 20, 41, and 65 stand out as the most common. It simply shows that the dataset covers a broad age range, with a somewhat balanced representation of customers in different age brackets.

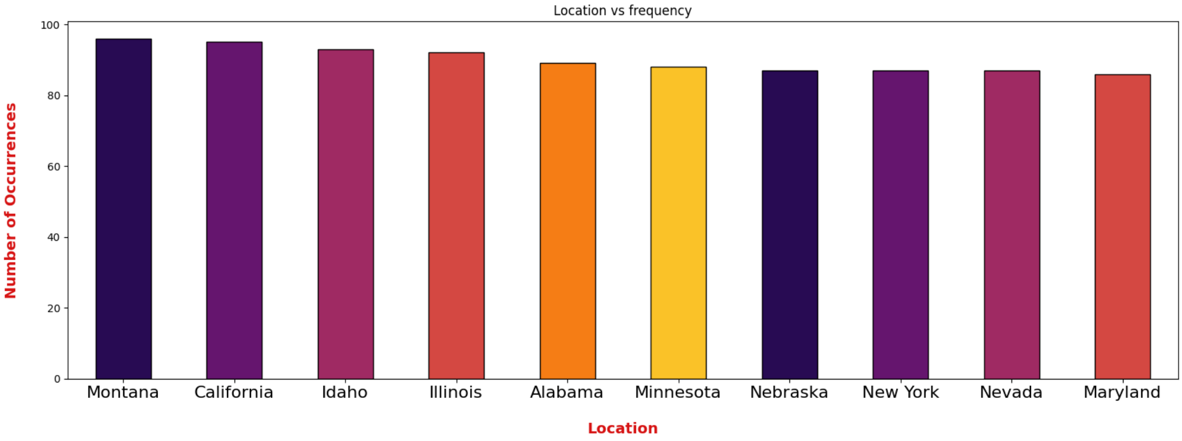


FIG. 6 Location vs Frequency

This bar graph shows where customers visit the most, with different locations on the x-axis and how often they're visited on the y-axis. Montana stands out as the most popular spot, while Maryland seems to be the least visited among them all. Interestingly, most locations have a similar number of visitors, making the graph pretty even across the board. It's like Montana's the hot spot and Maryland's the quieter one in this crowd of places customers like to go.

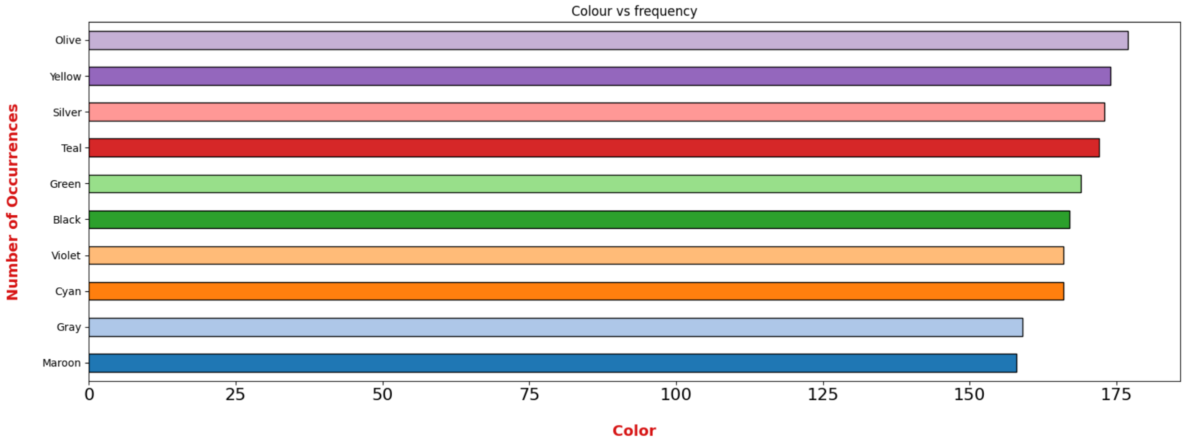


FIG. 7 Colour vs Frequency

The bar graph shows which colours customers bought the most, and it's clear that olive was the top-selling colour. On the flip side, maroon was the least popular choice among buyers. This data gives a quick view of which colours flew off the shelves and which ones quite caught people's eyes.

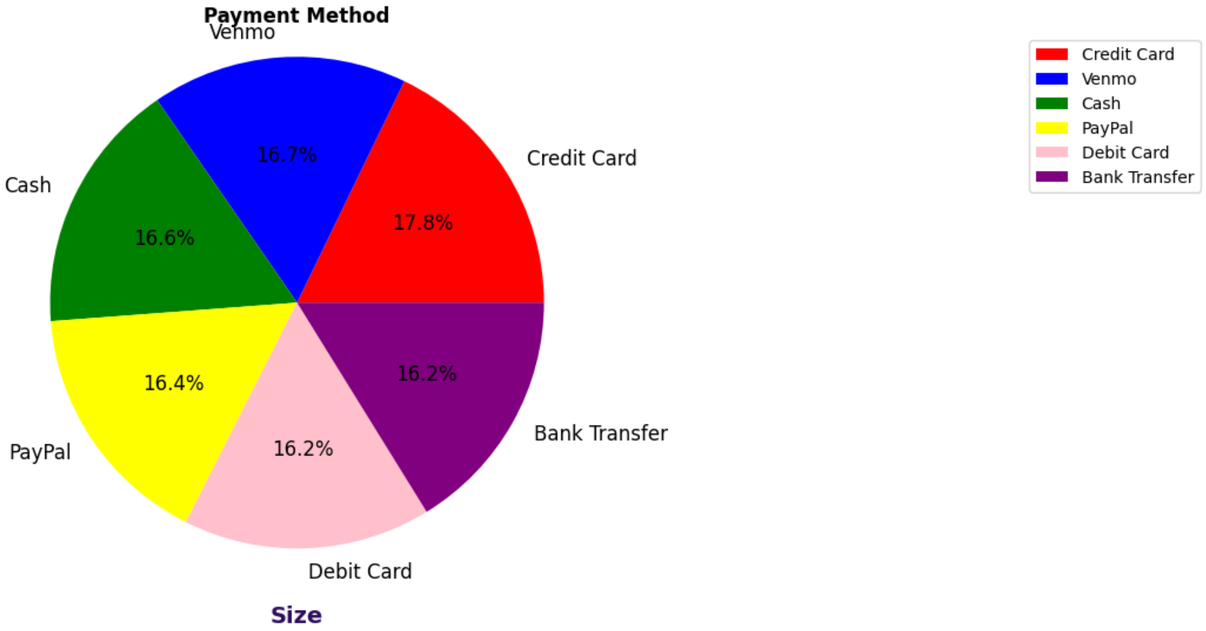


FIG. 8 Payment Method

This pie chart breaks down customer preferences for payment methods. Venmo stands out as the top choice, followed closely by credit cards. On the flip side, cash is the less favoured option, accounting for only 16.6% of customer preferences. The chart provides a clear snapshot of the payment landscape, showcasing the popularity of digital options like Venmo and credit cards, while emphasising the relatively lower preference for cash transactions. This insight can guide businesses in tailoring their payment processes to align with customer preferences and enhance overall satisfaction.

