Math 170S HW2

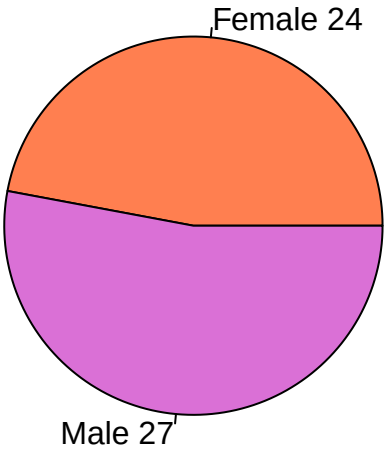
Jun Ryu

2023-01-25

```
df <- read.table("170s_hw2_data.txt", header = TRUE)
```

(1) exploratory analysis using sex

```
table <- table(df$Sex)
lbls <- paste(names(table), table, sep=" ")
pie(table, labels = lbls, col = c("coral", "orchid"))
```



(2) exploratory analysis using age

```
summary(df$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   22.00   24.00   24.51   27.00   33.00
```

```
IQR(df$Age)
```

```
## [1] 5
```

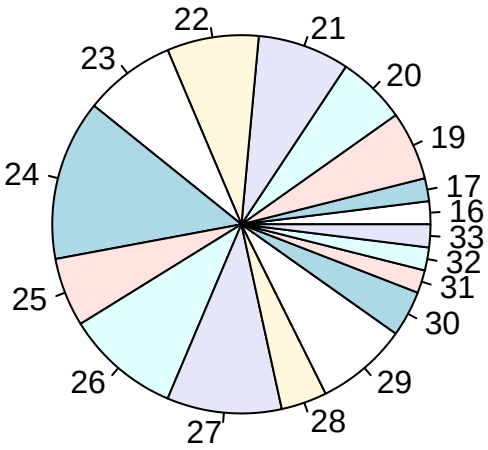
```
sd(df$Age)
```

```
## [1] 3.859391
```

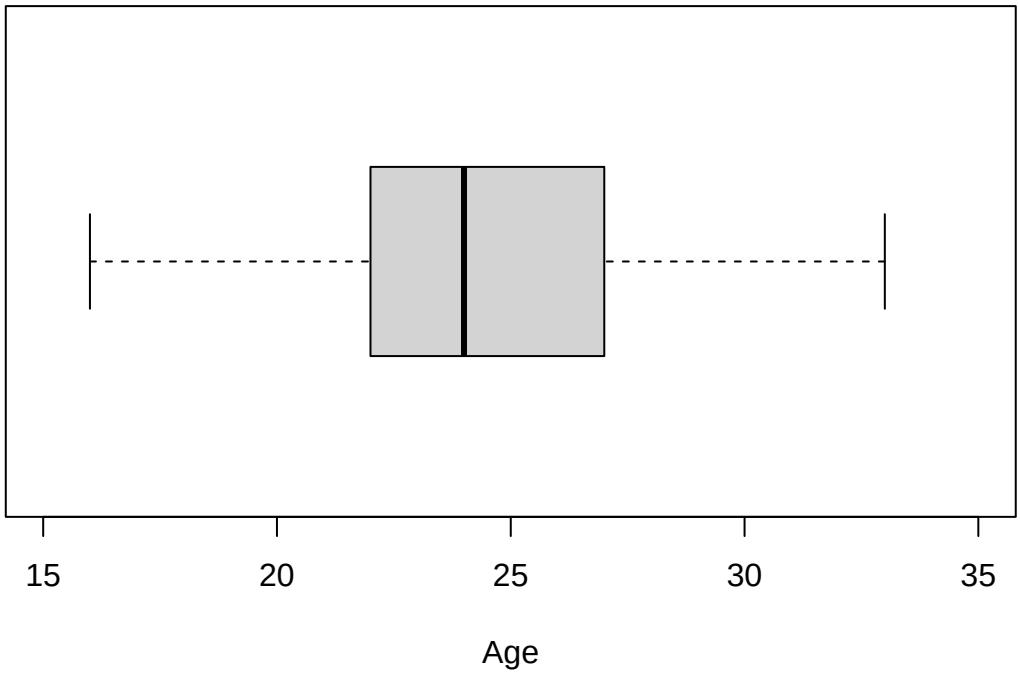
```
var(df$Age)
```

```
## [1] 14.8949
```

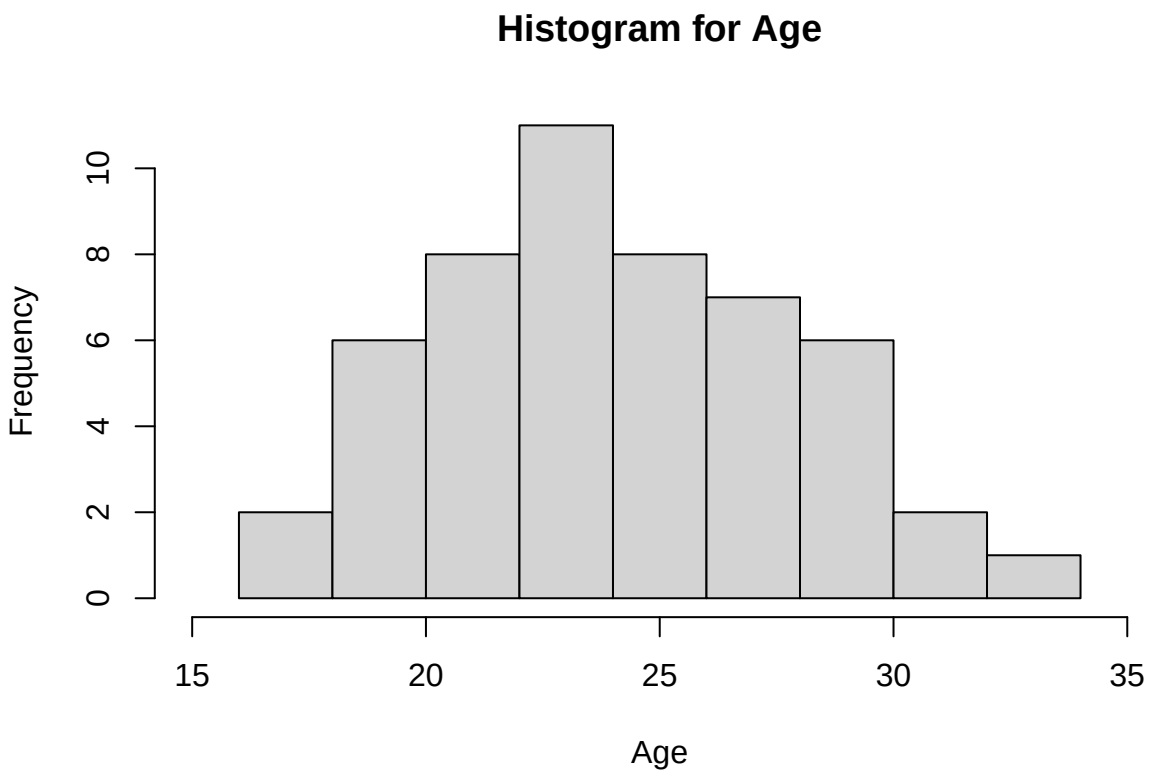
```
pie(table(df$Age), main = "Pie chart for Age")
```



