CKME 136 X10-DATA ANALYTICS: CAPSTONE PROJECT

Ekaete Czub

February 10, 2019

## Installation of relevant packages.

#install.packages("ISwR")  
#install.packages("caret")  
#install.packages("MASS")  
#install.packages("lubridate")  
#install.packages("stringr")  
#install.packages("corrplot")  
#install.packages("ggcorrplot")  
#install.packages("Metrics")  
library(ISwR)  
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(MASS)  
library(lubridate)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

library(stringr)  
library(corrplot) #### <--------------- Added for correlation plot

## corrplot 0.84 loaded

library(ggcorrplot) #### <--------------- Added for correlation plot  
library(Metrics)

##   
## Attaching package: 'Metrics'

## The following objects are masked from 'package:caret':  
##   
## precision, recall

## Read “Avocado.csv file” from folder or website:

Avocado <- read.csv(file.choose(), header = T, sep = ",", stringsAsFactors = F, na.strings = c(NA,"","?"))  
avobk <- Avocado

## Cleaning <data:correcting> column names

#names(Avocado)   
#names(Avocado)[1]<- "No"  
names(Avocado)[names(Avocado) =="X"]<- "No"  
names(Avocado)[names(Avocado) =="X4046"]<-"PLU4046"  
names(Avocado)[names(Avocado) =="X4225"]<-"PLU4225"  
names(Avocado)[names(Avocado) =="X4770"]<-"PLU4770"  
#head(Avocado)  
str(Avocado)

## 'data.frame': 18249 obs. of 14 variables:  
## $ No : int 0 1 2 3 4 5 6 7 8 9 ...  
## $ Date : chr "2015-12-27" "2015-12-20" "2015-12-13" "2015-12-06" ...  
## $ AveragePrice: num 1.33 1.35 0.93 1.08 1.28 1.26 0.99 0.98 1.02 1.07 ...  
## $ Total.Volume: num 64237 54877 118220 78992 51040 ...  
## $ PLU4046 : num 1037 674 795 1132 941 ...  
## $ PLU4225 : num 54455 44639 109150 71976 43838 ...  
## $ PLU4770 : num 48.2 58.3 130.5 72.6 75.8 ...  
## $ Total.Bags : num 8697 9506 8145 5811 6184 ...  
## $ Small.Bags : num 8604 9408 8042 5677 5986 ...  
## $ Large.Bags : num 93.2 97.5 103.1 133.8 197.7 ...  
## $ XLarge.Bags : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ type : chr "conventional" "conventional" "conventional" "conventional" ...  
## $ year : int 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...  
## $ region : chr "Albany" "Albany" "Albany" "Albany" ...

any(is.na(Avocado))

## [1] FALSE

# column: “No” - Categorical column <— This column will be removed …..

## Removing column 1(No)  
Avocado<- Avocado[, -1]   
str(Avocado)

## 'data.frame': 18249 obs. of 13 variables:  
## $ Date : chr "2015-12-27" "2015-12-20" "2015-12-13" "2015-12-06" ...  
## $ AveragePrice: num 1.33 1.35 0.93 1.08 1.28 1.26 0.99 0.98 1.02 1.07 ...  
## $ Total.Volume: num 64237 54877 118220 78992 51040 ...  
## $ PLU4046 : num 1037 674 795 1132 941 ...  
## $ PLU4225 : num 54455 44639 109150 71976 43838 ...  
## $ PLU4770 : num 48.2 58.3 130.5 72.6 75.8 ...  
## $ Total.Bags : num 8697 9506 8145 5811 6184 ...  
## $ Small.Bags : num 8604 9408 8042 5677 5986 ...  
## $ Large.Bags : num 93.2 97.5 103.1 133.8 197.7 ...  
## $ XLarge.Bags : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ type : chr "conventional" "conventional" "conventional" "conventional" ...  
## $ year : int 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...  
## $ region : chr "Albany" "Albany" "Albany" "Albany" ...

# column:Date - Categorical column

### names(Avocado)[2]  
names(Avocado)[1] ### <----- Date becomes the first column because the first column is removed..

## [1] "Date"

### Converting the date column from character to date  
Avo\_Date <- Avocado$Date  
Avo\_Date <- as.Date(Avo\_Date, format= "%Y-%m-%d")  
Avocado$Date <- Avo\_Date  
Avocado$Month <- month(Avocado$Date)  
Avocado$Month <- as.factor(Avocado$Month)  
Avocado$Day <- day(Avocado$Date)  
Avocado$Day <- as.factor(Avocado$Day)  
Avocado$year <- as.factor(Avocado$year)  
##Avocado$Week <- as.factor(week(Avocado$Date))  
Avocado <- Avocado[-1]

## column: AveragePrice - Numeric column

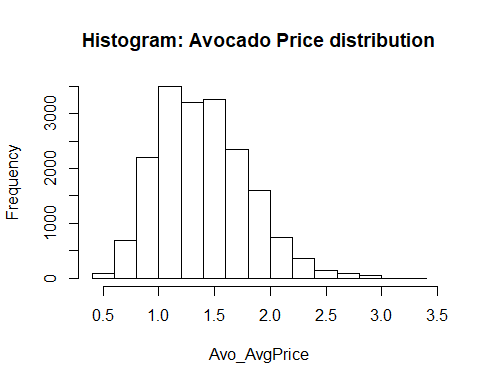
names(Avocado)[1]

## [1] "AveragePrice"

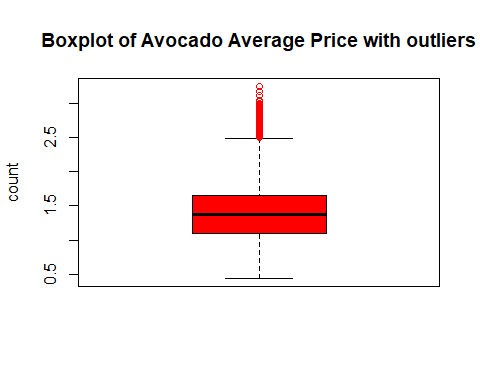
Avo\_AvgPrice <- Avocado$AveragePrice  
summary(Avo\_AvgPrice)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.440 1.100 1.370 1.406 1.660 3.250

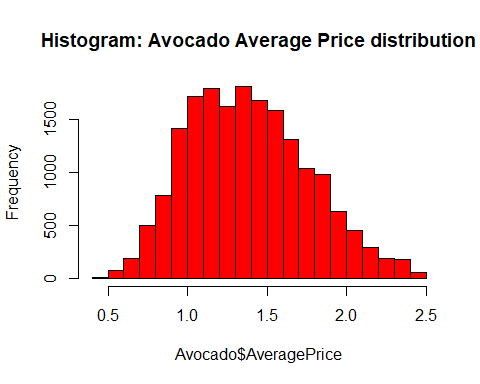
hist(Avo\_AvgPrice, breaks = 20, main="Histogram: Avocado Price distribution")### ----> This goes to the INITIAL analysis <---



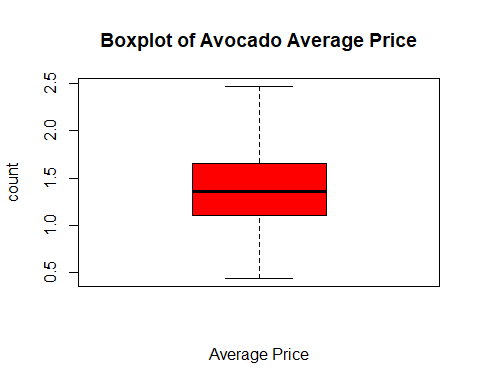
boxplot(Avo\_AvgPrice, main = "Boxplot of Avocado Average Price with outliers", ylab="count",horizontal = F, outcol = "red", col="red") ### ----> This goes to the INITIAL analysis <---



pout <- boxplot.stats(Avo\_AvgPrice)$out  
while (length(pout)>0) {  
 outrecs <- which(Avo\_AvgPrice %in% pout)  
 Avo\_AvgPrice[outrecs] <- NA  
 Avo\_AvgPrice[is.na(Avo\_AvgPrice)] <- median(Avo\_AvgPrice, na.rm = T)  
 pout <- boxplot.stats(Avo\_AvgPrice)$out  
}  
Avocado$AveragePrice<-Avo\_AvgPrice  
hist(Avocado$AveragePrice, main= "Histogram: Avocado Average Price distribution",col = "red") ### ----> This goes to the FINAL analysis <---



boxplot(Avocado$AveragePrice, main ="Boxplot of Avocado Average Price", ylab="count",xlab="Average Price", col="red") ### ----> This goes to the FINAL analysis <---



summary(Avo\_AvgPrice)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.44 1.10 1.36 1.39 1.65 2.47