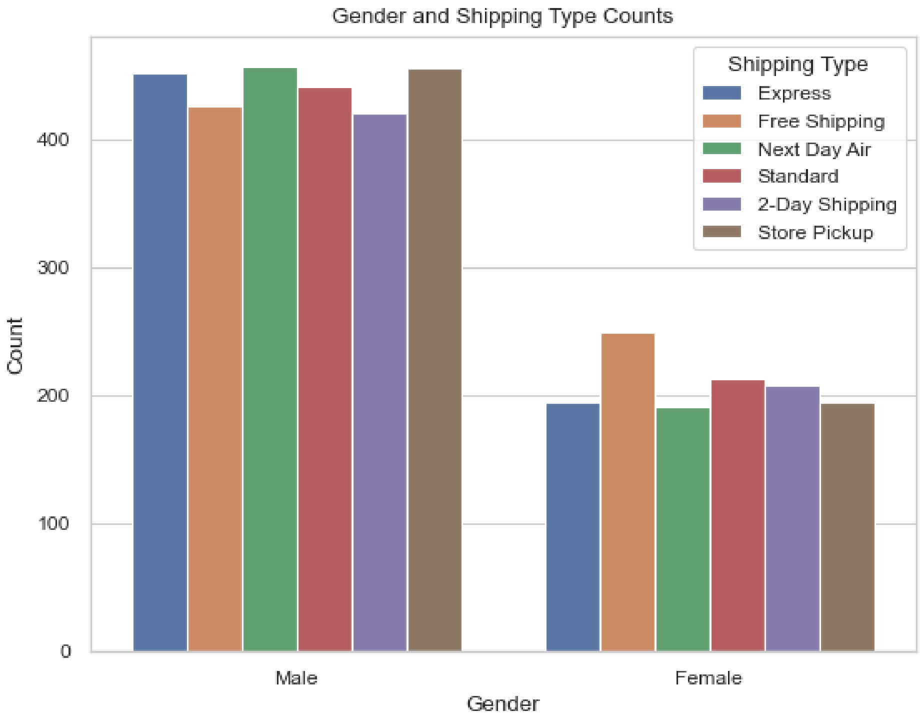


FIG. 9 Gender and Shipping Type Counts

The bar graph displays how different shipping options are preferred by customers based on their gender. There are six shipping methods: Express, free shipping, next-day air, standard, two-day shipping, and store pickup. What stands out is that guys seem to lean towards more purchases compared to gals. It's interesting to see that across these shipping choices, males seem to have a higher preference, indicating they might be more active in selecting and utilising these shipping methods than females.

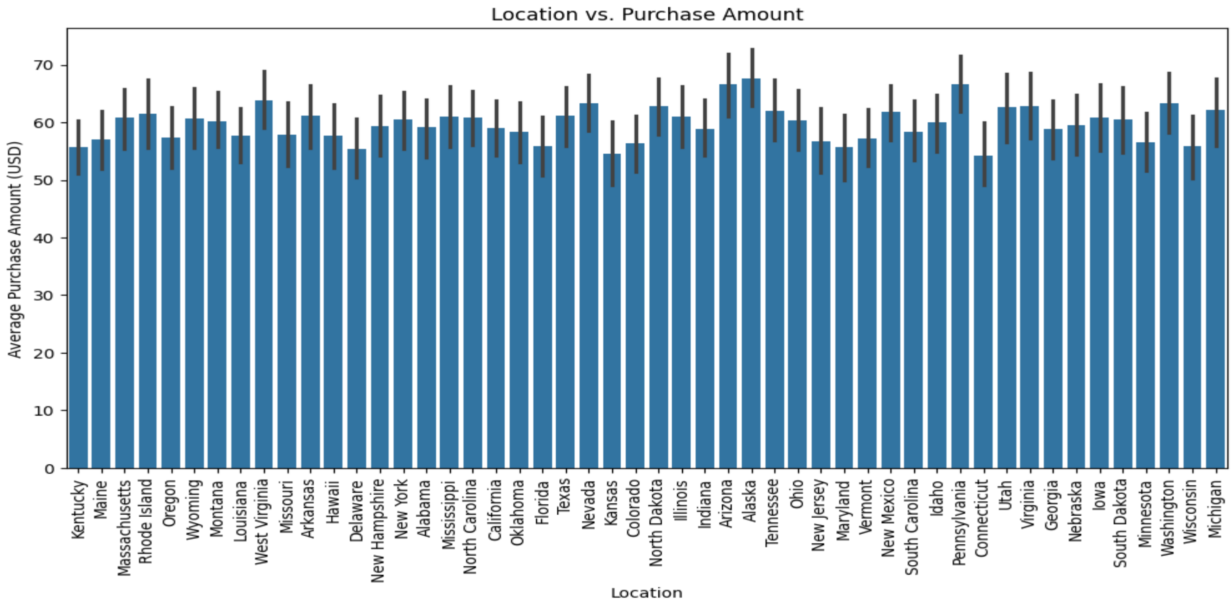


FIG. 10 Location vs Purchase Amount

This bar graph showcases how much customers spend on purchases in various locations. It highlights that in places like Arizona and Alaska, customers tend to spend the most on average for their purchases. Meanwhile, Connecticut stands out with the lowest average purchase amount among the locations studied. Understanding these spending patterns across different locations can help businesses tailor their strategies to better serve customers in those areas, ensuring they meet their preferences and needs.

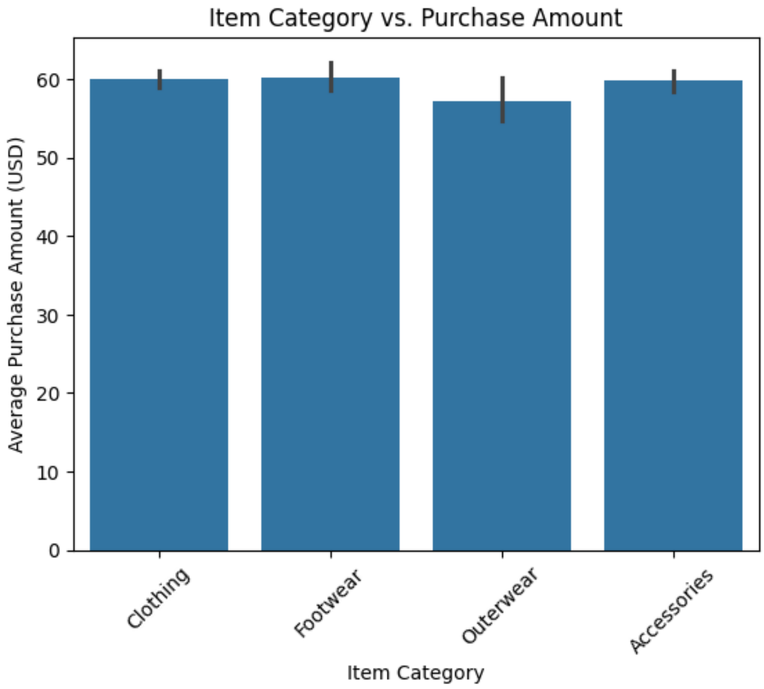


FIG. 11 Item category vs Purchase amount

In this bar graph, we see the connection between item categories and the average money spent by customers. The data reveals that customers typically spend around $60 on clothing, footwear, and accessories. Interestingly, outerwear also hovers around the $60 mark, indicating a similar average purchase amount. The graph offers a clear snapshot of how different item categories correlate with the average spending habits of customers, emphasising the consistent trend around the $60 range for various products.

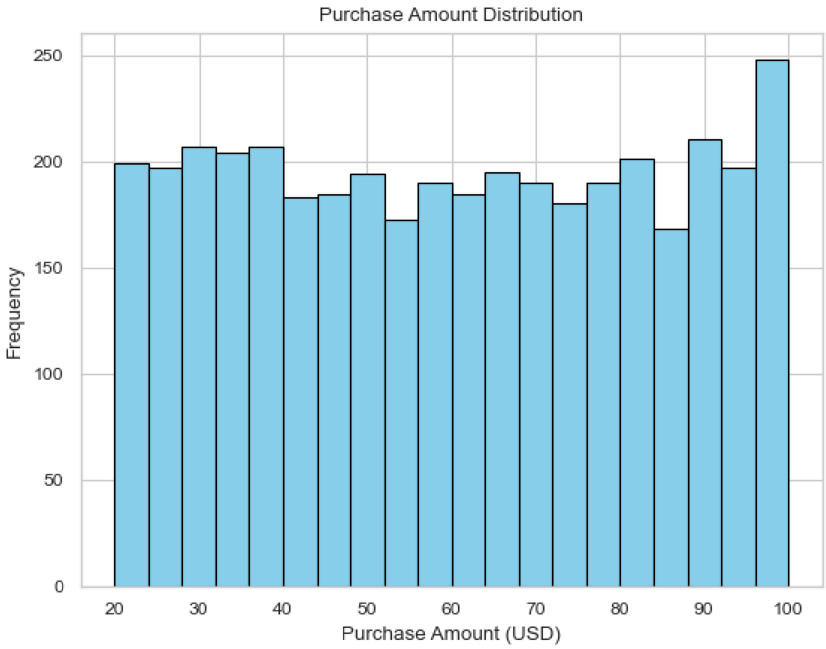


FIG. 12 Purchase amount Distribution

This bar graph shows how much money customers are spending and how frequently. The majority, around 250 customers, are spending $100. The average spending appears to fall between $70 to $100. The graph gives us a quick snapshot of the distribution of purchase amounts, indicating a concentration around the $100 mark. This suggests that a significant portion of customers tend to spend within this range, while fewer customers spend more or less than that amount.

