project-1-final

Rohith Kanakagiri

28/10/2021

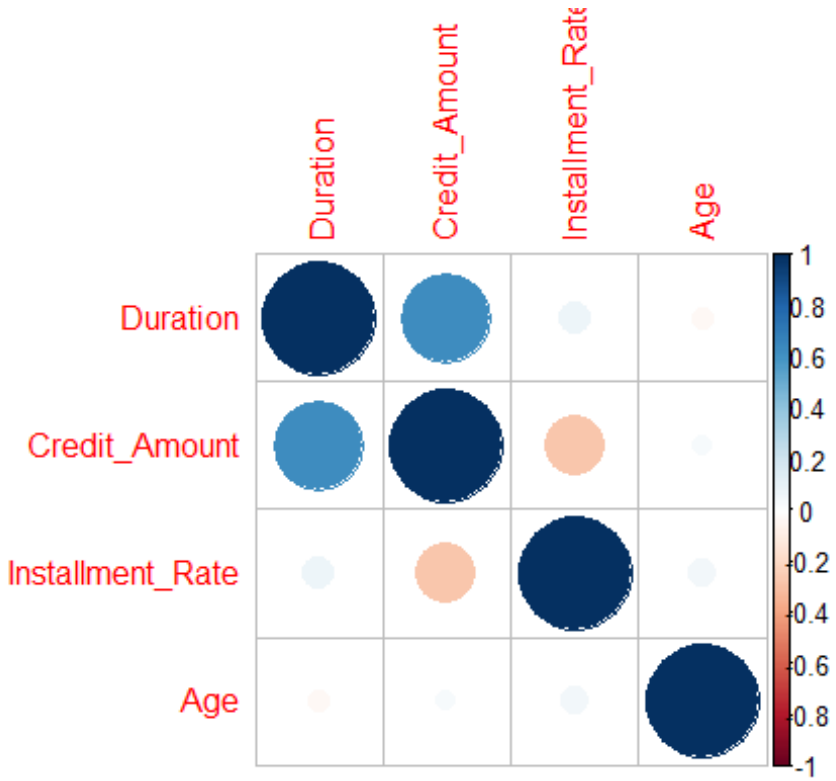
data preprocessing

```
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","Pur
pose","Credit_Amount","Savings_Account","Present_Employement","Installment_Ra
te","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$Personal_S
tatus == "A93" | german_data$Personal_Status == "A94"] <- "M"

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

german_data$Sex[german_data$Personal_Status == "A92" | german_data$Personal_S
tatus == "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$Age < 41] <- "Midd
le_Age"
german_data$Age_Category[german_data$Age >40 & german_data$Age < 61] <- "Adul
ts"
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

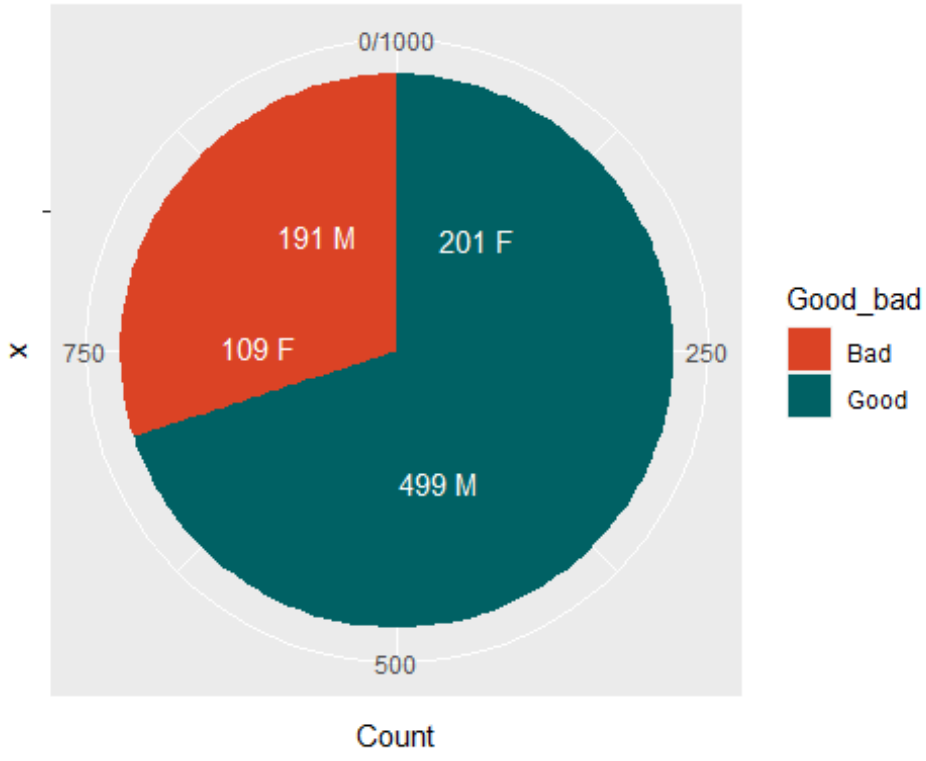
```
library(corrplot)
df <- select(german_data, Duration, Credit_Amount, Installment_Rate, Age)
corrplot(cor(df[,1:4]), method = "circle")
```



