Homework 5

Peyton Hall

Homework 5

$H_0 : \mu_{\text{Johns Hopkins}} = \mu_{\text{Rancho Los Amigos}} = \mu_{\text{St. Louis}}$ vs. H_a : At least one μ is different among the three medical centers
Question 1

```
johns_hopkins <- c(3.23, 3.47, 1.86, 2.47, 3.01,
                   1.69, 2.10, 2.81, 3.28, 3.36)

rancho_los_amigos <- c(3.22, 2.88, 1.71, 2.89, 3.77,
                       3.29, 3.39, 3.86, 2.64, 2.64)

st_louis <- c(2.79, 3.22, 2.25, 2.98, 2.47,
              2.77, 2.95, 3.56, 2.88, 2.88)

fev_data <- data.frame(
  center = factor(rep(c("Johns Hopkins", "Rancho Los Amigos", "St. Louis"), each = 10)),
  FEV = c(johns_hopkins, rancho_los_amigos, st_louis))

# Fit one-way ANOVA
fit <- aov(FEV~center, data = fev_data)
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## center       2  0.453  0.2265   0.709  0.501
## Residuals   27  8.621  0.3193
```

```
# Extract test statistic and p-value
anova_table <- summary(fit)[[1]]
F_stat <- anova_table$`F value`[1]
p_val <- anova_table$`Pr(>F)`[1]

F_stat
```

```
## [1] 0.7094999
```

```
p_val
```

```
## [1] 0.5008321
```

f = 0.709; p-value = 0.501 Fail to reject H_0. p-value (0.501) is greater than the significance level (0.05). There is not sufficient evidence to conclude that the mean FEV levels are significantly different among the three medical centers.

$H_{0,1}$: $\mu_{\text{Low}} = \mu_{\text{Medium}} = \mu_{\text{High}}$ vs. $H_{a,1}$: At least one mean is different
$H_{0,2}$: $\mu_{\text{Short}} = \mu_{\text{Long}}$ vs. $H_{a,2}$: $\mu_{\text{Short}} \neq \mu_{\text{Long}}$
$H_{0,3}$: No interaction effect between temperature and presoaking vs. $H_{a,3}$: There is an interaction effect between temperatu
Question 2

```
short_low <- c(8, 5, 9)
short_medium <- c(18, 15, 17)
short_high <- c(12, 11, 14)
long_low <- c(11, 13, 15)
long_medium <- c(20, 19, 20)
long_high <- c(15, 16, 12)

taste_data <- data.frame(
  presoaking = factor(rep(c("Short", "Long"), each = 9)),
  temperature = factor(rep(c("Low", "Medium", "High"), each = 3, times = 2)),
  score = c(short_low, short_medium, short_high,
            long_low, long_medium, long_high))
taste_data
```

```
##    presoaking temperature score
## 1       Short         Low     8
## 2       Short         Low     5
## 3       Short         Low     9
## 4       Short      Medium    18
## 5       Short      Medium    15
## 6       Short      Medium    17
## 7       Short        High    12
## 8       Short        High    11
## 9       Short        High    14
## 10       Long         Low    11
## 11       Long         Low    13
## 12       Long         Low    15
## 13       Long      Medium    20
## 14       Long      Medium    19
## 15       Long      Medium    20
## 16       Long        High    15
## 17       Long        High    16
## 18       Long        High    12
```

```
fit2 <- aov(score~temperature*presoaking, data = taste_data) # Two-way ANOVA
summary(fit2)
```

```
##                          Df Sum Sq Mean Sq F value   Pr(>F)
## temperature               2 194.78   97.39   33.08 1.31e-05 ***
## presoaking                1  56.89   56.89   19.32 0.000872 ***
## temperature:presoaking    2  10.78    5.39    1.83 0.202430
## Residuals                12  35.33    2.94
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova_table <- summary(fit2)[[1]] # extract test statistics and p-values
```

```r
# test statistic and p-value for temperature effect
F_temp <- anova_table$`F value`[1]
p_temp <- anova_table$`Pr(>F)`[1]

# test statistic and p-value for presoaking effect
F_presoak <- anova_table$`F value`[2]
p_presoak <- anova_table$`Pr(>F)`[2]

# test statistic and p-value for interaction
F_interaction <- anova_table$`F value`[3]
p_interaction <- anova_table$`Pr(>F)`[3]

F_temp; p_temp
```

```
## [1] 33.07547
```

```
## [1] 1.310636e-05
```

```r
F_presoak; p_presoak
```

```
## [1] 19.32075
```

