Stat501_Homework4

Kelby Kies

3/13/2021

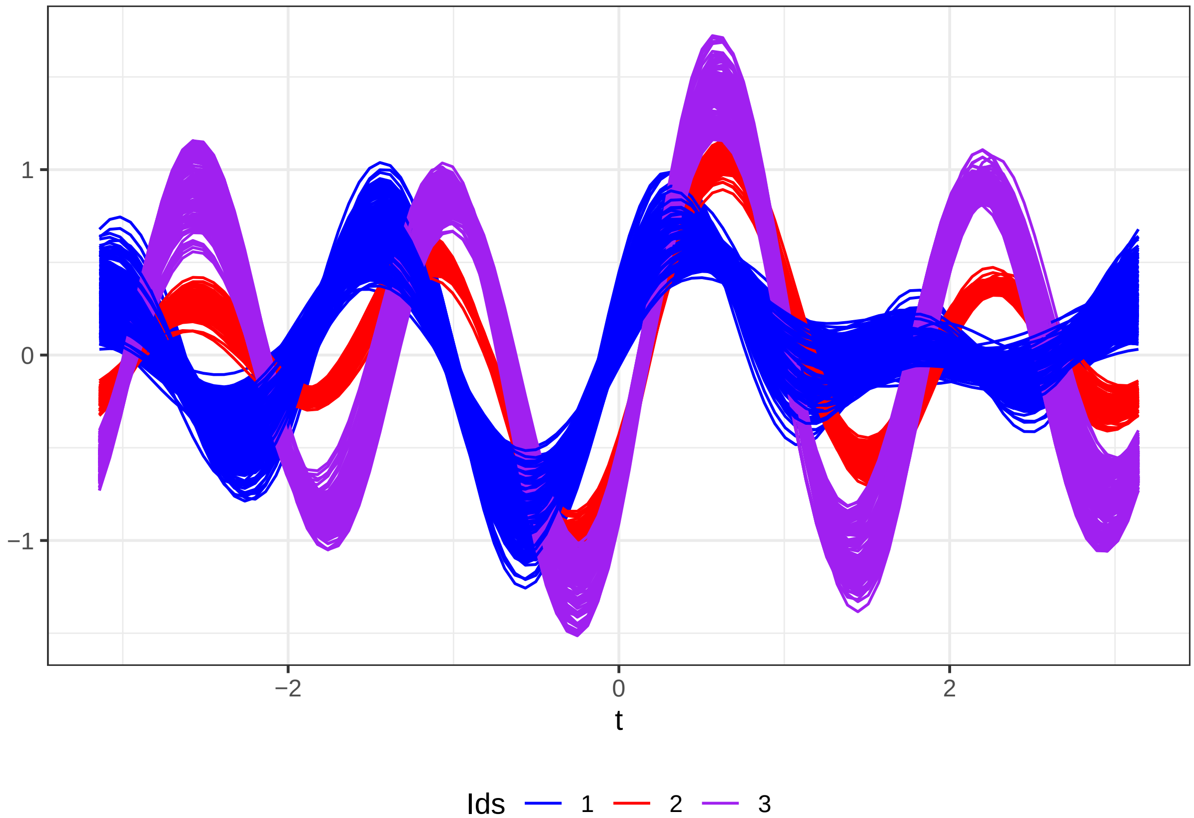
Question1:

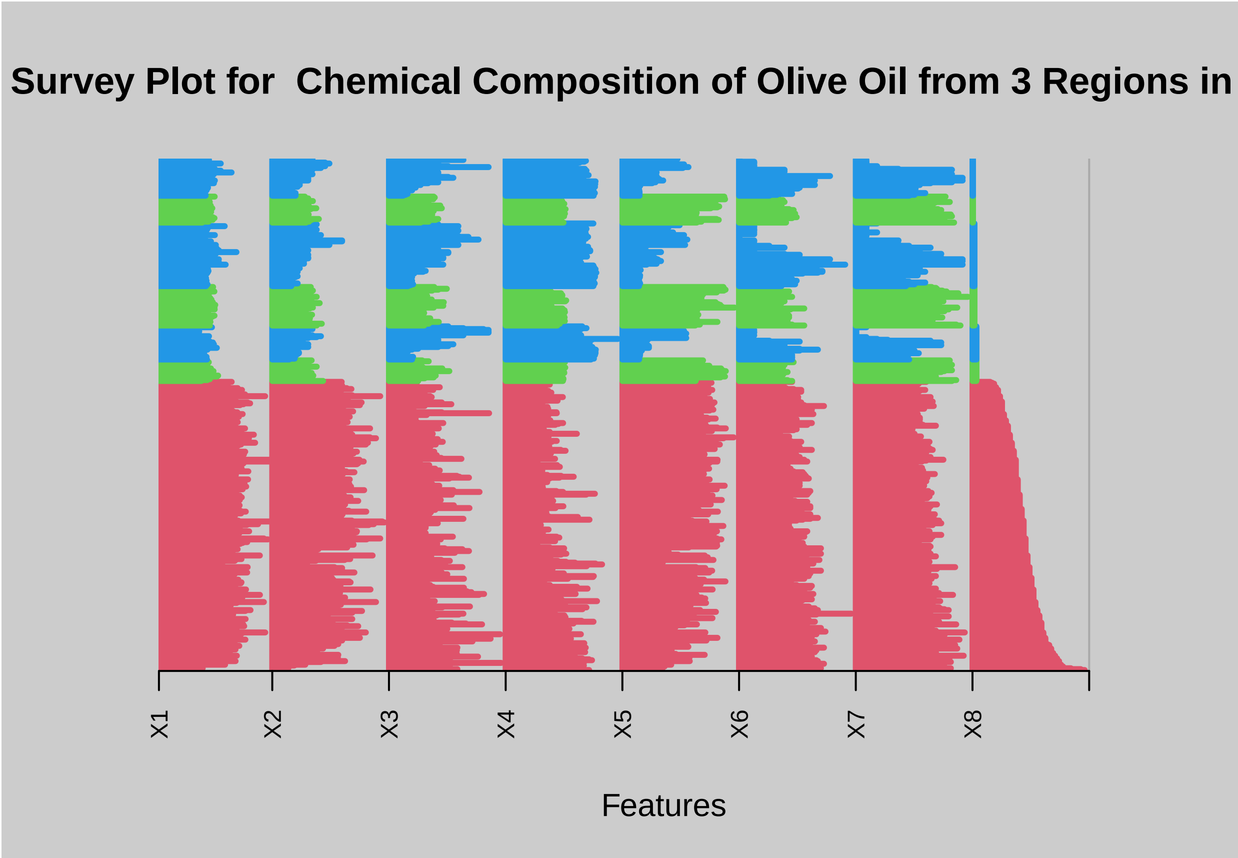
areas A1–A4 belong to region R1, areas A5 and A6 belong to region R2, and areas A7–A9 belong to region R3. ## Part A: Provide appropriate graphical summaries for the dataset. You may plot all nine regions (using color) in one figure, and also the individual sub-regions separately to better understand the dataset. Comment on the distinctiveness between the chemical composition of the olive oils in the three main regions, as well as individually between the sub-regions. [1 page; 20 points]

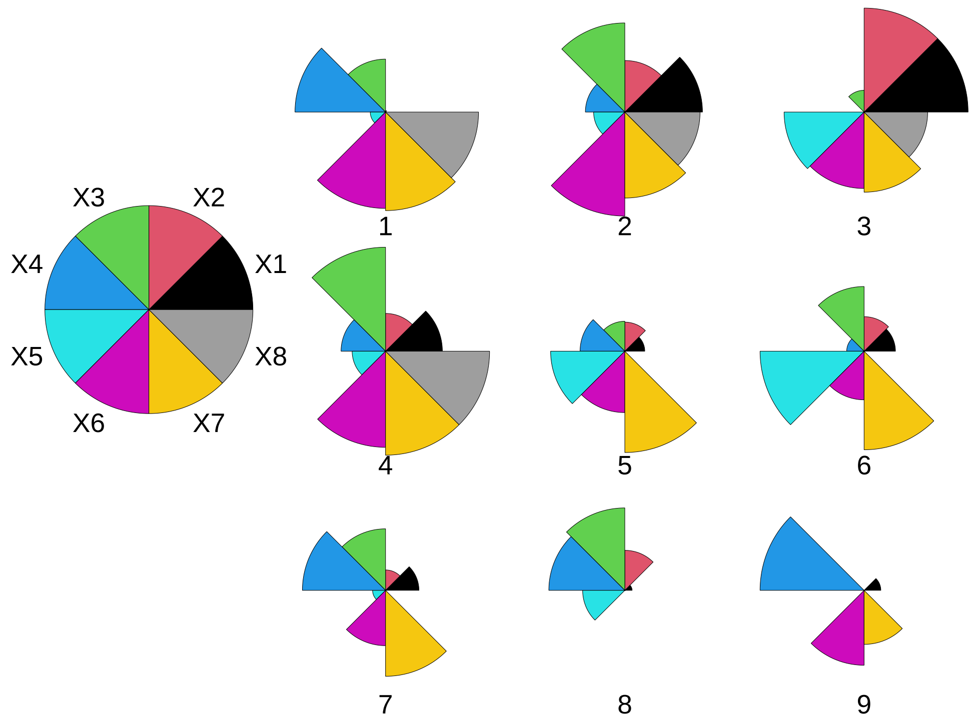
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```


