Final Project
Reza Saffari

1- This is a Summary of all the variables: (code: summary(dm), #dm is the name of my data frame)

Age	Gender	OwnHome	Married	Location
Length:1000	Length:1000	Length:1000	Length:1000	Length:1000
Class :character	Class :character	Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character	Mode :character	Mode :character

Salary	Children	History	Catalogs	AmountSpent
Min. : 10100	Min. :0.000	Length:1000	Min. : 6.00	Min. : 38.0
1st Qu.: 29975	1st Qu.:0.000	Class :character	1st Qu.: 6.00	1st Qu.: 488.2
Median : 53700	Median :1.000	Mode :character	Median :12.00	Median : 962.0
Mean : 56104	Mean :0.934		Mean :14.68	Mean :1216.8
3rd Qu.: 77025	3rd Qu.:2.000		3rd Qu.:18.00	3rd Qu.:1688.5
Max. :168800	Max. :3.000		Max. :24.00	Max. :6217.0

2- How much percent of customers are women? %50.6

Code:

```
sum(dm$Gender == "Female")/nrow(dm)*100
```

At first we calculate the sum of “Female”s in column “Gender”, then we can find the ratio of the achieved value with respect to the number of rows in data frame; while there is no “Not Available” value in this column.

3- How much percent of married men have salaries more than 50K? %87.36

Code:

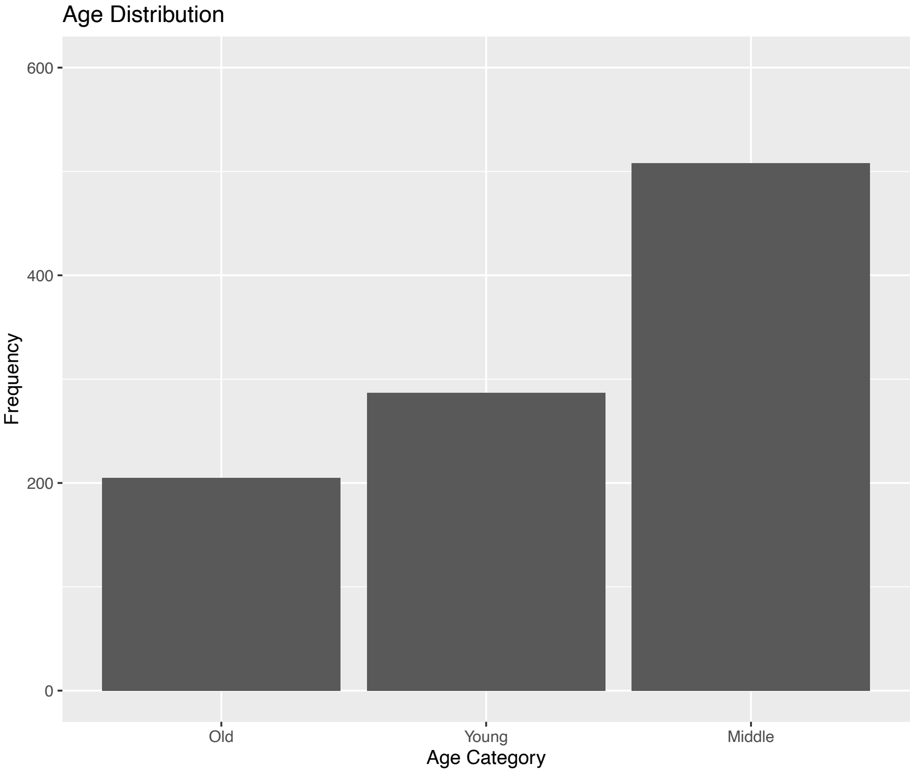
```
sum(dm$Gender == "Male" & dm$Married == "Married" & dm$Salary > 50000)/sum(dm$Gender == 
"Male" & dm$Married == "Married")*100
```

The problem here is that we must have two filters such as their gender and material status for our selected population while must one more filter for salary of them.