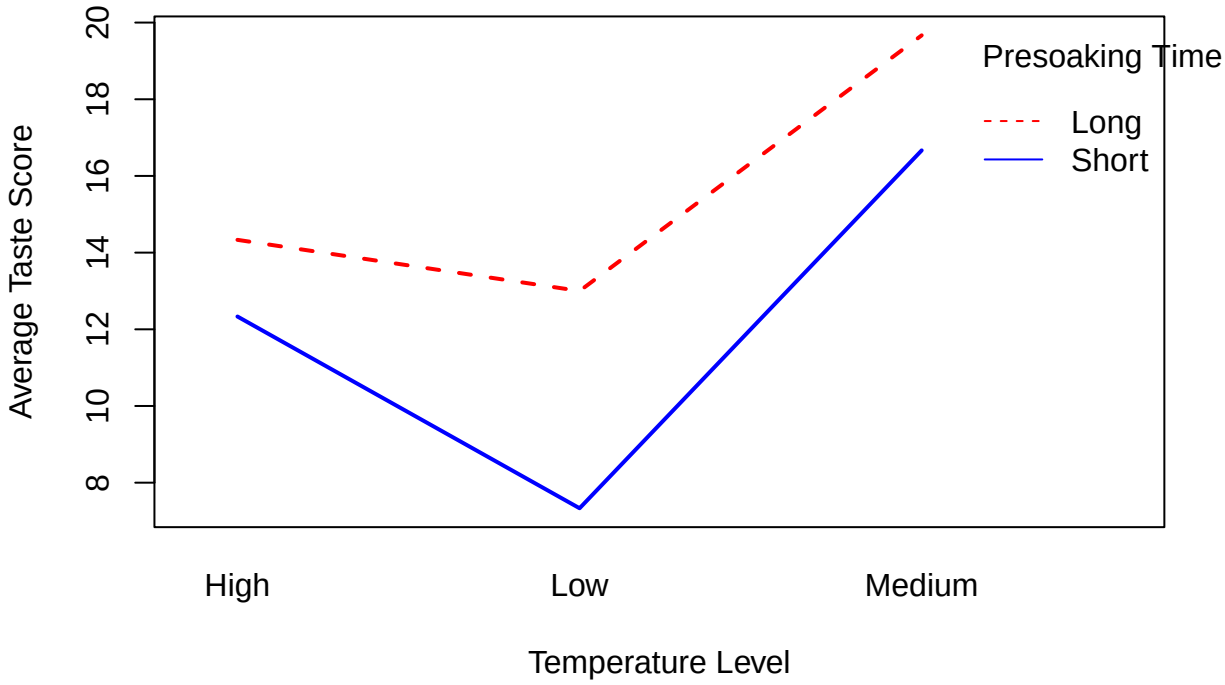
```
## [1] 0.0008721909
```

```r
F_interaction; p_interaction
```

```
## [1] 1.830189
```

```
## [1] 0.2024297
```

```r
interaction.plot(taste_data$temperature, taste_data$presoaking,
                 taste_data$score,
                 xlab = "Temperature Level",
                 ylab = "Average Taste Score",
                 trace.label = "Presoaking Time",
                 col = c("red","blue"),
                 lwd = 2)
```



temperature:
F = 33.08; p-value = 1.31e-05
Reject H0. There is a significant effect of temperature presoaking:
F = 19.32; p-value = 0.000872
Reject H0. There is a significant effect of presoaking interaction:
F = 1.83; p-value = 0.202
Fail to reject H0. No significant interaction effect between temperature and presoaking