# column: Total.Volume - Numeric column

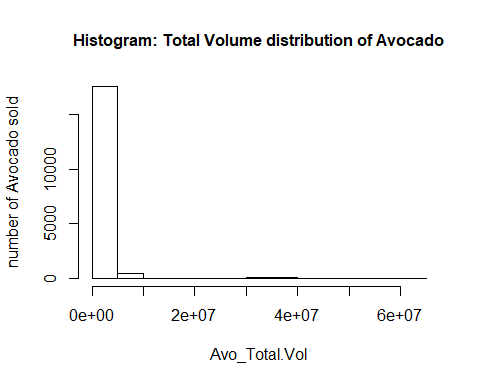
names(Avocado)[2]

## [1] "Total.Volume"

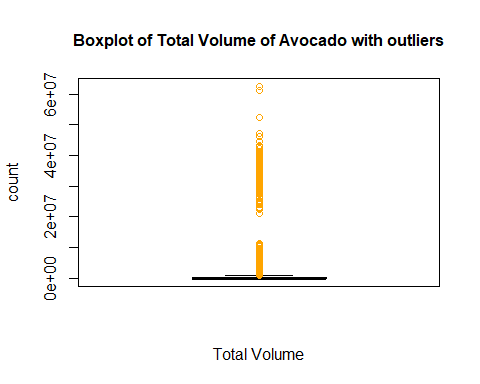
Avo\_Total.Vol <- Avocado$Total.Volume  
summary(Avo\_Total.Vol)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 85 10839 107377 850644 432962 62505647

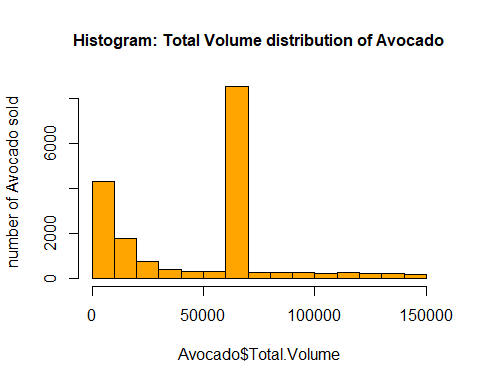
hist(Avo\_Total.Vol, main= "Histogram: Total Volume distribution of Avocado", cex.main=1, ylab = "number of Avocado sold",cex.lab=1)### ----> This goes to the INITIAL analysis <---



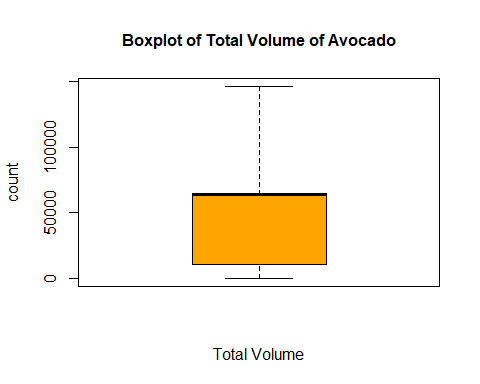
boxplot(Avo\_Total.Vol, main = "Boxplot of Total Volume of Avocado with outliers", cex.main=1, ylab = "count",cex.axis=1, xlab="Total Volume", cex.lab=1, horizontal = F, outcol = "orange", col= "orange") ### ----> This goes to the INITIAL analysis <---



poutTV<-boxplot.stats(Avo\_Total.Vol)$out  
while (length(poutTV)>0) {  
 outrecsTV <- which(Avo\_Total.Vol %in% poutTV)  
 Avo\_Total.Vol[outrecsTV] <- NA  
 Avo\_Total.Vol[is.na(Avo\_Total.Vol)] <- median(Avo\_Total.Vol, na.rm = T)  
 poutTV <- boxplot.stats(Avo\_Total.Vol)$out  
}  
Avocado$Total.Volume <- Avo\_Total.Vol  
hist(Avocado$Total.Volume, main= "Histogram: Total Volume distribution of Avocado",cex.main=1,ylab= "number of Avocado sold",cex.axis=1, col= "orange") ### ----> This goes to the FINAL analysis <---



boxplot(Avocado$Total.Volume, main = "Boxplot of Total Volume of Avocado", cex.main=1, ylab="count",cex.axis=1, xlab = "Total Volume",cex.lab=1, col="orange") ### ----> This goes to the FINAL analysis <---



summary(Avo\_Total.Vol)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 84.56 10838.58 63902.90 47243.99 65098.19 146422.18

# column: PLU4046- numeric column

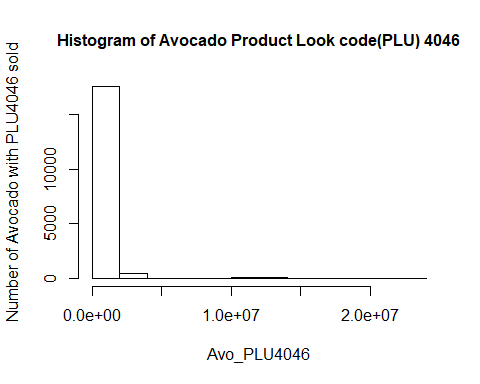
names(Avocado)[3]

## [1] "PLU4046"

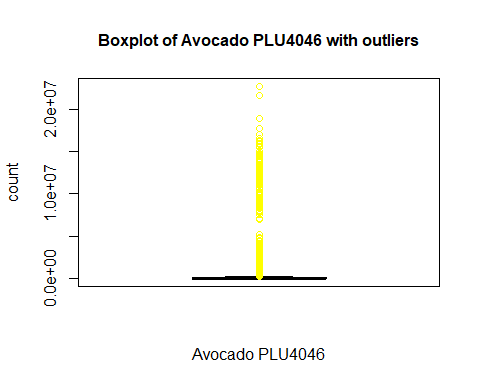
Avo\_PLU4046 <- Avocado$PLU4046  
summary(Avo\_PLU4046)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 854 8645 293008 111020 22743616

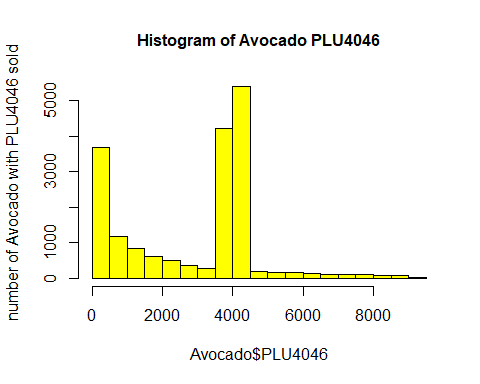
hist(Avo\_PLU4046, main = "Histogram of Avocado Product Look code(PLU) 4046", cex.main=1, ylab = "Number of Avocado with PLU4046 sold", cex.axis=1)### ----> This goes to the INITIAL analysis <---



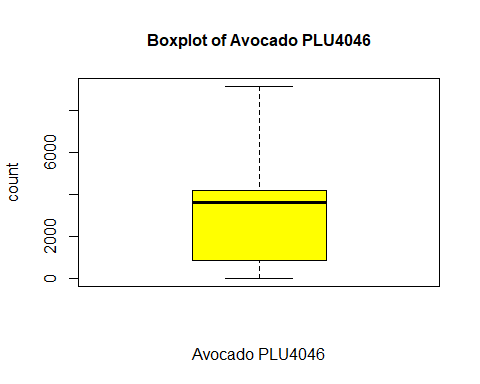
boxplot(Avo\_PLU4046, main = "Boxplot of Avocado PLU4046 with outliers", cex.main=1, ylab="count", cex.axis=1,xlab= "Avocado PLU4046", cex.lab=1,horizontal = F, outcol="yellow", col="yellow")### ----> This goes to the INITIAL analysis <---



poutC1<-boxplot.stats(Avo\_PLU4046)$out  
while (length(poutC1)>0) {  
 outrecsC1 <- which(Avo\_PLU4046 %in% poutC1)  
 Avo\_PLU4046[outrecsC1] <- NA  
 Avo\_PLU4046[is.na(Avo\_PLU4046)] <- median(Avo\_PLU4046, na.rm = T)  
 poutC1 <- boxplot.stats(Avo\_PLU4046)$out  
}  
Avocado$PLU4046<-Avo\_PLU4046  
hist(Avocado$PLU4046,main= "Histogram of Avocado PLU4046", cex.main=1,ylab= "number of Avocado with PLU4046 sold",cex.axis=1, col="yellow")### ----> This goes to the FINAL analysis <---



boxplot(Avocado$PLU4046, main="Boxplot of Avocado PLU4046", cex.main=1,ylab="count",cex.axis=1, xlab="Avocado PLU4046", cex.lab=1,col = "yellow")### ----> This goes to the FINAL analysis <---



summary(Avo\_PLU4046)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 854.1 3606.8 2865.0 4163.6 9127.1

