Exploring and visualizing data

After imputing the missing values, one should perform an exploratory analysis, which involves using a visualization plot and an aggregation method to summarize the data characteristics. The result helps the user gain a better understanding of the data in use. The following recipe will introduce how to use basic plotting techniques with a view to help the user with exploratory analysis.

Getting ready

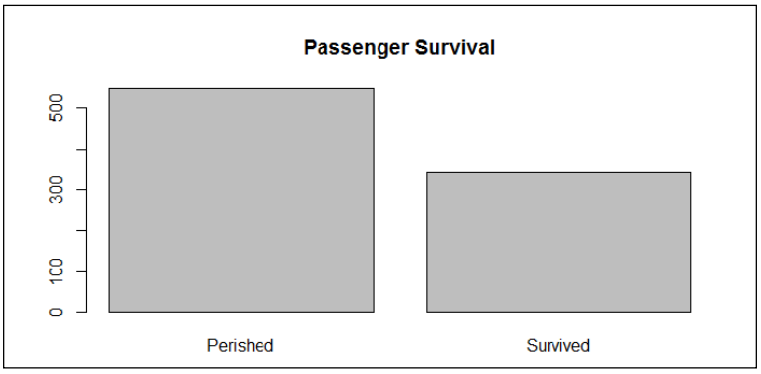
This recipe needs the previous recipe to be completed by imputing the missing value in the `age` and `Embarked` attribute.

How to do it...

Perform the following steps to explore and visualize data:

1. First, you can use a bar plot and histogram to generate descriptive statistics for each attribute, starting with passenger survival:

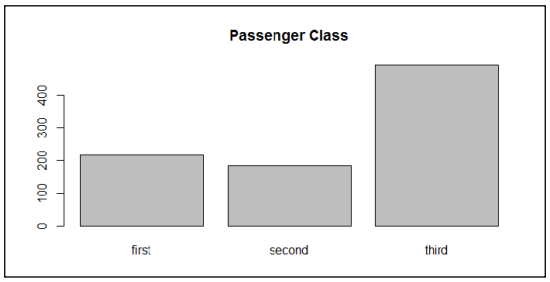
```
> barplot(table(train.data$Survived), main="Passenger Survival",
names= c("Perished", "Survived"))
```



Passenger survival

2. We can generate the bar plot of passenger class:

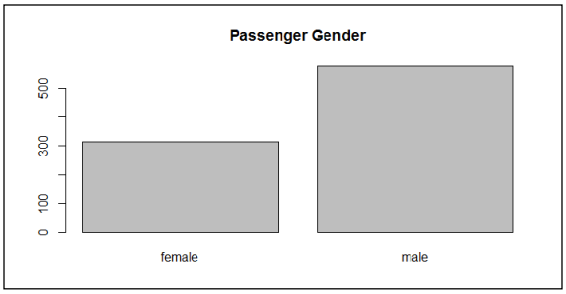
```
> barplot(table(train.data$Pclass), main="Passenger Class",
names= c("first", "second", "third"))
```



Passenger class

3. Next, we outline the gender data with the bar plot:

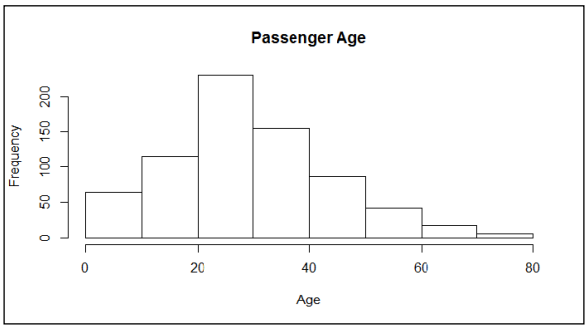
```
> barplot(table(train.data$Sex), main="Passenger Gender")
```



