Hw 3

Asahi (Ash) Kuroki

11/17/2021

```
### import libraries
library(stats)
library(factoextra)
```

```
## Loading required package: ggplot2

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

```
library(ggplot2)
```

```
telco_data <- read.csv("TelcoData.csv")
var_names <- c("Age", "Tenure.Months", "Monthly.Charges")
summary(telco_data[,var_names])
```

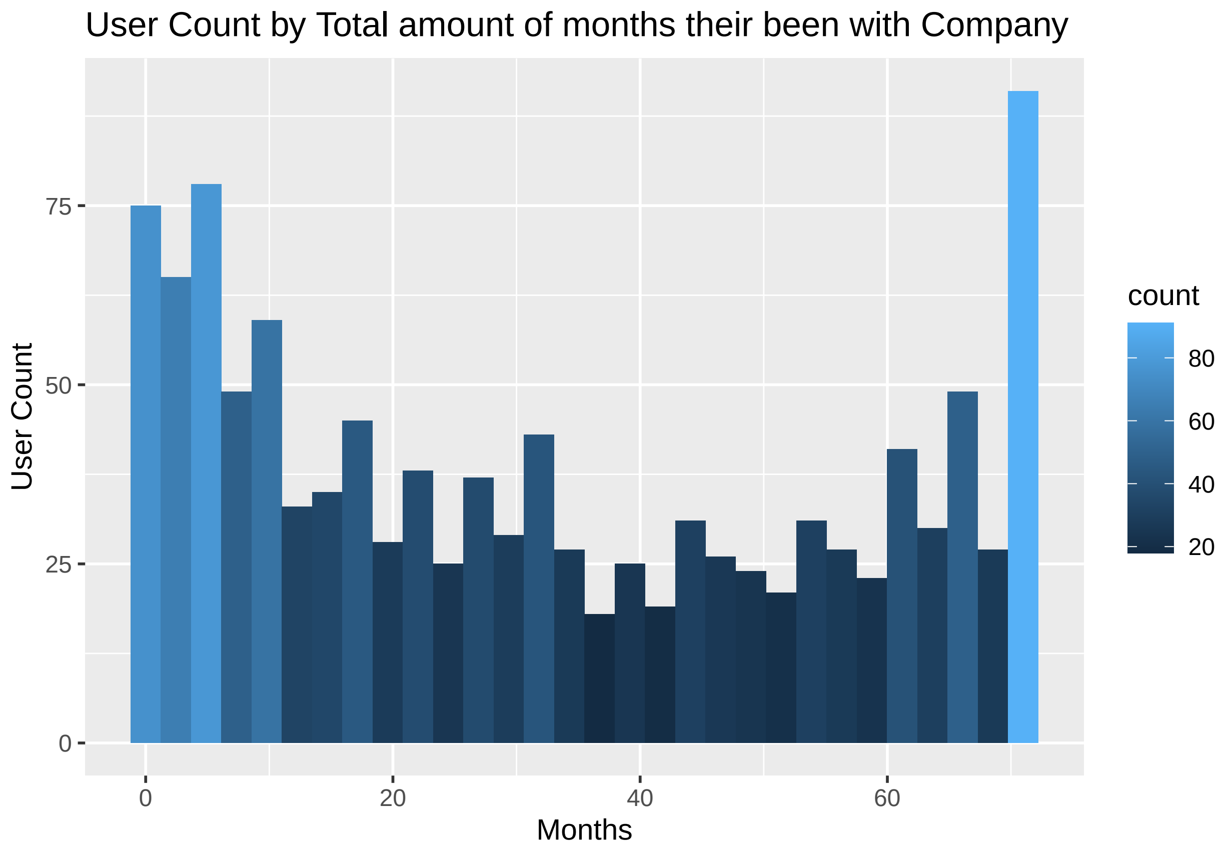
```
##       Age        Tenure.Months   Monthly.Charges
##  Min.   :19.00   Min.   : 1.00   Min.   : 68.95
##  1st Qu.:35.00   1st Qu.: 9.00   1st Qu.: 80.60
##  Median :51.00   Median :29.00   Median : 91.25
##  Mean   :50.48   Mean   :32.36   Mean   : 91.34
##  3rd Qu.:67.00   3rd Qu.:56.00   3rd Qu.:100.55
##  Max.   :80.00   Max.   :72.00   Max.   :118.35
```

Age -> Mean: 50.48, Min: 19, Max:80
Age is skewed toward Max, which makes sense because adults and parents pay bills most of the time
Tenure.Months -> Mean: 32.36, Min: 1, Max: 72
Average months are 32 months which means about 3 years Longest months are 72 which is about 6 years. It is shorter that I expected.
Montly.Charges -> Mean: 91.34, Min: 68.95, Max: 118.35
Mean for Monthly charges is $91 which is pretty much in between min and max.

b) use ggplot to produce histogram for Tenure