# column: PLU4225 - numeric column

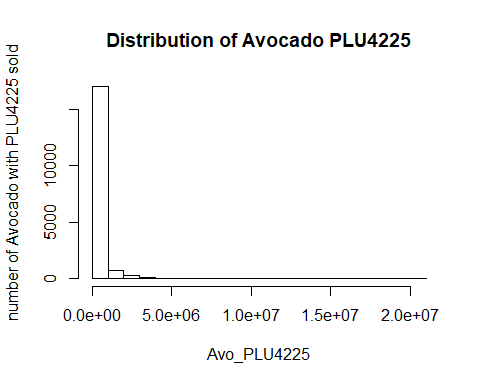
names(Avocado)[4]

## [1] "PLU4225"

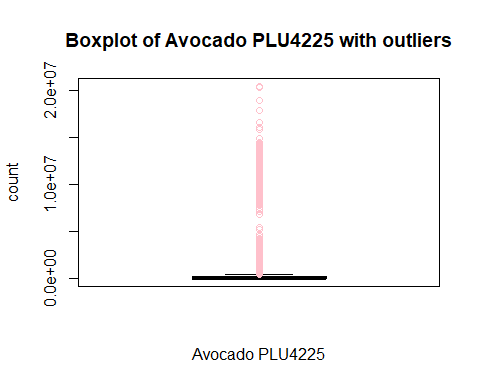
Avo\_PLU4225 <- Avocado$PLU4225  
summary(Avo\_PLU4225)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 3009 29061 295155 150207 20470573

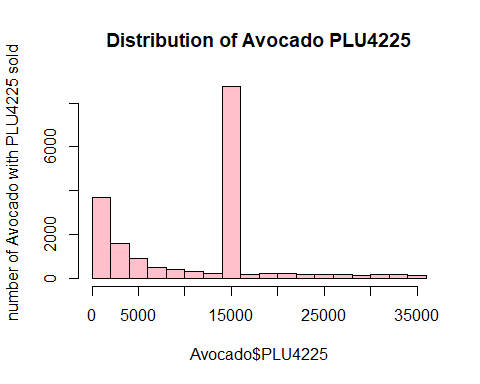
hist(Avo\_PLU4225, main="Distribution of Avocado PLU4225", ylab="number of Avocado with PLU4225 sold")### ----> This goes to the INITIAL analysis <---



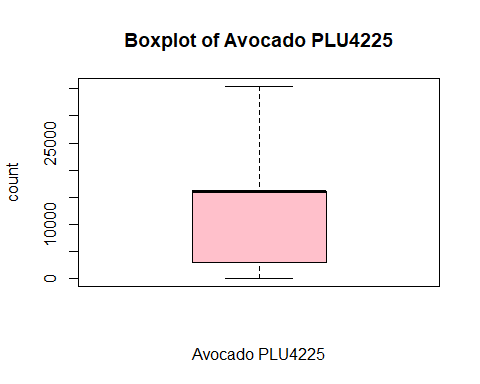
boxplot(Avo\_PLU4225, main = "Boxplot of Avocado PLU4225 with outliers",ylab="count", xlab= "Avocado PLU4225", horizontal = F, outcol="pink", col="pink")### ----> This goes to the INITIAL analysis <---



poutC2<-boxplot.stats(Avo\_PLU4225)$out  
while (length(poutC2)>0) {  
 outrecsC2 <- which(Avo\_PLU4225 %in% poutC2)  
 Avo\_PLU4225[outrecsC2] <- NA  
 Avo\_PLU4225[is.na(Avo\_PLU4225)] <- median(Avo\_PLU4225, na.rm = T)  
 poutC2 <- boxplot.stats(Avo\_PLU4225)$out  
}  
Avocado$PLU4225<-Avo\_PLU4225  
hist(Avocado$PLU4225, main="Distribution of Avocado PLU4225", ylab= "number of Avocado with PLU4225 sold",col="pink")### ----> This goes to the FINAL analysis <---



boxplot(Avocado$PLU4225, main="Boxplot of Avocado PLU4225", ylab="count", xlab= "Avocado PLU4225", col="pink")### ----> This goes to the FINAL analysis <---



summary(Avo\_PLU4225)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 3009 15989 11550 15989 35457

# column: PLU4770 - numeric column

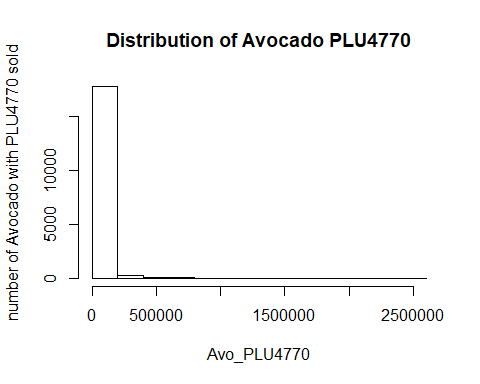
names(Avocado)[5]

## [1] "PLU4770"

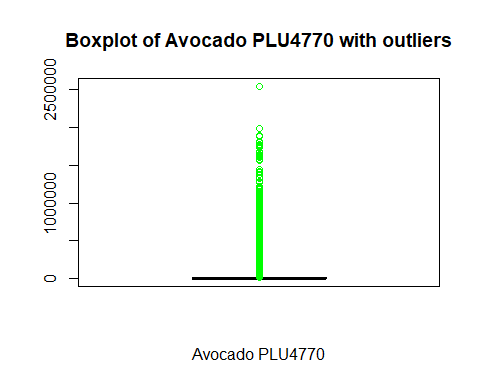
Avo\_PLU4770 <- Avocado$PLU4770  
summary(Avo\_PLU4770)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 185 22840 6243 2546439

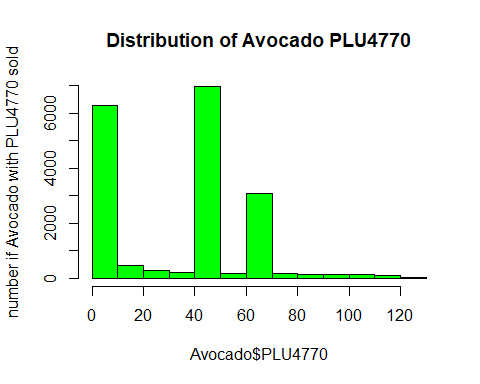
hist(Avo\_PLU4770, main="Distribution of Avocado PLU4770", ylab="number of Avocado with PLU4770 sold")### ----> This goes to the INITIAL analysis <---



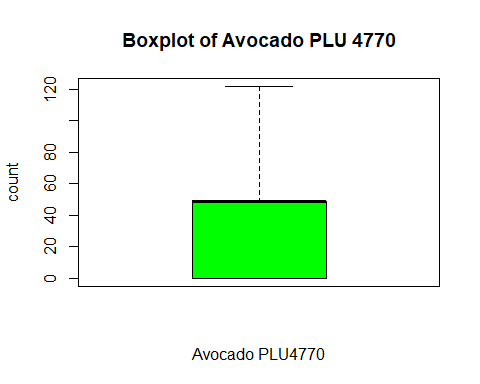
boxplot(Avo\_PLU4770, main = "Boxplot of Avocado PLU4770 with outliers", xlab= "Avocado PLU4770", horizontal = F, outcol="green", col="green")### ----> This goes to the INITIAL analysis <---



poutC3<-boxplot.stats(Avo\_PLU4770)$out  
while (length(poutC3)>0) {  
 outrecsC3 <- which(Avo\_PLU4770 %in% poutC3)  
 Avo\_PLU4770[outrecsC3] <- NA  
 Avo\_PLU4770[is.na(Avo\_PLU4770)] <- median(Avo\_PLU4770, na.rm = T)  
 poutC3 <- boxplot.stats(Avo\_PLU4770)$out  
}  
Avocado$PLU4770<-Avo\_PLU4770  
hist(Avocado$PLU4770, main="Distribution of Avocado PLU4770", ylab="number if Avocado with PLU4770 sold",col="green")### ----> This goes to the FINAL analysis <---



boxplot(Avocado$PLU4770, main = "Boxplot of Avocado PLU 4770", ylab="count",xlab="Avocado PLU4770", col="green")### ----> This goes to the FINAL analysis <---



summary(Avo\_PLU4770)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 48.81 35.01 48.81 121.93

