

# Worksheet 5

Peyton Hall

09/26/2025

$$H_0 : \mu_{low} = \mu_{med} = \mu_{high} \quad \text{vs.} \quad H_a : \text{at least two means differ}$$

In Class Coding One-Way ANOVA

```
library(readxl)
IQ_Data <- read_excel("~/Desktop/DATA 499/Week 5/IQ Data.xlsx")
# IQ_Data
model1 <- aov(iqf ~ lead_grp, data = IQ_Data)
summary(model1)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## lead_grp      2     711    355.4    1.734   0.181
## Residuals    121   24808    205.0
```

```
TukeyHSD(model1) # pairwise comparison
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = iqf ~ lead_grp, data = IQ_Data)
##
## $lead_grp
##              diff              lwr              upr              p adj
## Low-High    5.657343   -2.544799   13.859484   0.2342647
## Med-High    1.522727   -8.506047   11.551502   0.9309908
## Med-Low   -4.134615  -12.065709    3.796478   0.4337876
```

Fail to reject  $H_0$ . There is no evidence that the mean IQ scores are significantly different between lead groups.

$$H_0 : \mu_{Males} = \mu_{Females} \quad \text{vs.} \quad H_a : \mu_{Males} \neq \mu_{Females}$$

In Class Coding Two-Way ANOVA

```
# see slide 9 of week 5 review ANOVA
gender <- rep(c("Male", "Female"), each = 15)
treatment <- rep(rep(c("A", "B", "C"), each = 5), times = 2)
relieftime <- c(12, 15, 16, 17, 14, 14, 17, 19, 20, 17, 25, 27, 29, 24, 22,
               21, 19, 18, 24, 25, 21, 20, 23, 27, 25, 37, 34, 36, 26, 29)
```

```
myrelief <- data.frame(gender, treatment, relieftime)
```

