CLUSTERING:
K-MEANS VS HIERARCHICAL CLUSTERING

UNIVERSITÀ
DEGLI STUDI
DI MILANO

STUDENT NAME: **SAINEY MANGA**
MAT: **943874**
PROGRAM: **MASTER'S IN DATA SCIENCE AND ECONOMICS**
COURSE CORDINATOR: **PROFESSOR SILVIA SALLINI**

CLUSTERING:
K-MEANS VS HIERARCHICAL CLUSTERING

ABSTRACT

This report seeks to analyze the performances of the two algorithms of clustering; k-means and hierarchical clustering.

In the first section, we applied the k-means algorithm to compare customer's annual income and there spending score on the mall-customer dataset. We notice that the distribution of annual income is skewed to the left with a mean of 60.56 and a maximum of 137.00 whereas the spending score is normally distributed with a mean of 50.20 and a maximum of 99.00.

On the k-means clustering, we use the elbow method in order to determine the optimal number of clusters we used. We prefer to use five (5) clusters as the slope at the fifth elbow and the sixth elbow did not change much.

With the five clusters, there are two groups of customers who have a high spending rate and two with the lowest spending. One group of customers is in the medium category in both spending and annual income. In the high spending customers, one cluster earns below 50 thousand dollars annually whilst the other cluster earns above 75.

On the hierarchical clustering section, we choose six (6) optimal clusters from the dendrogram.

From the clusters shown, both cluster one, two and three are low income earners although cluster one spends more than both two and three. Cluster four falls under medium income earners and spends moderately whilst the remaining two clusters fall under high income earners although cluster five has a low spending rate.

STATEMENT OF THE PROBLEM OR GOAL:

Identifying the income class of customers that has the highest spending rate at the Mall. The goal of this report is to find segments of different groups of customers who have similar spending rate base on their income level, the idea is to presumably advise the Mall's management on how to strategies their marketing foresight and direction.

The dataset named Mall-customer dataset was downloaded from this link https://www.kaggle.com/shwetabh123/mall-customers/data . It has five variables with 200 rows and no missing instances. The variables are "customer id" representing the identification of a customer entering the Mall, gender of the customer, customer's age, Annual income representing the customers yearly earnings and spending score as a weight allocated on a customer's spending at the Mall. Our main focus is on the spending score and Annual income earning of the customer.

KEY FINDINGS:

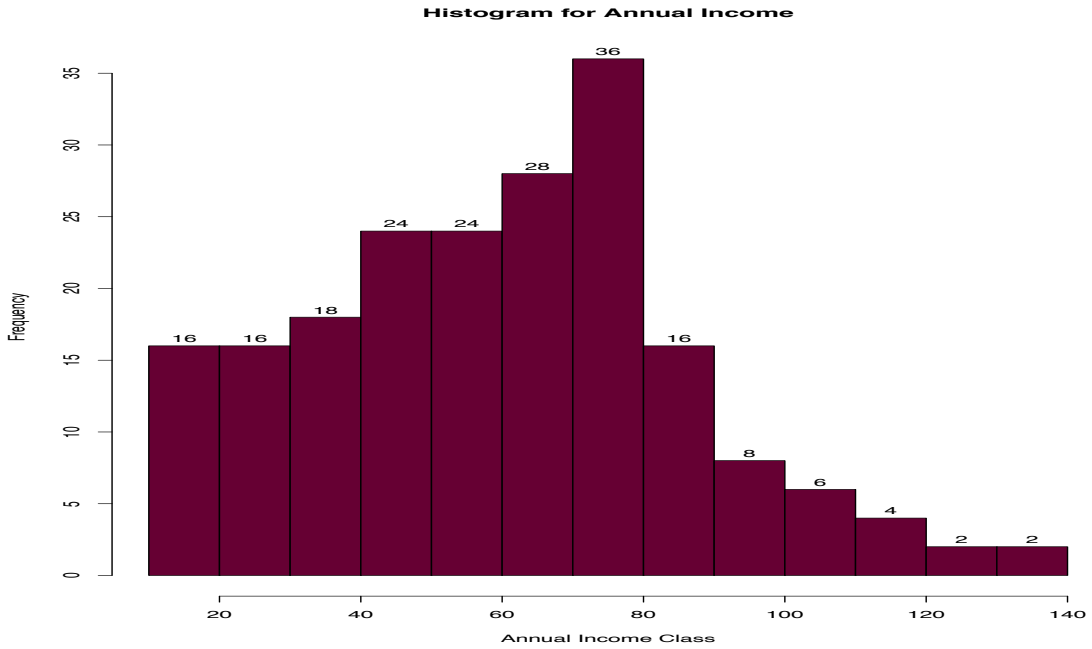
I. The k-means clustering algorithm shows that customers whose earnings is below 50 thousand and above 75 thousand are those who have the highest spending score.
II. The hierarchical clustering algorithm slightly differs, it shows the highest earning customers having the highest spending score.
III. Both algorithms do not have any major differences in terms of providing interpretable clusters of segments.

