Assingment\_1

# Loading Libraries

library(tidyverse)

## -- Attaching packages -------------------------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 2.2.1 v purrr 0.2.4  
## v tibble 1.4.1 v dplyr 0.7.4  
## v tidyr 0.7.2 v stringr 1.2.0  
## v readr 1.1.1 v forcats 0.2.0

## -- Conflicts ----------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(readxl)  
  
platelet <- read\_excel("Dataset 2.xlsx")  
platelet<- as.data.frame(platelet)

# Change necessary columns to factor

col\_names <- names(platelet)  
for (i in col\_names) {  
 if(length(unique(platelet[,i])) <= 4)  
 {  
 platelet[,i] <- as.factor(platelet[,i])  
 }  
}  
names(platelet)[2]<-"Study\_Group"

# Binning

#Binning Continuos Columns  
for (col in 3:ncol(platelet)) {  
 column = platelet[,col]  
 if(is.numeric(column))  
 {  
 range=(max(column,na.rm = T)-min(column,na.rm = T))/5  
 range=round(range)  
 min\_val<-min(column,na.rm = T)-range  
 max\_val=max(column,na.rm = T)  
 bin=seq(from=min\_val,to=max\_val,by=range)  
 temp <- cut(column,bin)  
   
 platelet<-cbind(platelet,temp)  
 names(platelet)[match("temp",names(platelet))]<-paste0(names(platelet)[col],"\_bins")   
 }  
}

# Questions:-

## Question 1 Construct frequency distribution of all variables according to Group 1 and Group 2

freq\_dist<- function(clm\_to\_dist)  
{  
 dataset <- platelet  
 m <<- m+1  
 freq\_db <- data.frame(cat="a",group="b")  
 num\_col<-ncol(dataset)  
 if(is.numeric(clm\_to\_dist))  
 {  
   
 range=(max(clm\_to\_dist,na.rm = T)-min(clm\_to\_dist,na.rm = T))/5  
 range=round(range)  
 min\_val<-min(clm\_to\_dist,na.rm = T)-range  
 max\_val=max(clm\_to\_dist,na.rm = T)  
 bin=seq(from=min\_val,to=max\_val,by=range)  
 dataset[,num\_col+1] <- cut(clm\_to\_dist,bin)  
  
 t= dataset %>% group\_by(dataset[,(num\_col+1)]) %>% summarise(Group\_1=sum(Study\_Group=="Group 1"),Group\_2=sum(Study\_Group=="Group 2"))  
 names(t)[1] = names(dataset)[m-1]  
   
 }  
 else if(is.factor(clm\_to\_dist))  
 {  
 fac\_col<-names(dataset)[m-1]  
 t = dataset %>% group\_by(dataset[,c(fac\_col)]) %>% summarise(Group\_1 = sum(Study\_Group=="Group 1"),Group\_2=sum(Study\_Group=="Group 2"))  
 names(t)[1] <- fac\_col  
 }  
   