4- Abundance of customers with respect to their ages:

Code:

```
library("ggplot2")
agesum <- table(dm$Age)
agesum
class(agesum)

agesum <- as.data.frame(agesum)
agesum

colnames(agesum) <- c("age", "count")
agesum <- agesum[order(agesum$count),]
agesum

ggplot(agesum, aes(x= reorder(age,count),
y= count)) +
  geom_bar(stat = "identity") +
  labs(
    title = "Age Distribution",
    x = "Age Category",
    y = "Frequency") +
  ylim(0, 600)
```

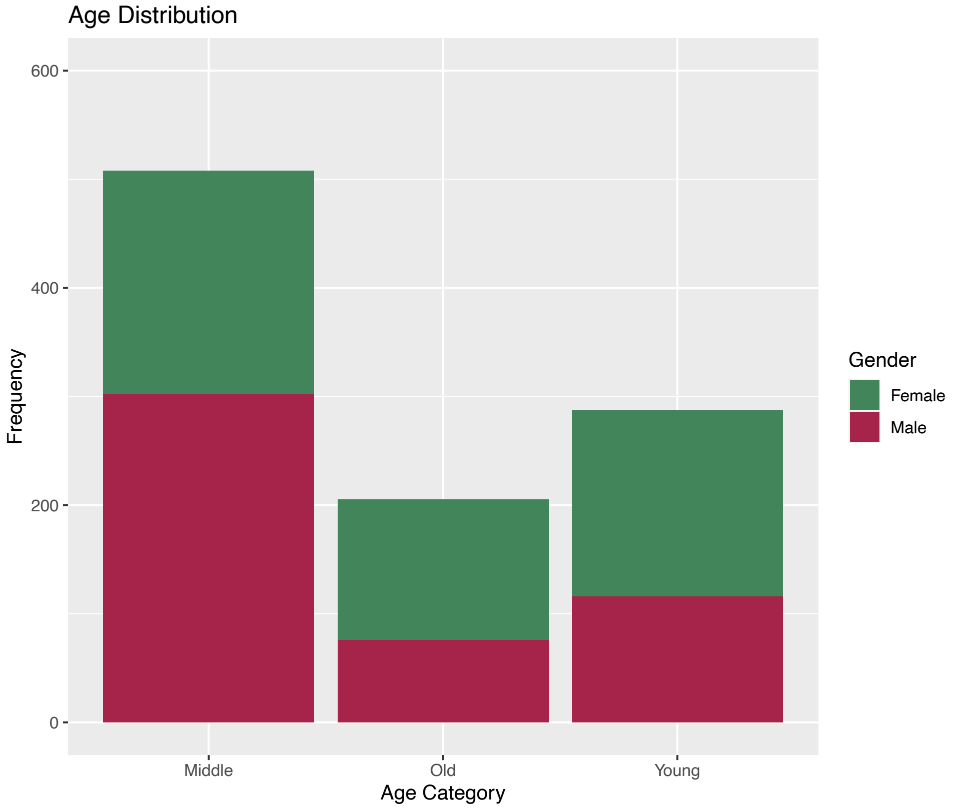


5- Other form of the previous question:

Code:

```
ggplot(dm, aes(Age, fill = Gender)) +
  geom_bar() +
  labs(
    title = "Age Distribution",
    x = "Age Category",
    y = "Frequency") +
  ylim(0, 600) +
  scale_fill_manual(values = c('#42855B',
‘#A62349'))
```

Here we see a cumulative age distribution of customers who are classified by gender.