```
ggplot(data = telco_data, aes(x= Tenure.Months, fill = ..count..)) +
  geom_histogram(alpha=1) +
  ggtitle("User Count by Total amount of months their been with Company") +
  labs(x = "Months", y = "User Count")
```

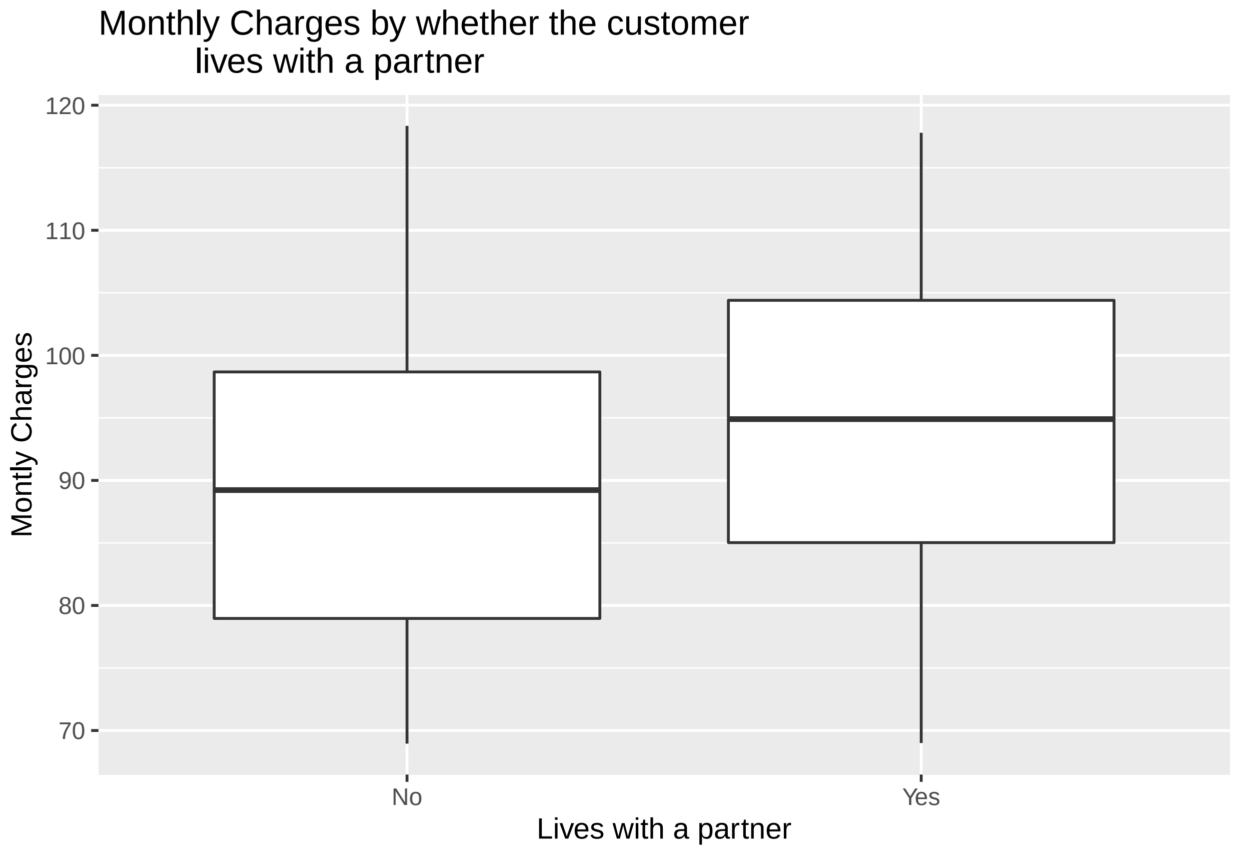
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The histogram is skewed toward the both side of the graph. There is a lot of user count in the first couple of months but gradually decrease when we go towards the middle, and go back up again as we pass the midpoint.

c) Use ggplot2 to produce a box-plot of Montlhy.Charges

