BANL6625- Mid Exam

Prashanth

library(ggplot2)  
setwd("D:/Data Mining")  
my\_data=read.csv("my\_data.csv")  
my\_data\_df=my\_data[,c(1,2,3,4,5,6)]  
summary(my\_data\_df)

## customer location soap\_type quantity   
## Min. : 1 Length:5000 Length:5000 Min. : 1.000   
## 1st Qu.:1251 Class :character Class :character 1st Qu.: 3.000   
## Median :2500 Mode :character Mode :character Median : 6.000   
## Mean :2500 Mean : 5.524   
## 3rd Qu.:3750 3rd Qu.: 8.000   
## Max. :5000 Max. :10.000   
## NA's :500   
## age payment\_type   
## Min. :18.00 Length:5000   
## 1st Qu.:38.00 Class :character   
## Median :58.00 Mode :character   
## Mean :58.56   
## 3rd Qu.:79.00   
## Max. :98.00   
## NA's :500

sum(is.na(my\_data\_df))

## [1] 1500

#removing NA's from the dataset  
  
clean\_data=na.omit(my\_data\_df)  
sum(is.na(clean\_data))

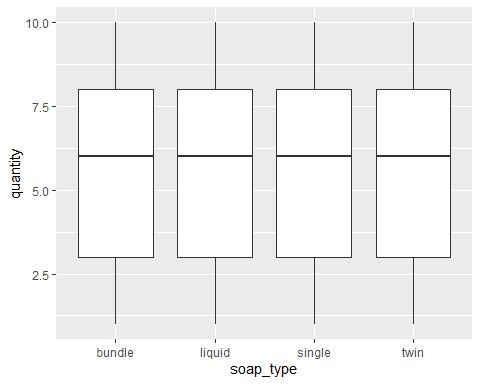
## [1] 0

#Binning the age into age\_groups  
clean\_data$age\_group=cut(clean\_data$age,breaks = c(0,10,30,50,100),labels = c("baby","young","Adult","Old"))  
summary(clean\_data)

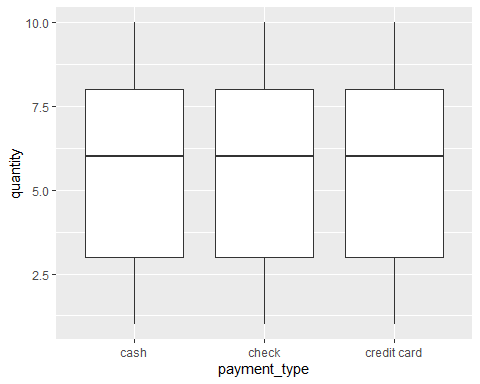
## customer location soap\_type quantity   
## Min. : 1 Length:4500 Length:4500 Min. : 1.000   
## 1st Qu.:1261 Class :character Class :character 1st Qu.: 3.000   
## Median :2504 Mode :character Mode :character Median : 6.000   
## Mean :2499 Mean : 5.524   
## 3rd Qu.:3735 3rd Qu.: 8.000   
## Max. :5000 Max. :10.000   
## age payment\_type age\_group   
## Min. :18.00 Length:4500 baby : 0   
## 1st Qu.:38.00 Class :character young: 673   
## Median :58.00 Mode :character Adult:1131   
## Mean :58.56 Old :2696   
## 3rd Qu.:79.00   
## Max. :98.00

## Box plot and Scatter plot

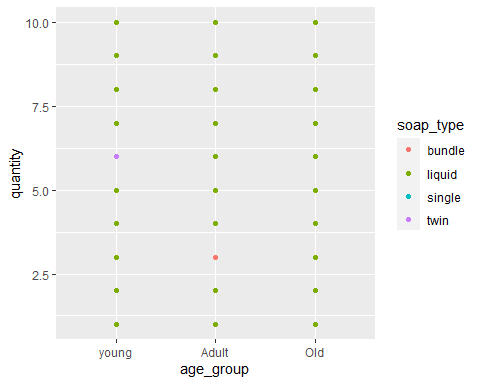
clean\_data$location=as.factor(clean\_data$location)  
clean\_data$soap\_type=as.factor(clean\_data$soap\_type)  
clean\_data$payment\_type=as.factor(clean\_data$payment\_type)  
  
boxpl1=ggplot(data=clean\_data,aes(x=soap\_type,y=quantity))+geom\_boxplot()  
boxpl2=ggplot(data=clean\_data,aes(x=payment\_type,y=quantity))+geom\_boxplot()  
Scatter1=ggplot(data=clean\_data,aes(x=age\_group,y=quantity,col=soap\_type))+geom\_point()  
Scatter2=ggplot(data=clean\_data,aes(x=age\_group,y=quantity,col=location))+geom\_point()  
boxpl1



