Weekly Quiz 3

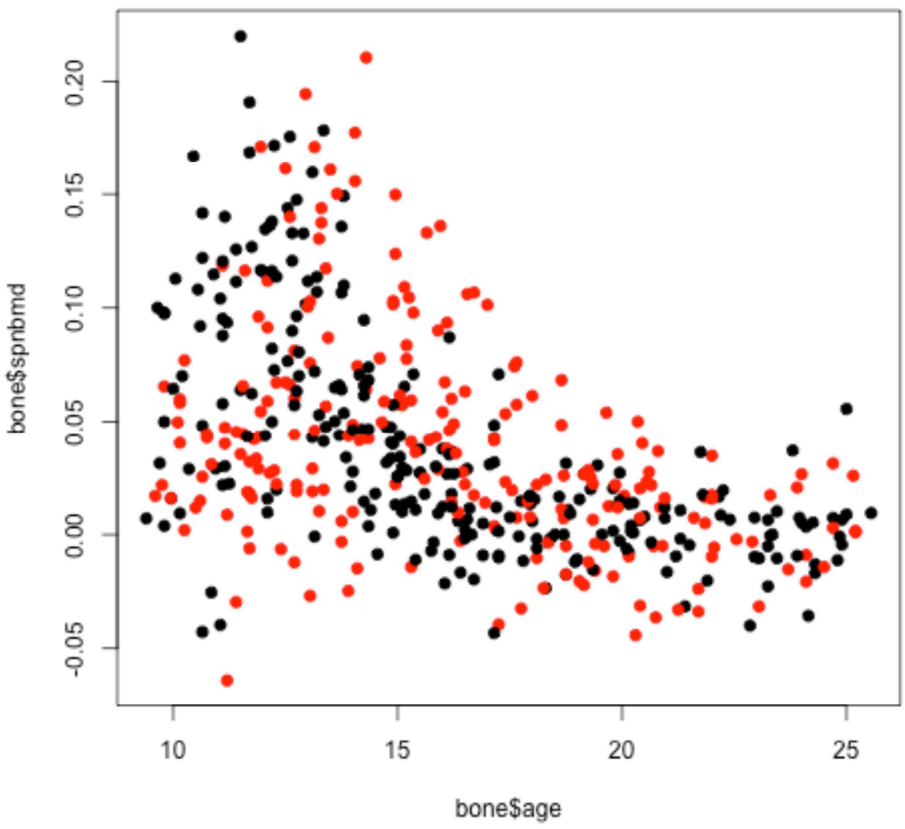
The **due date** for this quiz is **Sun 10 Feb 2013 8:30 PM PST**.

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



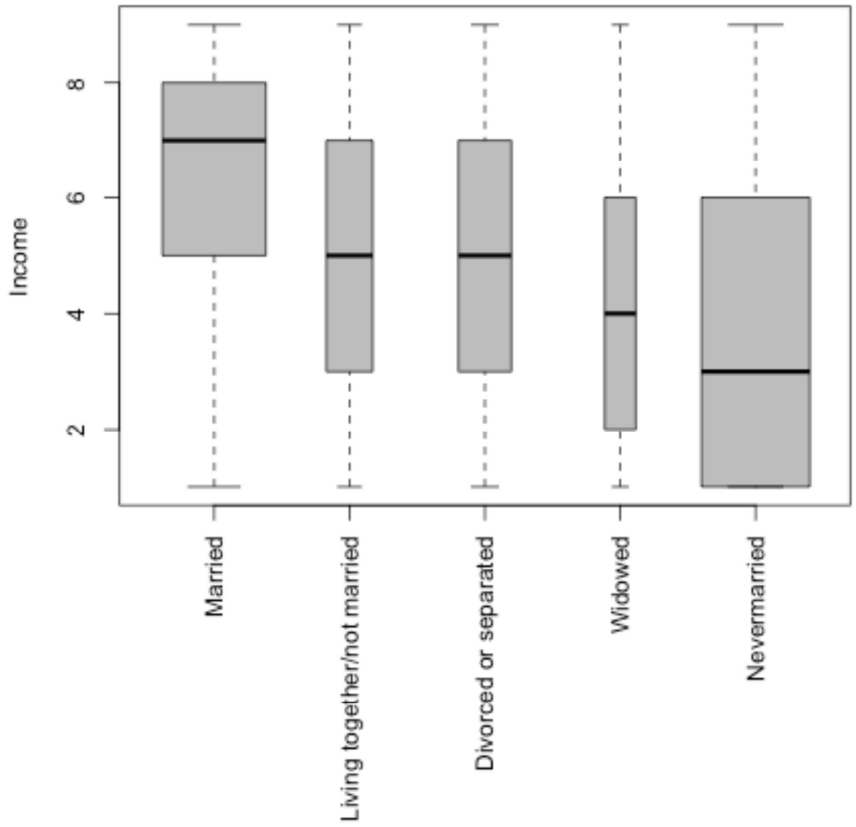
- [ ] The plot is made in R.

- [x] There plot does not have a legend.
- [ ] The plot uses color to make the figure "pretty"
- [x] The axis labels are R variables

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 marrie
d","Divorced or separated","Widowed","Nevermarried"),las=2)
```



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

- [x] The 25h percentile of the income for married individuals is almost the same as the median for individuals living together but not married.
- [ ] The income values are measured in dollars.
- [x] There are more individuals who were never married than divorced in this data set.
- [ ] The medians for all individuals who are not currently married are almost the same.

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?

- No clusters
- 1 cluster
- 4 clusters
- 5 clusters

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?

- There are two obvious clusters. The k-means algorithm does not converge to a solution because the clusters wrap around each other.
- 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.
- There are two obvious clusters.Despite the wrapped clusters, the k-means algorithm converges to the correct clusters when you use multiple starts of the algorithm.

There is one obvious cluster. The k-means algorithm does not work because there is only one cluster in the data.

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?

The first singular vector explains 99% of the variance for row 8 and 44% 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 larger percentage of variance.

The first singular vector explains 99% of the variance for row 8 and 44% for row 18. The reason the singular vector explains less variance for the 18th row is because the 8th row has higher average values.

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.

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 because the 8th row has higher average values.

In accordance with the Honor Code, I certify that my answers here are my own work.

Submit Answers Save Answers