 return(t)  
}  
  
m=1  
a<-lapply(platelet,freq\_dist)  
a

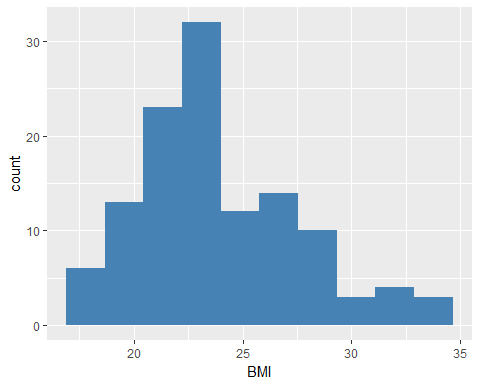
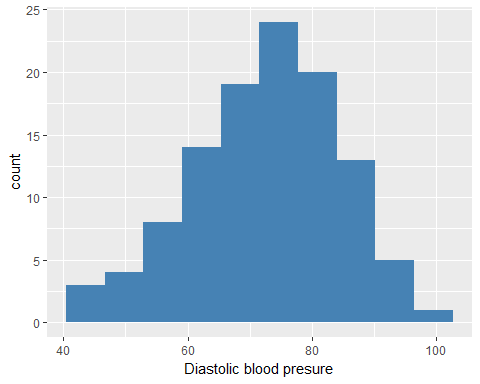
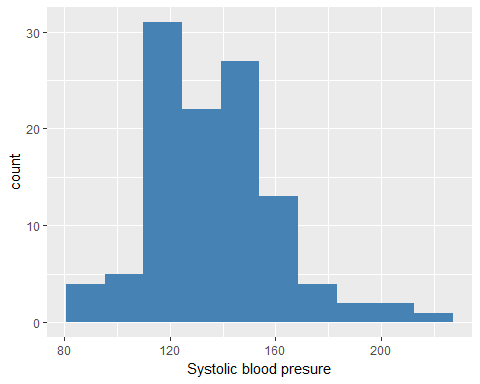
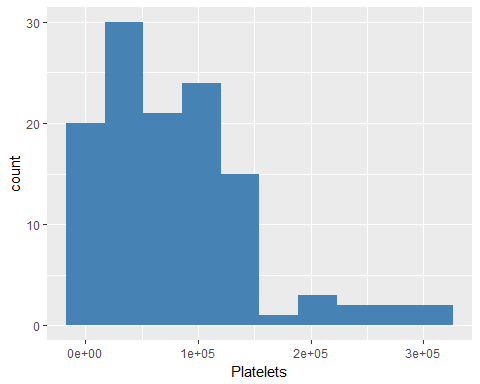
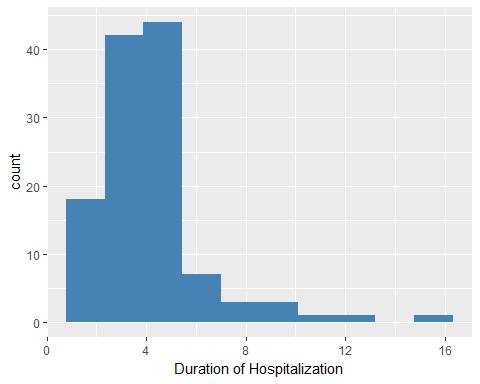
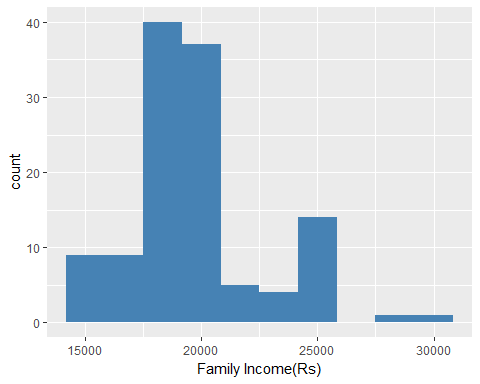
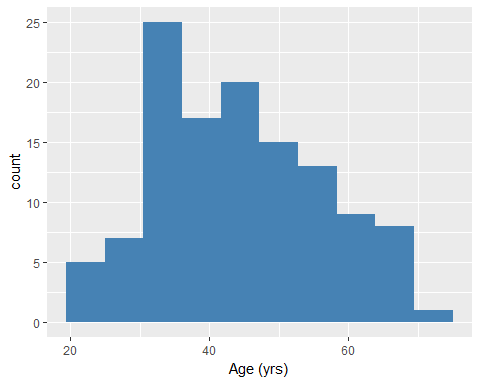
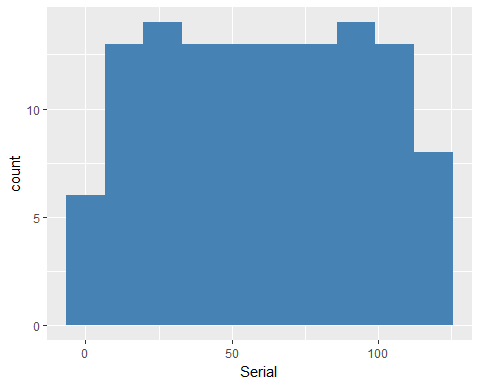
## $Serial  
## # A tibble: 6 x 3  
## Serial Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 (-23,1] 1 0  
## 2 (1,25] 24 0  
## 3 (25,49] 24 0  
## 4 (49,73] 11 13  
## 5 (73,97] 0 24  
## 6 <NA> 0 23  
##   
## $Study\_Group  
## # A tibble: 2 x 3  
## Study\_Group Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 Group 1 60 0  
## 2 Group 2 0 60  
##   
## $`Age (yrs)`  
## # A tibble: 6 x 3  
## `Age (yrs)` Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 (11,21] 0 1  
## 2 (21,31] 9 7  
## 3 (31,41] 19 18  
## 4 (41,51] 16 17  
## 5 (51,61] 9 9  
## 6 (61,71] 7 8  
##   
## $Sex  
## # A tibble: 2 x 3  
## Sex Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 Female 30 36  
## 2 Male 30 24  
##   
## $`Family Income(Rs)`  
## # A tibble: 6 x 3  
## `Family Income(Rs)` Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 (1.2e+04,1.5e+04] 6 3  
## 2 (1.5e+04,1.8e+04] 27 17  
## 3 (1.8e+04,2.1e+04] 18 26  
## 4 (2.1e+04,2.4e+04] 4 3  
## 5 (2.4e+04,2.7e+04] 4 10  
## 6 (2.7e+04,3e+04] 1 1  
##   
## $`Duration of Hospitalization`  
## # A tibble: 6 x 3  
## `Duration of Hospitalization` Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 (-2,1] 4 3  
## 2 (1,4] 50 34  
## 3 (4,7] 6 14  
## 4 (7,10] 0 6  
## 5 (10,13] 0 2  
## 6 <NA> 0 1  
##   
