Code ▾

Project 3 : Supermarket sale

TUAN LINH DAO

8/16/2021

Project 3 : Supermarket sale

A. Branch A :

Load package:

Hide

```
library(dplyr)
library(ggplot2)
library(ggthemes)
library(ggpubr)
library(RColorBrewer)
```

I/ Explore data :

1) Import dataset :

Hide

```
data <- read.csv(file.choose(), header = T)
data
```

Invoice.ID	Bra…	City	Customer.type	Gen…	Product.line	Unit.pric
<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl
750-67-8428	A	Yangon	Member	Female	Health and beauty	74.6

Invoice.ID <chr>	Bra… <chr>	City <chr>	Customer.type <chr>	Gen… <chr>	Product.line <chr>	Unit.pric <dbl
226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.2
631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.3
123-19-1176	A	Yangon	Member	Male	Health and beauty	58.2
373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.3
699-14-3026	C	Naypyitaw	Normal	Male	Electronic accessories	85.3
355-53-5943	A	Yangon	Member	Female	Electronic accessories	68.8
315-22-5665	C	Naypyitaw	Normal	Female	Home and lifestyle	73.5
665-32-9167	A	Yangon	Member	Female	Health and beauty	36.2
692-92-5582	B	Mandalay	Member	Female	Food and beverages	54.8

1-10 of 1,000 rows | 1-8 of 17 columns Previous **1** 2 3 4 5 6 ... 100 Next

2) Summarize data :

a.1) Summary table :

Hide

```
summary(data)
```

```
 Invoice.ID           Branch              City          
 Length:1000        Length:1000        Length:1000       
 Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character  



 Customer.type         Gender          Product.line      
 Length:1000        Length:1000        Length:1000       
 Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character  



   Unit.price       Quantity        Tax.5.       
 Min.   :10.08   Min.   : 1.00   Min.   : 0.5085  
 1st Qu.:32.88   1st Qu.: 3.00   1st Qu.: 5.9249  
 Median :55.23   Median : 5.00   Median :12.0880  
 Mean   :55.67   Mean   : 5.51   Mean   :15.3794  
 3rd Qu.:77.94   3rd Qu.: 8.00   3rd Qu.:22.4453  
 Max.   :99.96   Max.   :10.00   Max.   :49.6500  
     Total             Date               Time          
 Min.   :  10.68   Length:1000        Length:1000       
 1st Qu.: 124.42   Class :character   Class :character  
 Median : 253.85   Mode  :character   Mode  :character  
 Mean   : 322.97                                        
 3rd Qu.: 471.35                                        
 Max.   :1042.65                                        
   Payment               cogs        gross.margin.percentage
 Length:1000        Min.   : 10.17   Min.   :4.762          
 Class :character   1st Qu.:118.50   1st Qu.:4.762          
 Mode  :character   Median :241.76   Median :4.762          
                    Mean   :307.59   Mean   :4.762          
                    3rd Qu.:448.90   3rd Qu.:4.762          
                    Max.   :993.00   Max.   :4.762          
  gross.income         Rating      
 Min.   : 0.5085   Min.   : 4.000  
 1st Qu.: 5.9249   1st Qu.: 5.500  
 Median :12.0880   Median : 7.000  
 Mean   :15.3794   Mean   : 6.973  
 3rd Qu.:22.4453   3rd Qu.: 8.500  
 Max.   :49.6500   Max.   :10.000  
```

a.2) Dimension of dataset :

Hide

```
dim(data)
```

```
[1] 1000   17
```

b) Unique value :

b.1) Using “unique” function :

Hide

```
# Branch :
a <- unique(data$Branch)
a
```

```
[1] "A" "C" "B"
```

Hide

```
# City :
b <- unique(data$City)
b
```

```
[1] "Yangon"    "Naypyitaw" "Mandalay"
```

Hide

```
# Customer type :
c <- unique(data$Customer.type)
c
```

```
[1] "Member" "Normal"
```

Hide

```
# Gender :
d <- unique(data$Gender)
d
```

```
[1] "Female" "Male"
```

Hide

```
# Product line :
e <- unique(data$Product.line)
e
```

```
[1] "Health and beauty"      "Electronic accessories"
[3] "Home and lifestyle"     "Sports and travel"
[5] "Food and beverages"     "Fashion accessories"
```

Hide

```
# Payment :
f <- unique(data$Payment)
f
```

```
[1] "Ewallet"     "Cash"        "Credit card"
```

b.2) Using "distinct" function with %>% operator :

Hide

```
# Branch :
a_1 <- data %>%
  distinct(data$Branch)
a_1
```

data$Branch <chr>
A
C
B

3 rows

Hide

```
# City :
b_1 <- data %>%
  distinct(data$Customer.type)
b_1
```

data$Customer.type <chr>

data$Customer.type
<chr>
Member
Normal

2 rows

Hide

```
# Customer type :
c_1 <- data %>%
  distinct(data$Customer.type)
c_1
```

data$Customer.type
<chr>
Member
Normal

2 rows

Hide

```
# Gender :
d_1 <- data %>%
  distinct(data$Gender)
d_1
```

data$Gender
<chr>
Female
Male

2 rows

Hide

```
# Product line :
e_1 <- data %>%
  distinct(data$Product.line)
e_1
```

data$Product.line
<chr>
Health and beauty

data$Product.line <chr>
Electronic accessories
Home and lifestyle
Sports and travel
Food and beverages
Fashion accessories

6 rows

Hide

```
# Payment :
f_1 <- data %>%
  distinct(data$Payment)
f_1
```

data$Payment <chr>
Ewallet
Cash
Credit card

3 rows

II/ Member table and non-member with product line:

3) Member table and non-member table of branch A:

Sort data for branch A uniquely :

Hide

```
m_table <- data[data$Branch == "A", ]
m_table
```

	Invoice.ID <chr>	Bra... <chr>	City <chr>	Customer.type <chr>	Gen... <chr>	Product.line <chr>	Unit.pric <dbl
1	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.6
3	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.3
4	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.2
5	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.3
7	355-53-5943	A	Yangon	Member	Female	Electronic accessories	68.8
9	665-32-9167	A	Yangon	Member	Female	Health and beauty	36.2
13	365-64-0515	A	Yangon	Normal	Female	Electronic accessories	46.9
14	252-56-2699	A	Yangon	Normal	Male	Food and beverages	43.1
15	829-34-3910	A	Yangon	Normal	Female	Health and beauty	71.3
17	656-95-9349	A	Yangon	Member	Female	Health and beauty	68.9

1-10 of 340 rows | 1-9 of 17 columns Previous **1** 2 3 4 5 6 ... 34 Next

a) Count the quantity for each product line based on customer type, gender :

Hide

```
m <- m_table %>%
  count(m_table$Customer.type,m_table$Gender,m_table$Product.line)
m
```

m_table$Customer.type <chr>	m_table$Gender <chr>	m_table$Product.line <chr>	n <int>
Member	Female	Electronic accessories	12
Member	Female	Fashion accessories	12
Member	Female	Food and beverages	10
Member	Female	Health and beauty	12
Member	Female	Home and lifestyle	16

m_table$Customer.type <chr>	m_table$Gender <chr>	m_table$Product.line <chr>	n <int>
Member	Female	Sports and travel	18
Member	Male	Electronic accessories	19
Member	Male	Fashion accessories	10
Member	Male	Food and beverages	19
Member	Male	Health and beauty	10

1-10 of 24 rows Previous **1** 2 3 Next

b) Table for member of branch A:

*) Member table :

Hide

```
m1 <- m[m$`m_table$Customer.type` == "Member", ]
m1
```

	m_table$Customer.type <chr>	m_table$Gender <chr>	m_table$Product.line <chr>	n <int>
1	Member	Female	Electronic accessories	12
2	Member	Female	Fashion accessories	12
3	Member	Female	Food and beverages	10
4	Member	Female	Health and beauty	12
5	Member	Female	Home and lifestyle	16
6	Member	Female	Sports and travel	18
7	Member	Male	Electronic accessories	19
8	Member	Male	Fashion accessories	10
9	Member	Male	Food and beverages	19
10	Member	Male	Health and beauty	10

1-10 of 12 rows Previous **1** 2 Next

Hide

```
# Rename all columns of member table :
colnames(m1) <- c("Customer type","Gender","Product_line","Quantity")
colnames(m1)
```

```
[1] "Customer type" "Gender"        "Product_line"
[4] "Quantity"
```

