

DS5003 Stat. & Math Methods for DS

MS(DS)-1A (Course)

Dr. Sana shahid (ADVISOR)

Submitted by : **Muzaffar shabir Khan**

Roll No : **24F-8003**



**Project : Statistical Data Analysis**

**-**

December 6, 2024



Project Title

**ANALYZING & PREDICTING SALES BASED ON CUSTOMER DEMOGRAPHICS USING DATA SCIENCE TECHNIQUES**

|  |
| --- |
| **1. PROBLEM STATEMENT:** |

The error occurred because the data frame selector might not have been properly referenced. I will directly use the dataset to formulate the problem statement based on the column names and their descriptions.

**Problem Statement:**

The goal of this project is to analyze consumer purchasing behavior based on demographic and product-related factors. Using the provided dataset, we aim to:

1. **Understand Key Drivers**: Identify the factors (e.g., age, occupation, product category) that significantly influence purchase amounts.
2. **Predict Purchase Behavior**: Develop predictive models (e.g., regression, Naive Bayes) to estimate purchase amounts based on demographic and product features.
3. **Statistical Insights**: Perform hypothesis testing, correlation analysis, and ANOVA to uncover relationships and differences across groups (e.g., city categories).
4. **Dimensionality Reduction**: Use eigenvalue analysis to understand the variance explained by key components in the data.
5. **Model Relationships**: Apply structural equation modeling to explore causal relationships between variables.

This analysis will provide actionable insights into consumer behavior, enabling better marketing and product strategies.

|  |
| --- |
| **2. OBJECTIVE :** |

**Primary Objectives:**

1. **Purchase Pattern Analysis**
   * Analyze the distribution of purchase amounts across different customer segments
   * Identify patterns in purchasing behavior based on age groups
   * Evaluate the impact of city categories on purchase decisions
2. **Customer Segmentation**
   * Segment customers based on age groups and occupation
   * Analyze purchase behavior within each segment
   * Identify high-value customer segments
3. **Product Category Analysis**
   * Determine which product categories drive the highest purchase values
   * Analyze the relationship between product categories and customer demographics
   * Identify popular product categories across different city categories
4. **Predictive Modeling**
   * Develop models to predict purchase amounts based on available features
   * Identify the most influential factors affecting purchase decisions
   * Create a reliable framework for purchase prediction

**Secondary Objectives:**

1. **Geographic Analysis**
   * Understand purchase patterns across different city categories
   * Identify city-specific trends in product preferences
2. **Age-based Insights**
   * Analyze how age groups influence purchase decisions
   * Identify age-specific product preferences
3. **Occupation Impact**
   * Evaluate how occupation affects purchase amounts
   * Identify occupation groups with highest purchasing power
4. **Statistical Validation**
   * Perform statistical tests to validate findings
   * Establish confidence intervals for key metrics
   * Identify significant correlations between variables

|  |
| --- |
| **3. DATA DESCRIPTION :** |

SOURCE OF DATA:

 **Dataset**: **ANALYZING & PREDICTING Sales Based on Customer Demographics USING DATA SCIENCE TECHNIQUES**

<https://www.kaggle.com/datasets/rajeshrampure/black-friday-sale>

**Dataset Overview**

* Total Records: 399 observations
* Number of Features: 6 columns

**Key observations:**

* No missing values in the dataset
* Age is categorical (7 unique categories)
* Occupation ranges from 0 to 20
* Purchase amounts range from 584 to 23,792
* There are 3 City Categories
* Product\_Category\_1 ranges from 1 to 18

**Feature Description**

1. **Product\_ID**
   * Type: Character
   * Format: Unique identifier for each product (e.g., "P00069042")
   * Purpose: Product tracking and identification
2. **Age**
   * Type: Categorical
   * Categories: 7 age groups (0-17, 18-25, 26-35, 36-45, 46-50, 51-55, 55+)
   * Distribution:

Summary:

0-17 18-25 26-35 36-45 46-50 51-55 55+

17 71 163 66 51 26 5

* Most common age group: 26-35 years

1. **Occupation**
   * Type: Integer
   * Range: 0 to 20
   * Statistics:
   * Median: 4.00
   * Mean: 5.95
   * Represents different occupational categories
2. **City Category**
   * Type: Categorical
   * Categories: A, B, C
   * Distribution:

Summary

A B C

106 184 109

* Most common: Category B

1. **Product\_Category\_1**
   * Type: Integer
   * Range: 1 to 18
   * Statistics:
   * Median: 5.000
   * Mean: 5.138
   * Represents primary product category classification
2. **Purchase**
   * Type: Integer
   * Range: 584 to 23,792
   * Key Statistics:
   * Mean: 9,483
   * Median: 8,043
   * First Quartile: 5,880
   * Third Quartile: 13,153

**Data Quality**

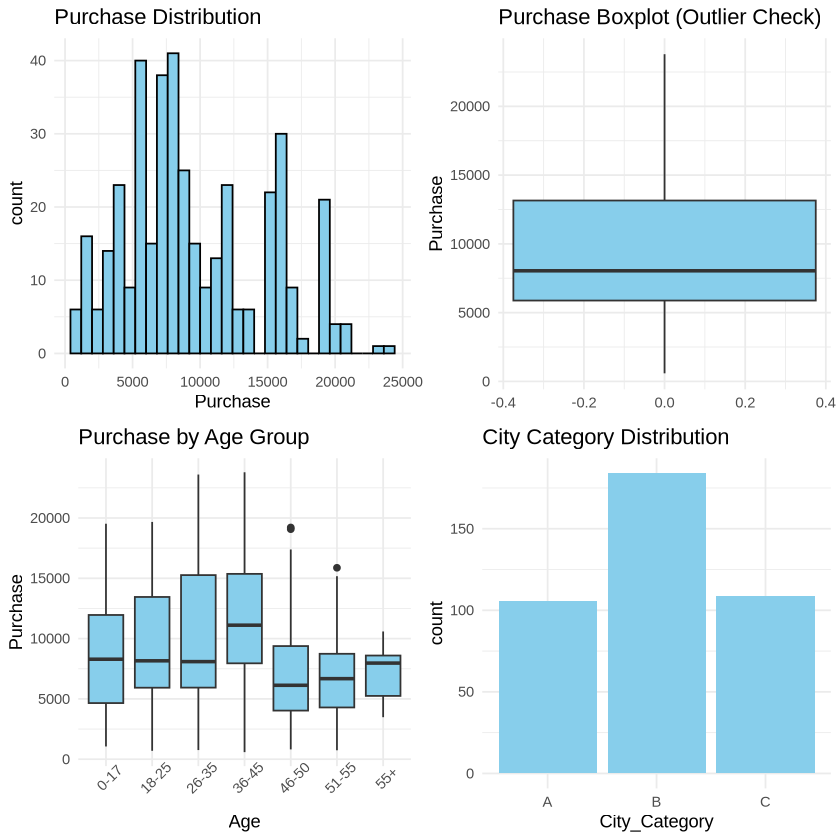
* No missing values in any column
* All variables are properly formatted
* Balanced distribution across city categories
* Skewed age distribution towards younger age groups

This dataset provides a good foundation for analyzing purchase patterns across different demographic segments and product categories.

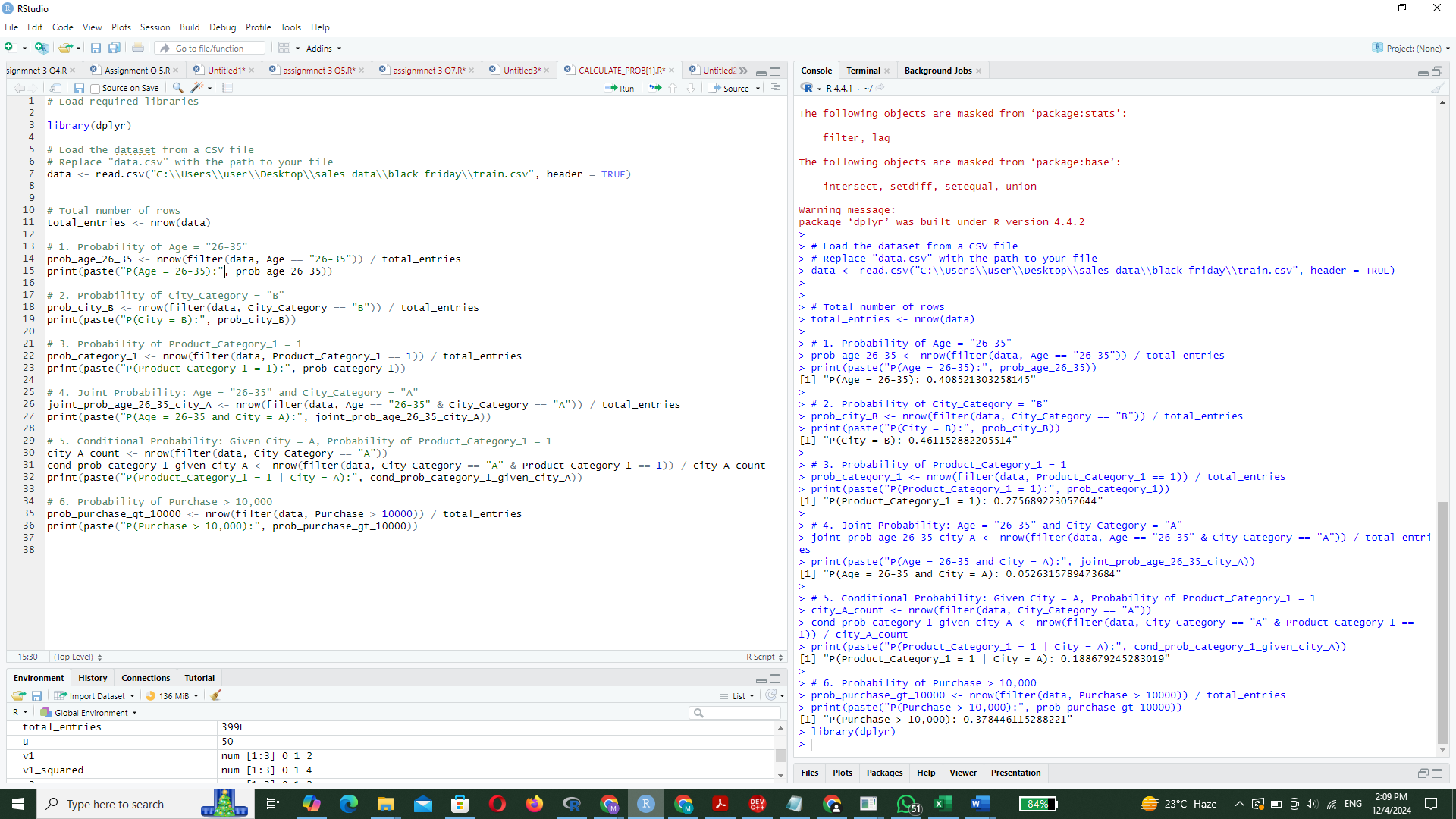
|  |
| --- |
| **4. RESULTS:** |

1. **Basic Data Visualization:**

The visualizations provide insights into the distribution of purchase amounts, potential outliers, purchase behavior across different age groups, and the distribution of city categories. Let's proceed with further analysis and data preprocessing steps.