```
german_data$Installment_Rate <- paste("B",german_data$Installment_Rate_Catego
ry)

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

Pie Chart depicting number of good bad male female

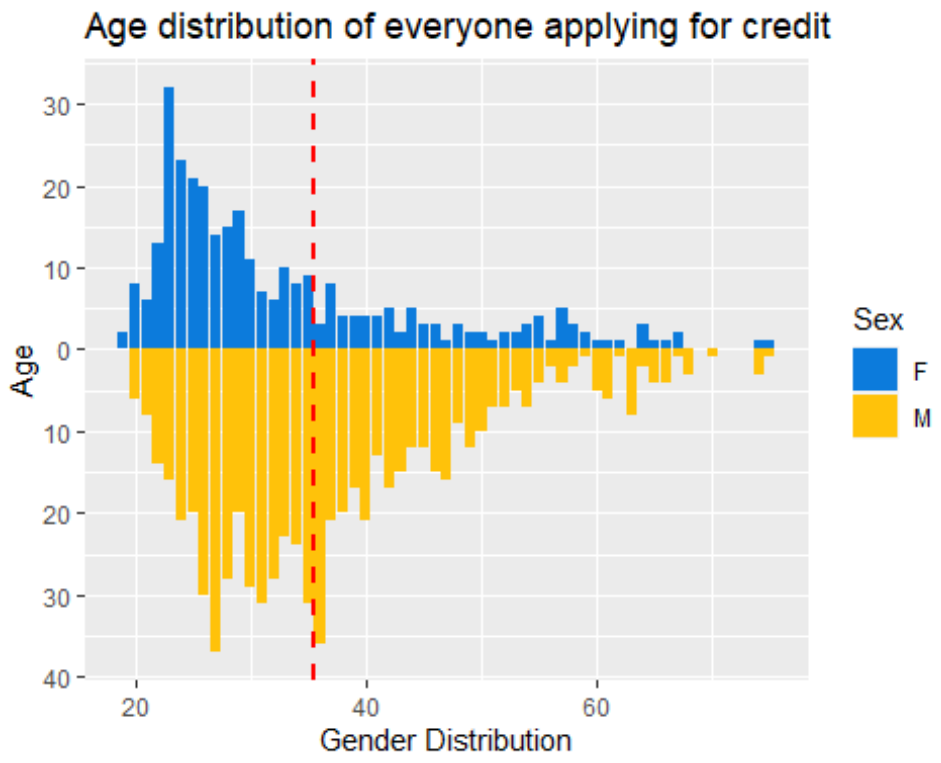
```
df <- german_data %>%
   group_by(Good_bad, Sex) %>%
   summarise(Count = n())

## `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")
```