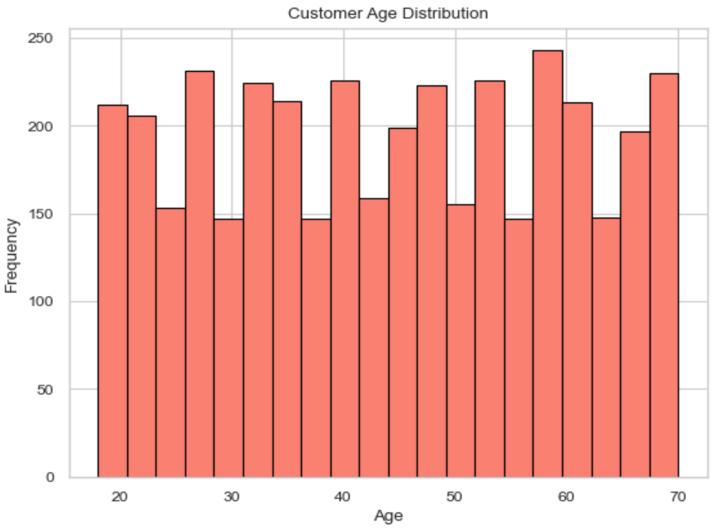


FIG. 13 Customer age distribution

This bar plot showcases the age distribution of customers in the dataset. The majority of customers fall within the age range of 20 to 70. Interestingly, those in their 10s exhibit the highest frequency of visits among all age groups. The graph suggests a diverse age representation, with a concentration in the 20 to 70 range. This implies a broad appeal of the products or services to a wide demographic. Understanding this age distribution can guide marketing strategies and product offerings to cater to the varied preferences and needs of customers across different age brackets.

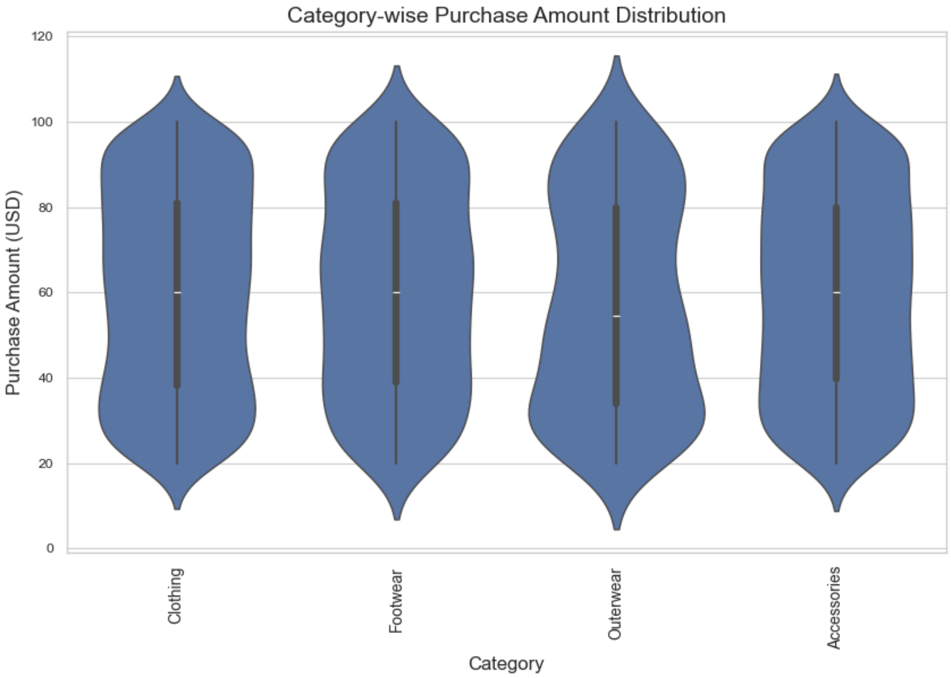


FIG. 14 Category-wise Purchase Amount Distribution

We distributed the purchase amount across various categories, as showcased by this violin graph. At the $100 mark, clothing emerges as the most popular choice, evident from the widened section of the graph. On the other hand, outerwear finds favour among buyers in the $20 range, indicating a common price point for such items. The visualisation provides insights into customer preferences, revealing that a significant number opt for clothing at higher price points, while outerwear attracts more purchases at a more affordable range around $20. This analysis helps understand the spending patterns and popular price ranges within each category.

SMART Questions

Question 1: Which factors (such as customer gender and location) have the most significant impact on the purchase amount?

Features Used:

T-Test:

A t-test is a statistical tool for determining if there is a significant difference in the means of two groups. It's especially useful when dealing with tiny sample numbers or an unknown population standard deviation.

T-tests are classified into two types: independent samples t-tests and paired samples t-tests. The independent samples t-test compares the means of two distinct and independent groups, such as test results from two separate courses. The paired samples t-test, on the other hand, evaluates the means of two related groups, similar to before-and-after measurements in an experiment on the same individuals.

The t-test computes a t-value based on the difference in group means and the variability within groups. A higher t-value implies that the difference between groups is less likely to be caused by chance.

The t-value is then compared to a critical value from a t-distribution, taking into account the degrees of freedom and the significance threshold (typically 0.05). If the estimated t-value is more than the crucial value, it indicates that the groups differ significantly.

Overall, t-tests assist researchers and analysts in determining whether reported differences between groups are actual or the product of random fluctuation in the data.

Linear Regression:

Linear regression is a fundamental statistical technique used for modeling the relationship between a dependent variable and one or more independent variables. It assumes that there's a linear relationship between the predictor variables (also called features or independent variables) and the outcome variable (also known as the dependent variable).

The model takes the form of a linear equation:

$$y = mx + c$$

Where:

y is the dependent variable (the variable you want to predict).

x is the independent variable (the variable used to make predictions).

m is the slope of the line (how much y changes when x changes).

c is the y-intercept (the value of y when x is 0).

In multiple linear regression, where there are multiple predictors, the equation becomes

$$y = b\_0 + b\_1x\_1 + b\_2x\_2 + \ldots.. + b\_nx\_n$$

Where:

y is still the dependent variable.

x_1, x_2, ..., x_n are the independent variables.

b_0 is the intercept.

b_1, b_2,...., b_n are the coefficients for each independent variable.

The goal of linear regression is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the difference between the predicted and actual values of the dependent variable. We often do this by minimizing the sum of the squared differences between the observed and predicted values (this method is known as ordinary least squares).

We widely used linear regression in various fields for predictive analysis, understanding relationships between variables, and making forecasts based on historical data.

Modelling:

Relationship between "Gender" and "Purchase Amount":

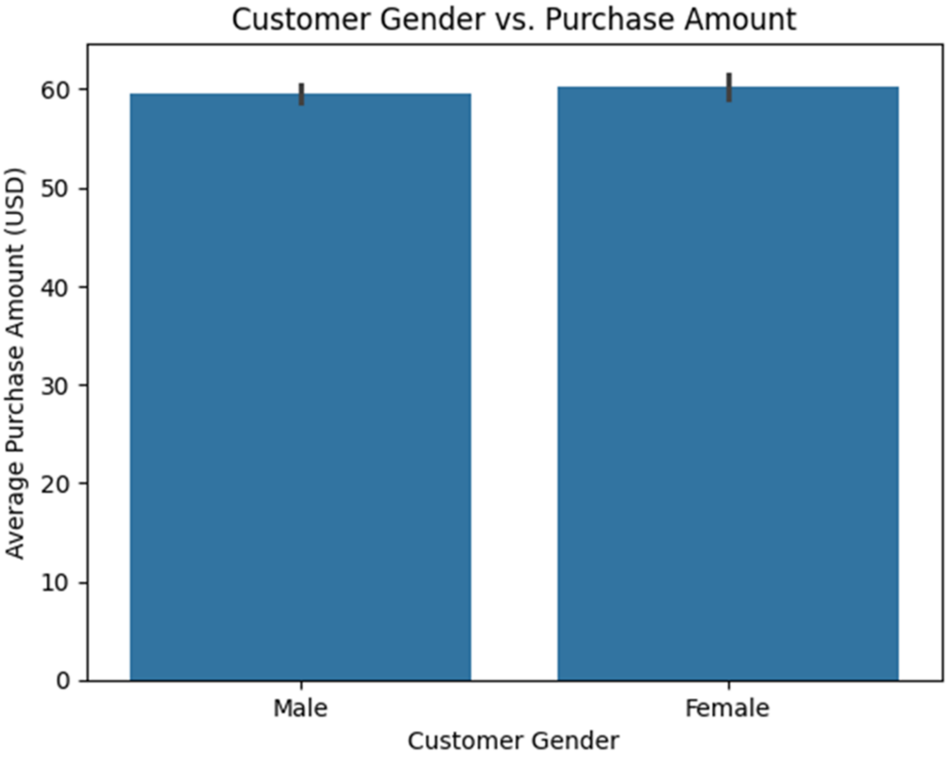
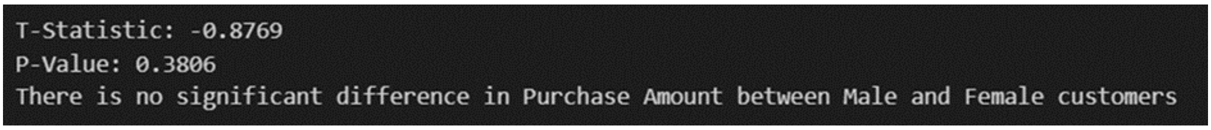


