AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES

(AIMS RWANDA, KIGALI)

Name: Jean de Dieu NGIRINSHUTI
Assignment Number: 1
Course: Statistical Regression
Date: November 16, 2024

Introduction

This report analyzes customer purchase behavior using a dataset containing information about , Gender, City Category, Stay in Current City Years, Marital Status, Age (approximate), and Purchase Amount. The objectives are to summarize key patterns through descriptive statistics and visualizations, examine the relationship between Purchase and Age using linear regression, and assess the impact of Gender on Purchase with statistical and graphical methods.

Descriptive Analysis

Central Tendency, Dispersion, and Frequency Distribution

Table 1 presents the summary statistics for numerical variables, while Table 2show frequency distributions for categorical variables , which is essential for further analysis.

Table 1: Summary Statistics for Numerical Variables

Variable	**Min**	**1st Qu.**	**Median**	**Mean**	**Max**
Age (Years)	18	26	35	34.8	70
Purchase Amount	185	5894	8047	9269.8	23961

Table 2: Frequency of Categorical Variables

Variable	Gender		City Category			Marital Status		Stay in Current City Years			
Category	Female	Male	A	B	C	Unmarried	Married	1	2	3	4
Frequency	39,374	119,626	46,784	83,860	28,556	93,654	65,346	100,573	32,960	15,219	16,194

The fugure 1 shows that most purchases range between 5,000 and 15,000, with minimal differences in spending patterns between genders. The average purchase amount remains consistent across age groups, with a slight but negligible positive correlation between age and spending. The distribution of customers across city categories and marital status highlights demographic variation but does not strongly influence purchase amounts.

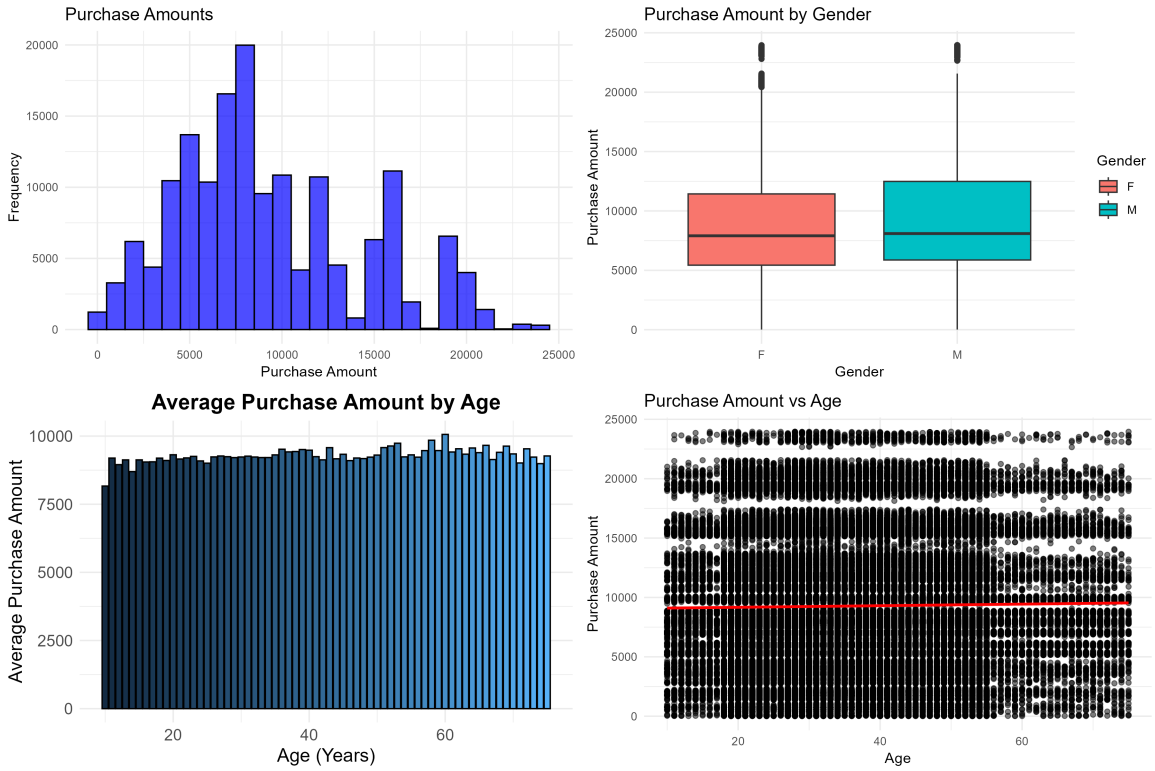

Figure 1: Descriptive Visualizations

Linear Regression of Purchase Amount vs. Age

To investigate the dependence of 'Purchase' on 'Age', a simple linear regression model was fit:

$$\text{Purchase} = \beta_0 + \beta_1(\text{Age}) + \epsilon$$

Results

Intercept ($\hat{\beta_0} = 9037.936$), Slope($\hat{\beta_1} = 6.681$), Residual Standard Error = 5026, $R^2 = 0.0002$ (low explanatory power).

The fitted model is given as :

$$\text{Purchase} = 9037.936 + 6.681(\text{Age})$$

For Uncertainty, $\text{Var}(\hat{\beta_0}) = 1549.938$, $\text{Var}(\hat{\beta_1}) = 1.148278$ and $\text{Cov}(\hat{\beta_0}, \hat{\beta_1}) = -39.9667$.

Interpretation

Age has a statistically significant but minimal impact on purchase amounts, increasing purchases by 6.68 units for every additional year of age. However, the model explains almost none of the variability in purchase amounts (R-squared = 0.0002), indicating other factors are likely more important.

Model Usefulness and Limitations

The model is statistically valid ($p - value < 0.001$) but practically ineffective. It explains less than 0.03% of the variation in purchase amounts, with a large residual standard error (5026), making

predictions highly unreliable. Including additional predictors and exploring non-linear relationships could improve its utility.

Suggestions for Improvement

1. We can improve our model by adding other variables like `Gender`, `City Category`, and `Marital_Status`.
2. Another way is to investigate whether `Age_num` has a non-linear effect on `Purchase` (e.g., quadratic or piecewise regression).

Association Between Purchase and Gender

The relationship between 'Purchase' and 'Gender' was explored using Descriptive Statistics, where Males have a slightly higher mean purchase amount (9445.14) compared to females (8740.00) and as shown on table 3 below.

Two-Sample t-Test, The p-value was significant (< 0.05), indicating that mean purchase amounts differ significantly by gender.

The 95% confidence interval for the difference in means is:

$$[-760.7066, -650.0360]$$

This interval suggests that males spend between 650 and 760 units more than females on average.

The results suggest that gender is a statistically significant factor in explaining purchase behavior, with males spending more on average.

Table 3: Summary Statistics by Gender

Gender	Mean Purchase	Median Purchase	SD Purchase
Female	8740.00	7908.00	4780.00
Male	9445.14	8096.00	5093.00

Table 4: Coefficients from Linear Regression Model

Variable	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	8739.76	25.29	345.6	$< 2e-16$ ***
Gender	705.37	29.15	24.2	$< 2e-16$ ***

The results of the regression analysis are:

$$\hat{\beta}_0 = 8739.76, \quad \hat{\beta}_1 = 705.37$$

This means that the average purchase amount for males is 705.37 units higher than for females.

End.