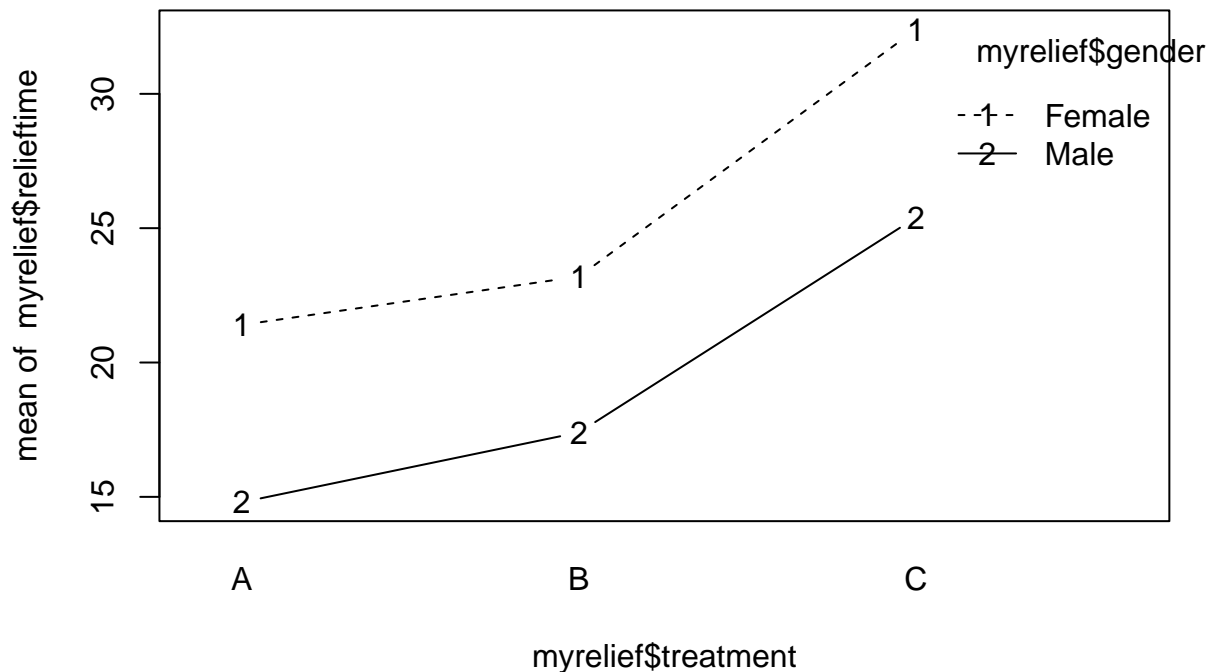
```
model2 <- aov(relieftime ~ gender + treatment + gender*treatment, data = myrelief)
summary(model2)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## gender          1  313.6    313.6    33.54 5.7e-06 ***
## treatment        2  651.5    325.7    34.84 8.0e-08 ***
## gender:treatment  2    1.9     0.9     0.10  0.905
## Residuals       24  224.4     9.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(model2) # pairwise comparison
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = relieftime ~ gender + treatment + gender * treatment, data = myrelief)
##
## $gender
##              diff            lwr            upr      p adj
## Male-Female -6.466667 -8.771096 -4.162237 5.7e-06
##
## $treatment
##              diff            lwr            upr      p adj
## B-A    2.2 -1.214989  5.614989 0.2613818
## C-A   10.8  7.385011 14.214989 0.0000001
## C-B    8.6  5.185011 12.014989 0.0000049
##
## $'gender:treatment'
##              diff            lwr            upr      p adj
## Male:A-Female:A  -6.6 -12.57951 -0.6204899 0.0245686
## Female:B-Female:A  1.8  -4.17951  7.7795101 0.9345076
## Male:B-Female:A   -4.0  -9.97951  1.9795101 0.3360790
## Female:C-Female:A 11.0   5.02049 16.9795101 0.0000978
## Male:C-Female:A   4.0  -1.97951  9.9795101 0.3360790
## Female:B-Male:A    8.4   2.42049 14.3795101 0.0026899
## Male:B-Male:A     2.6  -3.37951  8.5795101 0.7580309
## Female:C-Male:A   17.6  11.62049 23.5795101 0.0000000
## Male:C-Male:A    10.6   4.62049 16.5795101 0.0001624
## Male:B-Female:B  -5.8 -11.77951  0.1795101 0.0609450
## Female:C-Female:B  9.2   3.22049 15.1795101 0.0009715
## Male:C-Female:B   2.2  -3.77951  8.1795101 0.8608062
## Female:C-Male:B   15.0   9.02049 20.9795101 0.0000008
## Male:C-Male:B     8.0   2.02049 13.9795101 0.0044533
## Male:C-Female:C  -7.0 -12.97951 -1.0204899 0.0152655
```

```
interaction.plot(myrelief$treatment, myrelief$gender, myrelief$relieftime, fun = mean, type = "b")
```



$f = 33.54$ ;  $p\text{-value} = 0$  Reject  $H_0$ . There is significant mean heart rate difference between males and females. The pairwise comparison test shows that treatment C leads to significantly longer relief times than treatments A and B. The interaction plot shows that treatment C produces the highest relief times, especially for females. The gap between males and females widens with treatment C, showing a significant interaction.

In Class Coding Component Analysis

```
factory_and_substances <- read_excel("~/Desktop/DATA 499/Week 5/factory and substances.xlsx")
model4 <- prcomp(factory_and_substances[, -1], scale. = TRUE)
result4 <- summary(model4)
result4
```

