assignment 1

Load data

avocado=read.csv(file =“https://github.com/KazuMaeshima/Group-9-/raw/main/avocado.csv”, header =TRUE) ##Provide a introduction of your analysis in the .RMD file so it can be produced in the output # this codes will introduce us how to use R Studio as part of our day to day data analysis. It will produce variables, mean,median,mode, show and manipulate data and plot graphs using ggplot2 ## head str(avocado) ## Print the structure of your dataset. print(avocado) ##List the variables in your dataset names(avocado) ##Print the top 15 rows of your dataset. head(avocado,15) ##Write a user defined function using any of the variables from the data set m <- c(45,34,34,34,67)

getmode <- function(m) { uniqv <- unique(m) uniqv[which.max(tabulate(match(m, uniqv)))] } getmode(m) ##Use data manipulation techniques and filter rows based on any logical criteria that exist in your dataset filter(Avocado,AveragePrice<1) ##Identify the dependent & independent variables and use reshaping techniques and create a new data frame by joining those variablesfrom your dataset.
# Create a new dataset with the selected columns bags <- as.data.frame(avocado %>% select(Total.Bags,Small.Bags,Large.Ba
##Remove missing values in your dataset. na.omit(Avocado) ##Identify and remove duplicated data in your dataset. avocado[!duplicated(avocado),] ##Reorder multiple rows in descending order avocado %>% arrange(desc(AveragePrice)) ##Rename some of the column names in your dataset. head(avocado) m <- avocado dim(m) col_name <- paste(“Col”, 1:14, sep = "") head(m) names(m) <-col_name head(m)

##Add new variables in your data frame by using a mathematical function (for e.g. –multiply an existing column by 2 and add it as a new variable to your data frame) #Create new variable by mutliplying an existing column by 2

avocado$Doubleyear = avocado$year*2

##Create a training set using random number generator engine. # Initiate random number generator engine

set.seed(1234)

Select 80% rows from the main dataset as the training set

training = avocado %>% sample_frac(0.8,replace=FALSE)

#Print the summary statistics of your dataset.

summary(avocado)

##Use any of the numerical variables from the dataset and perform the following statistical functions. Mean mean(avocado$Large.Bags)

##Median median(avocado$Total.Bags)

##Mode v <- c(avocado$AveragePrice) # Calculate the mode using the user defined function result <- getmode(v) print(result)

##Range range(avocado$Total.Bags)

##Plot a scatter plot for any 2 variables in your dataset. ggplot(data = avocado, aes(x = Total.Bags, y=AveragePrice))+geom_point()

##Plot a bar plot for any 1 variables in your dataset ggplot(data = avocado, aes(x = AveragePrice))+geom_bar()

##Find the correlation between any 2 variables by applying least square linear regression model. library(ggpubr)

ggscatter(avocado, x="Total.Bags", y="AveragePrice", add="reg.line", conf.int = TRUE, cor.coef = TRUE, cor.method = "pearson", xlab = "Total Bags",ylab = "Average Price")

##Provide a conclusion of your analysis if any in the .RMD file. #there are different types of avocados. Each avocados corresponds with their respective price based on their size and each country prices their avocados differently