# column: Total.Bags- numeric column

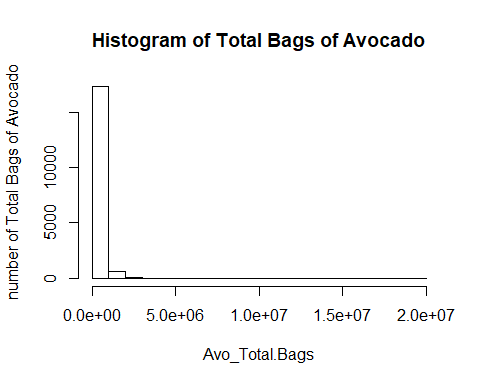
names(Avocado)[6]

## [1] "Total.Bags"

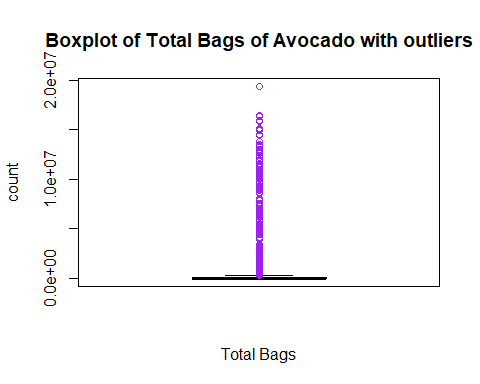
Avo\_Total.Bags <- Avocado$Total.Bags  
summary(Avo\_Total.Bags)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 5089 39744 239639 110783 19373134

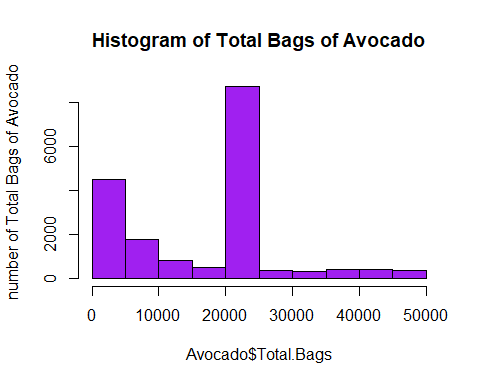
hist(Avo\_Total.Bags, main="Histogram of Total Bags of Avocado", ylab="number of Total Bags of Avocado")### ----> This goes to the INITIAL analysis <---



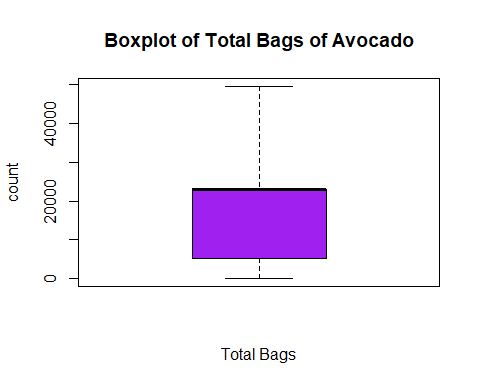
boxplot(Avo\_Total.Bags, main = "Boxplot of Total Bags of Avocado with outliers", ylab="count", xlab="Total Bags", horizontal = F, outcol = "purple", col="purple")### ----> This goes to the INITIAL analysis <---



poutTB<-boxplot.stats(Avo\_Total.Bags)$out  
while (length(poutTB)>0) {  
 outrecsTB <- which(Avo\_Total.Bags %in% poutTB)  
 Avo\_Total.Bags[outrecsTB] <- NA  
 Avo\_Total.Bags[is.na(Avo\_Total.Bags)] <- median(Avo\_Total.Bags, na.rm = T)  
 poutTB <- boxplot.stats(Avo\_Total.Bags)$out  
}  
Avocado$Total.Bags<-Avo\_Total.Bags  
hist(Avocado$Total.Bags, main="Histogram of Total Bags of Avocado", ylab="number of Total Bags of Avocado",col = "purple")### ----> This goes to the FINAL analysis <---



boxplot(Avocado$Total.Bags, main="Boxplot of Total Bags of Avocado", ylab="count", xlab="Total Bags", col="purple")### ----> This goes to the FINAL analysis <---



summary(Avo\_Total.Bags)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 5089 22905 17054 22905 49622

# column: Small.Bags- numeric column

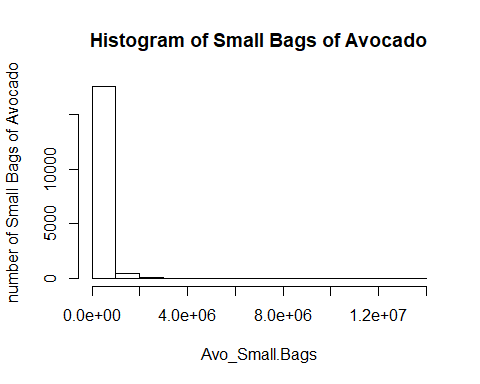
names(Avocado)[7]

## [1] "Small.Bags"

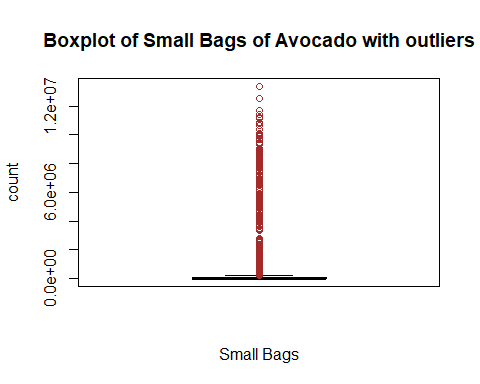
Avo\_Small.Bags <- Avocado$Small.Bags  
summary(Avo\_Small.Bags)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 2849 26363 182195 83338 13384587

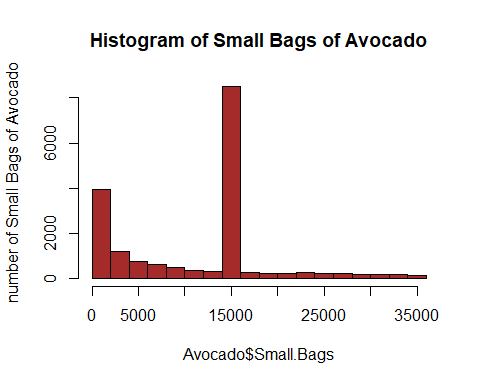
hist(Avo\_Small.Bags, main="Histogram of Small Bags of Avocado", ylab="number of Small Bags of Avocado")### ----> This goes to the INITIAL analysis <---



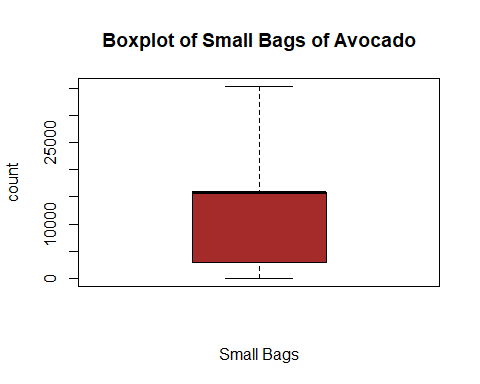
boxplot(Avo\_Small.Bags, main = "Boxplot of Small Bags of Avocado with outliers",ylab="count", xlab="Small Bags", horizontal = F, outcol = "brown", col="brown")### ----> This goes to the INITIAL analysis <---



poutSB<-boxplot.stats(Avo\_Small.Bags)$out  
while (length(poutSB)>0) {  
 outrecsSB <- which(Avo\_Small.Bags %in% poutSB)  
 Avo\_Small.Bags[outrecsSB] <- NA  
 Avo\_Small.Bags[is.na(Avo\_Small.Bags)] <- median(Avo\_Small.Bags, na.rm = T)  
 poutSB <- boxplot.stats(Avo\_Small.Bags)$out  
}  
Avocado$Small.Bags<-Avo\_Small.Bags  
hist(Avocado$Small.Bags, main="Histogram of Small Bags of Avocado" , ylab="number of Small Bags of Avocado", col="brown")### ----> This goes to the FINAL analysis <---



boxplot(Avocado$Small.Bags, main="Boxplot of Small Bags of Avocado", ylab="count", xlab="Small Bags", col="brown")### ----> This goes to the FINAL analysis <---



summary(Avo\_Small.Bags)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 2849 15847 11645 15847 35322