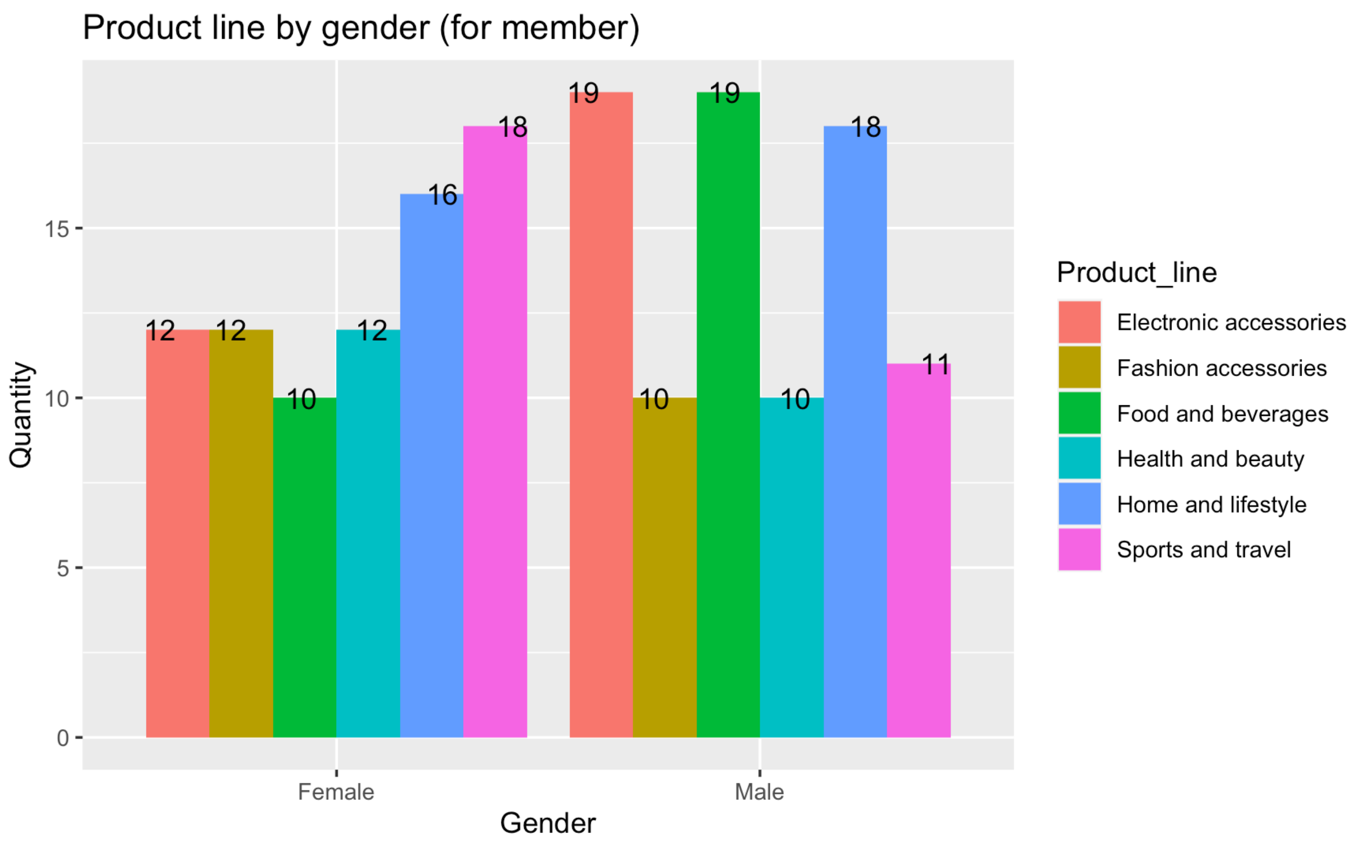
Hide

```
View(m1)
```

*) Barplot using ggplot package:

Hide

```
m <- ggplot(m1,aes(x=Gender,y=Quantity,fill=Product_line)) +
  geom_bar(stat='identity',position = 'dodge') +
  labs(title = "Product line by gender (for member)",
       y="Quantity",
       x="Gender") +
  geom_text(m1,mapping = aes(label=Quantity),stat = 'identity',position_dodge(width =
1)) #+
  #theme(legend.title = element_blank()) +
  #theme_fivethirtyeight()
m
```



c) Table for non-member of branch A:

*) Non-member table :

Hide

```
n_table <- m[m$`m_table$Customer.type`=="Normal", ]
n_table
```

	m_table$Customer.type <chr>	m_table$Gender <chr>	m_table$Product.line <chr>	n <int>
13	Normal	Female	Electronic accessories	16
14	Normal	Female	Fashion accessories	16
15	Normal	Female	Food and beverages	13
16	Normal	Female	Health and beauty	9
17	Normal	Female	Home and lifestyle	16
18	Normal	Female	Sports and travel	11
19	Normal	Male	Electronic accessories	13

	m_table$Customer.type <chr>	m_table$Gender <chr>	m_table$Product.line <chr>	n <int>
20	Normal	Male	Fashion accessories	13
21	Normal	Male	Food and beverages	16
22	Normal	Male	Health and beauty	16

1-10 of 12 rows Previous **1** 2 Next

Hide

```
colnames(n_table) <- c("Customer type","Gender","Product_line","Quantity")
colnames(n_table)
```

```
[1] "Customer type" "Gender"        "Product_line"  "Quantity"
```

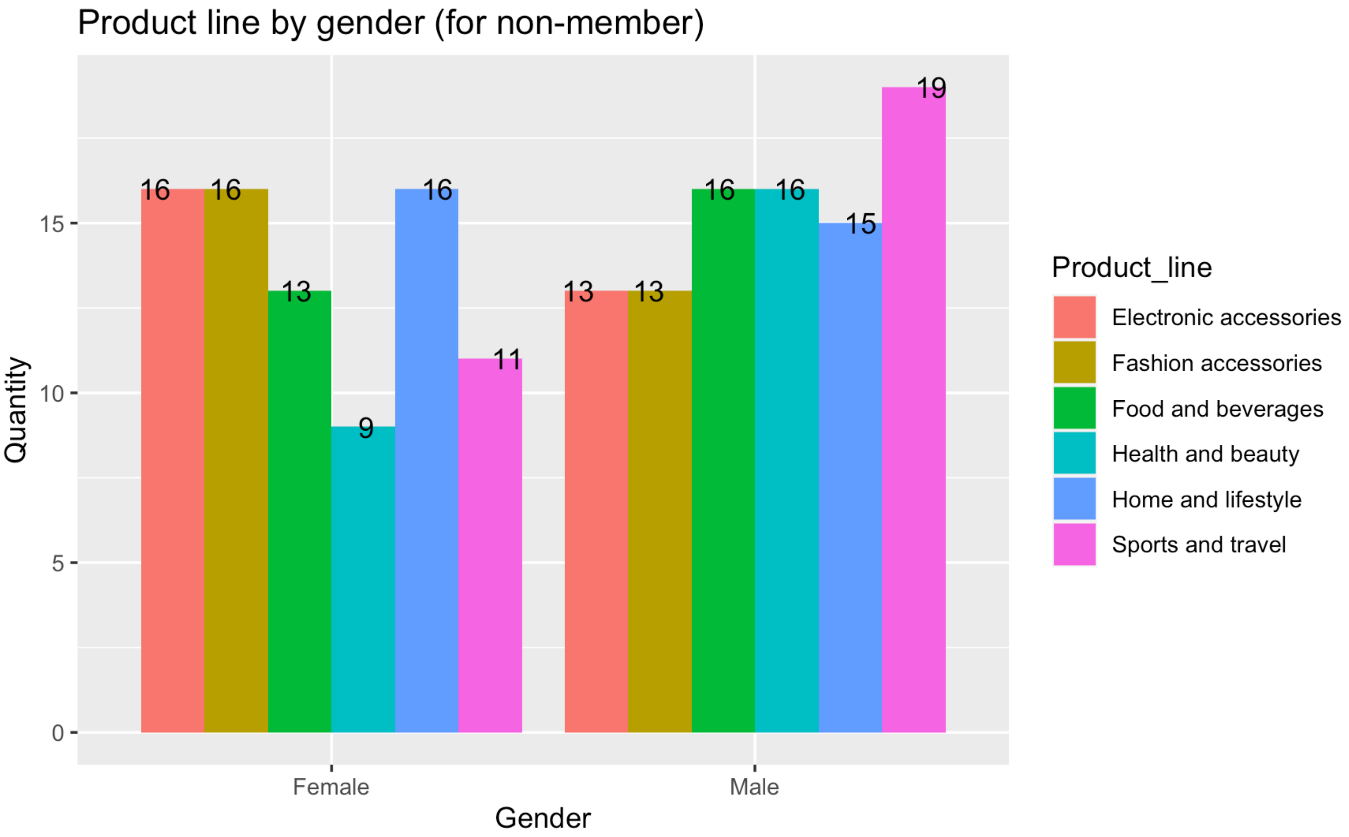
Hide

```
View(n_table)
```

*) Barplot using ggplot package :