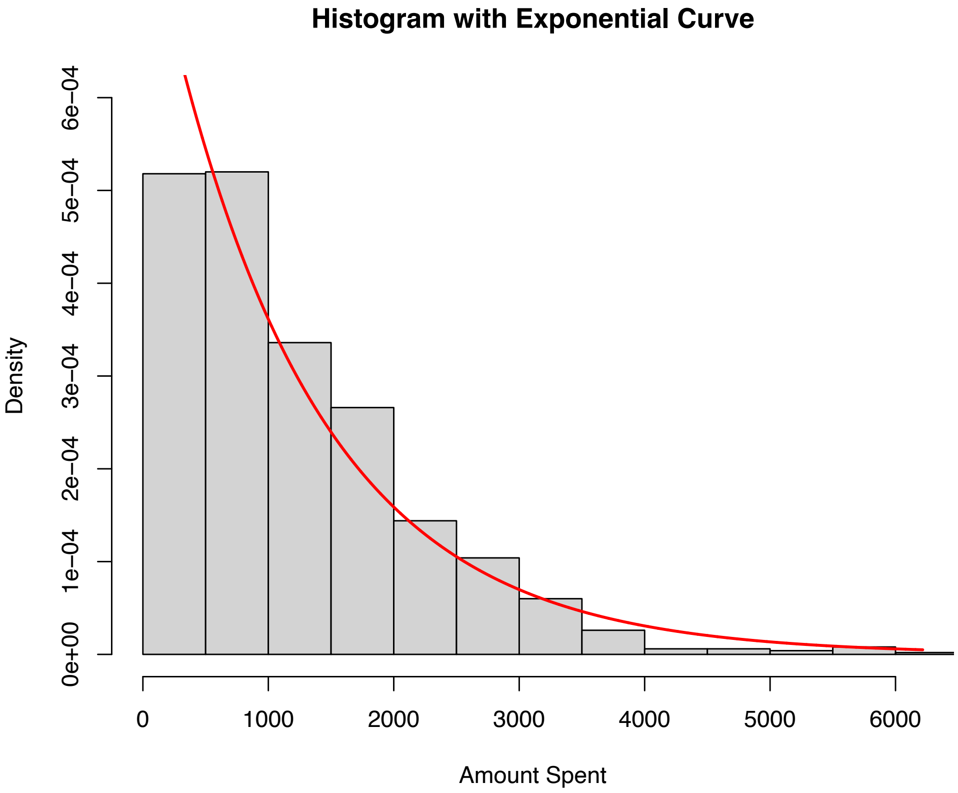


6- Spending Amount Density Distribution

Code:

```
x <- dm$AmountSpent
hist(x, freq = F, xlim = c(0, max(x)), ylim
= c(0, 6e-4), xlab = "Amount Spent", main =
"Histogram with Exponential Curve")

lambda <- 1/mean(x)
xfit <- seq(0, max(x), length = 10000)
yfit <- dexp(xfit, rate = lambda)
lines(xfit,yfit, col = "Red" ,lwd = 2)
```

Decreasing value of spending amount density what is doped by an exponential curve.

