Feedback — Weekly Quiz 3

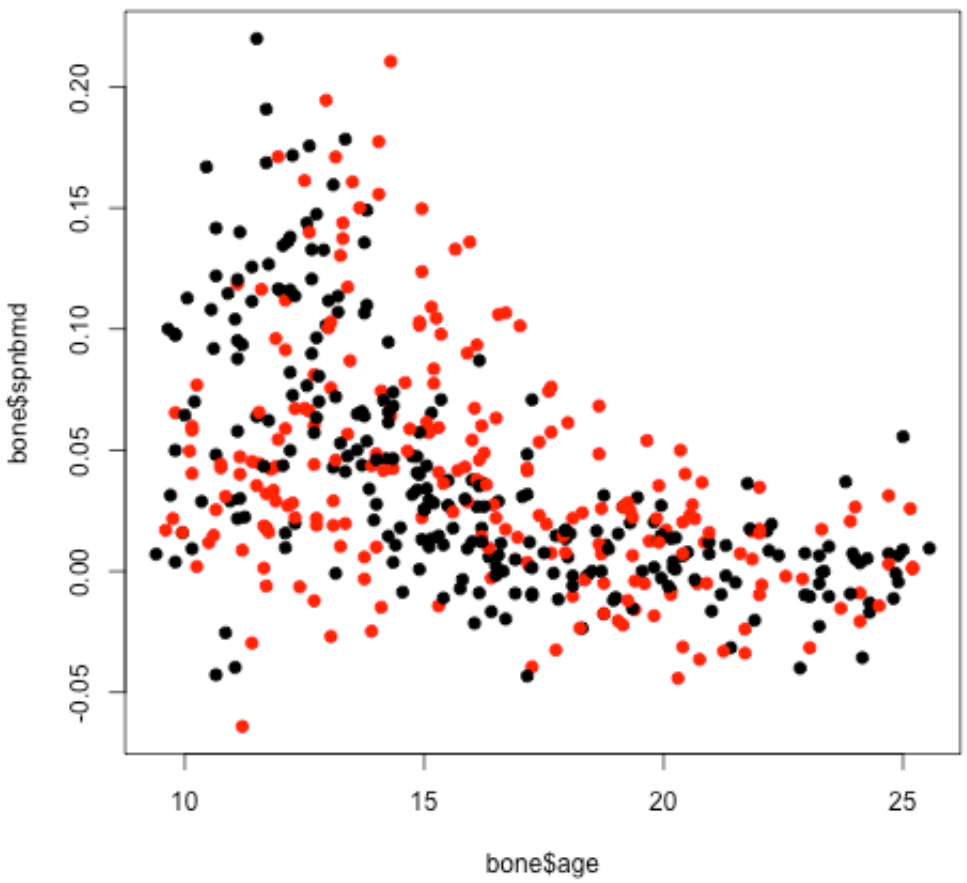
You submitted this quiz on **Wed 6 Feb 2013 8:05 PM PST**. You got a score of **9.00** out of **9.00**.

Question 1

Below is a plot of bone density versus age. It was created using the following code in R:

```
library(ElemStatLearn)
data(bone)
plot(bone$age,bone$spnbmd,pch=19,col=((bone$gender=="male")+1))
```

Males are shown in black and females in red. What are the characteristics that make this an exploratory graph? Check all correct options