Hide

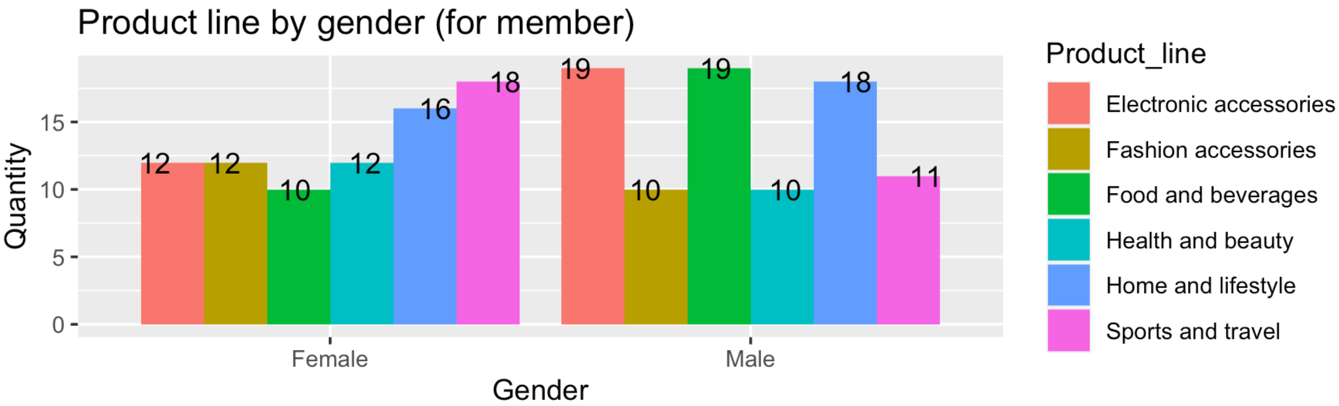
```
n <- ggplot(n_table,aes(x=Gender,y=Quantity,fill=Product_line)) +
  geom_bar(stat='identity',position = 'dodge') +
  labs(title = "Product line by gender (for non-member)",
       y="Quantity",
       x="Gender") +
  geom_text(n_table,mapping = aes(label=Quantity),stat = 'identity',position_dodge(wi
dth = 1)) #+
  #theme(legend.title = element_blank()) +
  #theme_fivethirtyeight()
n
```

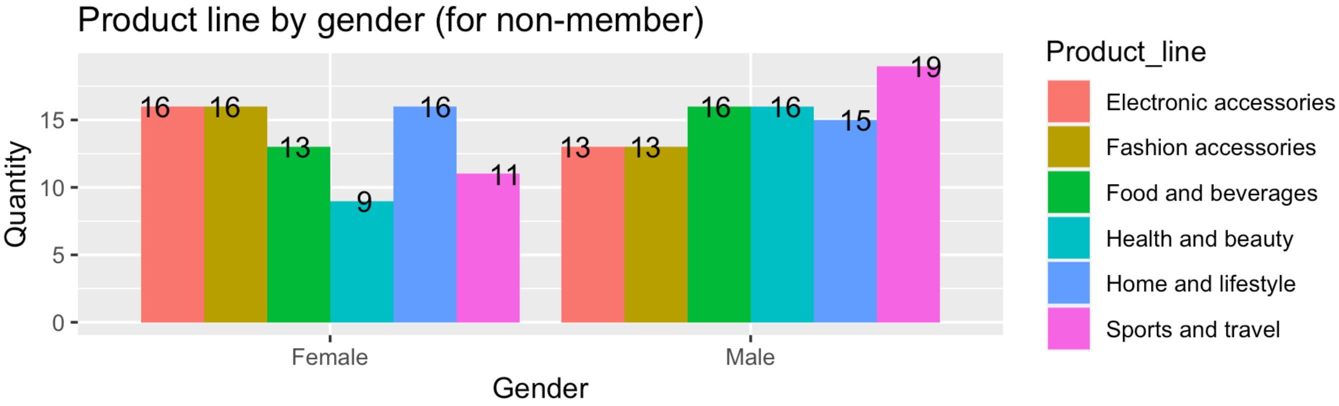


*) Combining 2 chart:

Hide

```
figure <- ggarrange(m, n,
                    labels = c("Member", "Non-member(normal)"),
                    ncol = 1, nrow = 2, hjust = 4)
figure
```


Product line by gender (for member)


Product line by gender (for non-member)

III/ Member table and non-member with rating(shopping experience on scale from 1 to 10):

*) Create a dataframe :

Hide

```
data1 <- data.frame(data$Customer.type,data$Gender,data$Rating,data$gross.income)
data1
```

data.Customer.type <chr>	data.Gender <chr>	data.Rating <dbl>	data.gross.income <dbl>
Member	Female	9.1	26.1415
Normal	Female	9.6	3.8200
Normal	Male	7.4	16.2155
Member	Male	8.4	23.2880

data.Customer.type <chr>	data.Gender <chr>	data.Rating <dbl>	data.gross.income <dbl>
Normal	Male	5.3	30.2085
Normal	Male	4.1	29.8865
Member	Female	5.8	20.6520
Normal	Female	8.0	36.7800
Member	Female	7.2	3.6260
Member	Female	5.9	8.2260

1-10 of 1,000 rows Previous **1** 2 3 4 5 6 ... 100 Next

a) “Member” :

*) Table :

Hide

```
m_table1 <- data1[data1$data.Customer.type == "Member", ]
m_table1
```

	data.Customer.type <chr>	data.Gender <chr>	data.Rating <dbl>	data.gross.income <dbl>
1	Member	Female	9.1	26.1415
4	Member	Male	8.4	23.2880
7	Member	Female	5.8	20.6520
9	Member	Female	7.2	3.6260
10	Member	Female	5.9	8.2260
11	Member	Female	4.5	2.8960
12	Member	Male	6.8	5.1020
16	Member	Female	4.5	28.1160
17	Member	Female	4.6	24.1255
21	Member	Male	4.8	21.5100

1-10 of 501 rows Previous **1** 2 3 4 5 6 ... 51 Next

Hide

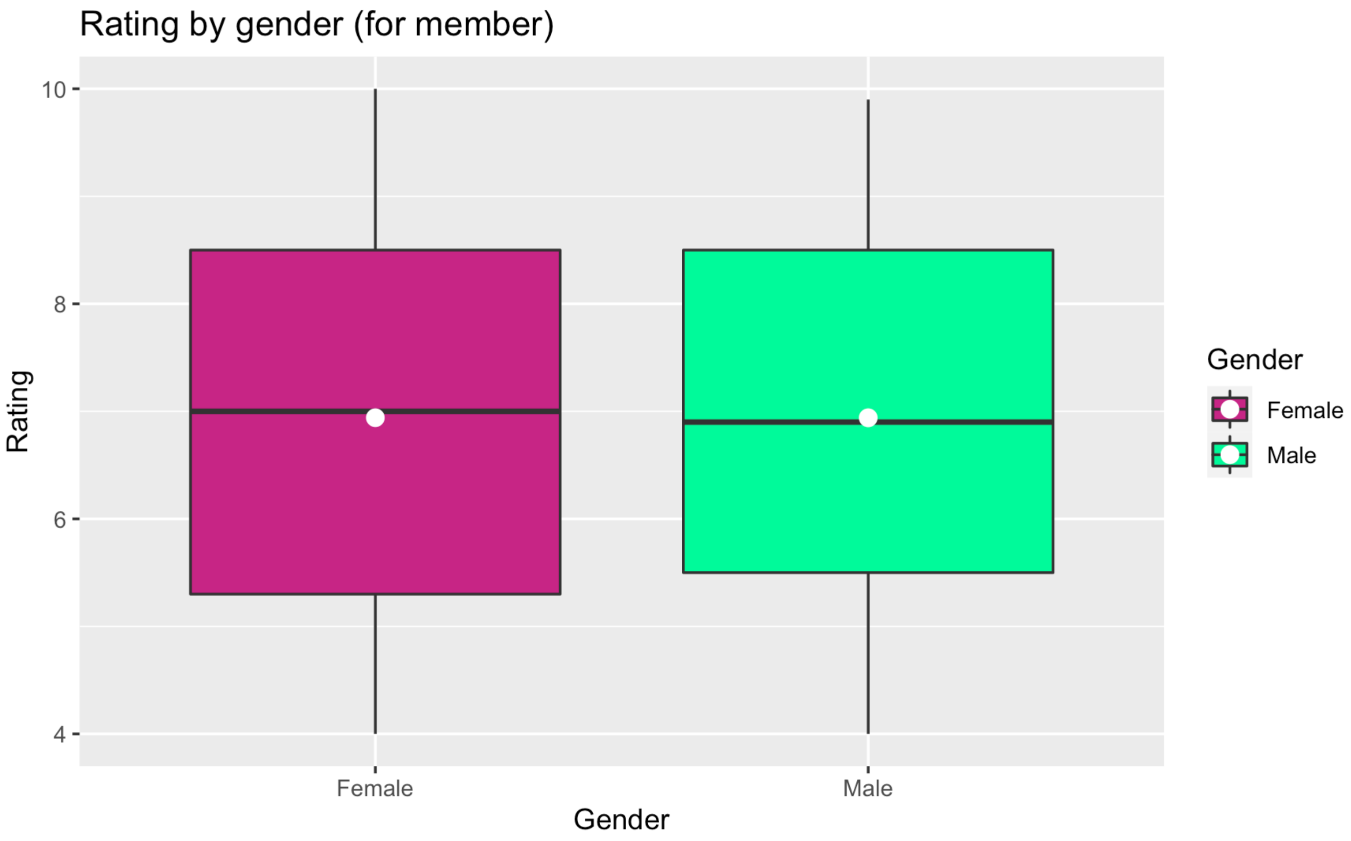
```
colnames(m_table1) <- c("Customer type","Gender","Rating","Gross_income")
colnames(m_table1)
```

```
[1] "Customer type" "Gender"        "Rating"       
[4] "Gross_income" 
```