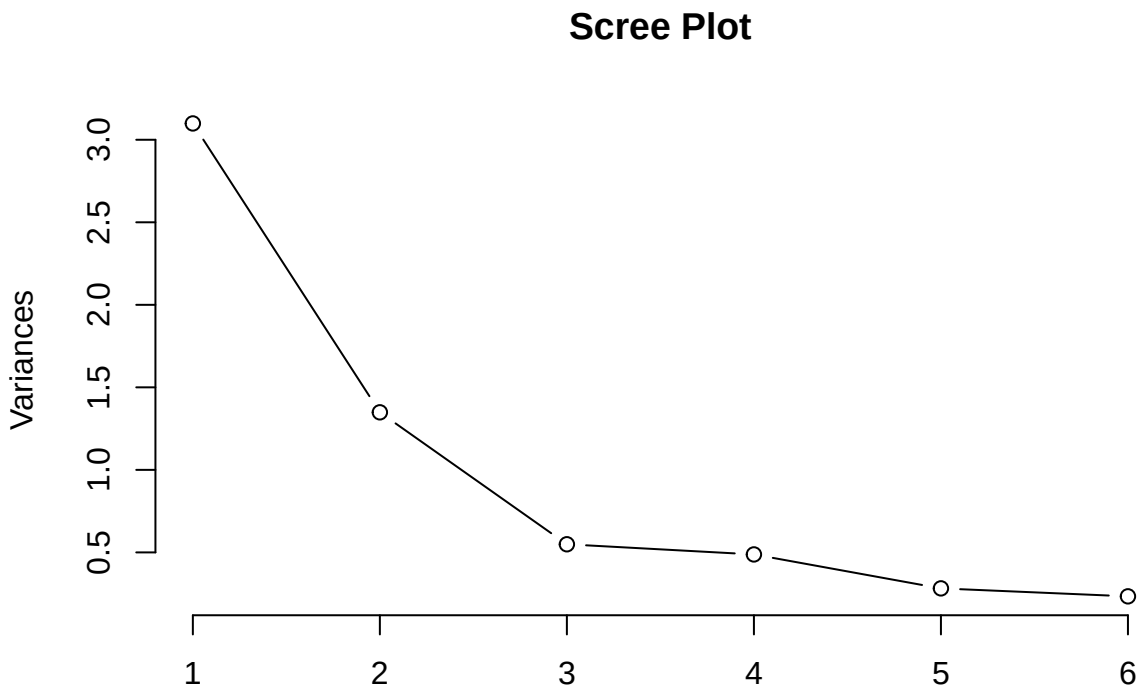
Question 3

```
library(readxl)
psychological_test_data <- read_excel("~/Desktop/DATA 499/Week 5/psychological test data.xlsx")
psychological_test_data
```

```
## # A tibble: 73 x 6
##    visperc cubes lozenges paragrap sentence wordmean
##      <dbl> <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
##  1      33    22       17        8       17       10
##  2      30    25       20       10       23       18
##  3      36    33       36       17       25       41
##  4      28    25        9       10       18       11
##  5      30    25       11       11       21        8
##  6      20    25        6        9       21       16
##  7      17    21        6        5       10       10
##  8      33    31       30       11       23       18
##  9      30    22       20        8       17       20
## 10      36    28       22       13       24       36
```

```
## # i 63 more rows
```

```
psych_pca <- prcomp(psychological_test_data, scale. = TRUE) # PCA
plot(psych_pca, type = "l", main = "Scree Plot")
```




```
eigenvalues <- psych_pca$sdev^2
eigenvalues
```

```
## [1] 3.0988862 1.3486867 0.5491883 0.4876152 0.2819058 0.2337179
```

a) Keep 2 PCAs
b) PC1 explains about 60–65% of the variance, PC2 explains about 15–20%. Together about 80–85%.
c) Each PC is a linear combination of the six variables.
d) From the biplot, the six variables can be grouped into Group 1: visperc, cubes, lozenges (visual/spatial tests) and Group 2: paragrap, sentence, wordmean (verbal/linguistic tests).