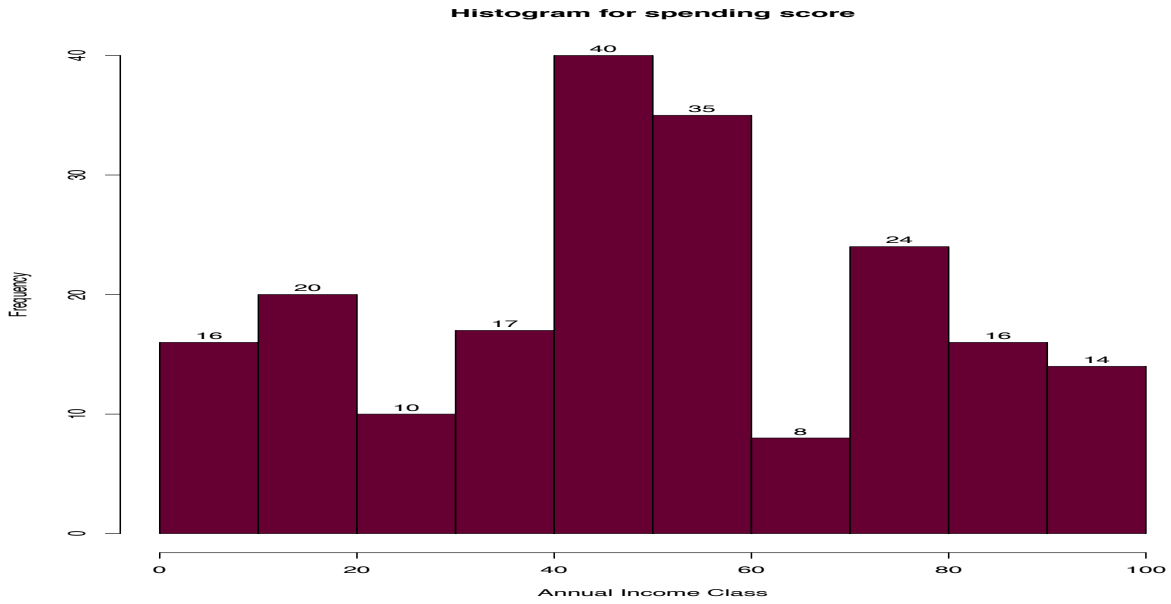
CLUSTERING:
K-MEANS VS HIERARCHICAL CLUSTERING

In the section below, we would show results corresponding to our key findings and the techniques employed:

- **MAIN RESULTS:**

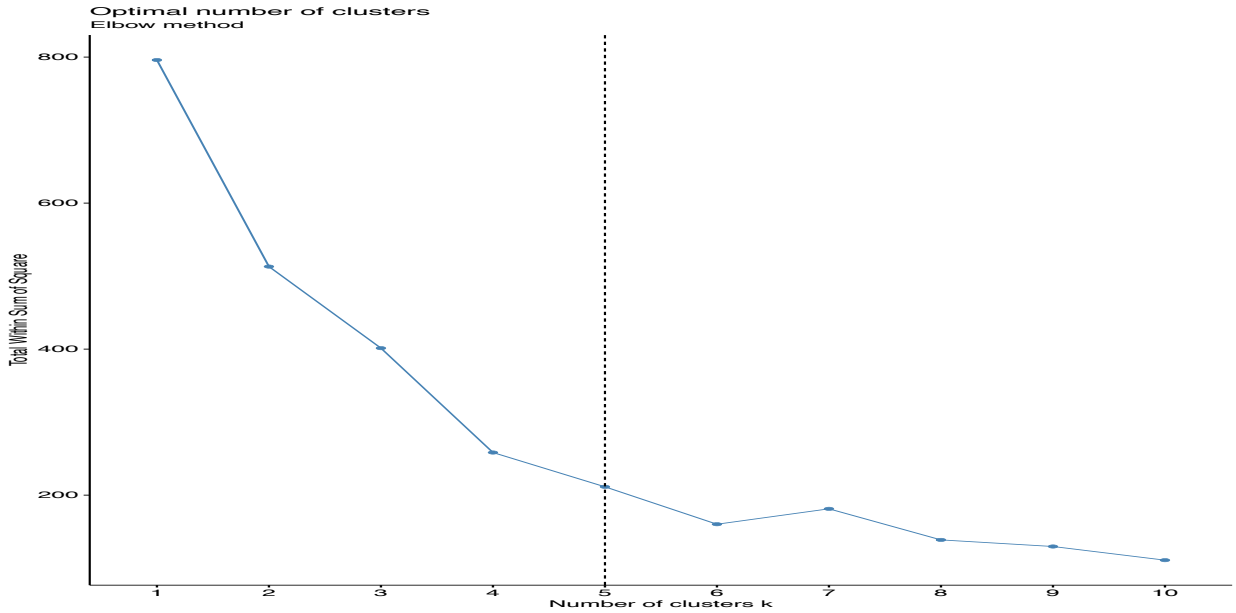
On our variables of interest, Annual income's distribution is skewed to the left with a mean of 60.56 and a maximum of 137.00, whilst the spending score variable is normally distributed with a mean of
50.20 and a maximum of 99.00.

