Hepatitis C Prediction

Sunil Kumar P

Dec 25 2021

Contents

1 Overview

1.1 About the Dataset

The data set contains laboratory values of blood donors and Hepatitis C patients and demographic values like age. The data was obtained from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/HCV+data

Attributes 1 to 4 refer to the data of the patient: 1) X (Patient ID/No.) 2) Category (diagnosis) (values: '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis') 3) Age (in years) 4) Sex (f,m)

Attributes 5 to 14 refer to laboratory data: 5) ALB 6) ALP 7) ALT 8) AST 9) BIL 10) CHE 11) CHOL 12) CREA 13) GGT 14) PROT

Quick reference on Hepatitis C:

Hepatitis C – an infection that can cause severe liver damage.

Fibrosis – build-up of collagen and other fibrous scar tissue, leading to a 'stiff' liver.

Cirrhosis – serious scarring that blocks blood flow through the liver, kills liver cells and interferes with liver function.

1.2 Objective

The target attribute for classification is Category (2): blood donors vs. Hepatitis C patients

The goal of the project is to identify Hepatitis C patients(including its progress ('just' Hepatitis C, Fibrosis, Cirrhosis)

1.3 Key Steps

Below are the Key steps that were performed for this project.

- Data Analysis of the Hepc dataset
  - Individual attributes
  - Each attribute's correlation with Category
  - Correlation of attributes among each other
  - Principal Component Analysis to identify Variance
- Partition Hepc into test and train sets
  - Data Cleansing
  - Applying various classification methods to train the train set, followed by predictions on the test set
  - Employing an Ensemble as the final approach to make a prediction

2 Analysis

We will begin our analysis by installing the pre-requisite libraries and loading the HepatitisC data from the web

2.1 Data Analysis

Let's Analyze the data and attributes available in the Hepc dataset. We will be using common machine learning techniques/algorithms to train a sample of data (called the train set) to generate predictions. These predictions are then compared against the remaining sample of data (called the test set).

To help us make a decision on the best model/method to predict, Accuracy of prediction will be compared across the methods.

Hepc is a data frame with 615 observations and 14 attributes.
Attributes 5 to 14 are the laboratory data and 1 to 4 are the patient's data

```
## [1] "Sample Observations from the Hepc Dataset"
```

```
##   X      Category Age Sex  ALB  ALP  ALT  AST  BIL   CHE CHOL CREA  GGT PROT
## 1 1 0=Blood Donor  32   m 38.5 52.5  7.7 22.1  7.5  6.93 3.23  106 12.1 69.0
## 2 2 0=Blood Donor  32   m 38.5 70.3 18.0 24.7  3.9 11.17 4.80   74 15.6 76.5
## 3 3 0=Blood Donor  32   m 46.9 74.7 36.2 52.6  6.1  8.84 5.20   86 33.2 79.3
## 4 4 0=Blood Donor  32   m 43.2 52.0 30.6 22.6 18.9  7.33 4.74   80 33.8 75.7
## 5 5 0=Blood Donor  32   m 39.2 74.1 32.6 24.8  9.6  9.15 4.32   76 29.9 68.7
## 6 6 0=Blood Donor  32   m 41.6 43.3 18.5 19.7 12.3  9.92 6.05  111 91.0 74.0
```

2.1.1 Data Analysis: Category

Category attribute is not numerical

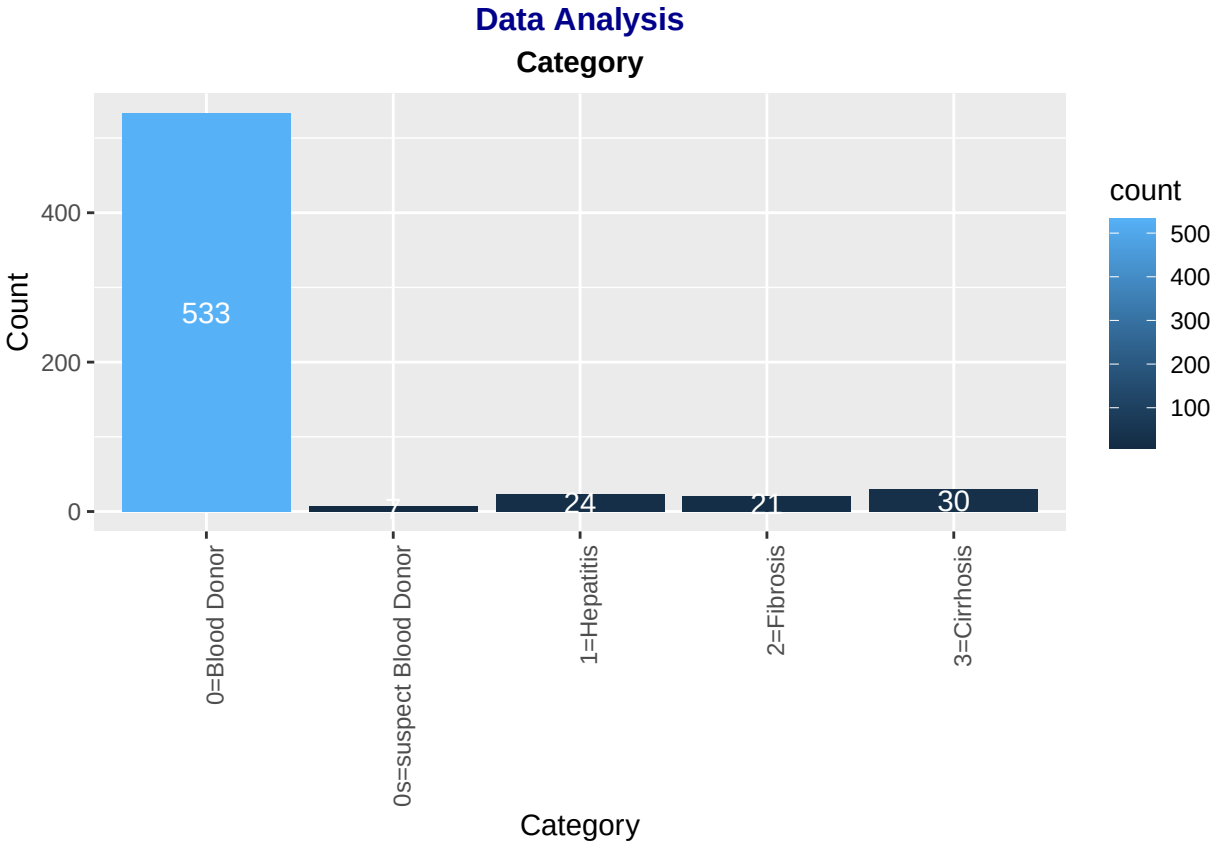
Category	Count
0=Blood Donor	533
0s=suspect Blood Donor	7
1=Hepatitis	24
2=Fibrosis	21
3=Cirrhosis	30

From the frequency analysis, we could gather the following Category attribute is composed of 5 categories as listed above

There are a total of 615 observations 540 samples belong to Blood Donors whereas 75belong to Hepatitis C patients 88% are Blood Donors as against 12% Hepatitics C patients

Blood Donors make up most of the samples Category

Visually, here is the histogram that explains the frequencies across categories



2.1.2 Data Analysis: Age

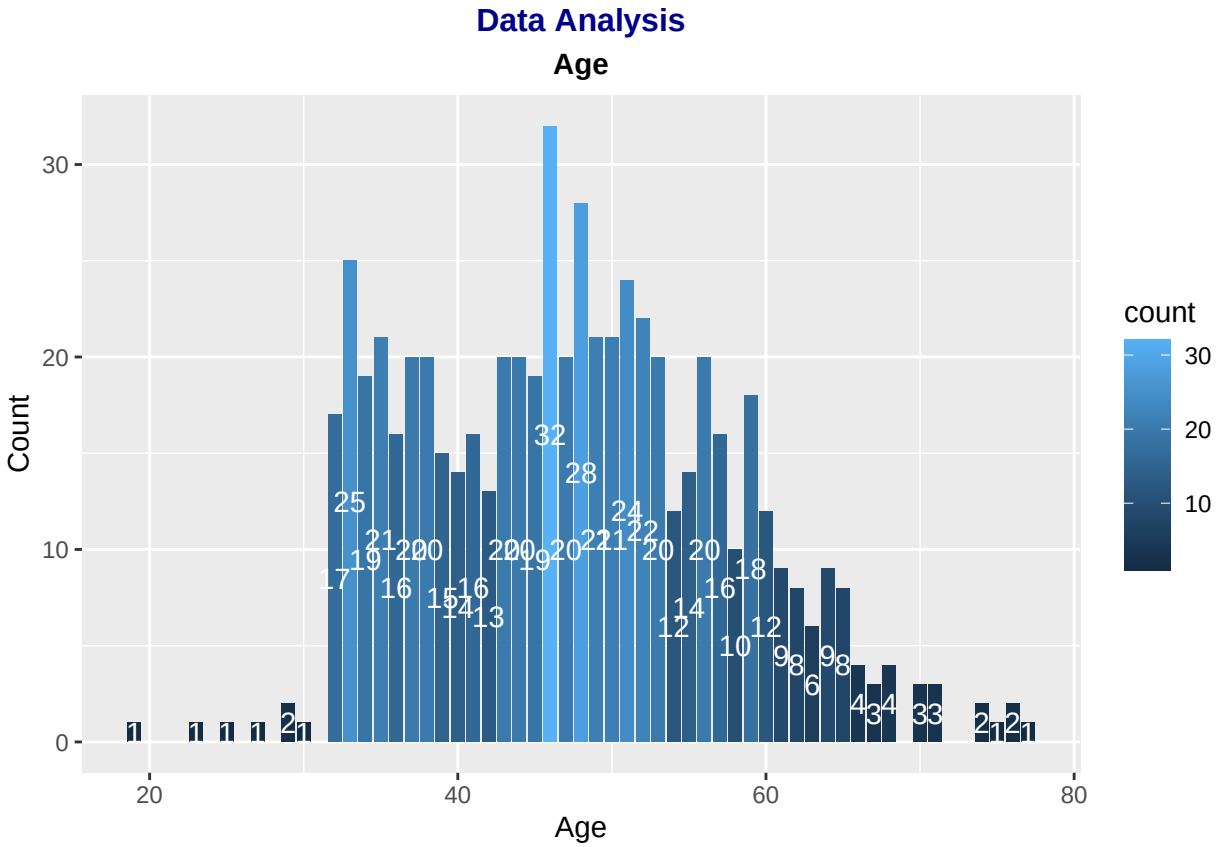
Below is the Analysis of the Mean, Median and SD of the Age attribute

```
## [1] "Class: "
```

```
## [1] "integer"
```

Median	Mean	SD
47	47.40813	10.05511

Histogram of the Age attribute:



Exploring the quartile ranges and values

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.00   39.00   47.00   47.41   54.00   77.00
```

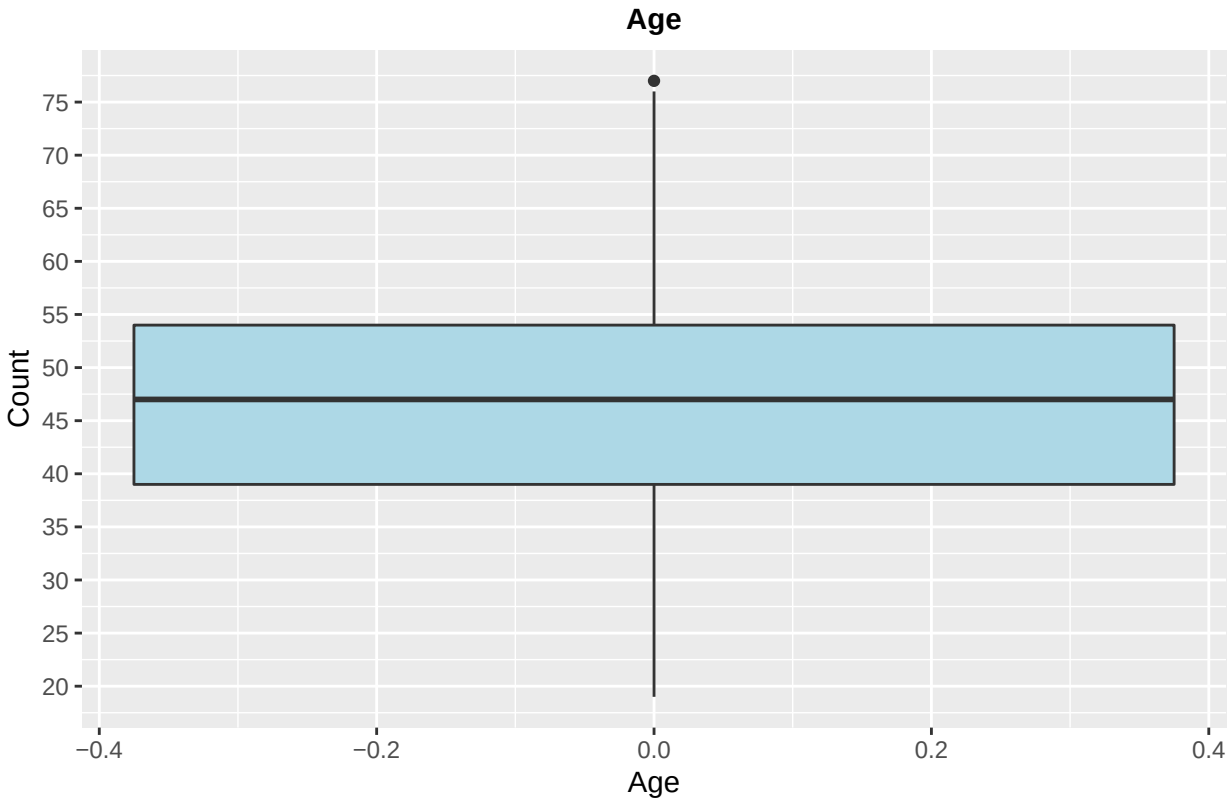
```
## [1] "IQR: "
```

```
## [1] 15
```

As seen from the summary above, 1st and 3rd Quartiles are 39 and 54 respectively The inter quartile range between 1st and 3rd quartiles is 15

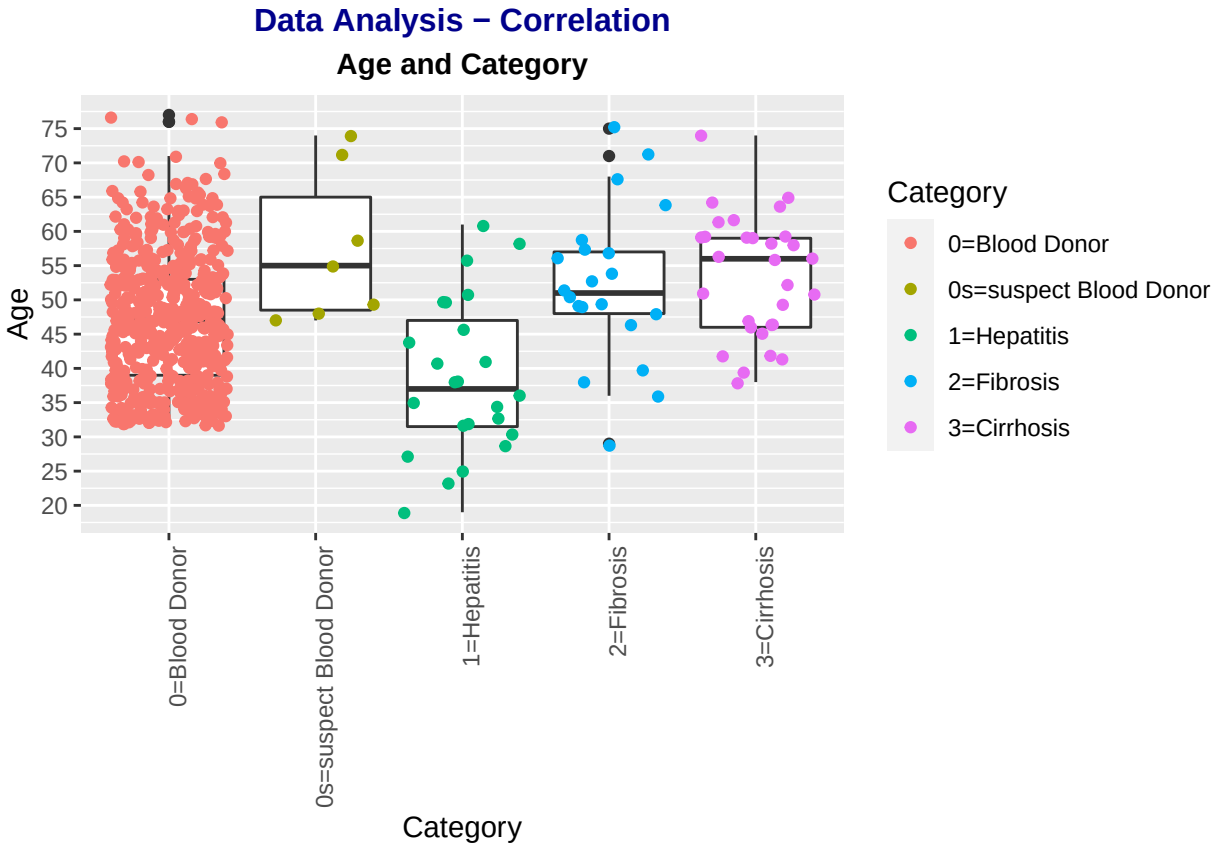
Below is a boxplot to visualize the IQR

Data Analysis



50% of the samples are in the age group 39 to 54 - 25th(Q1) percentile to 75th(Q3) percentile

Age's co-relation with Category demonstrated visually via boxplot

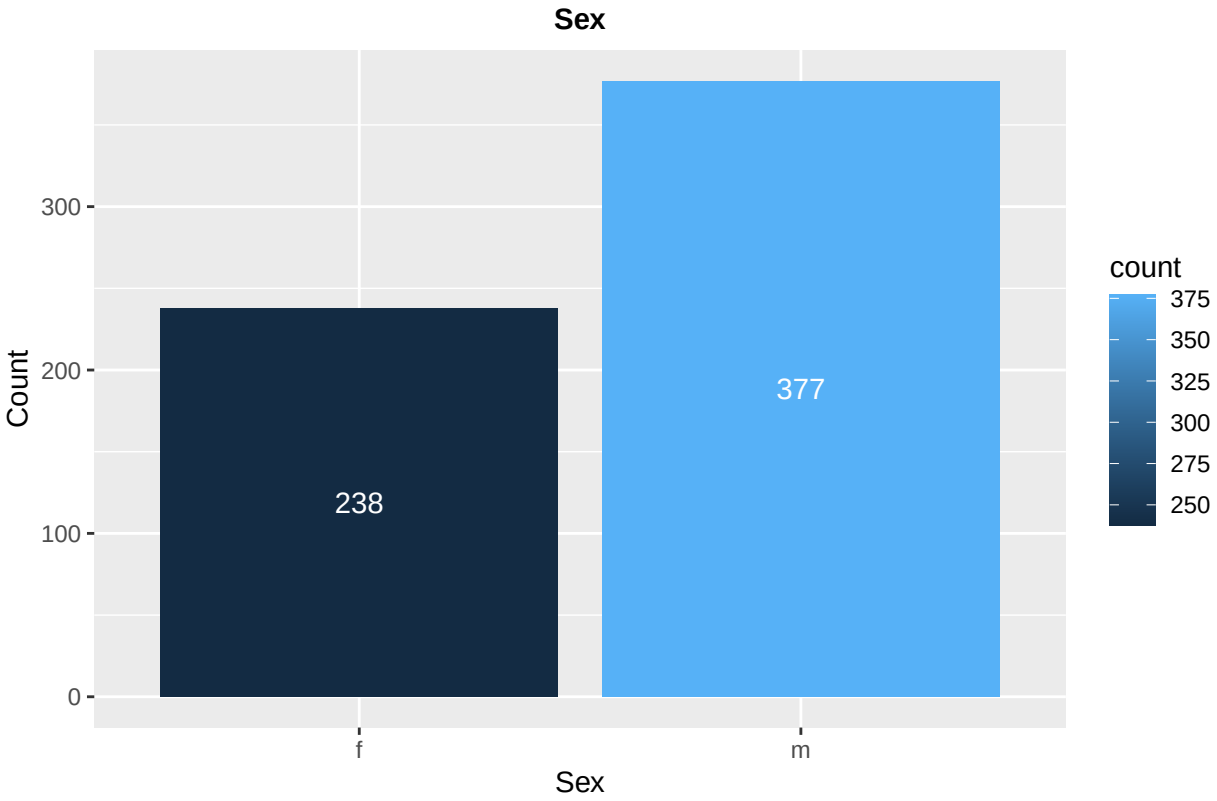


There are overlaps between all the categories w.r.t age. Density of blood donor category is much higher than the others, while Median for 1=Hepatitis is lower than the others

2.1.3 Data Analysis: Sex

Sex is a non-numerical attribute

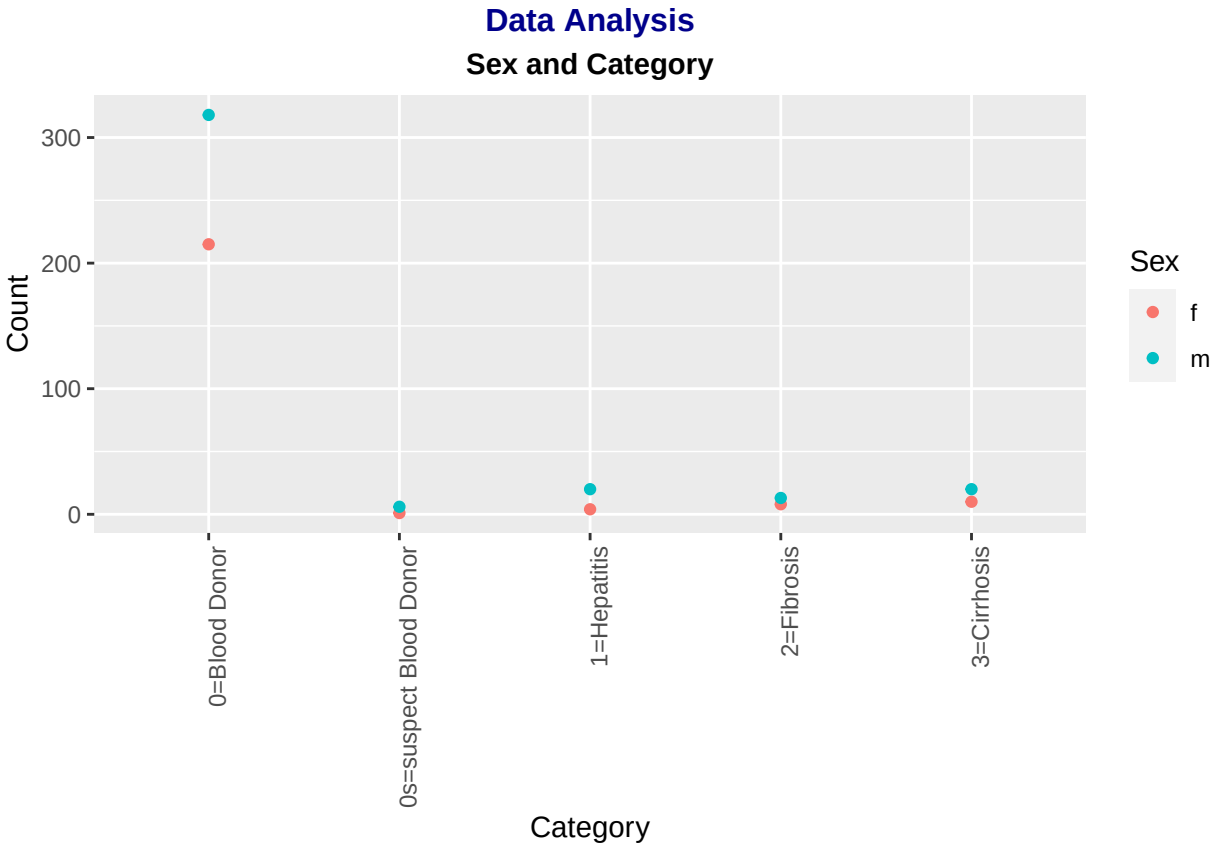
Data Analysis



Sample has more males than females. 377 males and 238 females

Distribution of Sex across Category via boxplot

