OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

J.E. CAIRNES SCHOOL OF BUSINESS & ECONOMICS

INDIVIDUAL ASSIGNMENT COVER PAGE

Title: Individual Assignment 3

Module Name and Code: Applied Customer Analytics MS5108

Student Name and ID: Dilip Venkatesan Sankar 22225743

In submitting this assignment, I am aware that it is my responsibility to adhere to the submission guidelines. Please tick (double click… or Yes/No) for the following:

	Yes	No
I am aware of what the UOG plagiarism policy entails.	☒	☐
I have named the assignment file (MS Word docx , doc, or jar) to contain my student ID and the module code and assignment number (e.g., 1187404_MS220_A1.docx or 1187404_MS220_A1.jar).	☒	☐

Declaration for this Assignment Submission*:*

In submitting this work, I confirm that it is entirely my own. I acknowledge that I may be invited to interview if there is any concern in relation to the integrity, and I am aware that any breach will be subject to the University's Procedures for dealing with plagiarism (http://www.universityofgalway.ie/plagiarism).

Data preparation: The dataset was imported as a csv file and the dataset contains the survey details about the satisfaction level of members based on the services provided by CrossFit Ridgeline. The dataset contains 24 columns and 117 rows including the header.
The preliminary analysis included data cleansing of NULL/NA values, checking the class and datatype.

Question 1: Boxplot with 200 words describing the analysis.

The box plot is a graphical representation and summary of numerical data used to examine outliers, skewness, median, minimum, and maximum values. The box plot of **Membership Tenure** based on **Gender (Male and Female)** and **Age (<30 and 30+)** is constructed. Below is the output and analysis,

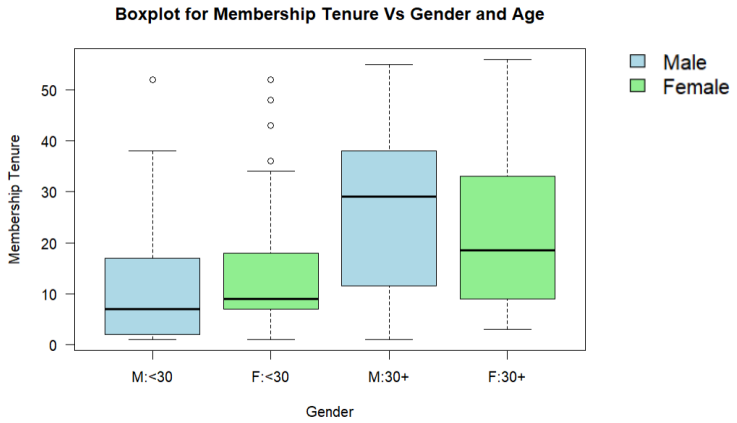


Figure 1: Boxplot for Membership Tenure for Males and Females with Age <30 and

1. Boxplot length – The length of the boxplot for Males is longer compared to the Females indicating the number of males is higher compared to the Females for both age groups (<30 and 30+)
2. Outliers – We can observe that both male and female in the Age group < 30 has outliers indicating a small percentage of members have a longer membership tenure greater than 35 months
3. Median Line – In the age group < 30, the median line for women is slightly higher than that of men, indicating that women have a slightly higher median score than men. In the age group > 30, the median line for men is significantly higher than that for women, indicating that men have a higher median score than women.
4. Skewness - We can determine whether or not a distribution is skewed based on the location of the median value in the box plot.

- The median is closer to the bottom of the box and the whisker is shorter for males and females under the age of 30, demonstrating positive skewness.
- The median is further toward the top and the whisker is shorter in males with 30+ years of age, showing negative skewness, but the median is nearly equal on both sides in females with 30+ years of age, indicating symmetrical distribution.

Conclusion: Based on the boxplot and the above data, we can infer that the median and membership tenure of Males is greater than Females.