```
ggplot(data = telco_data, aes(x= Partner_cat, y = Monthly.Charges,
                              group = Partner_cat )) +
  geom_boxplot(outlier.colour="orange", outlier.shape=2, outlier.size=3) +
  ggtitle("Monthly Charges by whether the customer
          lives with a partner") +
  labs(x = "Lives with a partner", y = "Montly Charges")
```



The mean value for monthly charges is higher for customers who lives with his or her partner. There are no outliers.   d) Asses which variables you should include in your cluster analysis A. Include everything but customer id and partner_cat for the cluster analysis. This is because it does make sense for us to use customer_id for our analysis since they all have unique ids. Also we only include binary variables for partner because we want to be able to calculate the distance.   e) Asses whether you need to normalize the data. A. Yes we need to normalize our data. Because, 1. We are dealing with different data types 2. Total Charges have larger numbers than any other ones.

```
min_max_norm = function(x){
  return ((x - min(x))/(max(x) - min(x)))}
telco_norm <-telco_data
norm_column <- c(2:15)
telco_norm[, norm_column] <- apply(telco_norm[, norm_column], MARGIN = 2, FUN = min_max_norm)
```

Hierarchical Clustering f) Generate a distance matrix using Euclidian distance

```
dist_matrix_norm <- dist(telco_norm[, norm_column], method = "euclidian")
# View((as.matrix(dist_matrix_norm)[1:5, 1:5]))
dist_matrix_print <- (as.matrix(dist_matrix_norm)[1:5, 1:5])
print(dist_matrix_print)
```

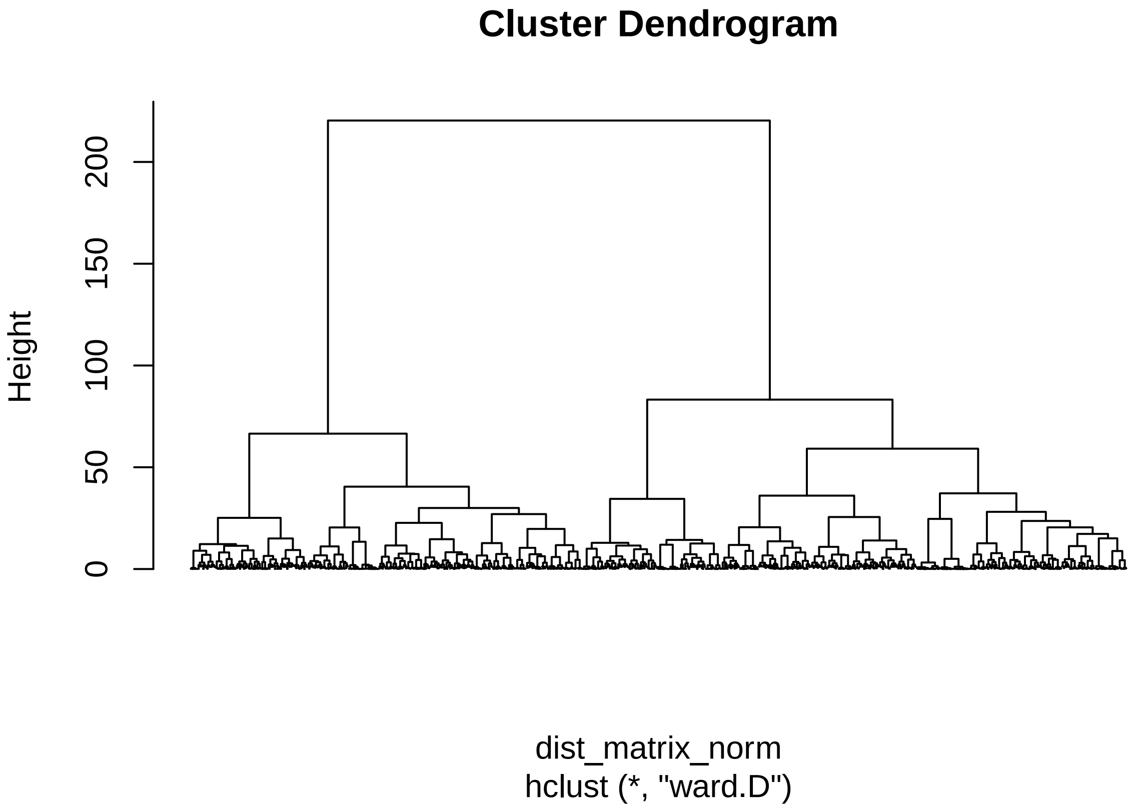
```
##          1        2        3        4        5
## 1 0.000000 2.993884 2.015139 2.391619 2.472560
## 2 2.993884 0.000000 3.370387 2.146703 1.916271
## 3 2.015139 3.370387 0.000000 2.411053 2.515465
## 4 2.391619 2.146703 2.411053 0.000000 1.728955
## 5 2.472560 1.916271 2.515465 1.728955 0.000000
```

g) Run Hierarchical clustering

```
hc <- hclust(dist_matrix_norm, method = "ward.D")
```

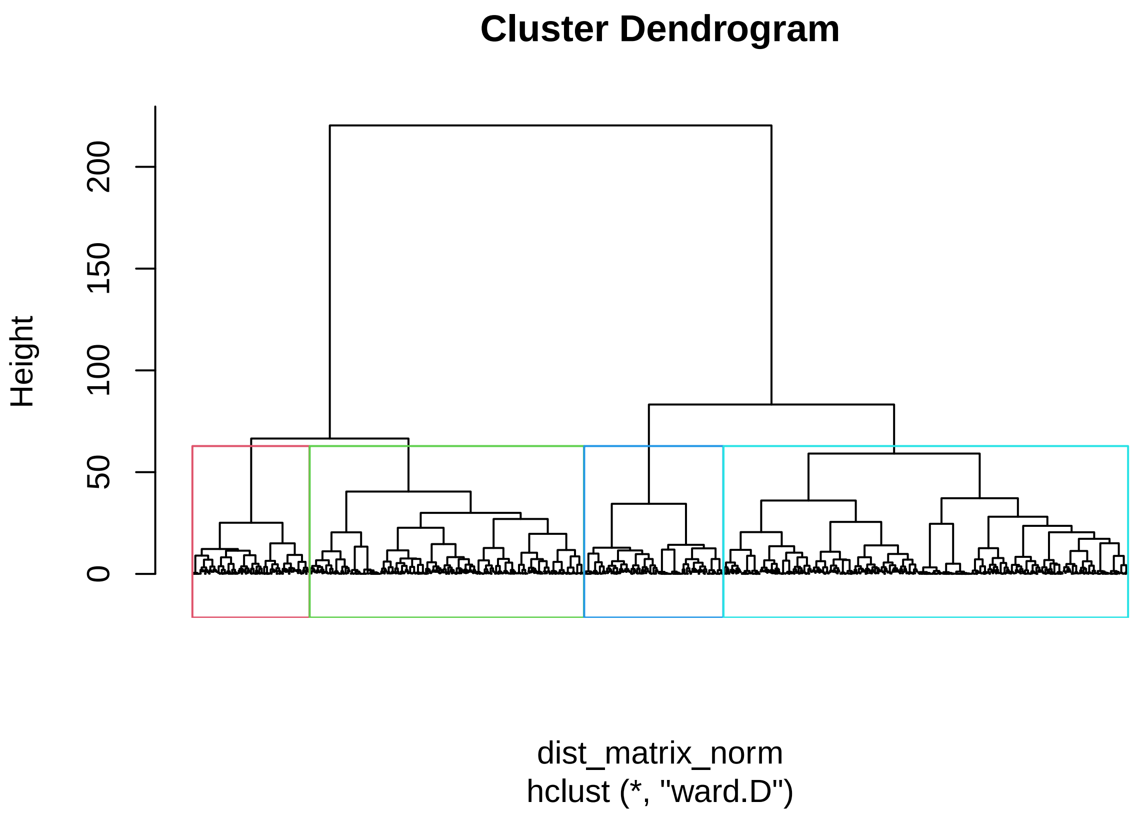
h) Plot the dendrogram: labels = False, hang = 0