```
## `summarise()` has grouped output by 'Category'. You can override using the `.groups` argument.
```



2.1.4 Data Analysis: ALB

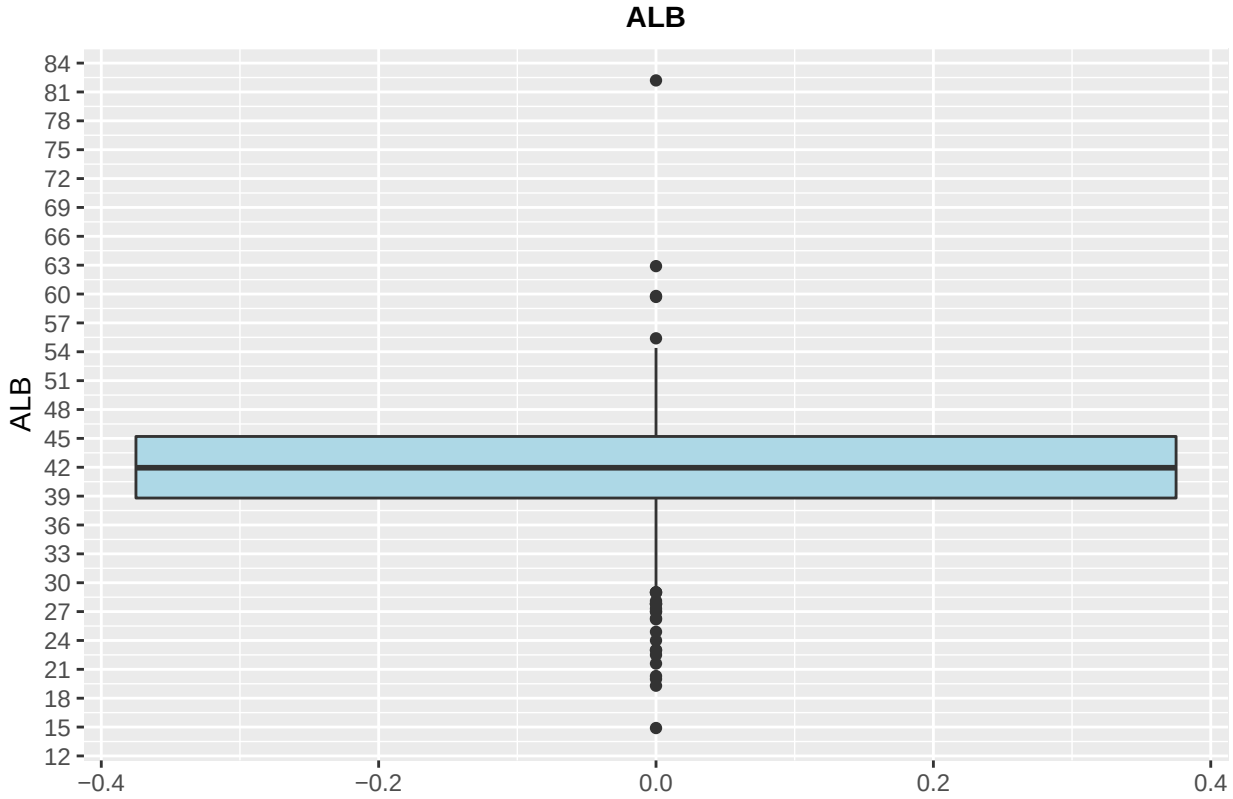
Below is the Analysis of the Mean, Median and SD of the Age attribute

```
## [1] "Class: "
## [1] "numeric"
```

Median	Mean	SD
41.95	41.6202	5.780629

Histogram of the ALB attribute:

Data Analysis



Exploring the quartile ranges and values

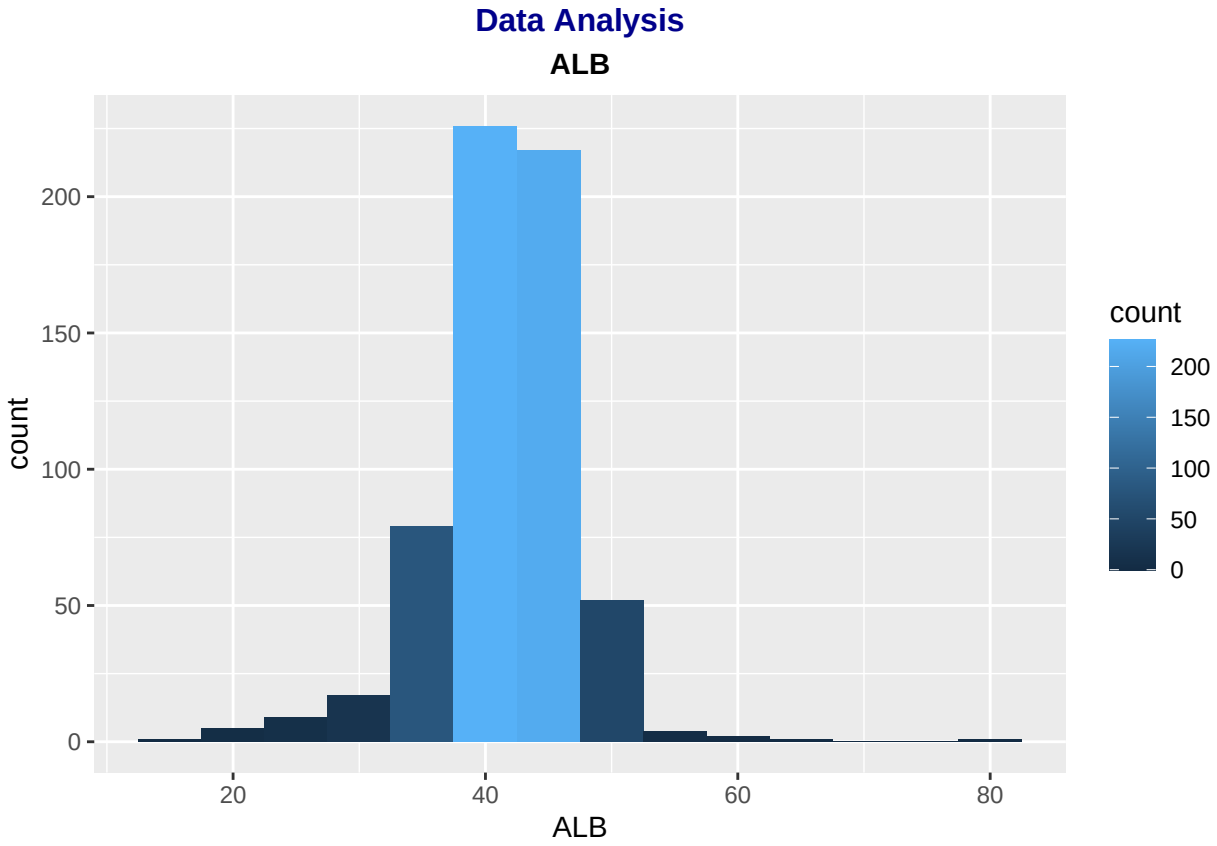
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   14.90   38.80   41.95   41.62   45.20   82.20       1
```

```
## [1] "IQR: "
```

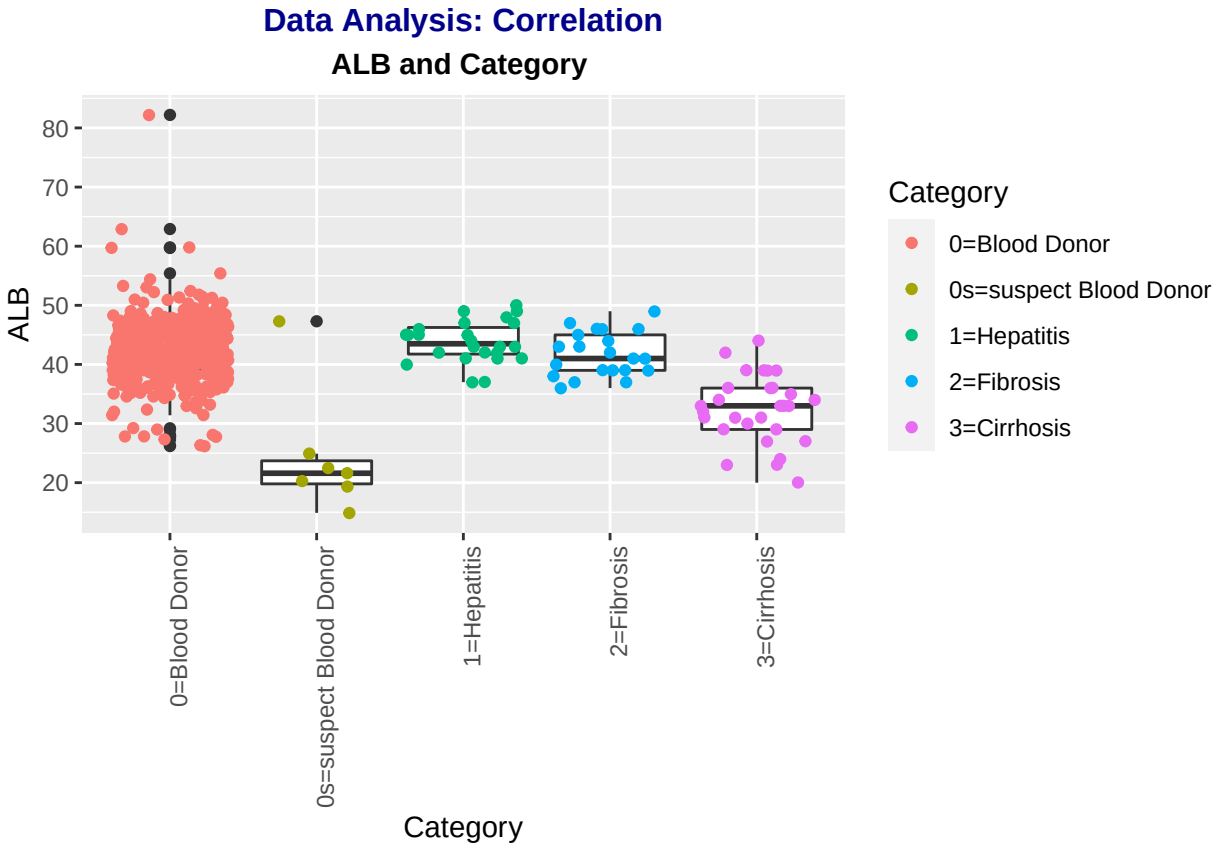
```
## [1] 6.4
```

As seen from the summary above, 1st and 3rd Quartiles are 38.8 and 45.2 respectively The inter quartile range between 1st and 3rd quartiles is 6.4 There is one observation where ALB is NA

Below is a boxplot to visualize the IQR



ALB's co-relation with Category demonstrated visually via boxplot



Os=Blood donor suspect has a very low ALB interquartile range that is different from the other categories
3=Cirrhosis has an IQR that doesn't overlap with the other categories

2.1.5 Data Analysis: ALP

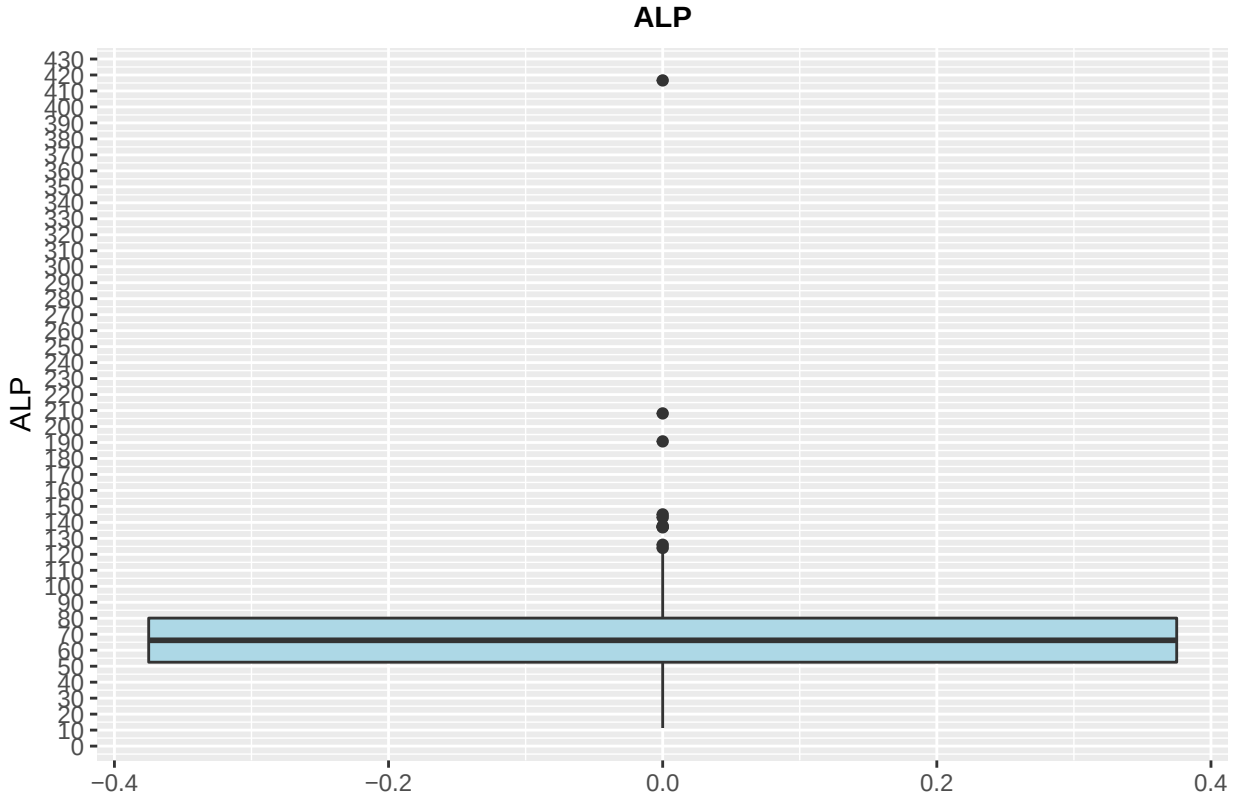
ALP is a numeric field and is a continuous variable Below is the Analysis of the Mean, Median and SD of the ALP attribute

```
## [1] "Class: "
## [1] "integer"
```

Median	Mean	SD
47	47.40813	10.05511

Histogram of the ALP attribute:

Data Analysis



Exploring the quartile ranges and values

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   11.30   52.50   66.20   68.28   80.10  416.60      18
```

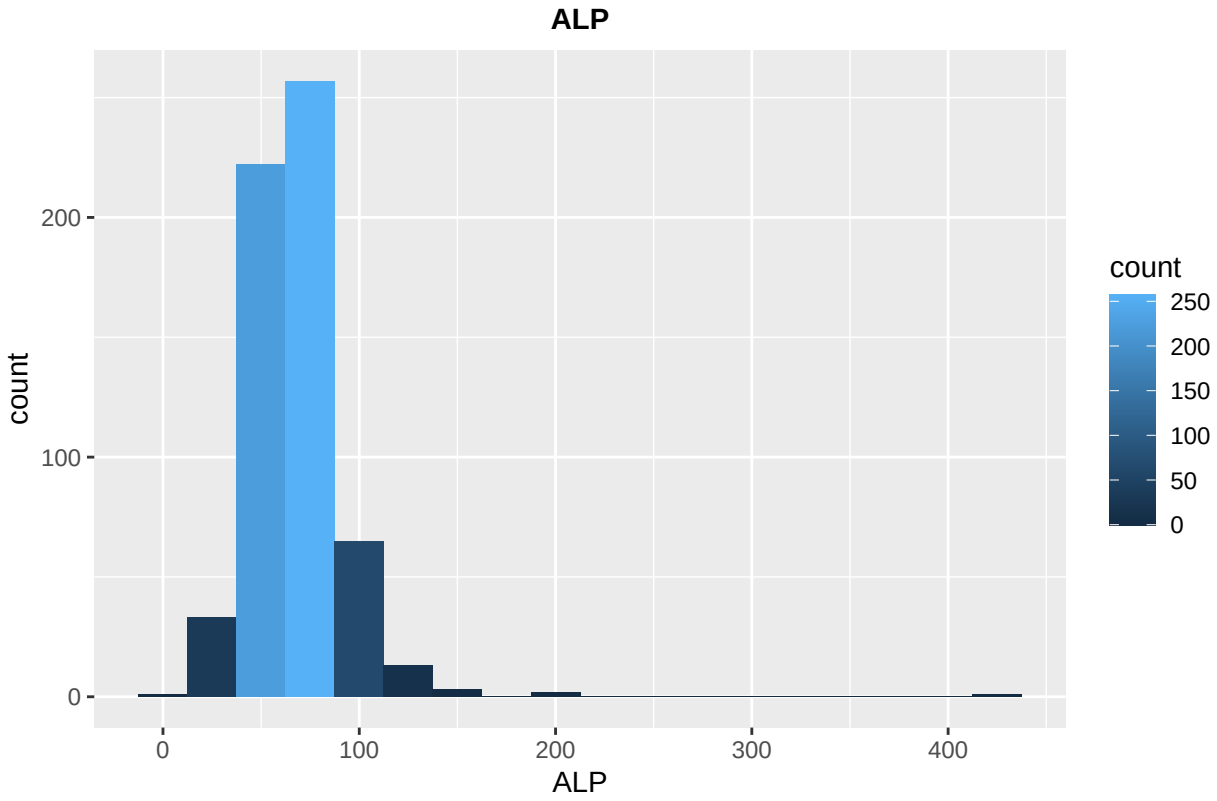
```
## [1] "IQR: "
```

