HCV Dataset Analysis

Department of Information Technology, York University

AP/ITEC 2600 Section A: Introduction to Analytical Programming (Fall 2024-2025)

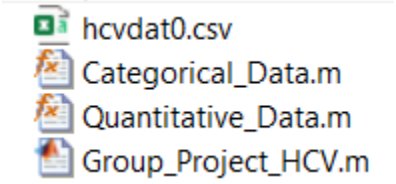
Tuesday, April 4, 2025

Introduction

After a careful overview of all three dataset options, We decided to work on the Hepatitis C virus (HCV) data because, aside from information technology, the medical science field is our interest. Following up on the decision, we conducted additional research on what exactly this Hepatitis C virus entails because we were unfamiliar with many of the column header names within the Excel file (ALB, BIL,etc). The Hepatitis C Virus is a serious health issue that affects the liver and can progressively worsen into severe conditions like cirrhosis and fibrosis; we noticed that in the dataset, there are specific biochemical markers (certain enzymes found in liver) that can directly track these conditions, such as AST or ALT. These biochemical markers names can be found at the column headers within the HCV dataset and all relate to liver-related diseases (however, some are more related to having an unhealthy pair of kidneys) and overall liver functionality. It's also important to note that HCV is most commonly contracted from infected blood, but can also be spread through sexual contact, so we will use the following dataset to try and track some of these values (age, category, and sex) to see any possible correlations or general relationships between this information. Before analyzing the dataset, we predict that the "Age" category may positively correlate with liver disease because as humans get older, our bodies tend to have a harder time protecting or adapting our systems to viruses and diseases. We also predict that the "Sex" category may show significantly different results when comparing males to females as the biological factors may skew or offset the data representation. Ultimately, our goal is to discover as many key patterns, trends, and relationships as possible after completing a thorough analysis of the Hepatitis C Virus dataset.

Data Analysis

When it comes to the Analysis of the data, we have to check how we are going to process it. For that, we coded a traversal loop through different columns, and then analyzed each column one by one first. The dataset consists of 615 people (a few are patients with a record of having infectious liver) and 15 subclasses, such as categorical data (categories of people, sex) and numerical data (Age, ALB , ALT , AST). From the lectures, we figured out the difference between the two types of data, categoricals are those we cannot measure but we could count them, whereas numerical ones could have more statistics applied to them due to the fact that they are numeric so we should compute different relations and trends in between these values.
By noticing this, we thought that it would be a good idea to divide these types into two separate user-defined functions so the analyzing of the two types would not be a burden anymore.
Thus, we defined those two functions as '*Quantitative_Data'* and *'Categorical_Data'*.
In addition, we define the main script ('*Group_Project_HCV'*) as a gateway to firstly traverse through the columns of data and secondly guide each traverse through one of the user defined functions.



These scripts create efficient and robust code so that we can easily access every cell of data and visually represent those quantities through MATLAB's graphical interfaces. As a law from mathematics, we have specific figures to display both categorical and Quantitative (numeric)

