**Ramaiah Institute of Technology**

(An Autonomous Institute, Affiliated to VTU)

MSR Nagar, MSRIT post, Bangalore-54

A Dissertation Report on

Liver Function Test Results

Submitted by

Rachana H 1MS15CS096

Sadhana Ravoori 1MS15CS104

Shravani Nindra 1MS15CS116

*Bachelor of Engineering in Computer Science & Engineering*

Under the guidance of

Sowmya B J

Assistant Professor

CSE department

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**M.S.RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

**BANGALORE-560054**

www.msrit.edu, **2017**

**INTRODUCTION**

K-means is one of the simplest unsupervised learning algorithms. This algorithm is used when we have unlabelled data that is data without defined categories. K-means follows an easy and simple way to classify the given data set through a certain number of clusters(k clusters) which can be determined by the analyst. The main idea is to define k centroids one for each cluster. The centroids should be placed in such a way that they are as much as possible far away from each other because different locations causes different results. Next step is to take each point or tuple belonging to the given data set and then associate it to the nearest centroid. Each time a point is inserted into a cluster the k centroid values are recalculated. When there is no point pending this step is complete and grouping is done.

K-means algorithm has its own advantages and disadvantages. K-means is computationally faster compared to the hierarchical clustering methods and it also provides tighter clusters. The drawbacks of this algorithm is that it is very hard to decide on a numerical value for ‘k’. K-means algorithm is primarily based on iteratively calculating the cluster means and it is known to us that mean value is very sensitive to outliers present in the dataset. Hence, K-means does not provide with an accurate clustering methodology.

Our main objective through this project is to apply k-means algorithm on the liver data set which we have procured and cluster the dataset using various parameters. The liver data set has attributes where each of them indicate the Liver Function Test results on various test subjects ,both male and female, across all age groups. Post clustering, graphs are plotted against any of the various attributes present and meaningful inferences can be drawn out of it.

**DATA SET DESCRIPTION**

Source of Dataset

We found this dataset from Kaggle website.

URL: https://www.kaggle.com/uciml/indian-liver-patient-records

Attributes Description

AGE

Age attribute describes the age of the subject. Age ranges from 4-85 years old in this dataset.

GENDER

Gender attribute describes whether the subject is male or female.

TOTAL BILIRUBIN

The total bilirubin test is used to diagnose conditions like liver disease, haemolytic anaemia and blocked bile ducts. Most bilirubin comes from the body’s normal process of breaking down old red blood cells. A healthy liver can normally get rid of bilirubin. Normal values of total bilirubin range from 0.3-1.0 mg/dL.

DIRECT BILIRUBIN

Conjugated, or direct, bilirubin travels freely through your bloodstream to your liver. Most of this bilirubin passes into the small intestine. A very small amount passes into your kidneys and is excreted in your urine. This bilirubin also gives urine its distinctive yellow colour. This test is usually done to look for liver problems, such as hepatitis, or blockages, such as gallstones.

ALKALINE PHOSPHATASE

An alkaline phosphatase level test measures the amount of alkaline phosphatase enzyme in your bloodstream. Abnormal levels of ALP in your blood most often indicate a problem with your liver, gallbladder, or bones. The normal range of ALP varies from person to person and depends on your age, blood type, gender, and whether you’re pregnant.

ALANINE AMINOTRANSFERASE

Alanine aminotransferase (ALT) is an enzyme found primarily in theliver and kidney.The reference range for ALT is 20-60 IU/L. ALT is an enzyme made by cells in your liver. Your body uses ALT to break down food into energy. Normally, ALT levels in the blood are low. If your liver is damaged, it will release more ALT into your blood and levels will rise.

ASPARTATE AMINOTRANSFERASE

Aspartate transaminase or aspartate aminotransferase is a pyridoxal phosphate (PLP)-dependent transaminase enzyme. AST is found in the liver, heart, skeletal muscle, kidneys, brain, and red blood cells. AST is an enzyme produced by the liver and injected into the bloodstream when liver is damaged.

TOTAL PROTEINS

The total protein test measures the total amount of two classes of proteins found in the fluid portion of your blood. The two classes are albumin and globulin. Proteins are important parts of all cells and tissues. Albumins helps prevent fluid from leaking out of blood vessels .Globulins are an important part of your immune system.