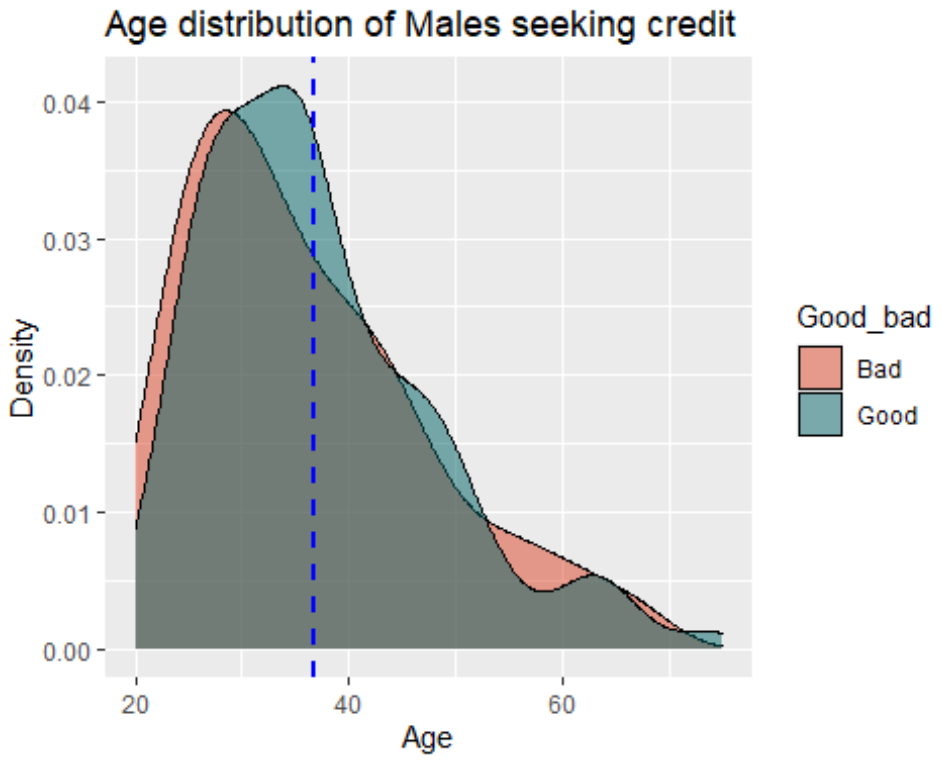
DENSITY PLOT - 2 BOX PLOT 1

```
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")) +
```

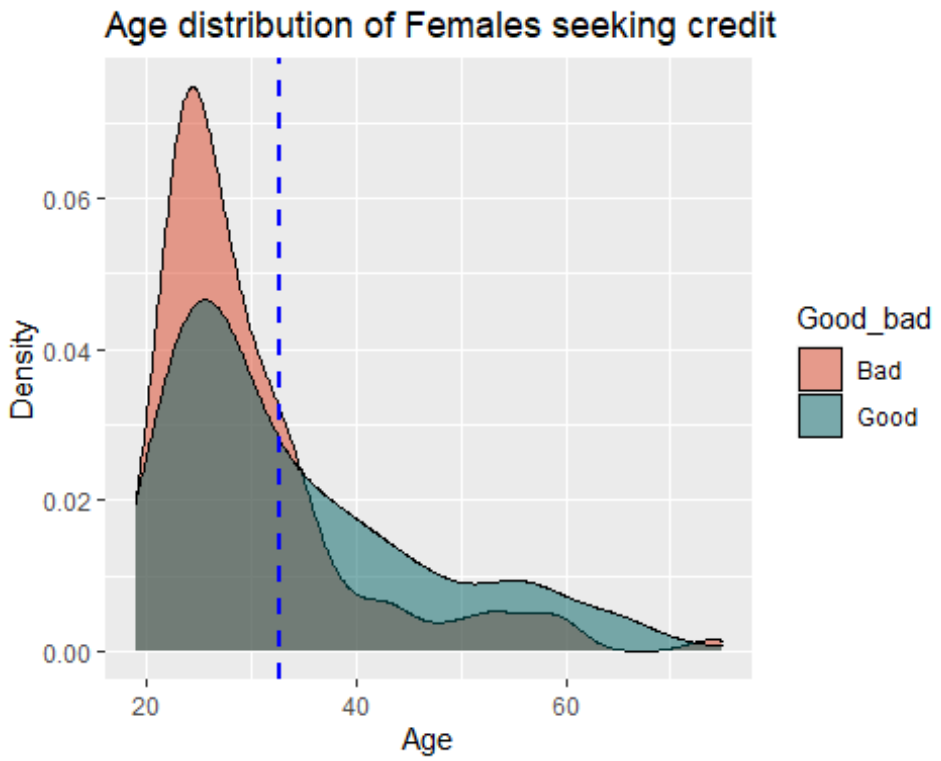
```
  labs(x = "Age", y = "Density", title = "Age distribution of Males seeking c
redit") +
  geom_vline(aes(xintercept=mean(Age)), color="blue", linetype="dashed", size
=1)

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 eve
ryone applying for credit") +
  geom_vline(aes(xintercept=mean(Age)), color="Red", linetype="dashed", size=
1)
p
```




```
pm
```