Question 2: Histogram with 200 words describing the analysis

The histogram shows the frequency or count along the y-axis and groups the numerical values into bins along the x-axis to indicate the frequencies of numerical values in a dataset. Plotting the histogram allows

us to see how many people of all ages are participating in CrossFit Ridgeline. Below is the output and analysis,

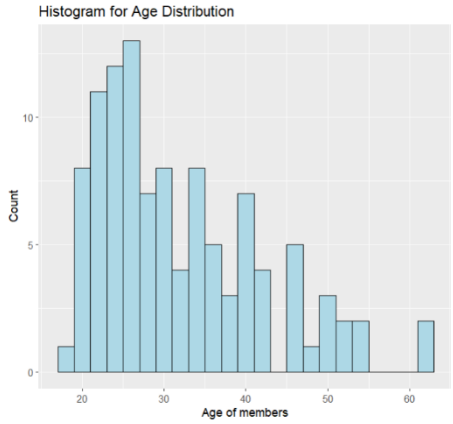


Figure 2: Histogram to display Age distribution

- The count or frequency of the age group from 20 to 25 appears to be the greatest with a count larger than 10 suggesting that younger people under the age of 25 tend to have more gym memberships and engage in regular exercise.
- We may infer from the histogram that fewer people join gyms as they age. The graph shows that as people age, the number of gym memberships drops and appears to reach its lowest point with fewer than three members who are over 30.
- It is also evident that there is an unequal distribution of gym memberships among people between the ages of 40 and 55, which suggests that as people become older, fewer and fewer people are still interested in working out at CrossFit Ridgeline.
- Lastly, we may see an outlier that represents a group older than 60 years old.

Conclusion: The bulk of CrossFit Ridgeline's members tend to be younger people under the age of 25, and as people get older, their numbers appear to be declining, according to the histogram representation.

Question 3: Regression model – With one Independent and three dependent variables

The linear regression model is carried out to check the significant relationship between the variables. From the dataset, we are choosing **Membership tenure** as the dependent variable and **COASAT1(Friendliness and Courteous services by Coach), CLIM1(Comfortability around other gym members), and COND6(Music loudness)** as the independent variables.

The output of the general linear model is stored in **linear_model**. Below are the output and analysis,

```
Call:
lm(formula = Membership.Tenure ~ COASAT1 + CLIM1 + COND6, data = dataset_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-22.817 -11.331  -3.472   6.091  33.528

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   35.393     17.175   2.061  0.04188 *
COASAT1       -2.345      3.007  -0.780  0.43719
CLIM1          8.563      2.015   4.251 4.73e-05 ***
COND6         -9.002      2.675  -3.365  0.00108 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.8 on 102 degrees of freedom
Multiple R-squared:  0.2168,    Adjusted R-squared:  0.1938
F-statistic: 9.413 on 3 and 102 DF,  p-value: 1.51e-05
```

Residuals:

The residuals are denoted as the distance between the actual data point and the regression line and below are the details,

1. Min = -22.81 and Max = 33.52 do not reflect each other
2. 1st Quartile = -11.33 and 3rd Quartile = 6.09 do not reflect each other
3. Median = -3.47 which is very much less than zero

Based on the above values and upon checking the histogram of the residuals in the linear model we can conclude; the residuals are not normally distributed as they did not produce a perfect **bell-shaped curve**.

Coefficients and P-values:

The Estimate column shows y intercept = 35.39 which is the point where the regression line crosses the y-axis and slope (**COASAT1 + CLIM1 + COND6**), and below is the general linear model equation,

y = m.x +b

Where y is the predicted value, m is the slope, x is the independent value (COASAT1 + CLIM1 + COND6) and b is the intercept. So, in our case, the equation can be modified as follow,

Membership Tenure (y) = 35.39 + (-2.34* COASAT1) + (8.56 * CLIM1) + (-9.0* COND6)

This meant that there would be a comparable fluctuation in the length of Membership for every increase or reduction in the independent variable.

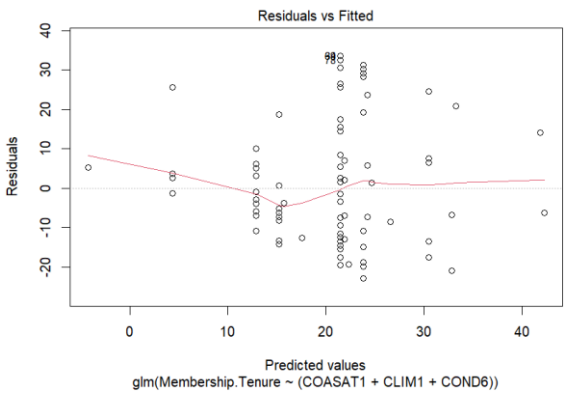
The p-value of the intercept = 0.041which is < 0.05 and there was a significant relationship between **membership tenure with CLIM1 & COND6** where p-value < 0.05 but there was no significant difference between the **intercept and COASAT1** (p=0.43) which does not support the alternate hypothesis.