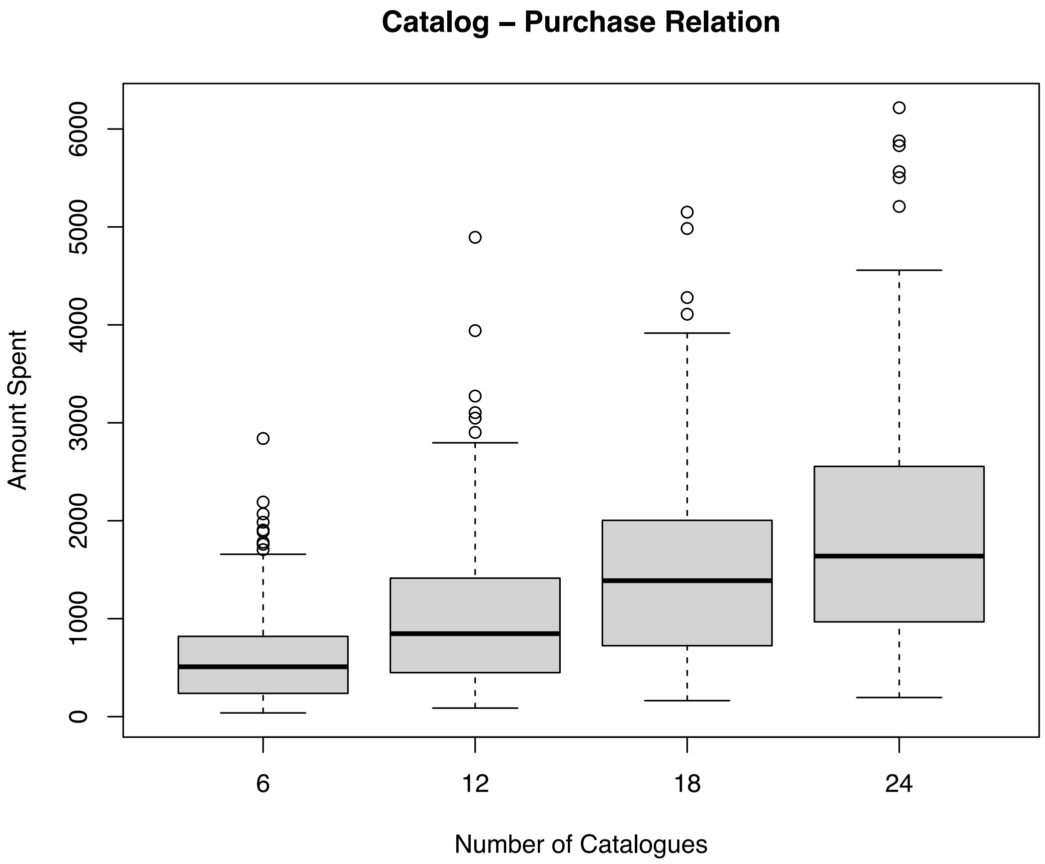


7- Spending Amount - Received Catalog Relation

Code:

```
boxplot(AmountSpent~Catalogs, data =
dm, main = "Catalog - Purchase
Relation",
        xlab = "Number of Catalogues"
, ylab = "Amount Spent”)
```

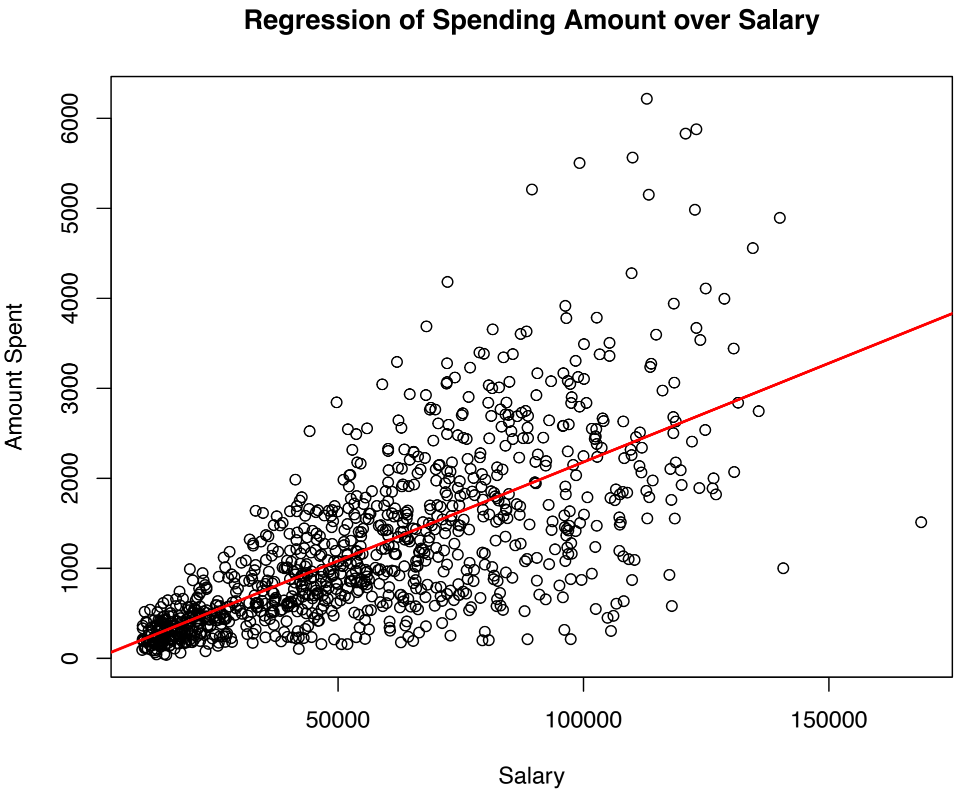
Here we see an amazing classification of customers who received different numbers of catalogs. More received catalogs spent more money.