```
## [1] 27.6
```

As seen from the summary above, 1st and 3rd Quartiles are 52 and 80 respectively The inter quartile range between 1st and 3rd quartiles is 28 There are 18 observations where ALP is NA

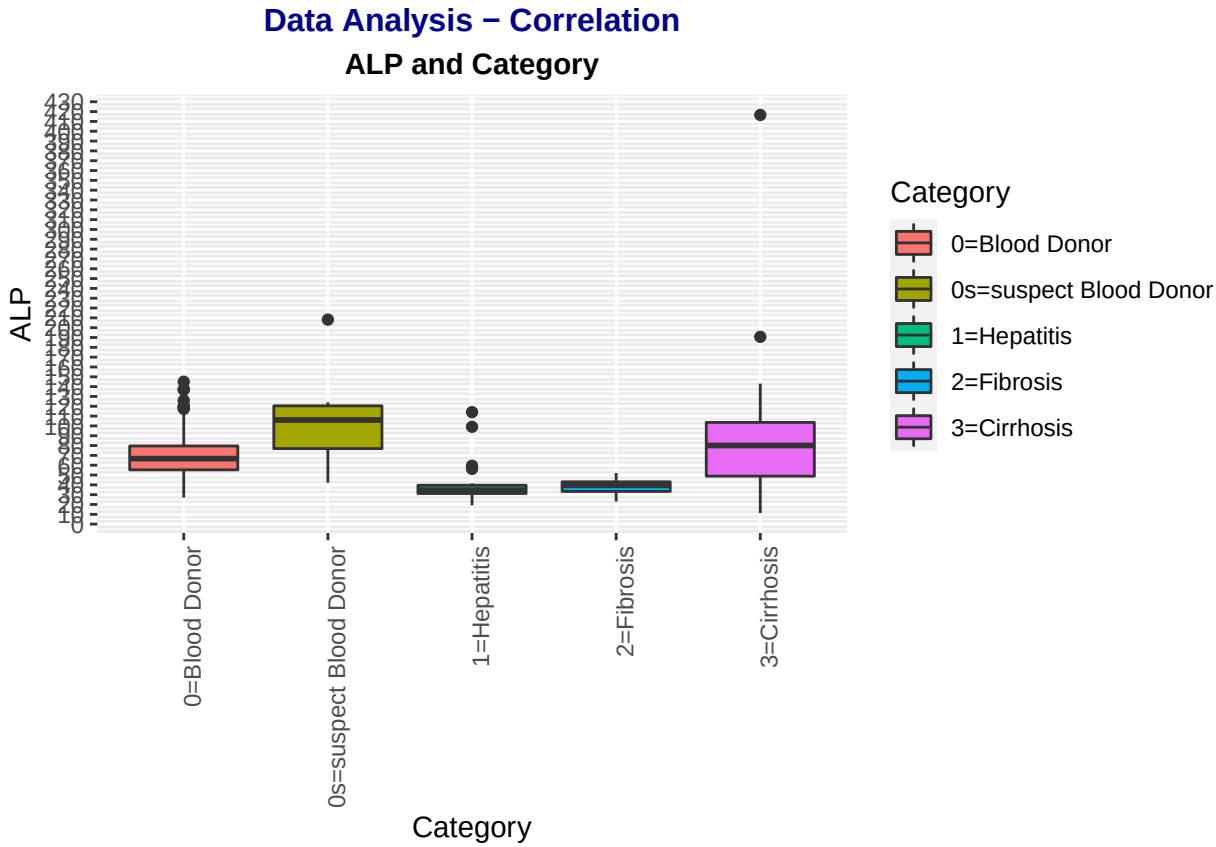
Below is a boxplot to visualize the IQR

Data Analysis



Histogram is left-skewed

ALP's co-relation with Category demonstrated visually via boxplot



There are overlaps across other categories 1=Hepatitis and 2=Fibrosis overlap each other but are distinguishable form the other categories

2.1.6 Data Analysis: ALT

ALT is a numeric field and is a continuous variable Below is the Analysis of the Mean, Median and SD of the ALT attribute

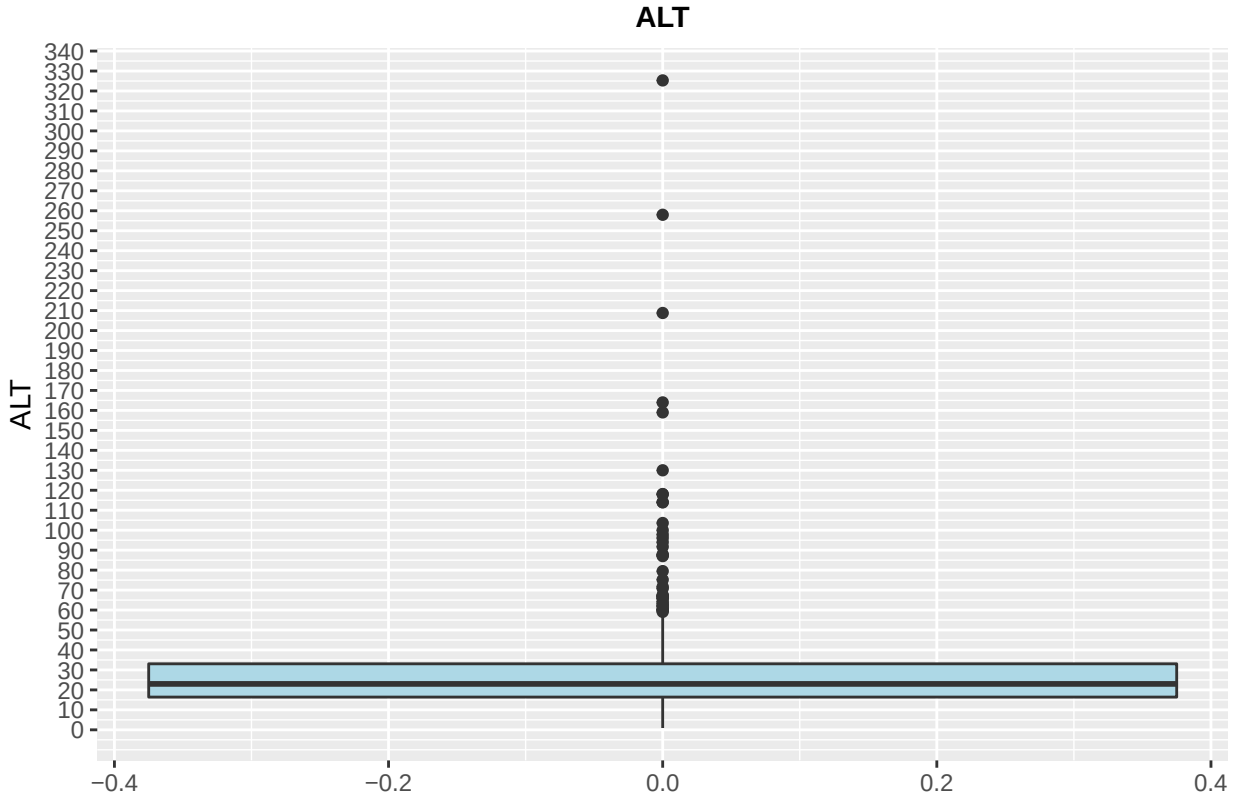
```
## [1] "Class: "
```

```
## [1] "numeric"
```

Median	Mean	SD
23	28.45081	25.46969

Histogram of the ALT attribute:

Data Analysis



Exploring the quartile ranges and values

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.90   16.40   23.00   28.45   33.08  325.30       1
```

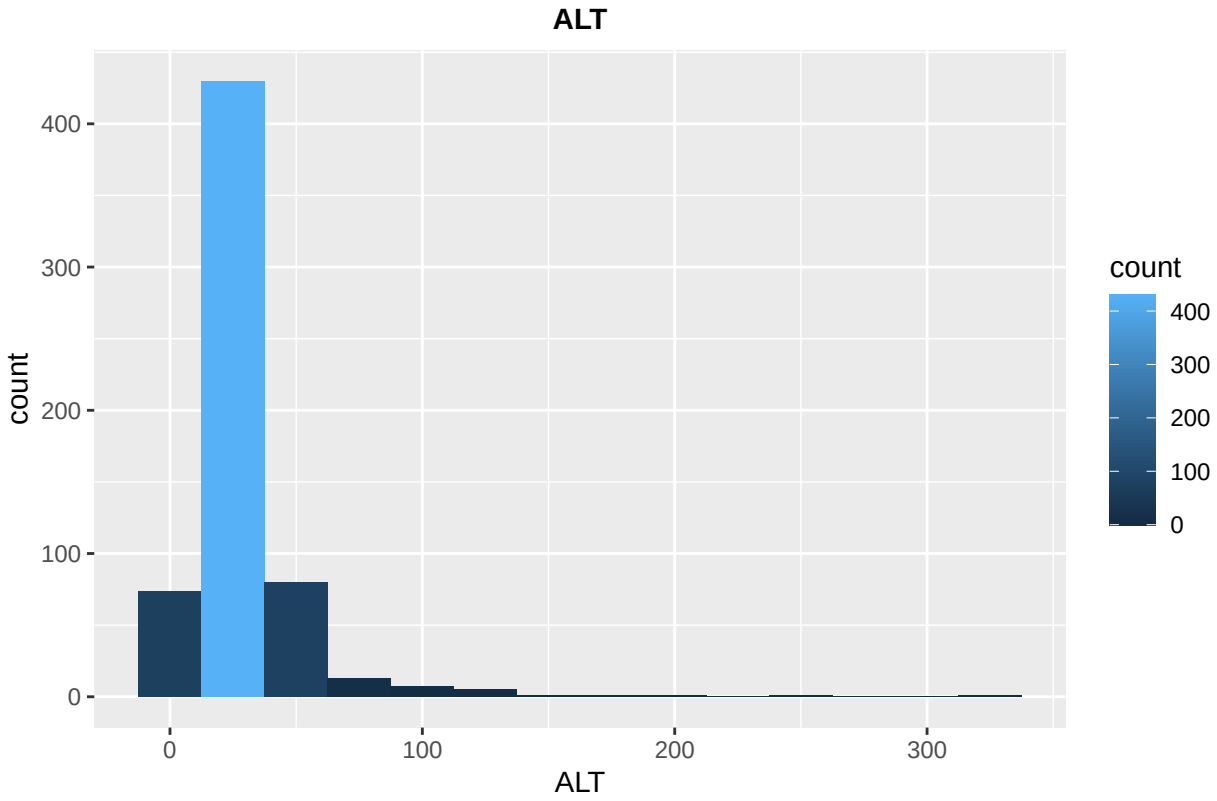
```
## [1] "IQR: "
```

```
## [1] 16.675
```

As seen from the summary above, 1st and 3rd Quartiles are 16 and 33 respectively. The inter quartile range between 1st and 3rd quartiles is 16.7 However, the Min and Ma are 1 and 325 respectively There is 1 observation where ALT is NA

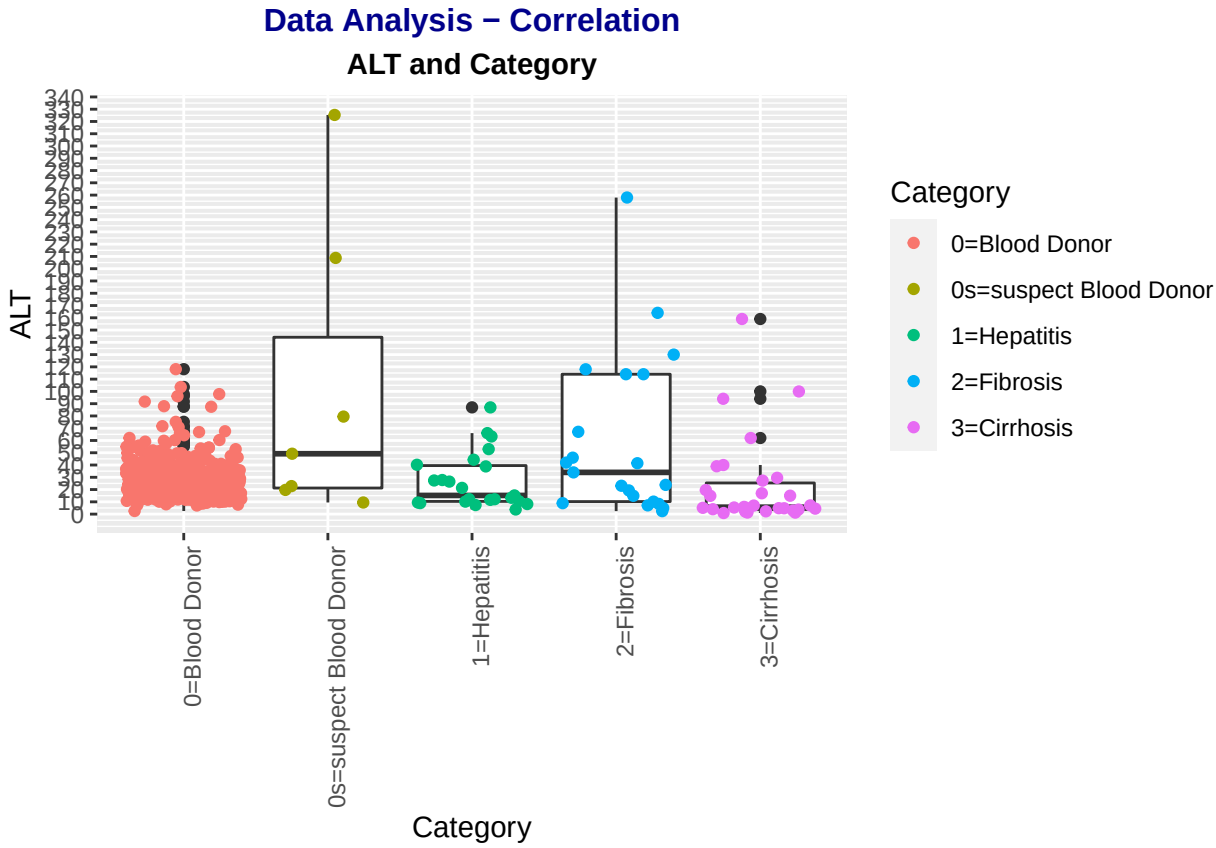
Below is a boxplot to visualize the IQR

Data Analysis



Histogram is left-skewed

ALT's co-relation with Category demonstrated visually via boxplot



There are overlaps across all categories

2.1.7 Data Analysis: AST

AST is a numeric field and is a continuous variable Below is the Analysis of the Mean, Median and SD of the AST attribute

```
## [1] "Class: "
```

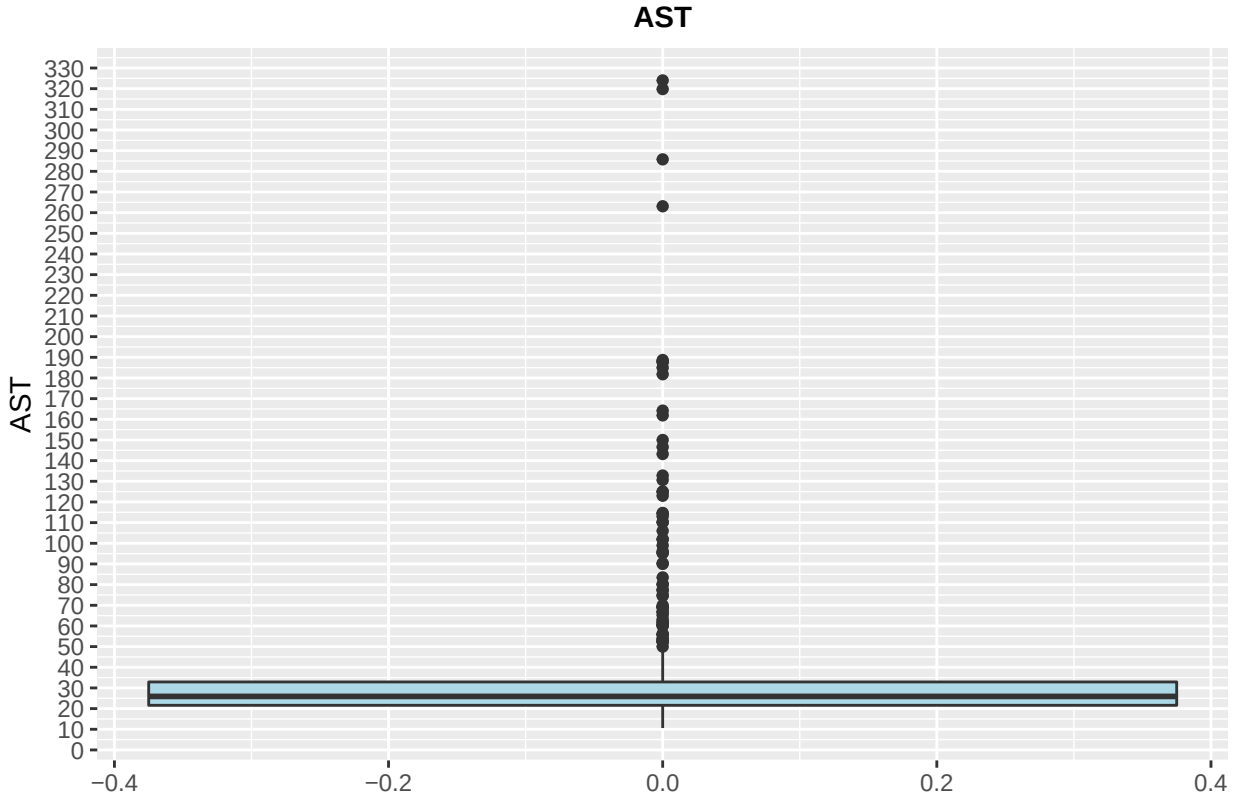
```
## [1] "numeric"
```

Median	Mean	SD
25.9	34.78634	33.09069

SD is 33.1, which explains the observations that are outliers

Histogram of the AST attribute:

Data Analysis



Exploring the quartile ranges and values

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.60   21.60   25.90   34.79   32.90  324.00
```

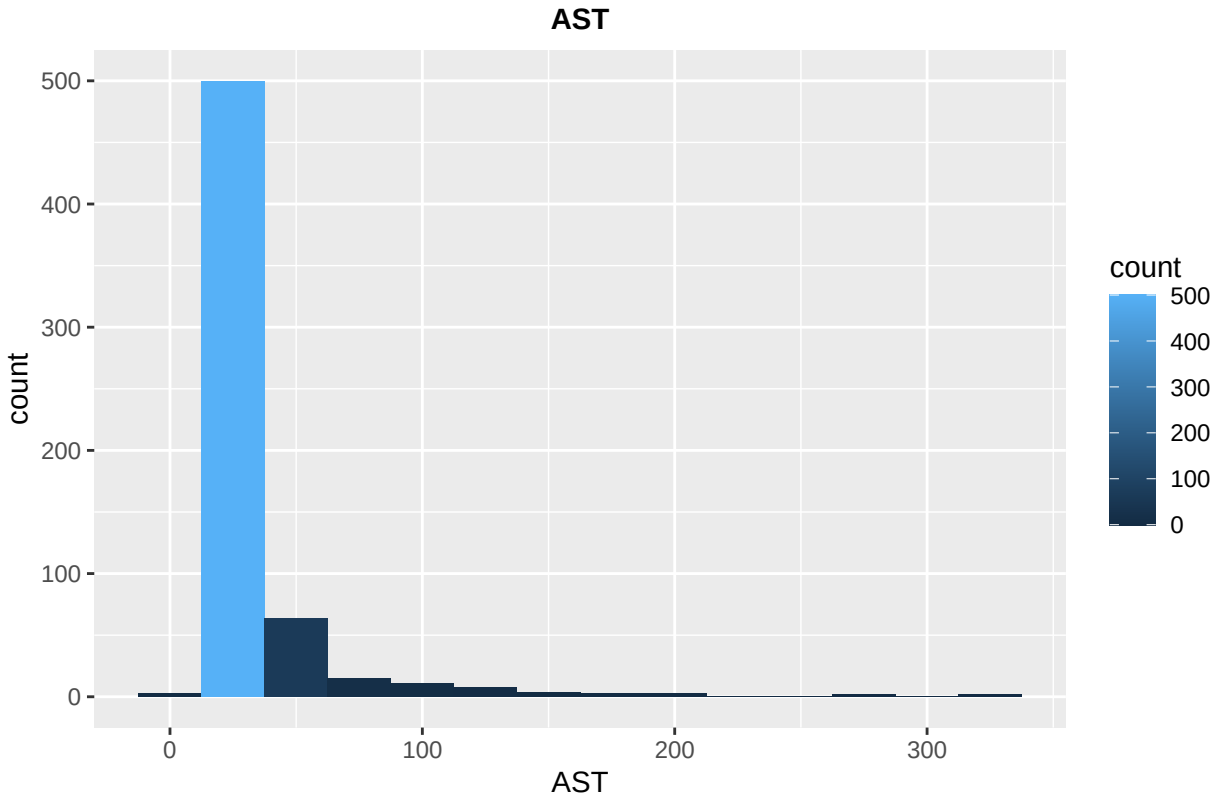
```
## [1] "IQR: "
```