ALBUMIN

The albumins are a family of globular proteins, the most common of which are the serum albumins. Albumin is made mainly in the liver. It helps keep the blood from leaking out of blood vessels.

ALBUMIN AND GLOBULIN RATIO

Laboratories report total protein, albumin, and the calculated ratio of albumin to globulins, termed the A/G ratio. The A/G ratio is calculated from measured total protein, measured albumin, and calculated globulin (total protein - albumin). Because disease states affect the relative amounts of albumin and globulin, the A/G ratio may provide a clue as to the cause of the change in protein levels. The albumin to globulin (A/G) ratio has been used as an index of disease state.

DATASET

This attribute is used to split the data into two sets i.e ones who have liver diseases and ones who do not have any liver diseases.

Data Set size in terms of Bytes and Number of Tuples

The following data set size is of size 23.3 kB. This dataset consists of liver function test results of 583 patients out of which 416 patients are known to have been diagnosed with some form of liver disease whereas the remaining 167 patients do not have any liver defects. This comprises of 441 male patient records and 142 female patient records.

**ALGORITHM DESCRIPTION**

K-Means is an iterative Algorithm where K represents the number of clusters the dataset is going to be partitioned into.

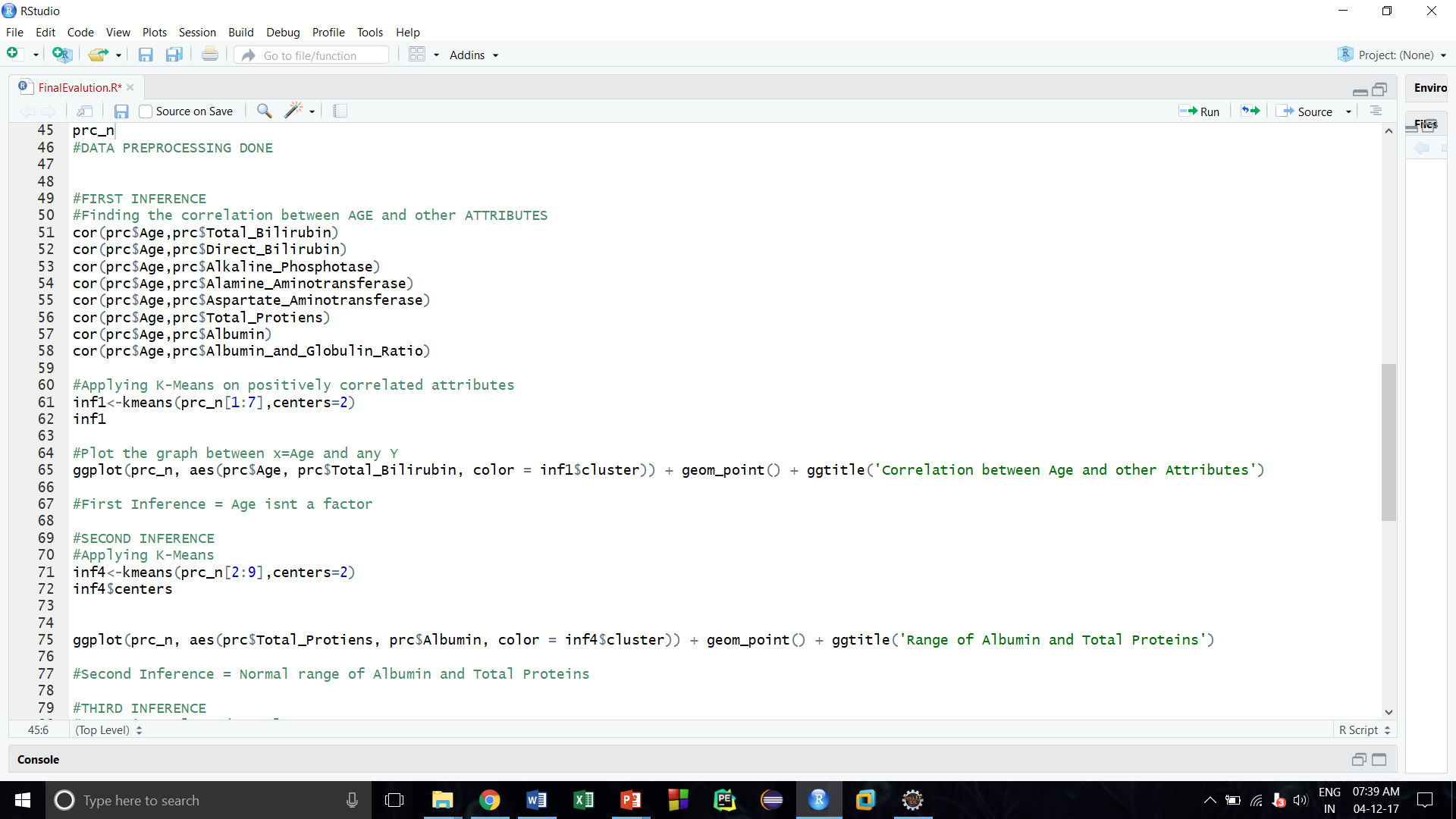
Working of k-means algorithm

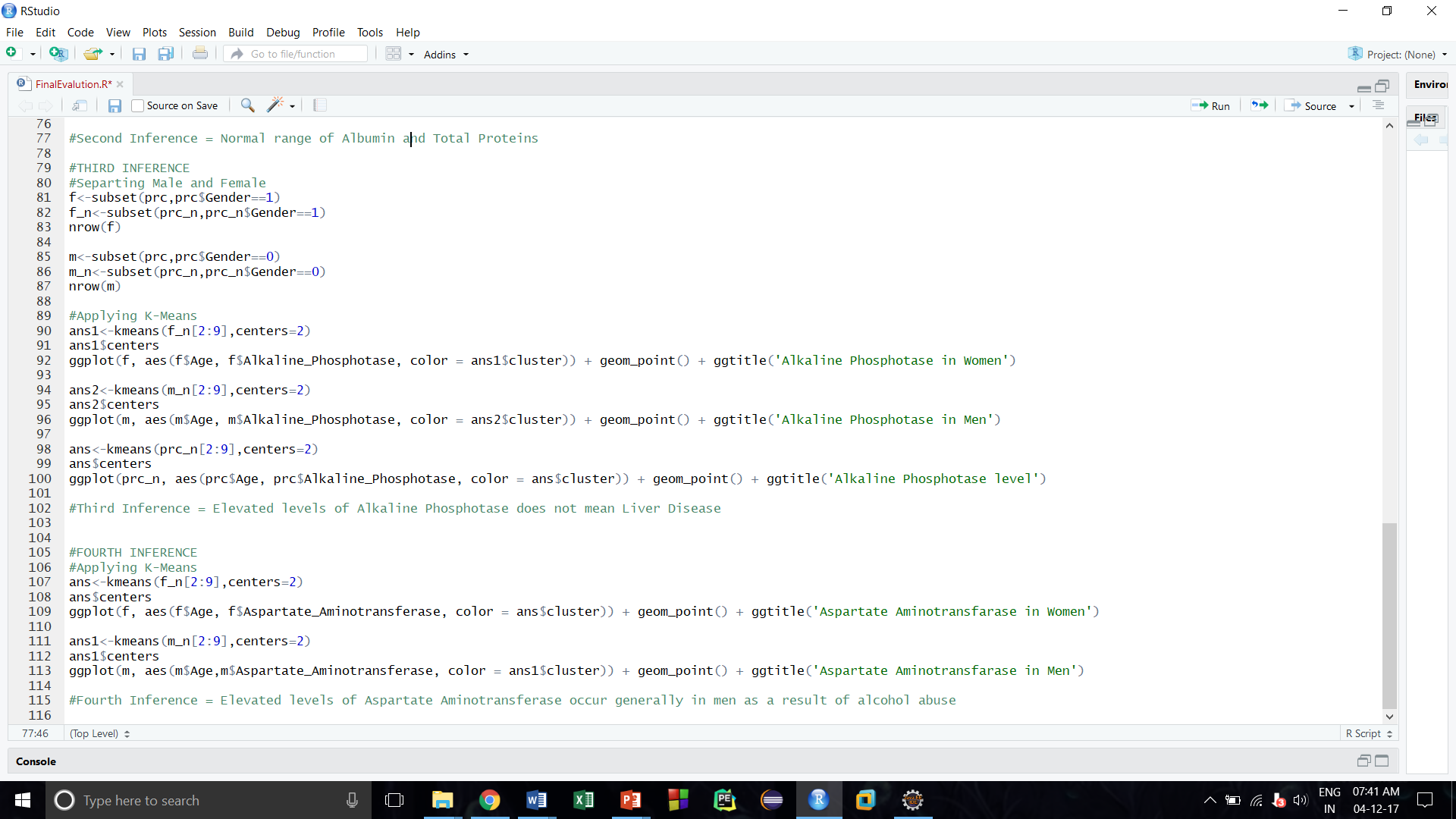
* Randomly choose centers for each of the k clusters.
* The distance between a data point and each of these k centers is found using the Euclidean distance formula.
* This is done for all the data points and the data points are assigned to the cluster to which it has the least distance from.
* The centers are recalculated for each of these k clusters and once again the distance between the data points and each of these k centers is calculated and assigned to a cluster.
* The algorithm stops when the data points are assigned to the same cluster in 2 consecutive iterations.