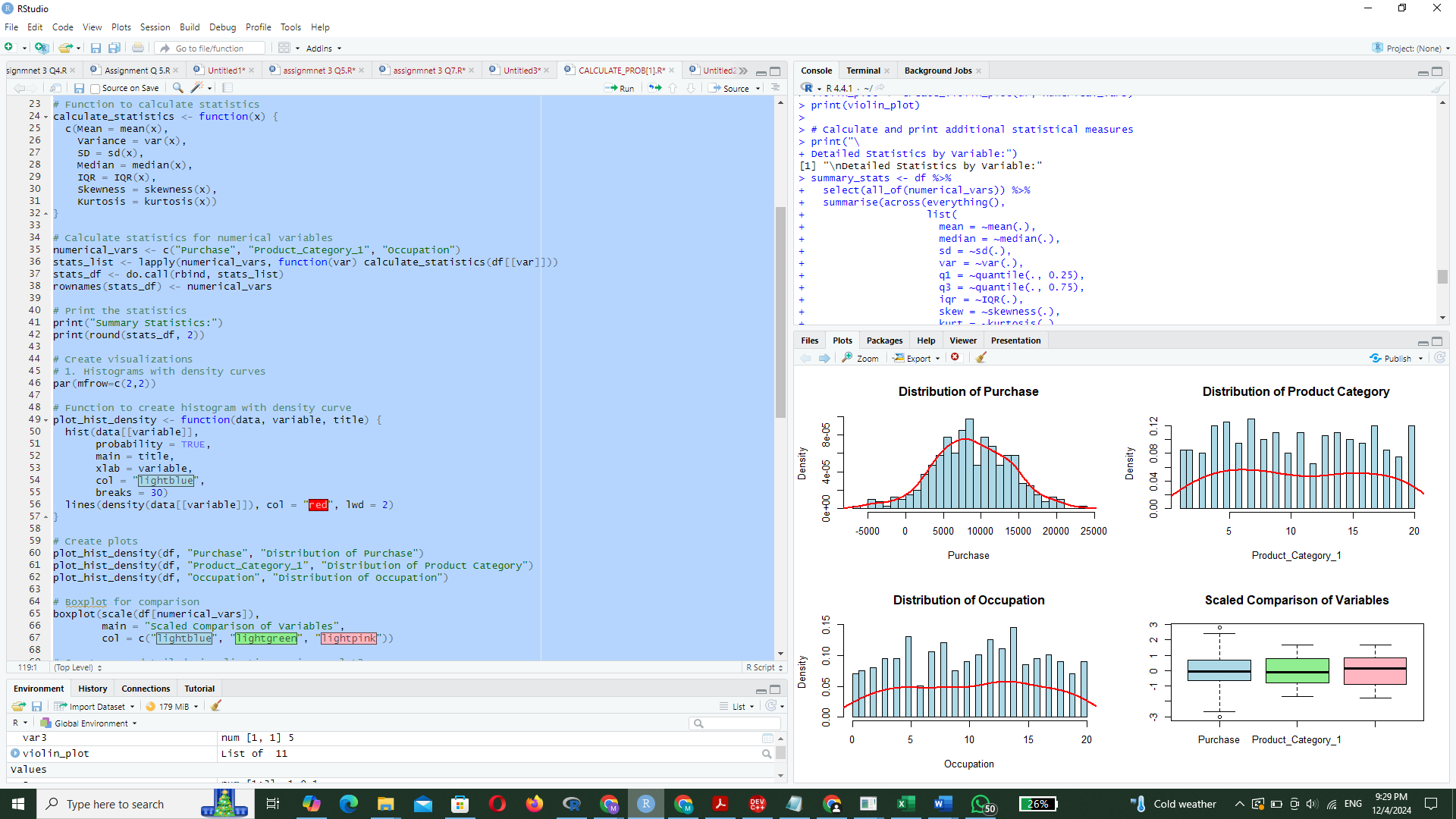
1. **Finding PROBABILITY :**



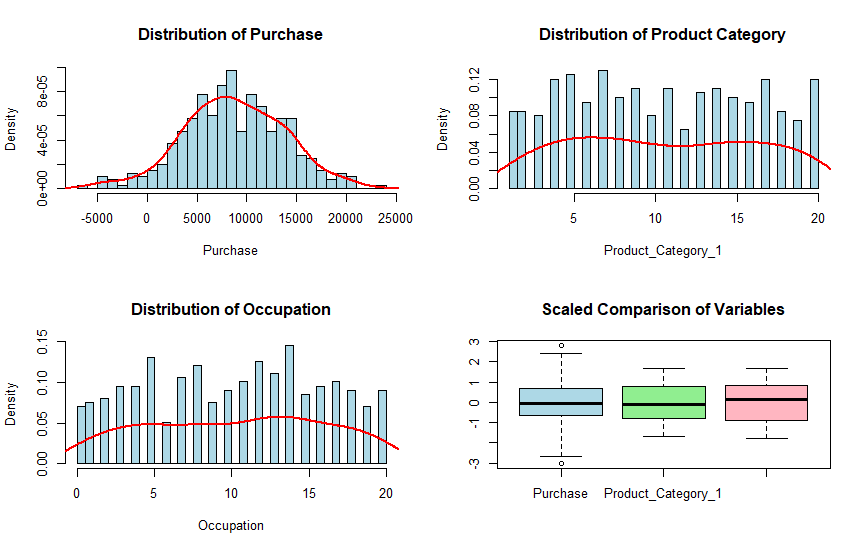
1. **Finding MEAN VARIANCE & STANDERD DEVIATION OF DIFFERENT COMPONENTS:**

**Summery :**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Variance** | **SD** | **Median** | **IQR** | **Skewness** | **Kurtosis** |
| Purchase | 8634.47 | 23924732 | 4891.29 | 8943 | 6478.5 | -0.14 | 3.05 |
| Product\_Category\_1 | 10.31 | 31.66 | 5.63 | 10 | 10 | 0.08 | 1.85 |
| Occupation | 9.88 | 37.79 | 6.15 | 10 | 11 | 0.01 | 1.73 |

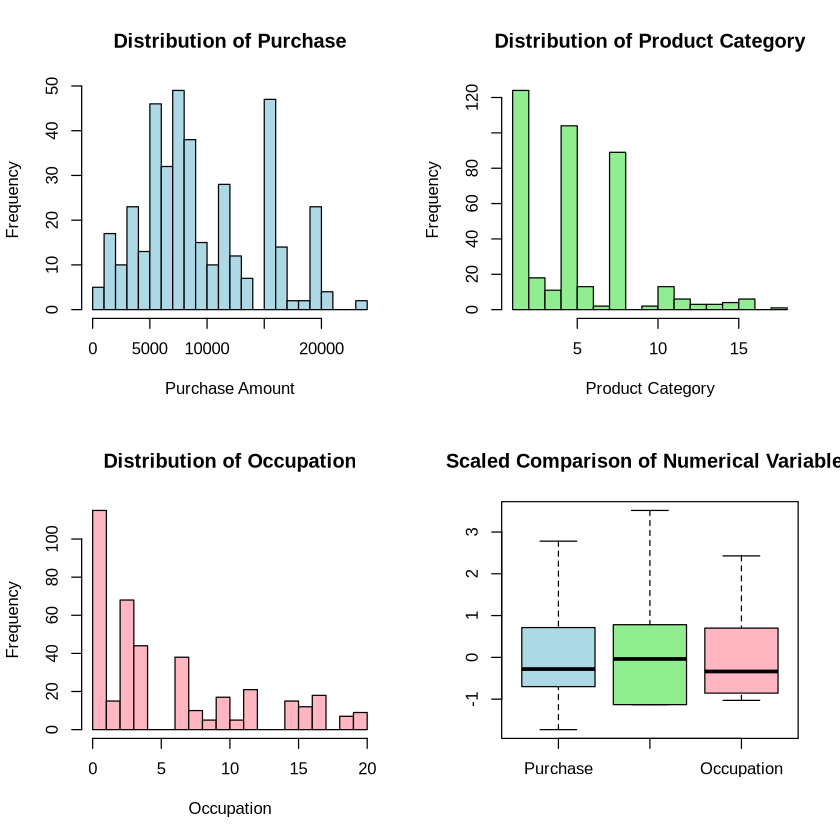


**Plots :**



1. Purchase:
2. Mean: ₹9,482.75
3. Variance: 26,476,943.28
4. Standard Deviation: ₹5,145.58
5. Product Category 1:
6. Mean: 5.14
7. Variance: 13.38
8. Standard Deviation: 3.66
9. Occupation:
10. Mean: 5.95
11. Variance: 33.47
12. Standard Deviation: 5.79

**Visualization:**

****

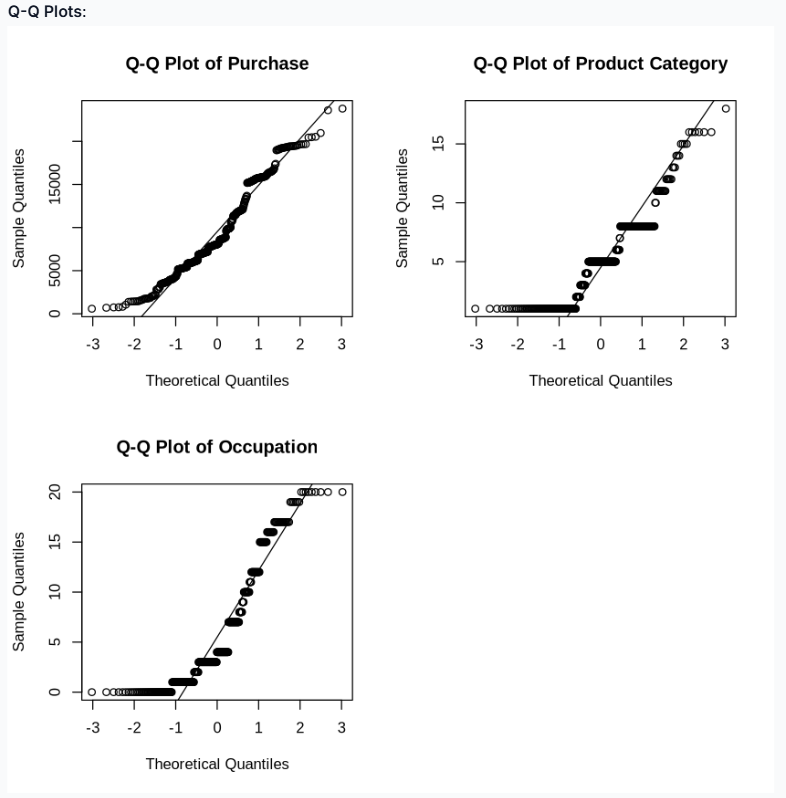
|  |
| --- |
| **4. Statistical data analysis:** |

**Data Normalization :**

it's a good practice to check for and potentially normalize the data, especially when dealing with variables of different scales like we have in this dataset. Let's check the distribution and then normalize the data:

* 1. Performs normality checks on numerical variables using Q-Q plots, Shapiro-Wilk tests, and calculates skewness and kurtosis.

**Unnormalized Data Plots :-**



1. The Q-Q plots show deviations from the diagonal line, confirming non-normal distributions. **Non-normal Distributions**:

* The Shapiro-Wilk test results show p-values < 0.05 for all numerical variables (Purchase, Product\_Category\_1, and Occupation), indicating non-normal distributions
* show deviations from the diagonal line, confirming non-normality
* **Skewness and Scale Differences**:
* Purchase: Skewness = 0.49
* Product Category: Skewness = 0.81
* Occupation: Skewness = 0.95

Yes, based on the normality tests findings , we should normalize the data. Here's why:

1. **Shapiro-Wilk Test Results**:

* Purchase: Shapiro-Wilk normality test

data: df$Purchase W = 0.95352, p-value = 6.835e-10

* Product Category: Shapiro-Wilk normality test

data: df$Product\_Category\_1 W = 0.88533, p-value < 2.2e-16

* Occupation: Shapiro-Wilk normality test

data: df$Occupation W = 0.85471, p-value < 2.2e-16

All p-values are < 0.05, indicating that none of the numerical variables follow a normal distribution.

