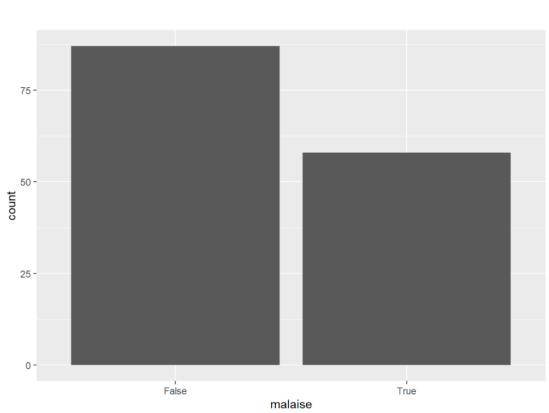
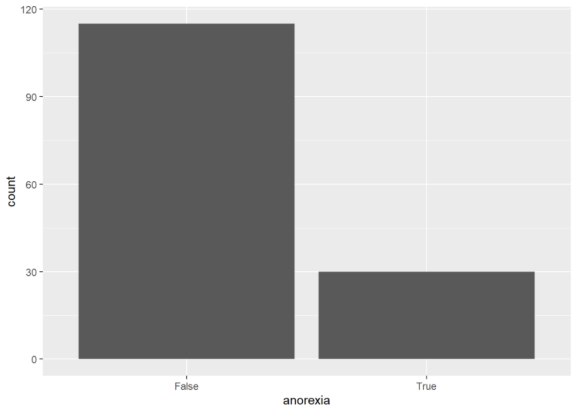
Sunniul Alam
SXA180118
CS4375.002
February 27th. 2022

CS 4375 - Assignment#1

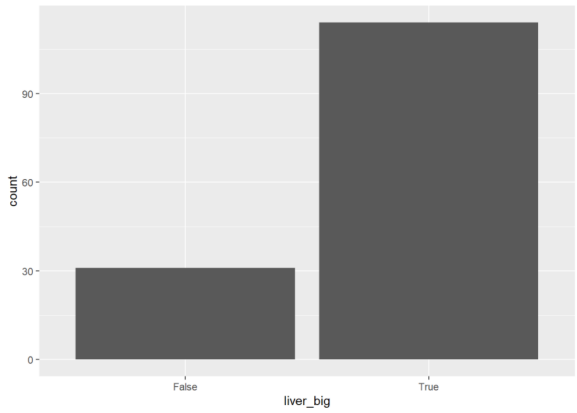
1.

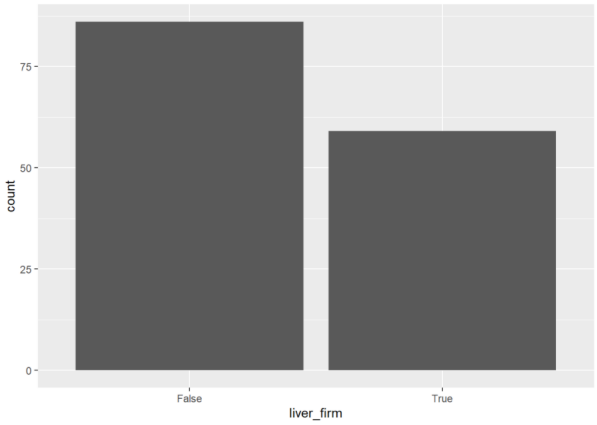
- There are 155 Objects and 20 Attributes.
- Type of Each Attribute
    - Age- Ratio
    - Sex- Nominal
    - Steroid- Nominal
    - Antivirals- Nominal
    - Fatigue- Nominal
    - Malaise- Nominal
    - Anorexia- Nominal
    - Big Liver- Nominal
    - Firm Liver- Nominal
    - Palpable Spleen- Nominal
    - Spider- Nominal
    - Ascites- Nominal
    - Varices- Nominal
    - Bilirubin- Interval
    - Alk_Phosphate- Interval
    - Sgot- Interval
    - Albumin- Interval
    - Protime- Ratio
    - Histology- Nominal
    - Class- Nominal
- Yes, there are several missing values in the dataset. In order to improve the quality of the dataset, the rows that are missing 3 or more continuous values are removed. Any remaining rows with missing values for continuous attributes are defaulted to the mean of that row. All discrete attributes have been defaulted to FALSE.