FIG. 15 Customer gender vs Purchase amount

The above bar graph describes the relation between customer gender and the average purchase amount. There is no significant difference in purchase amount between males and females. So from this graph, we got to know that males and females had spent almost equally that is around 60$.



```
T-Statistic: -0.8769
P-Value: 0.3806
There is no significant difference in Purchase Amount between Male and Female customers
```

FIG. 16 T-Test

This is a t-statistic of -0.8769 and a corresponding p-value of 0.3806 when comparing the purchase amounts between male and female customers.

In statistical hypothesis testing, the t-statistic measures the difference between the means of two groups relative to the variation within the groups. The p-value, on the other hand, represents the probability of observing the data (or something more extreme) if the null hypothesis (which typically assumes no difference between groups) were true.

In this case:

The t-statistic of -0.8769 suggests that the difference in purchase amounts between male and female customers is -0.8769 standard deviations.

The p-value of 0.3806 indicates the probability of observing such data if there were no differences in purchase amounts between male and female customers.

With a p-value of 0.3806, which is greater than common significance levels like 0.05 or 0.01, there isn't enough evidence to reject the null hypothesis. Therefore, based on these statistical results, we fail to find a significant difference in purchase amounts between male and female customers.

It's important to note that while these results suggest no significant difference, it's always beneficial to consider the context, the specific dataset, and the practical significance alongside statistical significance when interpreting the findings.

Relationship between "Location" and "Purchase Amount":

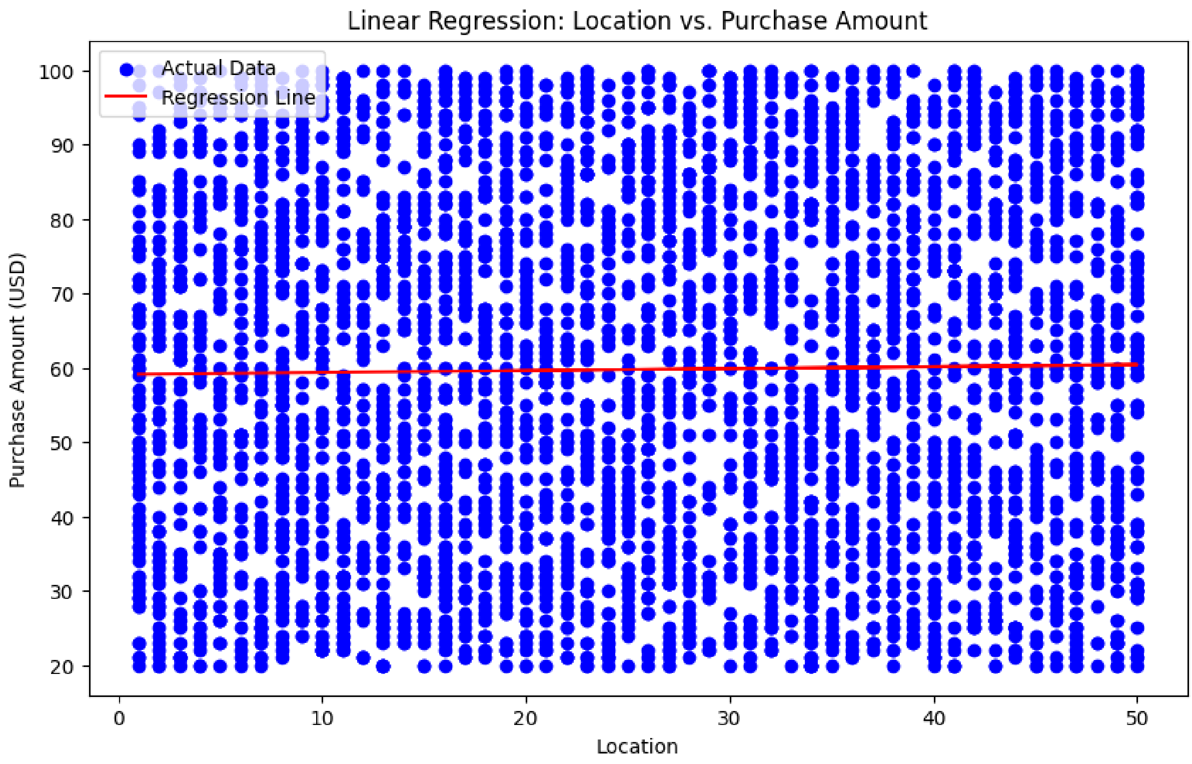
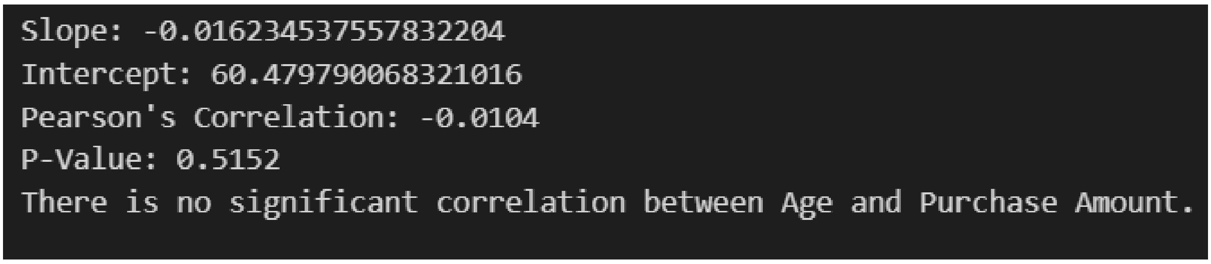


FIG. 17 Linear Regression

By using a linear regression model, we generated a scatter plot for the location and the purchase amount. From the graph, we can observe that the blue dots are actual data and the red line is the regression line. The line on the graph shows how buying changes as people get older. For every extra year someone ages, their spending drops by about 1.6 cents, but it's not a strong link.



```
Slope: -0.016234537557832204
Intercept: 60.479790068321016
Pearson's Correlation: -0.0104
P-Value: 0.5152
There is no significant correlation between Age and Purchase Amount.
```

FIG. 18 Linear Regression Model

Slope: Indicates the rate of change in "Purchase Amount" concerning "Location." Here, the slope is 0.026, suggesting a minimal change in "Purchase Amount" per unit change in "Location."

Intercept: Represents the value of "Purchase Amount" when "Location" is zero. Here, it's 59.097, indicating the expected "Purchase Amount" when "Location" has no impact.

Pearson's Correlation: Measures the linear relationship between variables. A value close to zero (0.0159 in this case) suggests a weak linear relationship between "Location" and "Purchase Amount."

P-Value: Indicates the significance of the correlation. A high p-value (0.3214) suggests that the observed correlation could likely be due to random chance rather than a genuine relationship between "Location" and "Purchase Amount."

Because of these facts, we can infer that there is no substantial relationship between "Location" and "Purchase Amount." The low correlation coefficient and high p-value imply that any observed association is unlikely to be essential or significant.

Question 2: What is the relationship between review ratings and repeat purchases?