*) Boxplot using ggplot :

Hide

```
m_1 <- ggplot(m_table1,aes(x=Gender,y=Rating,fill=Gender)) +
  geom_boxplot() +
  labs(title = "Rating by gender (for member)",
       y="Rating",
       x="Gender") +
  stat_summary(fun = mean, geom = "point", shape=16, size=3, color="white") +
  scale_fill_manual(values=c("mediumvioletred", "mediumspringgreen"))
m_1
```



b) “Member” :

*) Table :

Hide

```
n_table1 <- data1[data1$data.Customer.type == "Normal", ]
n_table1
```

	data.Customer.type	**data.Gender**	**data.Rating**	**data.gross.income**
	<chr>	<chr>	<dbl>	<dbl>
2	Normal	Female	9.6	3.8200
3	Normal	Male	7.4	16.2155
5	Normal	Male	5.3	30.2085
6	Normal	Male	4.1	29.8865
8	Normal	Female	8.0	36.7800
13	Normal	Female	7.1	11.7375
14	Normal	Male	8.2	21.5950
15	Normal	Female	5.7	35.6900
18	Normal	Male	6.9	21.7830
19	Normal	Male	8.6	8.2005

1-10 of 499 rows Previous **1** 2 3 4 5 6 ... 50 Next

Hide

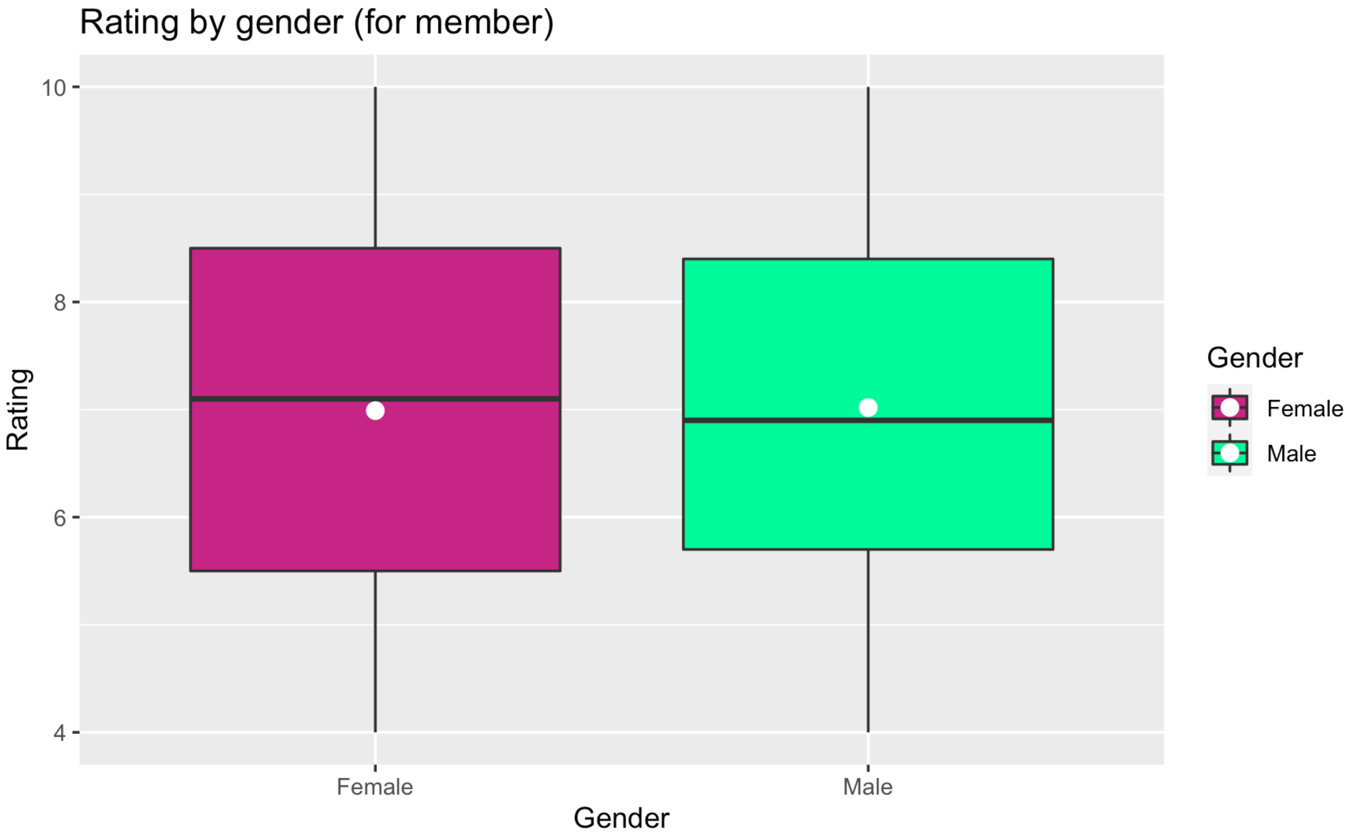
```
colnames(n_table1) <- c("Customer type","Gender","Rating","Gross_income")
colnames(n_table1)
```

```
[1] "Customer type" "Gender"        "Rating"
[4] "Gross_income"
```

*) Boxplot using ggplot :

Hide

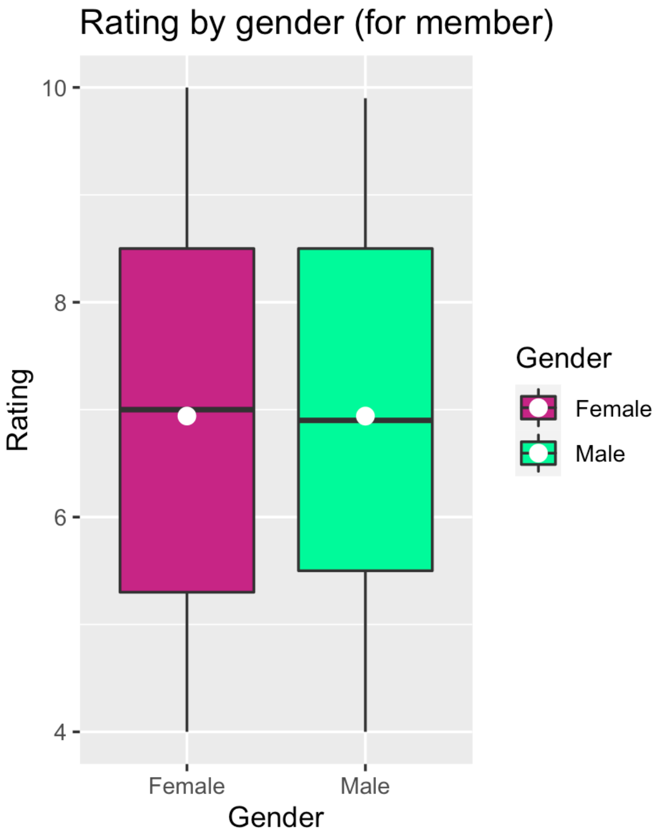
```
n_1 <- ggplot(n_table1,aes(x=Gender,y=Rating,fill=Gender)) +
  geom_boxplot() +
  labs(title = "Rating by gender (for member)",
       y="Rating",
       x="Gender") +
  stat_summary(fun = mean, geom = "point", shape=16, size=3, color="white") +
  scale_fill_manual(values=c("mediumvioletred", "mediumspringgreen"))
n_1
```