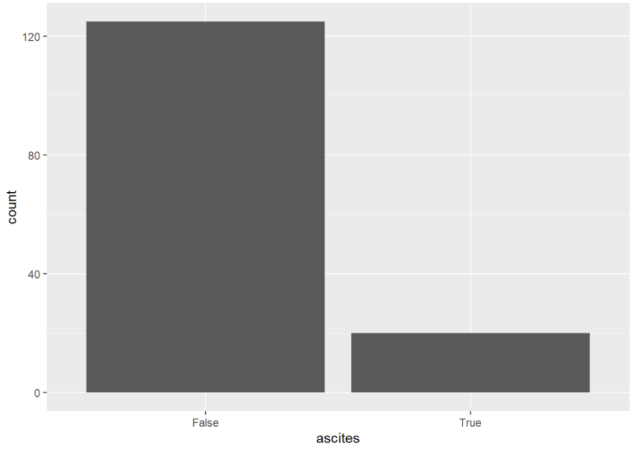




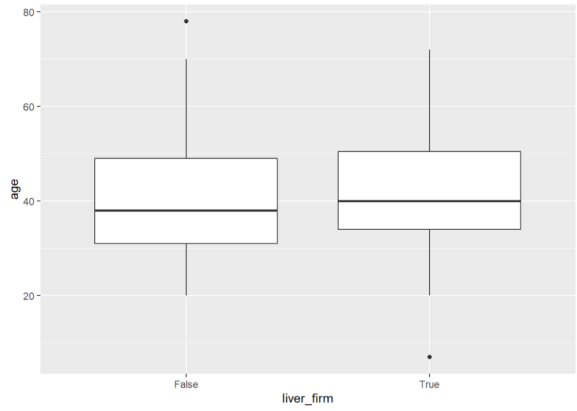
Sunniul Alam
SXA180118
CS4375.002
February 27th. 2022