# column: Large.Bags - numeric column

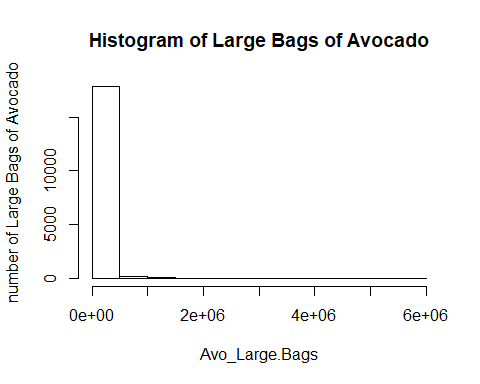
names(Avocado)[8]

## [1] "Large.Bags"

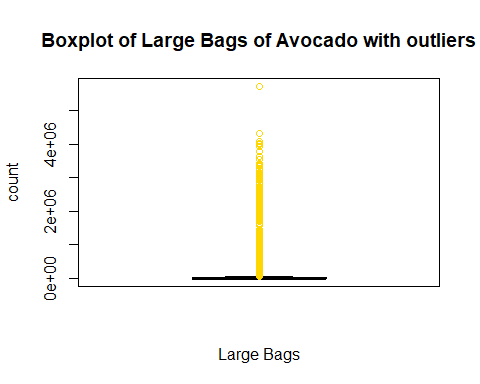
Avo\_Large.Bags <- Avocado$Large.Bags  
summary(Avo\_Large.Bags)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 127 2648 54338 22029 5719097

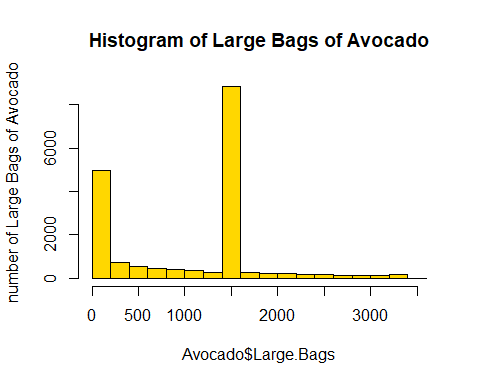
hist(Avo\_Large.Bags, main="Histogram of Large Bags of Avocado", ylab="number of Large Bags of Avocado")### ----> This goes to the INITIAL analysis <---



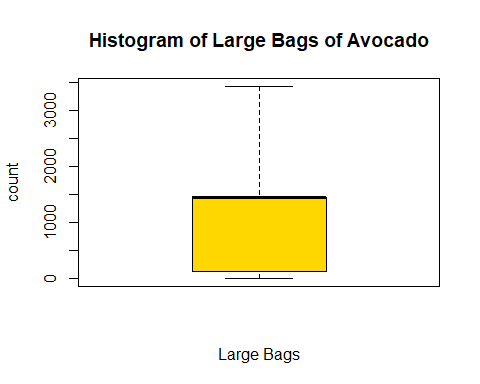
boxplot(Avo\_Large.Bags, main = "Boxplot of Large Bags of Avocado with outliers", ylab="count", xlab="Large Bags",horizontal = F, outcol = "gold", col="gold")### ----> This goes to the INITIAL analysis <---



poutLB<-boxplot.stats(Avo\_Large.Bags)$out  
while (length(poutLB)>0) {  
 outrecsLB <- which(Avo\_Large.Bags %in% poutLB)  
 Avo\_Large.Bags[outrecsLB] <- NA  
 Avo\_Large.Bags[is.na(Avo\_Large.Bags)] <- median(Avo\_Large.Bags, na.rm = T)  
 poutLB <- boxplot.stats(Avo\_Large.Bags)$out  
}  
Avocado$Large.Bags<-Avo\_Large.Bags  
hist(Avocado$Large.Bags, main="Histogram of Large Bags of Avocado", ylab="number of Large Bags of Avocado", col = "gold")### ----> This goes to the FINAL analysis <---



boxplot(Avocado$Large.Bags, main="Histogram of Large Bags of Avocado", ylab="count", xlab="Large Bags", col="gold")### ----> This goes to the FINAL analysis <---



summary(Avo\_Large.Bags)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 127.5 1446.0 1033.8 1446.0 3423.5

# column: XLarge.Bags - numeric column —-> THIS COLUMN WILL BE REMOVED <—–

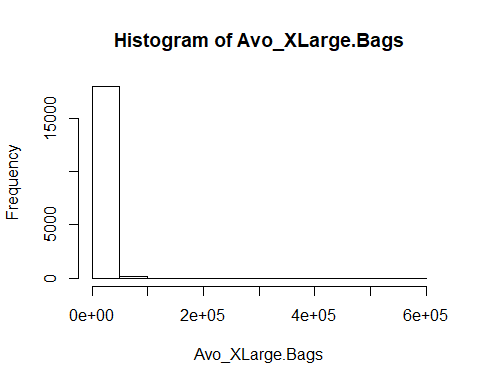
names(Avocado)[9]

## [1] "XLarge.Bags"

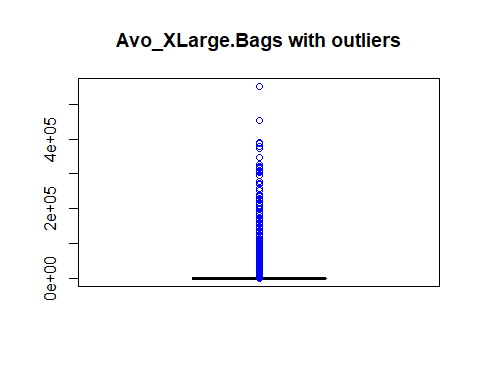
Avo\_XLarge.Bags <- Avocado$XLarge.Bags  
summary(Avo\_XLarge.Bags)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 0.0 0.0 3106.4 132.5 551693.7

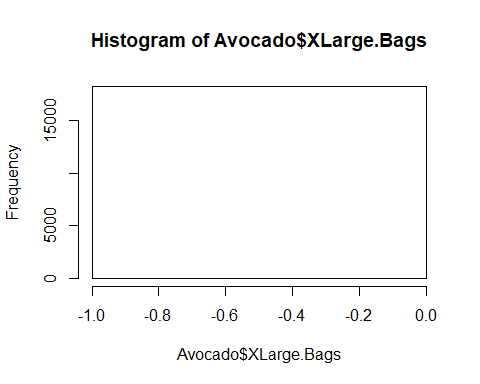
hist(Avo\_XLarge.Bags)### ----> This goes to the INITIAL analysis <---



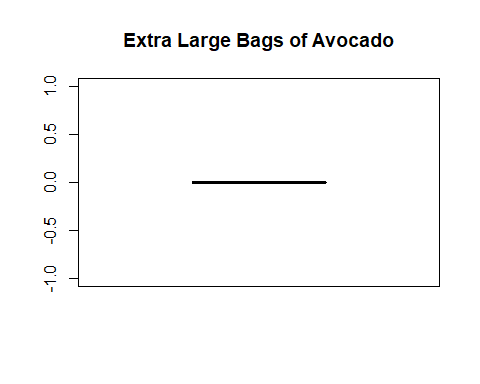
boxplot(Avo\_XLarge.Bags, main = "Avo\_XLarge.Bags with outliers", horizontal = F, outcol = "blue", col="blue")### ----> This goes to the INITIAL analysis <---



poutXLB<-boxplot.stats(Avo\_XLarge.Bags)$out  
while (length(poutXLB)>0) {  
 outrecsXLB <- which(Avo\_XLarge.Bags %in% poutXLB)  
 Avo\_XLarge.Bags[outrecsXLB] <- NA  
 Avo\_XLarge.Bags[is.na(Avo\_XLarge.Bags)] <- median(Avo\_XLarge.Bags, na.rm = T)  
 poutXLB <- boxplot.stats(Avo\_XLarge.Bags)$out  
}  
Avocado$XLarge.Bags<-Avo\_XLarge.Bags  
hist(Avocado$XLarge.Bags)



boxplot(Avocado$XLarge.Bags, main="Extra Large Bags of Avocado")