Objectives:

The objective of Smart Question 2 is to explore and uncover the connection between review ratings and repeat purchasing behavior. Specifically, this analysis aims to determine whether products or services with higher ratings are more likely to result in repeat purchases, which could indicate higher levels of customer satisfaction. Additionally, the analysis will investigate the impact of negative reviews on repeat purchases in comparison to the effect of positive reviews. This study seeks to understand how customer feedback, as reflected in review ratings, influences their decision to make repeat purchases of the same product or service.

EDA:

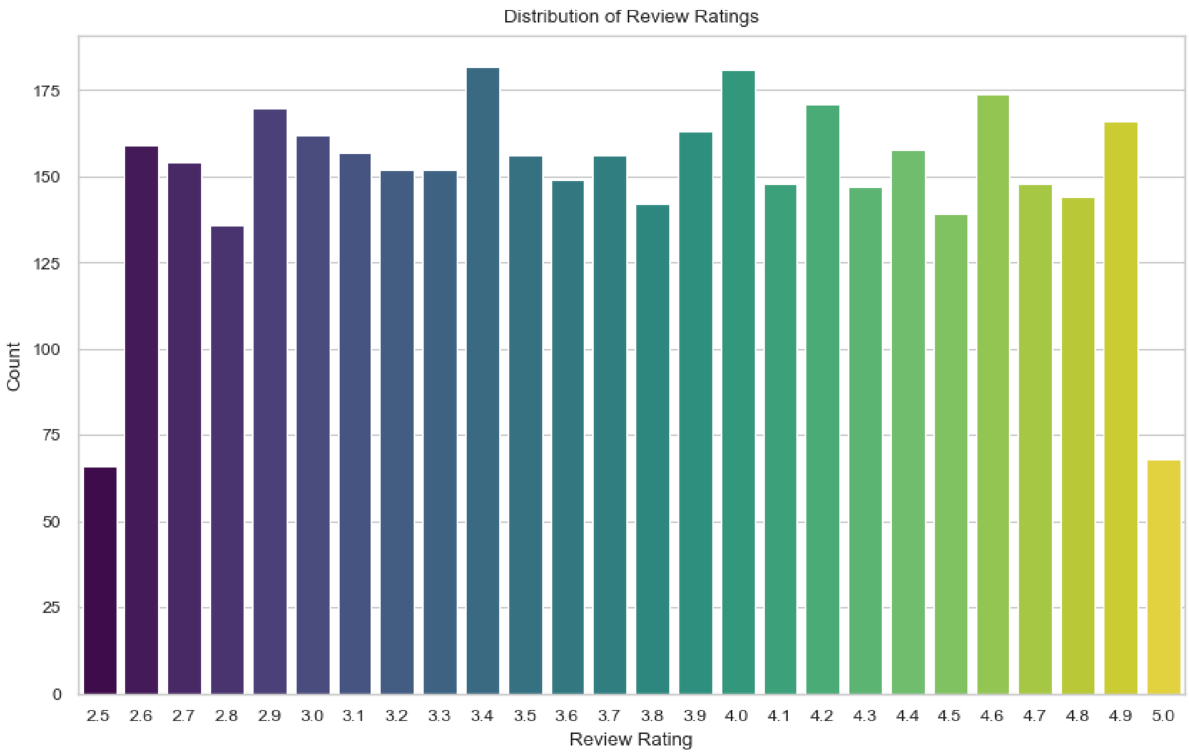


FIG. 19 Distribution of Review Ratings

As we can see from the plot, the distribution of review ratings is almost uniform among all the review groups. This means that our data has a good mix of review ratings from 2.5 to 5. The lowest groups among them (2.5 and 5) appear to have the fewest products associated with them.

Now, let's also visualise the "Previous Purchases" column in our dataset.

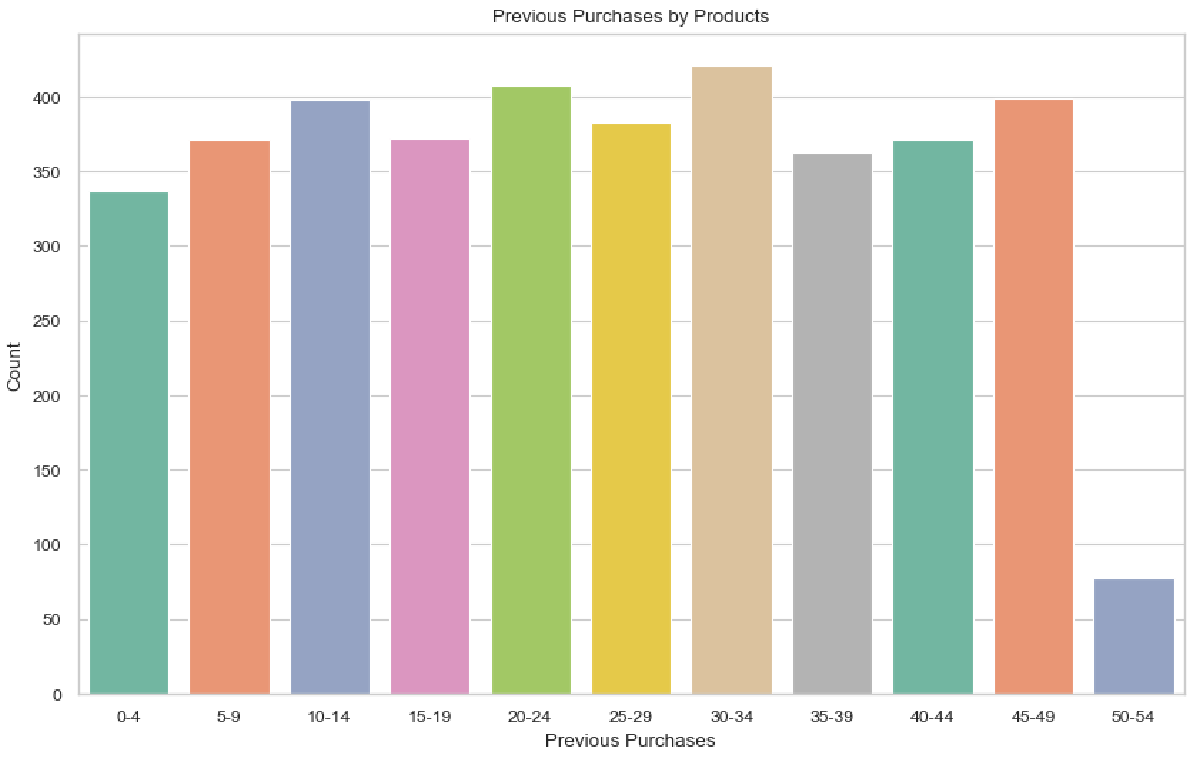


FIG. 20 Previous Purchases by Products

In the above plot, the y-axis (Count) is the number of products and the x-axis (Previous Purchases) means the number of times the product was purchased previously. We can see that the count of products that are re-purchased within 0-49 times is similar for each group. Only a small number of products are re-purchased more than 50 times.