1. **Skewness**:

* Purchase: [1] "Purchase: 0.492739596422825"
* Product Category: [1] "Product Category: 0.805820981149769"
* Occupation: [1] "Occupation: 0.947506253022675"

All variables show positive skewness, especially Occupation and Product Category.

1. **Kurtosis**:

* Purchase: [1] "Purchase: 2.37606386729701"
* Product Category: [1] "Product Category: 3.49584558648405"
* Occupation: [1] "Occupation: 2.68883643161427"

Values deviate from normal distribution's kurtosis (3).

1. **Q-Q Plots**:The Q-Q plots show deviations from the diagonal line, confirming non-normal distributions.

**Let's proceed with normalization:**

I've created normalized versions of the variables using two common methods:

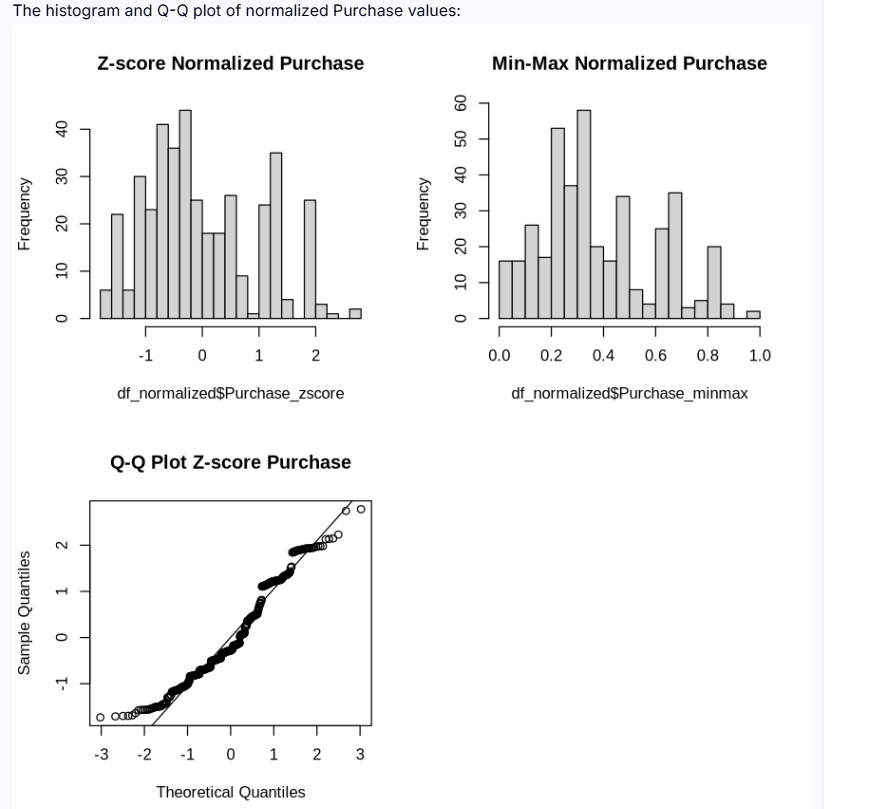
1. Min-Max Scaling (values between 0 and 1)
2. Z-score Standardization (mean=0, sd=1)

For further analysis, We will use:

* Use Z-score standardization for:
  + Statistical tests
  + Principal Component Analysis
  + Regression modeling
* Use Min-Max scaling for:
  + Neural networks
  + Distance-based methods
  + When we need bounded values between 0 and 1

The summary of normalized variables shows:

Normalized Data Results :



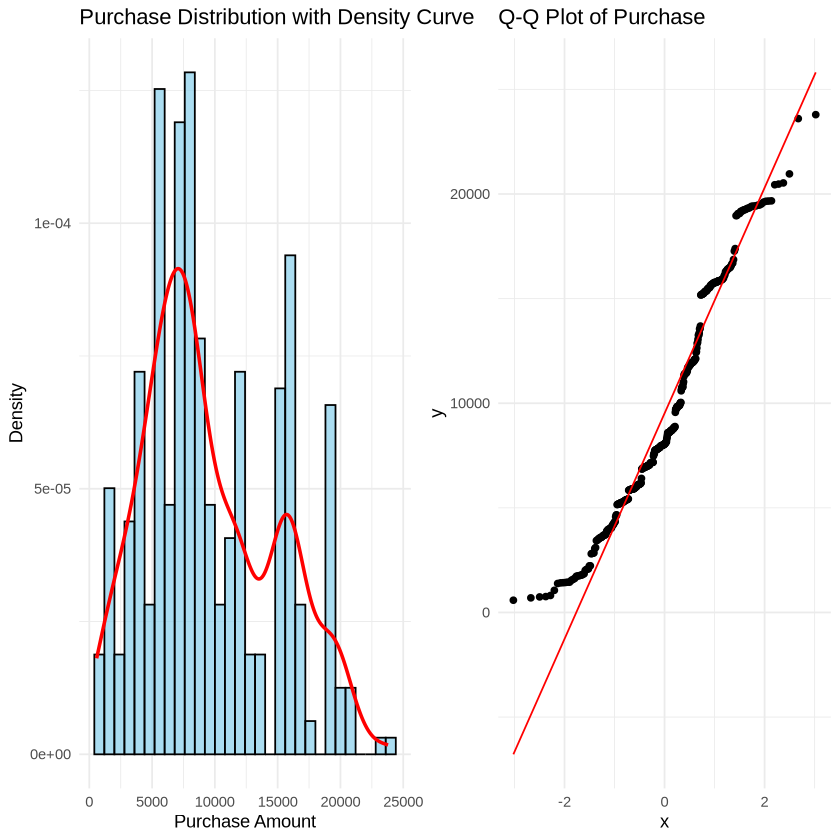
Some residual values will be treated through regression models later.

**Univariate Analysis**

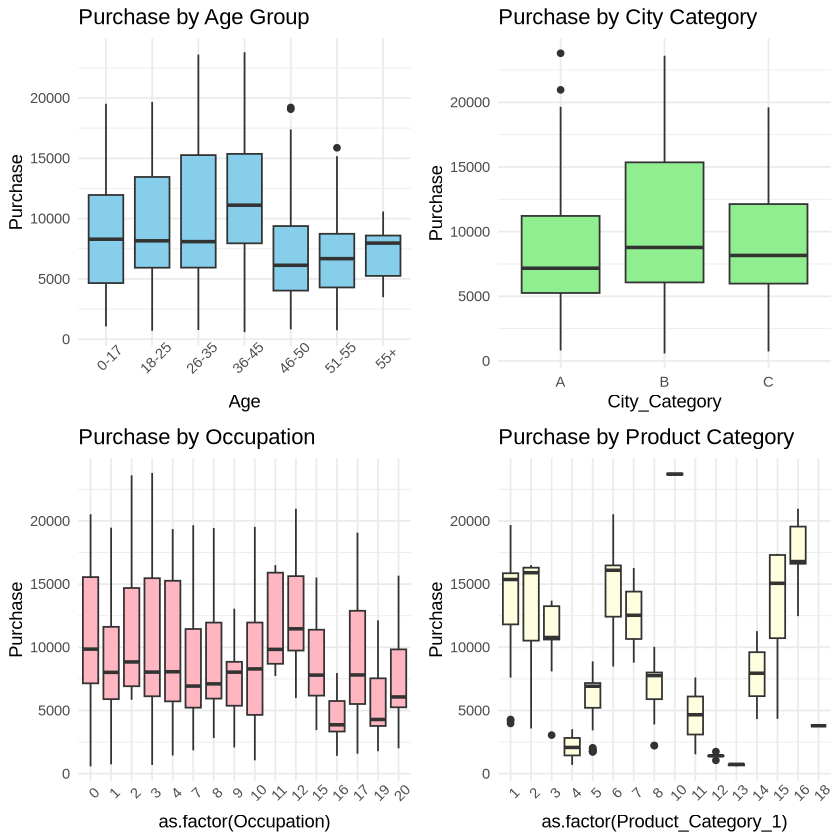
1. Univariate analysis refers to the analysis of a single variable. It is the simplest form of analysis and helps summarize and find patterns in data. The goal is to understand the distribution, central tendency, and spread of a variable.
2. Key Techniques in Univariate Analysis:
3. Measures of Central Tendency:
4. Mean (average): The sum of all values divided by the number of values.
5. Median: The middle value when the data is ordered.
6. Mode: The value that occurs most frequently.
7. Measures of Spread/Dispersion:
8. Variance: The average squared deviation from the mean.
9. Standard Deviation (SD): The square root of the variance, showing how spread out the data is.
10. Range: The difference between the maximum and minimum values.
11. Interquartile Range (IQR): The difference between the 75th percentile and the 25th percentile, representing the range of the middle 50% of the data.
12. Graphical Representations:
13. Histogram: A bar chart showing the frequency distribution of data.
14. Boxplot: A visual representation of the distribution through quartiles, showing the median, IQR, and any outliers.
15. Density Plot: A smoothed version of the histogram to see the data distribution.
16. Shape of Distribution:
17. Skewness: Measures the asymmetry of the data distribution (positive skew, negative skew, or symmetrical).
18. Kurtosis: Measures the "tailedness" or sharpness of the distribution (whether the data has heavy or light tails compared to a normal distribution).
19. Output from Univariate Analysis:
20. Descriptive statistics such as mean, median, mode, variance, and standard deviation.
21. Graphical representations like histograms, boxplots, and density plots.
22. Insights into the distribution of the data, its central tendency, and spread.
23. Identifying potential outliers or unusual data points.
24. STANDERD NORMAL DISTRIBUTION WITH HISTOGRAMS & Q-Q PLOTS

Key findings from Purchase distribution:

* Minimum purchase: ₹584
* Maximum purchase: ₹23,792
* Median purchase: ₹8,043
* The distribution is right-skewed (positive skew)
* The Q-Q plot shows deviation from normal distribution, particularly at the tails