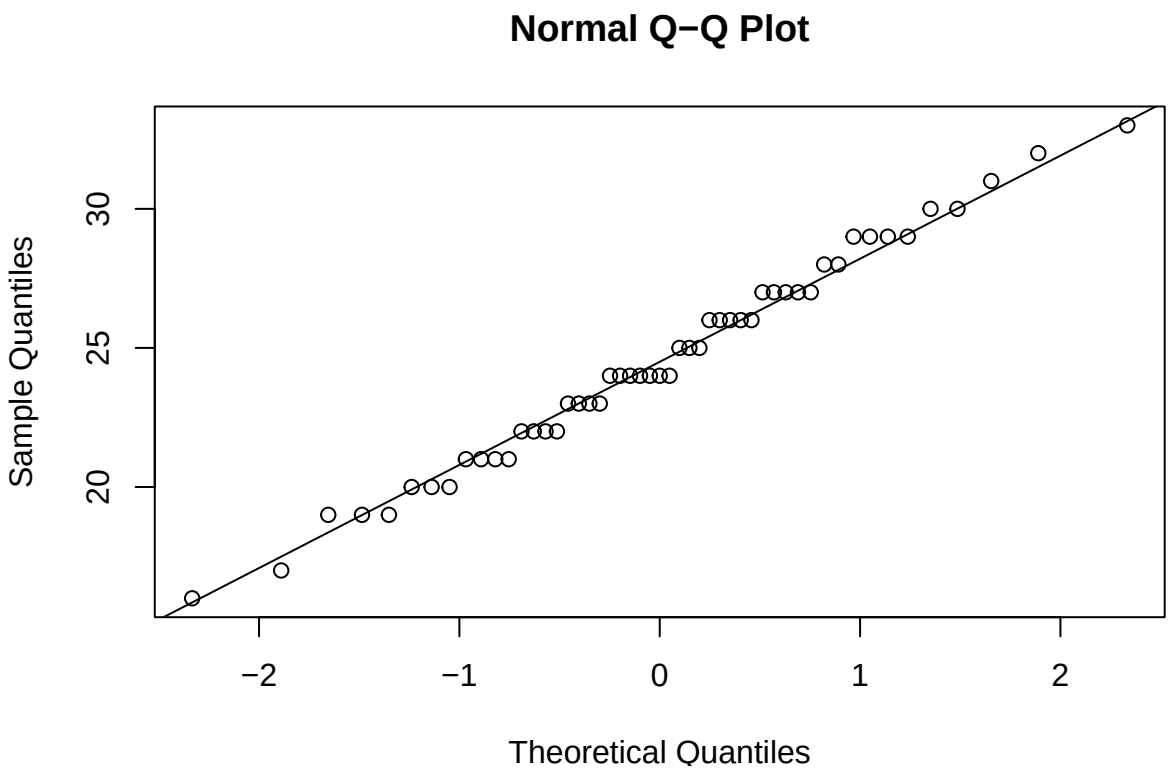

```
boxplot(df$Age, horizontal = TRUE, xlab = "Age", ylim = c(15,35))
```



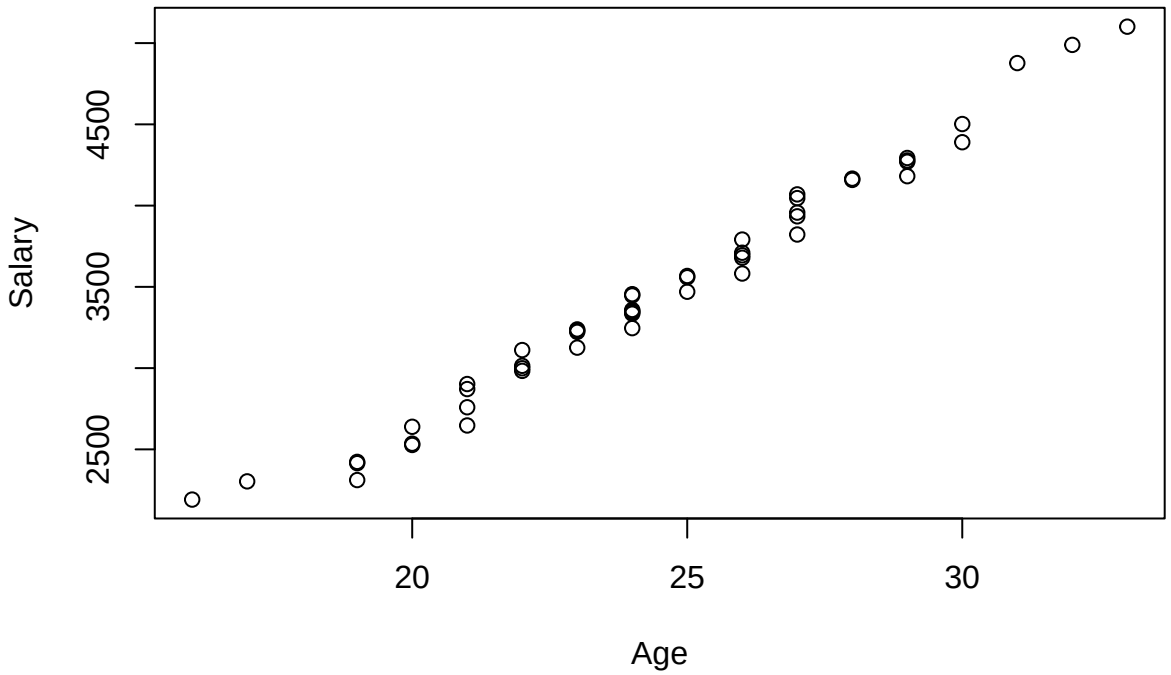

```
hist(df$Age, xlim = c(15,35), main = "Histogram for Age", xlab = "Age")
```