1
0
−1
−2
0
2
t
Ids
1
2
3


Survey Plot for Chemical Composition of Olive Oil from 3 Regions in I
X1
X2
X3
X4
X5
X6
X7
X8
Features



Comment on the distinctiveness between the chemical composition of the olive oils in the three main regions, as well as individually between the sub-regions. [1 page; 20 points]

For the first part of this question, I used a survey plot to show the distinctiveness between the chemical composition between the 3 main regions. The separation is clearested when I sorted my values by the 8th chemical. When looking at the individual sub-regions, I used a stars plot to display the means of each chemical within the region. I thought that this plot showed the variety of chemical composition between the 9 sub regions. The main similarity between most of the sub regions is that they have a good amount of chemical X6 and X7 except for sub-region 8.

Part B:

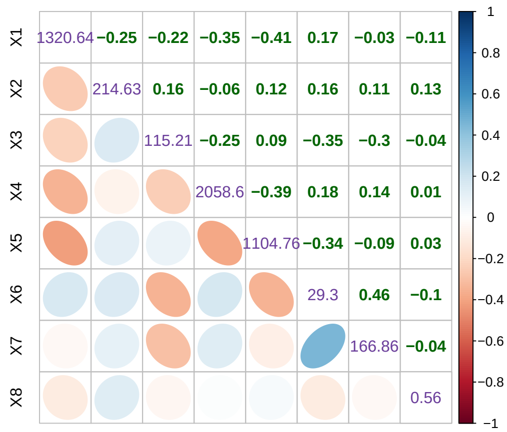
We will now focus on Region R2. Answer the following questions.

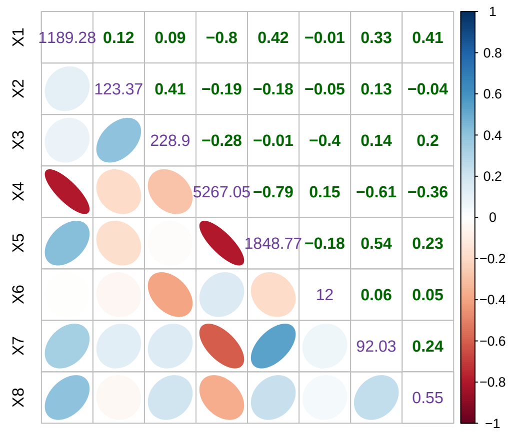
Part i:

Calculate the correlation matrix for the chemical composition in each of the two sub-regions. Display the correlation matrices for the two sub-regions side-by-side, and comment on possible differences. [1 page; 5 points]

```
## Loading required package: corrplot
```

```
## corrplot 0.84 loaded
```





Comment on possible differences: There are some differences found when comparing the sub-regions within Region 2. For region 5, the plot on the left, we can see that there is a lot of negatively correlated chemicals. Meaning that as 1 chemical is being added the composition of the corresponding chemical is going down. For region 6, the plot on the right, we can see that there is a lot of positively correlated chemicals. One big difference between the two regions is the correlation between Chemical 4 (X4) and Chemical 1(X1). In sub region 5 we can see that these two values have a moderate negative correlation, while in sub region 6 this negative correlation is very strong. Similarly, we see that Chemical 5(X5) correlated with Chemcial 4(X4) have a stronger negative correlation in sub region 6. For Chemical 7 (X7) and Chemical 4(X4) Correlation, sub region 6 has a very strong negative correaltion between the chemicals, while in sub-region 5 there is a slight positive correlation.