## $Platelets  
## # A tibble: 6 x 3  
## Platelets Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 (-5.58e+04,6e+03] 3 1  
## 2 (6e+03,6.78e+04] 30 23  
## 3 (6.78e+04,1.3e+05] 15 25  
## 4 (1.3e+05,1.91e+05] 6 8  
## 5 (1.91e+05,2.53e+05] 4 1  
## 6 (2.53e+05,3.15e+05] 2 2  
##   
## $`Systolic blood presure`  
## # A tibble: 7 x 3  
## `Systolic blood presure` Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 (64,90] 1 1  
## 2 (90,116] 10 6  
## 3 (116,142] 25 24  
## 4 (142,168] 14 21  
## 5 (168,194] 4 2  
## 6 (194,220] 1 1  
## 7 <NA> 5 5  
##   
## $`Diastolic blood presure`  
## # A tibble: 7 x 3  
## `Diastolic blood presure` Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 (33,44] 1 0  
## 2 (44,55] 2 5  
## 3 (55,66] 14 10  
## 4 (66,77] 21 19  
## 5 (77,88] 13 19  
## 6 (88,99] 4 2  
## 7 <NA> 5 5  
##   
## $BMI  
## # A tibble: 7 x 3  
## BMI Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 (14.6,17.6] 2 2  
## 2 (17.6,20.6] 12 5  
## 3 (20.6,23.6] 20 27  
## 4 (23.6,26.6] 15 13  
## 5 (26.6,29.6] 7 8  
## 6 (29.6,32.6] 3 3  
## 7 <NA> 1 2  
##   
## $`Culture 1`  
## # A tibble: 2 x 3  
## `Culture 1` Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 Negative 59 54  
## 2 Positive 1 6  
##   
## $`Culture 2`  
## # A tibble: 2 x 3  
## `Culture 2` Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 Negative 43 42  
## 2 Positive 17 18  
##   
## $`Age (yrs)\_bins`  
## # A tibble: 6 x 3  
## `Age (yrs)\_bins` Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 (11,21] 0 1  
## 2 (21,31] 9 7  
## 3 (31,41] 19 18  
## 4 (41,51] 16 17  
## 5 (51,61] 9 9  
## 6 (61,71] 7 8  
##   
## $`Family Income(Rs)\_bins`  
## # A