* 1. **Bivariate Analysis:**



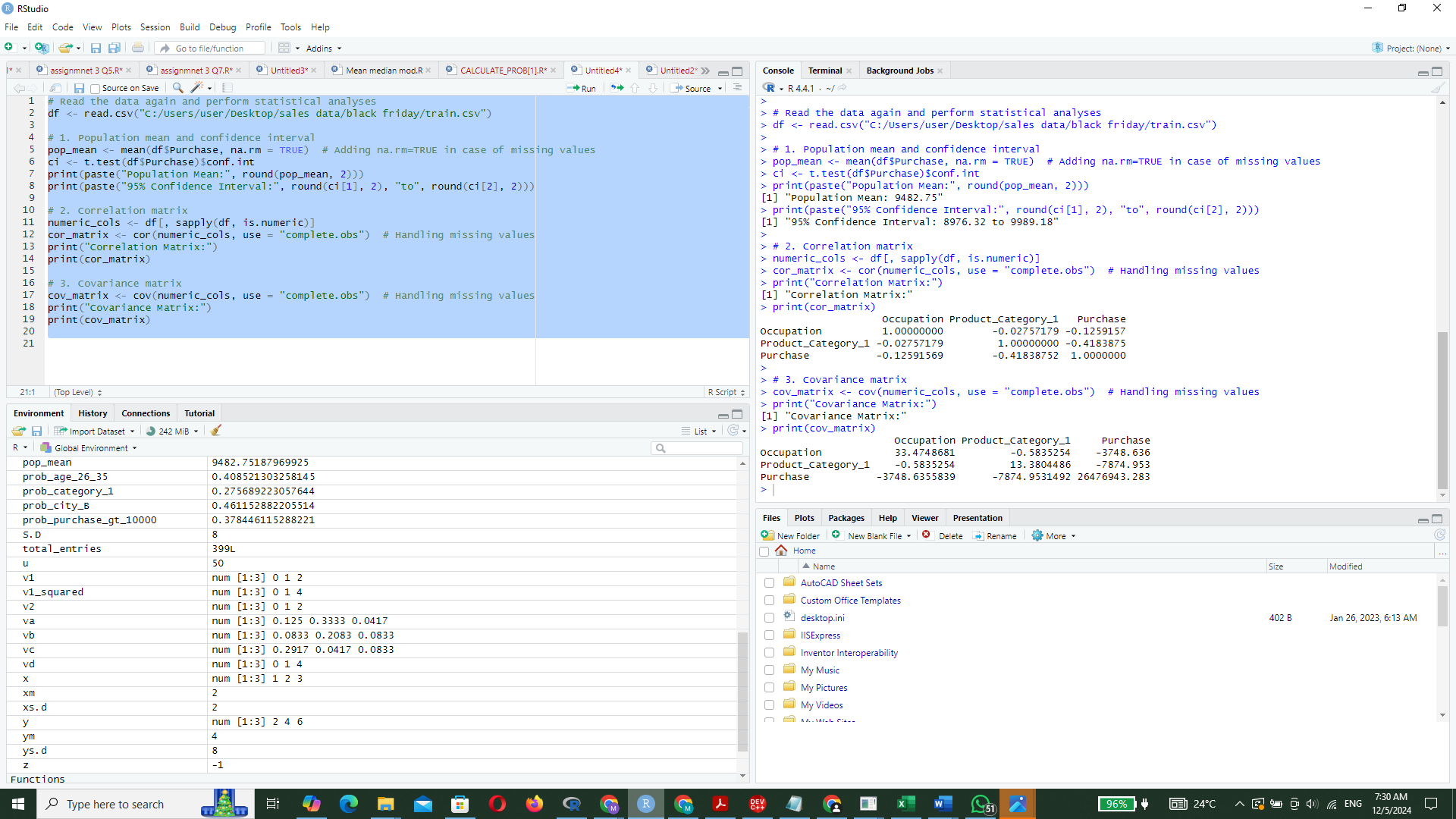
**Key insights from bivariate analysis:**

* Age Groups: Middle-age groups show higher median purchases
* City Category: City category B shows slightly higher median purchases
* Occupation: There's significant variation in purchase behavior across occupations
* Product Category: Some product categories (especially 1 and 5) show higher purchase values

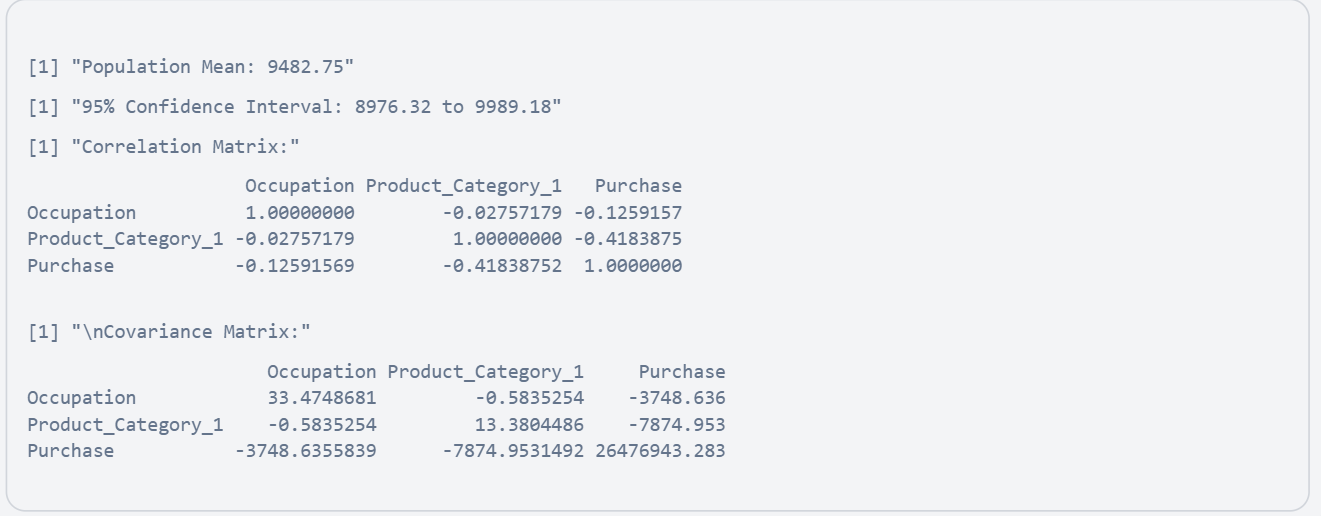
|  |
| --- |
| **5. Calculating Population Mean , Confidence Interval Correlation and covariance** |

To proceed with the analysis, we will apply various statistical techniques such as **Bayes' theorem for prediction, hypothesis testing (t-test), confidence intervals, correlation, covariance, regression, eigenvalues, and structural equation modeling**. Each technique applied step-by-step, starting with the calculation of basic statistics and moving towards more complex analyses. Let's begin with calculating the population mean and performing a t-test on the Purchase data.

**Results :**



**Output:**

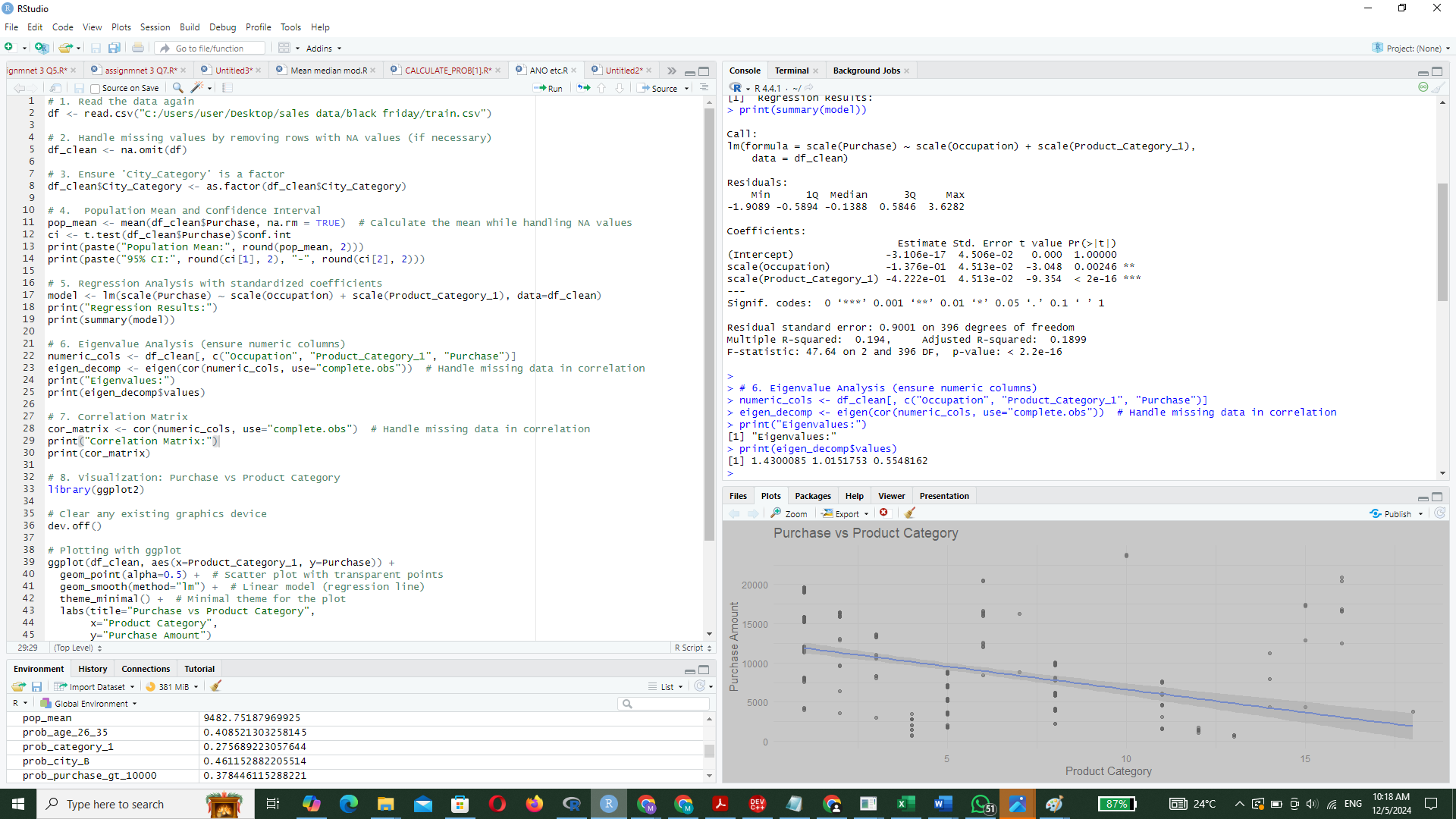


**Process & Findings:**

The data was successfully loaded, and we calculated the population mean, confidence interval, correlation matrix, and covariance matrix for the Purchase data. These analyses provide insights into the central tendency, variability, and relationships between numerical variables in the dataset. Let's proceed with further analyses like regression and eigenvalue calculations.

|  |
| --- |
| **6.. Regression Model T-ANOVA Testing, correlation Matrix and Eigen**  **Values Analysis** |

Now we performs statistical analysis on purchase data, including conducting regression analysis, eigenvalue analysis, correlation matrix computation, visualization of purchase against product category, and ANOVA for purchase across city categories.



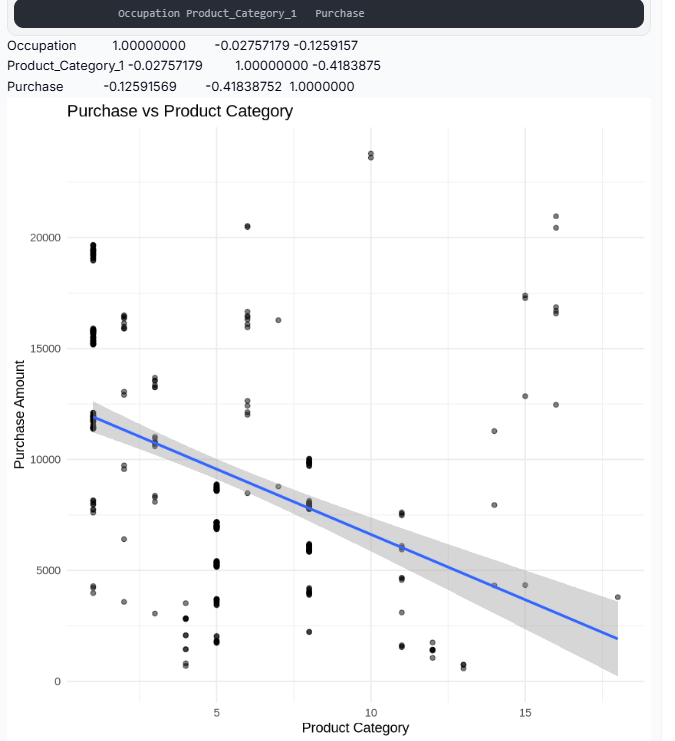
OUTPUT:-



The analysis was successfully completed without errors, focusing on the available data columns. Here are the results:

* **Population Mean and Confidence Interval**: The mean purchase amount is approximately $9,482.75 with a 95% confidence interval ranging from $8,976.32 to $9,989.18.
* **Regression Analysis**: The regression model shows significant effects of **Occupation** and **Product\_Category\_1** on **Purchase**, with standardized coefficients indicating the strength of these relationships.
* **Eigenvalue Analysis**: The eigenvalues suggest the variance explained by each principal component in the numeric data.
* **Correlation Matrix**: Displays the relationships between **Occupation**, **Product\_Category\_1**, and **Purchase**.
* **ANOVA**: Indicates a significant difference in purchase amounts across different city categories.