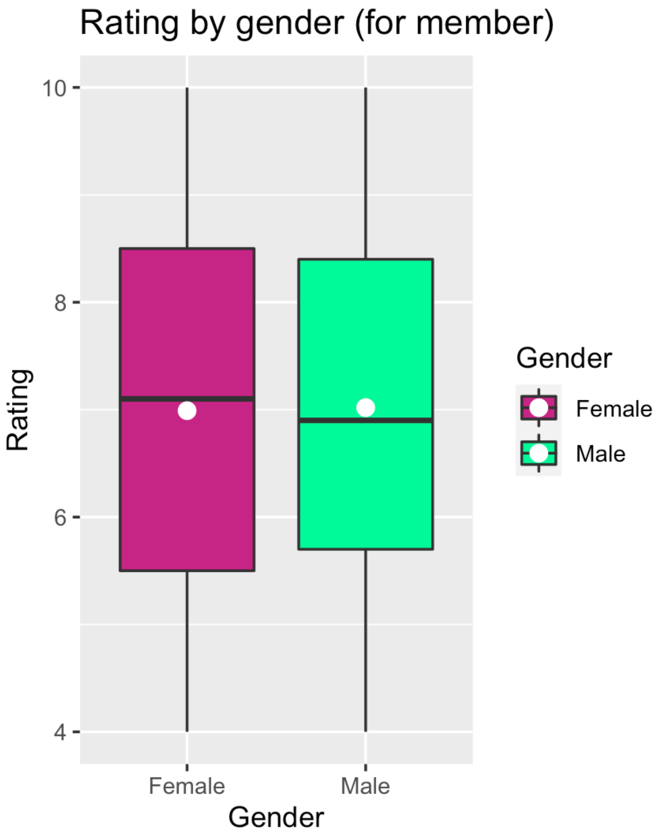


c) Combining 4 boxplots :

Hide

```
figure1 <- ggarrange(m_1, n_1,
                    labels = c("Member", "Non-member(normal)"),
                    ncol = 2, nrow = 1, hjust = 4)
figure1
```





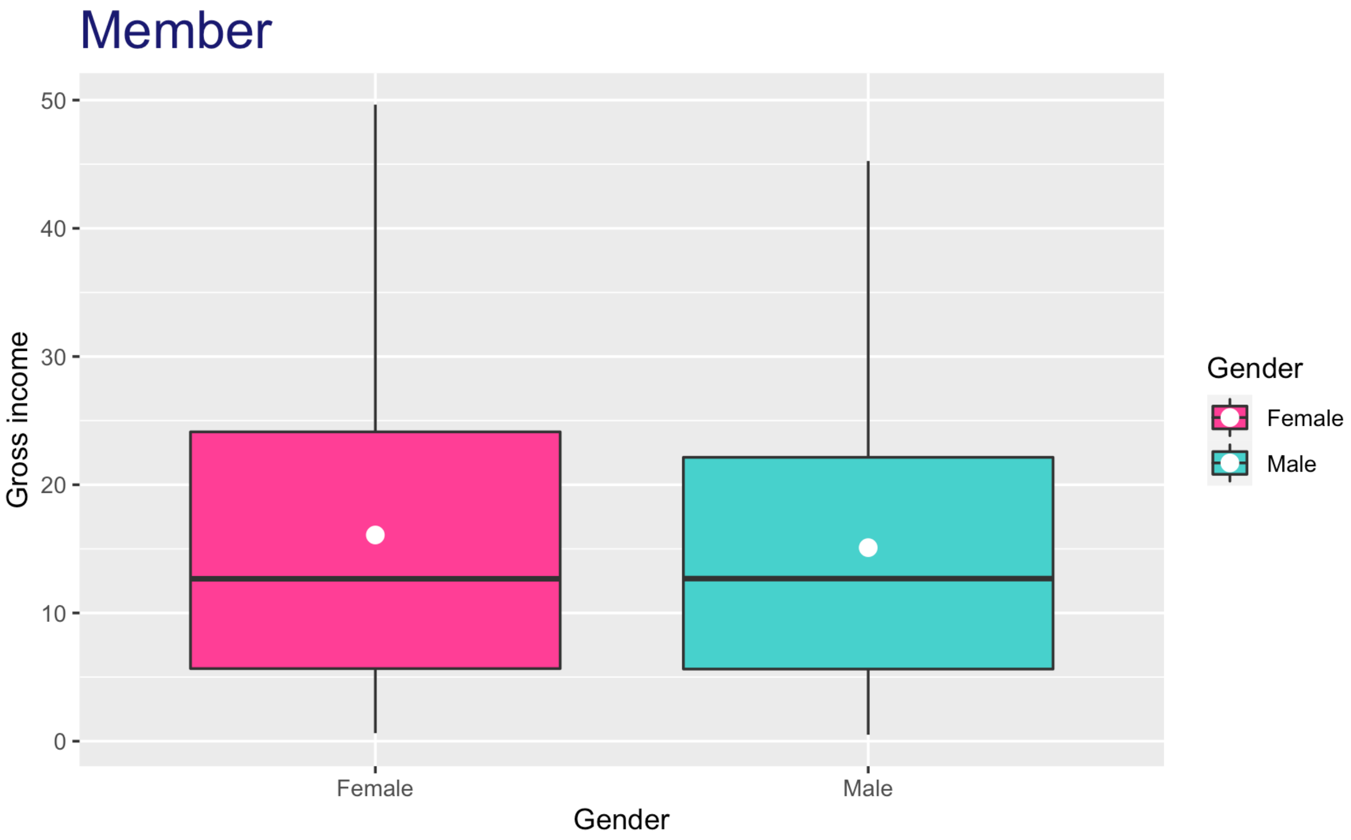
Comment: Both “member” or “non-member” have quite good shopping experience (almost 50% people in both genders, both types of customer have ranked their stratification 7/10).

IV/ Gross income:

a) Boxplot for “Member”:

Hide

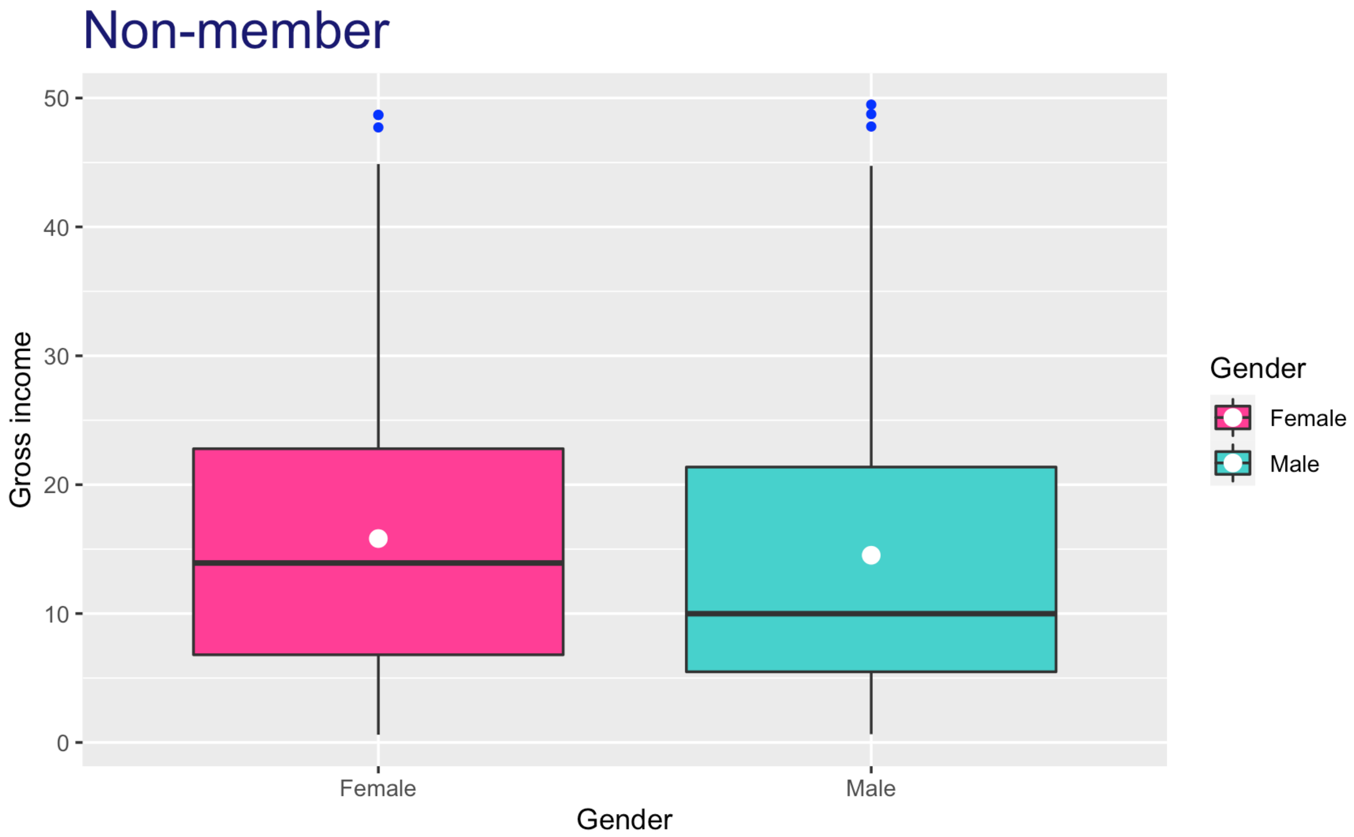
```
m_2 <- ggplot(m_table1,aes(x=Gender,y=Gross_income,fill=Gender)) +
  geom_boxplot() +
  labs(title = "Member",
       y="Gross income",
       x="Gender") +
  stat_summary(fun = mean, geom = "point", shape=16, size=3, color="white") +
  theme(plot.title = element_text(size = 20, color = "midnightblue")) +
  #stat_boxplot(geom = "errorbar", width=0.3) +
  scale_fill_manual(values=c("violetred1", "mediumturquoise"))
m_2
```



b) Boxplot for "Non-member":

