AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES

(AIMS RWANDA, KIGALI)

Task 1: Descriptive Analysis of the Data

The dataset *data_purchased_behaviour.csv"* comprises of 159000 observations and 7 variables, where only 6 are key variables. There are 4 numerical and 2 categorical variables. The numerical variables are: *Stay_In_Current_City_Years, Marital_Status, Age_num, Purchase* , while the categorical variables are *Gender* and *City_Category*.
58.9% of the customers are not married with the male having the highest proportion. This is evident from the fact that majority of the customers sampled are males. There are 39374 females and 119626 males. This imbalance might affect the trend of the purchase amount. Table 1 shows the cross frequency table for the gender and city category. City B has the highest number of customers, also with the highest proportion of males and females.

Table 1: Cross frequency table for Gender and City category

	A	B	C	Total
Female	10276	16776	12322	39374
Male	32092	50326	37208	119626
Total	42368	67102	49530	159000

The summary for the numerical variables is given by Table 2. The distribution of ages is relatively centered around the early 30s, this shows the customers is made up of mostly young people. Purchase amounts average 9270.46 units, with significant variability (SD = 5026.50), and a mode of 6950, suggesting some high-value purchases. Customers have stayed in their current city for 1.86 years on average. The mode of 1 year indicates that most customers have only been in their current city for a year.

Table 2: Summary of the numerical variables

	Mean	Median	Mode	SD
Age	34.81	33	31	11.76
Purchase	9270.46	8044	6950	5026.50
Stay in current city	1.86	2	1	1.29

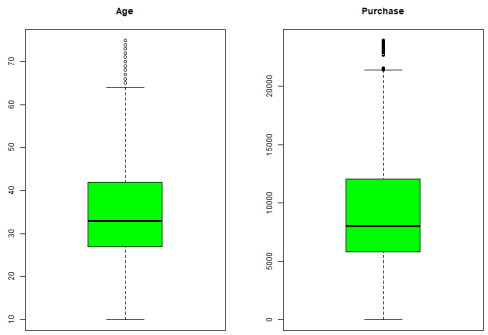

Figure 1: Box plots for Age and Purchase distribution

From the box plot shown in Figure 1, the age variable has some outliers. This outliers are the customers who are of old age. The distribution is approximately normal. Similarly, purchase has some outliers. It has a right-skewed distribution. This can be seen from summary in Table 2.

Task 2: Simple Linear Regression

The linear regression model to investigate the dependence of the purchase amount on age is given by the equation

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1}$$

where y_i, x_i, β_0, and β_1 represents the purchase amount, age, intercept and slope respectively. The estimators for the slope and intercept are found to be $\hat{\beta}_0 \approx 9037.93$ and $\hat{\beta}_1 \approx 6.68$ with standard errors 39.37 and 1.07 respectively.

(a) Uncertainties

The uncertainties for the model is given below.

$$\text{Var}(\hat{\beta}_0) \approx 1549.94, \text{Var}(\hat{\beta}_1) \approx 1.15, \text{Covar}(\hat{\beta}_0, \hat{\beta}_1) \approx -39.97$$

(b) Interpretation of the model

The intercept of the model is gotten to be $\hat{\beta}_0 \approx 9037.93$. This represents the predicted purchase amount when the age is zero. The slope on the other hand is $\hat{\beta}_1 \approx 6.68$. This indicate that a increase in age by 1 will cause the purchase amount to increase by 6.68. Since it is positive, then purchase and age have a positive linear relationship. The covariance of the estimators parameters is less than 0. The variance of $\hat{\beta}_0$ is very large. Thus, the estimate is not precise. Therefore, the parameters tends to move in different direction.

(c) Usefulness of model

The model shows how the age of customers can be a factor to look out for in marketing. We noticed that the increase in age will cause an increase in purchase amount. Understanding this positive linear relationship, marketing strategies can be put in place to target the older people. This kind of model help to look at the possible way to improve revenue.

(d) Limitations

This model has some limitations. Some of the assumptions are not met.

1. **Linear association:** From the plot of the residuals against the fitted values, we have that the residuals above the line does not balance the ones below. Thus, the linear association assumption is not satisfied.

2. **Normality:** From the histogram of the residuals and normal curve shown in Figure 2, we have that the histogram does not closely follow the shape of the normal curve. This is because the age variable over-estimate the purchase amount causing the change in the shape of the histogram. Overall, the residuals is not normal.

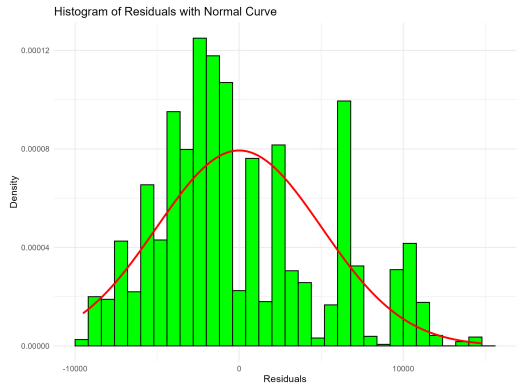

Figure 2: Histogram of Residuals with Normal Curve

(e) Improvement of model

The model can be improve in different ways. Two of them are:

1. **Transformation:** The variables can be transform to make it satisfy the normality assumption. This could be by using the log transformation.

2. **Add more variables:** We can add more covariate or independent variables to improve the model.

Task 3: Investigating association between the purchase amount and gender

Since the purchase amount is a numerical variable and gender is a categorical variable, we can investigate the association between them by carrying out hypothesis testing using the Student's t test. We will group the purchase amount by gender and test the mean of each group. The null hypothesis is $H_0 : \mu_i - \mu_j = 0$ and the alternative hypothesis is $H_a : \mu_i - \mu_j \neq 0$. The test is valid only if the numerical variable is normally distributed in each group (normality). and the variance of the numerical variable is the same in both groups (homogeneity).