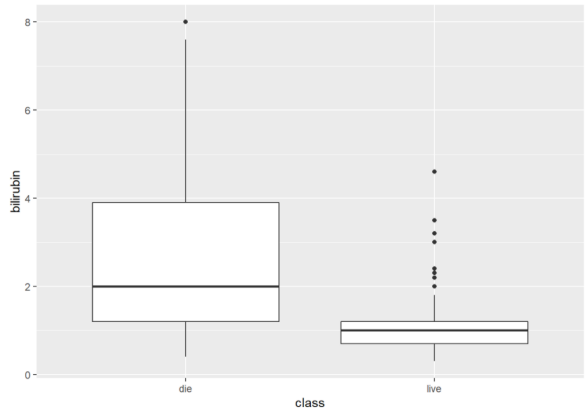




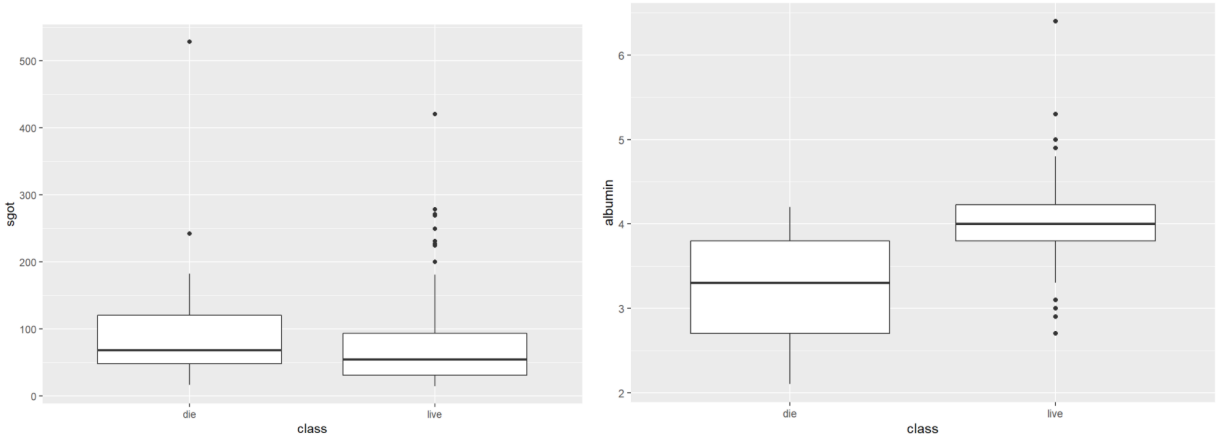


- The bar graph of malaise amongst hepatitis patients shows that a slight majority do not experience it. The bar graph of anorexia amongst hepatitis patients shows that a majority of hepatitis patients are not anorexic. The bar graph of liver size amongst hepatitis patients shows that the majority have big livers. However, a slight majority of hepatitis patients did not have firm livers. A majority of hepatitis patients did not have ascites, as seen through the bar graph.





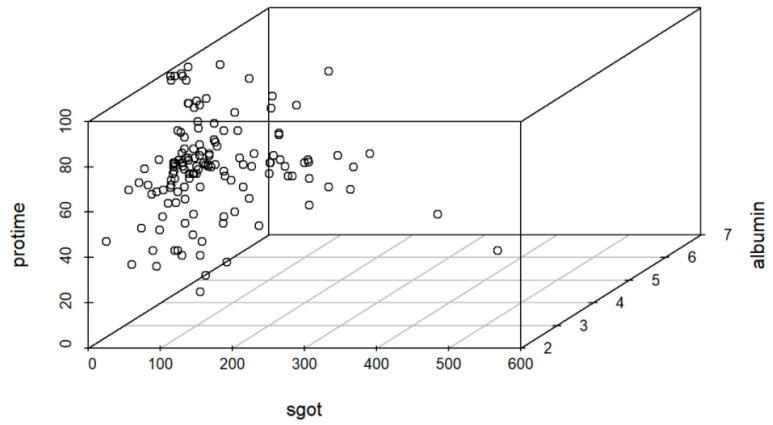
Sunniul Alam
SXA180118
CS4375.002
February 27th. 2022

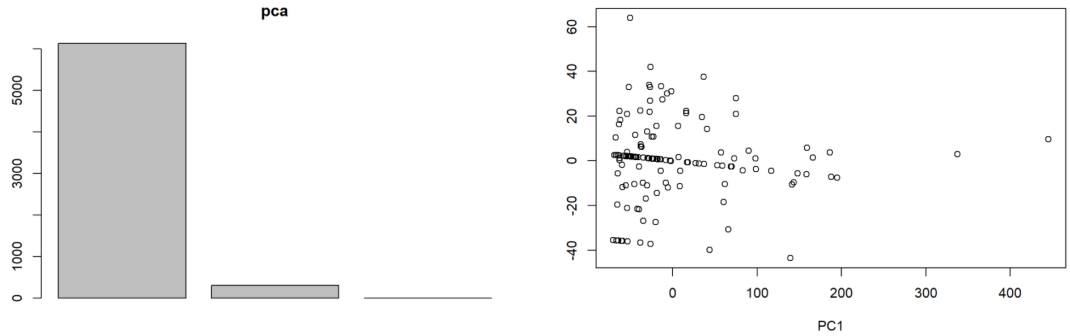


- The boxplots for liver firmness vs age shows that the attributes are independent from one another, as seen through the overlapping interquartile ranges. The boxplots for class vs bilirubin show a higher bilirubin distribution amongst dead patients, with evident outliers on both boxplots. The boxplots for class vs sgot show no distinguishing factors between the two attributes. The boxplots for class vs albumin, however, show a lower albumin level in patients who are deceased.