**INFERENCES**

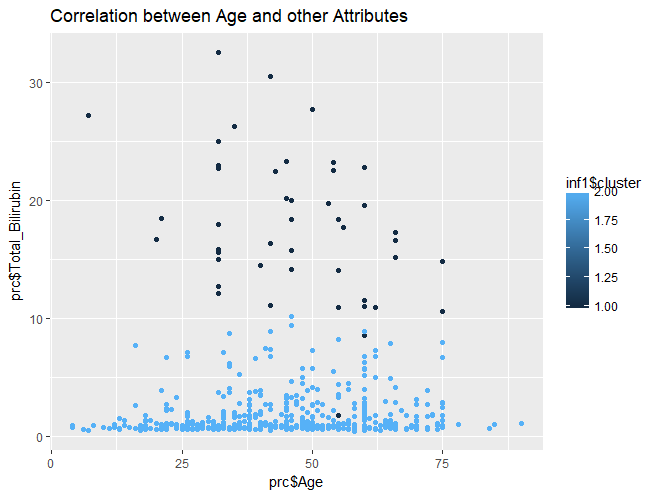
* The first inference could be drawn by plotting Age against all the other attributes present in the dataset. It is clearly seen that Age plays no role in the determination of the prevalence of any sort of liver disease in a subject. A liver is said to be not functioning properly due to the excess secretion of certain enzymes which leads to the variation in the levels in the blood. Age does not determine the excess secretion of enzymes. Hence, any person across all ages has a good probability of being diagnosed with a liver defect.
* The second inference is by plotting two strongly correlated attributes namely the albumin and total proteins. Through this it is seen that patients with albumin levels from 3.2-5 and total proteins levels from 6-9 have been grouped into one cluster signifying the patients who do not have liver disease whereas the other cluster depicts patients with levels outside the above mentioned range and have been diagnosed with liver disease.
* The third inference is obtained by splitting the dataset gender-based. It is known that elevated levels of Alkaline Phosphatase(ALP) is present in females. The reasons for this could be attributed to reduced bone density in females and pregnancies undergone. We plot a graph to indicate the ALP levels in the female subjects and it is clear that levels above the normal range does not indicate any liver defect whereas the same plot for male subjects speaks otherwise.
* The fourth inference is with respect to Aspartate Aminotransferase(AST) levels in the body. When liver damage is due to alcohol, it is seen that AST levels spike up and this could eventually lead to alcoholic hepatitis. From the graphs we can infer that the number of females affected are far less than the number of males affected. This implies that drinking habits prevail more amongst men than women.