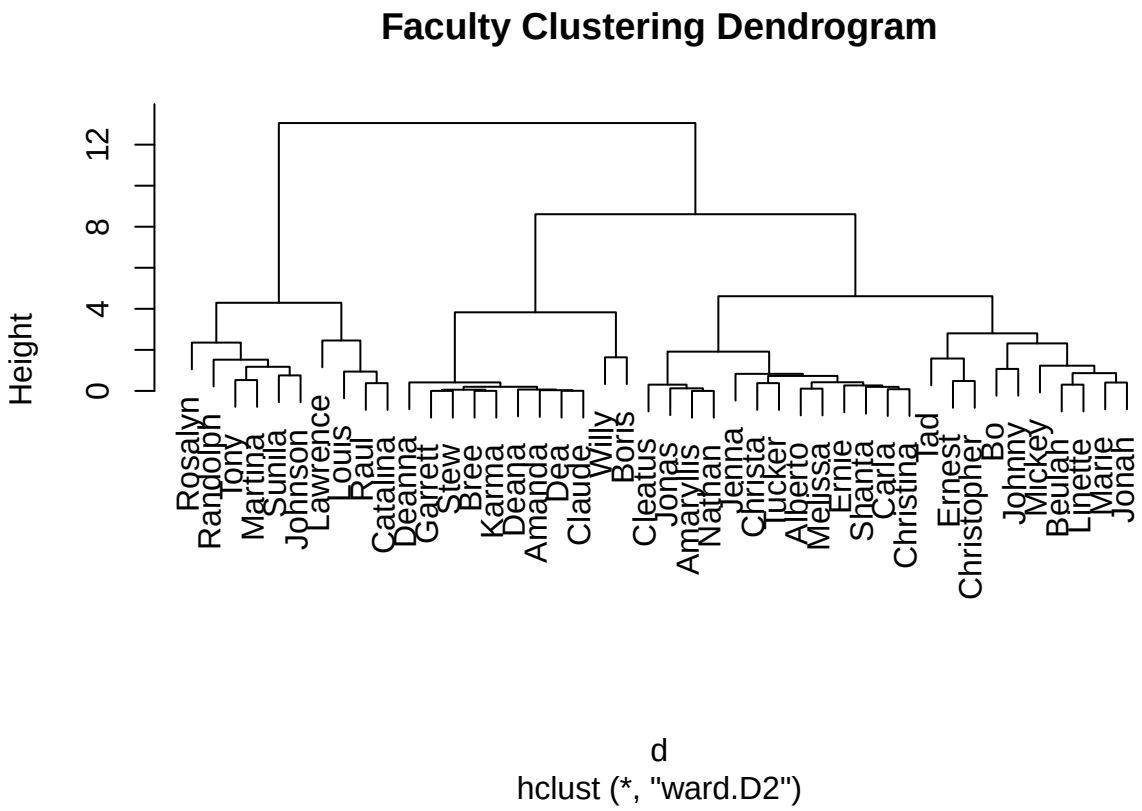
Question 4

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

```
## # A tibble: 44 x 5
```

```
##    Name      Salary  Rank Articles Experience
##    <chr>      <dbl> <dbl>    <dbl>      <dbl>
##  1 Rosalyn  123600      5       21         32
##  2 Lawrence  96800.     5       71         27
##  3 Sunila    83358      5       19         24
##  4 Randolph  83236.     5       17         38
##  5 Mickey    75041.     5       18         11
##  6 Louis     74957.     5       50         26
##  7 Tony      72226.     4       25         28
##  8 Raul      72056.     5       41         18
##  9 Catalina  68128.     5       40         22
## 10 Johnson   66158.     5       26         24
## # i 34 more rows
```

```r
# remove the Name column for clustering (keep it separate for labeling later)
faculty_names <- ClusterFaculty$Name
cluster_vars <- ClusterFaculty[, -1]
# scale the variables so they are on the same scale
cluster_scaled <- scale(cluster_vars)
# perform hierarchical clustering using Ward's method
d <- dist(cluster_scaled, method = "euclidean")
hc <- hclust(d, method = "ward.D2")
# dendrogram
plot(hc, labels = faculty_names, main = "Faculty Clustering Dendrogram")
```



```r
clusters <- cutree(hc, k = 4) # cut into 4 clusters
table(clusters) # find the smallest cluster
```

```
## clusters
##  1  2  3  4
## 10 10 13 11
```

```r
smallest_cluster <- which.min(table(clusters))
# faculty names in the smallest cluster
faculty_names[clusters == smallest_cluster]
```

```
##  [1] "Rosalyn"  "Lawrence" "Sunila"   "Randolph" "Louis"    "Tony"
##  [7] "Raul"     "Catalina" "Johnson"  "Martina"
```

a) Rosalyn, Lawrence, Sunila, Randolph, Louis, Tony, Raul, Catalina, Johnson, Martina.