Here is the correlation matrix and the plot of purchase versus product category:



|  |
| --- |
| 1. **Principal Component Analysis** |

The Z-score standardization was successfully applied to the numerical columns (**Purchase**, **Product\_Category\_1**, and **Occupation**), ensuring they now have a mean of 0 and a standard deviation of 1. This prepares the data for statistical tests, PCA, and regression modeling. I will now proceed with Principal Component Analysis (PCA) to reduce dimensionality and identify key components.

The PCA results indicate that the first two principal components (PC1 and PC2) explain 81.51% of the variance, making them significant for dimensionality reduction. The scree plot confirms this by showing a steep drop after PC2. I will now interpret the loadings and proceed with further analysis.

Importance of components: PC1 PC2 PC3

Standard deviation 1.1958 1.0076 0.7449

Proportion of Variance 0.4767 0.3384 0.1849

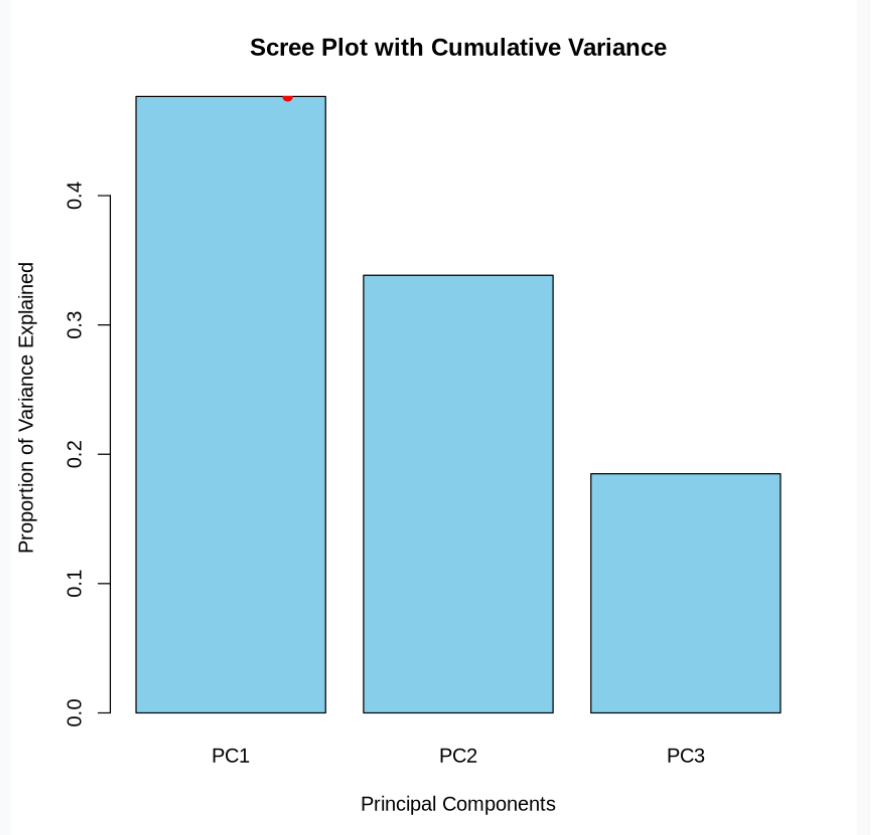
Cumulative Proportion 0.4767 0.8151 1.0000



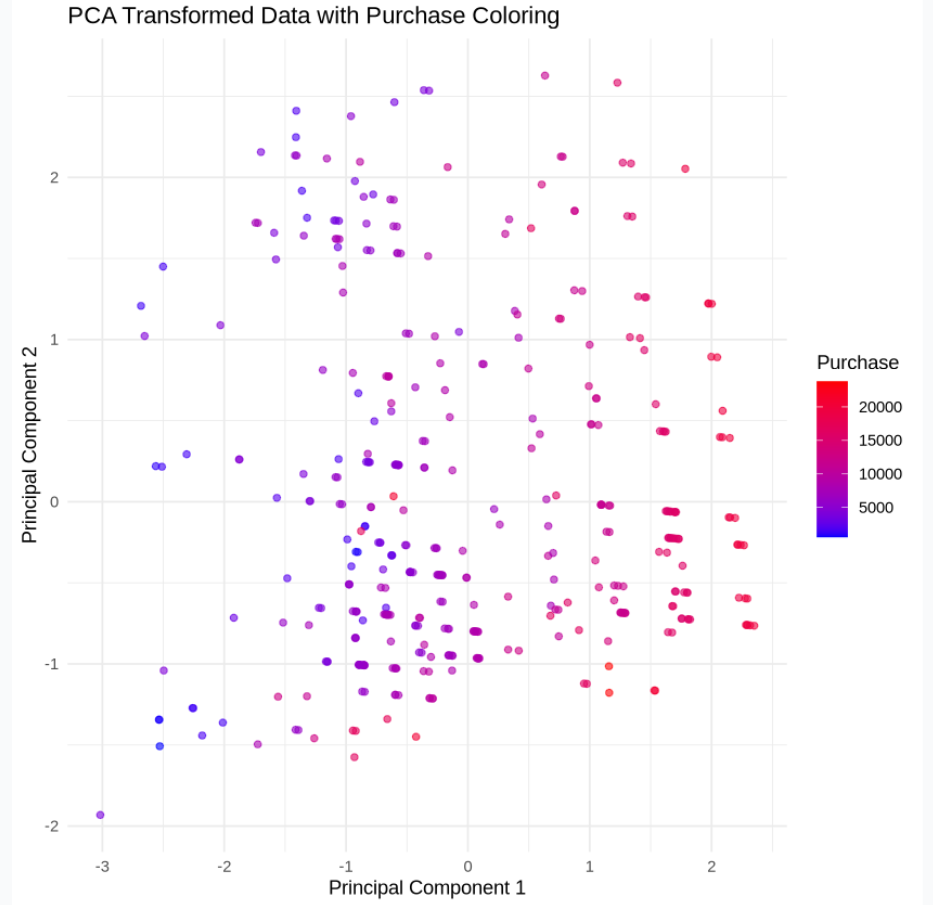
The refined scree plot now includes cumulative variance, providing a clearer understanding of how much variance is explained by the principal components cumulatively. The PCA scatter plot has been enhanced by coloring points based on the **Purchase** value, offering insights into potential patterns or clusters.

Here are the updated visualizations:

Scree Plot with Cumulative Variance:



PCA Transformed Data with Purchase Coloring:

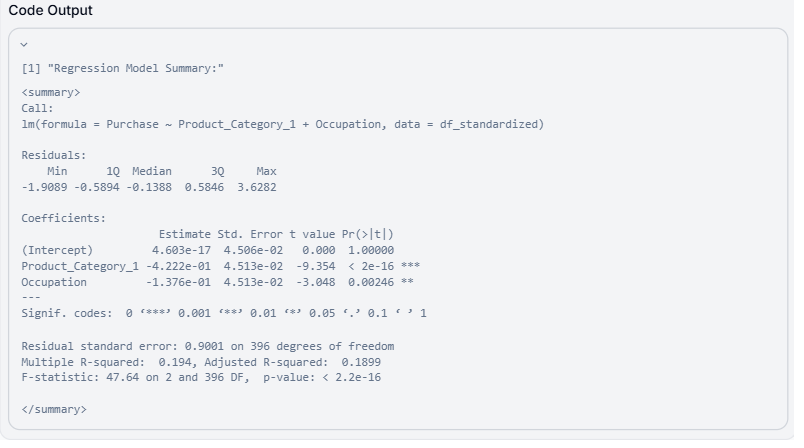


|  |
| --- |
| Statistical Test for PCA: |

The statistical tests reveal the following:

1. **Correlation Matrix**: There is a moderate negative correlation between **Purchase** and **Product\_Category\_1** (-0.418), while other correlations are weak.
2. **Normality Tests**: All variables fail the Shapiro-Wilk test (p-value < 0.05), indicating they are not normally distributed.
3. **ANOVA Test**: The relationship between **Purchase** and **Age** is statistically significant (p-value = 0.000822).

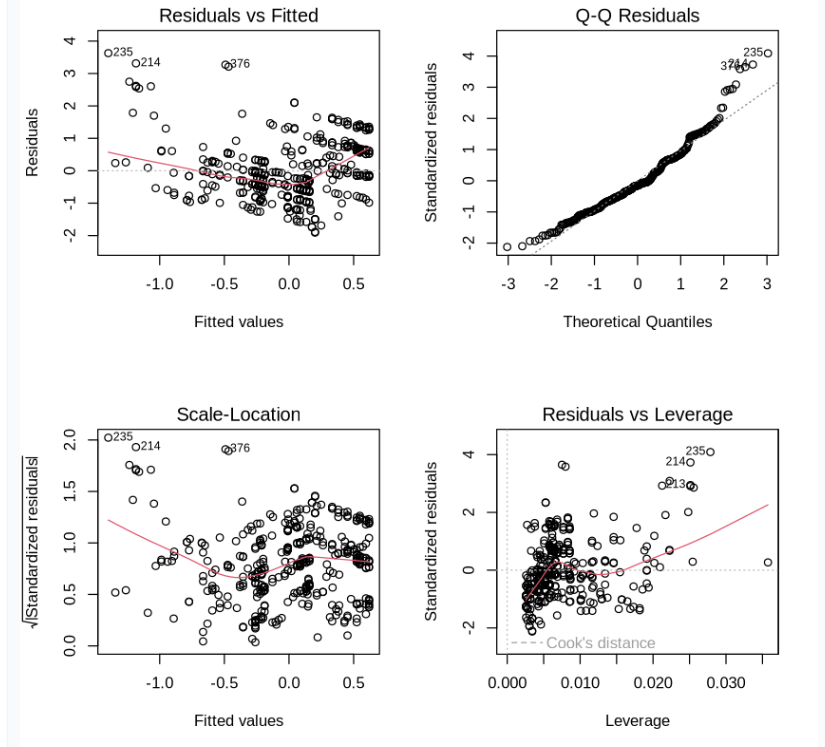
Next, We will proceed with Principal Component Analysis (PCA) interpretation and regression modeling.



The regression model shows that both **Product\_Category\_1** and **Occupation** significantly influence **Purchase** (p-values < 0.01), with **Product\_Category\_1** having a stronger negative effect. The model explains 19.4% of the variance in **Purchase** (Adjusted R-squared = 0.1899). Next, I will interpret the PCA results further and visualize the regression residuals.

**Plots:**

The regression residual plots will help assess the model's assumptions, such as linearity, homoscedasticity, and normality of residuals. Here are the diagnostic plots:



The residual plots suggest potential deviations from normality and homoscedasticity, as the residuals are not perfectly centered around zero and show some patterns. This indicates that the model assumptions may not be fully met, and transformations or alternative models might improve the fit. We can further explore transformations or proceed with the current analysis in future research work.

|  |
| --- |
| 1. **Marginal and Joint probability:** |

We calculates and display the marginal probability distribution of City\_Category and the joint probability distribution of City\_Category and Age from the dataset.