8- Scatter Plot of Spending Amount with respect to Annual Salary and its Regression

Code:

```
plot(dm$Salary, dm$AmountSpent, xlab =
"Salary" , ylab = "Amount Spent")
abline(lm(dm$AmountSpent~dm$Salary), col
= "red", lw = 2)
title("Regression of Spending Amount
over Salary")
```

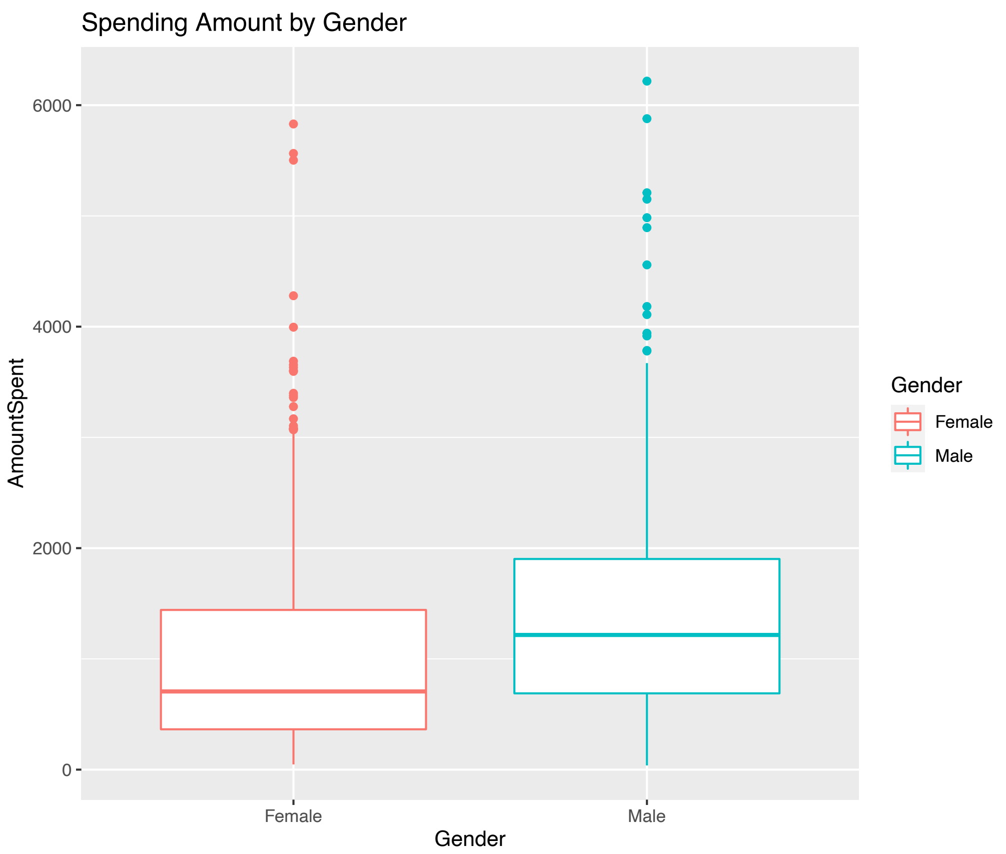


9. a) Spending Amount Distribution by Gender

Code:

```
d <- data.frame(dm$Gender, dm$Salary,
dm$AmountSpent)
head(d)
library(ggplot2)

ggplot(dm, aes(x= Gender, y= AmountSpent,
col = Gender)) +
  geom_boxplot() +
  labs(
    title = "Spending Amount by Gender”)
```

Here we see Males spend more than Females.
Even in their average amount:



```
SA_M <- mean(dm$AmountSpent[dm$Gender=="Male"])
> SA_M
[1] 1412.85
> SA_F <- mean(dm$AmountSpent[dm$Gender=="Female"])
> SA_F
[1] 1025.34
```