Hide

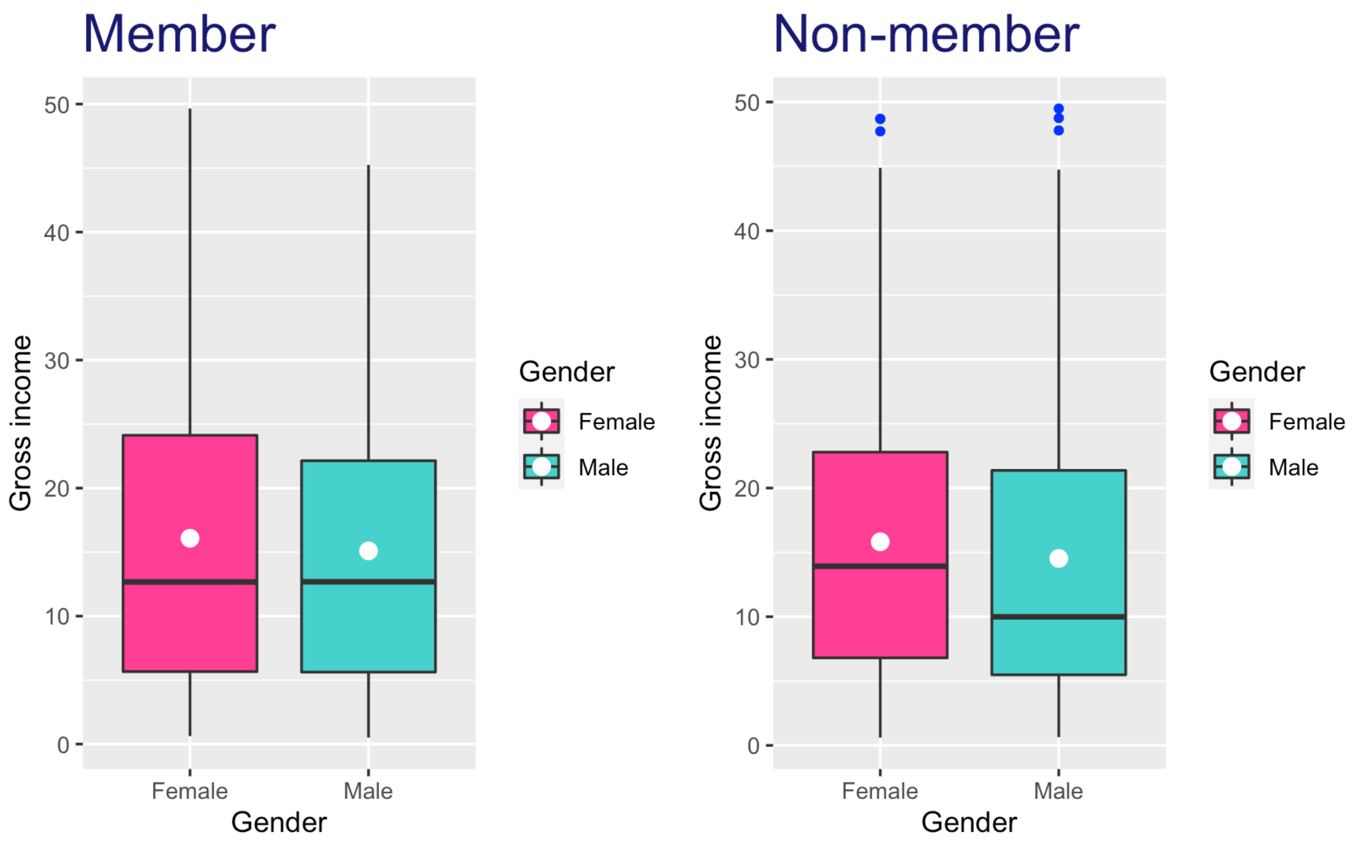
```
n_2 <- ggplot(n_table1,aes(x=Gender,y=Gross_income,fill=Gender)) +
  geom_boxplot(outlier.colour = "blue",outlier.shape = 16) +
  labs(title = "Non-member",
       y="Gross income",
       x="Gender") +
  stat_summary(fun = mean, geom = "point", shape=16, size=3,color="white") +
  theme(plot.title = element_text(size = 20, color = "midnightblue")) +
  #stat_boxplot(geom = "errorbar", width=0.3) +
  scale_fill_manual(values=c("violetred1", "mediumturquoise"))
n_2
```



c) Combining 4 boxplots:

Hide

```
figure2 <- ggarrange(m_2,n_2,
                     labels = c("Member", "Non-member(normal)"),
                    ncol = 2, nrow = 1, hjust = 4)
figure2
```



Comment:

) “Member”: #### 50% of female and 50% male have the the same median in gross income(about 12), female’s 3rd quartile value is greater than male’s 3rd quartile value(almost 25 > 22). Both genders also have the same 1st quartile value(slightly more than 5).For female, the max value even reaches to 50, for male it’s just 45.There are no outlier here. #### ###) “Non-member”:

Female has greater median value than male(14 > 10),it seems like there isn’t much difference in 3rd quartile value for both genders.Both genders have the same max value.Notice that there is even some outliers in this chart for “non-member”, male just has one more outlier than female(almost reach to 50).

V/Payment:

*) Dataframe :

Hide

```
data2 <- data.frame(data$Branch,data$Customer.type,data$Gender,data$Payment)
data2
```

data.Branch <chr>	data.Customer.type <chr>	data.Gender <chr>	data.Payment <chr>
A	Member	Female	Ewallet
C	Normal	Female	Cash
A	Normal	Male	Credit card
A	Member	Male	Ewallet
A	Normal	Male	Ewallet
C	Normal	Male	Ewallet
A	Member	Female	Ewallet
C	Normal	Female	Ewallet
A	Member	Female	Credit card
B	Member	Female	Credit card

1-10 of 1,000 rows Previous **1** 2 3 4 5 6 ... 100 Next

*) Sort data for branch A uniquely:

Hide

```
p_table <- data2[data2$data.Branch=="A", ]
p_table
```

	data.Branch <chr>	data.Customer.type <chr>	data.Gender <chr>	data.Payment <chr>
1	A	Member	Female	Ewallet
3	A	Normal	Male	Credit card
4	A	Member	Male	Ewallet
5	A	Normal	Male	Ewallet
7	A	Member	Female	Ewallet
9	A	Member	Female	Credit card

	data.Branch <chr>	data.Customer.type <chr>	data.Gender <chr>	data.Payment <chr>
13	A	Normal	Female	Ewallet
14	A	Normal	Male	Ewallet
15	A	Normal	Female	Cash
17	A	Member	Female	Credit card

1-10 of 340 rows Previous **1** 2 3 4 5 6 ... 34 Next

*) The amount of people in each type of payment based on customer type and gender:

Hide

```
p <- p_table %>%
  count(p_table$data.Customer.type,p_table$data.Gender,p_table$data.Payment) %>%
  mutate(prop_payment = round((n/sum(n))*100,2))
p
```

p_table$data.Customer.type <chr>	p_table$data.Gender <chr>	p_table$data.Payment <chr>	n <int>	prop_pay
Member	Female	Cash	34	
Member	Female	Credit card	20	
Member	Female	Ewallet	26	
Member	Male	Cash	22	
Member	Male	Credit card	29	
Member	Male	Ewallet	36	
Normal	Female	Cash	25	
Normal	Female	Credit card	27	
Normal	Female	Ewallet	29	
Normal	Male	Cash	29	

1-10 of 12 rows Previous **1** 2 Next

Hide

```
sum(p$n)
```

```
[1] 340
```

Hide

```
#View(p)
```

*) Payment table for “Member”:

Hide

```
p_m <- p[p$`p_table$data.Customer.type`=="Member", ]
p_m
```

	p_table$data.Customer.type <chr>	p_table$data.Gender <chr>	p_table$data.Payment <chr>	n <int>	prop_pa
1	Member	Female	Cash	34	
2	Member	Female	Credit card	20	
3	Member	Female	Ewallet	26	
4	Member	Male	Cash	22	
5	Member	Male	Credit card	29	
6	Member	Male	Ewallet	36	