**SNAPSHOT OF THE CODE**

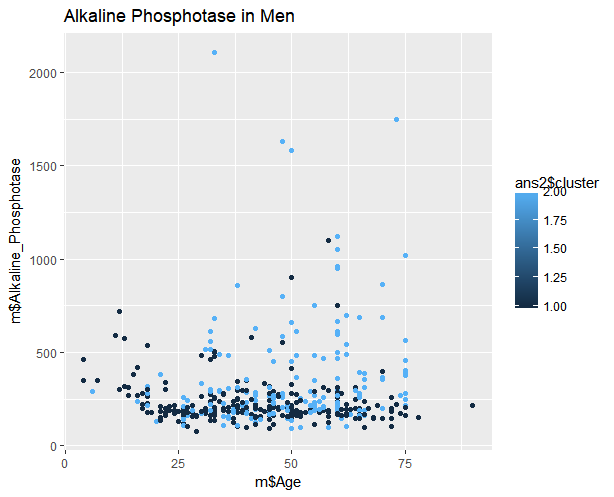




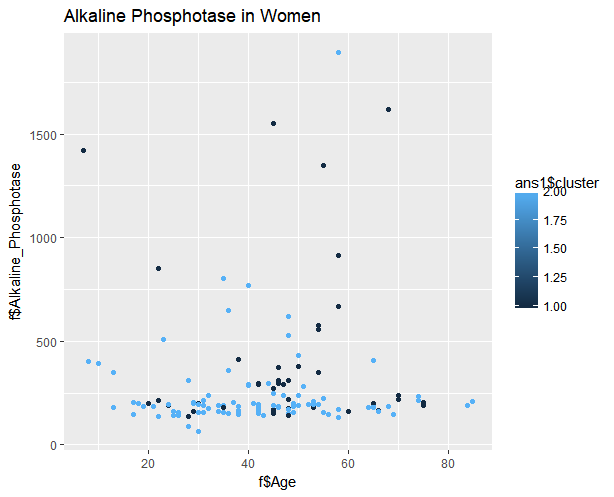
**RESULT SNAPSHOT AND ITS DESCRIPTION**



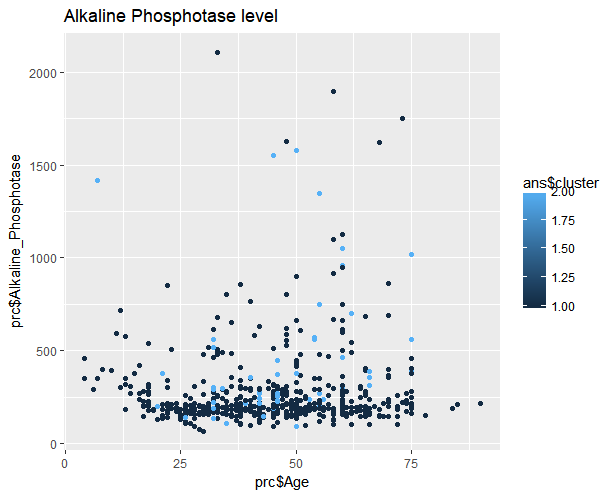
This graph depicts the plot of Age with Total Bilirubin which clearly shows that people of all ages are put into both the clusters which shows that age has no significance on the secretion of enzymes in the liver that is people of all ages secrete the same level of enzymes.



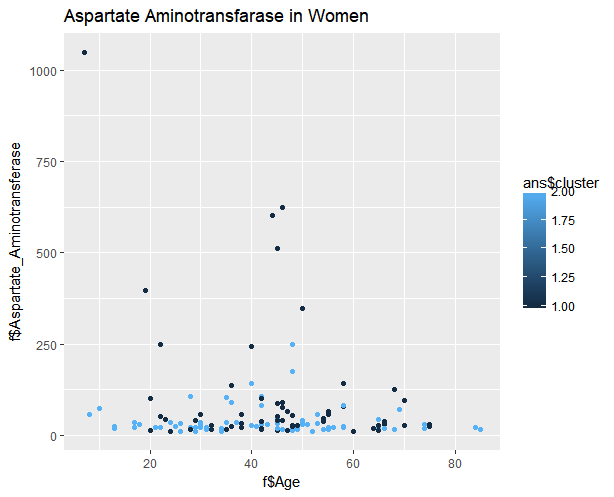
The above graph depicts the Alkaline Phosphatase levels against Age for all of the male subjects in the dataset. Though the Alkaline Phosphatase level is high between the ages of 0-25, it does not indicate that the person has liver disease as alkaline phosphatase is also secreted during growth spurts.