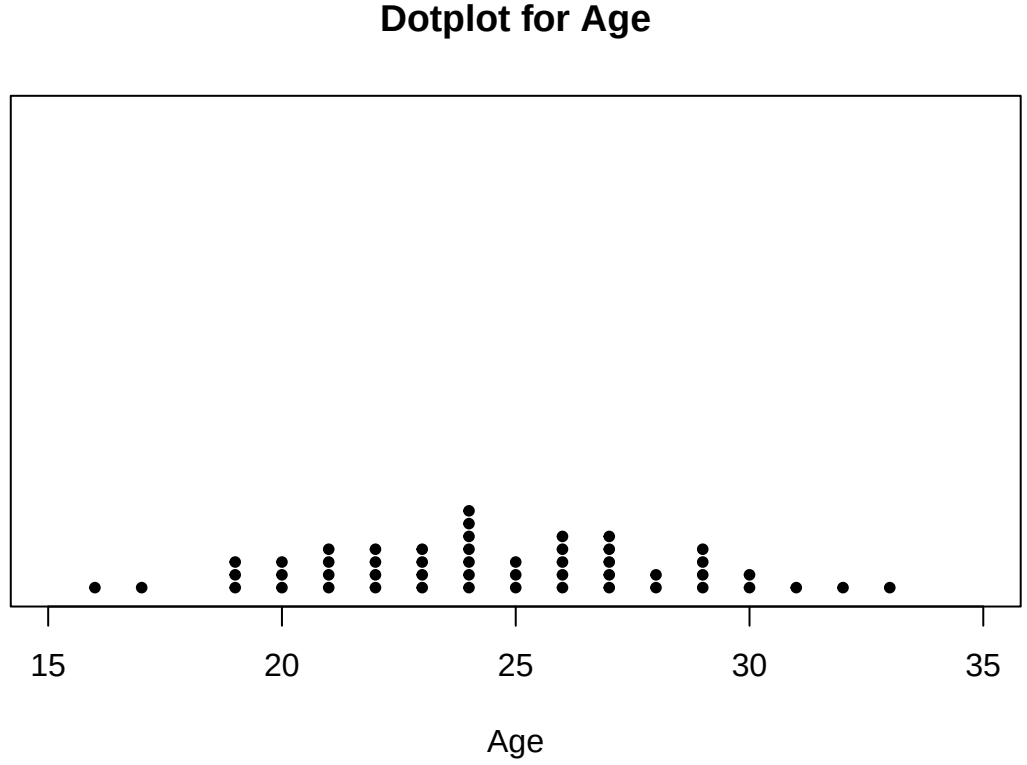

```
qqnorm(df$Age)
qqline(df$Age)
```



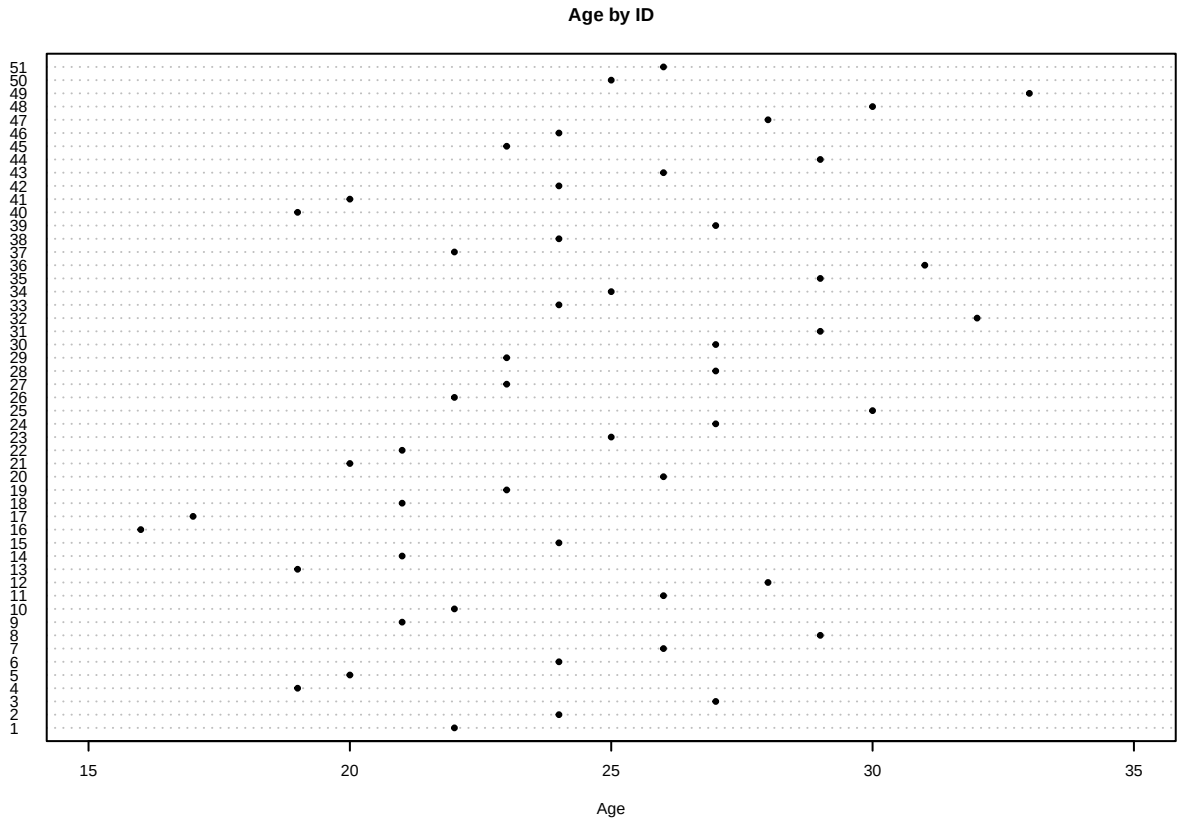

```
qqplot(df$Age, df$Salary, xlab = "Age", ylab = "Salary") # age vs. salary
```




```
stripchart(df$Age, method = "stack", at = 0, xlab = "Age", xlim = c(15,35),
           pch = 20, main = "Dotplot for Age")
```




```
dotchart(df$Age, labels = row.names(df), xlab = "Age", xlim = c(15,35),
         pch = 20, cex = 0.5, main = "Age by ID")
```



(3) exploratory analysis using salary

```
summary(df$Salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2191    2991    3447    3472    4002    5101
```

```
IQR(df$Salary)
```

```
## [1] 1010.5
```

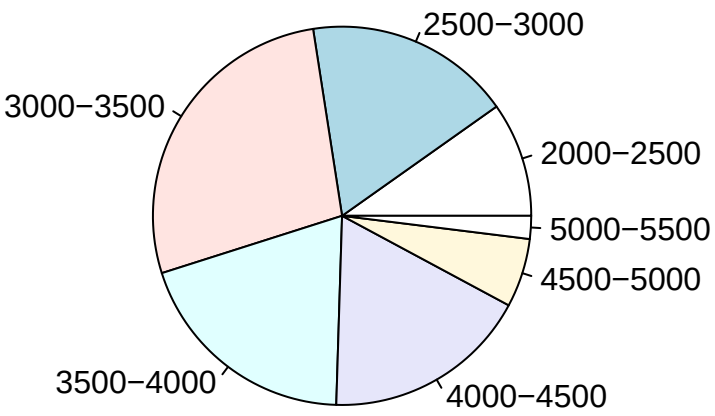
```
sd(df$Salary)
```

```
## [1] 719.8625
```

```
var(df$Salary)
```

```
## [1] 518202.1
```