```
## [1] 11.3
```

As seen from the summary above, 1st and 3rd Quartiles are 22 and 33 respectively. The inter quartile range between 1st and 3rd quartiles is 11.3 However, the Min and Ma are 11 and 324 respectively

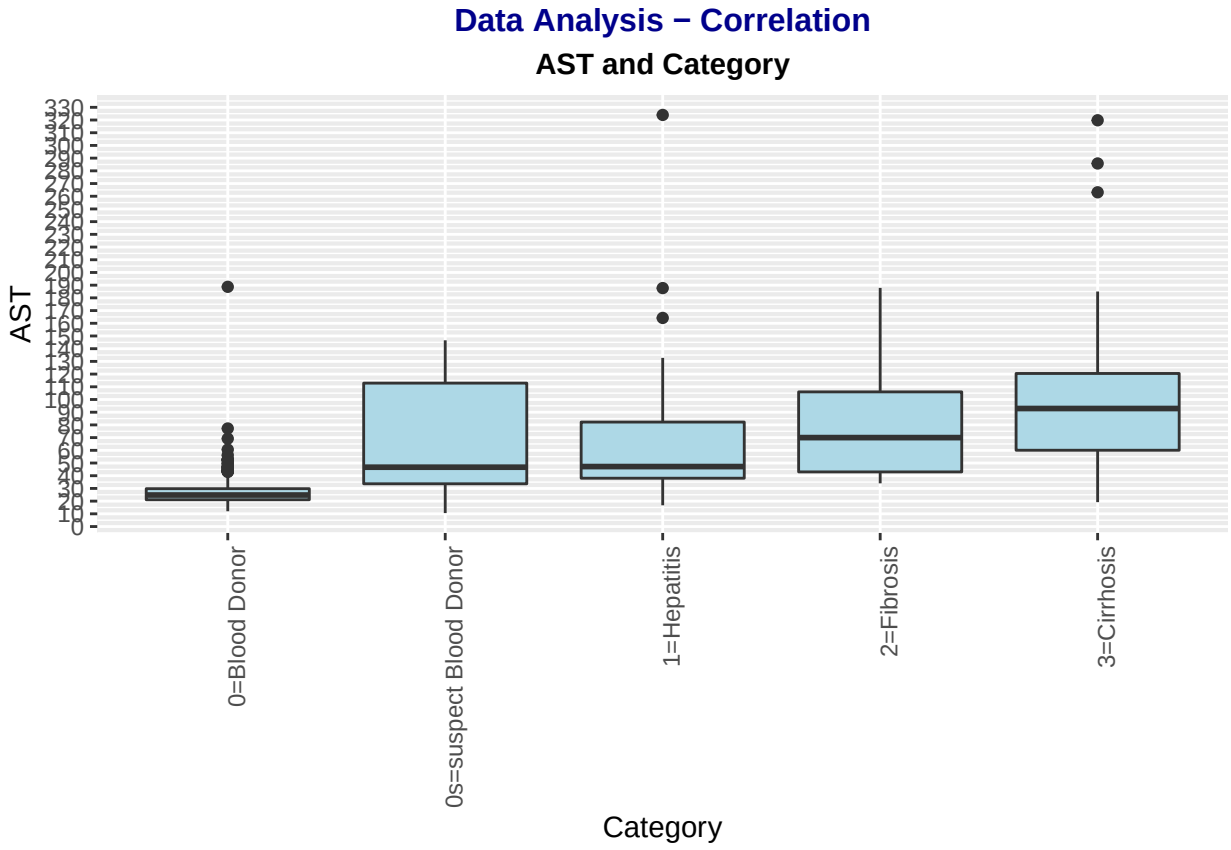
Below is a boxplot to visualize the IQR

Data Analysis



Histogram is left-skewed

AST's co-relation with Category demonstrated visually via boxplot



0-Blood donor's IQR doesn't overlap with the other categories. The remaining categories overlap amongst each other

2.1.8 Data Analysis: BIL

BIL is a numeric field and is a continuous variable Below is the Analysis of the Mean, Median and SD of the BIL attribute

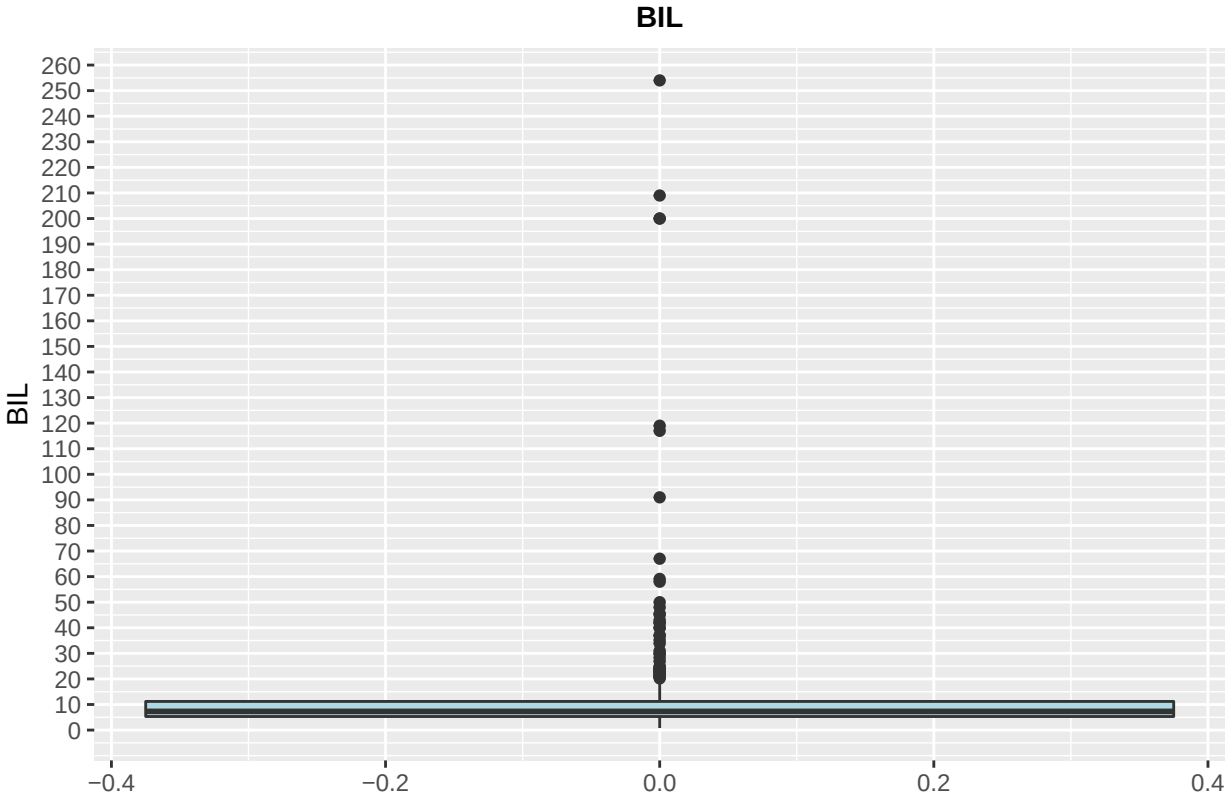
```
## [1] "Class: "
```

```
## [1] "numeric"
```

Median	Mean	SD
7.3	11.39675	19.67315

Histogram of the BIL attribute:

Data Analysis



Exploring the quartile ranges and values

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.8     5.3     7.3    11.4    11.2   254.0
```

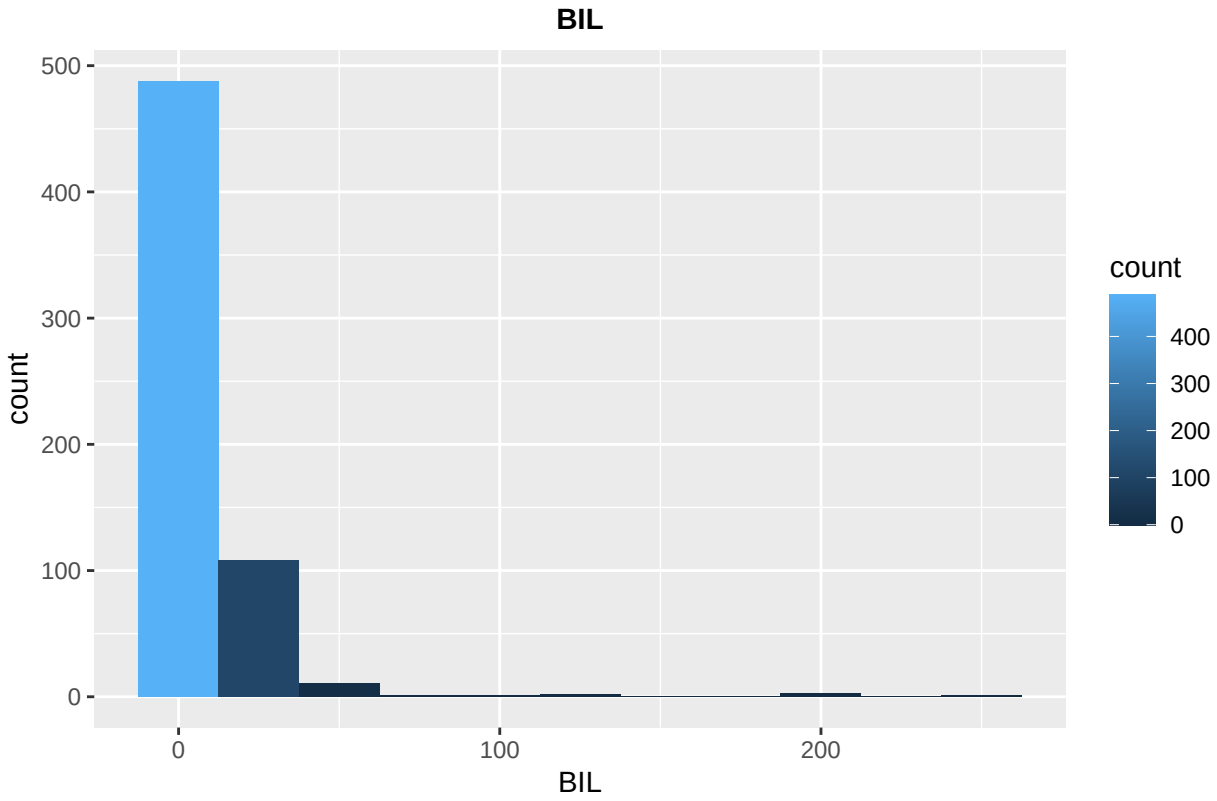
```
## [1] "IQR: "
```

```
## [1] 5.9
```

As seen from the summary above, 1st and 3rd Quartiles are 5.3 and 11.2 respectively. The inter quartile range between 1st and 3rd quartiles is 5.9 However, the Min and Max are 0.8 and 254 respectively

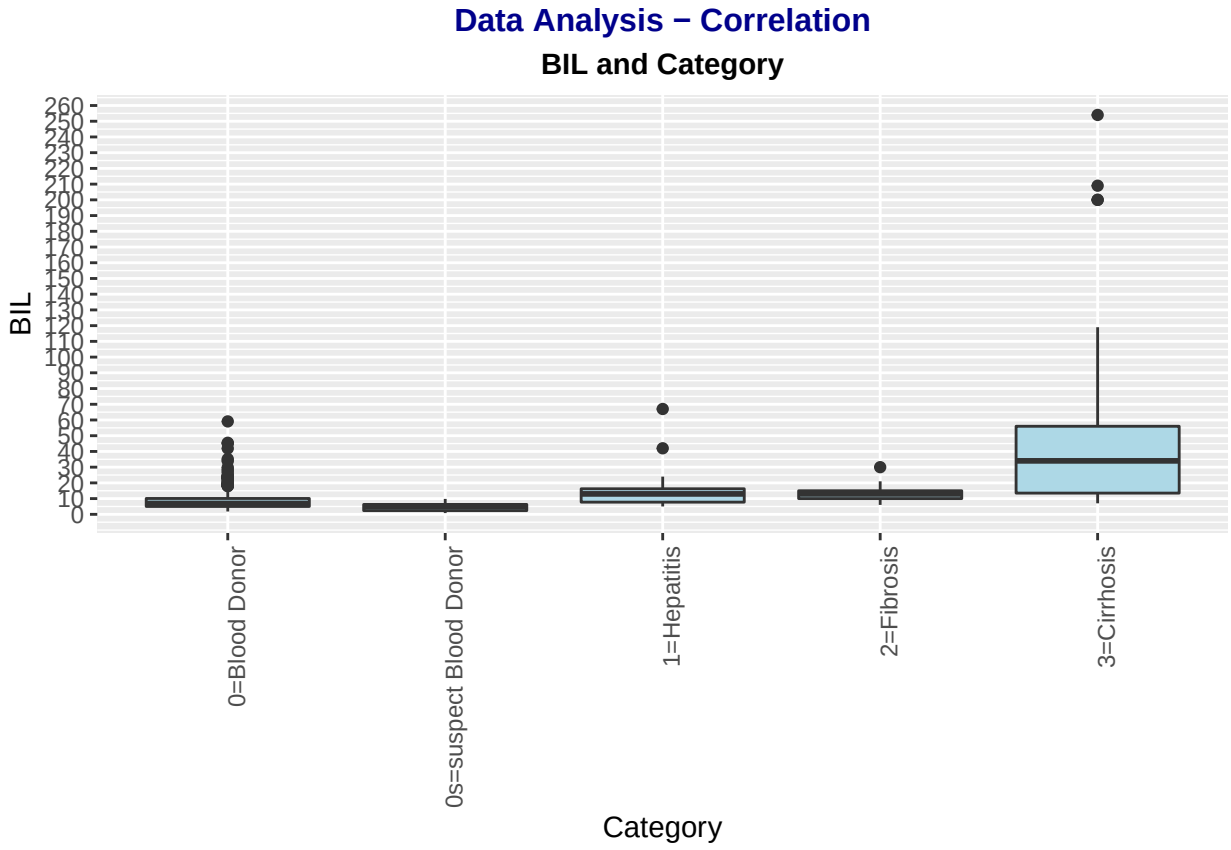
Below is a boxplot to visualize the IQR

Data Analysis



Histogram is left-skewed

BIL's co-relation with Category demonstrated visually via boxplot



3=Cirrhosis median is high compared to the other categories

2.1.9 Data Analysis: CHE

CHE is a numeric field and is a continuous variable Below is the Analysis of the Mean, Median and SD of the CHE attribute

```
## [1] "Class: "
```

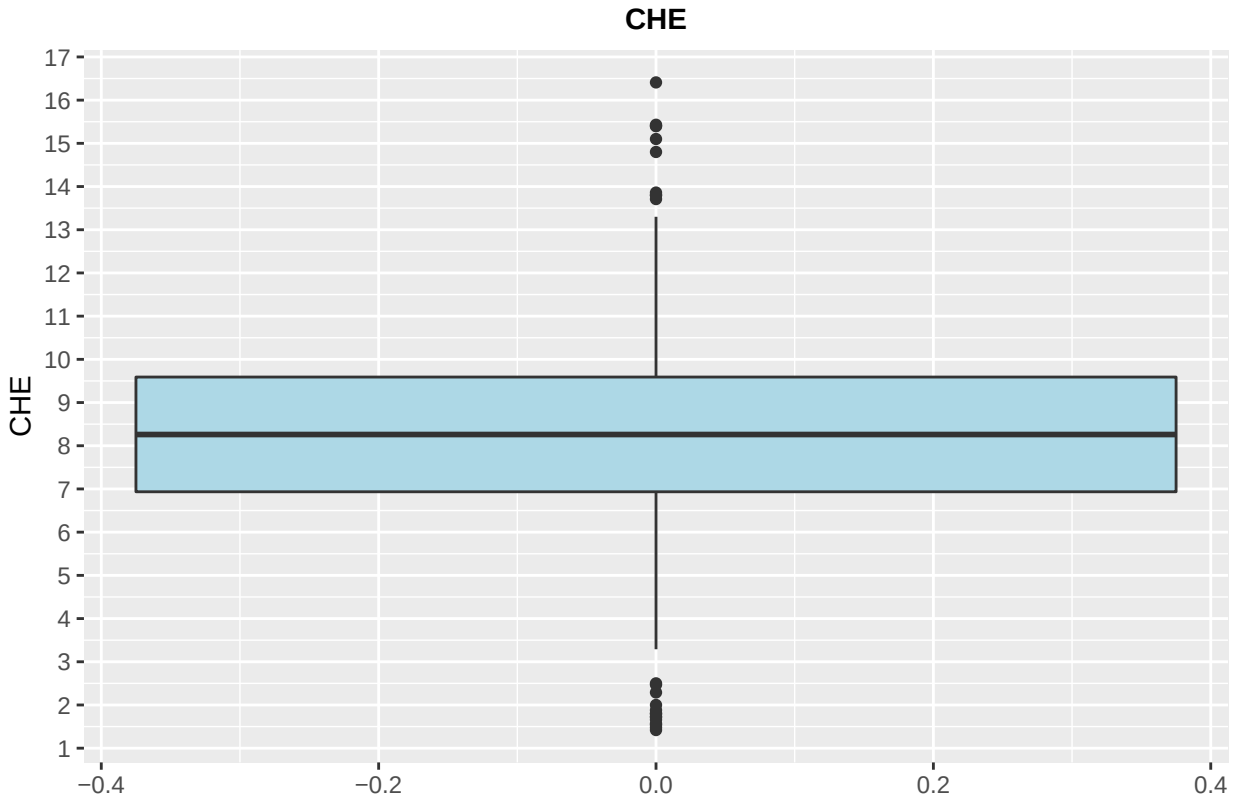
```
## [1] "numeric"
```

Median	Mean	SD
8.26	8.196634	2.205657

Standard Deviation is 2.21, which is small compared to the previous predictors

Histogram of the CHE attribute:

Data Analysis



Exploring the quartile ranges and values

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.420   6.935   8.260   8.197   9.590  16.410
```

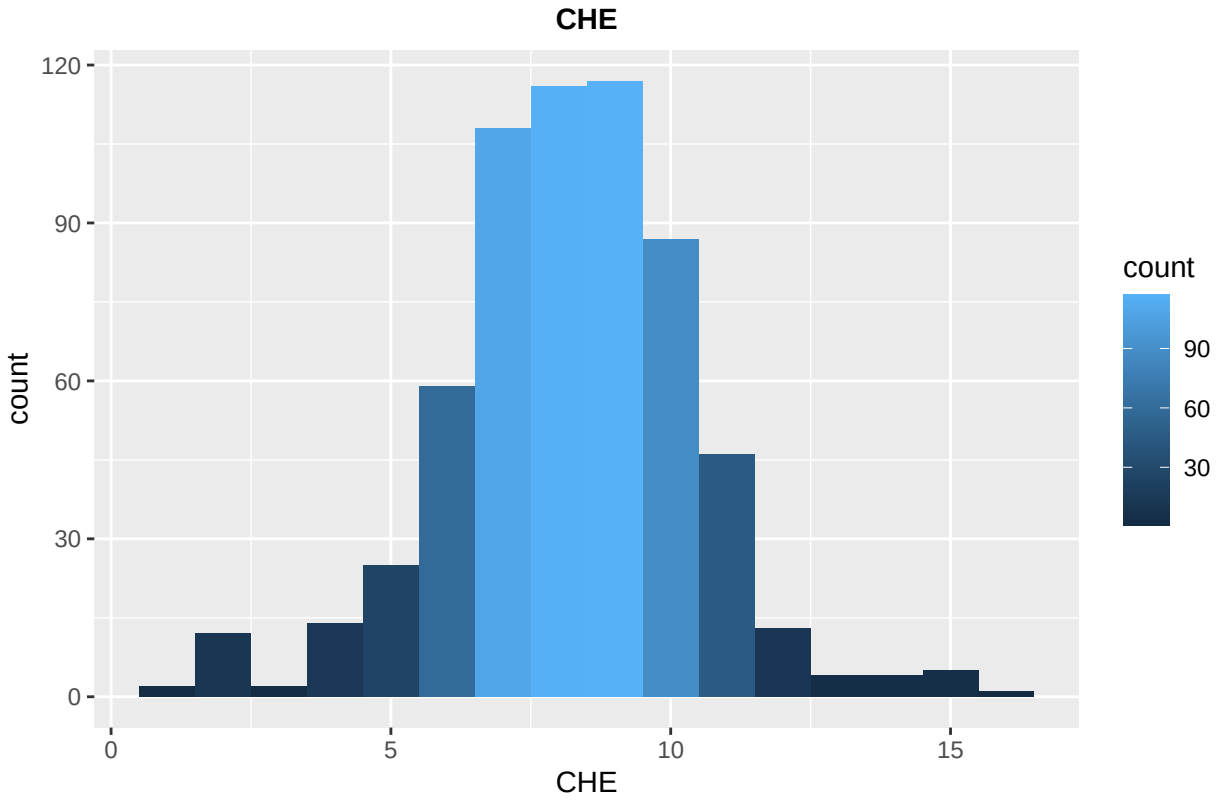
```
## [1] "IQR: "
```

```
## [1] 2.655
```

As seen from the summary above, 1st and 3rd Quartiles are 6.94 and 9.59 respectively. The inter quartile range between 1st and 3rd quartiles is 2.65

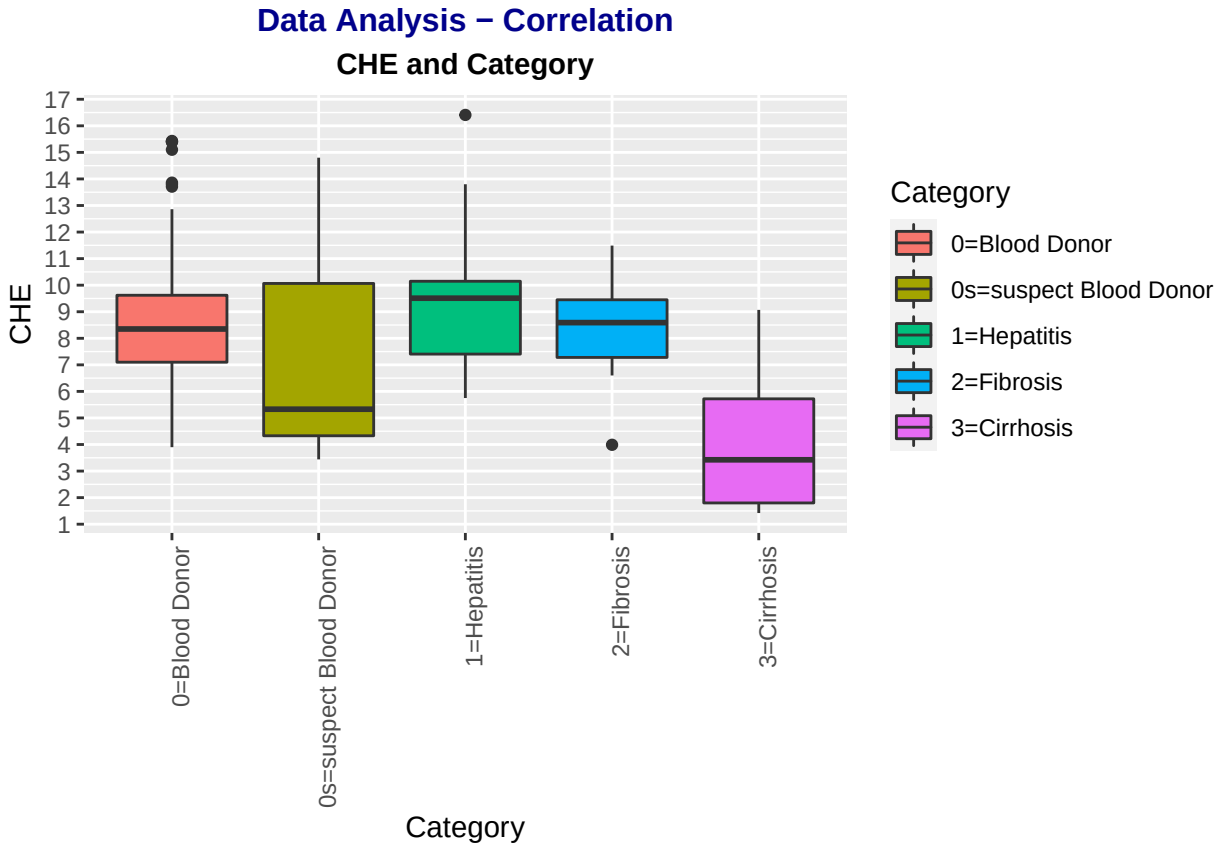
Below is a boxplot to visualize the IQR

Data Analysis



Histogram resembles a normal distribution

CHE's co-relation with Category demonstrated visually via boxplot



3=Cirrhosis median is low compared to the other categories

2.1.10 Data Analysis: CHOL

CHOL is a numeric field and is a continuous variable Below is the Analysis of the Mean, Median and SD of the CHOL attribute

```
## [1] "Class: "
```

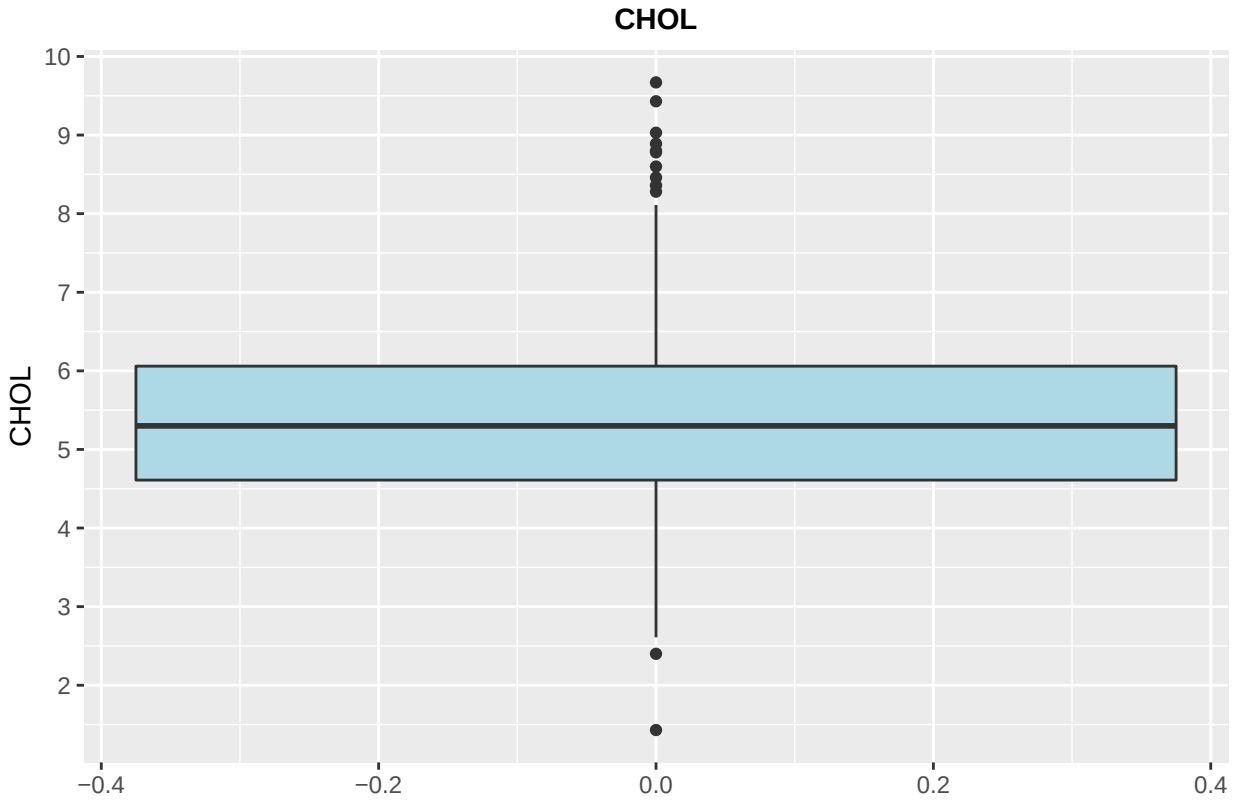
```
## [1] "numeric"
```

Median	Mean	SD
5.3	5.368099	1.132728

Standard Deviation is 1.13, which is small compared to most of the previous predictors

Histogram of the CHOL attribute:

Data Analysis



Exploring the quartile ranges and values

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.430   4.610   5.300   5.368   6.060   9.670      10
```

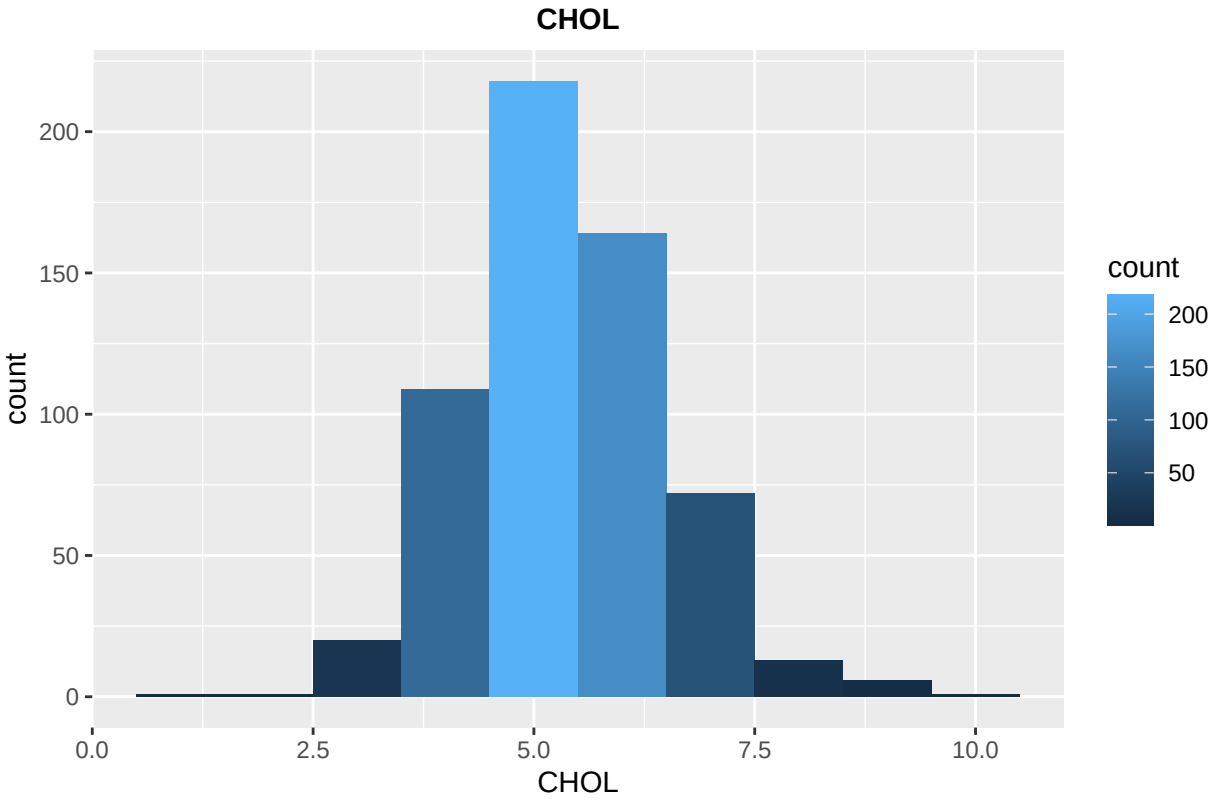
```
## [1] "IQR: "
```

```
## [1] 1.45
```

As seen from the summary above, 1st and 3rd Quartiles are 4.61 and 6.06 respectively. The inter quartile range between 1st and 3rd quartiles is 1.45. There are 10 observations that are NAs

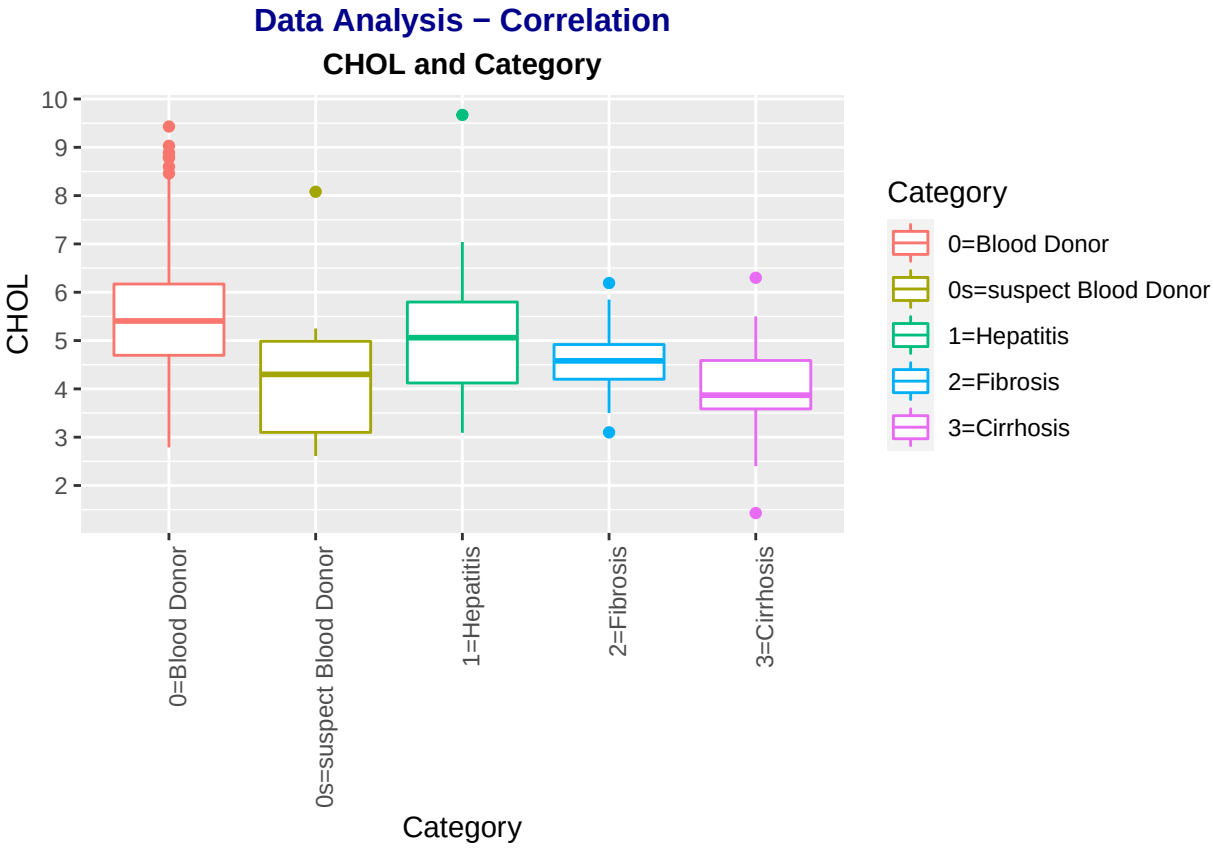
Below is a boxplot to visualize the IQR

Data Analysis



Histogram resembles a normal distribution

CHOL's co-relation with Category demonstrated visually via boxplot



There are overlaps across all categories

2.1.11 Data Analysis: CREA

CREA is a numeric field and is a continuous variable Below is the Analysis of the Mean, Median and SD of the CREA attribute

```
## [1] "Class: "
```

```
## [1] "numeric"
```

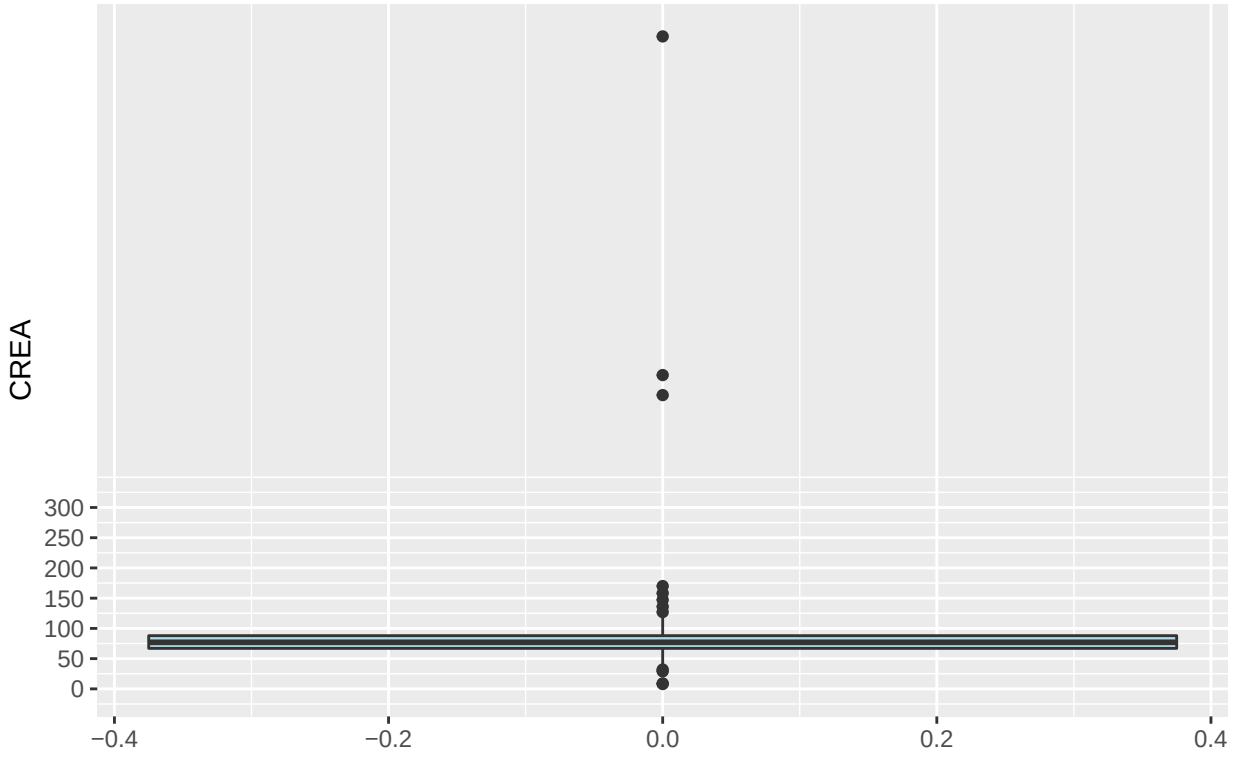
Median	Mean	SD
77	81.2878	49.75617

Standard Deviation is 49.8

Histogram of the CREA attribute:

Data Analysis

CREA



Exploring the quartile ranges and values

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   67.00   77.00   81.29   88.00 1079.10
```

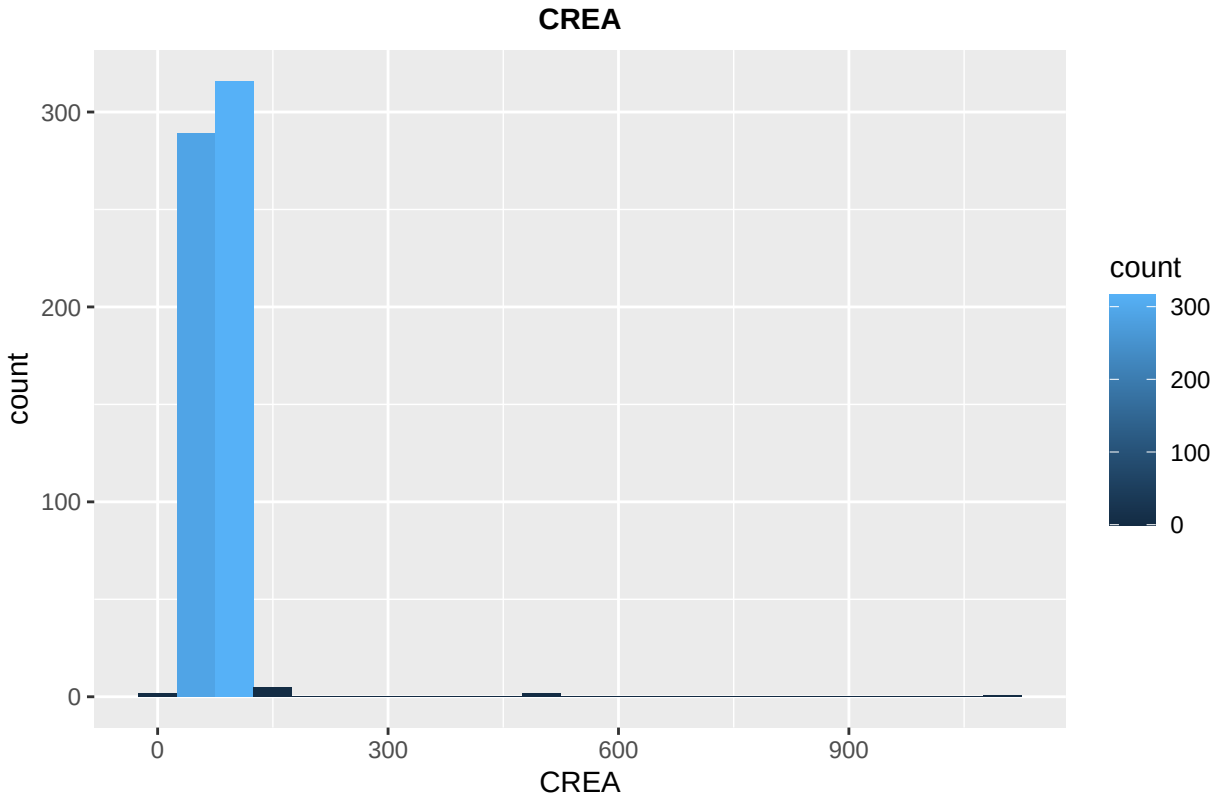
```
## [1] "IQR: "
```

```
## [1] 21
```

As seen from the summary above, 1st and 3rd Quartiles are 67 and 88 respectively. The inter quartile range between 1st and 3rd quartiles is 21 Max is huge with 1079, whereas Min is 8

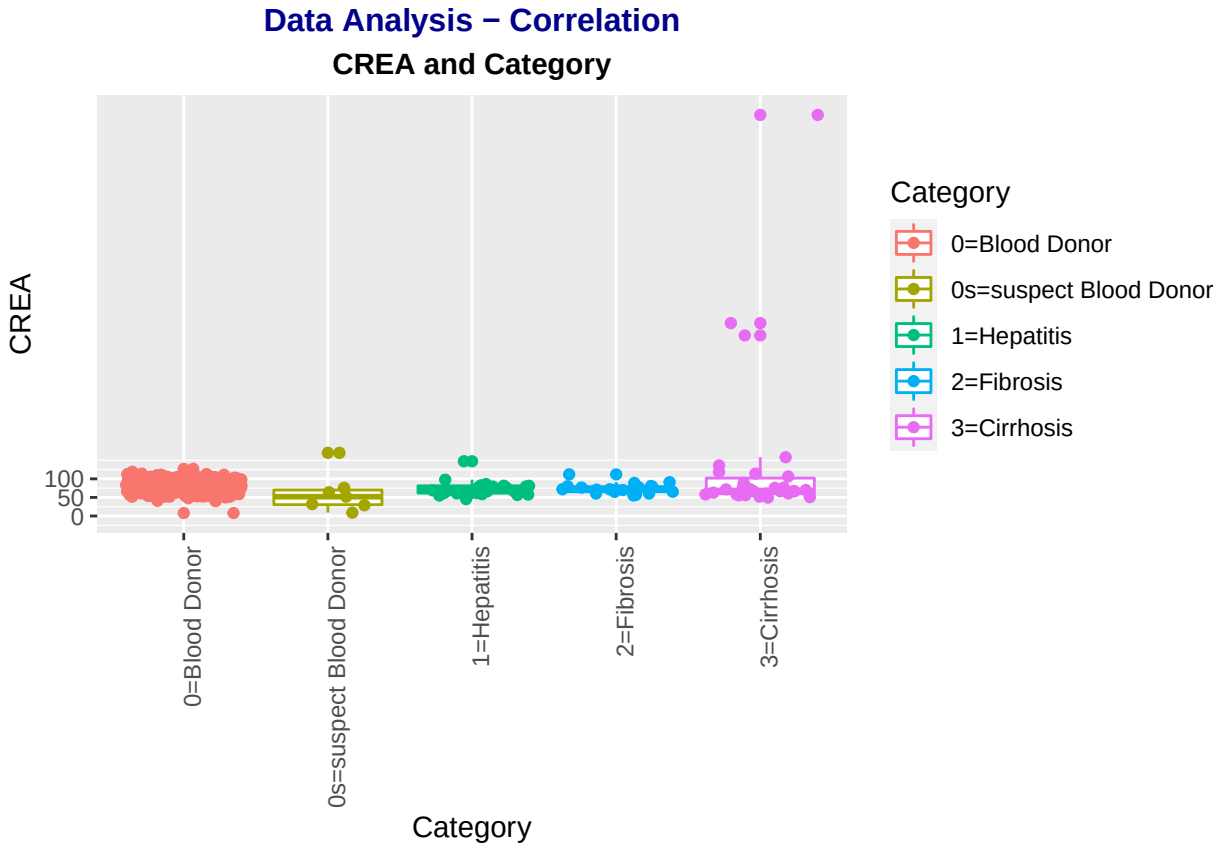
Below is a boxplot to visualize the IQR

Data Analysis



Histogram is left skewed

CREA's co-relation with Category demonstrated visually via boxplot



There are overlaps across all categories

2.1.12 Data Analysis: GGT

GGT is a numeric field and is a continuous variable Below is the Analysis of the Mean, Median and SD of the GGT attribute

```
## [1] "Class: "
```

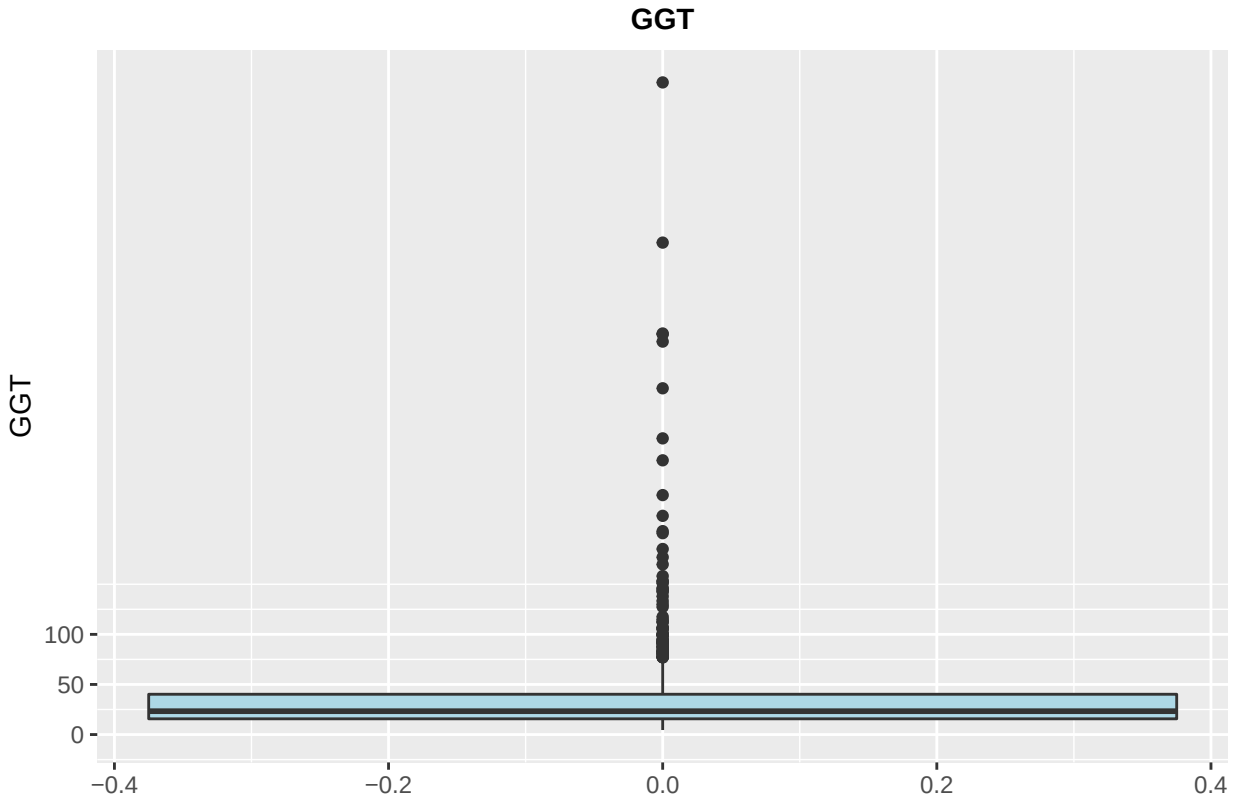
```
## [1] "numeric"
```

Median	Mean	SD
23.3	39.53317	54.66107

Standard Deviation is 54.7

Histogram of the GGT attribute:

Data Analysis



Exploring the quartile ranges and values

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.50   15.70   23.30   39.53   40.20  650.90
```

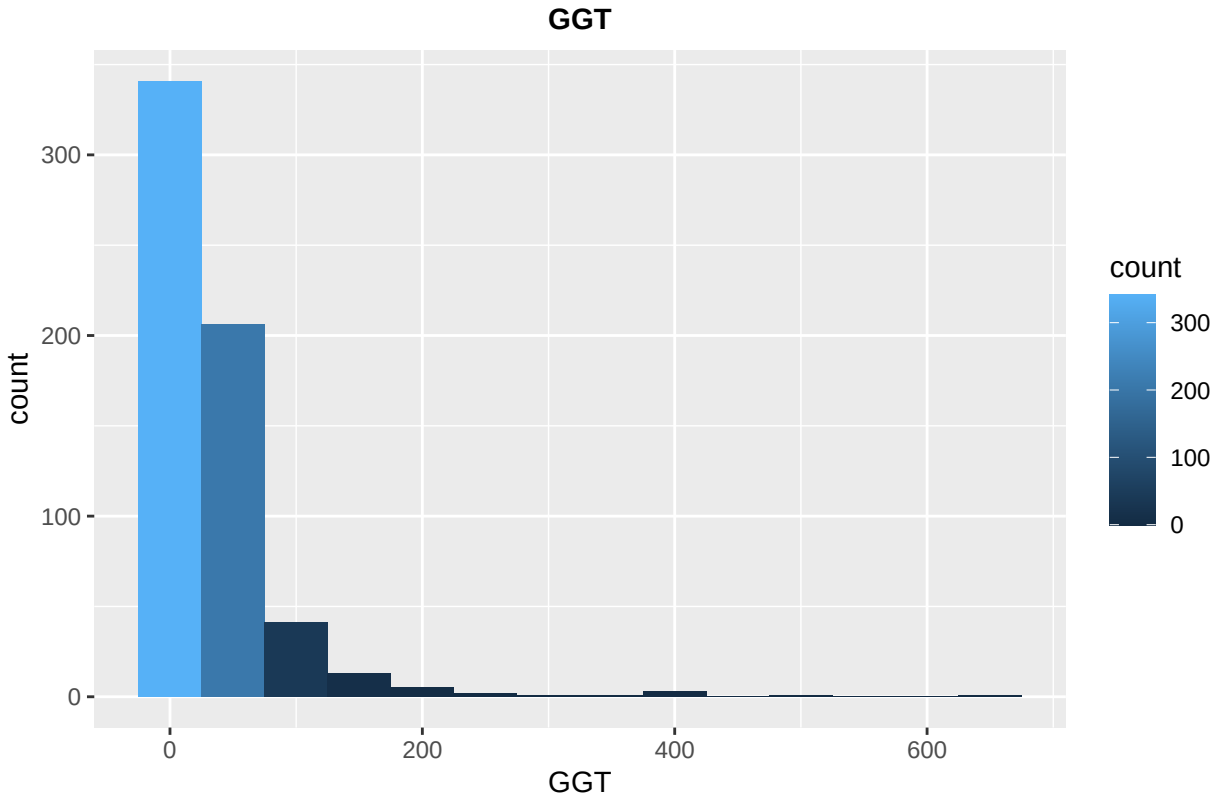
```
## [1] "IQR: "
```

```
## [1] 24.5
```

As seen from the summary above, 1st and 3rd Quartiles are 16 and 40 respectively. The inter quartile range between 1st and 3rd quartiles is 24.5 Max is huge with 651, whereas Min is 4

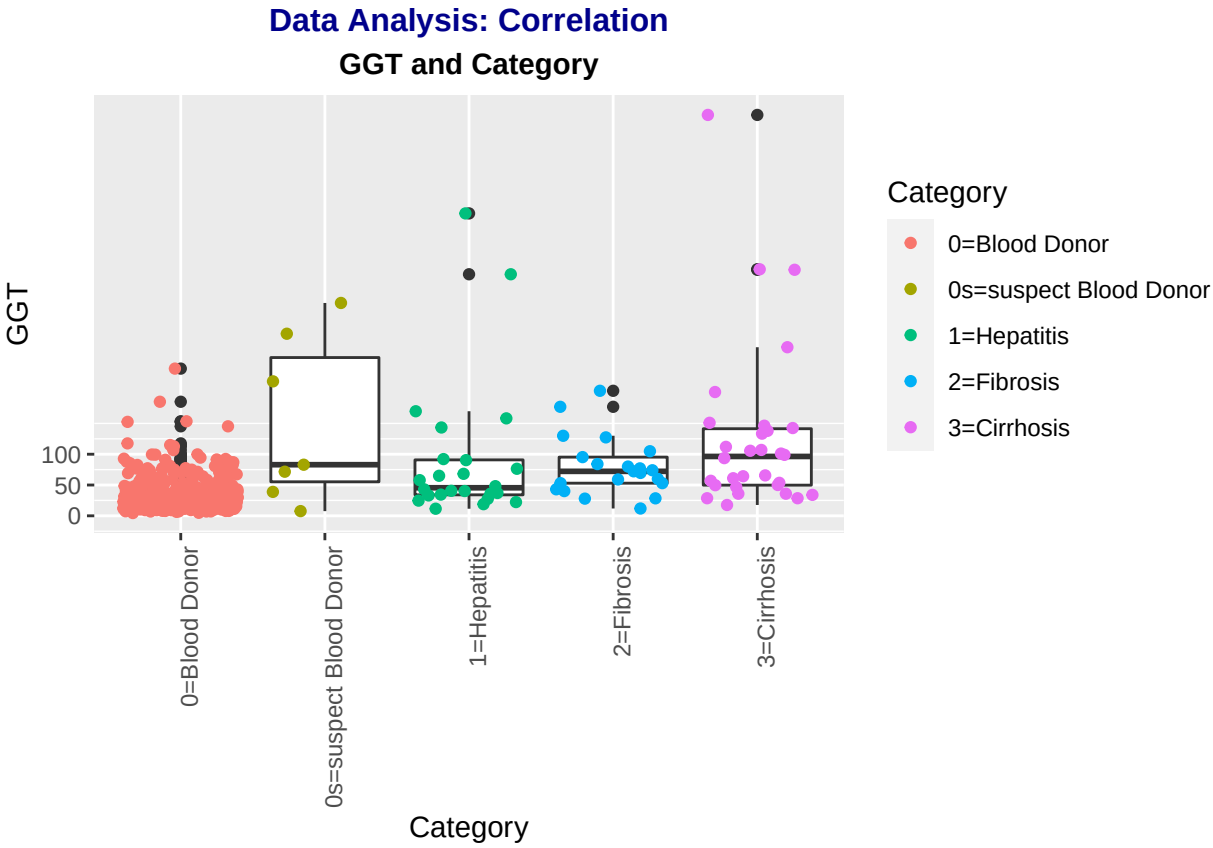
Below is a boxplot to visualize the IQR

Data Analysis



Histogram is left skewed

GGT's co-relation with Category demonstrated visually via boxplot



0=Blood donor's median doesn't overlap with the other category IQRs

2.1.13 Data Analysis: PROT

PROT is a numeric field and is a continuous variable Below is the Analysis of the Mean, Median and SD of the PROT attribute

```
## [1] "Class: "
```

```
## [1] "numeric"
```

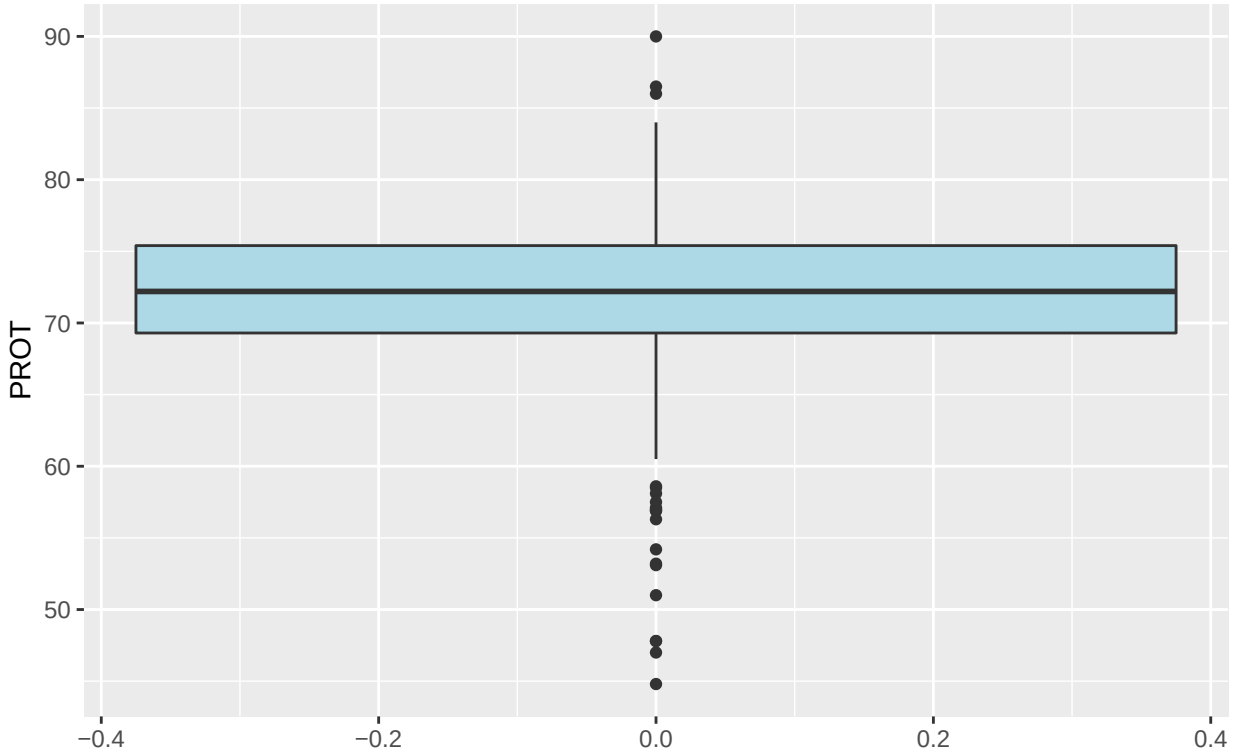
Median	Mean	SD
72.2	72.04414	5.402636

Standard Deviation is 5.4

Histogram of the PROT attribute:

Data Analysis

PROT



Exploring the quartile ranges and values

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   44.80   69.30   72.20   72.04   75.40   90.00       1
```

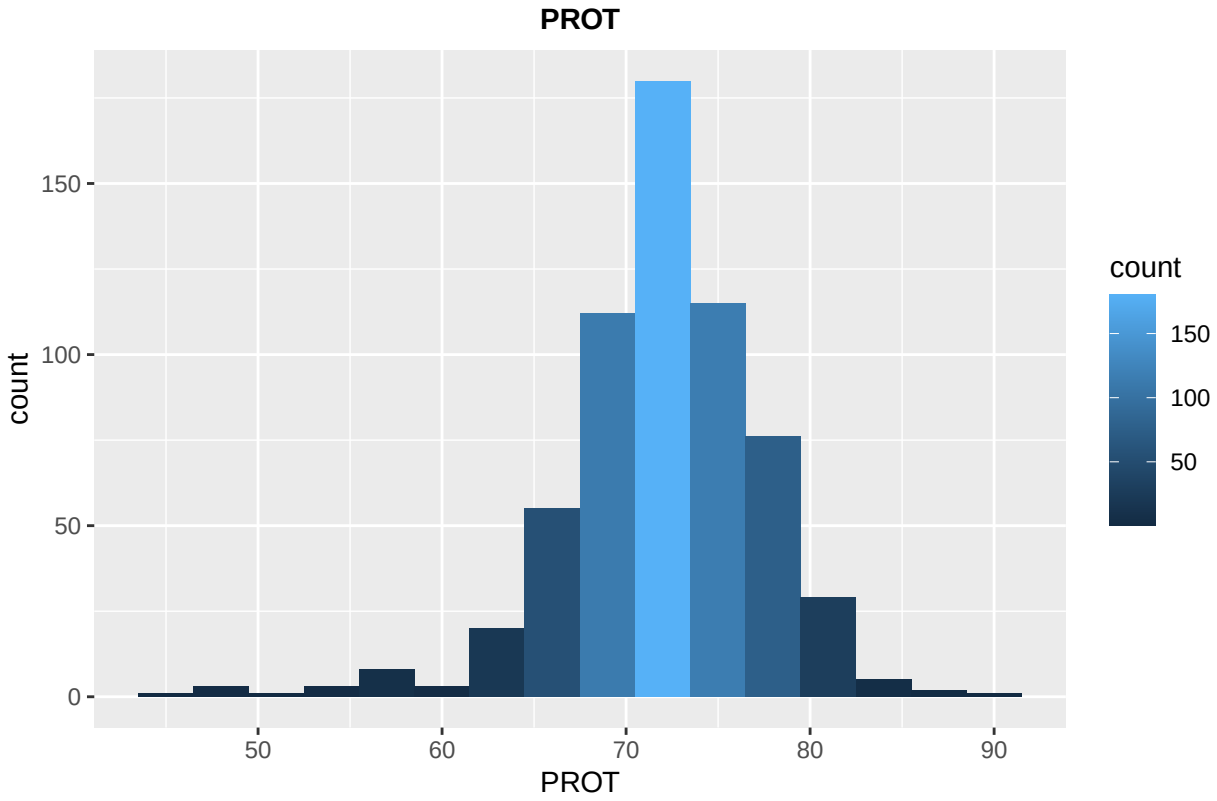
```
## [1] "IQR: "
```

```
## [1] 6.1
```

As seen from the summary above, 1st and 3rd Quartiles are 69.3 and 75.4 respectively. The inter quartile range between 1st and 3rd quartiles is 6.1 Max is 90, whereas Min is 44.8 There is 1 observation with NA

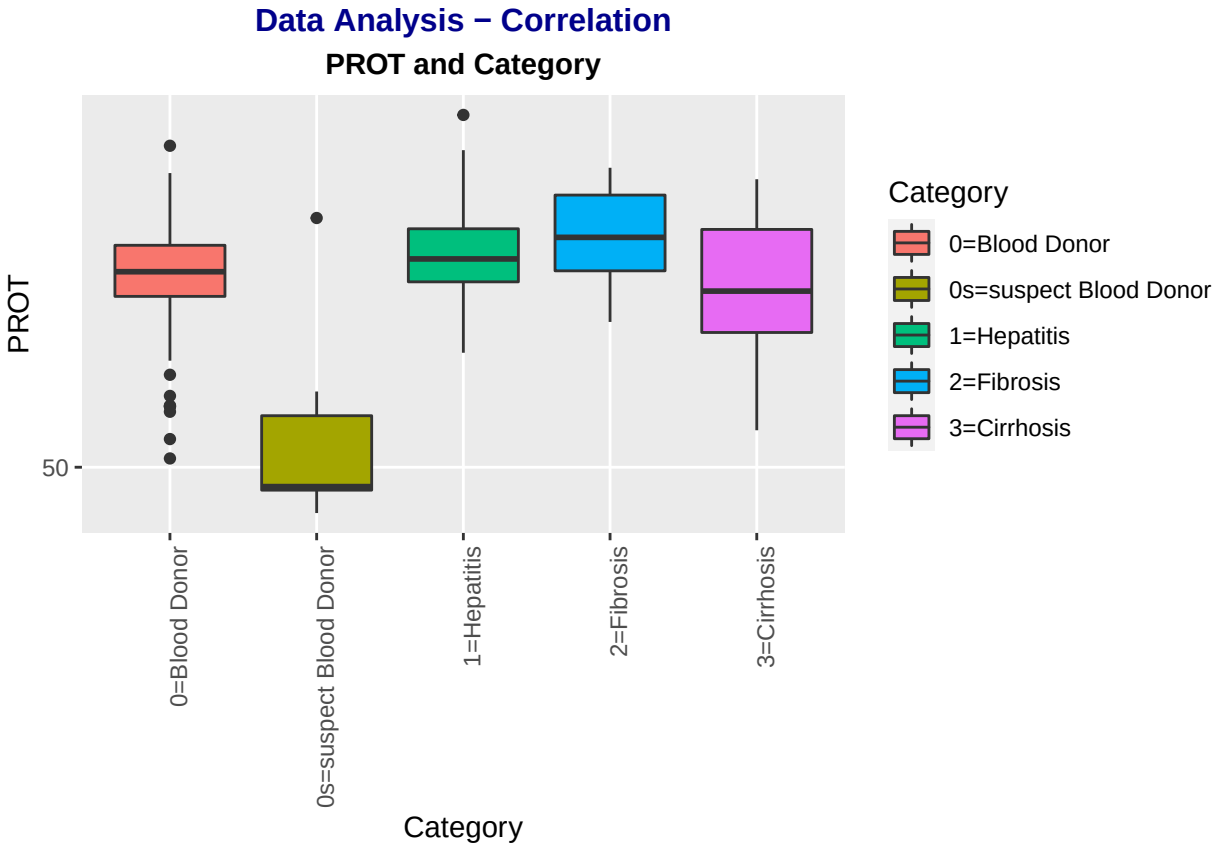
Below is a boxplot to visualize the IQR

Data Analysis



Histogram resembles a normal distribution

PROT's co-relation with Category demonstrated visually via boxplot



0s=Blood donor suspect is completely distinguishable from the other categories

2.2 Data Cleansing

Missing data such as the NAs observed above can deteriorate the statistical power of a study and can produce biased estimates. This in turn can lead to invalid inferences

Each row may have more than 1 NAs From Hepc, we will be omitting rows with atleast 1 NA

```
## [1] "No. of rows before cleansing:"
## [1] 615
## [1] "No. of rows after cleansing:"
## [1] 589
## [1] "No. of rows dropped:"
## [1] 0.04227642
```

26 rows were dropped that may have had atleast 1 NA Percentage of dropped rows is 4%

2.3 Correlation among the 10 predictors

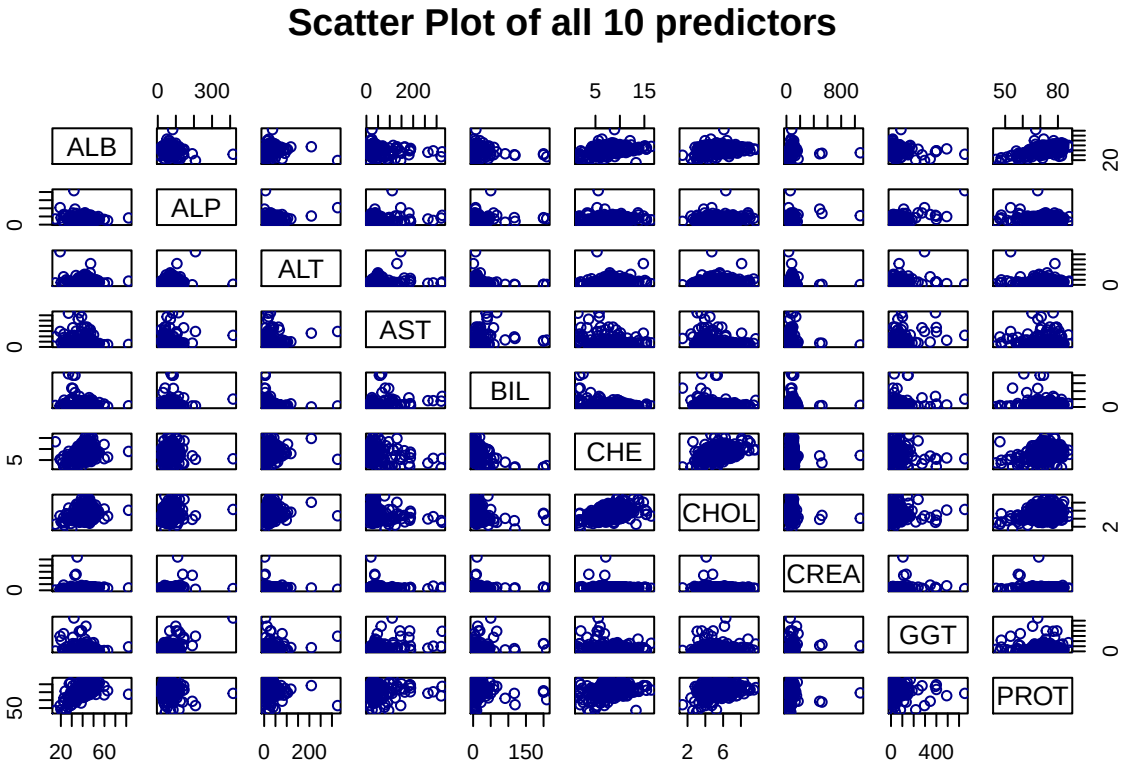
Correlation doesn't imply causation, however below table is used to understand the magnitude of correlation

very highly correlated - 0.9 to 1.0 highly correlated - 0.7 to 0.9 moderately correlated - 0.5 to 0.7 low correlation - 0.3 to 0.5 very less correlation - less than 0.3

Only two features are moderately correlated -> ALB & PROT have correlation of .571

The following pairs have low correlation (While the rest of the pairs have very little correlation) ALB<>CHE ALP<>GGT AST<>BIL AST<>GGT BIL<>CHE CHE<>CHOL CHE<>PROT

Let us Visualize this Correlation via Scatter plot



2.4 Principal Component Analysis

Using PCA, we will describe variance in the data. It would help us find obvious clusters. Though PCA reduces dimensionality it won't reduce the features

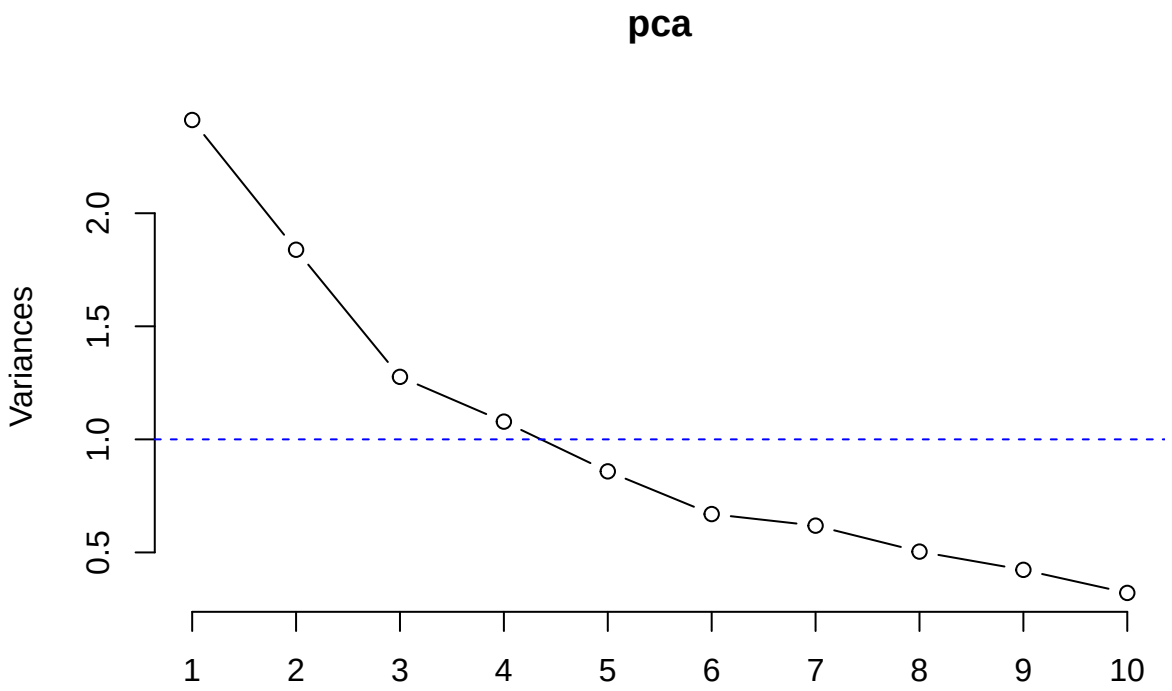
```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.5531 1.3559 1.1298 1.0385 0.92654 0.81835 0.78637
## Proportion of Variance 0.2412 0.1839 0.1277 0.1078 0.08585 0.06697 0.06184
## Cumulative Proportion  0.2412 0.4251 0.5527 0.6606 0.74642 0.81339 0.87522
##                           PC8    PC9   PC10
## Standard deviation     0.70971 0.6504 0.5666
## Proportion of Variance 0.05037 0.0423 0.0321
## Cumulative Proportion  0.92559 0.9679 1.0000
```

10 principal components are displayed in the decreasing order of the standard deviation (PC1 being 1.553 and PC10 being 0.5666)

PC1 accounts for >24% of total variance in data Using the first 8 components, we can account for >92% of total variance

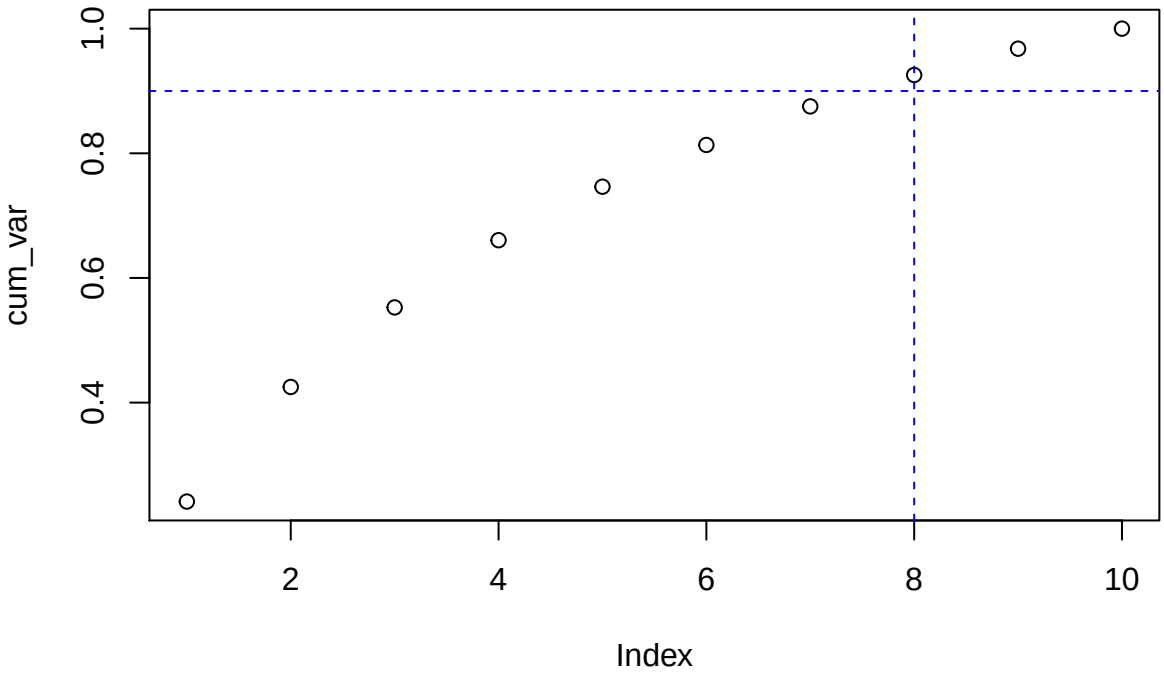
eigenvalues <1 would mean that the component explains less than a single explanatory variable, hence discarding them

Below is a Screeplot to visualize selection of factors



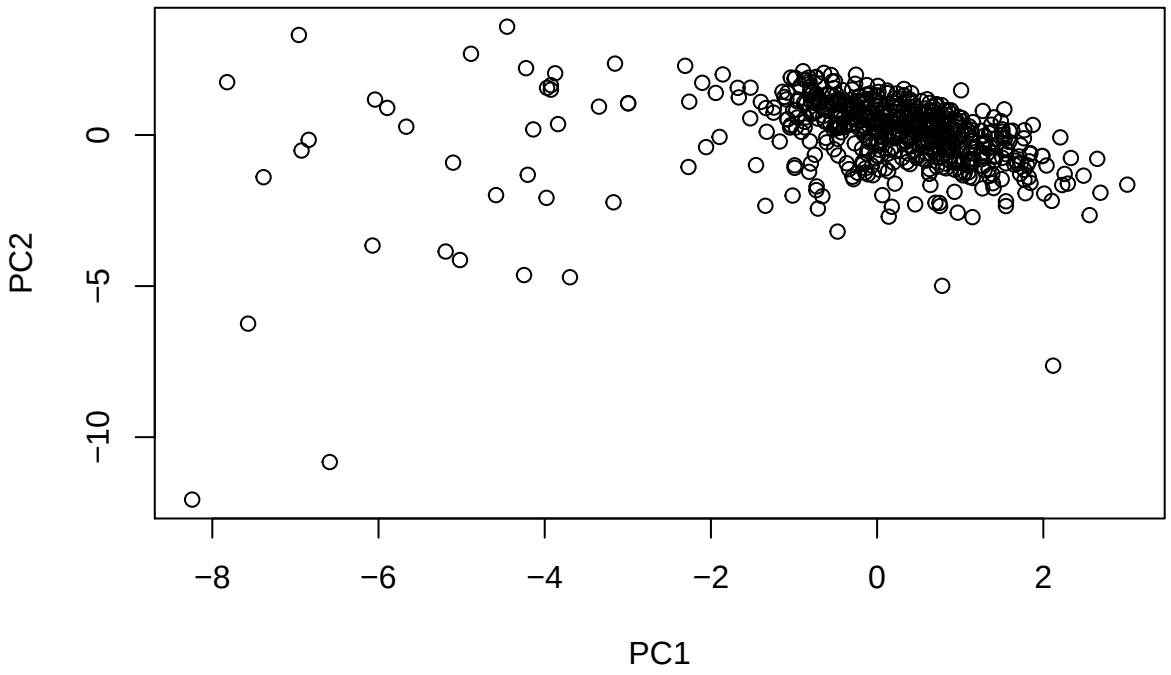
Till PC4, eigen values were >1

Plot to illustrate cumulative variance:

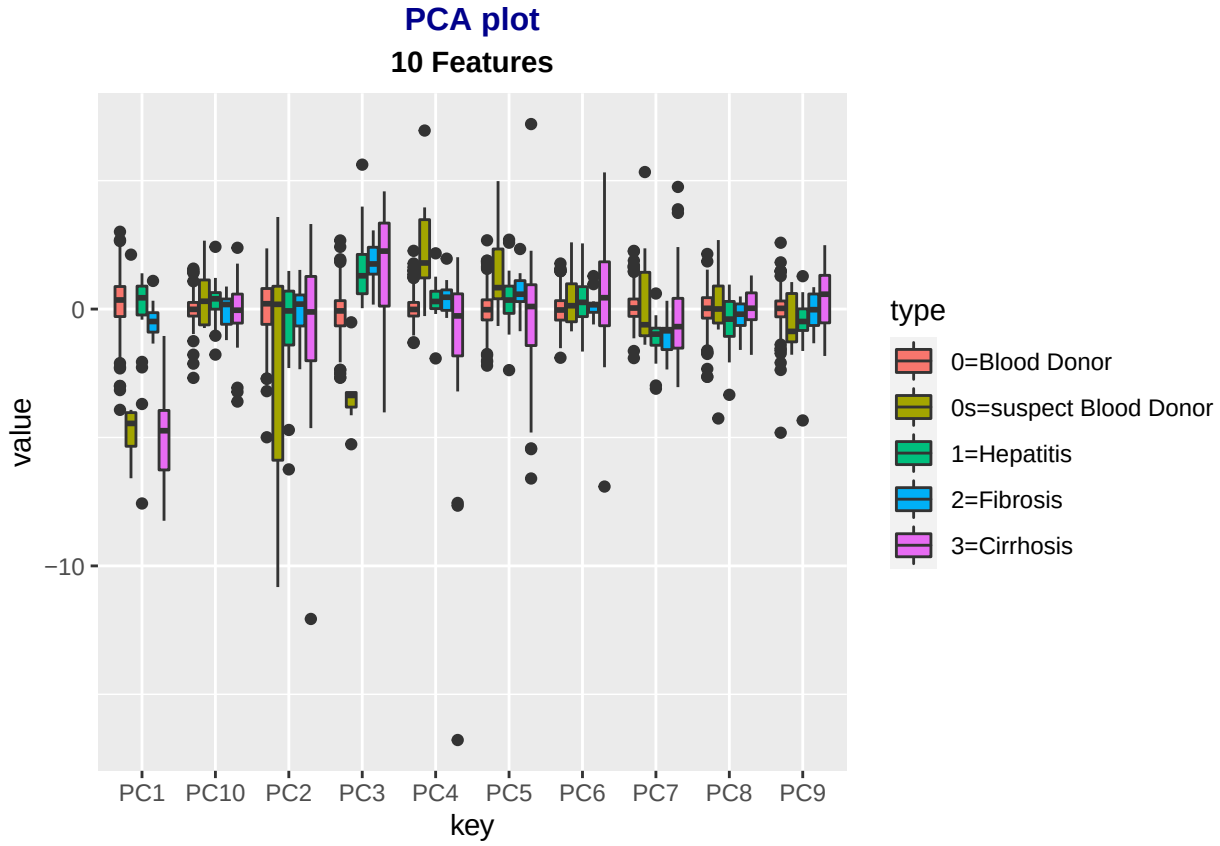


PC1 to PC6 explains >90% variance

Plotting the first 2 principal components to explain >42% variance



Plotting the PCs with Category



From the plot, we can see some separation to detect Hepatitis from a blood donor in the lower PCs

2.5 Data Partition

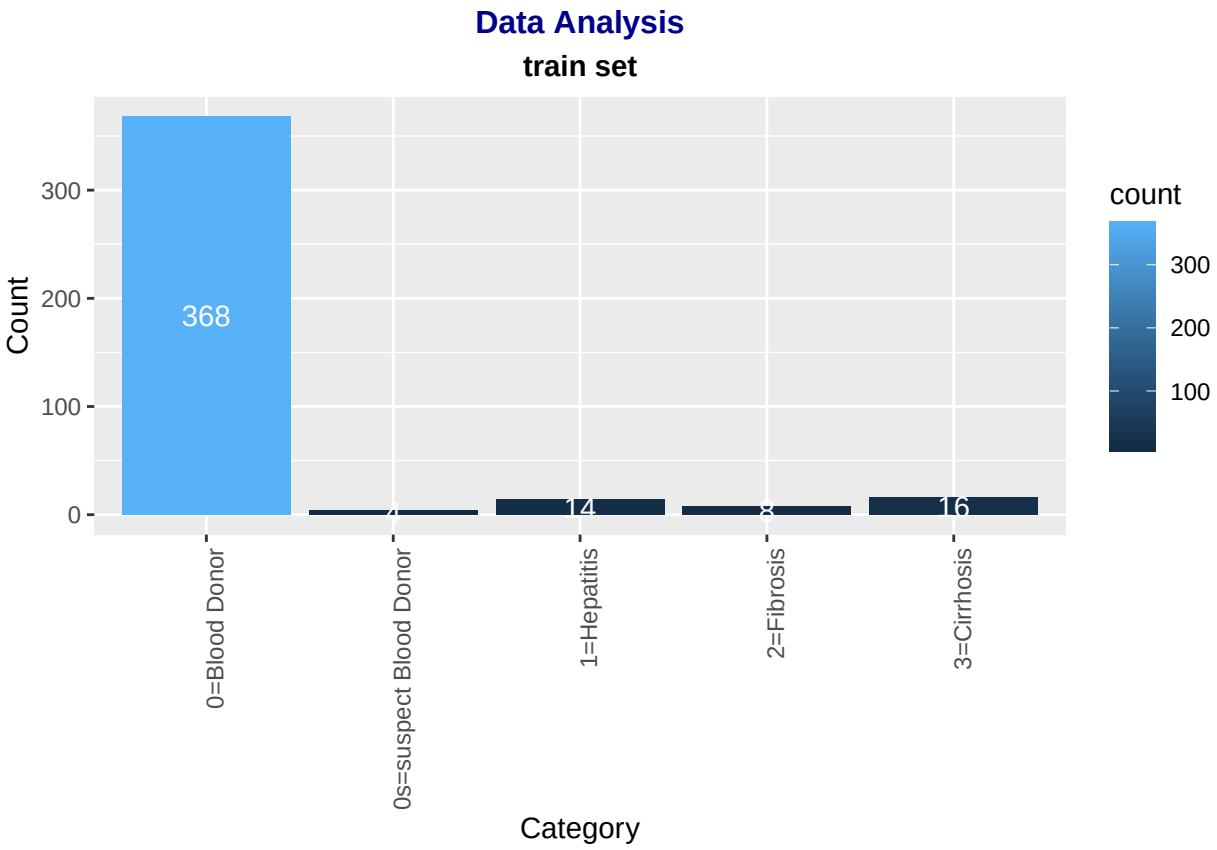
We will be Splitting the Hepc dataset into train and test sets

Owing to the smaller dataset size (600+ observations with about 10 predictors), it's a 70:30 split used here for training and test datasets respectively

Data check: train set, Category

Category	Count
0=Blood Donor	368
0s=suspect Blood Donor	4
1=Hepatitis	14
2=Fibrosis	8
3=Cirrhosis	16

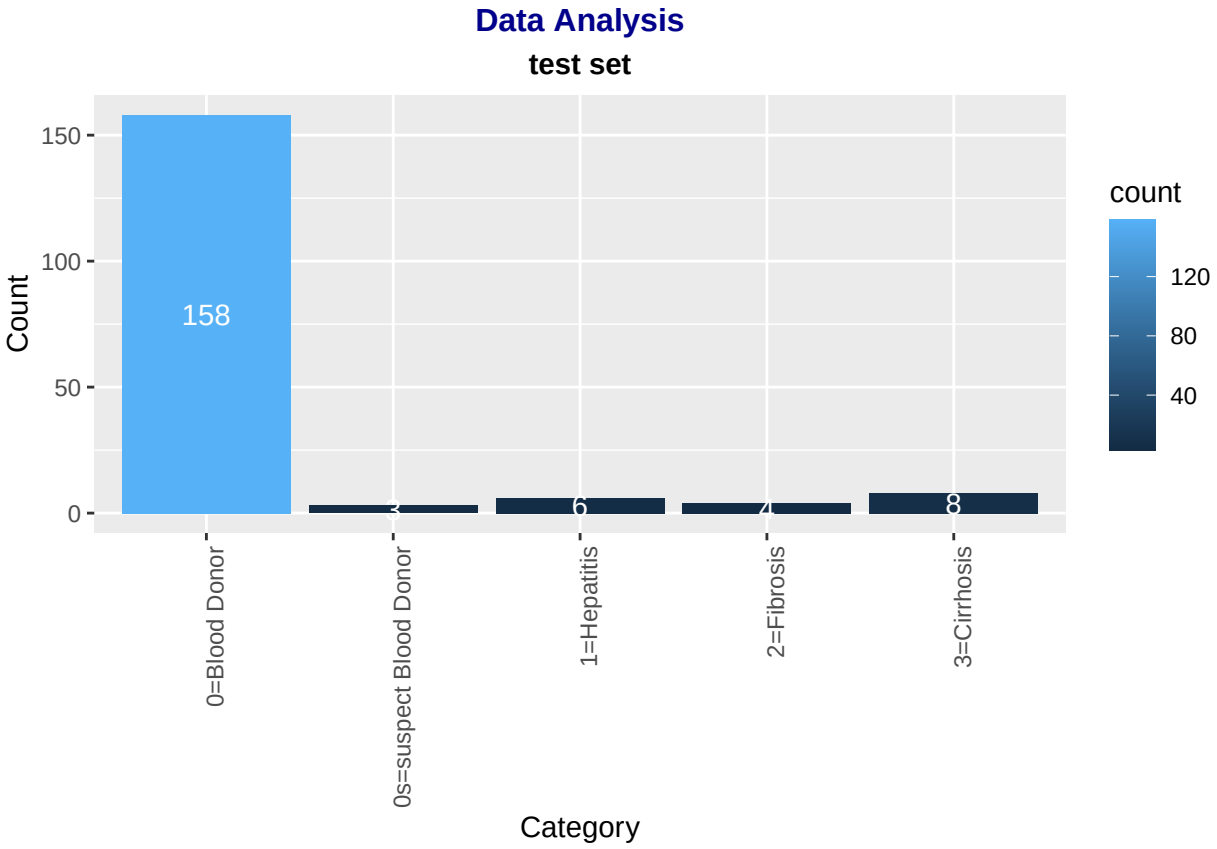
train_set is a dataframe with 410 observations
As can be seen from the dataset, Blood Donors make up most of the Category



Data check: test set, Category

Category	Count
0=Blood Donor	158
0s=suspect Blood Donor	3
1=Hepatitis	6
2=Fibrosis	4
3=Cirrhosis	8

test_set is a dataframe with 179 observations



2.6 Applying the models

Our Objective is to identify presence of Hepatitis in the sample based on the qualities/features being analysed We will be using classification models to arrive at a good model to predict

A classification model takes all the data points and the feature values associated with each feature and feeds these inputs to the Algorithm which in turn gives us an output in form of probabilities and the number of these probabilities is equal to number of classes

2.6.1 LDA

Linear Discriminant Analysis or LDA is a dimensionality reduction technique. LDA is a supervised classification technique.

The goal of LDA is to project the features in higher dimensional space onto a lower-dimensional space in order to avoid the curse of dimensionality and also reduce resources and dimensional costs.

This category of dimensionality reduction is used in areas like image recognition and predictive analysis in marketing.

