project-1-final

Rohith Kanakagiri

28/10/2021

## data preprocessing

library(readxl)  
library(dplyr)  
library(corrplot)

## corrplot 0.90 loaded

library(ggplot2)  
library(tidyr)  
  
german\_data <- read\_excel("german data.xlsx")  
  
colnames(german\_data) = c("Checking\_Account","Duration","Credit\_History","Purpose","Credit\_Amount","Savings\_Account","Present\_Employement","Installment\_Rate","Personal\_Status","Guarantors","Residence\_Since","Property","Age","Other\_Installment\_Plan","Housing","Existing\_Credit","Job","Dependents","Telephone","Foreign\_Worker","Good\_bad")  
  
  
german\_data$Good\_bad[german\_data$Good\_bad == 2] <- "Bad"  
german\_data$Good\_bad[german\_data$Good\_bad == 1] <- "Good"  
  
german\_data$Sex[german\_data$Personal\_Status == "A91" | german\_data$Personal\_Status == "A93" | german\_data$Personal\_Status == "A94"] <- "M"

## Warning: Unknown or uninitialised column: `Sex`.

german\_data$Sex[german\_data$Personal\_Status == "A92" | german\_data$Personal\_Status == "A95"] <- "F"  
  
#german\_data$Income = round((german\_data$Credit\_Amount/german\_data$Duration)\*(100/german\_data$Installment\_Rate))  
  
german\_data$Age\_Category[german\_data$Age < 31] <- "Young"

## Warning: Unknown or uninitialised column: `Age\_Category`.

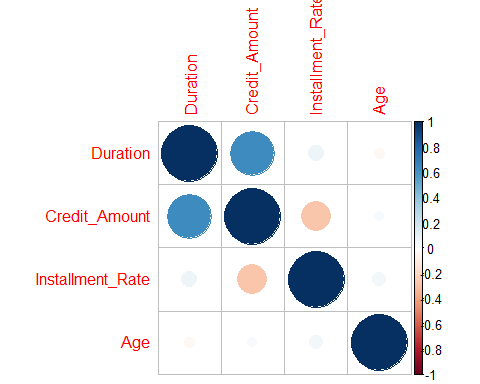
german\_data$Age\_Category[german\_data$Age >30 & german\_data$Age < 41] <- "Middle\_Age"  
german\_data$Age\_Category[german\_data$Age >40 & german\_data$Age < 61] <- "Adults"  
german\_data$Age\_Category[german\_data$Age >60] <- "Seniors"  
  
german\_data\_f <- subset(german\_data,Sex == "F")  
german\_data\_m <- subset(german\_data,Sex == "M")  
  
#converting numerical data to categorical data for box plot  
german\_data$Installment\_Rate\_Cat <- paste("B", german\_data$Installment\_Rate)  
  
  
german\_df <- as.data.frame(  
 cbind(  
 lapply(  
 lapply(german\_data, is.na), sum)  
 )  
 )  
colnames(german\_df) <- c('Number of Null Values in Column')  
rownames(subset(german\_df, german\_df$nullvalues != 0))

## character(0)

View(german\_df)

## Corr Plot to show co relation

library(corrplot)  
df <- select(german\_data, Duration, Credit\_Amount, Installment\_Rate, Age)  
corrplot(cor(df[,1:4]), method = "circle")



german\_data$Installment\_Rate <- paste("B",german\_data$Installment\_Rate\_Category)

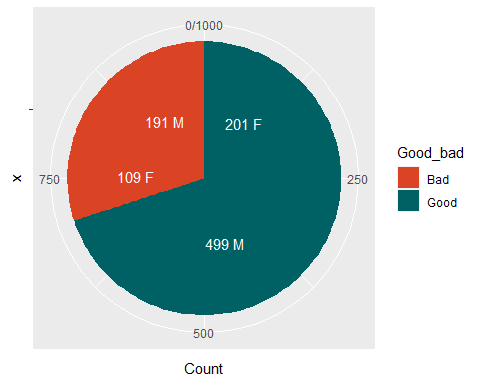
## Warning: Unknown or uninitialised column: `Installment\_Rate\_Category`.