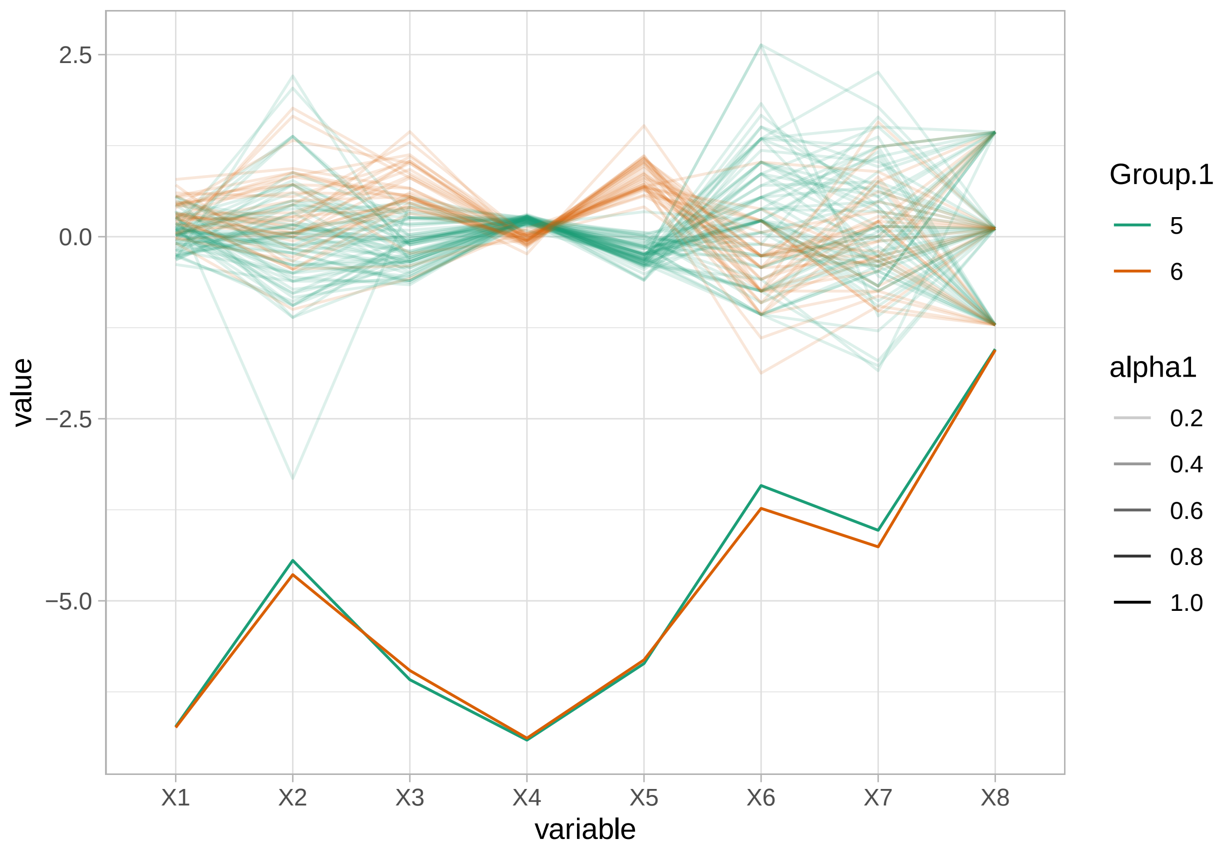
Part ii:

Also, compare the marginal standard deviations of the composition of the eight chemicals for the two regions. Once again, you may use parallel coordinate plots if it helps to easily compare the standard deviations for the two groups. [10 points]

```
## Loading required package: GGally

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

## Loading required package: RColorBrewer
```



Compare the marginal standard deviations of the composition of the eight chemicals for the two regions: We can see that the standard deviation each sub region follows the same general trend for all of the chemicals in both regions. This might be harder to see if we don't look at the standard deviation lines bolded for us. If we just look at the parallel coordinate plot along then we might conclude something otherwise.