```
plot(hc, hang = 0, labels = FALSE)
```



i) Use the rect.hclust() function to draw best 4 clusters

```
plot(hc, hang = 0, labels = FALSE)
rect.hclust(hc, k = 4, border = 2:5)
```



j) cut the dendrogram into 4 clusters

```r
hc_4 <- cutree(hc, k = 4)
table(hc_4)
```

```
## hc_4
##   1   2   3   4 
## 171 337 497 144
```

Cluster 1: 171 Cluster 2: 337 Cluster 3: 497 Cluster 4: 144

K) Add column to the original dataset

```r
telco_data$hc_4 <- hc_4
library(plyr)
ddply(telco_data, .(hc_4), summarize, n = length(CustomerID),
      Partner = sum(Partner) / n, senior_citizen = sum(Senior.Citizen) / n,
      online_backup = sum(Online.Backup) /n, tech_support = sum(Tech.Support) / n,
      streaming_movies = sum(Streaming.Movies) / n,
      streaming_tv = sum(Streaming.TV) /n,
      online_security = sum(Online.Security) / n,
      unlimitated_data = sum(Unlimited.Data) /n,
      Montly_mean=mean(Monthly.Charges), tenure_mean=mean(Tenure.Months),
      age_mean = mean(Age))
```

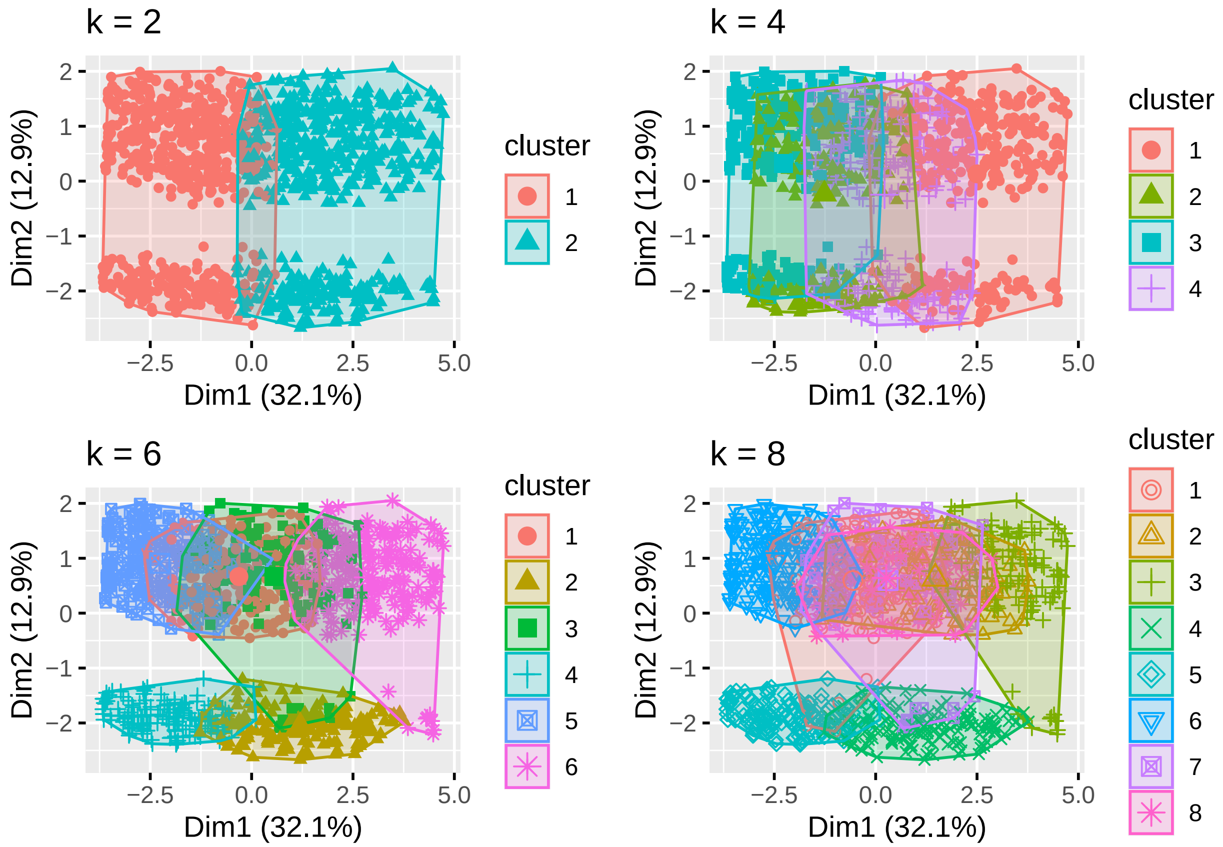
```
##   hc_4   n   Partner senior_citizen online_backup tech_support streaming_movies
## 1    1 171 0.3742690    0.988304094     0.2456140   0.05847953        0.3567251
## 2    2 337 0.6765579    0.005934718     0.7299703   0.48664688        0.7418398
## 3    3 497 0.3259557    0.034205231     0.2273642   0.18712274        0.4305835
## 4    4 144 0.6458333    0.986111111     0.6250000   0.36111111        0.7986111
##   streaming_tv online_security unlimitated_data Montly_mean tenure_mean
## 1    0.4210526      0.05263158        0.8713450    82.81959     18.10526
## 2    0.7982196      0.42729970        0.8872404   101.64332     51.38279
## 3    0.4144869      0.19114688        0.8370221    84.62746     19.43260
## 4    0.7222222      0.43750000        0.8055556   100.53611     49.36806
##   age_mean
## 1 72.12865
## 2 41.23739
## 3 42.97988
## 4 72.29861
```

K-means Clustering
I) Run four different versions of k-means clustering