#Removing column 11(XLarge.Bags)  
Avocado$Total.Bags <- Avocado$Small.Bags + Avocado$Large.Bags  
Avocado<- Avocado[, -9]  
## str(Avocado)

### ———————————————————————————

### —> COLUMN NUMBERS CHANGED AGAIN AFTER REMOVING the column XLarge.Bags <——-

### ———————————————————————————

# column: type - categorical column

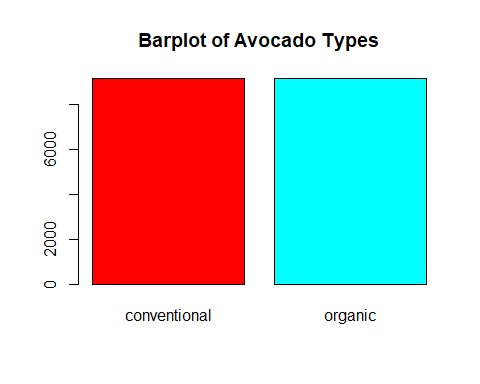
names(Avocado)[9]

## [1] "type"

Avo\_type <- Avocado$type  
table(Avo\_type)

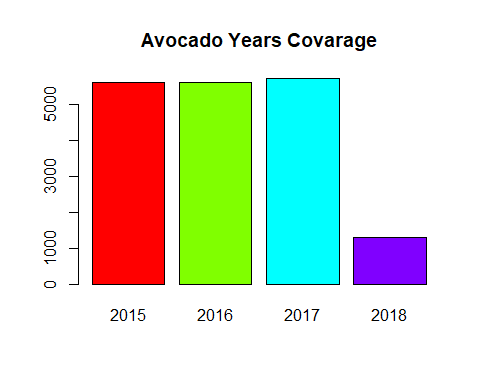
## Avo\_type  
## conventional organic   
## 9126 9123

type<-table(Avo\_type)  
barplot(type, main = "Barplot of Avocado Types", col=rainbow(2))



# column: year - categorical column

# names(Avocado)[10]  
# Avo\_year <- Avocado$year  
# Avo\_year <- factor(Avo\_year)  
#head(Avo\_year)  
Years <- table(Avocado$year)  
barplot(Years, main = "Avocado Years Covarage", col=rainbow(4))



# column: region - categorical column

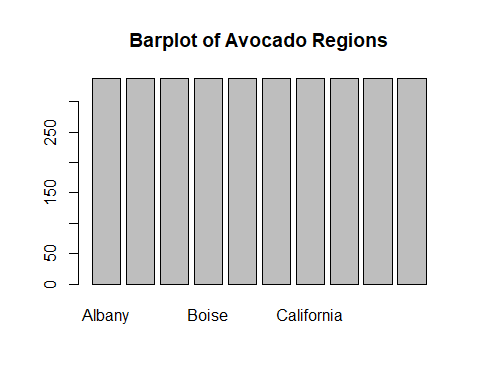
names(Avocado)[11]

## [1] "region"

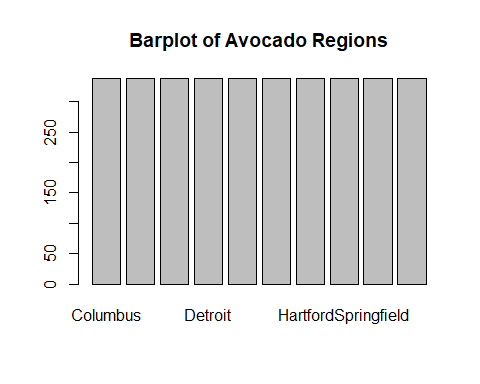
Avo\_region<- Avocado$region  
region<-table(Avo\_region, useNA = 'ifany')  
region

## Avo\_region  
## Albany Atlanta BaltimoreWashington   
## 338 338 338   
## Boise Boston BuffaloRochester   
## 338 338 338   
## California Charlotte Chicago   
## 338 338 338   
## CincinnatiDayton Columbus DallasFtWorth   
## 338 338 338   
## Denver Detroit GrandRapids   
## 338 338 338   
## GreatLakes HarrisburgScranton HartfordSpringfield   
## 338 338 338   
## Houston Indianapolis Jacksonville   
## 338 338 338   
## LasVegas LosAngeles Louisville   
## 338 338 338   
## MiamiFtLauderdale Midsouth Nashville   
## 338 338 338   
## NewOrleansMobile NewYork Northeast   
## 338 338 338   
## NorthernNewEngland Orlando Philadelphia   
## 338 338 338   
## PhoenixTucson Pittsburgh Plains   
## 338 338 338   
## Portland RaleighGreensboro RichmondNorfolk   
## 338 338 338   
## Roanoke Sacramento SanDiego   
## 338 338 338   
## SanFrancisco Seattle SouthCarolina   
## 338 338 338   
## SouthCentral Southeast Spokane   
## 338 338 338   
## StLouis Syracuse Tampa   
## 338 338 338   
## TotalUS West WestTexNewMexico   
## 338 338 335

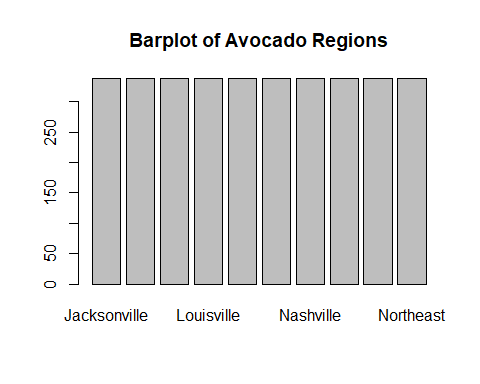
barplot(region[1:10], main = "Barplot of Avocado Regions")



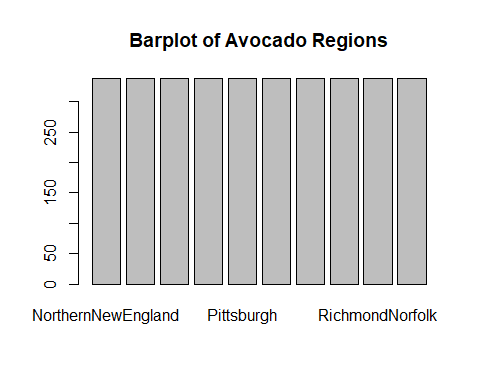
barplot(region[11:20], main = "Barplot of Avocado Regions")



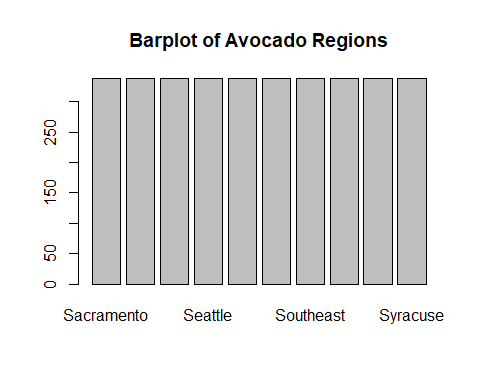
barplot(region[21:30], main = "Barplot of Avocado Regions")



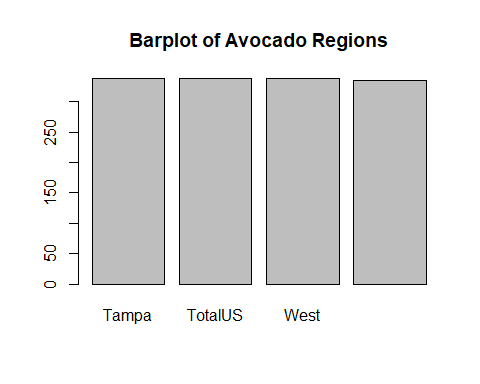
barplot(region[31:40], main = "Barplot of Avocado Regions")



barplot(region[41:50], main = "Barplot of Avocado Regions")