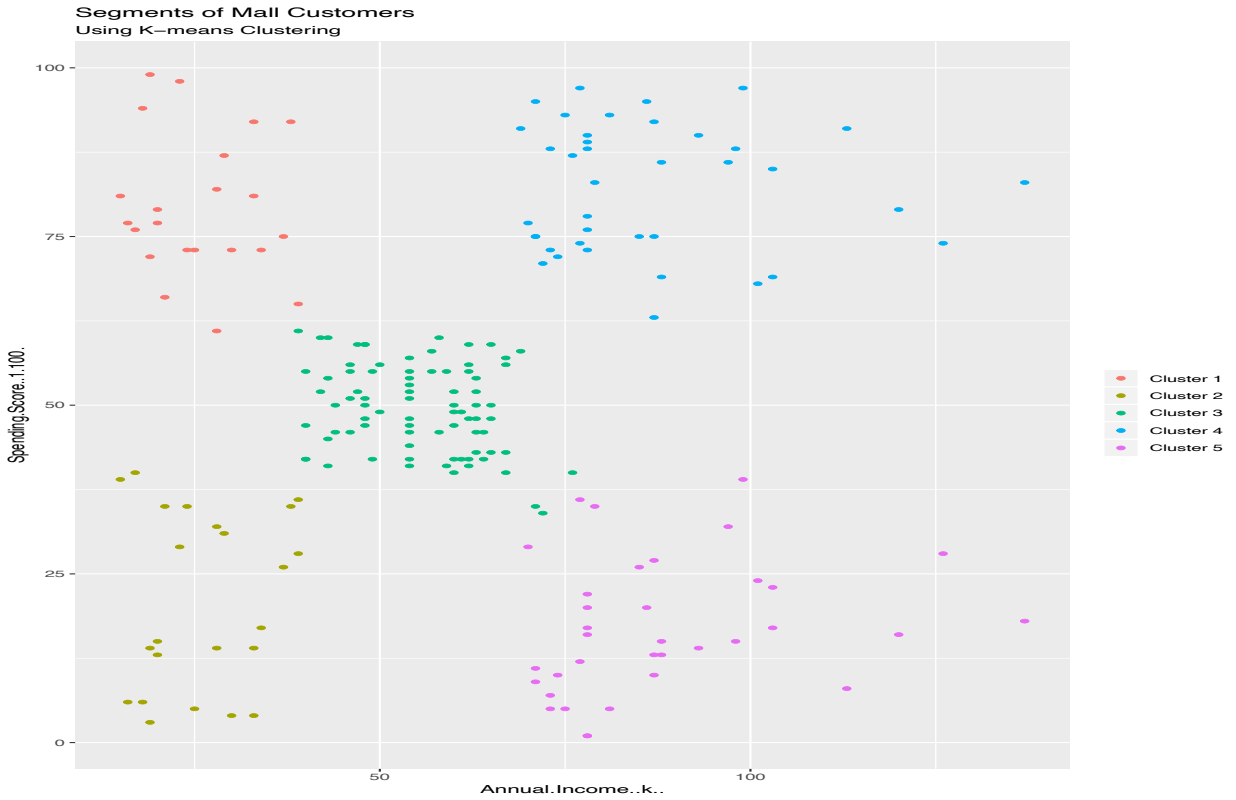




CLUSTERING:
K-MEANS VS HIERARCHICAL CLUSTERING

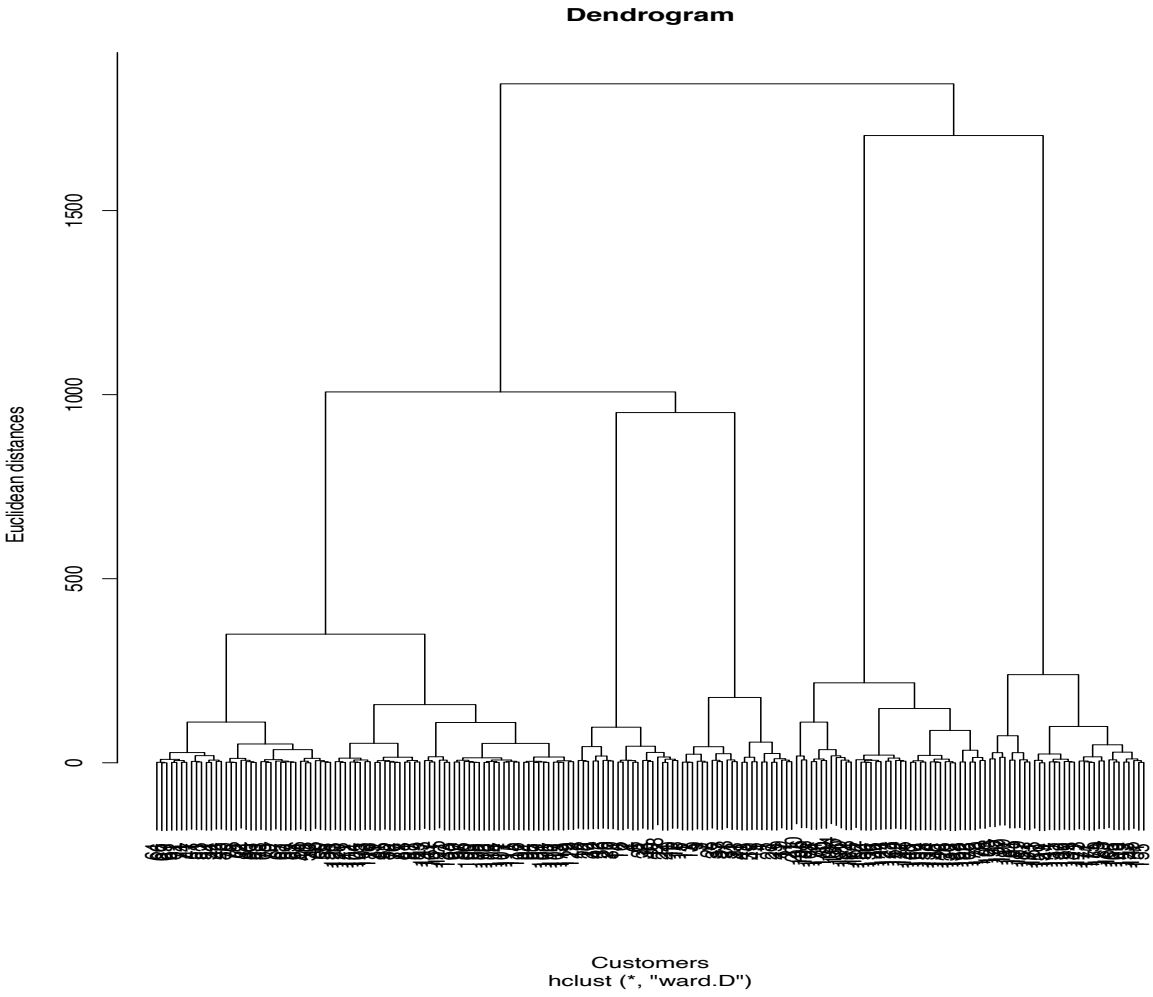
I. In applying the k-means clustering algorithm, we use the elbow method to determine the optimal number of clusters we would use to segment our data into clusters. Looking at the diagram, one could choose k = 6 clusters, but we prefer to take k=5 clusters as you can notice that the slope did not change much from elbow 5 to 6.