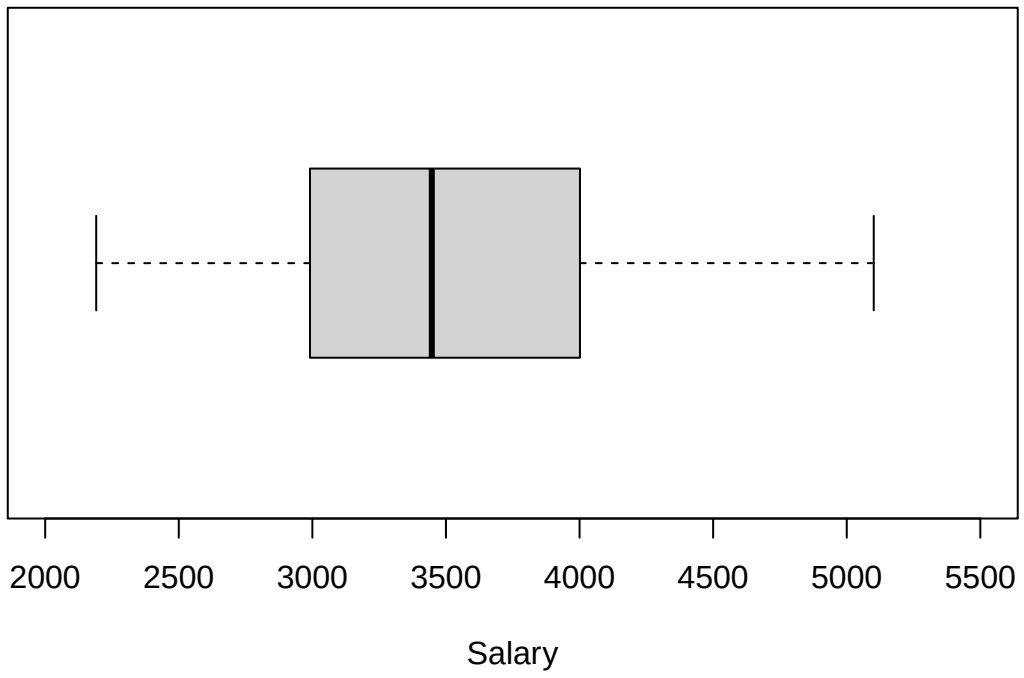
```
salary_cat <- cut(df$Salary, breaks = seq(2000,5500,500), labels =
                    c("2000-2500","2500-3000","3000-3500","3500-4000","4000-4500",
                      "4500-5000","5000-5500"))
pie(table(salary_cat), main = "Pie chart for Salary")
```




```
table(salary_cat)
```

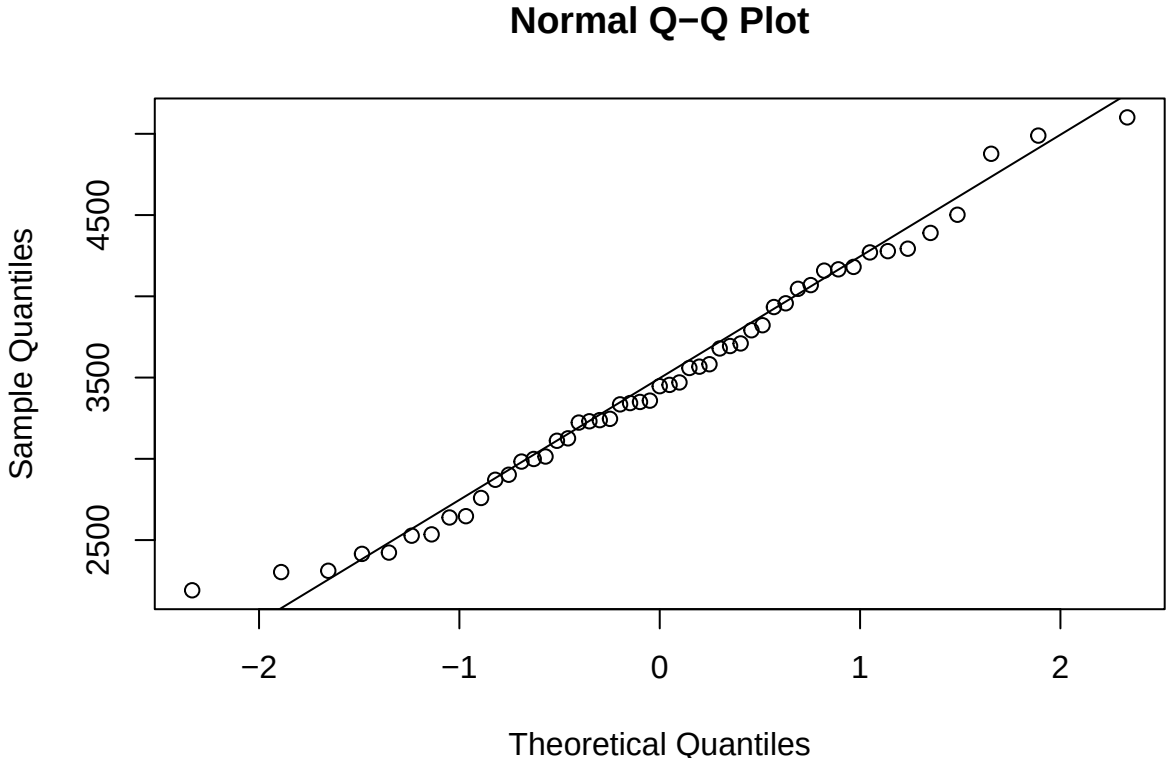
```
## salary_cat
## 2000-2500 2500-3000 3000-3500 3500-4000 4000-4500 4500-5000 5000-5500
##         5         9        14        10         9         3         1
```

```
boxplot(df$Salary, horizontal = TRUE, xlab = "Salary", ylim = c(2000,5500))
```

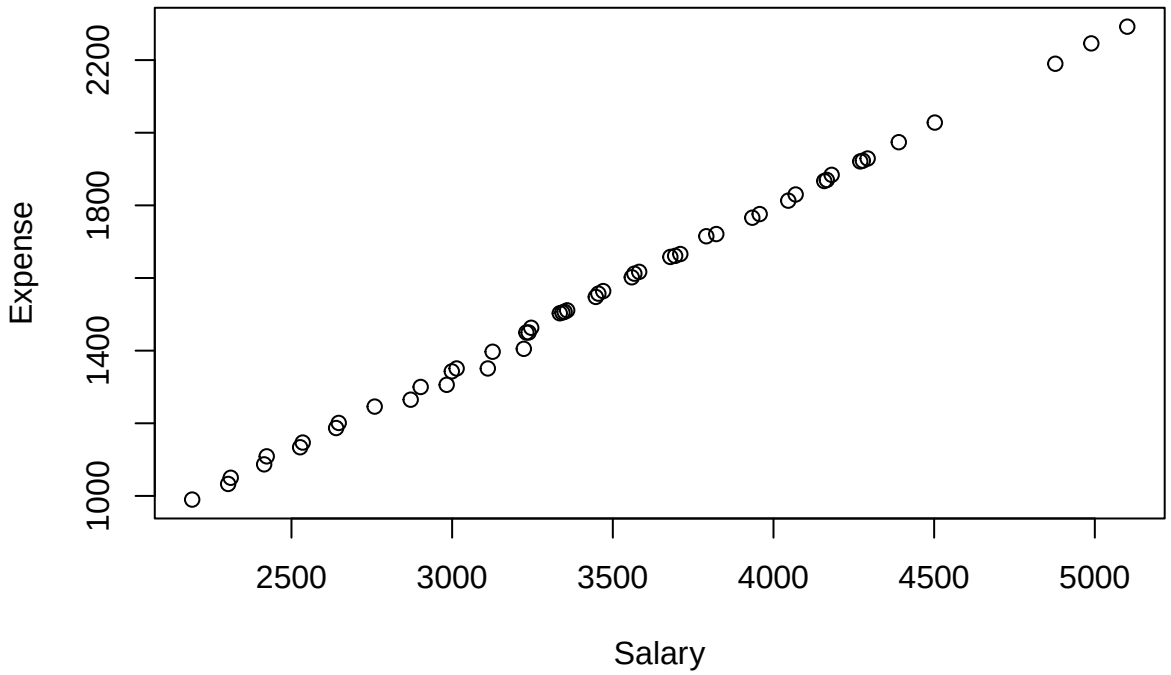



```
hist(df$Salary, xlim = c(2000,5500), main = "Histogram for Salary", xlab = "Salary")
```