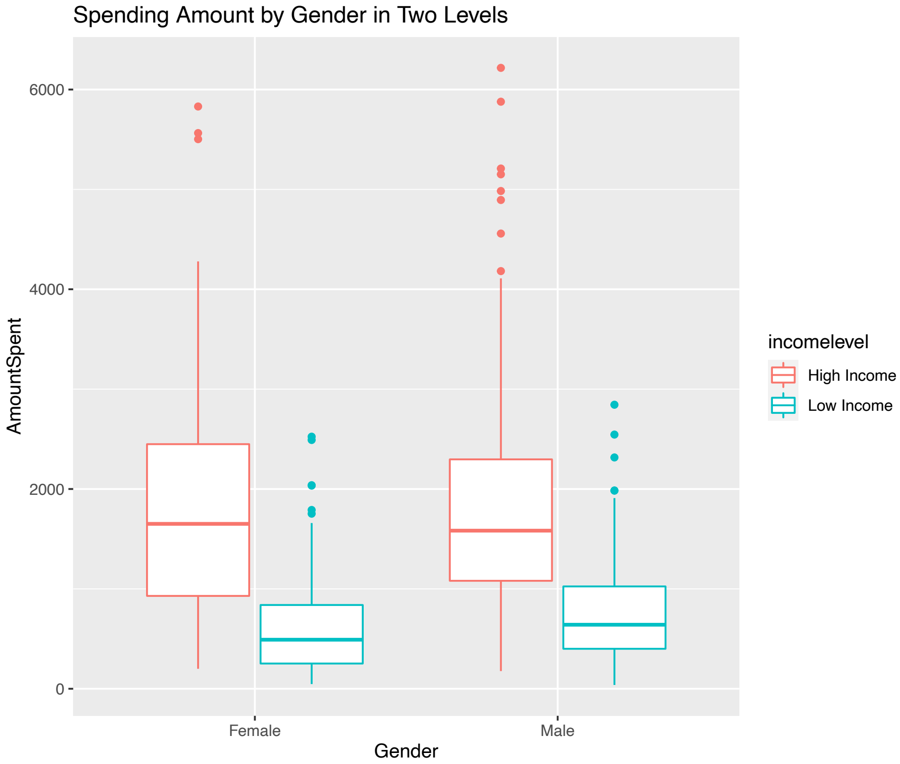
Is it really right result?

9. b) Spending Amount Distribution by Gender in Two Levels of Income

Code:

```
Sal_M <- mean(dm$Salary[dm$Gender=="Male"])
> Sal_M
[1] 64202.43
> Sal_F <- mean(dm$Salary[dm$Gender=="Female"])
> Sal_F
[1] 48197.43

dm$incomelevel <- ifelse(dm$Salary >
median(dm$Salary), "High Income","Low Income")

ggplot(dm, aes(x= Gender, y= AmountSpent, col =
incomelevel)) +
  geom_boxplot() +
  labs(
    title = "Spending Amount by Gender in Two
Levels”)
```

Here we hope to find more things about spending amount by dividing any group of genders into two categories of high income and low income. But there in no Information, unless it is clear that in each gender high income groups spend more money.