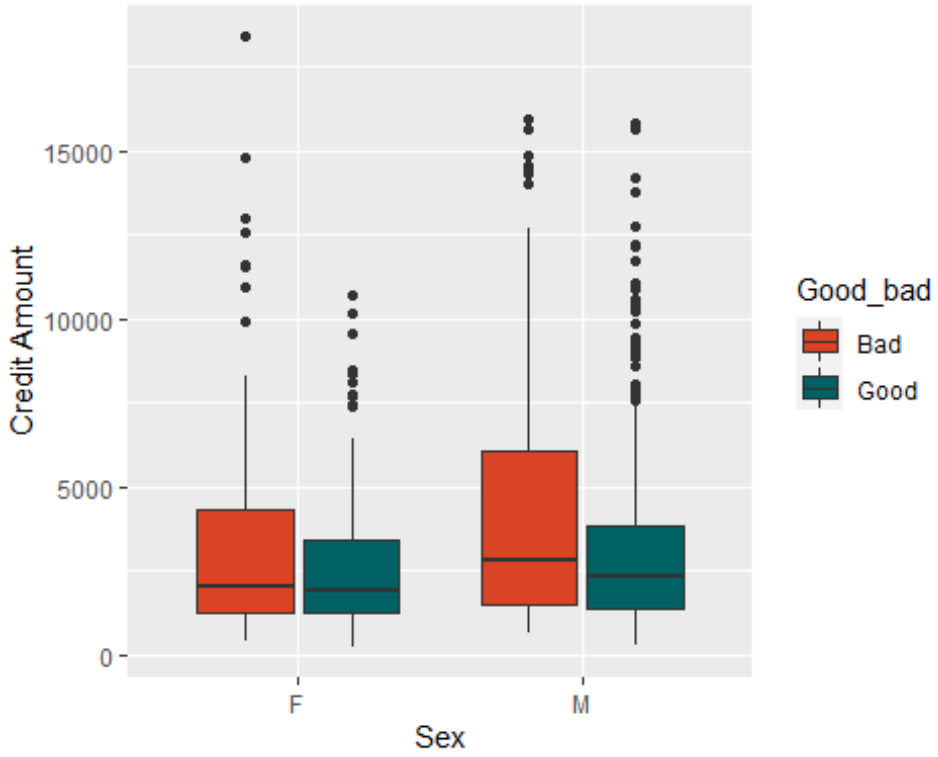

Age distribution of Males seeking credit
Density
0.04
0.03
0.02
0.01
0.00
20
40
60
Age
Good_bad
Bad
Good

```
pf
```


Age distribution of Females seeking credit
Density
0.06
0.04
0.02
0.00
20
40
60
Age
Good_bad
Bad
Good

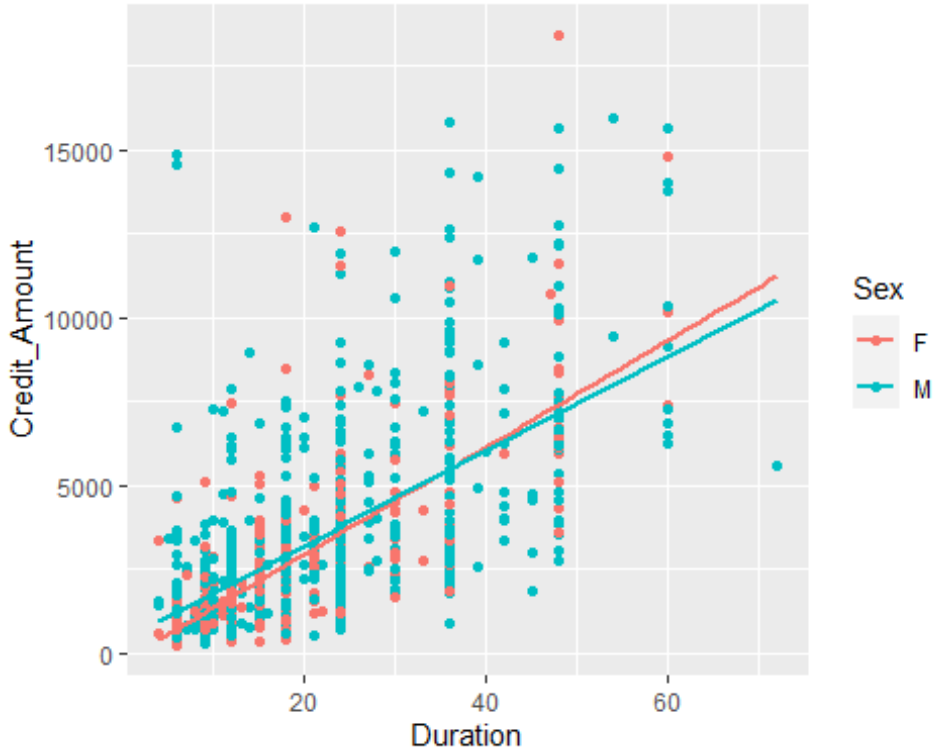
BOX PLOT for Credit Amount

```
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
```



