I have selected heart issues dataset from Kaggle.com. I have done initial analysis on the dataset and have implemented cluster and SVM models along with different plots and several other techniques on the dataset. The dataset has 319795rows and 18 columns. The dataset has several attributes like BMI, SMOKING, ALCOHOL, STROKE, PHYSICAL HEALTH, MENTAL HEALTH.

Analysis:

I have initially loaded the required libraries for the overall analysis. I have loaded the dataset and checked the structure of the dataset.

```
> str(heart_disease)
'data.frame':   319795 obs. of  18 variables:
 $ HeartDisease    : chr  "No" "No" "No" "No" ...
 $ BMI             : num  16.6 20.3 26.6 24.2 23.7 ...
 $ Smoking         : chr  "Yes" "No" "Yes" "No" ...
 $ AlcoholDrinking : chr  "No" "No" "No" "No" ...
 $ Stroke          : chr  "No" "Yes" "No" "No" ...
 $ PhysicalHealth  : num  3 0 20 0 28 6 15 5 0 0 ...
 $ MentalHealth    : num  30 0 30 0 0 0 0 0 0 0 ...
 $ DiffWalking     : chr  "No" "No" "No" "No" ...
 $ Sex             : chr  "Female" "Female" "Male" "Female" ...
 $ AgeCategory     : chr  "55-59" "80 or older" "65-69" "75-79" ...
 $ Race            : chr  "White" "White" "White" "White" ...
 $ Diabetic        : chr  "Yes" "No" "Yes" "No" ...
 $ PhysicalActivity: chr  "Yes" "Yes" "Yes" "No" ...
 $ GenHealth       : chr  "Very good" "Very good" "Fair" "Good" ...
 $ SleepTime       : num  5 7 8 6 8 12 4 9 5 10 ...
 $ Asthma          : chr  "Yes" "No" "Yes" "No" ...
 $ KidneyDisease   : chr  "No" "No" "No" "No" ...
 $ SkinCancer      : chr  "Yes" "No" "No" "Yes" ...
>
```

Using structure function we get to know, the various categorical and numerical variables and also the overall structure of the dataset.

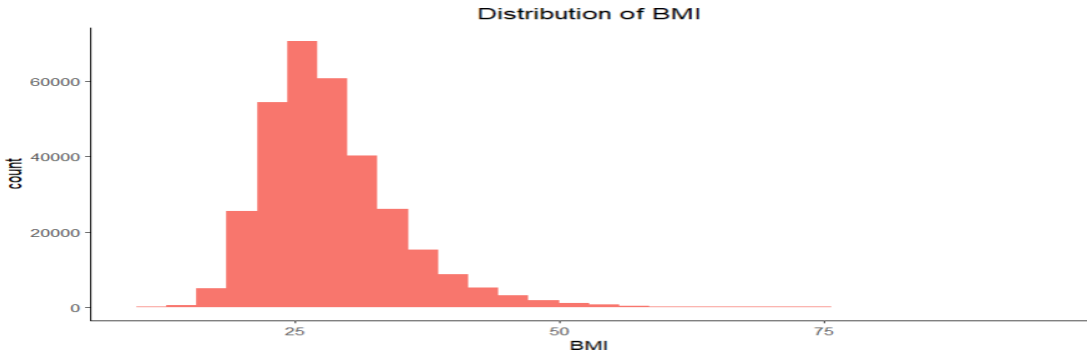
```
> summary(heart_disease)
 HeartDisease            BMI           Smoking           AlcoholDrinking       Stroke
 Length:319795      Min.   :12.02   Length:319795      Length:319795      Length:319795
 Class :character   1st Qu.:24.03   Class :character   Class :character   Class :character
 Mode  :character   Median :27.34   Mode  :character   Mode  :character   Mode  :character
                    Mean   :28.33
                    3rd Qu.:31.42
                    Max.   :94.85
 PhysicalHealth    MentalHealth    DiffWalking           Sex             AgeCategory
 Min.   : 0.000   Min.   : 0.000   Length:319795      Length:319795      Length:319795
 1st Qu.: 0.000   1st Qu.: 0.000   Class :character   Class :character   Class :character
 Median : 0.000   Median : 0.000   Mode  :character   Mode  :character   Mode  :character
 Mean   : 3.372   Mean   : 3.898
 3rd Qu.: 2.000   3rd Qu.: 3.000
 Max.   :30.000   Max.   :30.000
     Race             Diabetic          PhysicalActivity    GenHealth
 Length:319795      Length:319795      Length:319795      Length:319795
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character


   SleepTime         Asthma          KidneyDisease       SkinCancer
 Min.   : 1.000   Length:319795      Length:319795      Length:319795
 1st Qu.: 6.000   Class :character   Class :character   Class :character
 Median : 7.000   Mode  :character   Mode  :character   Mode  :character
 Mean   : 7.097
 3rd Qu.: 8.000
 Max.   :24.000
>
```