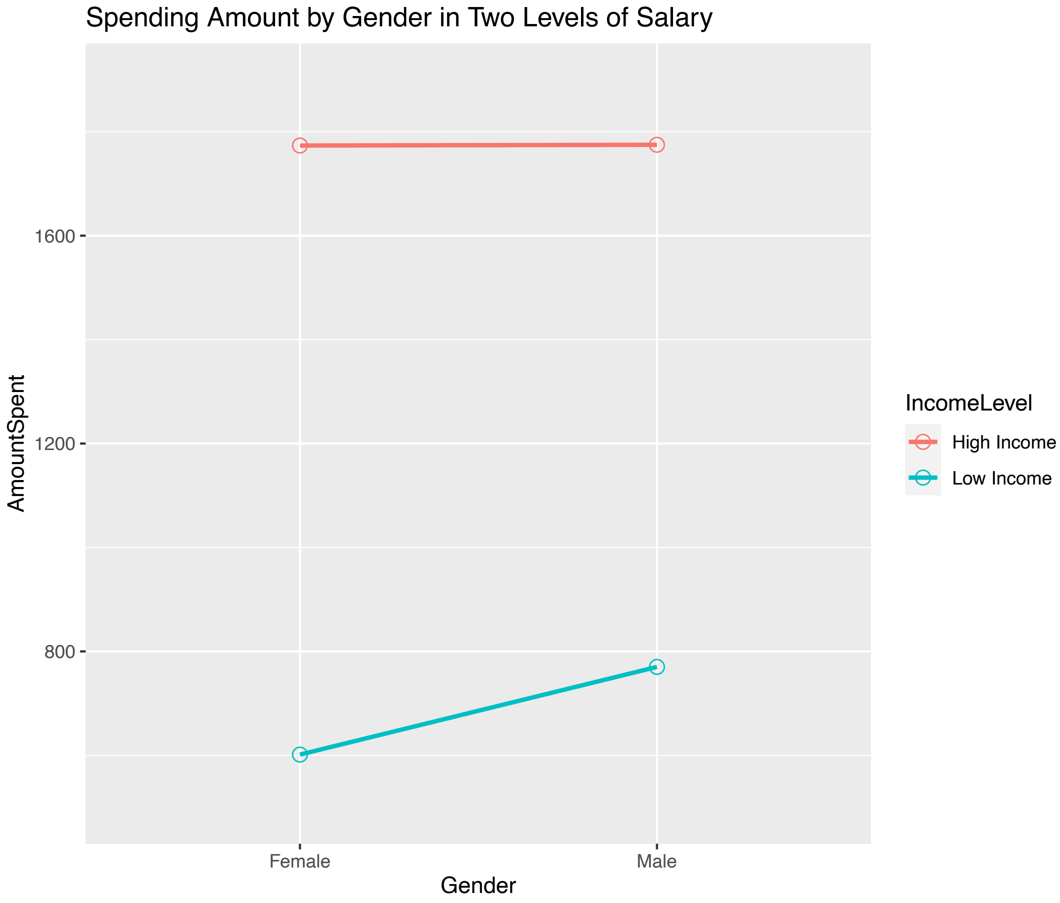
9. c) Reducing Income Effects on Spending Amount by Genders

Code:

```
median(dm$Salary)
[1] 53700

MH <- mean(new_dm$AmountSpent[new_dm$Gender == "Male" & new_dm$IncomeLevel == "High Income"])
> MH
[1] 1774.861
> ML <- mean(new_dm$AmountSpent[new_dm$Gender == "Male" & new_dm$IncomeLevel == "Low Income"])
> ML
[1] 770.1798
> FH <- mean(new_dm$AmountSpent[new_dm$Gender == "Female" & new_dm$IncomeLevel == "High Income"])
> FH
[1] 1773.475
> FL <- mean(new_dm$AmountSpent[new_dm$Gender == "Female" & new_dm$IncomeLevel == "Low Income"])
> FL
[1] 601.4737

Gender = c("Female","Female", "Male","Male")
IncomeLevel = c("High Income","Low Income","Low Income", "High Income")
AmountSpent = c(FH, FL, ML, MH)
df <- data.frame(Gender, IncomeLevel, AmountSpent)
head(df)

 Gender IncomeLevel AmountSpent
1 Female High Income   1773.4754
2 Female  Low Income    601.4737
3   Male  Low Income    770.1798
4   Male High Income   1774.86084
```