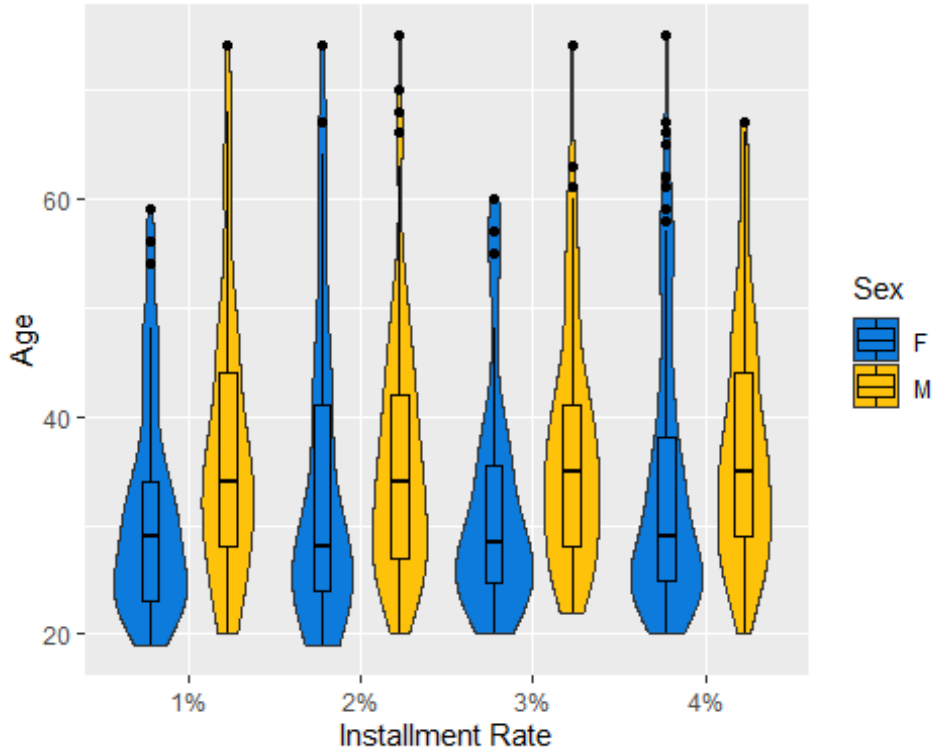
Scatter plot for credit amount and duration

```
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

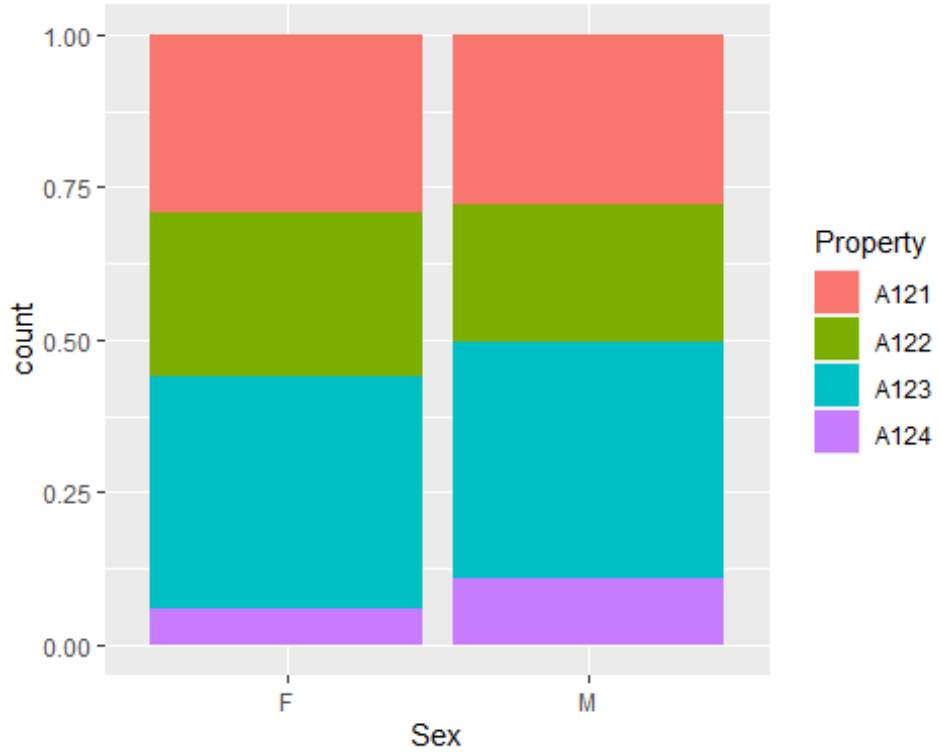
```
library(ggplot2)