```
## $lda
## Linear Discriminant Analysis
##
## 410 samples
##  10 predictor
##   5 classes: '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis'
##
```

```
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 410, 410, 410, 410, 410, 410, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.9296462  0.587979
```

2.6.2 Naive Bayes

Naive Bayes is a simple but surprisingly powerful probabilistic machine learning algorithm used for predictive modeling and classification tasks. It is a popular algorithm mainly because it can be easily written in code and predictions can be made real quick which in turn increases the scalability of the solution.

Some typical applications of Naive Bayes are spam filtering, sentiment prediction, classification of documents, etc.

```
## $naive_bayes
## Naive Bayes
##
## 410 samples
##  10 predictor
##   5 classes: '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 410, 410, 410, 410, 410, 410, ...
## Resampling results across tuning parameters:
##
##   usekernel  Accuracy   Kappa
##   FALSE      0.9284819  0.6217676
##    TRUE      0.9316441  0.6252101
##
## Tuning parameter 'laplace' was held constant at a value of 0
## Tuning
##  parameter 'adjust' was held constant at a value of 1
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were laplace = 0, usekernel = TRUE
##  and adjust = 1.
```

2.6.3 K-NN

K nearest neighbors or K-NN Algorithm is a simple algorithm which uses the entire dataset in its training phase. Whenever a prediction is required for an unseen data instance, it searches through the entire training dataset for k-most similar instances and the data with the most similar instance is finally returned as the prediction.

This algorithm suggests that if you're similar to your neighbours, then you are one of them. Let us consider a simple example, if apple looks more similar to peach, pear, and cherry (fruits) than monkey, cat or a rat (animals), then most likely apple is a fruit.

K-NN can be used for both regression and classification predictive problems. However, in the industry it is mostly used in classification problems.

One of the biggest applications of K-Nearest Neighbor search is Recommender Systems.

```
## $knn
```

```
## k-Nearest Neighbors
##
## 410 samples
##  10 predictor
##   5 classes: '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 410, 410, 410, 410, 410, 410, ...
## Resampling results across tuning parameters:
##
##   k  Accuracy   Kappa
##   5  0.9278273  0.5666865
##   7  0.9229610  0.5126851
##   9  0.9189212  0.4647283
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 5.
```

2.6.4 Multinom

Multinomial Logistic Regression is a classification algorithm used to do multiclass classification. Multinomial Logistic regression is nothing but K-1 logistic regression models combined together to predict a nominal labelled data for supervised learning

```
## $multinom
## Penalized Multinomial Regression
##
## 410 samples
##  10 predictor
##   5 classes: '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 410, 410, 410, 410, 410, 410, ...
## Resampling results across tuning parameters:
##
##   decay  Accuracy   Kappa
##   0e+00  0.9270802  0.6357275
##   1e-04  0.9289428  0.6307459
##   1e-01  0.9351463  0.6657403
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was decay = 0.1.
```

2.6.5 Random Forest

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

```
## $rf
## Random Forest
##
## 410 samples
```

```
##  10 predictor
##   5 classes: '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 410, 410, 410, 410, 410, 410, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.9397115  0.6075706
##    6    0.9410130  0.6464906
##   10    0.9320673  0.6139288
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 6.
```

2.6.6 CForest

The main idea behind CForest is that many trees are built in parallel between the same start and goal states.

Key concepts of CForest are:

Every time a tree finds a better solution, it is shared with all other trees so that all trees have the best solution found so far. Trees are expanded into regions that are known to be beneficial. Samples that cannot lead to a better solution are immediately discarded. Trees are pruned every time a better solution is found. Those states in the tree that do not help to find a better solution are removed from the tree. CForest is designed to be used with any random tree algorithm under the following assumptions: The search tree has almost sure convergence to the optimal solution. The configuration space obeys the triangle inequality. That is, there exists an admissible heuristic.

```
## $cforest
## Conditional Inference Random Forest
##
## 410 samples
##  10 predictor
##   5 classes: '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 410, 410, 410, 410, 410, 410, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.9246039  0.4347628
##    6    0.9289506  0.5403938
##   10    0.9264936  0.5545705
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 6.
```

2.6.7 Neural Networks

A neural network classifier is a software system that predicts the value of a categorical value. For example, a neural network could be used to predict a person's political party affiliation (Democrat, Republican, Other) based on the person's age, sex and annual income.

There are many ways to create a neural network. You can code your own from scratch using a programming language such as C# or R. Or you can use a tool such as the open source Weka or Microsoft Azure Machine Learning. The R language has an add-on package named nnet that allows you to create a neural network classifier.

```
## $nnet
## Neural Network
##
## 410 samples
##  10 predictor
##   5 classes: '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 410, 410, 410, 410, 410, 410, ...
## Resampling results across tuning parameters:
##
##   size  decay  Accuracy   Kappa
##   1     0e+00  0.9053484  0.1889662
##   1     1e-04  0.9049960  0.1178186
##   1     1e-01  0.9186400  0.5342278
##   3     0e+00  0.9086029  0.2728349
##   3     1e-04  0.9099204  0.3109421
##   3     1e-01  0.9311955  0.6141060
##   5     0e+00  0.9118728  0.2799255
##   5     1e-04  0.9127897  0.4008770
##   5     1e-01  0.9392460  0.6595461
##
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were size = 5 and decay = 0.1.
```

2.6.8 Predictions and Accuracy of the 7 models

Here are the accuracies of the 7 models trained here

Table 15: Accuracy of the various models

	Accuracy
lda	0.9217877
naive_bayes	0.9385475
knn	0.9106145
multinom	0.9273743
rf	0.9329609
cforest	0.9162011
nnet	0.9608939

Table 16: Mean Accuracy

Mean Accuracy
0.9297686

0.929 is the mean Accuracy

2.6.9 Final model

We will now improve the accuracy using an ensemble

2.6.10 Ensemble

Ensemble methods aim at improving the predictive performance of a given statistical learning or model fitting technique. The general principle of ensemble methods is to construct a linear combination of some model fitting method, instead of using a single fit of the method.

An ensemble is itself a supervised learning algorithm, because it can be trained and then used to make predictions. The main principle behind the ensemble model is that a group of weak learners come together to form a strong learner, thus increasing the accuracy of the model.When we try to predict the target variable using any machine learning technique, the main causes of difference in actual and predicted values are noise, variance, and bias. Ensemble helps to reduce these factors (except noise, which is irreducible error). The noise-related error is mainly due to noise in the training data and can't be removed. However, the errors due to bias and variance can be reduced. The total error can be expressed as follows:

Total Error = Bias + Variance + Irreducible Error

There are two families of ensemble methods which are usually distinguished:

Averaging methods. The driving principle is to build several estimators independently and then to average their predictions. On average, the combined estimator is usually better than any of the single base estimator because its variance is reduced.| Examples: Bagging methods, Forests of randomized trees.

Boosting methods. Base estimators are built sequentially and one tries to reduce the bias of the combined estimator. The motivation is to combine several weak models to produce a powerful ensemble. Examples: AdaBoost, Gradient Tree Boosting.

Basic Ensemble Techniques

Max Voting: Max-voting is one of the simplest ways of combining predictions from multiple machine learning algorithms. Each base model makes a prediction and votes for each sample. The sample class with the highest votes is considered in the final predictive class. It is mainly used for classification problems.

Averaging: Averaging can be used while estimating the probabilities in classification tasks. But it is usually used for regression problems. Predictions are extracted from multiple models and an average of the predictions are used to make the final prediction.

Weighted Average: Like averaging, weighted averaging is also used for regression tasks. Alternatively, it can be used while estimating probabilities in classification problems. Base learners are assigned different weights, which represent the importance of each model in the prediction.

Reporting Accuracy of the Ensemble(Improvement in accuracy)

Table 17: Accuracy of the Ensemble

Accuracy
0.972067

With an ensemble of the above models, we achieved over 97% accuracy in predicting the presence of Hepatitis in the sample

3 Results

We analysed the dataset and the attributes available. We studied the correlation between the predictors as well as the outcome. Though Prinicpal Component Analysis works really well to summarize large datasets, our usage of this procedure here enabled us to interpret and visualize the underlying information and the importance of the principal components. We then applied various classification methods with the objective of improving accuracy. All the methods performed reasonably well w.r.t accuracy, however an Ensemble with "max voting" gave us an improvement over all the other methods and thus is the final method that was chosen to predict the presence of Hepatitis C in the sample. The final reported accuracy using Ensemble is over **97%**

4 Conclusion

Classification algorithms belong to a type known as supervised Machine Learning. It involves a model building, based on which the input variables are used to predict the outcome class. In this project, we have used the classification methods to build such models of prediction. The most important aspect of developing or applying a method/model is to understand the dataset and the various attributes available. As we start analysing the various predictors available in the dataset, it became clear that there are certain predictors that can be leveraged to make better predictions of the outcome class. We used PCA to better understand the data

The final method that is selected is the **Ensemble** which yielded an accuracy of over **97%**

Tuning the methods further for best parameters would have further resulted in an even better accuracy

4.1 Constraints

Trying out the various models and learning about them was very inspiring. There were however some models that didn't behave the way they were intended, atleast with the default parameters(without tuning)

Tuning is a time-intensive activity. Since the dataset in this project is small, it may not have been as time-intensive as some of the larger datasets that we maylike to work on. In such scenarios, we would have to select appropriate methods and tuning for the intended accuracy. Since Ensemble combines such models, the number of models that we choose and the size of the dataset would have an impact on the performance(time and memory) of the approach

4.2 Future Work

It is important to find the best approach for the dataset at hand, as large datasets would require efficient implementations. There is future scope to learn and try out the various other ML algorithms, such as unsupervised learning and reinforcement learning

Choosing the right datasets and exploring other competitive models such as collaborative filtering and tensorflow recommendation systems would be something that would add a lot of value in understanding how personalized recommendation systems work. These methods can be leverages in the field of banking (that is currently undergoing a huge digital transformation) where product cross-selling and advisory based solutions for clients can be made effectively with data as the driver.

5 References

- Harvardx Data Science Professional certificate **(Instructor: Rafael Irizarry)**
- rdocumentation
- The Comprehensive R Archive Network

- kaggle dataset
- Hepatitis C
- Fibrosis & Cirrhosis
- Hepatitis C dataset
- Reserachgate
- towardsdatascience
- citeseerx
- stackexchange
- visualstudiomagazine
- aidsmap_1
- aidsmap_2
- stackoverflow