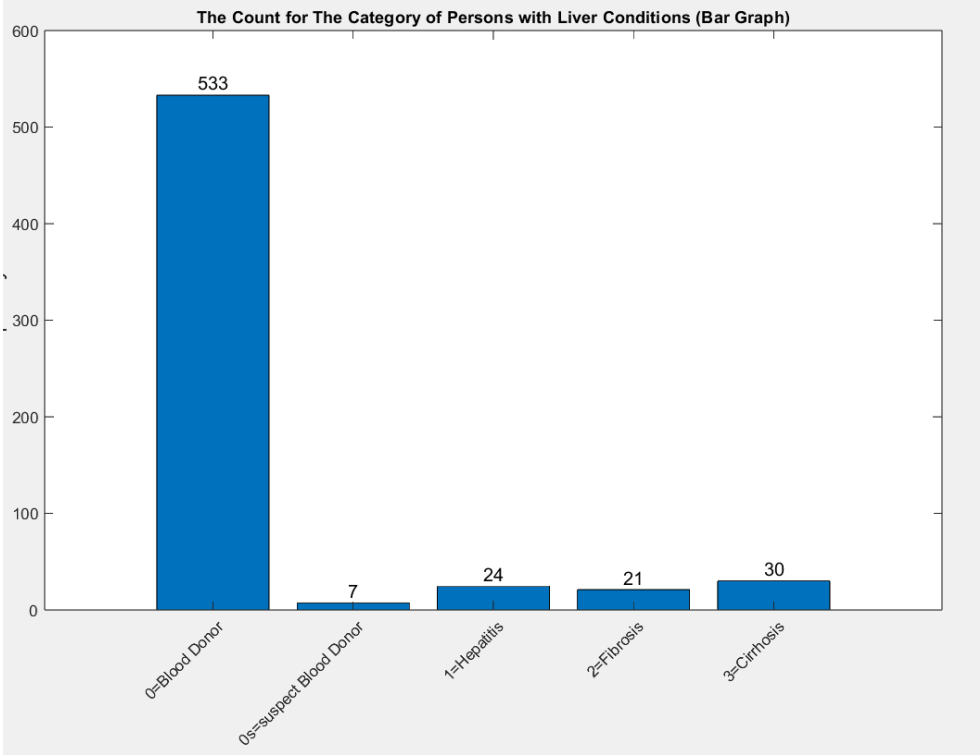
variables

Categorical columns of data: (*Categorical_Data* user-defined function in our script) For these types of values,we could use bar graphs to check for the quantity (count) of a categorical variable or a pie chart to check for the portion amount (percentage) of a categorical variable.

Bar Charts:

Going into the details of bar graph figures, we could clearly see that the height of each bar indicates their quantity or how many times there has been the very same categorical value.

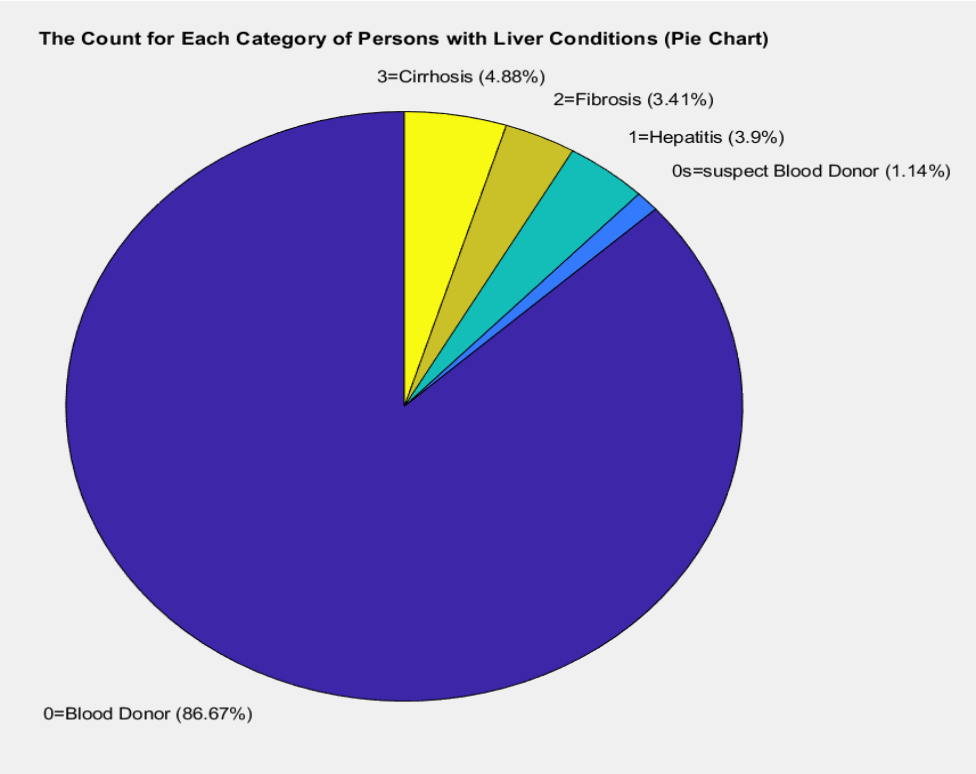
An Example on different types of people mentioned in a bar chart:



Pie Charts:

For the pie chart, it gathers all values of the categorical column, divides them by 615 (size of array) and finds out each one's percentage.