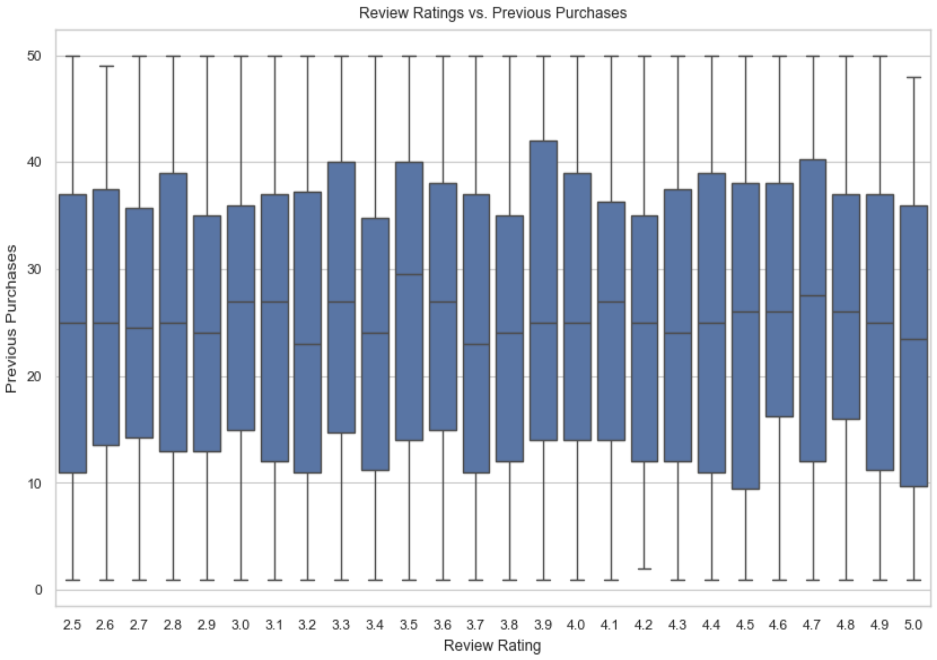


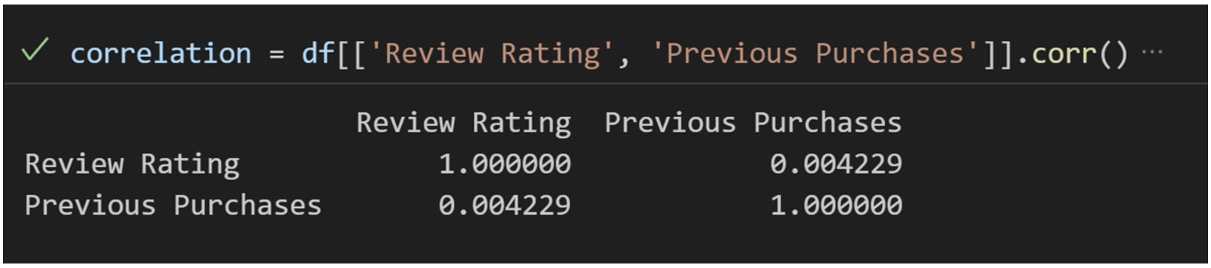
FIG. 21 Review Rating vs Previous Purchases

Contrary to our expectations, the data above reveals higher ratings don't necessarily correlate with repeat purchases. Products with a “3.9” review rating show the highest previous purchases.

STATISTICAL TEST:

CORRELATION MATRIX

Building on the objectives, we will use a statistical correlation matrix to further our analysis. This tool will help us quantify the linear relationship between 'Review Rating' and 'Repeat Purchases'. By examining this correlation, we can gain insights into how strongly these two variables are related. If the correlation is high, it suggests that higher review ratings might indeed lead to more repeat purchases. Conversely, a low or negative correlation would indicate a weaker or inverse relationship.



```
✓ correlation = df[['Review Rating', 'Previous Purchases']].corr() ···
```

```
                    Review Rating  Previous Purchases
Review Rating            1.000000            0.004229
Previous Purchases       0.004229            1.000000
```

FIG. 22 Correlation Matrix - Statistical Test

Since the correlation coefficient (0.004229) between 'Review Rating' and 'Previous Purchases' is close to 0, it means that there is virtually no linear relationship. In other words, review ratings don't appear to linearly predict previous purchase numbers.

Model Building

Linear Regression Model

Since the statistical test did not reveal much about the relation between the two variables, we now turn towards a linear regression model.

Here's a brief outline of the process we'll follow:

- Data Preparation: We start by organising our data. 'Review Rating' is set as the predictor (X), and 'Previous Purchases' is the target variable (y).

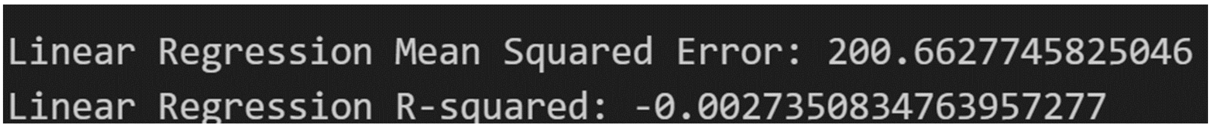
- Data Splitting: The data is then divided into training and testing sets.

- Model Building: We construct the Linear Regression Model using the training data.

- Model Prediction and Evaluation: After training, the model makes predictions on the test set. We evaluate its performance using two key metrics:

- Mean Squared Error (MSE): Measures the average squared difference between the predicted values and actual values, giving us insight into the model's accuracy.

- R-squared (R^2): Indicates the proportion of the variance in the dependent variable that is predictable from the independent variable.



```
Linear Regression Mean Squared Error: 200.6627745825046
Linear Regression R-squared: -0.0027350834763957277
```

FIG. 23 Linear Regression

The results from our linear regression model show a high Mean Squared Error (MSE) of 200 and a very low R-squared value of -0.00274. These outcomes indicate that the model is not effective in accurately predicting repeat purchases based on review ratings. The high MSE means the model's predictions are often far from the actual values. Furthermore, the negative R-squared value suggests there's little to no correlation between review ratings and repeat purchases, as per our dataset.

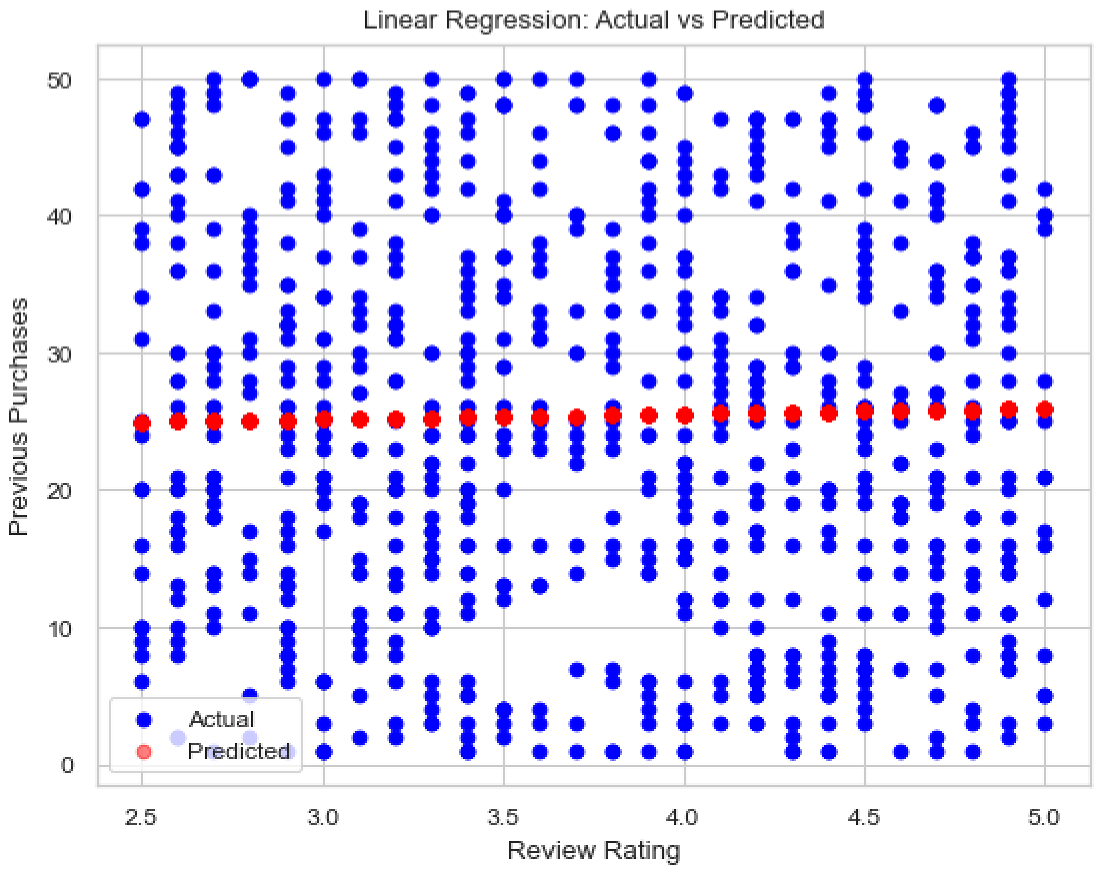
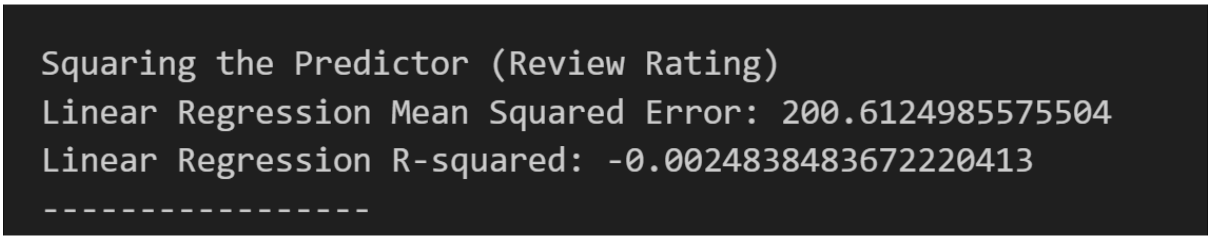