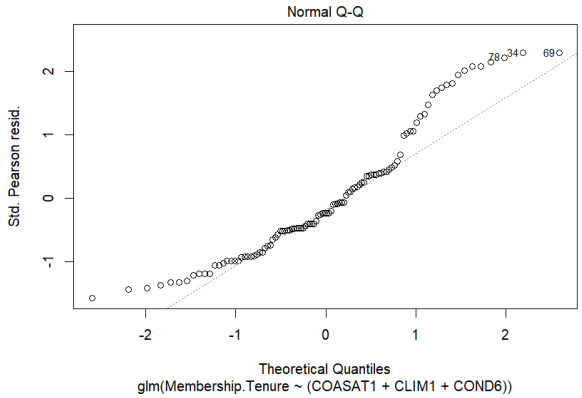
The t-value can be determined by dividing **Estimate Value/Standard Error**, where the t-value for COASAT1 = -2.34/3.0 = -0.78

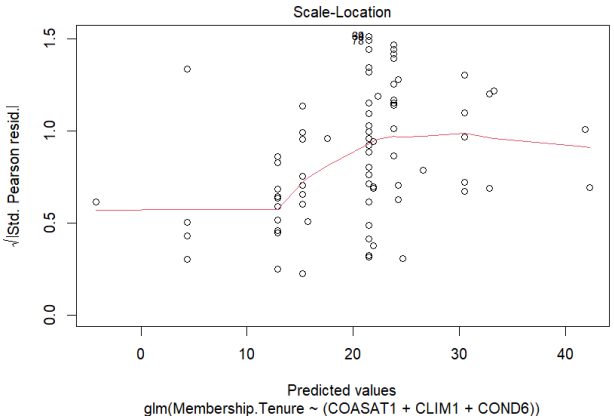
Other Observations:

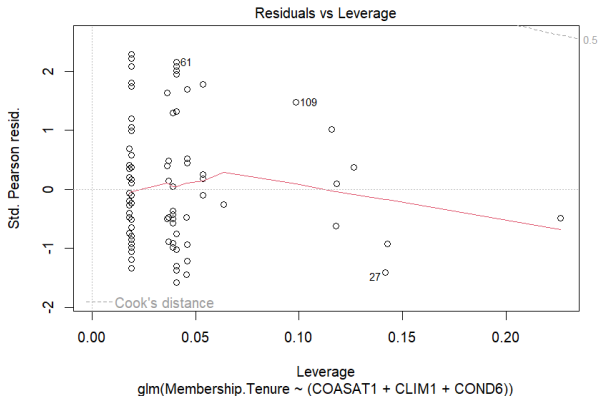
1. **Residual Standard error** denotes the average distance that the observed values fall from the regression line = 14.8 with 102 degrees of freedom
2. **Multiple and adjacent R-squared** values show how much variance the dependent variable can be accounted for by the independent variable which shows 0.21 and 0.19 (21% and 19%) of variance in y is accounted for by COASAT1, CLIM1, and COND6
3. Finally, the **F-Statistics = 9.4 and p-value = 1.15e-05** which means the dependent variables are significantly related to y.

Generate diagnostic plots to test assumptions of linear regression









1. **Residual-Fitted graph:** The points are not spread across the horizontal red line indicating the values do not have a linear relationship.
2. **Normal Q-Q plot and histogram:** Shows an S-shaped curve indicating an **excessive Kurtosis** and the residuals are not normally distributed.
3. **Scale-Location graph:** The residuals are not spread equally across the horizontal line indicating the model does not hold well for **Homoskedasticity.**
4. **Residuals Vs Leverage:** Based on **Cook's distance,** we can observe we have multiple outliers.

In order to determine if there is a significant relationship, ANOVA test was carried out and from the output, we can observe that p = 1.51e-05 which is < 0.05 concluding there is at least one significant relationship in the model.

```
> anova(linear_model, lm(Membership.Tenure ~ 1, data = dataset_clean))
Analysis of Variance Table

Model 1: Membership.Tenure ~ COASAT1 + CLIM1 + COND6
Model 2: Membership.Tenure ~ 1
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    102 22348
2    105 28535 -3   -6186.7 9.4126 1.51e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> confint(linear_model, conf.level=0.95)
                2.5 %    97.5 %
(Intercept)  1.326228 69.459529
COASAT1     -8.309492  3.618706
CLIM1        4.567371 12.559127
COND6      -14.308712 -3.695359
```

Finally, the confidence interval for the model coefficient shows that for the estimated slope value, the true slope is between 97.5% and 2.5%. For example, the **COASAT1** slope value = -2.34 in the model and the true slope is between -8.30 and 3.61. The VIF of a predictor is a measure of how easy it is to predict from a linear regression using the other predictors. All the predictor values are close to 1 showing sustainability.

