AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES

(AIMS RWANDA, KIGALI)

Name: Yuusf Brima | Assignment Number: 2

Course: Statistical Regression with R | Date: December 12, 2020

Task

(1) Undertake a descriptive analysis of the variables in this dataset.

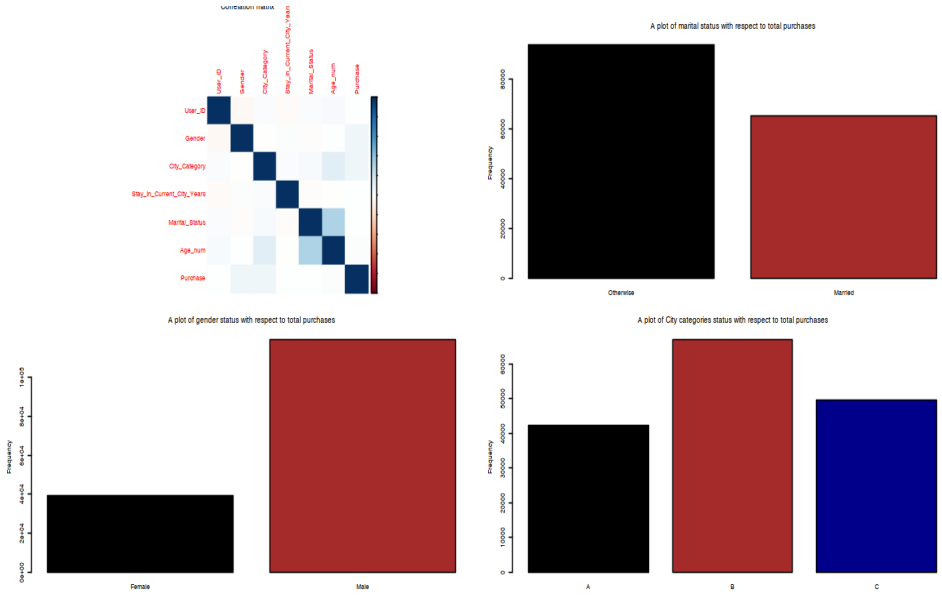

Figure 1: Basic visualization of the observed variables in the dataset

Basic Summary Statistics

```
Stay_In_Current_City_Years      Age_num         Purchase
Min.   :0.000               Min.   :10.00   Min.   :   12
1st Qu.:1.000               1st Qu.:27.00   1st Qu.: 5828
Median :2.000               Median :33.00   Median : 8044
Mean   :1.856               Mean   :34.81   Mean   : 9270
3rd Qu.:3.000               3rd Qu.:42.00   3rd Qu.:12059
Max.   :4.000               Max.   :75.00   Max.   :23961
```

```
Gender            City_Category       Marital_Status
Males:    119626  A      42368        Married:    65346
Females:  39374   B      67102        Unmarried:  93654
                  C      49530
```

From the summary statistics, we can observe the following: the distribution of age for shoppers is mostly clustered around 34 which indicates that youths account for more much purchases that other demographics and among them males tend to purchase more nearly four times compared to females. Furthermore, customer segment from city B do more purchase compared to city categories A and C which may be due to other indicators such as income level, purchasing power of the customers, education level which is a strong predictor of wealth, other reasons may be better marketing and sales strategy in city B as compared to the others. Another interesting insight from the data is that unmarried customers account for nearly two-thirds of all purchases this may largely be due to the fact that the mean age of observations is 34.81 largely youths in their early career stage. These phenomena understudy are presented in the visualizations in Figure (1).

(2) Use this dataset to build a linear regression model to predict the purchase amount. Justify all the choices you make to arrive at your final model.
The equation for a multiple linear regression model is stated in (1).

$$y = \beta_0 + \beta_1 x_1 + \beta_1 x_2 \cdots + \beta_{p-1} x_{p-1} \epsilon \tag{1}$$

The equation for the multiple linear regression model is thus:

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 \text{Age} + \hat{\beta}_2 + \hat{\beta}_3 \tag{2}$$

The estimates for intercept $\hat{\beta}_0 = 8271.658$ and slope of the line is $\hat{\beta}_1 = 3.274$.

$$\hat{\beta}_2 = \begin{cases} 708.019 & \text{Gender Male} \\ 0 & \text{Gender Female} \end{cases}$$

$$\hat{\beta}_2 = \begin{cases} 232.321 & \text{City Category B} \\ 815.757 & \text{City Category C} \\ 0 & \text{City Category A} \end{cases}$$

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    8271.658     48.229 171.509  < 2e-16 ***
GenderM         708.019     29.088  24.341  < 2e-16 ***
City_CategoryB  232.321     31.119   7.465 8.34e-14 ***
City_CategoryC  815.757     33.383  24.436  < 2e-16 ***
Age_num           3.274      1.076   3.044  0.00234 **

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5006 on 158995 degrees of freedom
```