Part iii:

Test for differences in dispersions among the two groups.[10points]

```
## [1] 2
## ------------------------------------------------
##  MBox Chi-sqr. df P
## ------------------------------------------------
##   109.5204    98.3550          36       0.0000
## ------------------------------------------------
## Covariance matrices are significantly different.

## $MBox
##        5
## 109.5204
##
## $ChiSq
##        5
## 98.35502
##
## $df
```

```
## [1] 36
##
## $pValue
##            5
## 1.069438e-07
```

Part iv:

For each of the two groups,is it reasonable to assume multivariate normality? [10points]

```
## [1] " Cramer Test Statistic"

## [1] " Sub-Region 5"

## Loading required package: parallel

## [1] 1.764453e-05

## [1] "Sub-Region 6"

## [1] 0.005056039
```

If the q-value returned is less than 0.05, then normality is not good and is rejected. For Sub Region 5, the q-value, 1.474649e-05 is $<$ 0.05 and normality is rejected. For Sub Region 6, the q-value, 0.3810251, is $>$ 0.05 and normality can be assumed.

Part v:

Regardless of how the tests in 1(b)iii and 1(b)iv above turnout,test for equality of the mean chemical compositions between the two groups. Specifically, report the Hotellings' T2-statistics and its p- value. [10 points]

```r
library(ICSNP)
```

```
## Loading required package: mvtnorm

## Loading required package: ICS
```

```r
y <- HotellingsT2(R2_subreg5, R2_subreg6)
# T.2 is not the Tstat. We have to convert.
print("We need to convert the T.2 statistic, because it represents the F test.")
```

```
## [1] "We need to convert the T.2 statistic, because it represents the F test."
```

```r
#n-1 * p /n - p
#n = # of observations; we have 98 observations
#p = 8
corrected_value <- y$statistic * 97 * 8 /90
print(paste0("The Corrected T2 statistic is ", corrected_value))
```

```
## [1] "The Corrected T2 statistic is 969.186841012365"
```

```r
print(paste0("The P-value is ", y$p.value))
```

```
## [1] "The P-value is 0"
```

Part vi:

Provide individual pairwise t-tests for the differences in the composition of the eight chemicals among the two groups, using Bonferroni and also after controlling false discovery rates, at the 5% level of significance. [5 points]

```r
#p.adj = "bonferroni")
x1 <- pairwise.t.test(R2$X1, R2$group.id, p.adj = "bonferroni")
x2 <- pairwise.t.test(R2$X2, R2$group.id, p.adj = "bonferroni")
x3 <- pairwise.t.test(R2$X3, R2$group.id, p.adj = "bonferroni")
x4 <- pairwise.t.test(R2$X4, R2$group.id, p.adj = "bonferroni")
x5 <- pairwise.t.test(R2$X5, R2$group.id, p.adj = "bonferroni")
x6 <- pairwise.t.test(R2$X6, R2$group.id, p.adj = "bonferroni")
x7 <- pairwise.t.test(R2$X7, R2$group.id, p.adj = "bonferroni")
x8 <- pairwise.t.test(R2$X8, R2$group.id, p.adj = "bonferroni")
chemical <- c('X1','X2','X3','X4','X5','X6','X7','X8')
p_values <- c( x1$p.value, x2$p.value, x3$p.value, x4$p.value, x5$p.value, x6$p.value, x7$p.value, x8$p

padj_pvalues <- p.adjust(p_values, method = "bonferroni", n = length(p_values))
rbind(chemical, padj_pvalues)
```

```
##              [,1]                   [,2]                [,3]
## chemical     "X1"                   "X2"                "X3"
## padj_pvalues "5.97332608884582e-06" "0.217621963432829" "9.17489048953503e-16"
##              [,4]                   [,5]
## chemical     "X4"                   "X5"
## padj_pvalues "6.63150308390845e-40" "1.02309903473137e-45"
##              [,6]                  [,7] [,8]
## chemical     "X6"                  "X7" "X8"
## padj_pvalues "3.8344399664852e-05" "1"  "1"
```

Above I have performed a pairwise t-test for the differences in the composition of the 8 chemicals using the Bonferroni method and also corrected the resulting p-values with the p.adjust() method in R. The table above are my final p-values for all of the pairwise t-tests between sub-region 5 and sub-region 6 for each chemical.