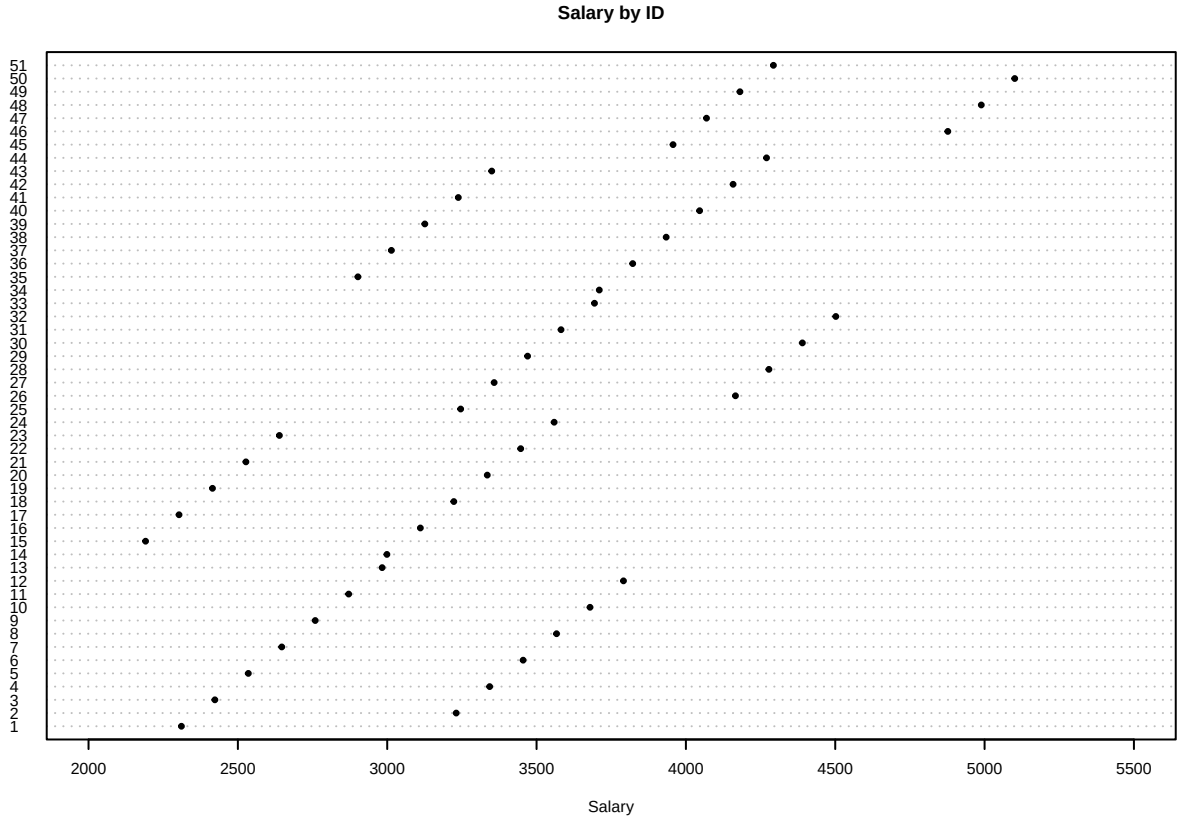

```
qqnorm(df$Salary)
qqline(df$Salary)
```




```
qqplot(df$Salary, df$Expense, xlab = "Salary", ylab = "Expense") # salary vs. expense
```




```
dotchart(df$Salary, labels = row.names(df), xlab = "Salary", xlim = c(2000,5500),
         pch = 20, cex = 0.5, main = "Salary by ID")
```



(4) exploratory analysis using expense

```
summary(df$Expense)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     990    1324    1548    1559    1794    2292
```

```
IQR(df$Expense)
```

```
## [1] 470
```

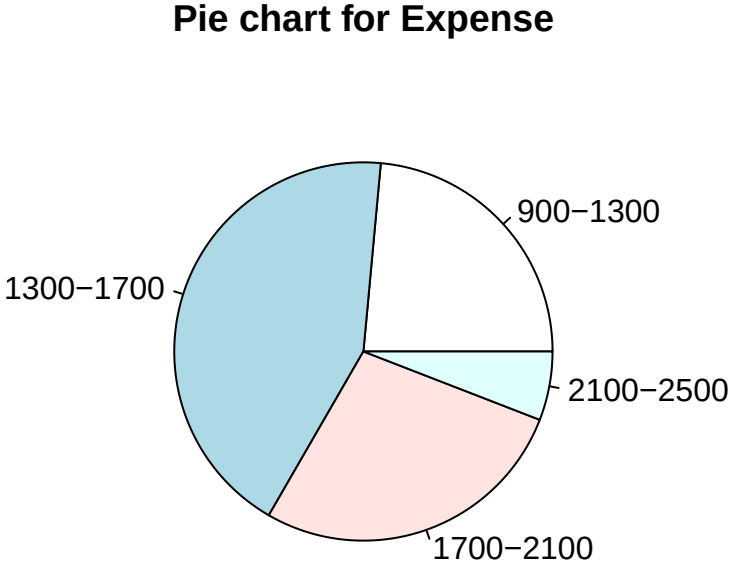
```
sd(df$Expense)
```

```
## [1] 324.3439
```

```
var(df$Expense)
```

```
## [1] 105199
```

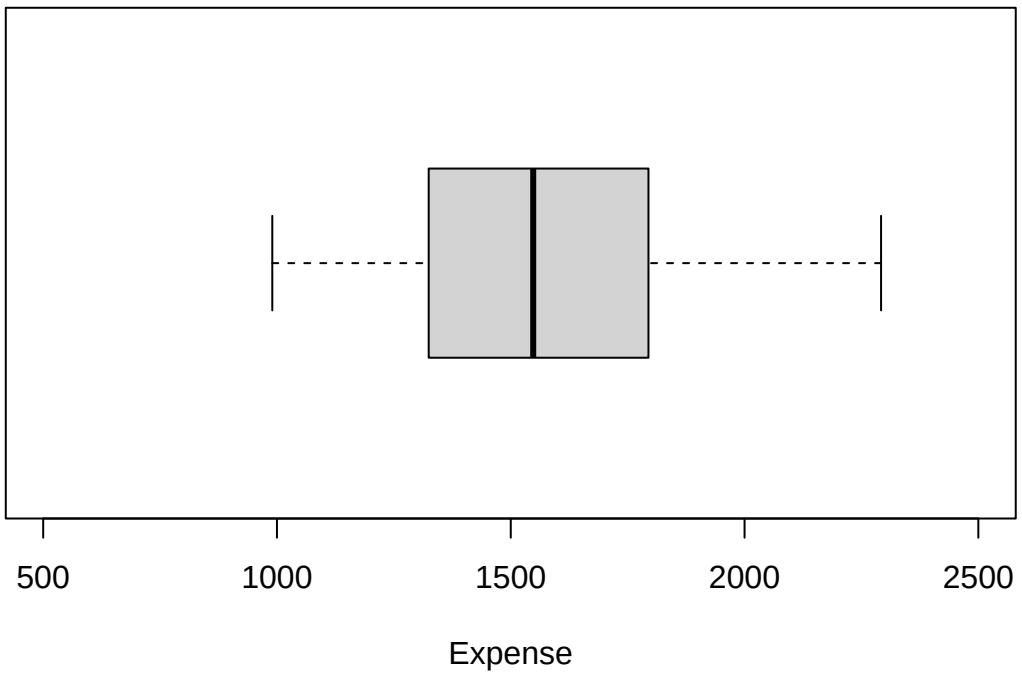
```
expense_cat <- cut(df$Expense, labels = c("900-1300","1300-1700","1700-2100",
                                          "2100-2500"), breaks = seq(900,2500,400))
pie(table(expense_cat), main = "Pie chart for Expense")
```