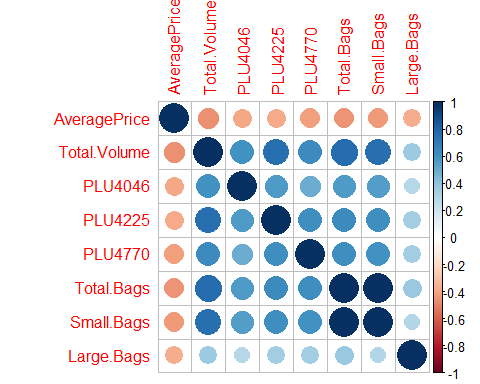


barplot(region[51:length(region)], main = "Barplot of Avocado Regions")

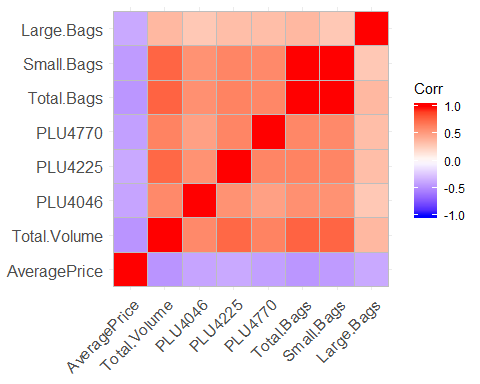


# correlation between numeric columns

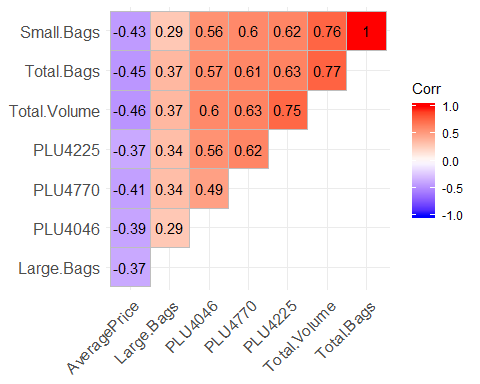
corr\_df<-Avocado[1:8]   
## plot(d) <--- Image is saved.   
corr\_mtrx <- cor(corr\_df)  
corrplot(corr\_mtrx)



ggcorrplot(corr\_mtrx)



ggcorrplot(corr\_mtrx, lab = T, hc.order = T, type = 'upper')



# Splitting dataset to 70% training and 30% test sets

Avocado$type <- as.factor(Avocado$type)  
Avocado$region <- as.factor(Avocado$region)  
Avocado\_Train <-sample(nrow(Avocado), floor(nrow(Avocado)\*0.7))  
train\_data <-Avocado[Avocado\_Train,]  
test\_data <-Avocado[-Avocado\_Train,]

# Building the Multiple Linear Regression Model

set.seed(123)  
Avocado\_Model <-lm(AveragePrice~., data=train\_data)  
summary(Avocado\_Model)

##   
## Call:  
## lm(formula = AveragePrice ~ ., data = train\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.94692 -0.13617 -0.00006 0.13742 0.92472   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.183e+00 2.109e-02 56.070 < 2e-16 \*\*\*  
## Total.Volume -4.362e-07 1.271e-07 -3.431 0.000603 \*\*\*  
## PLU4046 7.846e-06 1.759e-06 4.460 8.27e-06 \*\*\*  
## PLU4225 2.587e-06 4.605e-07 5.619 1.97e-08 \*\*\*  
## PLU4770 4.939e-04 1.324e-04 3.732 0.000191 \*\*\*  
## Total.Bags -5.486e-05 3.216e-06 -17.060 < 2e-16 \*\*\*  
## Small.Bags 5.207e-05 3.209e-06 16.226 < 2e-16 \*\*\*  
## Large.Bags NA NA NA NA   
## typeorganic 4.480e-01 7.892e-03 56.769 < 2e-16 \*\*\*  
## year2016 -4.267e-02 5.442e-03 -7.841 4.81e-15 \*\*\*  
## year2017 1.285e-01 5.435e-03 23.638 < 2e-16 \*\*\*  
## year2018 8.358e-02 1.065e-02 7.845 4.67e-15 \*\*\*  
## regionAtlanta -2.236e-01 2.145e-02 -10.423 < 2e-16 \*\*\*  
## regionBaltimoreWashington -2.093e-02 2.190e-02 -0.956 0.339311   
## regionBoise -1.936e-01 2.097e-02 -9.232 < 2e-16 \*\*\*  
## regionBoston 5.110e-05 2.162e-02 0.002 0.998114   
## regionBuffaloRochester -4.701e-02 2.118e-02 -2.219 0.026505 \*   
## regionCalifornia -1.467e-01 2.164e-02 -6.779 1.27e-11 \*\*\*  
## regionCharlotte 1.008e-02 2.155e-02 0.468 0.640044   
## regionChicago -1.538e-03 2.158e-02 -0.071 0.943185   
## regionCincinnatiDayton -2.894e-01 2.150e-02 -13.459 < 2e-16 \*\*\*  
## regionColumbus -2.754e-01 2.157e-02 -12.769 < 2e-16 \*\*\*  
## regionDallasFtWorth -4.622e-01 2.229e-02 -20.737 < 2e-16 \*\*\*  
## regionDenver -3.041e-01 2.235e-02 -13.606 < 2e-16 \*\*\*  
## regionDetroit -2.406e-01 2.144e-02 -11.224 < 2e-16 \*\*\*  
## regionGrandRapids -1.637e-02 2.081e-02 -0.787 0.431510   
## regionGreatLakes -1.763e-01 2.176e-02 -8.100 5.98e-16 \*\*\*  
## regionHarrisburgScranton -1.518e-02 2.136e-02 -0.711 0.477216   
## regionHartfordSpringfield 2.003e-01 2.153e-02 9.305 < 2e-16 \*\*\*  
## regionHouston -5.012e-01 2.189e-02 -22.896 < 2e-16 \*\*\*  
## regionIndianapolis -1.933e-01 2.146e-02 -9.010 < 2e-16 \*\*\*  
## regionJacksonville -4.835e-02 2.146e-02 -2.253 0.024290 \*   
## regionLasVegas -1.833e-01 2.145e-02 -8.547 < 2e-16 \*\*\*  
## regionLosAngeles -3.243e-01 2.164e-02 -14.988 < 2e-16 \*\*\*  
## regionLouisville -2.396e-01 2.146e-02 -11.167 < 2e-16 \*\*\*  
## regionMiamiFtLauderdale -1.349e-01 2.118e-02 -6.370 1.96e-10 \*\*\*  
## regionMidsouth -1.127e-01 2.224e-02 -5.066 4.12e-07 \*\*\*  
## regionNashville -3.229e-01 2.146e-02 -15.046 < 2e-16 \*\*\*  
## regionNewOrleansMobile -2.256e-01 2.126e-02 -10.608 < 2e-16 \*\*\*  
## regionNewYork 1.708e-01 2.175e-02 7.851 4.46e-15 \*\*\*  
## regionNortheast 6.237e-02 2.177e-02 2.865 0.004173 \*\*   
## regionNorthernNewEngland -7.119e-02 2.144e-02 -3.321 0.000900 \*\*\*  
## regionOrlando -4.704e-02 2.160e-02 -2.178 0.029435 \*   
## regionPhiladelphia 9.097e-02 2.129e-02 4.273 1.94e-05 \*\*\*  
## regionPhoenixTucson -3.458e-01 2.190e-02 -15.789 < 2e-16 \*\*\*  
## regionPittsburgh -1.526e-01 2.175e-02 -7.013 2.45e-12 \*\*\*  
## regionPlains -8.442e-02 2.203e-02 -3.832 0.000127 \*\*\*  
## regionPortland -2.818e-01 2.130e-02 -13.232 < 2e-16 \*\*\*  
## regionRaleighGreensboro -4.560e-02 2.143e-02 -2.128 0.033386 \*   
## regionRichmondNorfolk -2.352e-01 2.156e-02 -10.911 < 2e-16 \*\*\*  
## regionRoanoke -2.571e-01 2.123e-02 -12.112 < 2e-16 \*\*\*  
## regionSacramento -1.180e-02 2.164e-02 -0.545 0.585715   
## regionSanDiego -1.860e-01 2.173e-02 -8.563 < 2e-16 \*\*\*  
## regionSanFrancisco -2.614e-03 2.143e-02 -0.122 0.902899   
## regionSeattle -1.760e-01 2.146e-02 -8.203 2.56e-16 \*\*\*  
## regionSouthCarolina -1.278e-01 2.151e-02 -5.941 2.91e-09 \*\*\*  
## regionSouthCentral -3.921e-01 2.199e-02 -17.832 < 2e-16 \*\*\*  
## regionSoutheast -1.184e-01 2.205e-02 -5.367 8.15e-08 \*\*\*  
## regionSpokane -1.737e-01 2.136e-02 -8.134 4.55e-16 \*\*\*  
## regionStLouis -1.118e-01 2.157e-02 -5.184 2.20e-07 \*\*\*  
## regionSyracuse -1.680e-03 2.125e-02 -0.079 0.936973   
## regionTampa -1.509e-01 2.184e-02 -6.910 5.06e-12 \*\*\*  
## regionTotalUS -2.045e-01 2.180e-02 -9.381 < 2e-16 \*\*\*  
## regionWest -2.715e-01 2.242e-02 -12.110 < 2e-16 \*\*\*  
## regionWestTexNewMexico -2.989e-01 2.137e-02 -13.987 < 2e-16 \*\*\*  
## Month2 -2.614e-02 9.324e-03 -2.804 0.005060 \*\*   
## Month3 3.134e-02 9.413e-03 3.330 0.000871 \*\*\*  
## Month4 7.500e-02 1.064e-02 7.052 1.85e-12 \*\*\*  
## Month5 4.175e-02 9.595e-03 4.352 1.36e-05 \*\*\*  
## Month6 1.039e-01 1.047e-02 9.930 < 2e-16 \*\*\*  
## Month7 1.657e-01 1.030e-02 16.082 < 2e-16 \*\*\*  
## Month8 1.734e-01 1.064e-02 16.303 < 2e-16 \*\*\*  
## Month9 2.333e-01 1.064e-02 21.920 < 2e-16 \*\*\*  
## Month10 2.553e-01 1.006e-02 25.371 < 2e-16 \*\*\*  
## Month11 1.663e-01 1.028e-02 16.171 < 2e-16 \*\*\*  
## Month12 1.847e-02 1.022e-02 1.808 0.070611 .   
## Day2 4.130e-02 1.744e-02 2.368 0.017922 \*   
## Day3 1.846e-02 1.566e-02 1.179 0.238357   
## Day4 2.054e-02 1.503e-02 1.366 0.171961   
## Day5 -3.378e-02 1.539e-02 -2.195 0.028213 \*   
## Day6 6.693e-02 1.633e-02 4.098 4.19e-05 \*\*\*  
## Day7 4.548e-02 1.622e-02 2.804 0.005054 \*\*   
## Day8 3.283e-02 1.499e-02 2.190 0.028567 \*   
## Day9 1.647e-02 1.767e-02 0.932 0.351348   
## Day10 3.233e-02 1.561e-02 2.071 0.038406 \*   
## Day11 2.595e-02 1.524e-02 1.703 0.088599 .   
## Day12 -4.314e-03 1.534e-02 -0.281 0.778496   
## Day13 7.675e-02 1.632e-02 4.704 2.58e-06 \*\*\*  
## Day14 9.430e-02 1.633e-02 5.776 7.84e-09 \*\*\*  
## Day15 4.788e-02 1.496e-02 3.200 0.001376 \*\*   
## Day16 4.198e-02 1.738e-02 2.416 0.015711 \*   
## Day17 7.406e-02 1.570e-02 4.717 2.42e-06 \*\*\*  
## Day18 4.370e-02 1.516e-02 2.883 0.003948 \*\*   
## Day19 6.480e-03 1.533e-02 0.423 0.672433   
## Day20 5.611e-02 1.620e-02 3.464 0.000534 \*\*\*  
## Day21 9.303e-02 1.619e-02 5.745 9.38e-09 \*\*\*  
## Day22 1.761e-02 1.488e-02 1.183 0.236752   
## Day23 5.438e-02 1.749e-02 3.109 0.001879 \*\*   
## Day24 8.837e-02 1.563e-02 5.653 1.61e-08 \*\*\*  
## Day25 5.721e-02 1.507e-02 3.797 0.000147 \*\*\*  
## Day26 1.761e-02 1.525e-02 1.154 0.248320   
## Day27 7.224e-02 1.630e-02 4.431 9.47e-06 \*\*\*  
## Day28 9.031e-02 1.636e-02 5.521 3.44e-08 \*\*\*  
## Day29 -5.834e-03 1.575e-02 -0.370 0.711064   
## Day30 7.590e-02 1.762e-02 4.308 1.66e-05 \*\*\*  
## Day31 4.366e-02 1.742e-02 2.506 0.012233 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.227 on 12669 degrees of freedom  
## Multiple R-squared: 0.6438, Adjusted R-squared: 0.6409   
## F-statistic: 220.2 on 104 and 12669 DF, p-value: < 2.2e-16