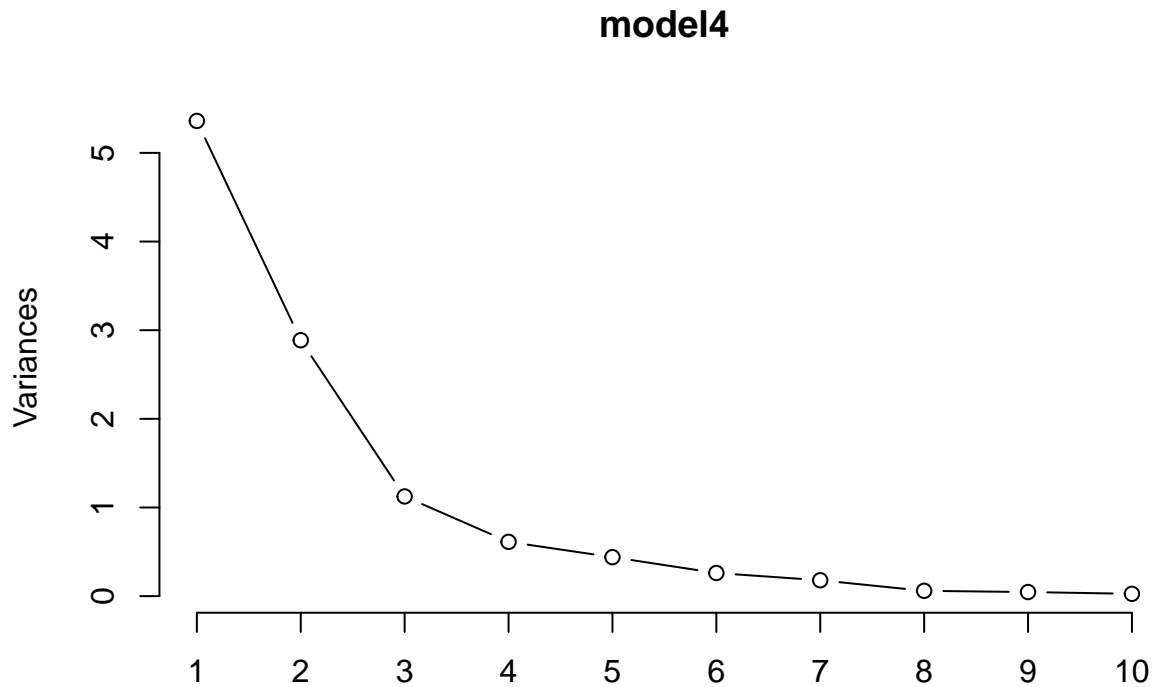
```
## Importance of components:
##               PC1    PC2    PC3    PC4    PC5    PC6    PC7
## Standard deviation  2.3152 1.6990 1.0605 0.78253 0.66303 0.51096 0.42331
## Proportion of Variance 0.4873 0.2624 0.1022 0.05567 0.03996 0.02373 0.01629
## Cumulative Proportion 0.4873 0.7497 0.8520 0.90764 0.94761 0.97134 0.98763
##               PC8    PC9    PC10    PC11
## Standard deviation  0.24525 0.21477 0.16148 0.06085
## Proportion of Variance 0.00547 0.00419 0.00237 0.00034
## Cumulative Proportion 0.99310 0.99729 0.99966 1.00000
```

```
model4$rotation
```