```r
set.seed(123)
library(gridExtra)
k2 <- kmeans(telco_norm[, norm_column], centers = 2, nstart = 15)
k4 <- kmeans(telco_norm[, norm_column], centers = 4, nstart = 15)
k6 <- kmeans(telco_norm[, norm_column], centers = 6, nstart = 15)
k8 <- kmeans(telco_norm[, norm_column], centers = 8, nstart = 15)
```

M) use fviz_cluster to visualize

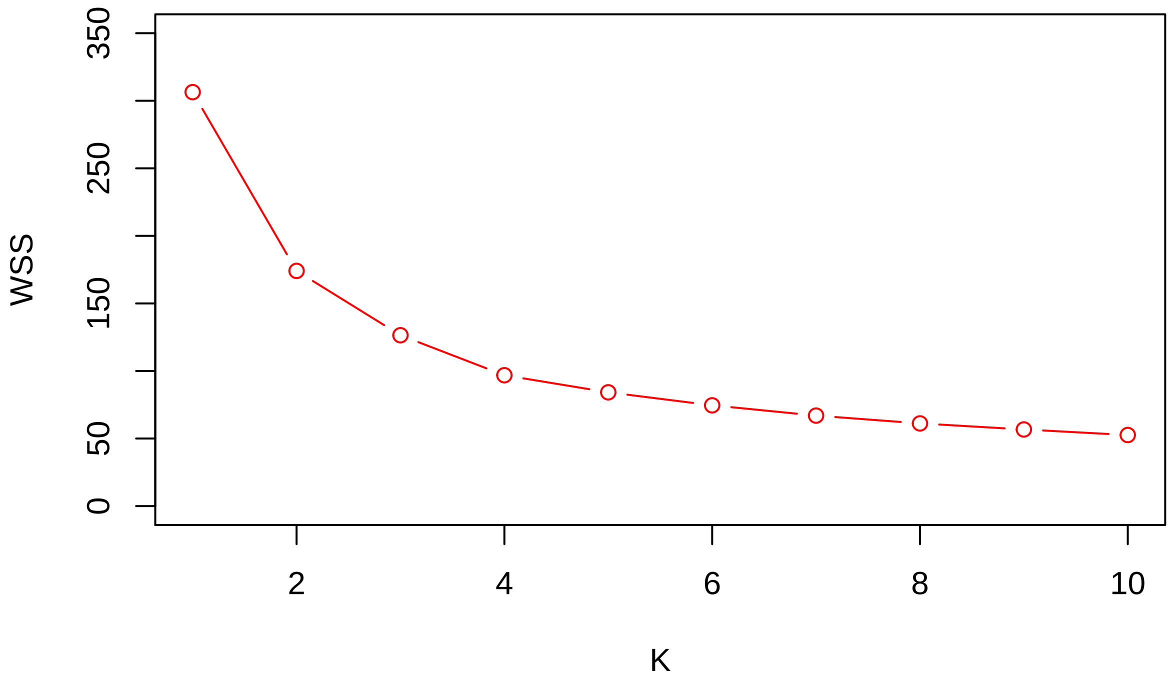
```r
p2 <- fviz_cluster(k2, geom = "point", data = telco_norm[, norm_column]) + ggtitle("k = 2")
p4 <- fviz_cluster(k4, geom = "point", data = telco_norm[, norm_column]) + ggtitle("k = 4")
p6 <- fviz_cluster(k6, geom = "point", data = telco_norm[, norm_column]) + ggtitle("k = 6")
p8 <- fviz_cluster(k8, geom = "point", data = telco_norm[, norm_column]) + ggtitle("k = 8")
grid.arrange(p2, p4, p6, p8, nrow = 2)
```



k = 2 looks the most appropriate because it has lowest overlap. When k > 2 clusters have too much overlap.

N) find out appropriate cluster by computing WSS

```
WSS_curve <- c()
for (n in 1:10){
  k = kmeans(telco_norm[,var_names], centers = n, nstart = 10)
  wss = k$tot.withinss
  WSS_curve[n] <- wss
}
plot(1:10, WSS_curve, type = "b", col = "red", ylab = "WSS", xlab = "K", ylim = c(0,350) )
```



The appropriate number of clusters might be 4. This is because there is a elbow point a t k = 4. After k = 4, the slope of the graph decreases a lot.

O) run k-menas using the k you got in part n