prediction\_prices<-predict(Avocado\_Model, interval="prediction", newdata=test\_data)

## Warning in predict.lm(Avocado\_Model, interval = "prediction", newdata =  
## test\_data): prediction from a rank-deficient fit may be misleading

summary(prediction\_prices)

## fit lwr upr   
## Min. :0.6114 Min. :0.1646 Min. :1.058   
## 1st Qu.:1.1628 1st Qu.:0.7160 1st Qu.:1.610   
## Median :1.3894 Median :0.9427 Median :1.836   
## Mean :1.3941 Mean :0.9473 Mean :1.841   
## 3rd Qu.:1.6222 3rd Qu.:1.1751 3rd Qu.:2.069   
## Max. :2.3137 Max. :1.8668 Max. :2.761

#par(mfrow=c(2,2))  
#plot(Avocado\_Model)  
#abline(Avocado\_Model, col="orange", lwd=4)

#red\_Avocado<-[,c("AveragePrice","Total.Volume","PLU4046","PLU4225","PLU4770", "type")]  
#plot(red\_Avocado)  
reducedModel<- lm(AveragePrice~Total.Volume+PLU4046+PLU4225+PLU4770+type, data = train\_data)#reduced model  
#par(mfrow=c(2,2))  
#plot(reducedModel)  
fullModel<-Avocado\_Model  
fullModel<-lm(AveragePrice~., data=train\_data)#full model  
anova(reducedModel, fullModel)

## Analysis of Variance Table  
##   
## Model 1: AveragePrice ~ Total.Volume + PLU4046 + PLU4225 + PLU4770 + type  
## Model 2: AveragePrice ~ Total.Volume + PLU4046 + PLU4225 + PLU4770 + Total.Bags +   
## Small.Bags + Large.Bags + type + year + region + Month +   
## Day  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 12768 1104.59   
## 2 12669 652.85 99 451.75 88.551 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#summary(reducedModel)

#Prediction  
prediction\_prices<-predict(Avocado\_Model, interval="prediction", newdata=test\_data)

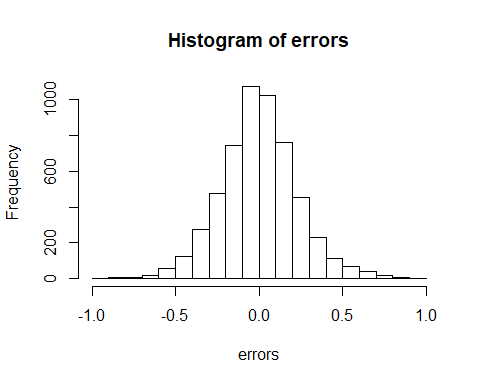
## Warning in predict.lm(Avocado\_Model, interval = "prediction", newdata =  
## test\_data): prediction from a rank-deficient fit may be misleading

summary(prediction\_prices)

## fit lwr upr   
## Min. :0.6114 Min. :0.1646 Min. :1.058   
## 1st Qu.:1.1628 1st Qu.:0.7160 1st Qu.:1.610   
## Median :1.3894 Median :0.9427 Median :1.836   
## Mean :1.3941 Mean :0.9473 Mean :1.841   
## 3rd Qu.:1.6222 3rd Qu.:1.1751 3rd Qu.:2.069   
## Max. :2.3137 Max. :1.8668 Max. :2.761

# Plotting errors on a histogram

#Errors  
errors<-prediction\_prices[,"fit"]-test\_data$AveragePrice  
#errors  
hist(errors)



#Model\_Performance  
#RMSE:  
rmse<-rmse(Avocado$AveragePrice, predict(Avocado\_Model, test\_data))

## Warning in predict.lm(Avocado\_Model, test\_data): prediction from a rank-  
## deficient fit may be misleading

## Warning in actual - predicted: longer object length is not a multiple of  
## shorter object length

rmse

## [1] 0.4797139