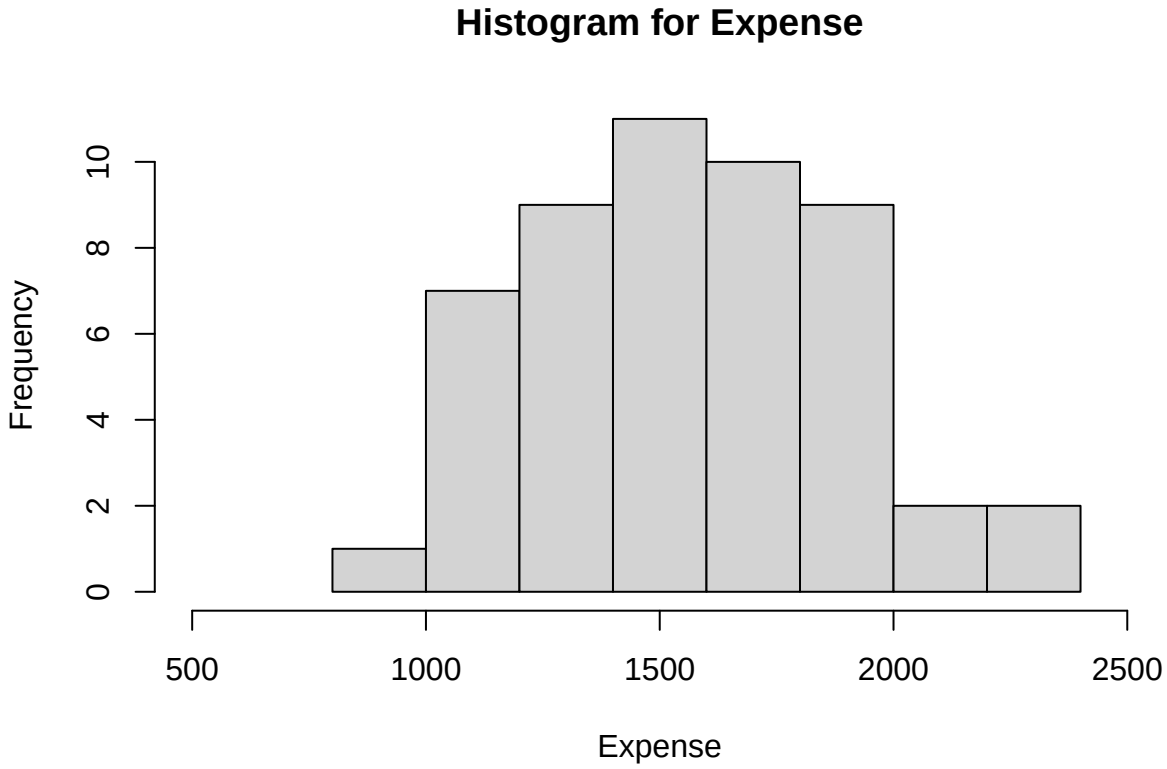

```
table(expense_cat)
```

```
## expense_cat
##  900-1300 1300-1700 1700-2100 2100-2500
##        12        22        14         3
```

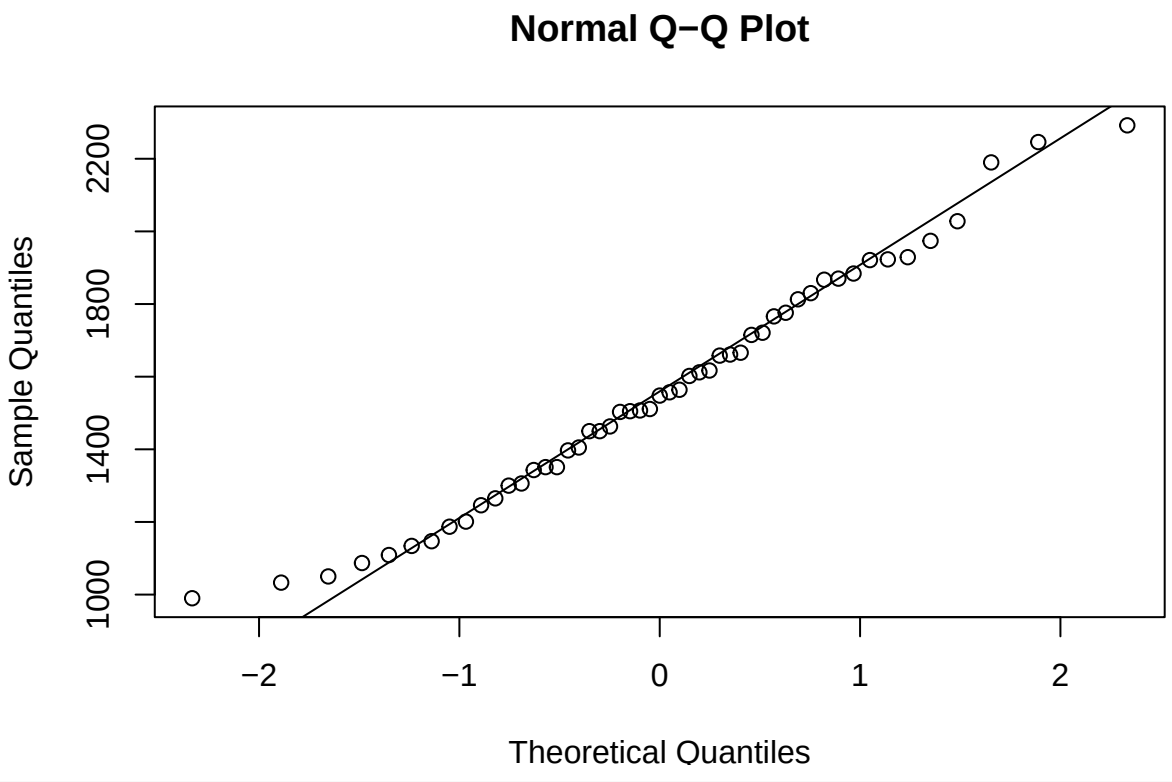
```
boxplot(df$Expense, horizontal = TRUE, xlab = "Expense", ylim = c(500,2500))
```