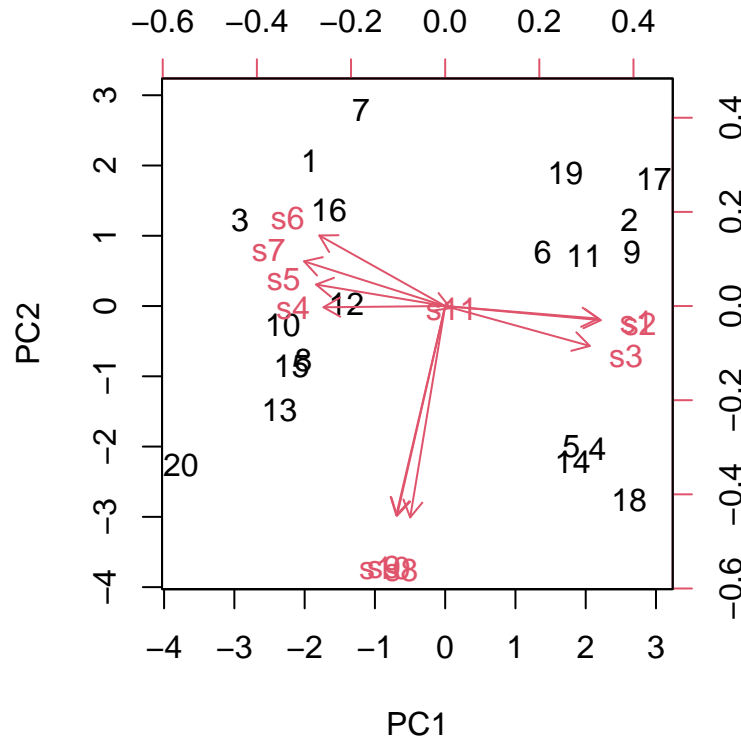
```
##           PC1           PC2           PC3           PC4           PC5           PC6
```

```
## s1  0.40694544 -0.03339569  0.052674064 -0.066028072 -0.03551565 -0.51176258
## s2  0.41330426 -0.03754534  0.050801746  0.102563045  0.27390157 -0.08786821
## s3  0.38407851 -0.10601586  0.189590673 -0.321921482  0.18245455 -0.32778269
## s4 -0.32301873 -0.00326990  0.318405346 -0.003715165  0.85784176 -0.02033177
## s5 -0.34281904  0.05710917 -0.060786864 -0.679784293 -0.03697320  0.07186318
## s6 -0.33423663  0.18741730  0.003968419  0.587849263 -0.03782930 -0.40883918
## s7 -0.37470747  0.11878231  0.124194228 -0.235968145 -0.19996940 -0.62330103
## s8 -0.09295179 -0.56131360 -0.056776497 -0.032707930 -0.05160594 -0.15254557
## s9 -0.12993068 -0.55544630  0.002562265  0.065622263 -0.02363896  0.02096085
## s10 -0.12791156 -0.55687055  0.022626837  0.103208673 -0.03643209 -0.03309875
## s11  0.01376587 -0.00554268  0.913477384  0.045731235 -0.32741305  0.19209467
##      PC7      PC8      PC9      PC10      PC11
## s1  0.29081675  0.30608020 -0.43769021  0.42427061 -0.1146552488
## s2  0.32770706  0.21738289  0.06969699 -0.71735705  0.2404599854
## s3 -0.11864324 -0.46821576  0.55126821  0.13988112 -0.0645718414
## s4 -0.07070950  0.06518500 -0.17895646  0.12273117 -0.0401159980
## s5  0.63091731 -0.03833668  0.04534017 -0.02050011  0.0698444472
## s6  0.45120110 -0.29731334  0.21014448  0.04369146 -0.0308268554
## s7 -0.40130139  0.27652641  0.03674219 -0.31963149  0.0697425640
## s8 -0.02836268 -0.54760321 -0.52140390 -0.27036753 -0.0206390958
## s9  0.11634688  0.34972095  0.30659220 -0.07950701 -0.6581222793
## s10 0.02590835  0.21422215  0.22133860  0.29220437  0.6921058066
## s11 0.10612952 -0.03446123 -0.07231563 -0.03829880 -0.0006572956
```

```
plot(model4, type = "l")
```



```
biplot(model4, scale = 0)
```



$H_0 : \mu_{DrugA} = \mu_{DrugB} = \mu_{DrugC}$  vs.  $H_a : \text{At least two drug means differ}$