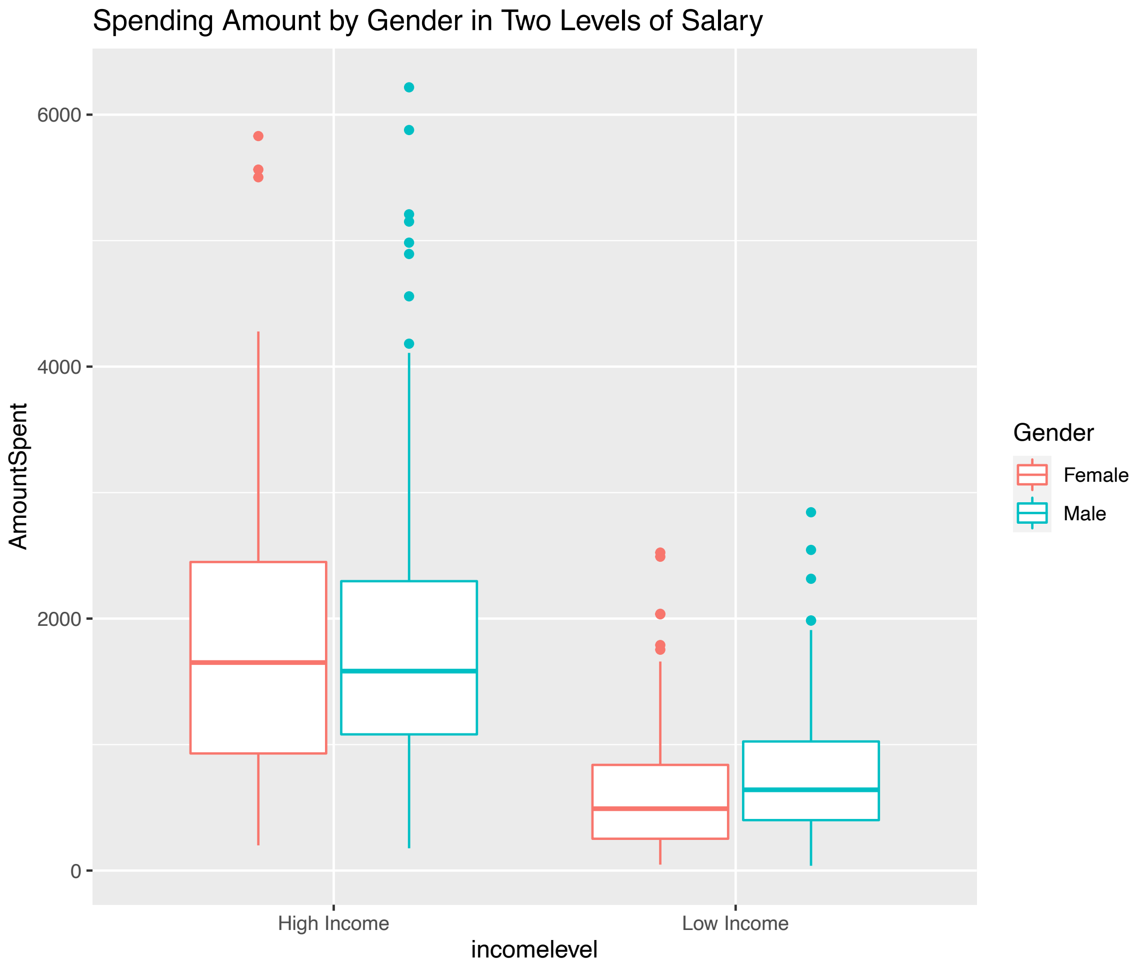
```
ggplot(df,aes(x = Gender , y = AmountSpent, group = IncomeLevel, col = IncomeLevel)) +
  geom_line(size = 1) +
  geom_point(shape = 1, size = 3) +
  ylim(500,1900) +
  labs(
    title = "Spending Amount by Gender in Two Levels of Salary”)
```

Here we see people with high income salary spend the same amount. But in the category of people with low income salary males spend more money than females.



```
ggplot(dm, aes(x= incomelevel, y= AmountSpent, col = Gender)) +
  geom_boxplot() +
  labs(
    title = "Spending Amount by Gender in Two Levels of Salary")
```

Here I have added another graph to see differences between various groups of people. In high income group females spent more money than males especially in average amount, while in low income group males spend more money than females.


Spending Amount by Gender in Two Levels of Salary
AmountSpent
High Income
Low Income
incomelevel
Gender
Female
Male