Part vii:

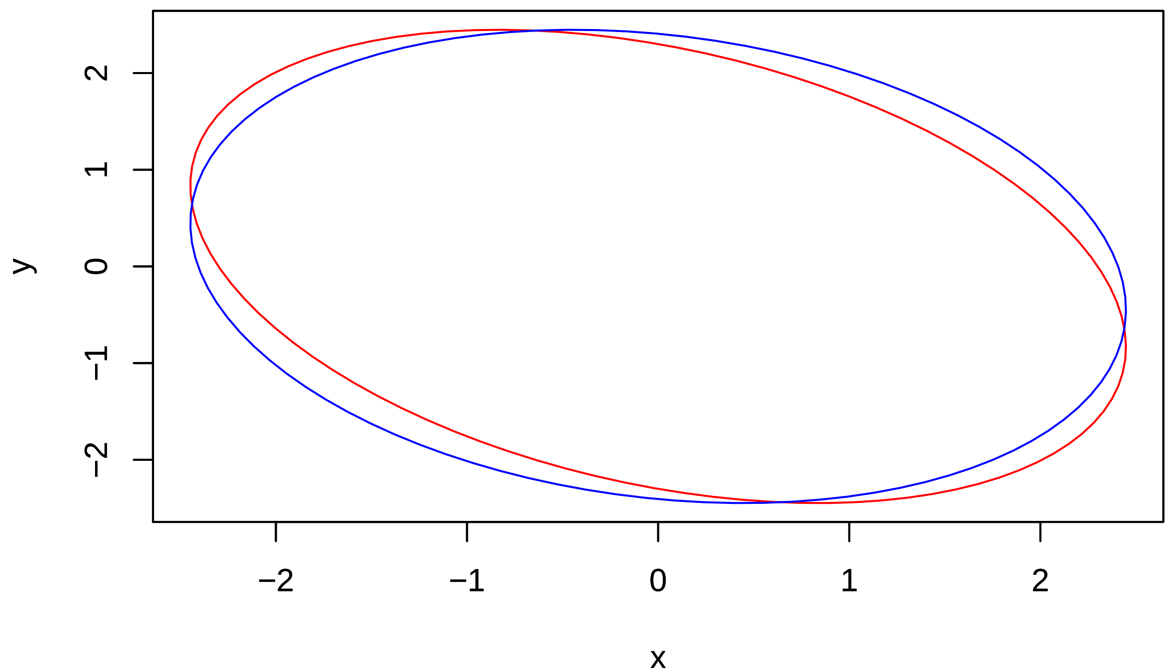
Restricting attention to the coordinates for the fifth and sixth chemicals (and on the same plot), draw the 95% confidence ellipses for the two groups. (You may use the ellipse package.) [5 points]

```
##
## Attaching package: 'ellipse'

## The following object is masked from 'package:graphics':
##
##     pairs
```

```
## [1] "Red Circle: Correlation between Chemical 5 and 6 in Sub-region 5."
```

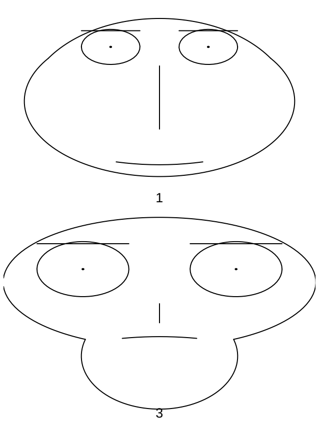
```
## [1] "Blue Circle: Correlation between Chemical 5 and 6 in Sub-region 6."
```



Part C:

Part i:

Display the three regional means of the chemical compositions for the olive oils in the three regions, using Chernoff faces.



Comment. [1 page; 5 points] I can see that the 3 regions show very different chemical compositions. Unfortunately, I can't figure out how to put a legend on this plot, so I can't tell what feature corresponds with which chemical. To me it looks like the chemical composition between region 2 an 3 are most similar.

Part ii:

Perform a one-way multivariate analysis of variance to test for differences in mean chemical composition among the three regions. Evaluate the assumptions made in making this inference? [0 points]

I am leaving question blank because I need to study for the exam tomorrow. I will revisit this question after the exam.