$H_0 : \mu_{Male} = \mu_{Female}$  vs.  $H_a : \mu_{Male} \neq \mu_{Female}$

Question 1

```
library(readxl)
DrugGenderCreatine <- read_excel("~/Desktop/DATA 499/Week 5/DrugGenderCreatine.xlsx")
# DrugGenderCreatine
# step 2
model1 <- aov(creatinine ~ drug + gender + drug*gender, data = DrugGenderCreatine)
# step 3
summary(model1)
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## drug         2 2704.1  1352.0   24.993 6.39e-06 ***
## gender        1 1134.4  1134.4   20.969 0.000233 ***
## drug:gender    2   60.7    30.4    0.561 0.580033
## Residuals    18  973.7    54.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$f = 24.993$ ;  $p\text{-value} = 6.39e-06$  Reject  $H_0$ . There is a significant mean creatinine difference between at least two drugs  $f = 20.969$ ;  $p\text{-value} = 0.000233$  Reject  $H_0$ . There is a significant mean creatinine difference

between males and females.  $f = 0.561$ ;  $p\text{-value} = 0.580033$  Fail to reject  $H_0$ . There is no interaction between gender and drug.

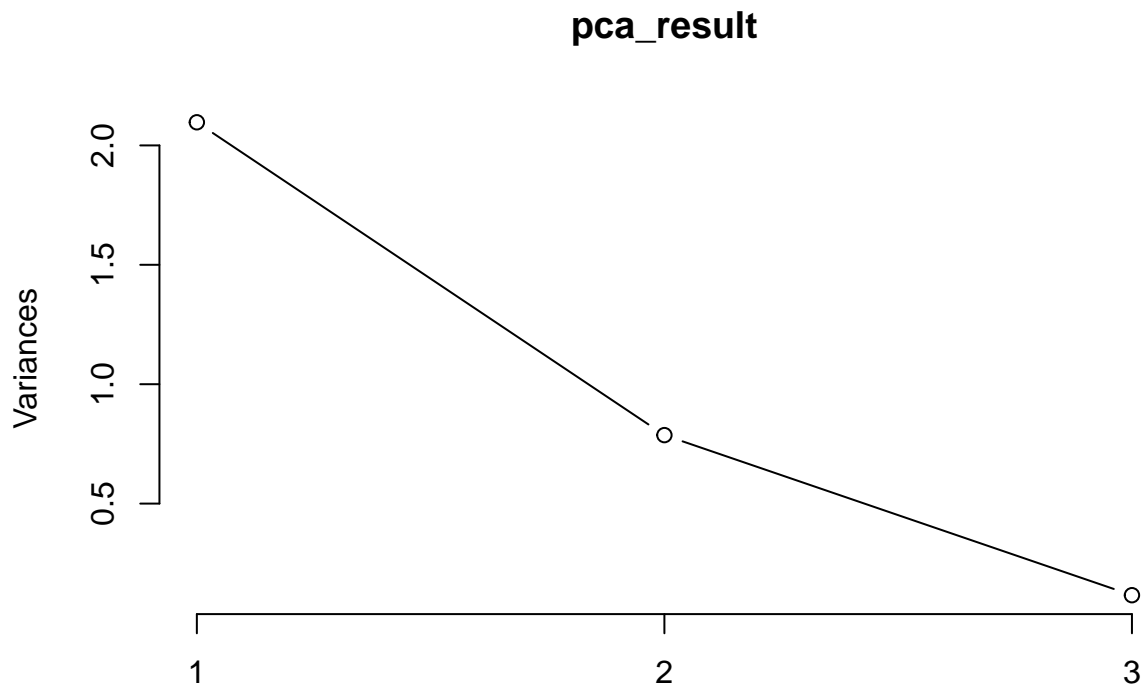
## Question 2

```
library(readxl)
Earthquake <- read_excel("~/Desktop/DATA 499/Week 5/Earthquake.xlsx")

# select only the dependent variables (longitude, depth, magnitude)
eq_data <- Earthquake[, c("longitude", "depth", "magnitude")]
pca_result <- prcomp(eq_data, scale. = TRUE) # perform PCA
summary(pca_result)
```

```
## Importance of components:
##              PC1      PC2      PC3
## Standard deviation    1.4480 0.8870 0.34129
## Proportion of Variance 0.6989 0.2623 0.03883
## Cumulative Proportion 0.6989 0.9612 1.00000
```

```
# plot to visualize how many components to keep
plot(pca_result, type = "l")
```



```
# loadings (coefficients for linear combinations)
pca_result$rotation
```

```
##           PC1           PC2           PC3
## longitude  0.4035810  0.9148345 -0.01414789
## depth      0.6454296 -0.2956244 -0.70429171
## magnitude -0.6484928  0.2751073 -0.70976970
```