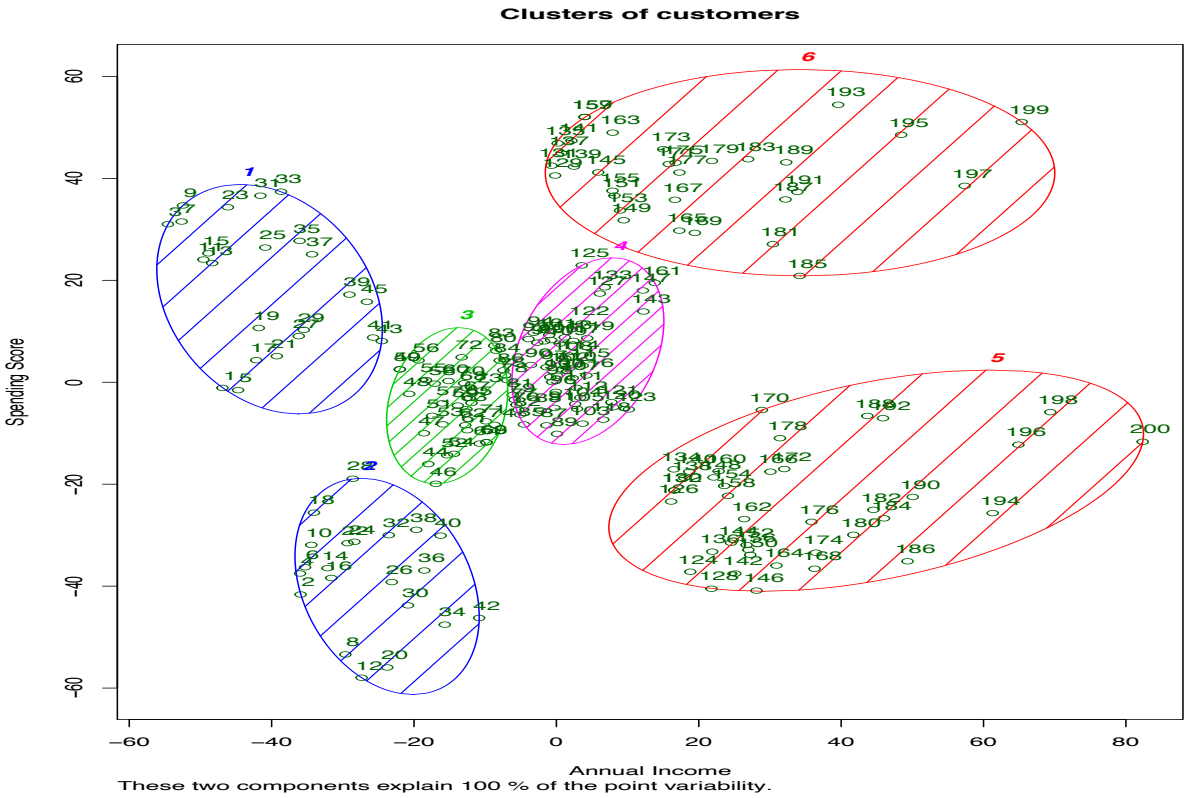
Having determine the number of optimal clusters, the diagram above shows the clusters obtained through the k-means algorithm. Both cluster 1 and cluster 2 are in the lowest income earning category, but cluster 1 has in the highest rank of spending score. Cluster 4 and 5 are the highest income category, but cluster 4 stands out better in terms of spending score. Cluster 3 stands out alone as it falls in medium category for both annual income and spending score. In this regard, the Mall would target cluster 1 customers, cluster 3 and cluster 4 with a market storm on cluster 4 as they have the highest annual earnings.

