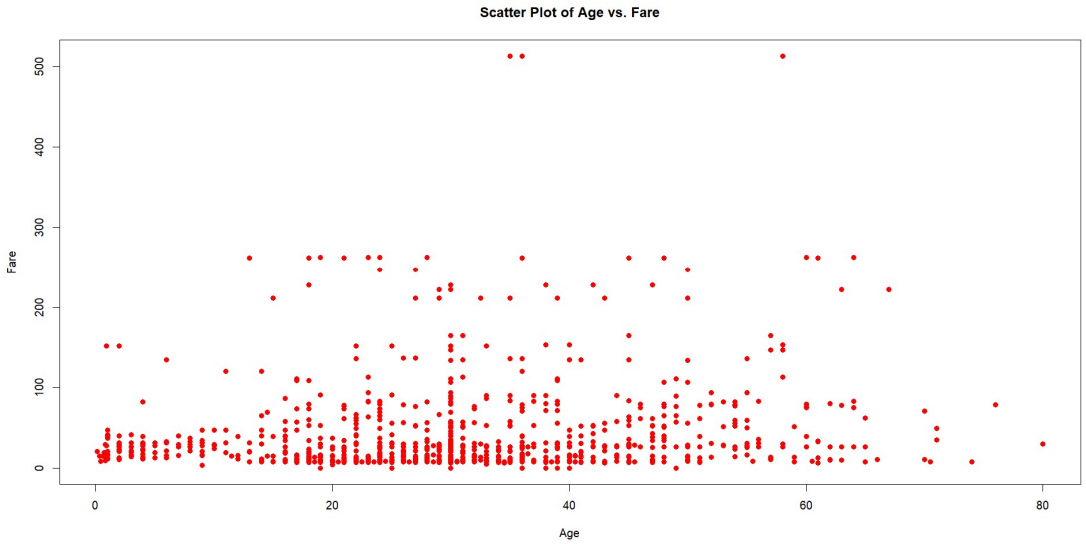
Q1-1. (1 point) Make a scatter plot to investigate the relationship between 'age' and 'fare' in titanic data. You should specify both of (1) R code and (2) scatter plot output (picture) here.

Answer:

```
titanic <- read.csv("titanic.csv", header=TRUE, stringsAsFactors=TRUE)
plot(titanic$age, titanic$fare, xlab = "Age", ylab = "Fare", main = "Scatter Plot of Age vs. Fare", col='red',pch=16)
```



1) No Clear Linear Relationship: The points on the scatter plot are spread out across the entire range of ages and fares. There is no strong linear relationship between age and fare, as there is no clear trend of points forming a straight line.

2) Variability in Fares: For passengers of various ages, fares vary widely. This suggests that other factors, such as ticket class, cabin location, or family size, may influence the fare paid, and age alone is not a significant determinant.

3) Concentration of Lower Fares: There is a concentration of points in the lower fare range, which may indicate that a significant number of passengers, regardless of their age, paid lower fares.

4) Outliers: There are a few outliers where passengers paid much higher fares, which may correspond to premium cabins or special circumstances.