```
Multiple R-squared:  0.008046,Adjusted R-squared:  0.008021
F-statistic: 322.4 on 4 and 158995 DF,  p-value: < 2.2e-16
```

Having experimented with Backward Pass mechanism, we therefore can research the conclusion that observational variables Gender,City Category and Age have higher explanatory power for the response variable Purchase because under the Null Hypothesis test with a significance level of 0.05, the P-values for these independent variables as show in the above model summary, are lower which indicates that explain the variation in the dependent variable in question.

(3) Interpret your final model.
From the model summary in (2), the parameter estimates for the observed variables indicate that the variables Gender, City Category and Age have high significance levels for the variability of the response variable Purchase. More importantly, having got positive beta estimates for the observed variables, it is a clear indication that purchases are strongly influenced by Gender cluster male followed by City Category C and the remaining observed variables. Given the P-values of the parameter estimates are lower than the significance level, we therefore reject the Null Hypotheses.

(4) Are all the assumptions about your model met?

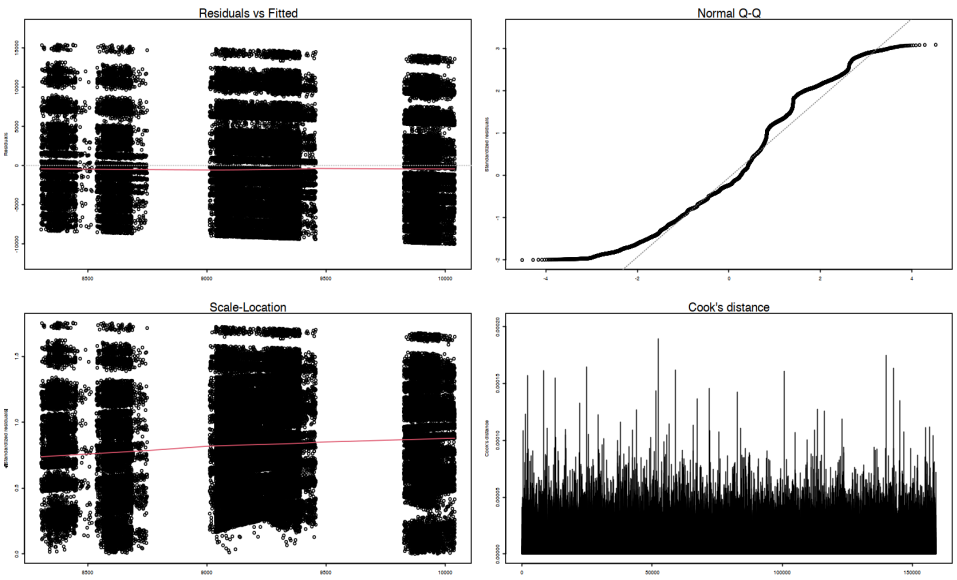

Figure 2: Basic visualization of the residuals vs. fitted under: Linearity, Normality, Scale-Location (homoscedasticity) and Cook's distance measure of outliers

(a) Fixed predictors: The independent variables are non-stochastic because they generated by an underlying probability process.

(c) Linear association: The relationship between the mean value of response variable Purchase and each observed variable is linear if the other explanatory variables are held fixed. However, from 2, the assumption does not hold because the residual vs. the fitted data does not have random variation.

(d) Homoskedasticity. The variance of the error term is constant. From the Scaled-Location plot in 2 (bottom left), the graph does not show random variation which means the assumption of constant variance of ϵ_i does not hold.

(e) Normality: The error term is normally distributed. Likewise, from the Normal Q-Q plot in 2, the standard residual does not follow a straight line which implies the data is not normally distributed.

(5) Explain how you can use your final model to help the company improve their sales?
From the initial exploratory data analysis and observations in (1), we can make the following recommendations to help the company improve their sales:

(a) Targeted advertisement and marketing to middle, elderly and young populations through promotions, discounts and other techniques.

(b) Expand more retail outlets in City Categories A and C as well as better customer relationship management in those cities possibly through marketing survey, sentiment analysis amongst other business analytics approaches.

(c) The company can invest in more female products to target that segment of the market.

(6) Suggest possible ways of improving your model.
Firstly, we may want to go back to the theory and hypotheses by asking the question: is there really a linear relationship between the predictors and the outcome? If the former is true, we may add more observational variables that have higher variability in the outcome (or response variable). For example, income level is a strong predictor of customer purchasing power as well as level of education with is a strong predictor of income level. Another strategy we can use is Principal Component Analysis (PCA) to select the principal components that explain the variability in the response variable. We can as well use different feature transformation techniques to improve the model performance. More importantly, if the data were systematically biased during the data collection process we may redesign data collection methods.

(7) Compare your final model to the model built on Gender and Age only.
The final model in (2) has an Adjusted R-squared of 0.008021 and an F-statistic of 322.4 compared to a model built using the predictors Gender and Age with an Adjusted R-squared of 0.003897 as well as an F-statistic of 312.1. This comparison indicates that the former is a better regressor compared to the latter because its Adjusted R-Squared is higher.