6 rows

Hide

```
colnames(p_m) <- c("Customer_type","Gender","Payment","Quantity","%")
colnames(p_m)
```

```
[1] "Customer_type" "Gender"        "Payment"
[4] "Quantity"      "%"
```

*) Payment table for “Non-member”:

Hide

```
p_n <- p[p$`p_table$data.Customer.type`=="Normal", ]
p_n
```

	p_table$data.Customer.type	p_table$data.Gender	p_table$data.Payment	n	prop_p
	<chr>	<chr>	<chr>	<int>	
7	Normal	Female	Cash	25	
8	Normal	Female	Credit card	27	
9	Normal	Female	Ewallet	29	
10	Normal	Male	Cash	29	
11	Normal	Male	Credit card	28	
12	Normal	Male	Ewallet	35	

6 rows

Hide

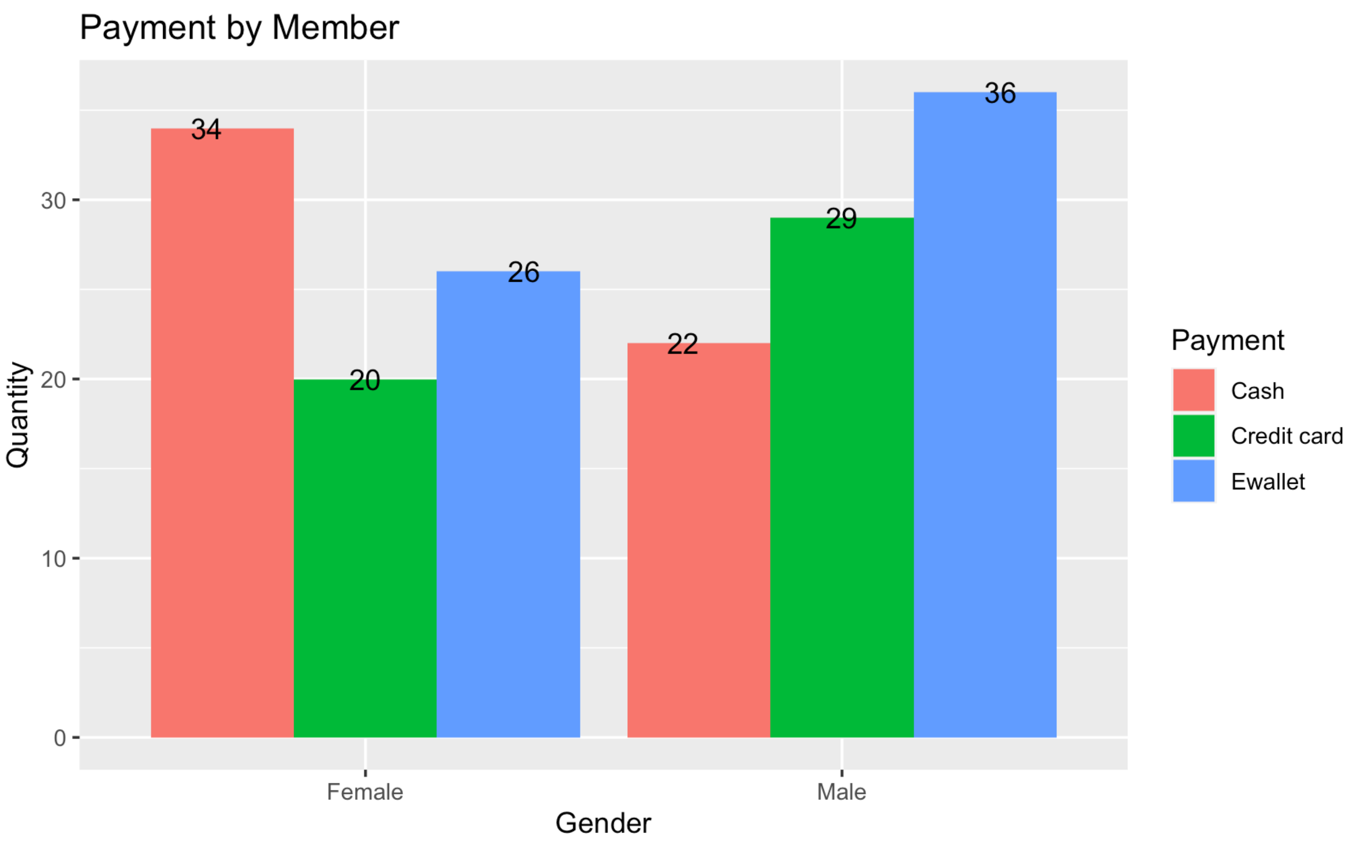
```
colnames(p_n) <- c("Customer_type","Gender","Payment","Quantity","%")
colnames(p_n)
```

```
[1] "Customer_type" "Gender"        "Payment"
[4] "Quantity"      "%"
```

*) ggplot chart: #### ### “Payment by Member”:

Hide

```
p1 <- ggplot(p_m,aes(x=Gender,y=Quantity,fill=Payment)) +
  geom_bar(stat='identity',position = 'dodge') +
  labs(title = ("Payment by Member"),
       y="Quantity",
       x="Gender") +
  geom_text(p_m,mapping = aes(label=Quantity),stat = 'identity',position_dodge(width
= 1))
p1
```

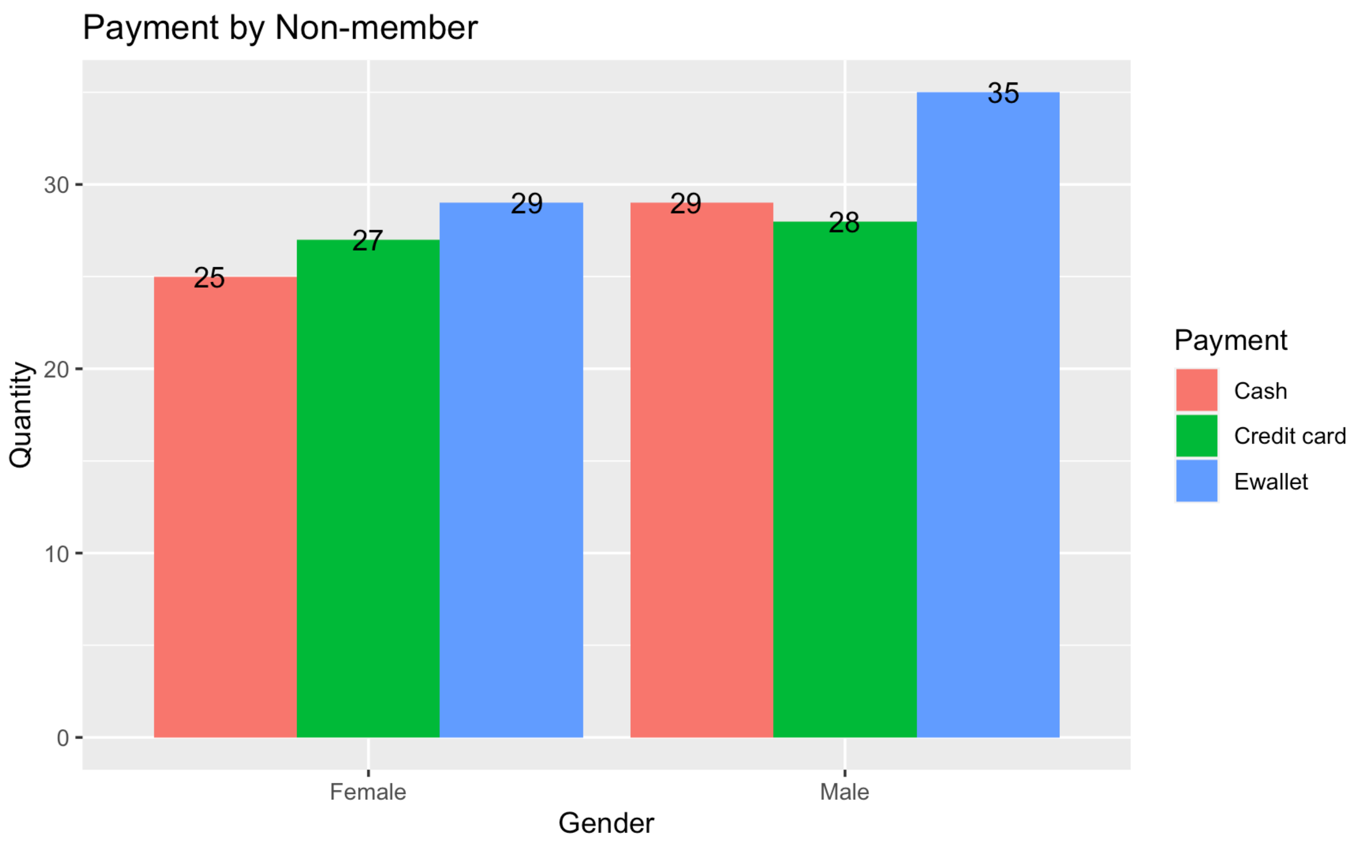


Payment by Member

“Payment by Non-member”:

Hide

```
p2 <- ggplot(p_n,aes(x=Gender,y=Quantity,fill=Payment)) +
  geom_bar(stat='identity',position = 'dodge') +
  labs(title = ("Payment by Non-member"),
       y="Quantity",
       x="Gender") +
  geom_text(p_n,mapping = aes(label=Quantity),stat = 'identity',position_dodge(width
= 1))
p2
```

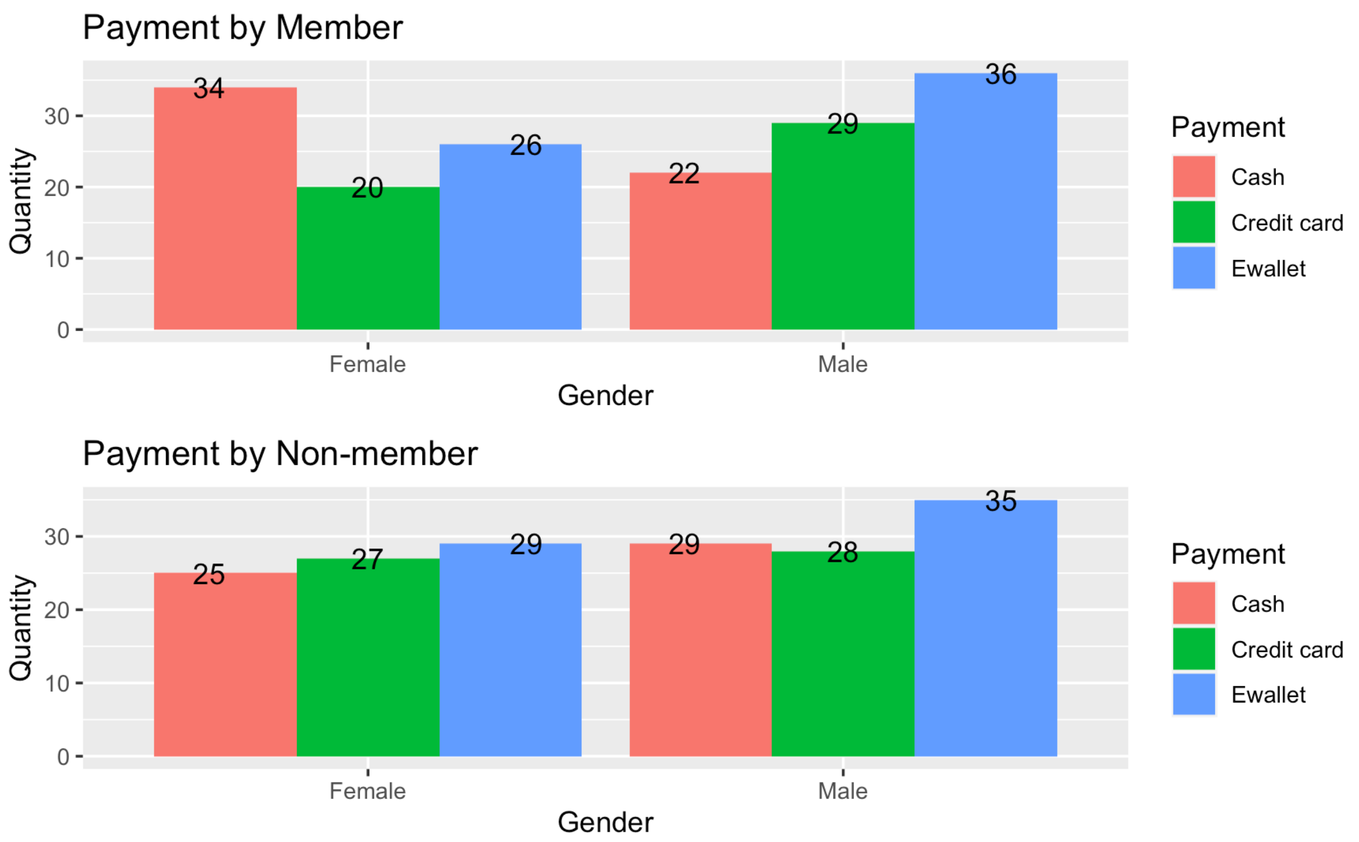


Combining chart:

a) Bar chart:

Hide

```
figure3 <- ggarrange(p1,p2,
                     labels = c("Member","Non-member"),
                     ncol=1,nrow = 2,hjust = 4)
figure3
```



b) Pie chart(for whole branch A):

*) Dataframe:

Hide

```
P_table <- p_table %>%
  count(p_table$data.Branch,p_table$data.Payment) %>%
  mutate(Percentage = (n/sum(n))*100)
P_table
```

p_table$data.Branch <chr>	p_table$data.Payment <chr>	n <int>	Percentage <dbl>
A	Cash	110	32.35294
A	Credit card	104	30.58824
A	Ewallet	126	37.05882

3 rows

Hide

```
colnames(P_table) <- c("Branch","Payment","Quantity","Percentage")
colnames(P_table)
```

```
[1] "Branch"     "Payment"    "Quantity"   "Percentage"
```

*) Pie chart using ggpie from ggpubr package:

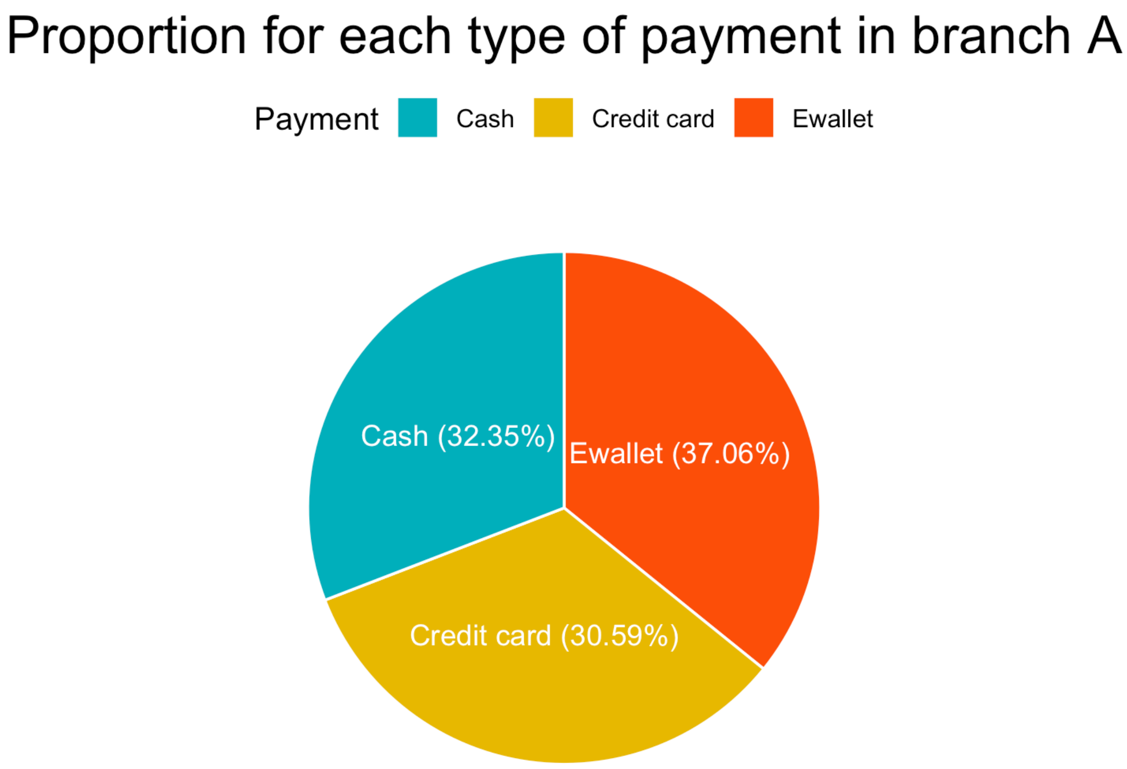
Hide

```
labs <- paste0(p_n1$Payment, " (", round(P_table$Percentage,2), "%)")
labs
```

```
[1] "Cash (32.35%)"        "Credit card (30.59%)"
[3] "Ewallet (37.06%)"
```

Hide

```
# Female member :
ppg_A <- ggpie(Py_nfm,"Prop_pay",label = labs,
   lab.pos = "in", lab.font = "white",
      fill = "Payment",
      color = "white",
       palette = c("#00AFBB", "#E7B800", "#FC4E07")) +
  ggtitle("Proportion for each type of payment in branch A") +
  theme(plot.title = element_text(hjust = 0.5,size = 20))
ppg_A
```


Proportion for each type of payment in branch A
Payment
Cash
Credit card
Ewallet
Cash (32.35%)
Ewallet (37.06%)
Credit card (30.59%)