FIG. 24 Linear Regression: Actual vs Predicted

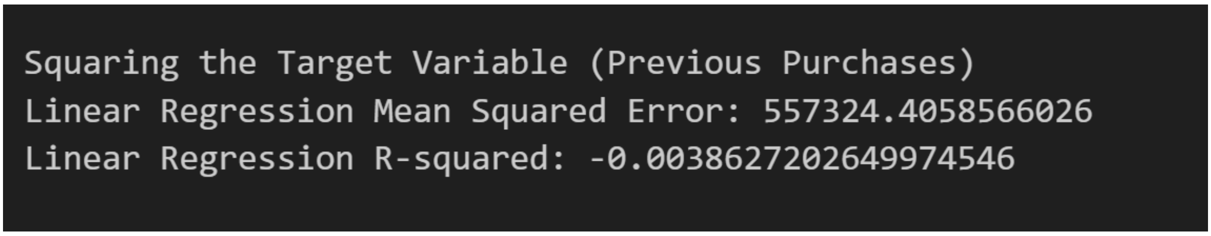
The scatter plot of actual versus predicted data shows a clear discrepancy. Actual data points are scattered widely, while the predicted values form a straight line, suggesting a constant prediction trend regardless of review ratings. This mismatch might be highlighting the linear model's limitations.

Now, let's try squaring the variables and trying the Linear model again.



```
Squaring the Predictor (Review Rating)
Linear Regression Mean Squared Error: 200.6124985575504
Linear Regression R-squared: -0.002483848367222041З
-----------------
```

FIG. 25 Linear Regression: Squaring the Predictor



```
Squaring the Target Variable (Previous Purchases)
Linear Regression Mean Squared Error: 557324.4058566026
Linear Regression R-squared: -0.0038627202649974546
```

FIG. 26 Linear Regression: Squaring the Target Variable

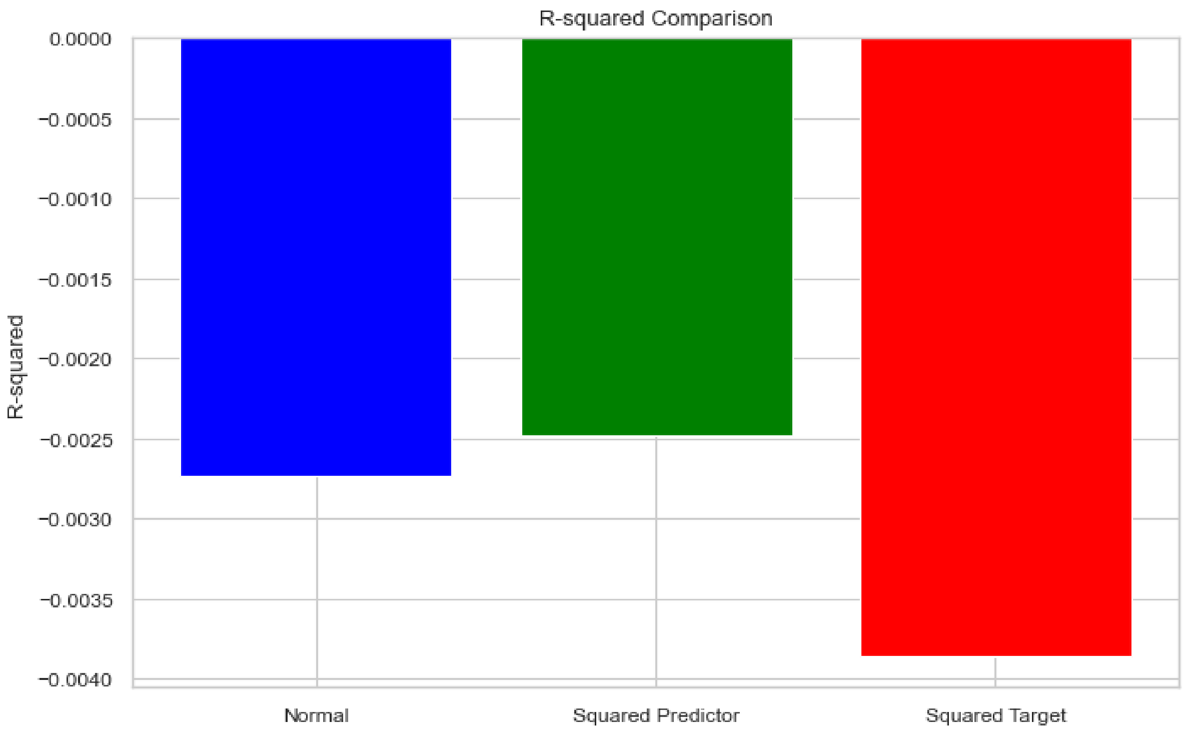


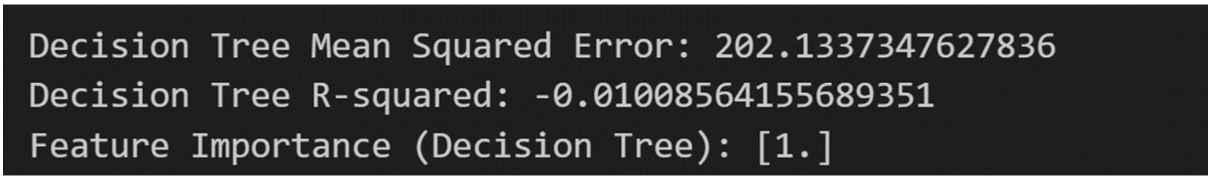
FIG. 27 Linear Regression: Comparing the R-squared values

For the first approach of squaring the predictor, the R-squared value is -0.0025.
This indicates a poor fit. A negative R-squared suggests that the model is worse than a horizontal line fit to the data.

For the second approach of squaring the target variable, the R-squared value is -0.0039.
These values are significantly worse than the first approach, indicating an even poorer model fit.

Since the Linear regression model wasn't able to accurately fit the data, we're now exploring a Decision Tree Regressor as an alternative model, hoping it better captures the varied patterns observed in the actual data.

Decision Tree regressor.



```
Decision Tree Mean Squared Error: 202.1337347627836
Decision Tree R-squared: -0.01008564155689351
Feature Importance (Decision Tree): [1.]
```

FIG. 28 Decision Tree Regressor

The Decision Tree Regressor yielded a Mean Squared Error (MSE) of 202, closely matching the Linear Model's performance. This similarity suggests limited improvement in predicting repeat purchases based on review ratings, with no notable enhancements in prediction accuracy.

Conclusion

In conclusion, our analysis using Linear Regression and Decision Tree regression models on a GPT-generated dataset revealed a minimal relationship between review ratings and repeat purchases. The high Mean Squared Error and low R-squared values in both models indicate a lack of predictive accuracy. This suggests that synthetic data, like that generated by GPT, may not align with real-world dynamics. Therefore, for more reliable insights into customer behavior, especially regarding the influence of review ratings on purchasing decisions, it's crucial to use real-world data. This approach will likely provide a more accurate and nuanced understanding of consumer behavior.

Question 3: Correlation between the Purchase Amount & customer's Repurchase Behaviour and the Product Category

Linear Regression:

We select a subset of potential predictor variables for the regression model We will include both numeric ('Age', 'Review Rating', 'Previous Purchases') and categorical variables('Gender', 'Category', 'Season', 'Subscription Status', 'Discount Applied', 'Promo Code Used') and will use one-hot encoding for the categorical variables.

Outputs:

MSE	R Square
561.8439280176617	-0.004041507268243372

The high Mean Squared Error (MSE) in the model indicates a significant deviation between the predicted and actual purchase amounts, suggesting a poor model fit. Additionally, the R^2 value is negative, implying that the model fails to effectively predict purchase amounts and is not better than a simplistic mean-based model. These outcomes suggest that the model's features lack a strong linear relationship with the 'Purchase Amount (USD)' and exploring non-linear models or incorporating additional factors might improve predictive accuracy.