Summary function gives us the minimum, maximum, 1st quartile, 3rd quartile and mode.

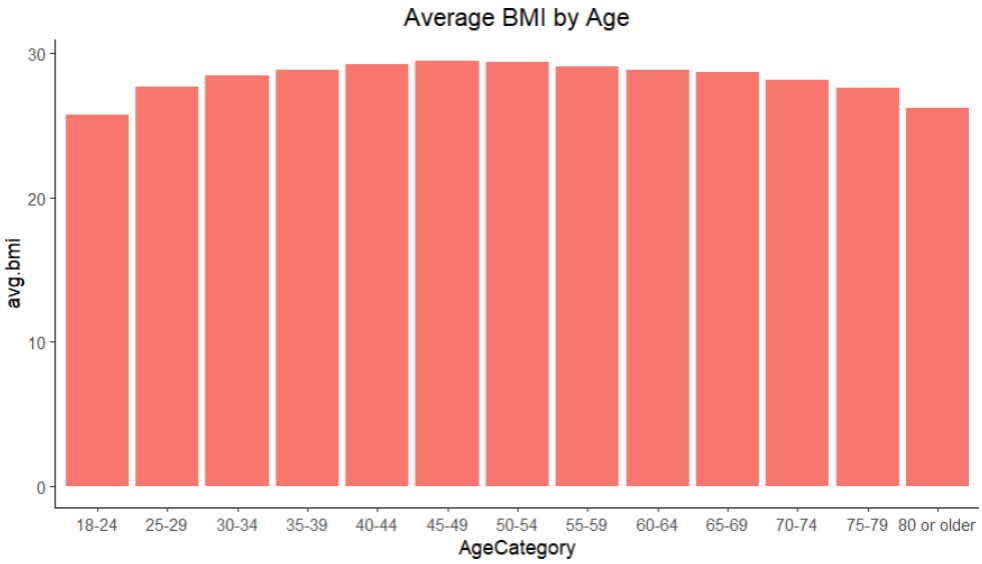
The rows and columns of all the dataset will be known to us by using nrow and ncol;

```
> nrow(heart_disease)
[1] 319795
> ncol(heart_disease)
[1] 18
> 
```

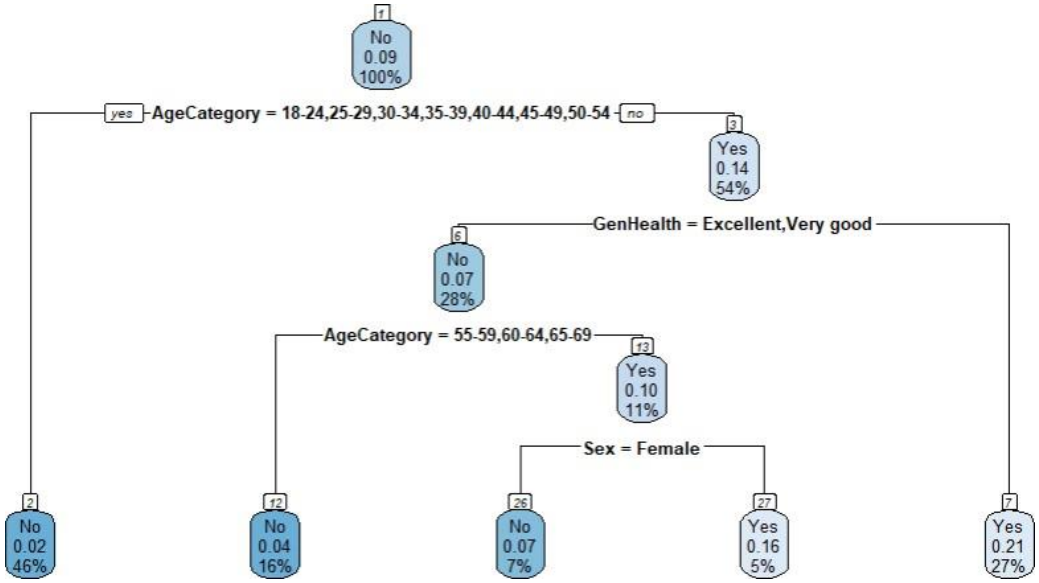
The plot of BMI and count tells us that most of the people lie between 18-30 BMI. At 25-26 BMI levels we can see the highest count.