2. In relation to bilirubin and sgot as a subject of sex, a higher level was found in females than males. For Alk_phosphate, a greater level was seen in males. In terms of albumin and protime, the data shows that there is no statistical difference between the two sexes for those two attributes.
3. The euclidean distance was calculated from a random sample of 10 from the data set. The same sample of 10 was used for both the standardized and unstandardized calculation of euclidean distance. However, due to the fact that the values are random, the results are subject to the compiler.
4. The sample set is based on the improved data set, which contains 142 objects. If we have a sample based on that data of a size of 500, that means that 358 new objects would need to be created. Since all of those objects are based on the original dataset, they would all be considered duplicate values. This means there are a total of 358 duplicate values in our sample.

Sunniul Alam
SXA180118
CS4375.002
February 27th. 2022





5. The scatterplot3d shows that there is a positive correlation between sgot and protime. The scatterplot3d the negative correlation between protime and albumin. It also shows that there exists no real linear correlation between sgot and albumin.
6. The interval width when using equal intervals was 17.75. There were 10 objects in [7,24.8), 77 objects in [24.8,42.5), 47 in [42.5, 60.2), and 11 in [60.2, 78]. The interval width when using equal frequencies was 25, 7, 11, and 28. There were 35 objects in [7,32), 35 objects in [32,39), 36 objects in [39,50), and 38 objects in [50, 78].
7. The correlation matrix is based upon the random sampling of 50 objects from the dataset. The correlation matrix gives us a Pearson Correlation Coefficient of 0.2431274. Since the coefficient is relatively low, there is not a linear relationship between the two attributes.
8.
Cosine Similarity = 1

Cos(x,y) = x . y / ||x|| * ||y||

Cos(x, y)= (1*1 + 1*1 + 0*0 + 0*0) / (sqrt(1^2 + 1^2 +0^2 +0^2)*sqrt(1^2 + 1^2 +0^2 +0^2))

Cos(x, y)= 2/sqrt(2)*sqrt(2)= 2/2= 1

Sunniul Alam
SXA180118
CS4375.002
February 27th. 2022

Correlation = undefined

corr(x,y)= covariance(x,y)/ (standard.deviation(x)*standard.deviation(y))

xbar= ¼ (1 + 1 + 0 + 0)= ½ ybar= ¼ (1 + 1+ 0 +0)= ½

cov(x,y)= ⅓((1-.5)*(1-.5) + (1-.5)*(1-.5) + (0-.5)(0-.5) + (0-.5)(0-.5))= ⅓(¼ +¼ -¼ -¼)
= 0

std.x=sqrt(⅓(1-0.5 + 1-0.5 + 0-0.5 + 0-0.5))= sqrt(⅓ (0.5+0.5 -0.5 -0.5)= sqrt(0)=0

std.y=sqrt(⅓(1-0.5 + 1-0.5 + 0-0.5 + 0-0.5))= sqrt(⅓ (0.5+0.5 -0.5 -0.5)= sqrt(0)=0

cov(x,y)/std.x*std.y= 0/0*0= 0/0= undefined

Euclidean Distance = 0

d(x, y) = ||x-y||= sqrt((1-1)^2 + (1-1)^2 + (0-0)^2 + (0-0)^2)
= sqrt(0 + 0 + 0 + 0)
= 0

Jaccard Coefficient = 1

J= f11/(f01 + f10 + f11) = 2/ (0 + 0 +2)= 1

9.

a. If duplicate objects were found in the data set, then it could falsify the K nearest neighbors due to the fact that the nearest neighbor to K could be a duplicate of K.
b. This problem can be solved by going through the data objects before the for loop and removing all duplicates.

10.

(a) Discrete, Quantitative, Nominal
(b) Continuous, Quantitative, Ratio
(c) Discrete, Qualitative, Ordinal
(d) Continuous, Quantitative, Ratio
(e) Discrete, Qualitative, Ordinal
(f) Continuous, Quantitative, Interval
(g) Discrete, Quantitative, Ratio
(h) Discrete, Qualitative, Nominal
(i) Discrete, Qualitative, Ordinal
(j) Discrete, Qualitative, Ordinal
(k) Continuous, Quantitative, Interval
(l) Continuous, Quantitative, Ratio
(m) Discrete, Qualitative, Nominal