Passenger gender

4. We then plot the histogram of the different ages with the `hist` function:

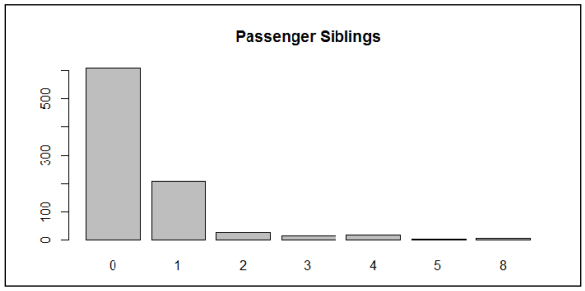
```
> hist(train.data$Age, main="Passenger Age", xlab = "Age")
```



Passenger age

5. We can plot the bar plot of sibling passengers to get the following:

```
> barplot(table(train.data$SibSp), main="Passenger Siblings")
```



Passenger siblings

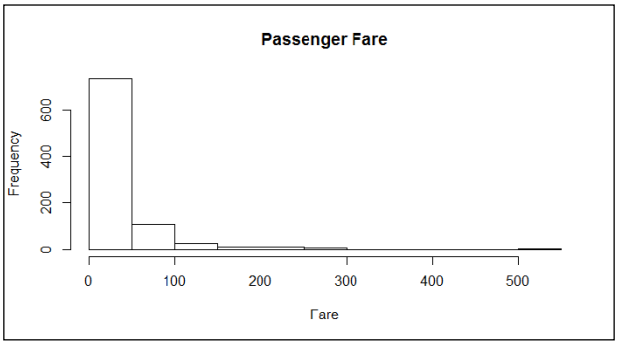
6. Next, we can get the distribution of the passenger parch:

```
> barplot(table(train.data$Parch), main="Passenger Parch")
```

Passenger parch

7. Next, we plot the histogram of the passenger fares:

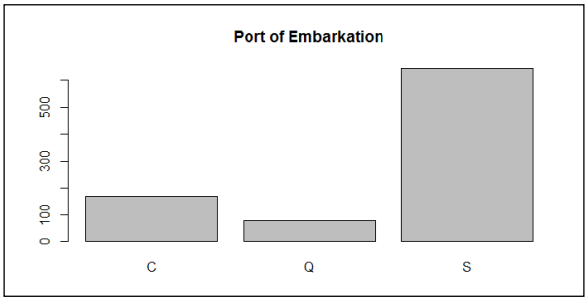
```
> hist(train.data$Fare, main="Passenger Fare", xlab = "Fare")
```



Passenger fares

8. Finally, one can look at the port of embarkation:

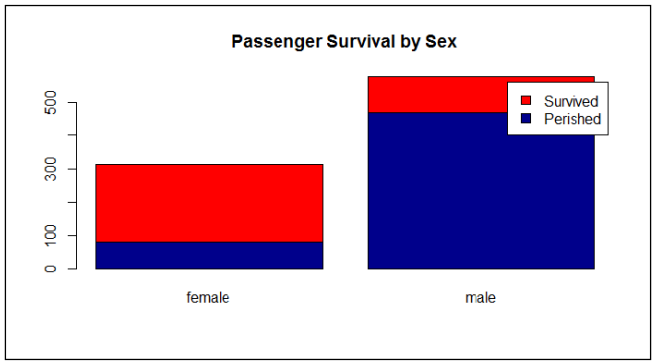
```
> barplot(table(train.data$Embarked), main="Port of Embarkation")
```



Port of embarkation

9. Use `barplot` to find out which gender is more likely to perish during shipwrecks:

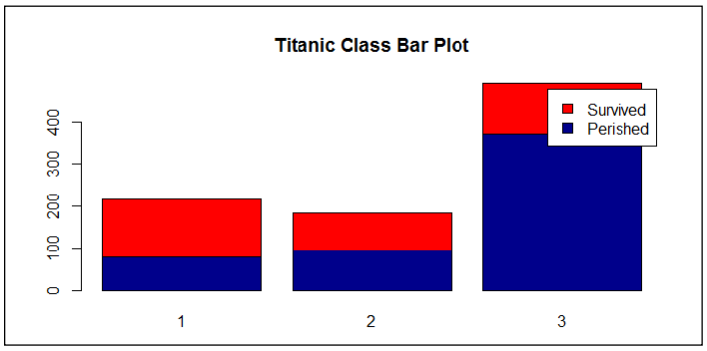
```
> counts = table( train.data$Survived, train.data$Sex)
> barplot(counts,  col=c("darkblue","red"), legend = c("Perished",
"Survived"), main = "Passenger Survival by Sex")
```