Question 5

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

```
## # A tibble: 20 x 3
##    CustermorID AnnualIncome SpendingScore
##          <dbl>        <dbl>         <dbl>
##  1           1          150            39
##  2           2          290            91
##  3           3          160            41
##  4           4          170            49
##  5           5          159            36
##  6           6          143            31
##  7           7          121            19
##  8           8          120            23
##  9           9          110            20
## 10          10           98            17
## 11          11          121            22
## 12          12          132            19
## 13          13          137            25
## 14          14          126            12
## 15          15          289            76
## 16          16          300            81
## 17          17          291            79
## 18          18          117            19
## 19          19          105            21
## 20          20          101            20
```

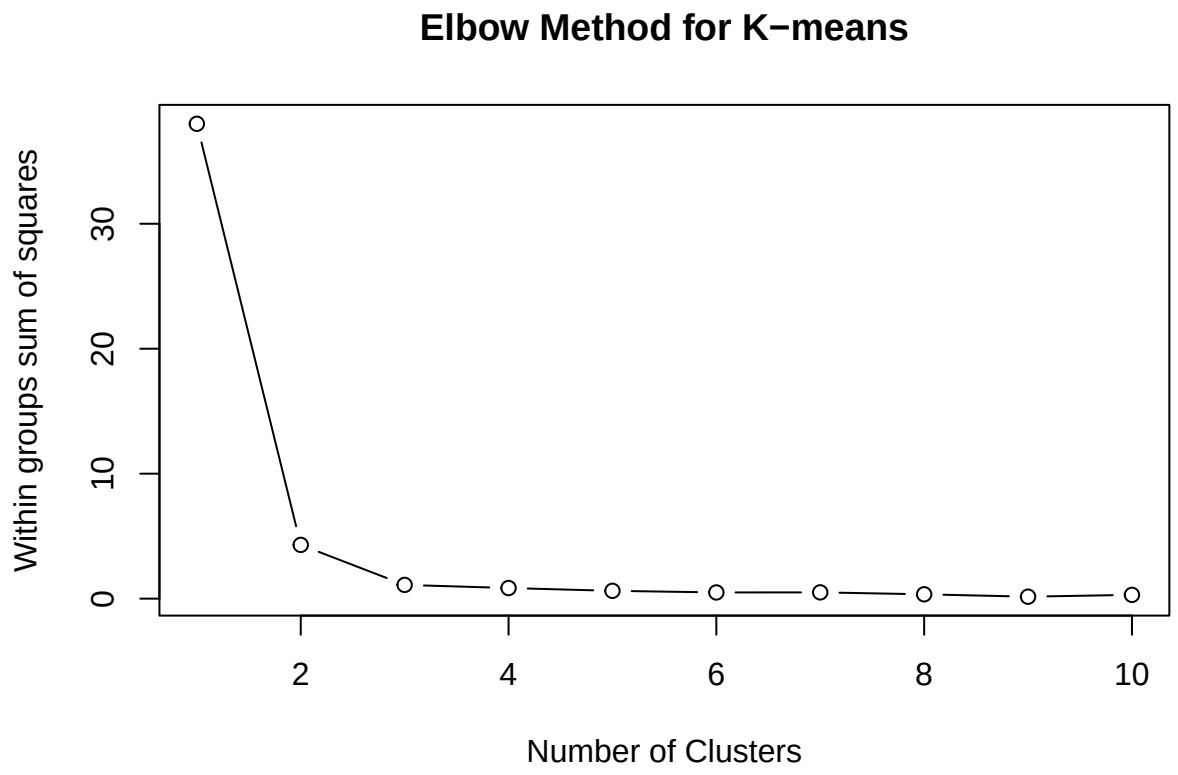
```r
# use only AnnualIncome and SpendingScore for clustering
spend_data <- SpendingScore[, c("AnnualIncome", "SpendingScore")]

spend_scaled <- scale(spend_data) # scale the data
```