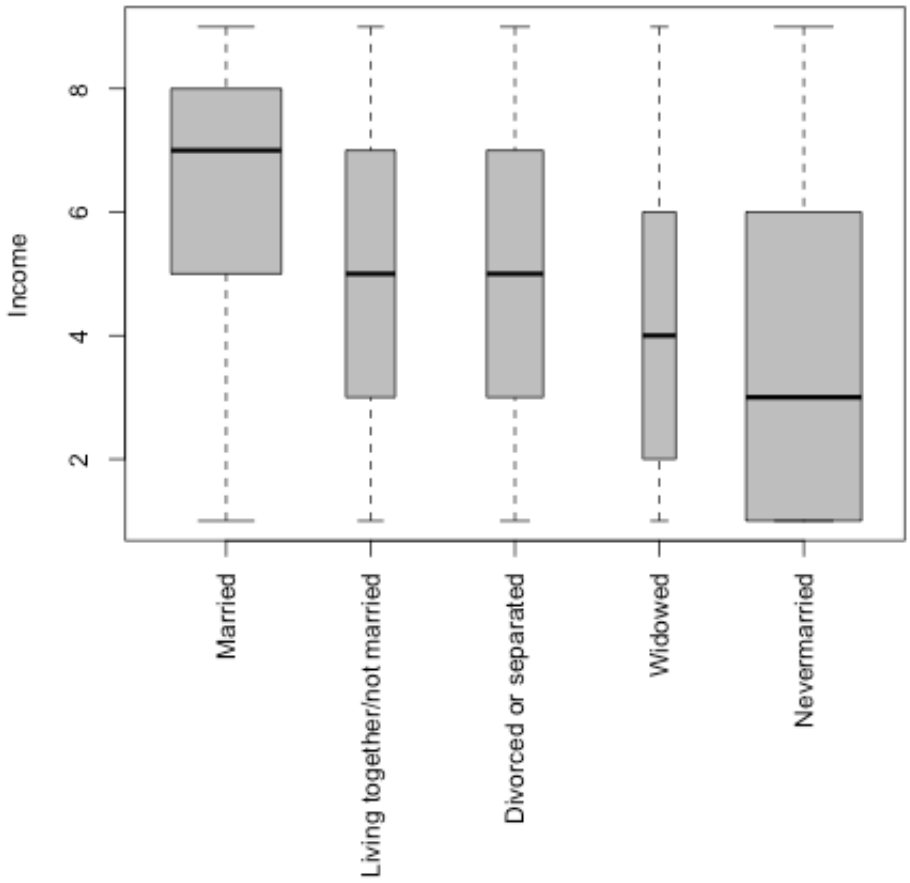


Your Answer		Score	Explanation
☑ There plot does not have a legend.	✔	1.00	
☐ The plot has multiple colors.	✘	0.00	
☑ The axis labels are R variables	✔	1.00	
☐ The plot uses color to make the figure "pretty"	✘	0.00	
Total		2.00 / 2.00	

Question 2

Below is a boxplot of yearly income by marital status for individuals in the United States. It was created using the following code in R:

```
library(ElemStatLearn)
data(marketing)
plot(bone$age,bone$spnbmd,pch=19,col=((bone$gender=="male")+1))
boxplot(marketing$Income ~ marketing$Marital,col="grey",xaxt="n",ylab="
Income",xlab="")
axis(side=1,at=1:5,labels=c("Married","Living together/not married","Di
vorced or separated","Widowed","Nevermarried"),las=2)
```



Which of the following can you conclude from the plot? (Check all that apply)

Your Answer	Score	Explanation

☑ The median income for individuals who are divorced is higher than the median for individuals who are widowed.	✔	0.50
☐ There are more individuals who are widowed than divorced in this data set.	✘	0.00
☑ There are more individuals who were never married than divorced in this data set.	✔	0.50
☐ The medians for all individuals who are not currently married are almost the same.	✘	0.00
Total		1.00 / 1.00

Question 3

Load the iris data into R using the following commands:

```
library(datasets)
data(iris)
```

Subset the iris data to the first four columns and call this matrix irisSubset. Apply hierarchical clustering to the irisSubset data frame to cluster the rows. If I cut the dendrogram at a height of 3 how many clusters result?

Your Answer		Score	Explanation
◉ 4 clusters	✔	2.00	
Total		2.00 / 2.00	

Question 4

Load the following data set into R using either the .rda or .csv file:

https://spark-public.s3.amazonaws.com/dataanalysis/quiz3question4.rda
https://spark-public.s3.amazonaws.com/dataanalysis/quiz3question4.csv

Make a scatterplot of the x versus y values. How many clusters do you observe? Perform k-means clustering using your estimate as to the number of clusters. Color the scatterplot of the x, y values by what cluster they appear in. Is there anything wrong with the resulting cluster estimates?

Your Answer		Score	Explanation
◉ There are two obvious clusters. The k-means algorithm does not assign all of the points to the correct clusters because the clusters wrap around each other.	✔	2.00	
Total		2.00 / 2.00	

Question 5

Load the hand-written digits data using the following commands:

```
library(ElemStatLearn)
data(zip.train)
```

Each row of the zip.train data set corresponds to a hand written digit. The first column of the zip.train data is the actual digit. The next 256 columns are the

intensity values for an image of the digit. To visualize the adigit we can use the zip2image() function to convert a row into a 16 x 16 matrix:

```
# Create an image matrix for the 3rd row, which is a 4
im = zip2image(zip.train,3)
image(im)
```

Using the zip2image file, create an image matrix for the 8th and 18th rows. For each image matrix calculate the svd of the matrix (with no scaling). What is the percent variance explained by the first singular vector for the image from the 8th row? What is the percent variance explained for the image from the 18th row? Why is the percent variance lower for the image from the 18th row?

Your Answer		Score	Explanation
◉ The first singular vector explains 98% of the variance for row 8 and 48% for row 18. The reason the first singular vector explains less variance for the 18th row is that the image is more complicated, so there are multiple patterns each explaining a large percentage of variance.	✔	2.00	
Total		2.00 / 2.00	