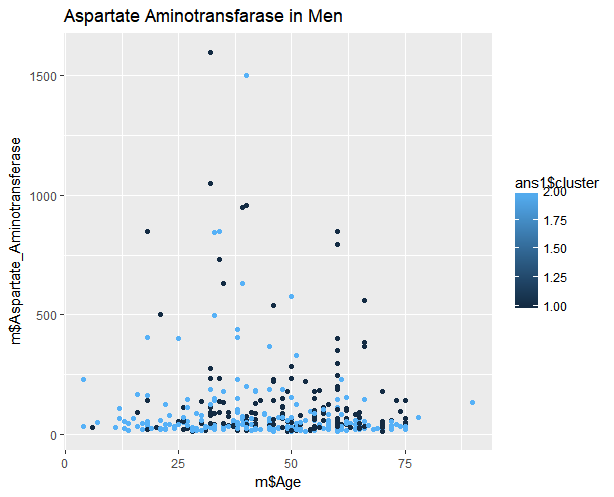
The above graph depicts the Alkaline Phosphotase levels against Age for all of the female subjects in the dataset. Though the Alkaline Phosphatase level is high between the ages of 20-40, it does not indicate that the person has liver disease as alkaline phosphatase is also secreted in women during pregnancy and even though the Alkaline Phosphatase level is high between the ages of 0-25, it does not indicate that the person has liver disease as alkaline phosphatase is also secreted during growth spurts.



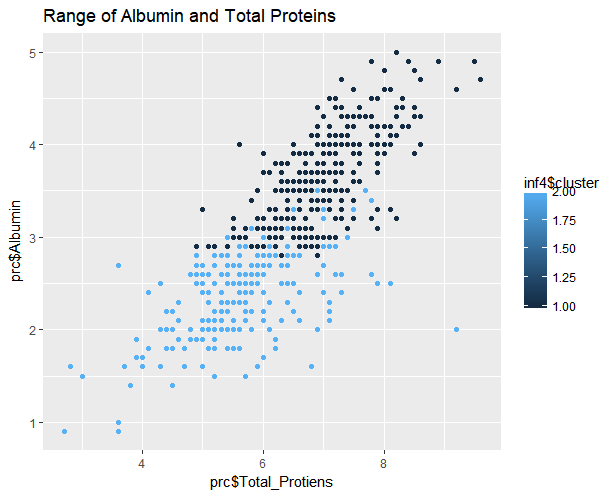
This graph depicts the plot of the Alkaline Phosphatase levels against Age for all test subjects (male and female). Thus Alkaline Phosphatase levels alone is not a clear indication as to whether the person has liver disease or not.



The above graph showcases the Aspartate Aminotransferase levels against Age for all the female subjects in the dataset.



The above graph showcases the Aspartate Aminotransferase levels against Age for all the male subjects in the dataset.



The above graph showcases the strongly correlated attributes namely the Albumin and Total Proteins against each other in the plot.

One of the clusters has a range of Total Proteins levels between 6 and 8.5 with Albumin levels between 3 and 5.5 while the other cluster has an indication of Total Proteins levels less than 6 with Albumin levels less than 3. Since total proteins as well as Albumin is secreted in the liver, the amount of its secretion decreases when the liver is damaged. Thus, the normal range of total proteins lies between 6 and 8.5 and Albumin levels between 3 and 5.5

**HOW TO IMPLEMENT IT?**

* The first step is to load the data set along with the necessary libraries like ggplot into the R workspace.
* The dataset undergoes data preprocessing before applying the algorithm.
* There are only a few missing data points in the attribute Albumin\_and\_Globulin\_Ratio. This is handled by inserting imputed values.
* The categorical attribute Gender is converted into numerical attribute by assigning 0 to Male and 1 to Female.
* The dataset is normalized by using Min-Max normalization.
* Outliers are removed by graphical method.
* Thus data preprocessing is complete.
* K-means algorithm is applied on a set of correlated attributes.
* Graphs are plotted using ggplot(x, y, color) where x and y are the attributes and color is the different colors assigned to the different clusters.
* Meaningful inferences are drawn from the graph.

**SOCIAL IMPACTS**

Through this project we are bringing to light the causes for liver defects in a certain group of people across various age groups. It can be clearly inferred that alcohol consumption plays a pivotal role in the elevated levels of certain enzymes in human body. This in turn could prove fatal leading to death.

* General public should be made aware of the root cause of various liver related diseases and should also be educated about the preventive measures regarding it.
* Government could also come up with certain measures to curb the alcoholism like open rehabilitation centres and provide necessary medical aid.