```
# variance explained by first two PCs
summary(pca_result)$importance[2, 1:2] # proportion of variance
```

```
##      PC1      PC2
## 0.69892 0.26226
```

```
summary(pca_result)$importance[3, 1:2] # cumulative variance
```

```
##      PC1      PC2
## 0.69892 0.96117
```

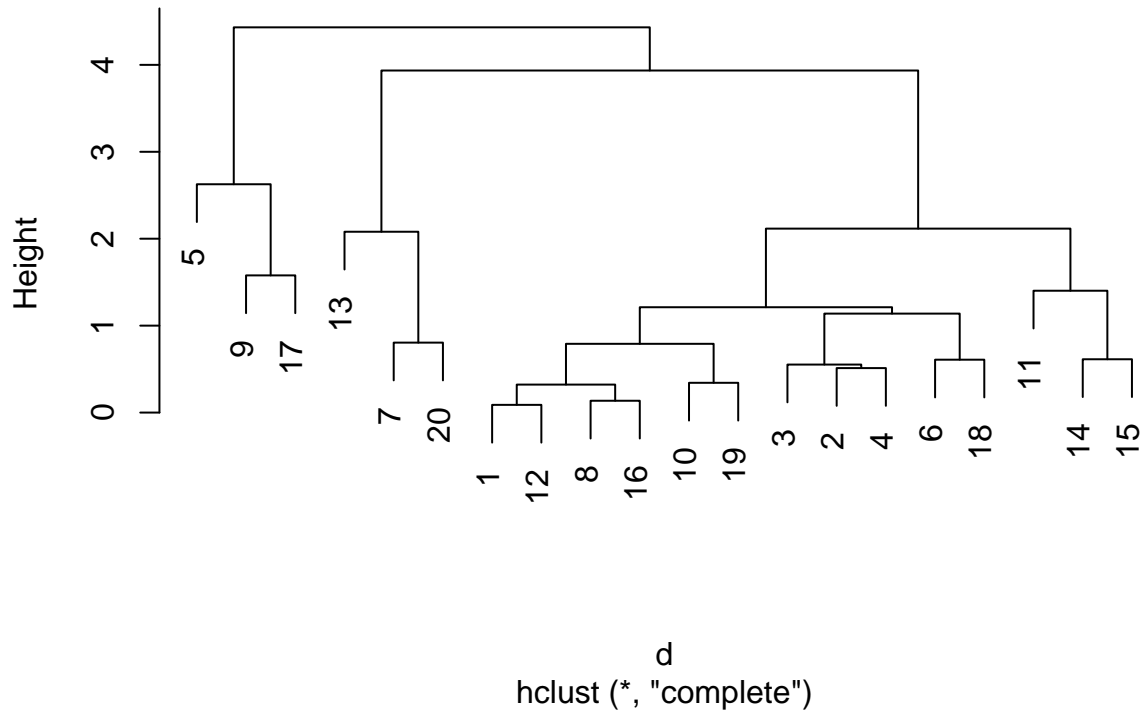
```
# equation of the first principal component
pc1 <- pca_result$rotation[,1]
pc1_equation <- paste("PC1 =",
                      round(pc1["Longitude"], 3), "* Longitude +",
                      round(pc1["Depth"], 3), "* Depth +",
                      round(pc1["Magnitude"], 3), "* Magnitude")
pc1_equation
```