Question 4: Regression model – with one dependent variable, one independent variable, and one interacting variable

The linear Regression model is carried out to predict **CLIM1(comfortability around other gym members)** based on **one independent variable (Gender) and one interacting variable (COASAT4 – Coach enthusiasm and energy level)**. The main objective is to determine how the comfortability of clients in the gym is dependent on the interacting variable.

In order to check the significance of the variables we have constructed models with and without interacting variables and below are the observations,

1. **Model without interacting:** Shows p = 0.64 which is greater than 0.05 showing the membership tenure (intercept) does not significantly depend on the Gender (slope).

2. **Model with interacting variable:** Shows p < 0.05 for both the variables showing the intercept is significantly dependent on the slope values. So after the introduction of interacting variables, p-values of **Gender: COASAT4 = 0.04** and **COASAT4 = 0.01** respectively.

```
Call:
lm(formula = CLIM1 ~ Gender, data = dataset_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4545 -0.3836  0.5454  0.6164  0.6164

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.45455    0.12710  35.048   <2e-16 ***
Gender2     -0.07098    0.15316  -0.463    0.644
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7301 on 104 degrees of freedom
Multiple R-squared:  0.002061,  Adjusted R-squared:  -0.007534
F-statistic: 0.2148 on 1 and 104 DF,  p-value: 0.644
```

```
Call:
lm(formula = CLIM1 ~ Gender * COASAT4, data = dataset_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7090 -0.4124  0.2910  0.5876  0.8508

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.2519     1.9680  -0.128   0.8984
Gender          2.1624     1.1217   1.928   0.0567 .
COASAT4         1.0515     0.4289   2.452   0.0159 *
Gender:COASAT4 -0.4918     0.2441  -2.015   0.0466 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7105 on 102 degrees of freedom
Multiple R-squared:  0.07306,   Adjusted R-squared:  0.0458
F-statistic:  2.68 on 3 and 102 DF,  p-value: 0.05087
```

The model comparison operation can be used to determine which model performed better in terms of prediction. The performance score, RMSE value, R2, and adjacent R2 value will determine the model performance, and below are the values obtained from model1 and model 2,

Name	Model	R2	R2 (adj.)	RMSE	Sigma	AIC weights	AICc weights	BIC weights	Performance-Score
lm_model2	lm	0.073	0.046	0.697	0.711	0.871	0.849	0.320	85.71%
lm_model1	lm	0.002	-0.008	0.723	0.730	0.129	0.151	0.680	14.29%

- From the table we can observe, the performance score for **lm_model2 = 85.71%** stating it is a better model in predicting the membership tenure with independent variables (Age and Gender) compared to **lm_model 1 = 14.29%** (only Gender)
- Lower RMSE score means better the prediction model and we can see the RMSE **of lm_model2 = 0.69** which is lower than the RMSE of **lm_model1 = 0.73**

A plot of the model performance and ANOVA test is further carried out to check the significance and below are the output and observation,

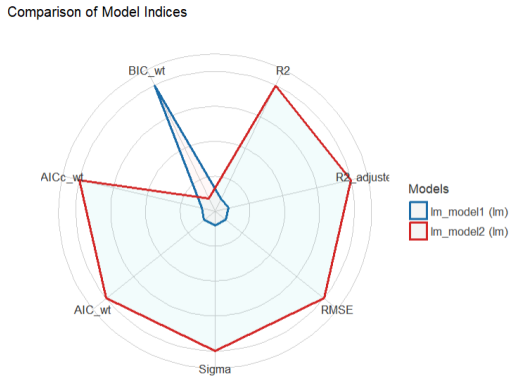


Figure 3: Visualization of Model Comparison

```
> summary(aov(lm_model2))
                Df Sum Sq Mean Sq F value Pr(>F)
Gender           1   0.11  0.1145   0.227 0.6349
COASAT4          1   1.90  1.8952   3.754 0.0555 .
Gender:COASAT4   1   2.05  2.0494   4.059 0.0466 *
Residuals      102  51.50  0.5049
```

Conclusion: We can see that Gender individually does not contribute a significant difference in predicting membership tenure because p > 0.05 but when the independent variable and the interacting variable are used together then p < 0.05 shows significance in predicting Membership Tenure.