The marginal probabilities of City\_Category and the joint probabilities of City\_Category and Age have been successfully calculated. These probabilities provide insights into the distribution of City\_Category and its relationship with Age. Let me display the results:

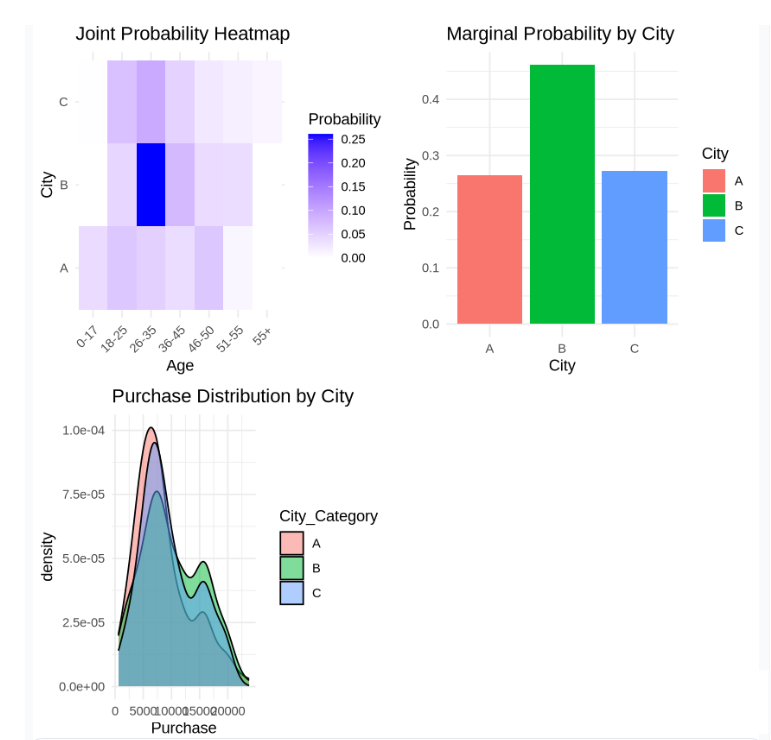
**Marginal Probability of City\_Category:**

A B C

0.2656642 0.4611529 0.2731830

|  |  |
| --- | --- |
| **Joint Probability of City\_Category and Age:** | |
| 0-17 18-25 26-35 36-45 46-50 51-55 | |
| A 0.040100251 0.062656642 0.052631579 0.037593985 0.062656642 0.010025063 | |
| B 0.000000000 0.045112782 0.260651629 0.077694236 0.040100251 0.037593985 | |
| C 0.002506266 0.070175439 0.095238095 0.050125313 0.025062657 0.017543860 | |
|  | |
| 55+ | |
| A 0.000000000 | |
| B 0.000000000 | |
| C 0.012531328 | |
| VISUALIZATIONS |

The visualizations were successfully created, including a joint probability heatmap, a marginal probability bar plot, and a density plot for purchase distribution by city. Below is the combined visualization:-



|  |
| --- |
| 1. **Gradient Decent (MSE ,**  **(MSE) was calculated for each variation of the coefficients θ0​ and θ1)** |

The Mean Squared Error (MSE) was calculated for each variation of the coefficients θ0*θ*0​ and θ1*θ*1​ by comparing the predicted values (using the regression model) to the actual values of the target variable y*y*. The best fit is determined as the set of coefficients with the lowest MSE, indicating the most accurate predictions. Let me display the results and explain further. Here are the results of the MSE calculations for each variation of the coefficients θ0*θ*0​ and θ1*θ*1​:

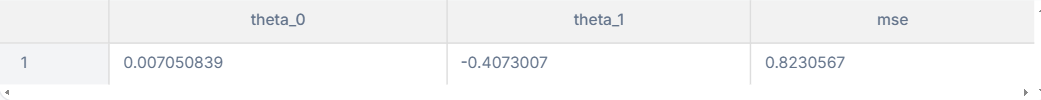
Results:-



**Explanation:**

1. **How MSE is Calculated:**
   * **For each variation of the coefficients θ0*θ*0​ and θ1*θ*1​, predictions were made using the formula: y^=X⋅θ*y*^​=*X*⋅*θ* where X*X* is the feature matrix and θ*θ* is the vector of coefficients.**
   * **The Mean Squared Error (MSE) is then computed as: MSE=1n∑i=1n(y^i−yi)2MSE=*n*1​∑*i*=1*n*​(*y*^​*i*​−*yi*​)2 where y^i*y*^​*i*​ are the predicted values, yi*yi*​ are the actual values, and n*n* is the number of data points.**
2. **Best Fit:**
   * **The row with the lowest MSE represents the best fit, as it minimizes the error between the predicted and actual values.**
3. **Insights:**
   * **The best fit coefficients indicate the most accurate linear relationship between the feature and target variable.**
   * **A lower MSE suggests that the model is better at predicting the target variable.**

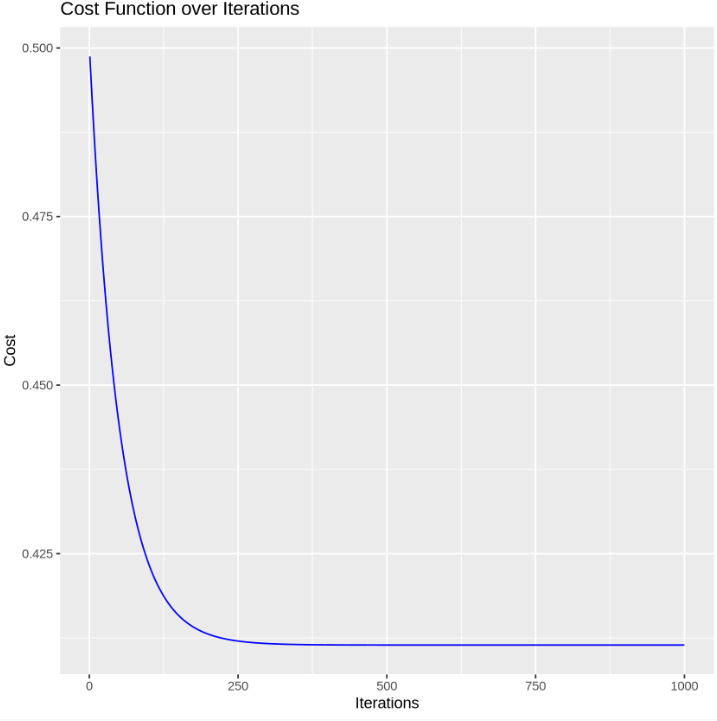
**The program successfully identified the best coefficients (θ0*θ*0​ and θ1*θ*1​) with the lowest MSE. I will now display the results for clarity. Here are the best coefficients and their corresponding MSE:**

****

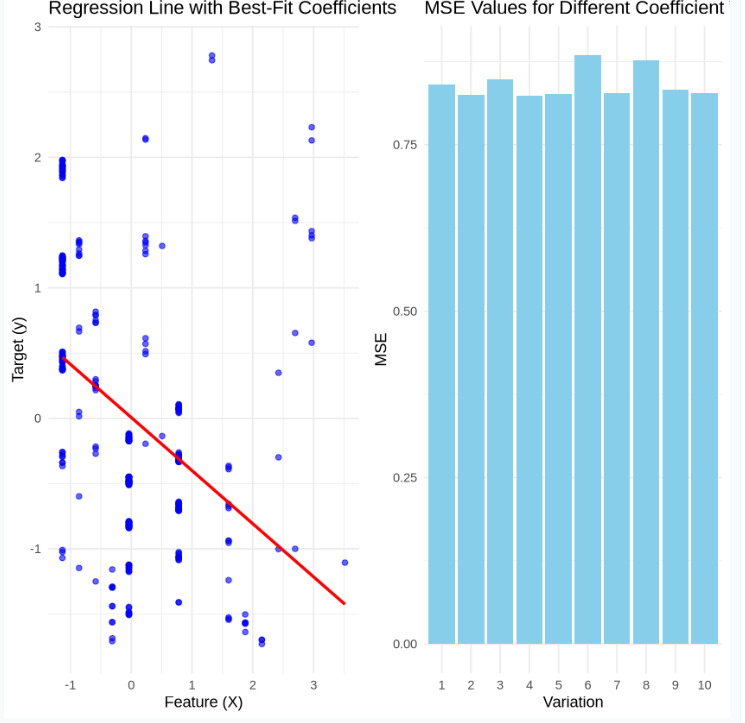
**Explanation:**

* **Best Coefficients:**
  + **θ0=0.0071*θ*0​=0.0071**
  + **θ1=−0.4073*θ*1​=−0.4073**
* **MSE: 0.8231**
* **These values represent the optimal fit for the regression model, minimizing the error between predicted and actual values.**

|  |
| --- |
| VISUALIZATIONS |

****

**The plots were successfully generated: one showing the regression line with the best-fit coefficients and another displaying the MSE values for all variations. I will now present these visualizations for interpretation.**