tibble: 6 x 3  
## `Family Income(Rs)\_bins` Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 (1.2e+04,1.5e+04] 6 3  
## 2 (1.5e+04,1.8e+04] 27 17  
## 3 (1.8e+04,2.1e+04] 18 26  
## 4 (2.1e+04,2.4e+04] 4 3  
## 5 (2.4e+04,2.7e+04] 4 10  
## 6 (2.7e+04,3e+04] 1 1  
##   
## $`Duration of Hospitalization\_bins`  
## # A tibble: 6 x 3  
## `Duration of Hospitalization\_bins` Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 (-2,1] 4 3  
## 2 (1,4] 50 34  
## 3 (4,7] 6 14  
## 4 (7,10] 0 6  
## 5 (10,13] 0 2  
## 6 <NA> 0 1  
##   
## $Platelets\_bins  
## # A tibble: 6 x 3  
## Platelets\_bins Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 (-5.58e+04,6e+03] 3 1  
## 2 (6e+03,6.78e+04] 30 23  
## 3 (6.78e+04,1.3e+05] 15 25  
## 4 (1.3e+05,1.91e+05] 6 8  
## 5 (1.91e+05,2.53e+05] 4 1  
## 6 (2.53e+05,3.15e+05] 2 2  
##   
## $`Systolic blood presure\_bins`  
## # A tibble: 7 x 3  
## `Systolic blood presure\_bins` Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 (64,90] 1 1  
## 2 (90,116] 10 6  
## 3 (116,142] 25 24  
## 4 (142,168] 14 21  
## 5 (168,194] 4 2  
## 6 (194,220] 1 1  
## 7 <NA> 5 5  
##   
## $`Diastolic blood presure\_bins`  
## # A tibble: 7 x 3  
## `Diastolic blood presure\_bins` Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 (33,44] 1 0  
## 2 (44,55] 2 5  
## 3 (55,66] 14 10  
## 4 (66,77] 21 19  
## 5 (77,88] 13 19  
## 6 (88,99] 4 2  
## 7 <NA> 5 5  
##   
## $BMI\_bins  
## # A tibble: 7 x 3  
## BMI\_bins Group\_1 Group\_2  
## <fct> <int> <int>  
## 1 (14.6,17.6] 2 2  
## 2 (17.6,20.6] 12 5  
## 3 (20.6,23.6] 20 27  
## 4 (23.6,26.6] 15 13  
## 5 (26.6,29.6] 7 8  
## 6 (29.6,32.6] 3 3  
## 7 <NA> 1 2