An Example on the percentage of different types of people mentioned in a pie chart:



Quantitative columns of data: (*Quantitative_Data* user-defined function in our script)

The numeric or quantitative types are those that need more computation and statistical figures as we are talking about two numeric columns of data here,

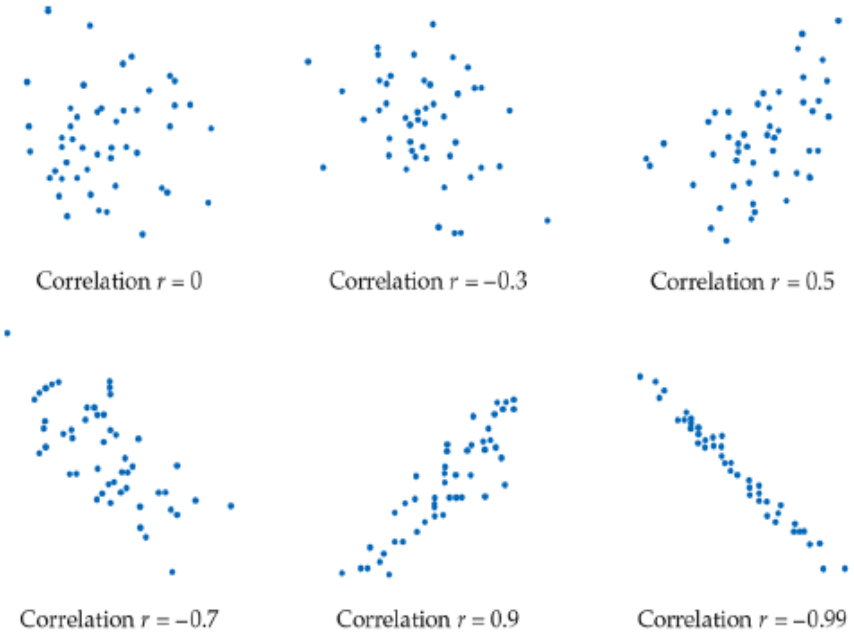
For these types of data, we could use countless types of figures and graphs. Despite that, we use the most known ones which are Scatterplots, Histograms, Boxplots, and Density curves with the use of histograms again. It is important to note that Any two numerical values in our data, so for simplicity reasons and avoidance of hundreds of figures, we decided to do those with a fixed numeric variable which is our index, and let the other variable be the other numeric variables. For example, we have a scatter plot for Index and Age to check the different patient's age distribution, whereas we don't have a scatterplot for Age and an enzyme like ALB, this could cause a massive overload for our system and matlab which is not what is intended to be accomplished for this course. (considering 4 figures at minimum, it would result in having 392 figures! 182 Scatterplots, 14 histograms, 14 Density curves, 182 Boxplots) for our numeric columns however, we have only 59 figures including 11 summary UI figures that we discuss about them shortly after in this report.

Scatterplots:

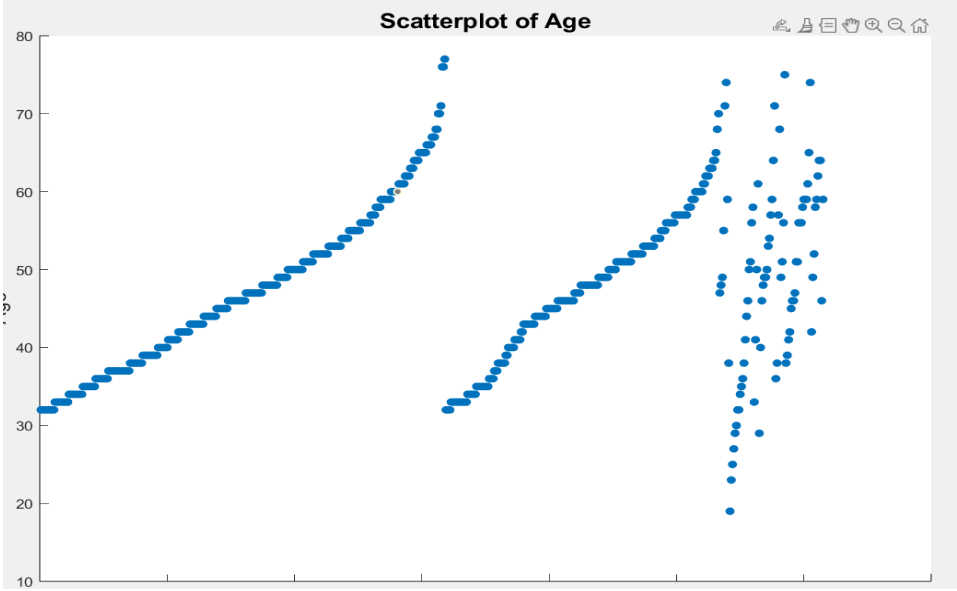
These figures help us understand the distribution of data across the range of our numerical values, any two numerical values could form a scatterplot, thus we used the index as the fixed variable on one side and then changed the other one so we could find out how data is distributed.The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis.