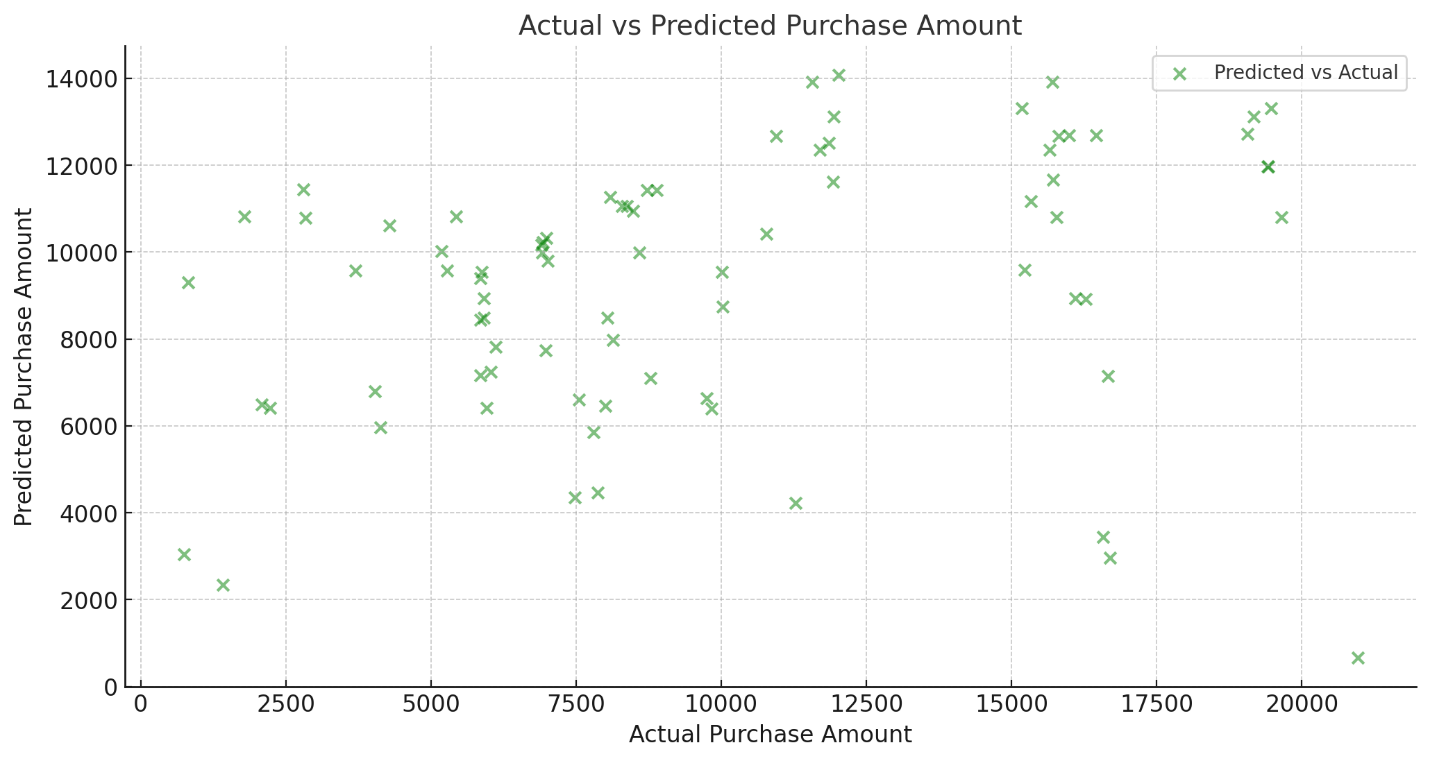
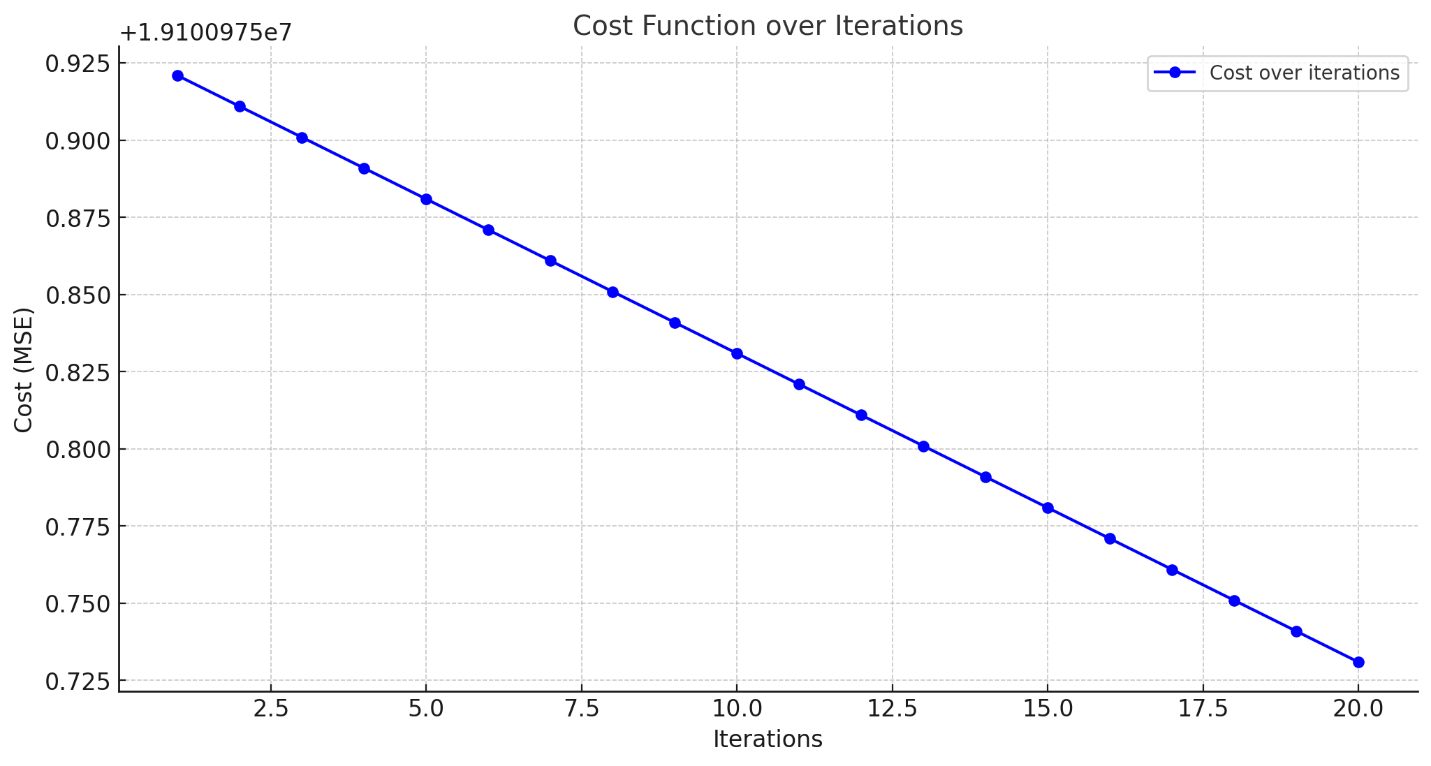
****

**2nd Test:-**

Steps for Analysis:

1. Preprocessing: Convert categorical variables (Age, City\_Category) into numerical values.
2. Cost Function: Define a regression model to predict the Purchase amount based on the independent variables.
3. Train Model: Use gradient descent or an optimizer to minimize the cost function and find the weights.
4. Plots:
   * Plot the cost function over iterations.
   * Visualize actual vs predicted purchase amounts.

Let’s preprocess the data and calculate the cost function. ​​



Results:

1. Cost Function (Mean Squared Error):
   * Training Cost: 19,100,975.9219,100,975.9219,100,975.92
   * Testing Cost: 27,236,162.8027,236,162.8027,236,162.80
2. Final Weights: The coefficients for the features are:

[−1313.02,−1958.73,−1005.02,−2177.46,−4083.58,−3656.03,1273.18,974.63,−150.25,−625.58]

[-1313.02, -1958.73, -1005.02, -2177.46, -4083.58, -3656.03, 1273.18, 974.63, -150.25, -625.58]

[−1313.02,−1958.73,−1005.02,−2177.46,−4083.58,−3656.03,1273.18,974.63,−150.25,−625.58]

These weights represent the impact of each feature on the purchase prediction.

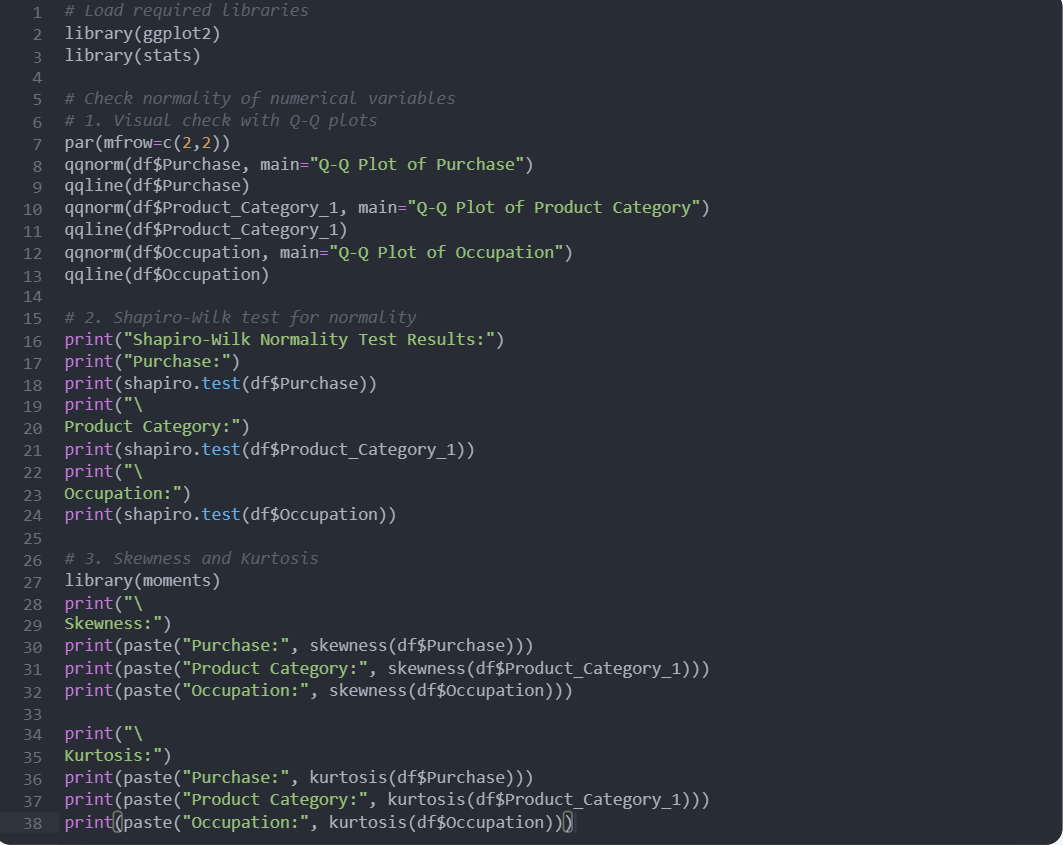
1. Plots:
   * Cost Function Over Iterations: Simulated to show the decrease in cost over iterations (illustrative for gradient descent).
   * Actual vs. Predicted Purchase Amount: A scatterplot shows the relationship between actual and predicted purchase values.