View(platelet)

## Question 2 Represent the all the given variables below using appropriate graphical presentation

### Numerical Column

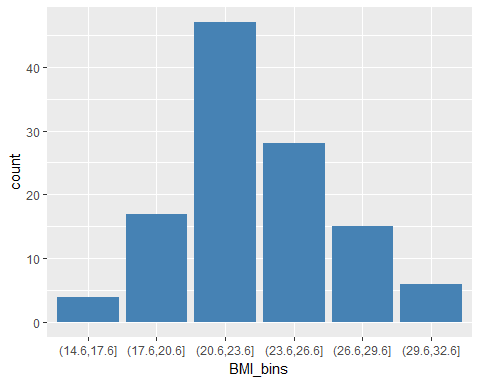
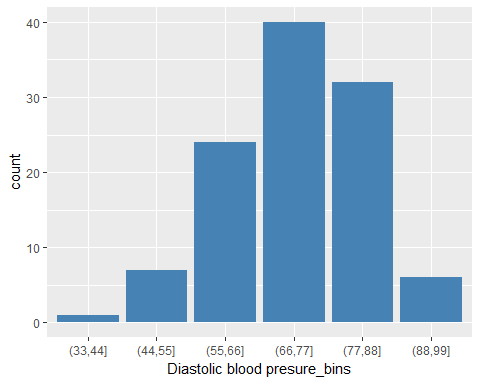
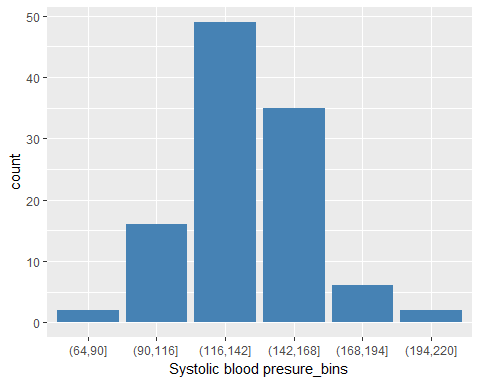
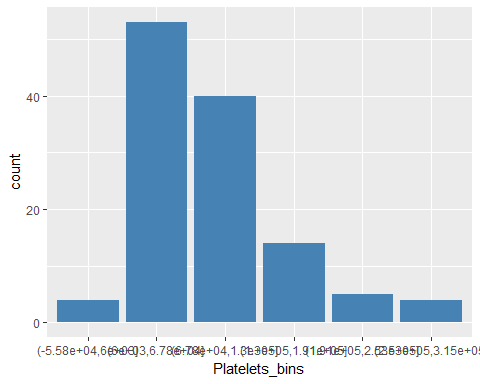
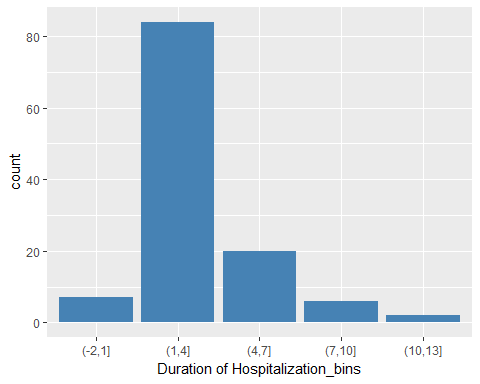
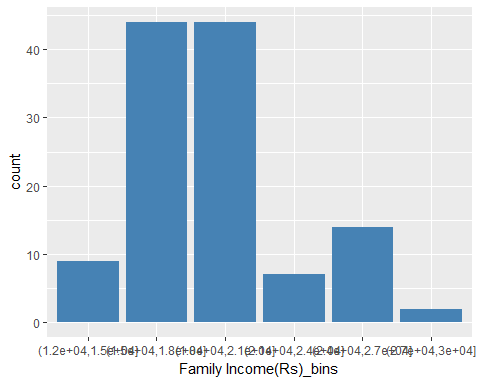
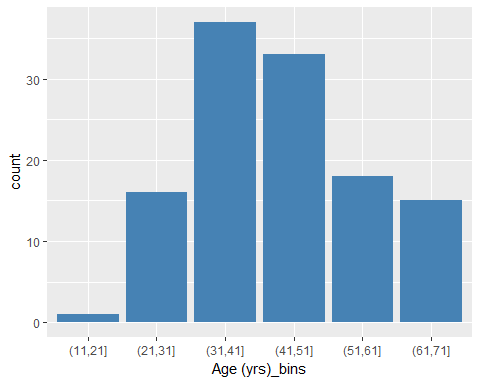
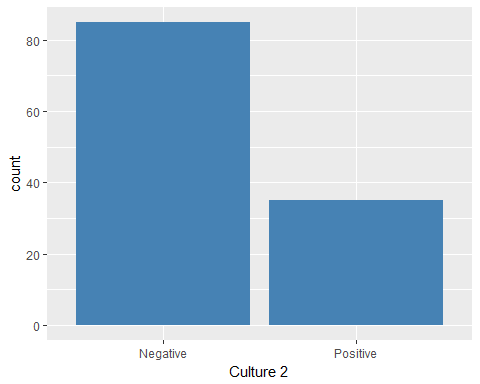
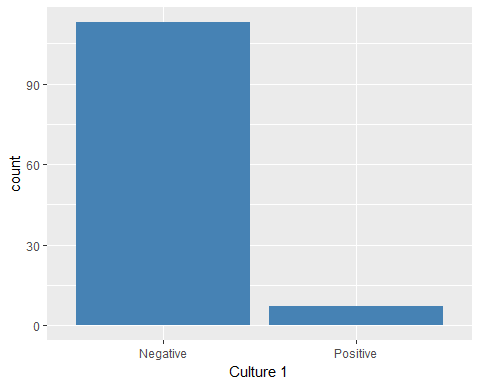
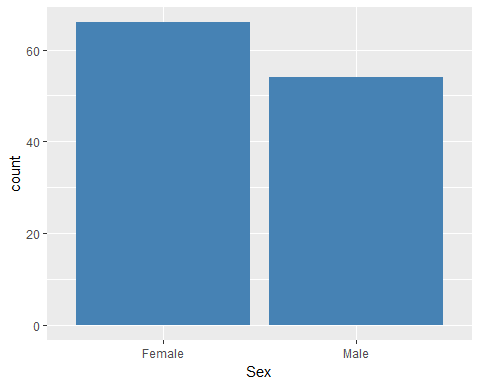
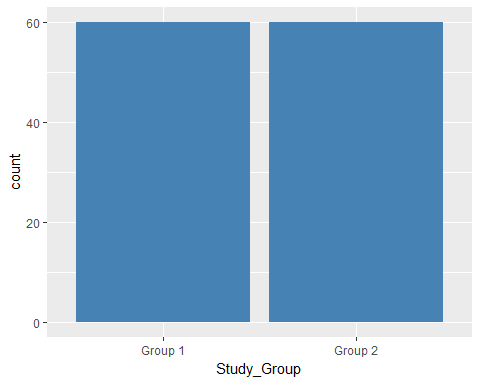
num = sapply(platelet, is.numeric)  
  