Here the age category and Average Bmi also shows the same results and we can identify most of the people from all age categories fall between 20-30 BMI.

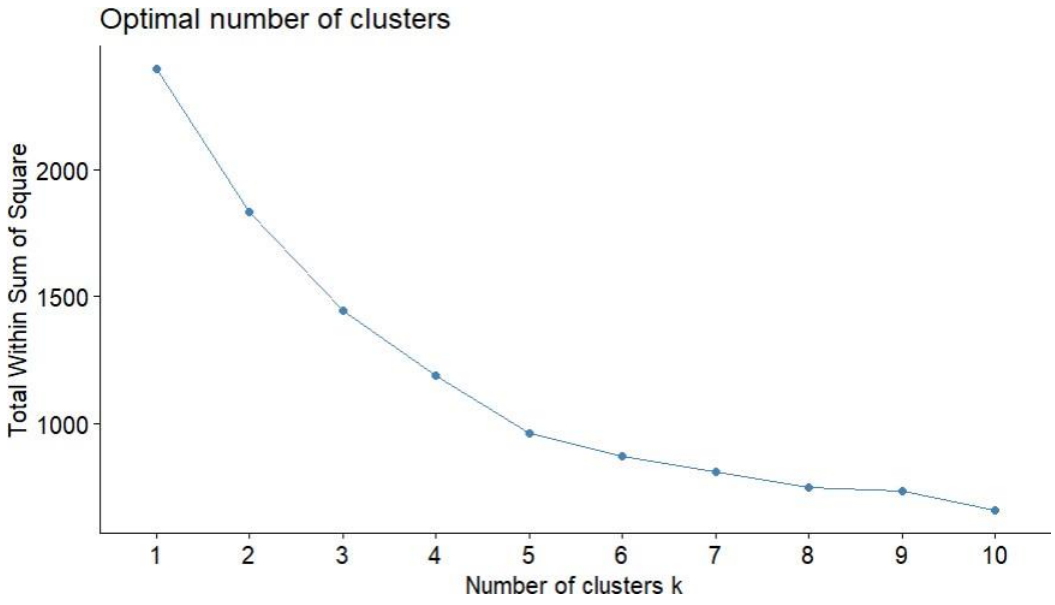


This is the decision tree for the dataset which explains that initially we have taken a large number of age category and check how many of them are in good health condition and further determine their sex and the percentage of people falling under the condition of health issues.

