Lab 1-Write Up and R Code

**Upload and Store:**

library(xlsx)

data<-read.xlsx("S3001 Lab 3 Dataset.xlsx",1) #reads data into a variable called data

**Descriptive Statistics:**

head(data) #gives a preview of the data

names(data) #lists the variables present

summary(data) #generates a 6-point summary of each independent variable

A screenshot of a computer

Description automatically generated with medium confidence

**Visualizing the Data:**

library(dplyr)

data2<-select(data,-Email,-Address,-Avatar) #this function will remove non-numerical data

cor(data2) #produces the result below

Text

Description automatically generated

* The “r” values were calculated in a pairwise fashion.
* The strongest relationships exist between Time on App, Average Session Length, Length of Membership and Yearly Amount Spent.
* The strongest relationship exists between Yearly Amount Spent and Length of Membership with a correlation co-efficient of 0.809.

library(car) #library contains scatterplot function

scatplot<- scatterplotMatrix(data2) #produces the matrix below

Diagram

Description automatically generated

* The scatterplot shows the distribution of the variables in pairs of 2. 4 of the 5 distributions are roughly normal.
* Strong positive correlation is also observed between Yearly Amount Spent and Length of Membership as mentioned above. This is further confirmed by the matrix. A simple linear model is expected to be sufficient using these two variables.

**Linear Model:**

lmod<-lm(Yearly.Amount.Spent~Length.of.Membership,data) #creates a linear model and stores it into the variable lmod

x11() #allows the coder to write code while visualizing it simultaneously

library(ggplot2)

plot(data2$Length.of.Membership, data2$Yearly.Amount.Spent, main=" Yearly Amount Spent Vs Length of Membership ", xlab=" Length.of.Membership ", ylab=" Yearly.Amount.Spent ") #plots a scatter plot of the selected variables of interest

abline(lm(Yearly.Amount.Spent ~ Length.of.Membership, data=data2), col='blue') #plots the linear model over the scatterplot

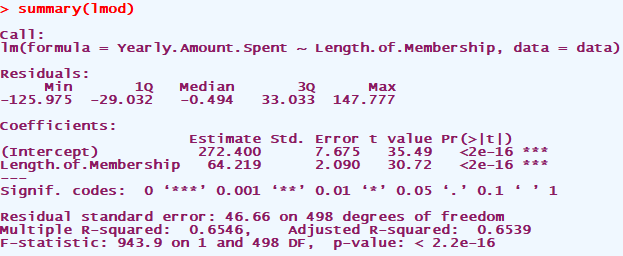
Chart, scatter chart

Description automatically generated

* Scatterplot of selected variables with the linear model layered on top. A linear model fits the data excellently. It’s time to examine the values from the summary.

**Analysing Model Characteristics:**

summary(lmod)



* Note that the least squares estimate for beta here (the slope) has a p-value of <2x10^-16 which is much less than at a 95% confidence level.
* The same can be said about the least square estimate for the intercept .
* The multiple R-squared value tells us that approximately 65% of the variation in the yearly amount spent is due to the length of time someone has been a member.
* Residuals are reasonable balanced on either side of the distribution(their sum appears to tend to 0), this supports the claim of normality on the part of the distribution.

**Finding the Confidence Interval For the Model Estimates:**

confint(lmod) #gives a 95% confidence interval for the least square model estimates

A picture containing company name

Description automatically generated

* The intercept lies between 257.32 and 287.45 at 95% confidence. This means that the cost of membership is those two costs.
* The slope of the solution (beta hat) lies between 60.11 and 68.32 with 95% confidence.