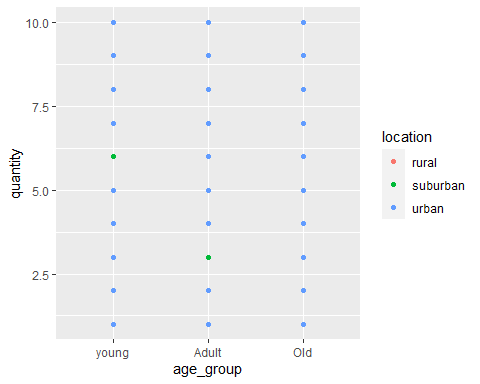
boxpl2



Scatter1



Scatter2



table(clean\_data$soap\_type)

##   
## bundle liquid single twin   
## 883 834 1668 1115

table(clean\_data$location)

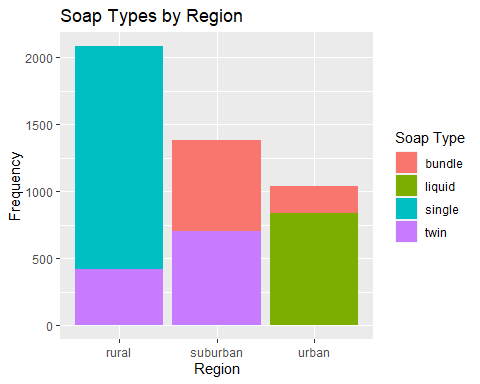
##   
## rural suburban urban   
## 2083 1381 1036

table(clean\_data$payment\_type)

##   
## cash check credit card   
## 1503 1475 1522

As we can see from the values in soap\_type single,in location rural, and in payment\_type credit card dominates the dataset.

ggplot(clean\_data, aes(x=location, fill=soap\_type)) +   
 geom\_bar(position="stack") +   
 labs(title="Soap Types by Region", x="Region", y="Frequency") +   
 scale\_fill\_discrete(name="Soap Type")



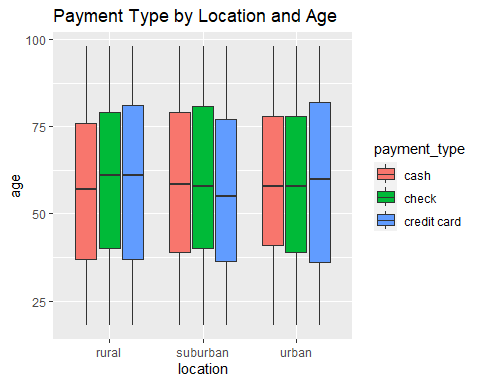
As we can see from the plot there are different type of soaps used in different locations.

In Rural Region - Single soap type is used

In Suburban Region - twin soap type is used

In Urban Region - Liquid soap type is used

ggplot(clean\_data, aes(x=location, y=age, fill=payment\_type)) +  
 geom\_boxplot() +  
 ggtitle("Payment Type by Location and Age")



As we can see from the boxplot, different age people from different location choose different type of payment method

library(class)

## Warning: package 'class' was built under R version 4.2.2

library(caret)

## Warning: package 'caret' was built under R version 4.2.3

## Loading required package: lattice

clean\_data$location=as.numeric(as.factor(clean\_data$location))  
clean\_data$payment\_type=as.numeric(as.factor(clean\_data$payment\_type))  
  
#Removed customer column for prediction of data  
clean\_data\_sub=subset(clean\_data,select=c("soap\_type","location","quantity","age","payment\_type"))  
  
#divided the data into 70:30 of train and test dataset  
train\_index <- sample(row.names(clean\_data\_sub), 0.7\*dim(clean\_data\_sub)[1])   
valid\_index <- setdiff(row.names(clean\_data\_sub), train\_index)   
train\_df <- clean\_data\_sub[train\_index, ]  
valid\_df <- clean\_data\_sub[valid\_index, ]  
train\_norm\_df=train\_df  
valid\_norm\_df=valid\_df  
#Standardizing the data  
norm\_values <- preProcess(train\_df[, c(2,3,4)], method=c("center", "scale"))  
train\_norm\_df[, c(2,3,4)] <- predict(norm\_values, train\_df[, c(2,3,4)])  
valid\_norm\_df[, c(2,3,4)] <- predict(norm\_values, valid\_df[, c(2,3,4)])  
  
new\_data=data.frame(location="urban",quantity=3,age=55,payment\_type="credit card")  
new\_data$location=as.numeric(as.factor(new\_data$location))  
new\_data$payment\_type=as.numeric(as.factor(new\_data$payment\_type))  
new\_data\_norm=predict(norm\_values,new\_data)  
  
predict\_soap\_type=knn(train=train\_norm\_df[,c(2,3,4,5)],test=new\_data\_norm,cl=train\_norm\_df$soap\_type,k=67)  
predict\_soap\_type

## [1] single  
## Levels: bundle liquid single twin

The predicted soap\_type for a person living in urban region, age of 55, quantity is 3 and payment type is credit card is “Single”