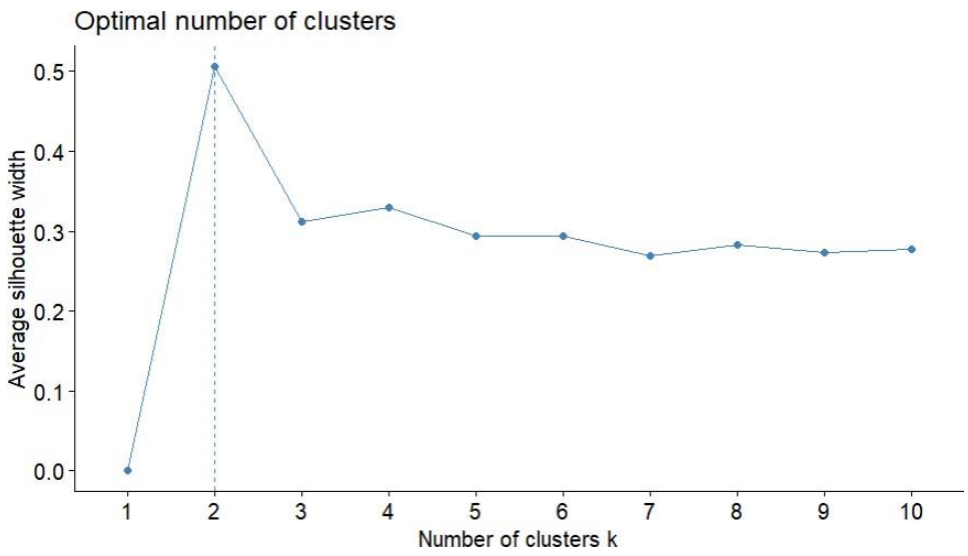


In order to do the clustering analysis, we have to know the optimal number of k for the clusters, and we use two methods one is elbow method and the other one being silhouette method.

The below is the elbow method curve:

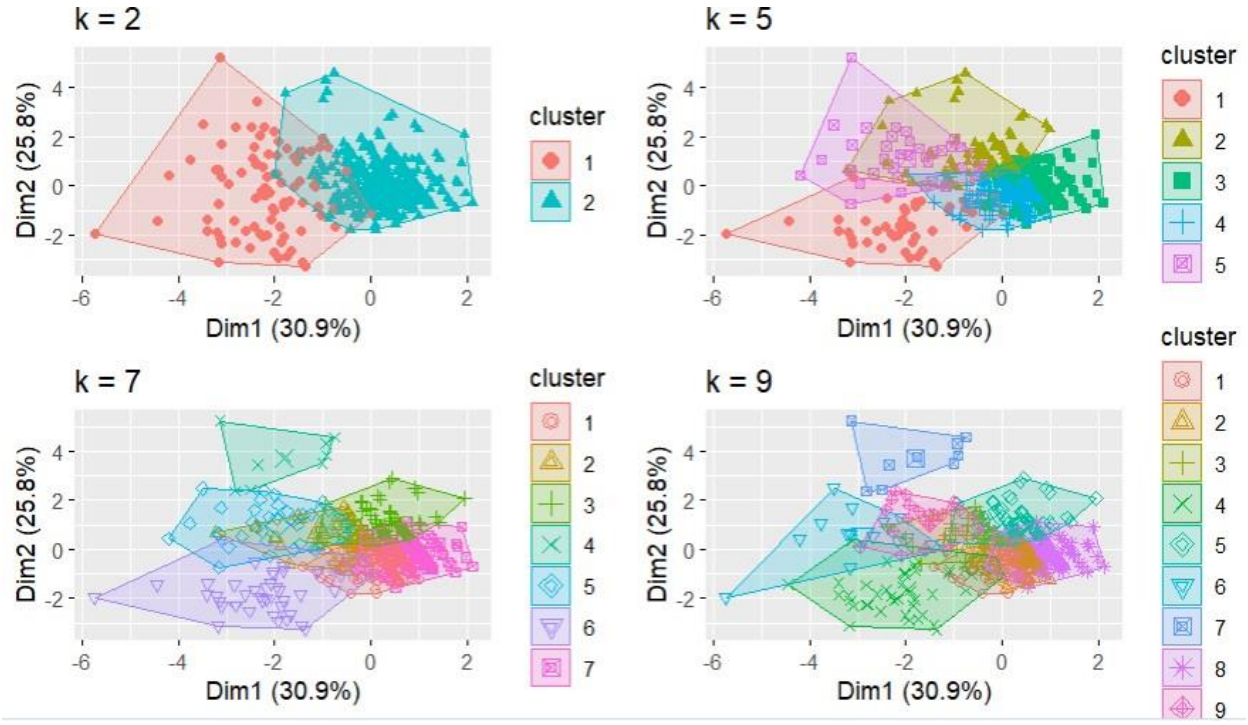


Silhoutte method graph is as below, the silhouette range varies between -1 to +1



The below is the clustering which is the conglomeration of 4 individual clusters with k = 2,5,7 and 9.

Here in our method, the clusters are overlapping there is no perfect distinction. Out of all the graphs we identify that k=2 has some clarity over the others.



The mean parameter across clusters for various attributes is as below, here we find that Sleep time and BMI does not have much difference but attributes like physical health and mental health are showing a huge variation.