## Pie Chart depicting number of good bad male female

df <- german\_data %>%   
 group\_by(Good\_bad, Sex) %>%   
 summarise(Count = n())

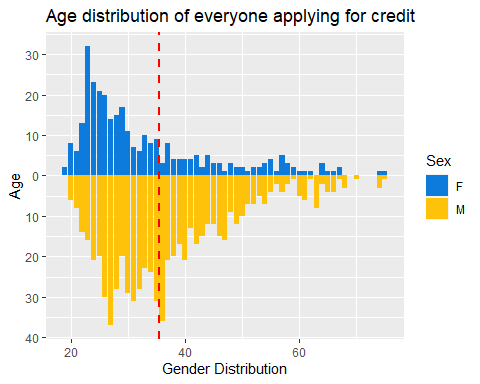
## `summarise()` has grouped output by 'Good\_bad'. You can override using the `.groups` argument.

ggplot(df, aes(x = "", y = Count ,fill = Good\_bad)) +   
 geom\_bar(width = 1, stat = "identity") +  
 geom\_text(aes(label = paste(Count,Sex)), position = position\_stack(vjust = 0.5), color = "white") +  
 scale\_fill\_manual(values = c("#DB4325","#006164")) +  
 coord\_polar(theta = "y")

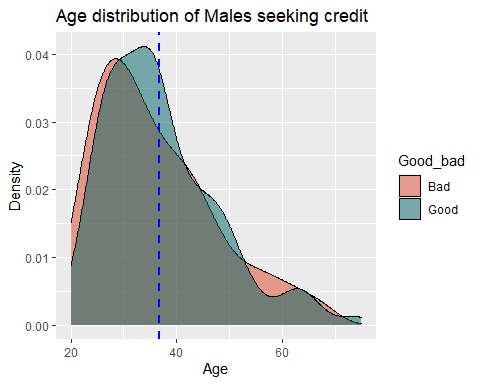


## DENSITY PLOT - 2 BOX PLOT 1

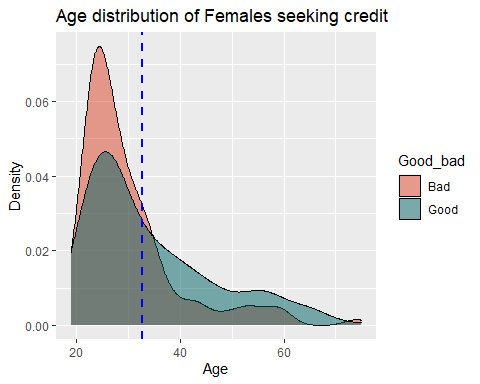
library(ggplot2)  
library(dplyr)  
  
pm <- ggplot(german\_data\_m, aes(Age, fill = Good\_bad)) +   
 geom\_density(alpha=.5) +  
 scale\_fill\_manual(values = c("#DB4325","#006164")) +  
 labs(x = "Age", y = "Density", title = "Age distribution of Males seeking credit") +   
 geom\_vline(aes(xintercept=mean(Age)), color="blue", linetype="dashed", size=1)  
  
pf <- ggplot(german\_data\_f, aes(Age, fill= Good\_bad)) +   
 geom\_density(alpha=.5) +   
 scale\_fill\_manual(values = c("#DB4325","#006164")) +  
 labs(x = "Age", y = "Density", title = "Age distribution of Females seeking credit") +   
 geom\_vline(aes(xintercept=mean(Age)), color="blue", linetype="dashed", size=1)  
  
p <- ggplot(german\_data,aes(x=Age ,fill = Sex)) +   
 geom\_bar(data=subset(german\_data, Sex == "F")) +   
 geom\_bar(data=subset(german\_data, Sex == "M"),aes(y=..count..\*(-1))) +   
 scale\_y\_continuous(breaks=seq(-40,40,10),labels=abs(seq(-40,40,10))) +   
 scale\_fill\_manual(values = c("#0C7BDC","#FFC20A")) +  
 labs( x= "Gender Distribution", y = "Age", title = "Age distribution of everyone applying for credit") +   
 geom\_vline(aes(xintercept=mean(Age)), color="Red", linetype="dashed", size=1)  
p