Non-Linear Regression:

We chose the Random Forest model:

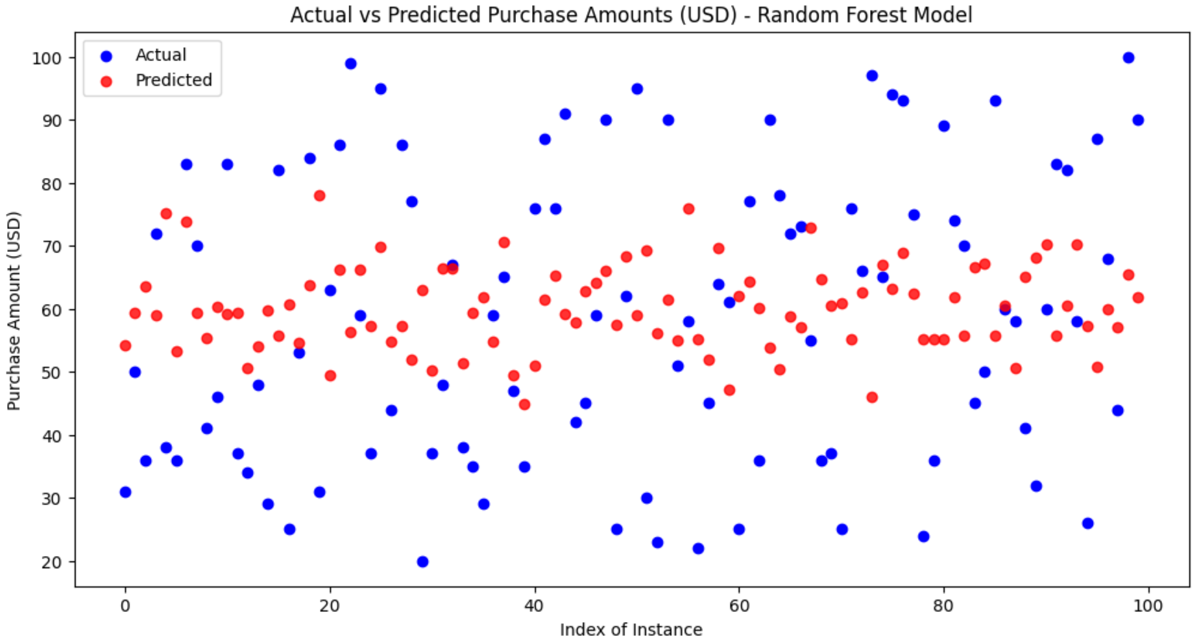


FIG. 29 Actual vs Predicted

In this plot, we only present 100 instances. The scatter plot shows a mismatch between the actual purchase amounts (blue points) and the model's predictions (red points), reflecting the poor model fit indicated by the negative R^2 value. This suggests the need for model adjustments or exploring different predictive approaches.

Because the output is not perfect, we try to square the target variable.

First, we try a linear regression model.

Output:

MSE	Rmse	R Square
21340474.34382176	4619.575125898675	-24.40616591062593

The MSE and RMSE values are quite high, which indicates a significant deviation between the predicted and actual squared purchase amounts. The RMSE provides a more interpretable value, as it is on the same scale as the original target variable (once you take the square root of the predictions). The R^2 score is substantially negative, this negative value

Then we try the Random Forest model

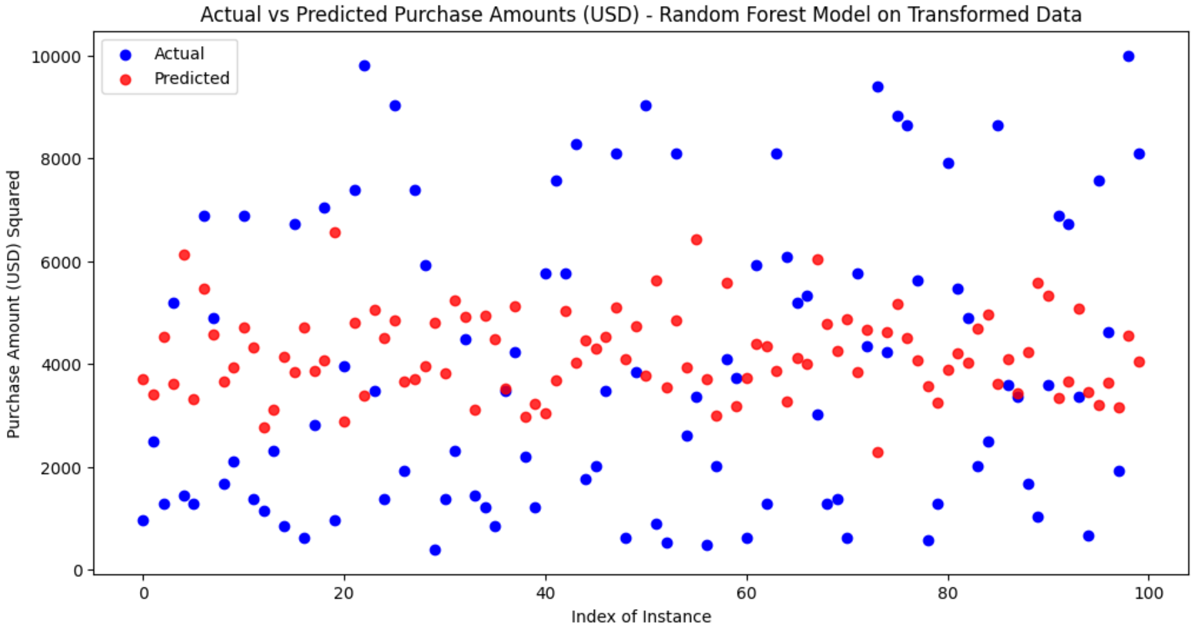


FIG. 30 Random Forest Model

From what is visible in the image, there seems to be a notable spread between the actual and predicted values, indicating that the Random Forest model may not be predicting the squared purchase amounts very accurately. This observation would be consistent with the negative R^2 value previously discussed, which suggests that the model does not explain the variance in the target variable well.

Summary of Results:

Question 1: Which factors (such as customer gender and location) have the most significant impact on the purchase amount?

After a comprehensive analysis, it seems that Age and Gender do not greatly affect how much individuals spend on items. Even after going deep into the data, there's no apparent correlation between these criteria and the amount paid. This implies that other factors may have a greater impact on how much people decide to buy.

Question 2: What is the relationship between review ratings and repeat purchases?

The analysis of both models suggests a weak relationship between 'Review Rating' and 'Previous Purchases', as indicated by the high MSEs and low R-squared values. Despite using advanced models, the predictive accuracy was low.
This challenges usual assumptions, highlighting differences between synthetic and real-world data. It emphasises the need for real-world data to gain accurate, practical insights.

Question 3: Correlation between the Purchase Amount & customer's Repurchase Behaviour and the Product Category

The analysis indicates a weak correlation between 'Review Rating' and 'Previous Purchases', as shown by high Mean Squared Errors and low R-squared values. Despite advanced modelling techniques, this relationship remained poorly defined.
The output highlights the limitations of synthetic data in accurately representing consumer behavior. This emphasises the importance of using real-world data for business analytics

Limitations & Other Considerations:

AI-Generated Data: It's crucial to recognize that the dataset used for this analysis is AI-generated. As a result, there may be disparities between this synthetic data and real-world scenarios. Variables in the dataset are artificially generated and may lack the interdependencies and nuances found in actual retail data.

Independence of Variables: The AI-generated data may exhibit a level of independence among variables, meaning that relationships and correlations observed in this dataset may not align with those in real-world retail environments.

AI Limitations: These limitations underscore the inherent constraints of AI-generated data. While it offers valuable insights, it is not a substitute for real-world data and may not fully capture the complexity of actual retail dynamics.

Conclusion:

According to our findings, "purchase amount" is unrelated to previously defined data items such as prior purchases or reviews. Attempts to relate purchasing habits or ratings to spending behavior have had no obvious effect. These observed spending patterns show fascinating independence from well-studied consumer behavior indexes. The amount spent appears to exist in a vacuum, confounding the conventional connections that firms use to predict consumer spending. Understanding the complexities of consumer decision-making processes makes forecasting or predicting future purchase decisions difficult.

References:

Data set: Customer Shopping Trends Dataset,

Visualisation: https://www.tableau.com/learn/articles/data-visualization,

Modeling: https://www.ibm.com/topics/linear-regression,

T-test: Scribbr - Scriptie laten nakijken,

Random Forest: https://www.ibm.com/.