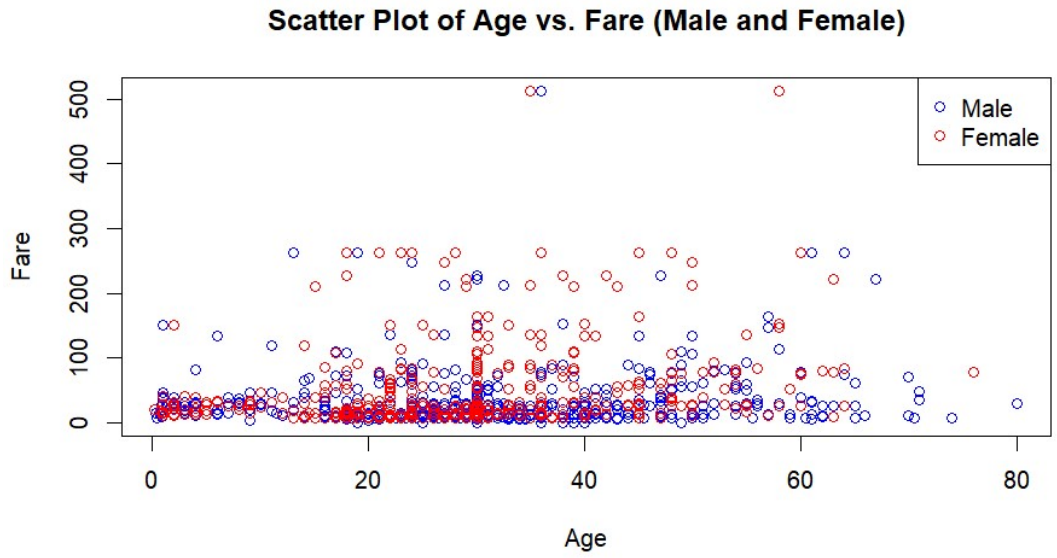
5) Age Distribution: The plot also shows the distribution of ages among passengers. Most passengers appear to be in the younger age range, but there are also passengers of older ages who paid varying fares.

Q1-2. (1 point) Make a scatter plot to investigate the relationship between 'age' and 'fare' in each sex (e.g., male and female). In a scatter plot, distinguish the male and female observations by color. You should specify both of (1) R code and (2) scatter plot output (picture) here.

```
plot(titanic[titanic$sex == "male", "age"], titanic[titanic$sex == "male", "fare"],
    xlab = "Age", ylab = "Fare", main = "Scatter Plot of Age vs. Fare (Male and Female)",
    col = "blue")


points(titanic[titanic$sex == "female", "age"], titanic[titanic$sex == "female", "fare"],
      col = "red")


legend("topright", legend = c("Male", "Female"), col = c("blue", "red"), pch = 1)
```



1) Male passengers (represented in blue) generally have fares concentrated in the lower range, especially for ages below 60.
2) Female passengers (represented in red) have a broader distribution of fare values across different age groups.
3) Both males and females have outliers who paid exceptionally high fares, primarily at older ages.
4) There is some overlap in fare values between males and females, particularly in the middle and lower fare ranges.
5) Female passengers exhibit greater variability in fares compared to males, whose fares are more tightly clustered in the lower range

Q1-3. (1point) What is the correlation coefficient (magnitude) in the relationship between 'age' and 'fare' in the male? You should specify both of (1) R code and (2) the value of correlation coefficient as the out of the code here.

correlation_coefficient_male <- cor(titanic[titanic$sex == "male", "age"], titanic[titanic$sex == "male", "fare"])

correlation_coefficient_male=0.1349

The correlation coefficient (magnitude) in the relationship between 'age' and 'fare' for male passengers is approximately 0.1349. This value indicates a weak positive correlation between age and fare for male passengers, suggesting that there is a slight tendency for fares to increase with age, but the relationship is not strong.

Q1-4. (1point) What is the correlation coefficient (magnitude) in the relationship between 'age' and 'fare' in the female? You should specify both of (1) R code and (2) the value of correlation coefficient as the out of the code here.

correlation_coefficient_female <- cor(titanic[titanic$sex == "female", "age"], titanic[titanic$sex == "female", "fare"])

correlation_coefficient_female= 0.2473

The correlation coefficient (magnitude) in the relationship between 'age' and 'fare' for female passengers is approximately 0.2473 This value indicates a weak positive correlation between age and fare for female passengers, suggesting that there is a slight tendency for fares to increase with age, but the relationship is not strong.