Passenger survival by sex

10. Next, we should examine whether the `Pclass` factor of each passenger may affect the survival rate:

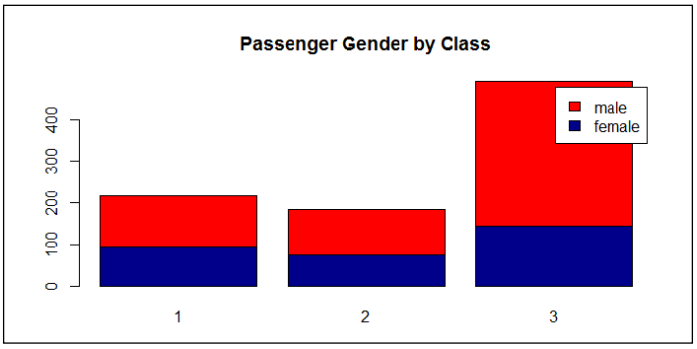
```
> counts = table( train.data$Survived, train.data$Pclass)
> barplot(counts,  col=c("darkblue","red"), legend =c("Perished",
"Survived"), main= "Titanic Class Bar Plot" )
```



Passenger survival by class

11. Next, we examine the gender composition of each `Pclass`:

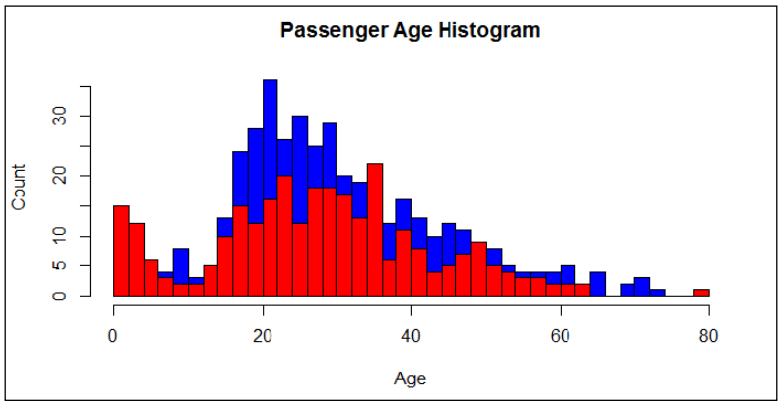
```
> counts = table( train.data$Sex, train.data$Pclass)
> barplot(counts,  col=c("darkblue","red"), legend =
rownames(counts), main= "Passenger Gender by Class")
```



Passenger gender by class

12. Furthermore, we examine the histogram of passenger ages:

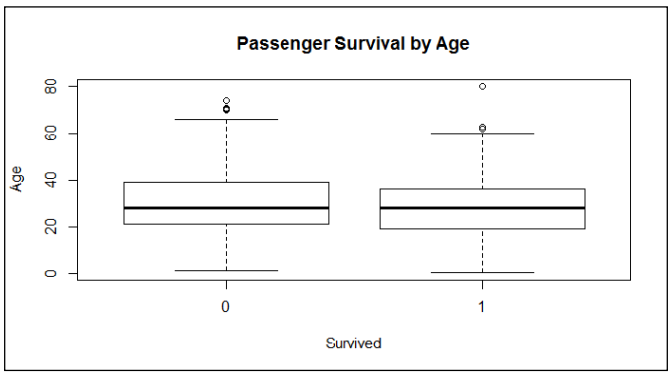
```
> hist(train.data$Age[which(train.data$Survived == "0")], main=
"Passenger Age Histogram", xlab="Age", ylab="Count", col ="blue",
breaks=seq(0,80,by=2))
> hist(train.data$Age[which(train.data$Survived == "1")], col
="red", add = T, breaks=seq(0,80,by=2))
```



Passenger age histogram

13. To examine more details about the relationship between the age and survival rate, one can use a `boxplot`:

```
> boxplot(train.data$Age ~ train.data$Survived,
+         main="Passenger Survival by Age",
+         xlab="Survived", ylab="Age")
```



Passenger survival by age