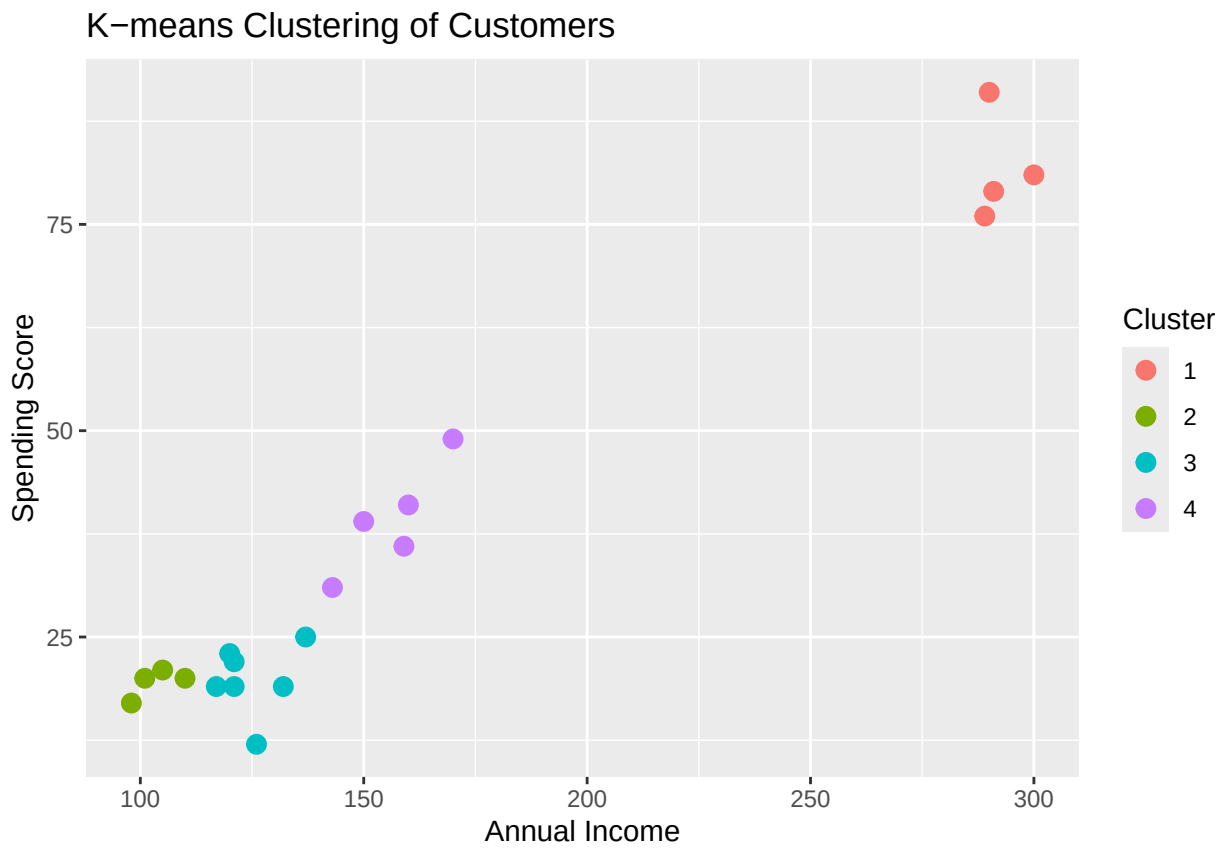
```r
# decide number of clusters using the elbow method
wss <- (nrow(spend_scaled)-1)*sum(apply(spend_scaled,2,var))
for (i in 2:10) {
  km <- kmeans(spend_scaled, centers=i)
  wss[i] <- km$tot.withinss
}
plot(1:10, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares",
     main="Elbow Method for K-means")

# from the elbow plot, pick k
set.seed(123)
kmeans_result <- kmeans(spend_scaled, centers=4)   # adjust 4 if elbow suggests different
# add cluster labels to data
SpendingScore$Cluster <- as.factor(kmeans_result$cluster)

library(ggplot2)
```




```r
ggplot(SpendingScore, aes(x=AnnualIncome, y=SpendingScore, color=Cluster)) +
  geom_point(size=3) +
  labs(title="K-means Clustering of Customers", x="Annual Income", y="Spending Score")
```



b) From the elbow plot, the sharp bend occurs at k = 4. Therefore, it was decided to use 4 clusters because adding more clusters does not significantly reduce the within-cluster variation.