Q2. Decision Tree Classification Model (R Practice)

Q2-1. (1 point) Split the titanic data into training and testing data sets. Please type 'set.seed(345)' at console display first. And then, you should randomly select (1) 2/3 of the rows as training data and (2) the remaining 1/3 of the rows as testing data. Specify the name of the training data and testing data as 'titanic.train' and 'titanic.test', respectively. You should provide R code here.

```
set.seed(345)

train = sample(1:nrow(titanic), nrow(titanic)*(2/3))

titanic.train=titanic[train,]
titanic.test = titanic[-train,]
```

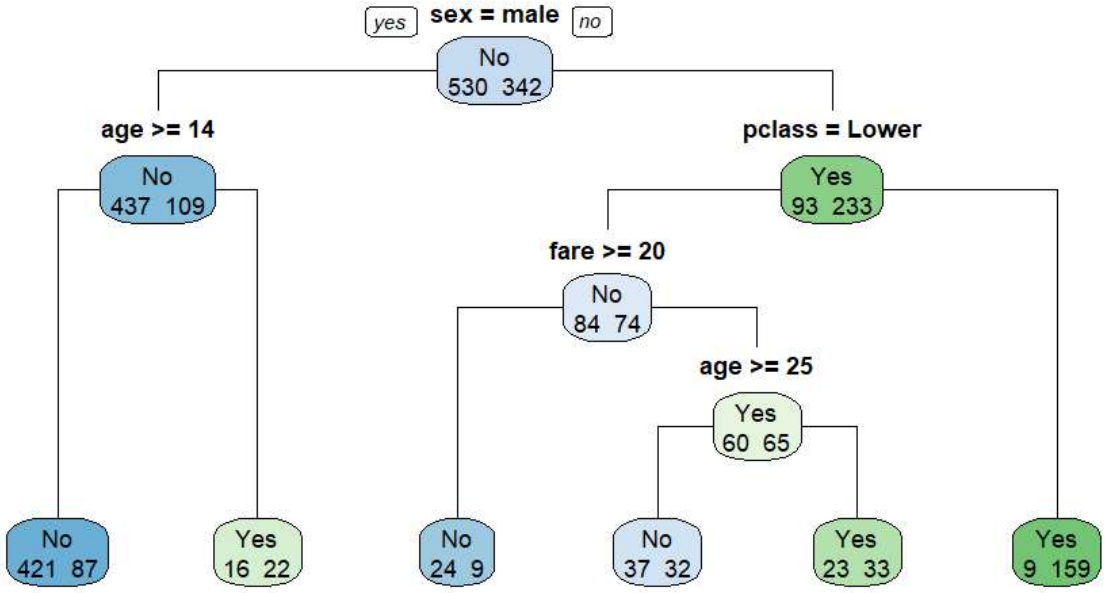
Q2-2. (1 point) Make a decision tree by using 'rpart' function. Set the dependent variable as 'survived' and the independent variables as remaining all other variables. Use 'titanic.train' as training data. Specify 'xval' value as 0 and 'minsplit' value as 100. Use 'gini' measure and name the decision tree as 'fit'. You should provide R code here.

```
install.packages("rpart")
 library(rpart)
fit = rpart(survived ~ .,
        data=titanic.train,
        method="class",
        control=rpart.control(xval=0, minsplit=100),

        parms=list(split="gini"))
```

Q2-3. (1point) Plot a tree using 'rpart.plot' package. In rpart.plot function, specify the type value as 1 and extra value as 1. You should specify both of (1) R code and (2) decision tree output (picture) here

```
install.packages("rpart.plot")
library(rpart.plot)
rpart.plot(fit, type = 1, extra = 1)
```



Q2-4. (1point) Extract the vector of 'predicted (dependent variable) class' and 'actual (dependent variable) class' of each observation in 'titanic.train' data. And then, build the confusion matrix using 'table' function. You should specify both of (1) R code and (2) confusion matrix (picture) here.

```
titanic.pred <- predict(fit, titanic.train, type="class")

titanic.actual <- titanic.train$survived

confusion.matrix <- table(titanic.pred, titanic.actual)
confusion.matrix
```

```
titanic.pred  No Yes
         No  482 128
         Yes  48 214
```

```
>
```