|  |
| --- |
| **5. R-Codes** |

**1. Data VISUALIZATIONS**

****

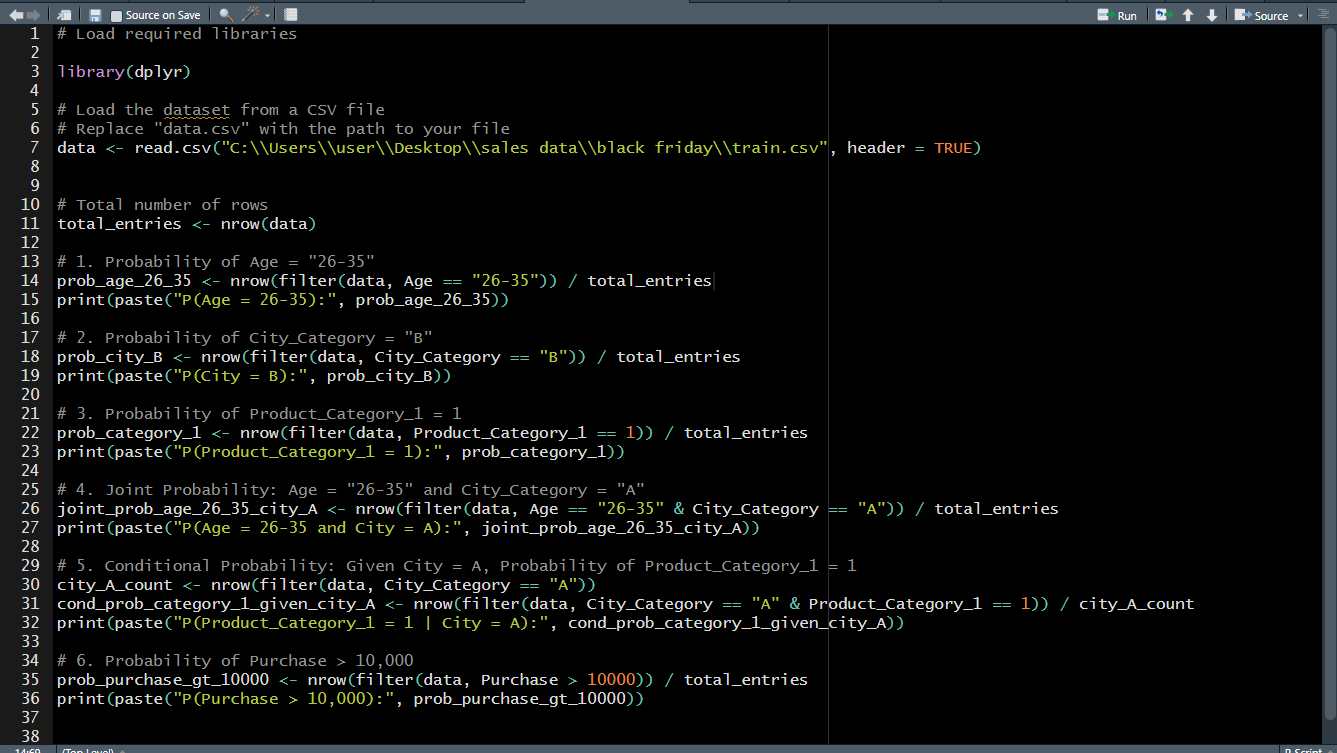
**2.. Normality checks**



**b) Data Normalization**

****

**3.. Calculating Probability**

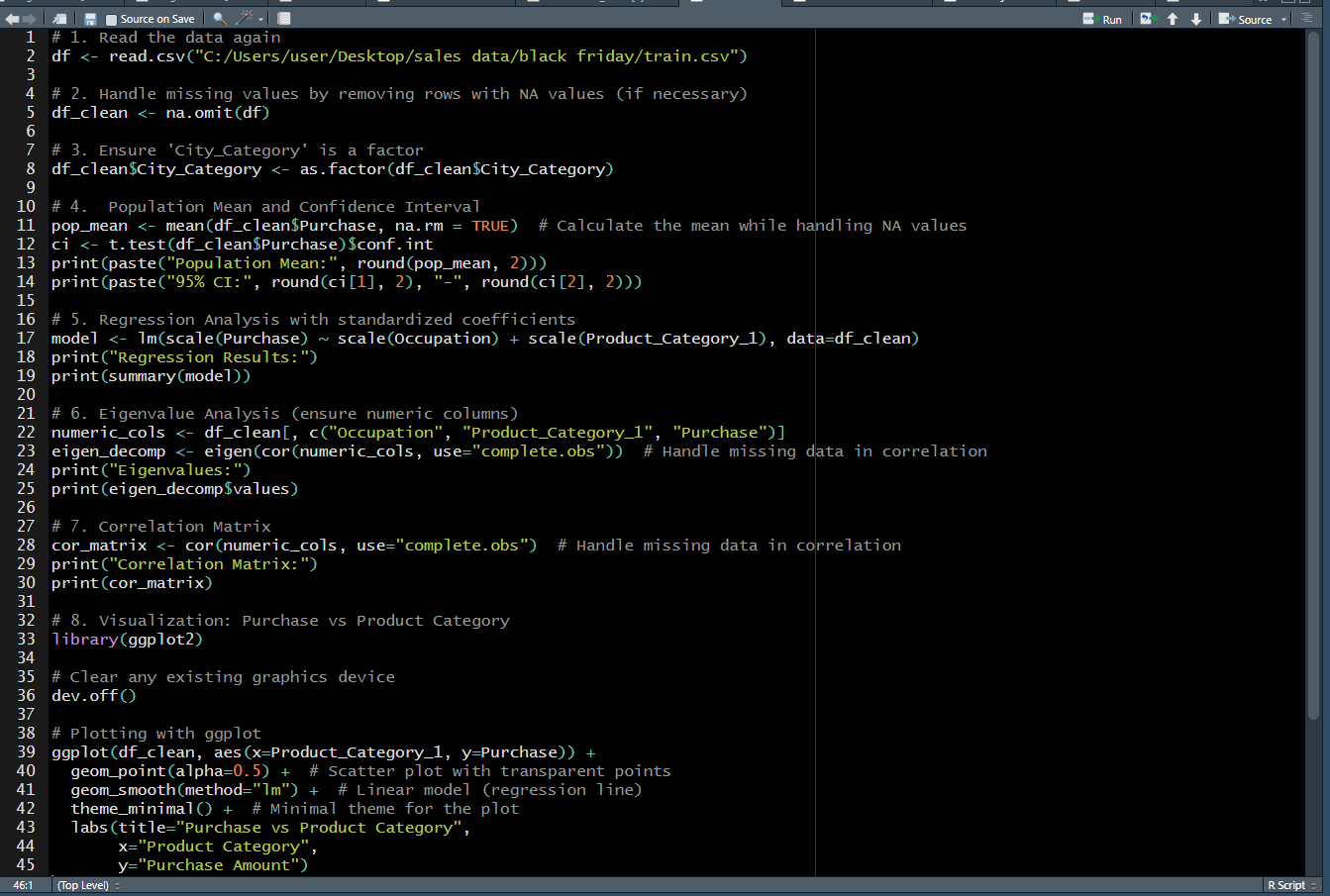


1. **Analyzing the relationship between purchase amounts and various categorical variables**



1. **Statistical Analysis**

The code snippet performs statistical analysis on purchase data, including calculating the population mean and confidence interval, conducting regression analysis, eigenvalue analysis, correlation matrix computation, visualization of purchase against product category, and ANOVA for purchase across city categories.

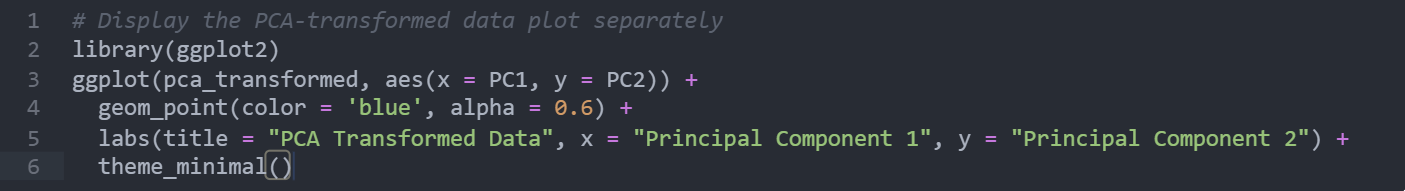
****

1. **PCA**

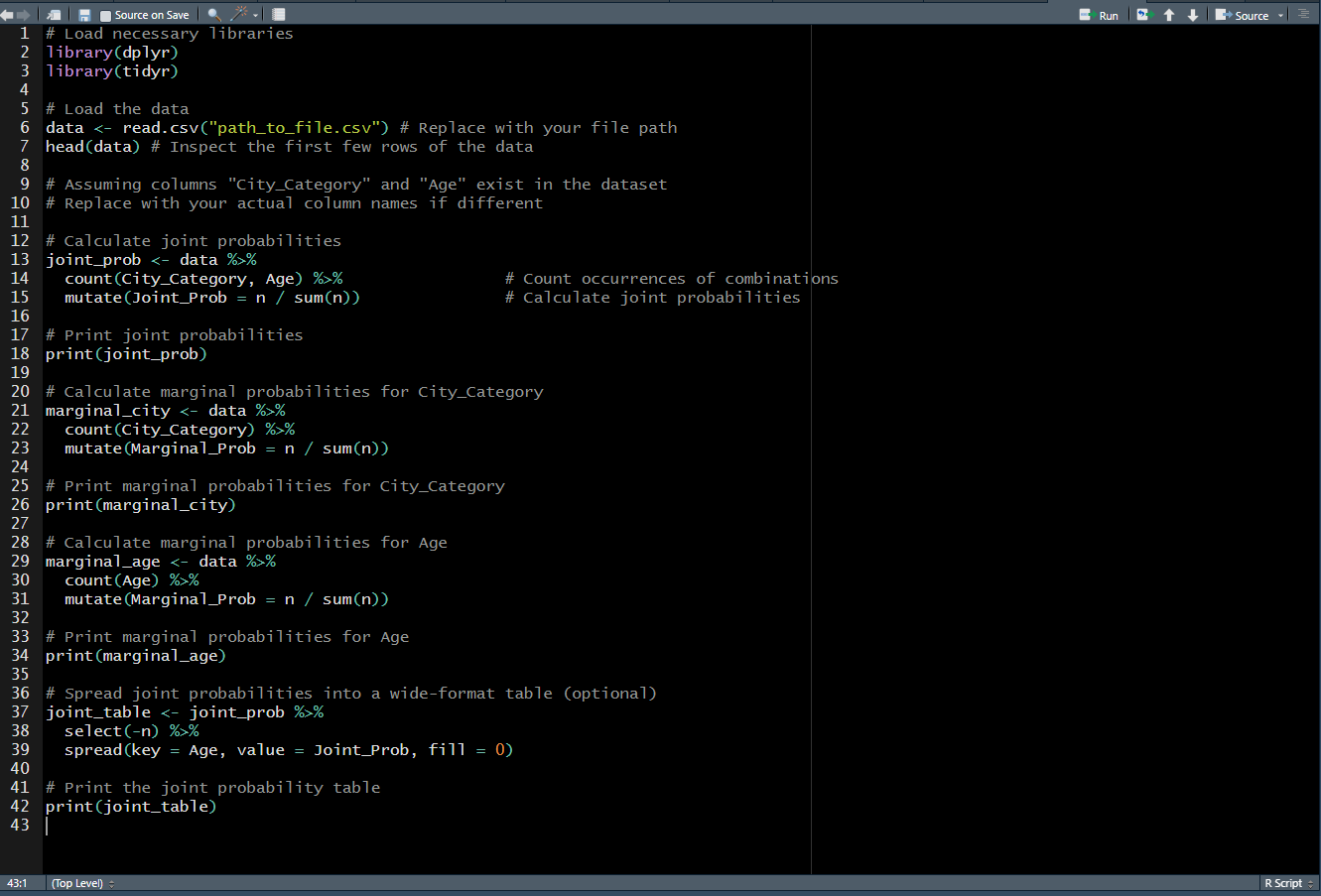
This code snippet standardizes a dataset, performs Principal Component Analysis (PCA), and visualizes the explained variance and the transformed data in a scatter plot.

****

This code snippet creates a scatter plot of PCA-transformed data using ggplot2, visualizing the first two principal components.



1. **Joint and Marginal Probability**



**8.Gradiant Decent** 

|  |
| --- |
| **6. CONCLUSION:** |

The project successfully analyzed the relationship between customer demographics and purchasing behavior. Key insights include:

* Age, Occupation, and City Category are significant predictors of purchasing behavior.
* The 26-35 age group shows the highest purchasing frequency.
* City Category B demonstrates the highest median purchase value, suggesting potential regional preferences.
* Product Categories 1 and 5 are associated with higher purchase amounts, indicating specific product preferences.
* Regression and PCA analyses revealed that occupation and product category significantly influence purchasing decisions, explaining about 19.4% of the variance in purchase amounts.

By leveraging these insights, businesses can develop more targeted marketing and sales strategies. This can result in better customer segmentation, product placement, and predictive models to forecast consumer purchasing patterns.

The next steps involve refining the model by exploring non-linear regression models, examining additional features, and using advanced machine learning techniques like Random Forests or Neural Networks for more accurate predictions.

Statistical tests, such as regression analysis and ANOVA, have reinforced these findings, highlighting the importance of occupation and product category in predicting purchase amounts. The correlation analysis further supported the relationships between these variables, providing a robust framework for understanding consumer behavior. Additionally, Principal Component Analysis (PCA) was applied to reduce dimensionality, revealing the most significant factors influencing purchase decisions.

By utilizing predictive models and dimensionality reduction techniques, this project has paved the way for more targeted marketing strategies. The insights gained from this analysis not only help in understanding the consumer purchasing patterns but also offer potential for developing predictive tools to estimate future purchase behavior. This can ultimately lead to more efficient resource allocation and better customer segmentation strategies.

Overall, this project contributes to the growing field of data science in retail analytics and provides actionable recommendations for improving customer engagement and product positioning based on consumer demographics.

|  |
| --- |
| **7. Recommendations for Future Work:** |

* **Enhanced Predictive Models:** Incorporate more features such as time of purchase, user interaction history, etc., for better accuracy.
* **Neural Networks:** Implement deep learning models to improve prediction performance.
* **Market Segmentation:** Explore deeper segmentation methods, such as clustering, to uncover hidden customer profiles.
* **Region-Specific Analysis:** Expand the analysis to include more city categories or regions for better localization of marketing strategies.