```
## [1] "PC1 = NA * Longitude + NA * Depth + NA * Magnitude"
```

Question 3

```
library(readxl)
CreditCard <- read_excel("~/Desktop/DATA 499/Week 5/CreditCard.xlsx")
# scale the data
c_data <- scale(CreditCard[, c("BALANCE", "PURCHASES", "PAYMENTS")])
d <- dist(c_data, method = "euclidean") # distance matrix
modelq3 <- hclust(d, method = "complete") # hierarchical clustering
plot(modelq3, main = "Credit Card Dendrogram") # dendrogram
```

## Credit Card Dendrogram



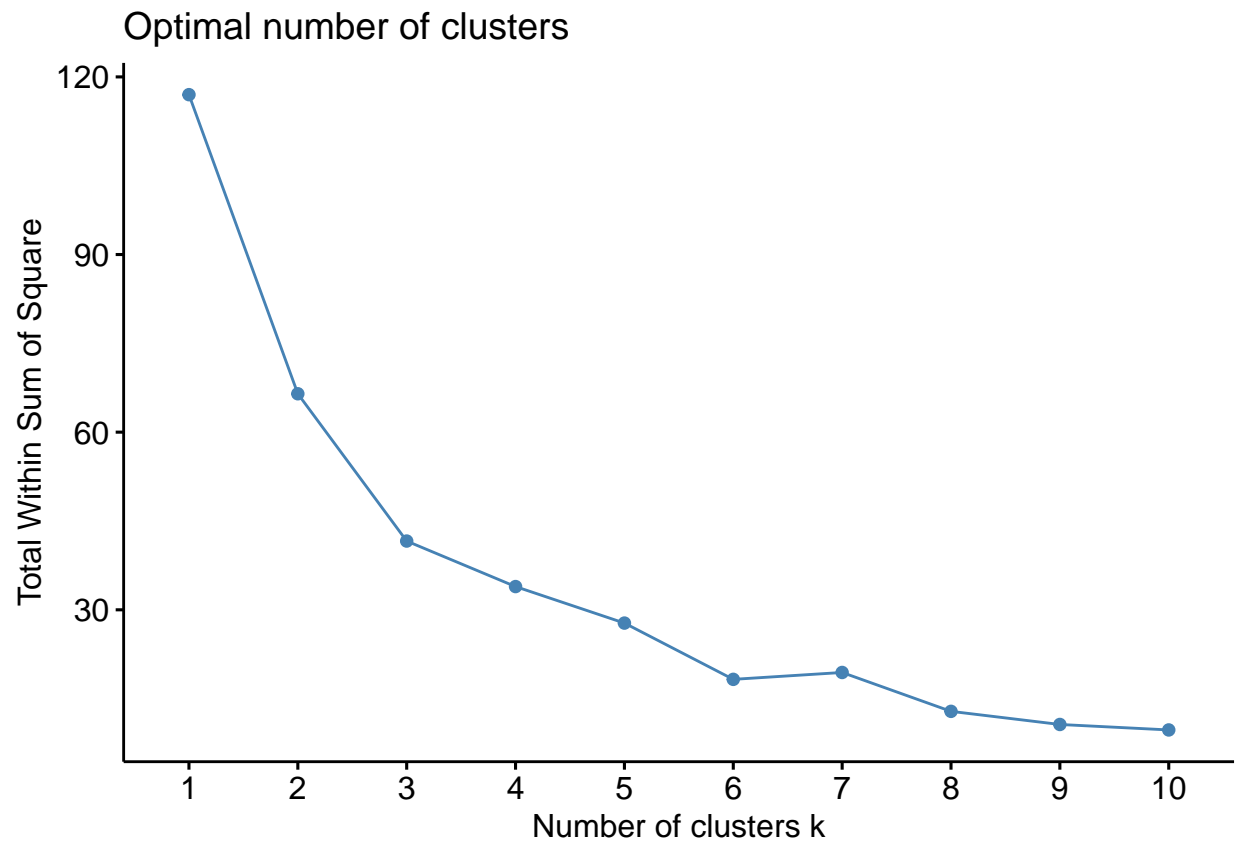
Question 4

```
library(readxl)
library(factoextra)
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
CAHousing <- read_excel("~/Desktop/DATA 499/Week 5/CAHousing.xlsx")
data2 <- scale(CAHousing[, c("total_rooms", "total_bedrooms", "median_value")])
fviz_nbclust(data2, kmeans, method = "wss")
```



```
model6 <- kmeans(data2, centers = 3, nstart = 40)
fviz_cluster(model6, data = data2, geom = "text", ellipse.type = "convex")
```