pm

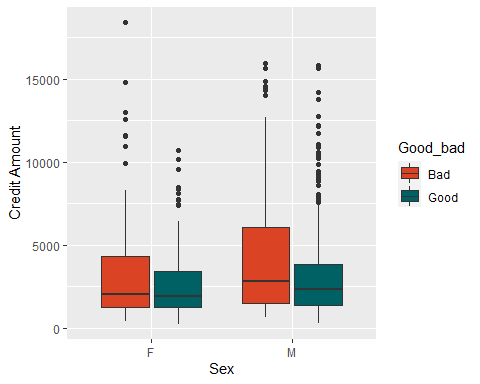


pf



## BOX PLOT for Credit Amount

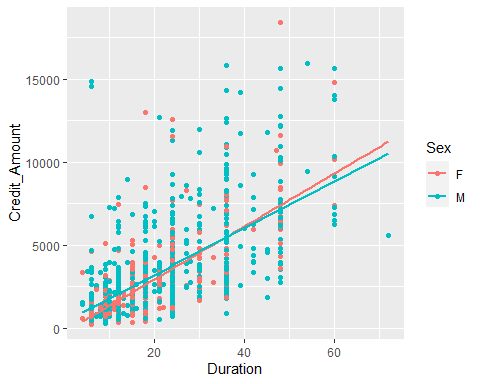
library(ggplot2)  
p2 <- ggplot(german\_data,aes(x = Sex, y = Credit\_Amount, fill = Good\_bad)) +   
 geom\_boxplot() +  
 labs(y = "Credit Amount") +  
 scale\_fill\_manual(values = c("#DB4325","#006164"))  
p2



## Scatter plot for credit amount and duration

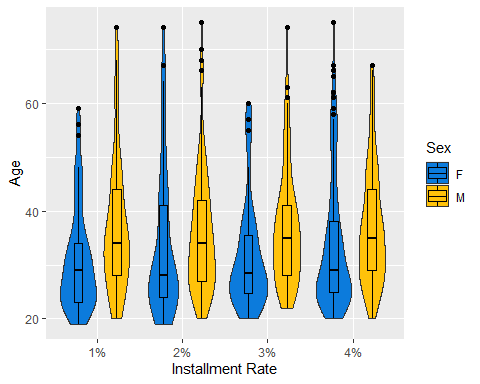
ggplot(german\_data, aes(x = Duration, y = Credit\_Amount, color = Sex)) +  
 geom\_point() +   
 #geom\_smooth(se = FALSE)  
 geom\_smooth(method=lm, se=FALSE, fullrange = TRUE)

## `geom\_smooth()` using formula 'y ~ x'



## Violin Box Plot

library(ggplot2)  
  
ggplot(german\_data, aes(x = Installment\_Rate\_Cat, y = Age, fill = Sex)) +   
 geom\_violin() +   
 geom\_boxplot(color="Black", width=0.2, position = position\_dodge(0.9)) +  
 scale\_fill\_manual(values = c("#0C7BDC","#FFC20A")) +  
 labs(x = "Installment Rate", y = "Age") +  
 scale\_x\_discrete(limits = c("B 1","B 2","B 3","B 4"),  
 labels = c("1%", "2%", "3%","4%"))



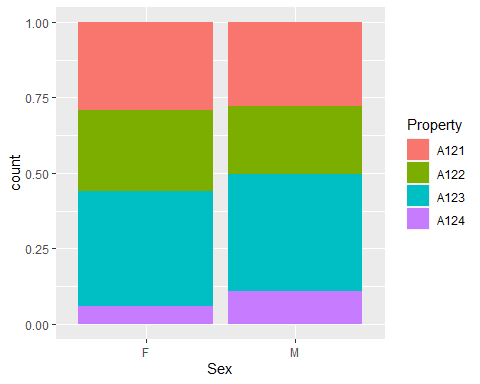
## Stacked bar plot gender vs Property

#property vs gender shows us that the percentage of young women having property or building and society savings agreements/life insurances is more than that of men but in the end they are still rejected more

library(ggplot2)  
library(dplyr)  
df <- german\_data%>%   
 subset(Age\_Category == "Young") %>%  
 group\_by(Property, Sex) %>%  
 dplyr::summarise(count = n())