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

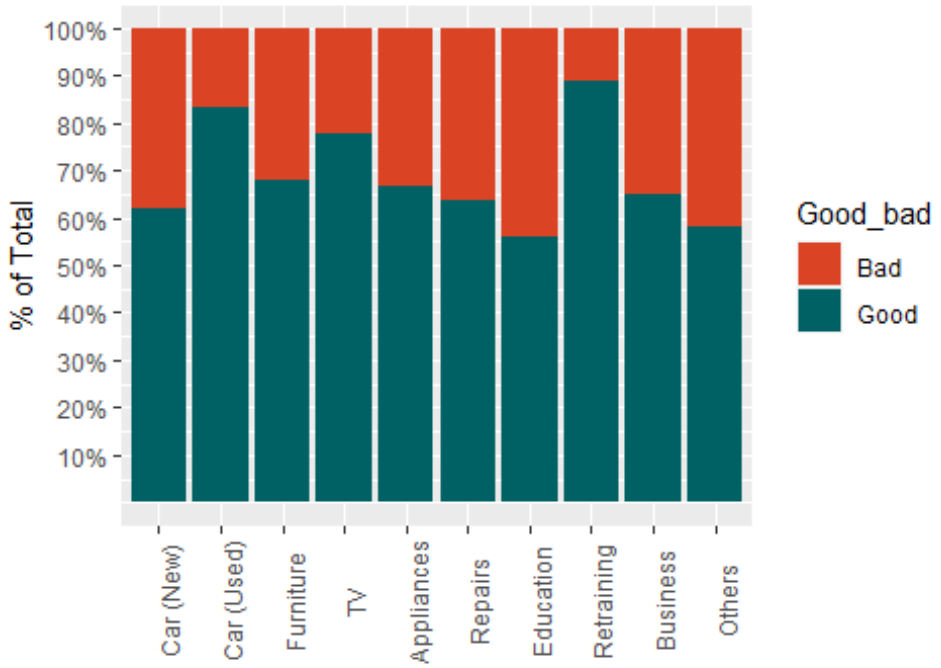
```
library(ggplot2)
library(dplyr)
df <- german_data%>%
  subset(Age_Category == "Young") %>%
  group_by(Property, Sex) %>%
  dplyr::summarise(count = n())

## `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 overrid
e 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","App
```

```
liances","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), l
abels = 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","A
ppliances","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%","8
0%","90%","100%")) +
  theme(axis.text.x = element_blank())+
  labs( y = "% of Total females", tag = "F")
```

```r
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","Ap
pliances","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%","8
0%","90%","100%")) +
  theme(axis.text.x = element_text(angle=90))+
  labs( y = "% of Total males", tag = "M")


grid.arrange(p4_2,p4_3)
```