plot\_graph\_num = function(col){  
 new = data.frame(col = platelet[,col])  
 print(new %>% na.omit() %>%   
 ggplot(aes(x=col)) + geom\_histogram(fill = 'steelblue',bins = 10) + xlab(col))  
}  
  
sapply(names(platelet)[num], plot\_graph\_num)



## Serial Age (yrs) Family Income(Rs) Duration of Hospitalization  
## data List,1 List,1 List,1 List,1   
## layout ? ? ? ?   
## plot List,9 List,9 List,9 List,9   
## Platelets Systolic blood presure Diastolic blood presure BMI   
## data List,1 List,1 List,1 List,1  
## layout ? ? ? ?   
## plot List,9 List,9 List,9 List,9

### Categorical Column

num = sapply(platelet, is.numeric)  
  
gen\_plot = function(col){  
 new = data.frame(col = platelet[,col])  
 print(new %>% na.omit() %>%   
 ggplot(aes(x=col)) + geom\_bar(fill = 'steelblue') + xlab(col))  
}  
  
sapply(names(platelet)[!num], gen\_plot)



## Study\_Group Sex Culture 1 Culture 2 Age (yrs)\_bins  
## data List,1 List,1 List,1 List,1 List,1   
## layout ? ? ? ? ?   
## plot List,9 List,9 List,9 List,9 List,9   
## Family Income(Rs)\_bins Duration of Hospitalization\_bins  
## data List,1 List,1   
## layout ? ?   
## plot List,9 List,9   
## Platelets\_bins Systolic blood presure\_bins  
## data List,1 List,1   
## layout ? ?   
## plot List,9 List,9   
## Diastolic blood presure\_bins BMI\_bins  
## data List,1 List,1   
## layout ? ?   
## plot List,9 List,9

## Question 3 Construct the cross tables of Age versus Sex, Culture 1 and Culture 2

cross\_tab <- function(versus=c("Sex","Culture 1","Culture 2"))  
{  
 cross\_tab\_list=list()  
 for (col in versus) {  
 cross\_tab\_list[[col]]=table(platelet[,"Age (yrs)\_bins"],platelet[,col])  
 }  
 return(cross\_tab\_list)  
}  
  
  
cross\_tab()