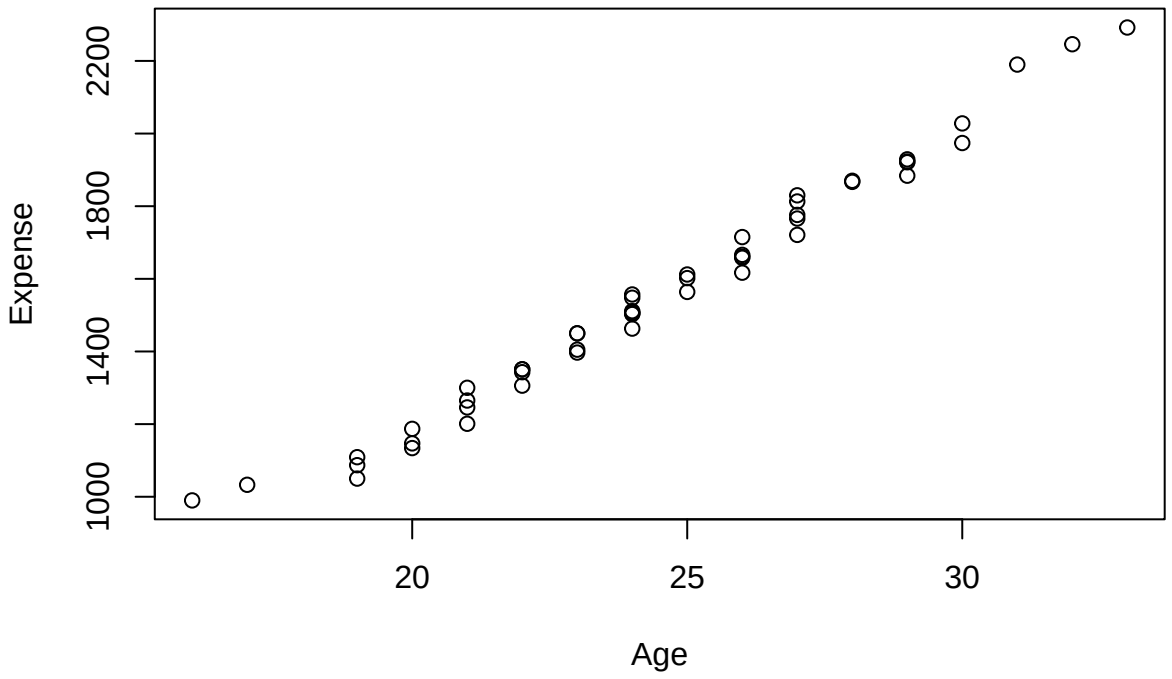

```
hist(df$Expense, xlim = c(500,2500), main = "Histogram for Expense", xlab = "Expense")
```



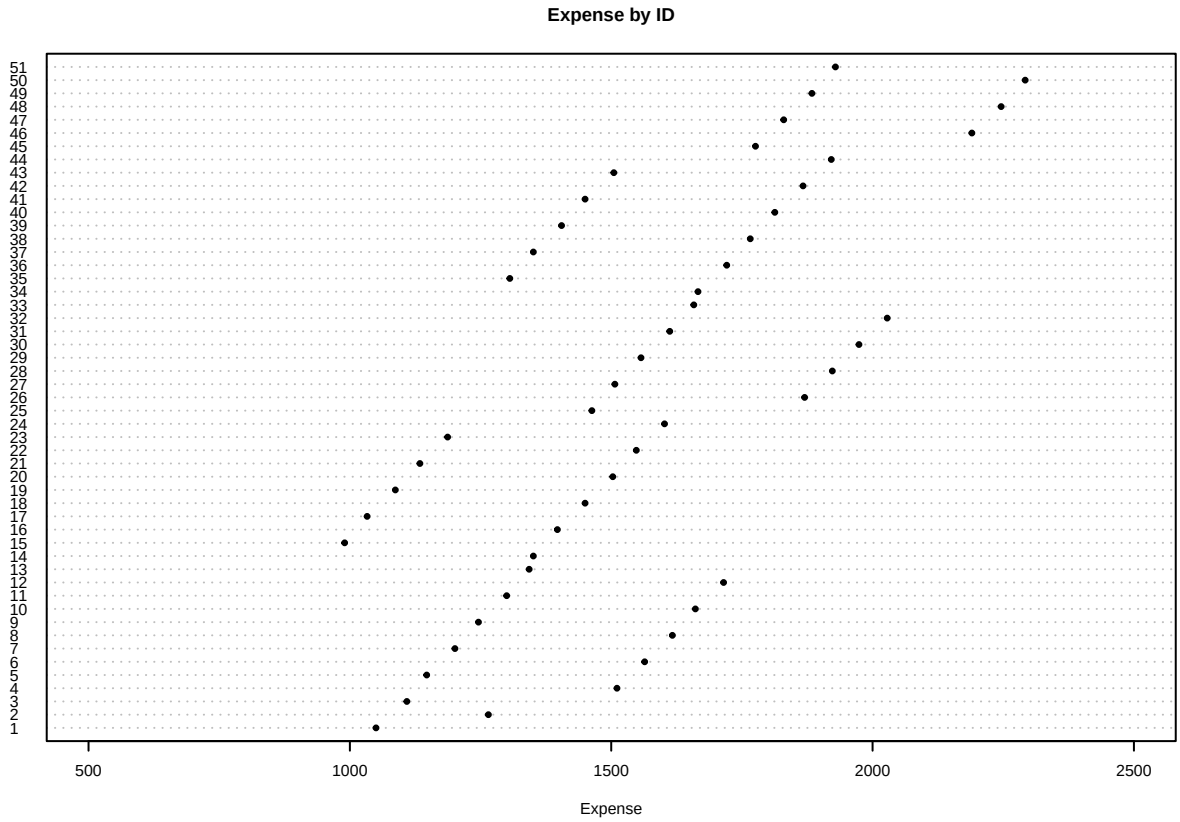

```
qqnorm(df$Expense)
qqline(df$Expense)
```




```
qqplot(df$Age, df$Expense, xlab = "Age", ylab = "Expense") # age vs. expense
```

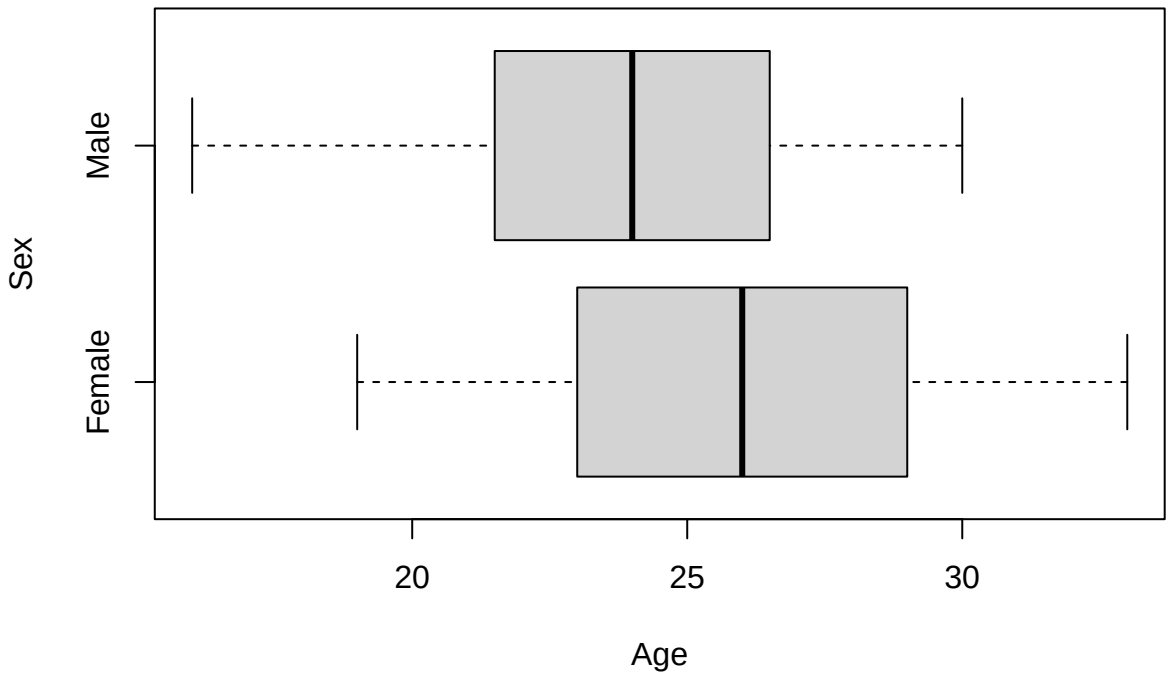



```
dotchart(df$Expense, labels = row.names(df), xlab = "Expense", xlim = c(500,2500),
         pch = 20, cex = 0.5, main = "Expense by ID")
```