```
k4 <- kmeans(telco_norm[, norm_column], centers = 4, nstart = 10)
telco_data$km_4 <- k4$cluster
ddply(telco_data, .(km_4), summarize, n = length(CustomerID),
      Partner = sum(Partner) / n, senior_citizen = sum(Senior.Citizen) / n,
      online_backup = sum(Online.Backup) /n, tech_support = sum(Tech.Support) / n,
      streaming_movies = sum(Streaming.Movies) / n,
      streaming_tv = sum(Streaming.TV) /n,
      online_security = sum(Online.Security) / n,
      unlimitated_data = sum(Unlimited.Data) /n,
      Montly_mean=mean(Monthly.Charges), tenure_mean=mean(Tenure.Months),
      age_mean = mean(Age))
```

```
##   km_4   n   Partner senior_citizen online_backup tech_support streaming_movies
## 1    1 314 0.2834395     0.00955414     0.2484076   0.16560510       0.69745223
## 2    2 312 0.3333333     0.32692308     0.2147436   0.09935897       0.01602564
## 3    3 312 0.7756410     0.04487179     0.7852564   0.59294872       0.78846154
## 4    4 211 0.5308057     1.00000000     0.4786730   0.24170616       0.80568720
##   streaming_tv online_security unlimitated_data Montly_mean tenure_mean
## 1   0.74203822       0.1369427        0.8312102    90.91736    20.60828
## 2   0.03205128       0.1666667        0.8653846    75.81651    15.86538
## 3   0.80128205       0.5320513         0.8621795  103.43446    56.44231
## 4   0.74881517       0.2369668         0.8530806   97.05521    38.61611
##   age_mean
## 1 43.53503
## 2 51.45513
## 3 41.76603
## 4 72.26540
```

Produce Insights

a) You need to decide whether to use the result from Hierarchical clustering or the results from K-menas.
A. I think the result we obtained from Hierarchical clustering is more meaningful and interesting. This is because in hierarchical clustering, the 4 different clusters represents different demographic and user information.

b) Explain in detail, how the individual customers contained in solution, differ from each other.
A. The 1st and 4th clusters have a mean age of over 72, but while the 1st one has tenure_mean of 18, the 4th one has tenure_mean of 50. By analyzing the differences between these two clusters, we might find the reason behind this big gap. This applies to clusters 2 and 3. While clusters 2 and 3 have a mean age of around 41-42, cluster 2 have tenure_mean of 51, and cluster 3 have a tenure mean of 19. Cluster 1-4 and 2-3 have differences in whether they have a partner too. While there are only 37% of people in cluster 1 with a partner, there are 64% of the people with a partner in cluster 4. This relationship also applies in clusters 2 and 3.

c) The only noticeable difference between clusters 2 and 4 is the age difference. While cluster 2 has a mean age of 41, cluster 4 has a mean age of 72. However, for other attributes, they are pretty similar in the way that they have been using this company for a long time, and they utilize the service provided by the company. It seems like a lot of individuals in these two clusters subscribe to online_backup and unlimited data. Also, they take advantage of free movies and tv streaming by the company. Therefore, the company could introduce a plan geared toward people who have been using the company for a more extended period. The plan could include online backup, movie and TV streaming, and unlimited data. I think the price for this plan could be higher because we know from the table that people in cluster 2 and cluster 4 pay more monthly fees than other people in different clusters.
The company could also introduce another plan for individuals in cluster 1 and cluster 3. These are the people who have not been using this company for a long time. The interesting thing about them is that only 35% - 43% of them take advantage of the company's free movie and tv streaming. Also, most of the people in the cluster do not pay for additional services besides unlimited data. Thus, the company could introduce a plan that only includes unlimited data and an optional movie and tv streaming service for people in clusters 1 and 3.