## $Sex  
##   
## Female Male  
## (11,21] 0 1  
## (21,31] 9 7  
## (31,41] 25 12  
## (41,51] 15 18  
## (51,61] 10 8  
## (61,71] 7 8  
##   
## $`Culture 1`  
##   
## Negative Positive  
## (11,21] 0 1  
## (21,31] 15 1  
## (31,41] 35 2  
## (41,51] 33 0  
## (51,61] 17 1  
## (61,71] 13 2  
##   
## $`Culture 2`  
##   
## Negative Positive  
## (11,21] 0 1  
## (21,31] 10 6  
## (31,41] 29 8  
## (41,51] 24 9  
## (51,61] 15 3  
## (61,71] 7 8

## Question 4 Compute the mean and standard deviation of data obtained in the age frequency distribution

freq\_dist\_age <- platelet %>% group\_by(`Age (yrs)\_bins`) %>% summarise(Total\_occurence=n())  
mid\_vector=c()  
for (row in 1:6) {  
 bined\_age<-freq\_dist\_age$`Age (yrs)\_bins`[row]  
 trimed\_range<-gsub("\\(|\\]","",bined\_age)  
 num\_min\_max<-as.numeric(unlist(strsplit(trimed\_range,",")))  
 mid = (num\_min\_max[2]-num\_min\_max[1])/2  
 mid\_vector <- append(mid\_vector,(mid+num\_min\_max[1]))   
}  
freq\_dist\_age<-cbind(freq\_dist\_age,mid\_vector)  
  
freq\_dist\_age<-freq\_dist\_age %>% mutate(F.X=Total\_occurence\*mid\_vector)  
mean\_Age\_Freq\_Distribution<-sum(freq\_dist\_age$F.X)/sum(freq\_dist\_age$Total\_occurence)  
  
SD\_Age\_Freq\_Didtribution <- sqrt(sum(((freq\_dist\_age$mid\_vector-mean\_Age\_Freq\_Distribution)^2)\*freq\_dist\_age$Total\_occurence))/sqrt(sum(freq\_dist\_age$Total\_occurence))  
  
mean\_Age\_Freq\_Distribution

## [1] 44

SD\_Age\_Freq\_Didtribution

## [1] 12.35584

## Question 5 Construct the cross tables between Culture 1 and Culture 2

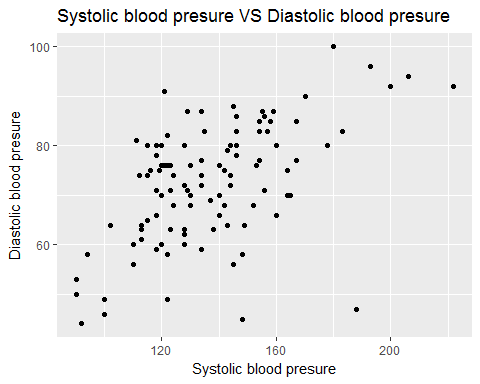
cross\_tab<- platelet %>% group\_by(`Culture 1`) %>% summarise(Culture2\_Negative=sum(`Culture 2`=="Negative"),Culture2\_positive=sum(`Culture 2`=="Positive"))

## Question 6 Present the summary statistics of all quantitative variables

n <- sapply(platelet, is.numeric)  
sum\_list<- lapply(platelet[,n], function(x){if(is.numeric(x)){summary(x)}else{return()}})

## Question 7 Draw a scattered diagram between Systolic blood pressure and Diastolic blood pressure

ggplot(platelet,aes(x=`Systolic blood presure`,y=`Diastolic blood presure`))+geom\_point() +ggtitle("Systolic blood presure VS Diastolic blood presure")

 ##Question 8 Present all the quantitative data using box-and Whisker plot

n=0  
plot1 <- function(col) {  
 n<<-n+1  
 if(is.numeric(col))  
 {  
 col\_name <- names(platelet)[n]  
 ggplot(platelet,aes(x=col\_name,y=col)) + geom\_boxplot() +ggtitle(paste0(col\_name," BOX AND WISKER PLOT"," "))  
 }  
}  
sapply(platelet[,n], plot1)

## named list()