(5) splitting the boxplots based on sex

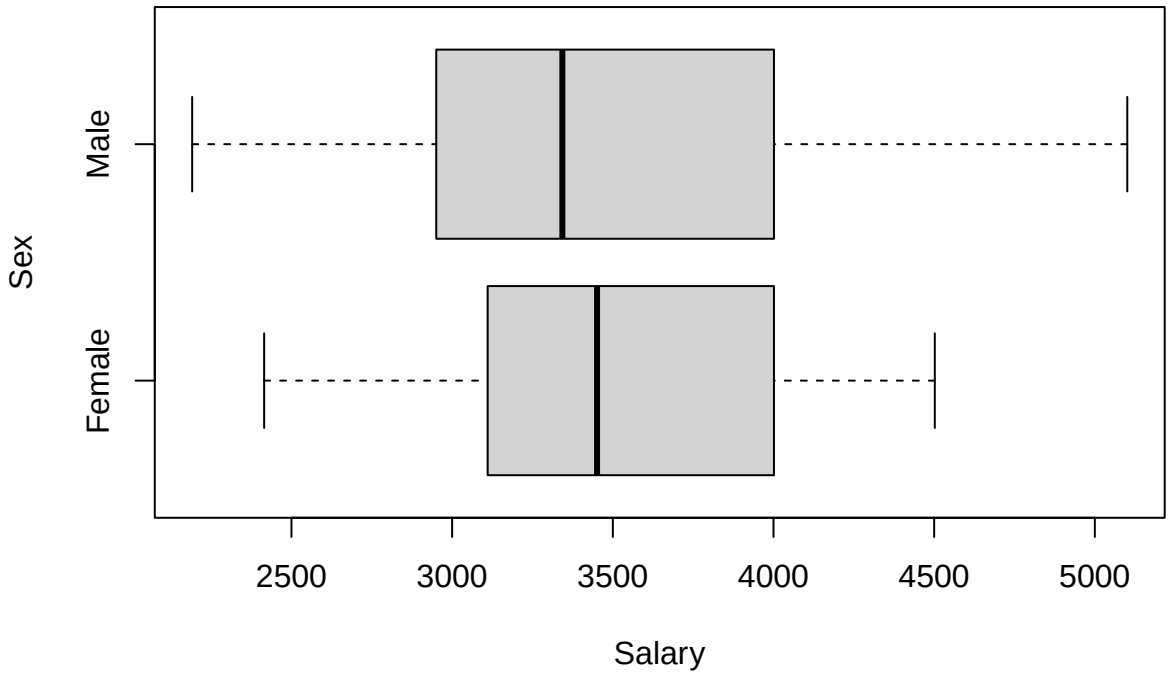
```
boxplot(df$Age ~ df$Sex, horizontal = TRUE, xlab = "Age", ylab = "Sex")
```



From the above boxplot, we can deduce that, from our sample, the female population is older in general. All

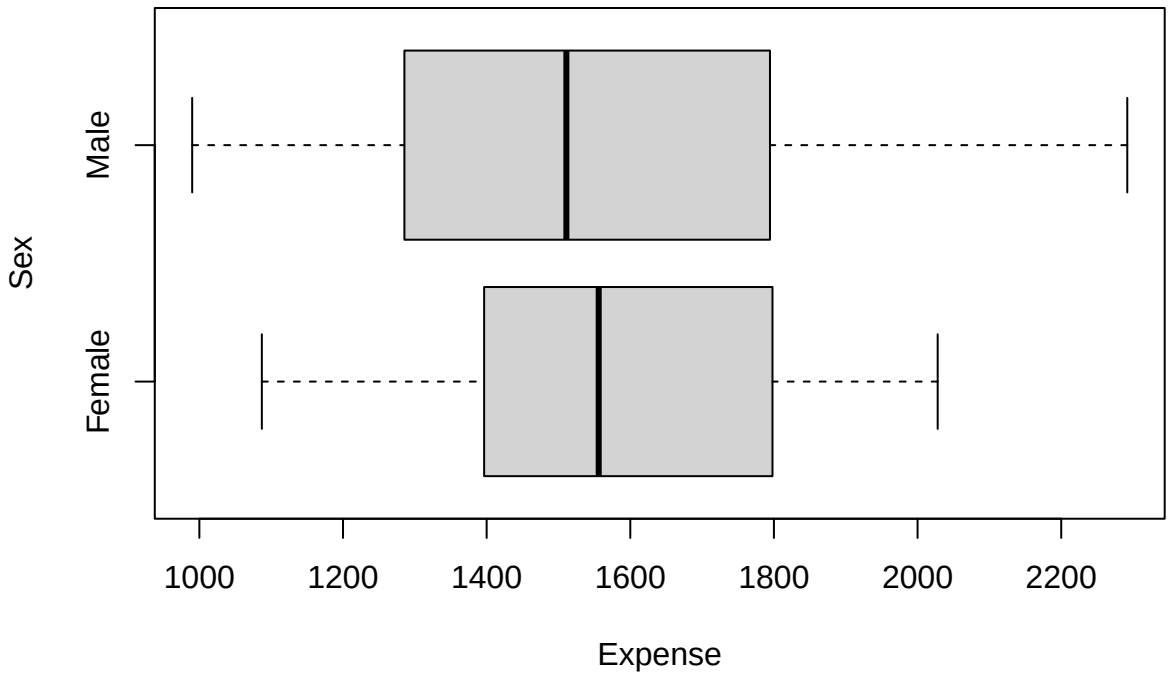
of the summary statistics (min, 1st Q, median, 3rd Q, and max) for the female population lie to the right when compared to the males' summary statistics.

```
boxplot(df$Salary ~ df$Sex, horizontal = TRUE, xlab = "Salary", ylab = "Sex")
```



From the above boxplot, we can deduce that the median salary for the female population is greater than that of the male population. This, in addition to the fact that the median age was greater for the female population as well, perhaps indicates a positive correlation between age and salary. However, the male population shows a wider distribution of salary, containing a far less minimum value and far greater maximum value.

```
boxplot(df$Expense ~ df$Sex, horizontal = TRUE, xlab = "Expense", ylab = "Sex")
```



Again, similar to the salary boxplot above, the median expense for the female population is greater than that of the male's. Also, the same wide distribution can be seen in the male subplot, where it contains both the minimum and the maximum values of the whole sample. Looking at this plot and the salary plot above, this also hints a positive correlation between salary and expense, where the more you earn, the more you are able to spend.