II. In order to apply the hierarchical clustering algorithm, we produce a dendrogram from the dataset using the ward.D method to determine the optimal number of clusters. The ward.D method as we know tries to minimize the within cluster variance. Like the elbow method, the dendrogram shows a similar number of optimal clusters, we choose 6 six clusters as could be seen below



Now that we have determine the number of optimal clusters to use, we now fit our dataset in to our hierarchical clusters as can be seen in the next page:

CLUSTERING:
K-MEANS VS HIERARCHICAL CLUSTERING



The clusters produce from the hierarchical algorithm slightly differs from the k-means clustering algorithm. Cluster 6 is found in the uppermost segment in the diagram, it composes of customers with the highest annual income and the highest spending score. Cluster 1 is at the bottom of income level with a high spending score just below cluster 6. Clusters 2 and 5 are at the bottom of the spending score with cluster 5 at the highest level of annual income. Both cluster 3 and 4 are in the medium annual income level and medium spending score. In light of the information above, the Mall would prefer customers of clusters 6, 1, 4 and 3 respectively.

CONCLUTION:

Base on the information generated from both algorithms, we would conclude that a larger chunk of customers between medium to high income earning level actually spend more than the lower earning class. However, those in the lower income level also have high spending score, but as any business owner would argue, the higher income earning class would contribute a larger magnitude of revenue to the Mall.

In terms of performances, the report holds the opinion that both algorithms did not produce any major difference in the results, the end product is similar. Therefore, if we were to advise the Mall's marketing department, we outline that they do more marketing on the medium and higher earning customers.

CLUSTERING:
K-MEANS VS HIERARCHICAL CLUSTERING

APPENDICES:

```
>mall_customer<-
read.csv('/Users/saineymanga/Desktop/DSE/datasets_7721_10938_Mall_Customers.csv')
>head(mall_customer)
>attach(mall_customer)
>any(is.na(mall_customer))
>install.packages("factoextra")
>library(factoextra)
>install.packages("NbClust")
>library(dplyr)
>library(stats)
>library(NbClust)
>library(ggplot2)
>library(cluster)
### SHOW THE DISTRIBUTION OF SPENDING SCORE AND ANNUAL INCOME
>hist(mall_customer$Annual.Income..k..,
    main="Histogram for Annual Income",
    col="#660033",
    xlab="Annual Income Class",
    ylab="Frequency",
    labels=TRUE)

>hist(mall_customer$Spending.Score..1.100.,
    main="Histogram for spending score",
    col="#660033",
    xlab="Annual Income Class",
    ylab="Frequency",
    labels=TRUE)

## DROP THE GENDER COLUMN TO REMOVE NON NUMERIC INSTANCES
>my_data<- select(mall_customer,c(1,3,4,5))
## SCALING THE DATA
df<-scale(my_data)

## FINDING THE NUMBER OF OPTIMAL CLUSTERS USING THE ELBOW METHOD
>fviz_nbclust(df, kmeans, method = "wss") +
  geom_vline(xintercept = 5, linetype = 2) + # add line for better visualisation
  labs(subtitle = "Elbow method") # add subtitle

### PLOTING THE CLUSTERS
>set.seed(1)
>k5<-kmeans(my_data[3:4],5,iter.max=100,nstart=50,algorithm="Lloyd")
>ggplot(my_data, aes(x =Annual.Income..k.., y = Spending.Score..1.100.)) +
  geom_point(stat = "identity", aes(color = as.factor(k5$cluster))) +
```

```
 scale_color_discrete(name=" ",
                      breaks=c("1", "2", "3", "4", "5"),
                      labels=c("Cluster 1", "Cluster 2", "Cluster 3", "Cluster 4", "Cluster 5")) +
 ggtitle("Segments of Mall Customers", subtitle = "Using K-means Clustering")

###### HIERARCHICAL CLUSTERING
## FINDING THE OPTIMALS NUMBER OF CLUSTERS USING THE DENDOGRAM
>dataset<-my_data[3:4]
>dendrogram = hclust(d = dist(dataset, method = 'euclidean'), method = 'ward.D')
>plot(dendrogram,
    main = paste('Dendrogram'),
    xlab = 'Customers',
    ylab = 'Euclidean distances')
## PLOTTING THE CLUSTERS
>hc = hclust(d = dist(dataset, method = 'euclidean'), method = 'ward.D')
y_hc = cutree(hc, 6)
clusplot(dataset,
       y_hc,
       lines = 0,
       shade = TRUE,
       color = TRUE,
       labels= 2,
       plotchar = FALSE,
       span = TRUE,
       main = paste('Clusters of customers'),
       xlab = 'Annual Income',
       ylab = 'Spending Score')
```