## `summarise()` has grouped output by 'Property'. You can override using the `.groups` argument.

ggplot(df, aes(x = Sex, y = count, fill = Property)) +   
geom\_bar(stat = "identity", position = "fill")



## Stacked bar plot for purpose vs gender and Good\_bad

library(ggplot2)  
library(dplyr)  
library(gridExtra)

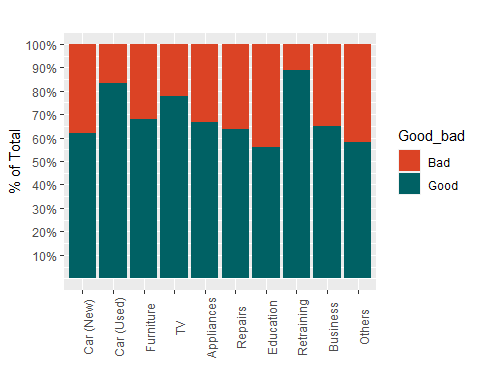
##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

df <- german\_data %>%  
 group\_by(Purpose, Good\_bad, Sex) %>%  
 dplyr::summarise(count = n())

## `summarise()` has grouped output by 'Purpose', 'Good\_bad'. You can override using the `.groups` argument.

p4\_1 <- ggplot(df, aes(x = Purpose, y = count, fill = Good\_bad)) +   
 geom\_bar(stat = "identity", position = "fill") +  
 scale\_fill\_manual(values = c("#DB4325","#006164")) +  
 labs( y = "% of Total", title = "Total") +  
 labs( x = "", title = "")+  
 scale\_x\_discrete(limits = c("A40","A41","A42","A43","A44","A45","A46","A48","A49","A410"),  
 labels = c("Car (New)", "Car (Used)", "Furniture","TV","Appliances","Repairs", "Education", "Retraining", "Business","Others")) +  
 scale\_y\_continuous(breaks = c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0), labels = c("10%","20%","30%","40%","50%","60%","70%","80%","90%","100%")) +  
 theme(axis.text.x = element\_text(angle=90))  
p4\_1



df<- german\_data %>%   
 subset(Sex == "F") %>%  
 group\_by(Purpose,Good\_bad) %>%  
 dplyr::summarise(count = n())

## `summarise()` has grouped output by 'Purpose'. You can override using the `.groups` argument.

p4\_2 <- ggplot(df, aes(x = Purpose, y = count, fill = Good\_bad)) +   
 geom\_bar(stat = "identity", position = "fill") +  
 scale\_fill\_manual(values = c("#DB4325","#006164")) +  
 #scale\_x\_discrete(limits = c("A40","A41","A42","A43","A44","A45","A46","A48","A49","A410"),  
 #labels = c("Car (New)", "Car (Used)", "Furniture","TV","Appliances","Repairs", "Education", "Retraining", "Business","Others")) +  
 scale\_y\_continuous(breaks = c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0),   
 labels = c("10%","20%","30%","40%","50%","60%","70%","80%","90%","100%")) +  
 theme(axis.text.x = element\_blank())+  
 labs( y = "% of Total females", tag = "F")  
  
  
df<- german\_data %>%   
 subset(Sex == "M") %>%  
 group\_by(Purpose, Good\_bad) %>%  
 dplyr::summarise(count = n())

## `summarise()` has grouped output by 'Purpose'. You can override using the `.groups` argument.

p4\_3 <- ggplot(df,aes(x = Purpose, y = count, fill = Good\_bad)) +   
geom\_bar(stat = "identity", position = "fill") +  
 scale\_fill\_manual(values = c("#DB4325","#006164")) +   
 scale\_x\_discrete(limits = c("A40","A41","A42","A43","A44","A45","A46","A48","A49","A410"),  
 labels = c("Car (New)", "Car (Used)", "Furniture","TV","Appliances","Repairs", "Education", "Retraining", "Business","Others")) +  
 scale\_y\_continuous(breaks = c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0),   
 labels = c("10%","20%","30%","40%","50%","60%","70%","80%","90%","100%")) +  
 theme(axis.text.x = element\_text(angle=90))+  
 labs( y = "% of Total males", tag = "M")  
  
  
grid.arrange(p4\_2,p4\_3)