Each case corresponds to one point on the graph.

Scatterplots are a great tool to find out about outliers as they stand out from the rest of the data values in the Scatterplot. It's also possible to estimate the correlation coefficient of each variable using scatterplot. An ultimate guide is provided below one how to find out each scatterplot's correlation coefficient. (the amount is specified as the variable 'r')



An Example on the Distribution of different types of people's age mentioned in a scatterplot:



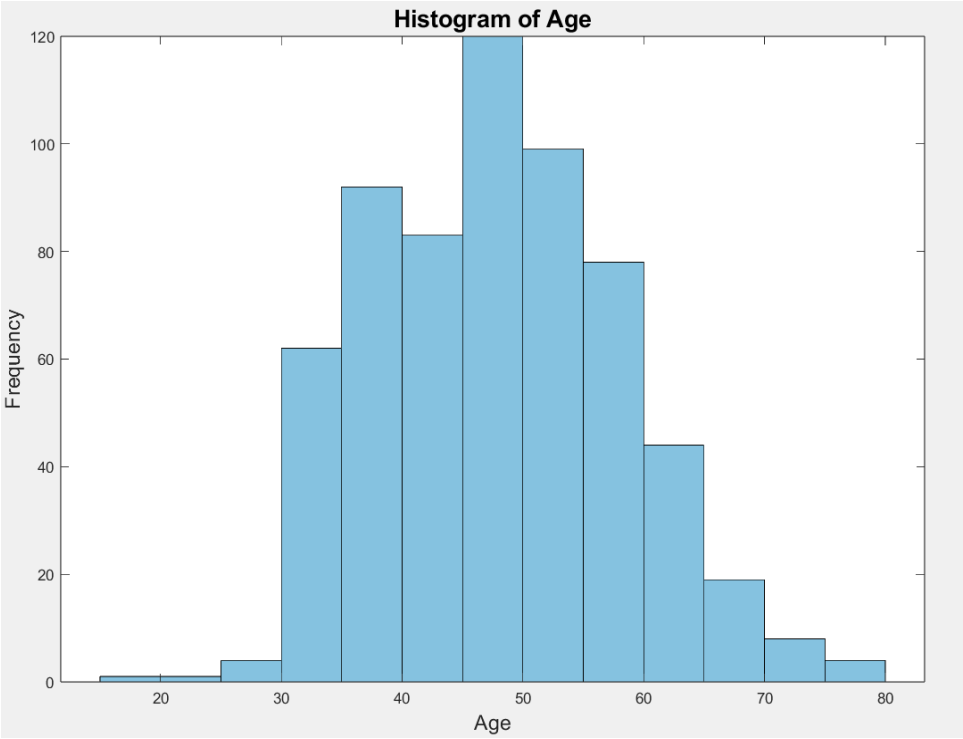
Histograms:

Another good way to find the distribution of the data values, but we categorize them into different bins (10 to 20, 20 to 30 , etc) automatically so then the histogram shows which range has the most amount of values.

Histograms also play a key factor on finding out the density curve to observe which parts reach to the pinnacle (the mean) of the density curve, how much percent of values lie within a certain range, etc.

An Example on the Distribution of different types of people's age mentioned in a Histogram:

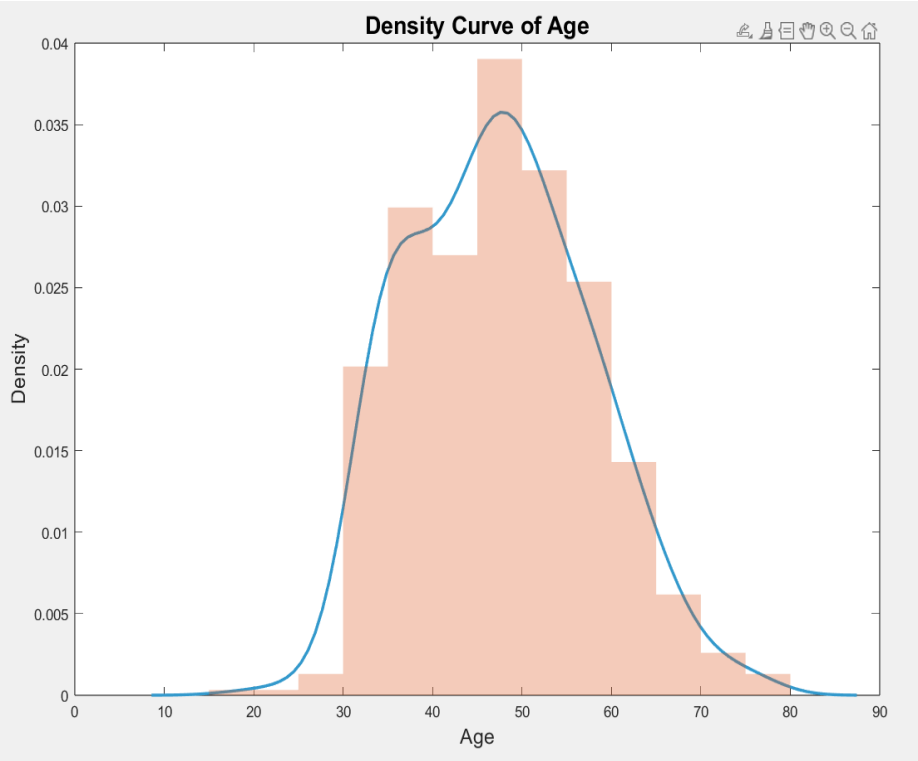
(Frequency is our indexes)



Density Curves: Histograms but instead of counting the values and putting them on different bin sections, we now make a curve that covers all the values we are having from the start to the end of the variable's range. A good way to find out how much percentage of values lie less or more than a certain value.

An Example on the Distribution of different types of people's age mentioned in a Density curve:

(Red figure is our beforehand computed Histogram)

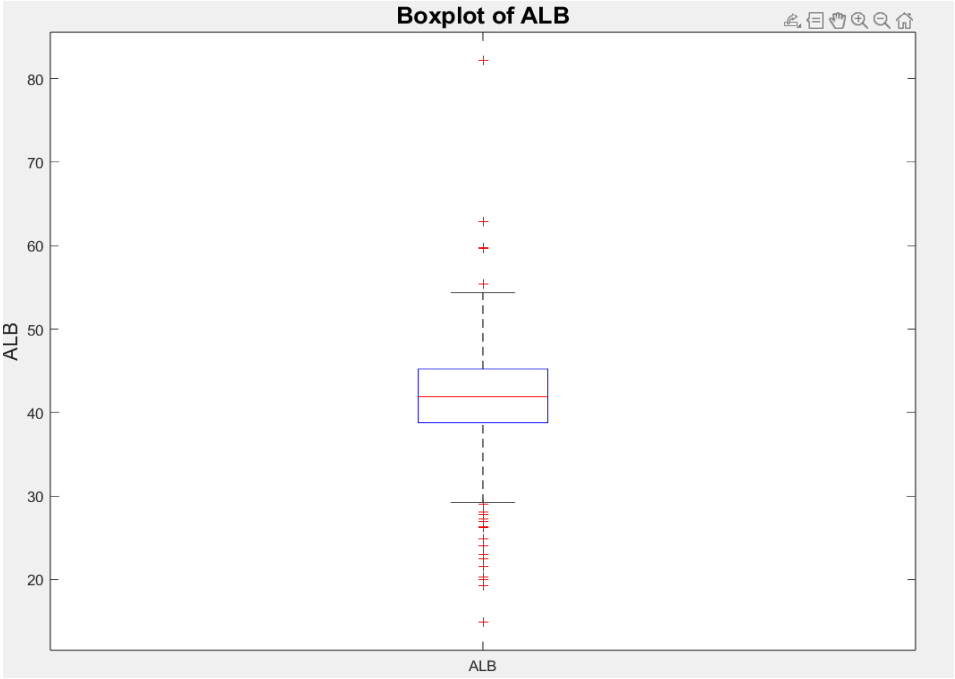


BoxPlots: a good way to implement a box of summary numbers (minimum, median, max. First and third quartile).

Every Number that is shown as a red cross could be a potential Outlier (Note: Mean value could also be counted as a potential outlier for this figures)

An Example on the Distribution of different types of people's ALB mentioned in a Box Plot:

(Red crosses could be potential outliers by a marginal difference)



Next step is to calculate statistical measures between the variables (variance, standard deviation, mean, median, correlation coefficients, etc) using MATLAB functions and put them separately into a UI figure which is basically a texture embedded in a window that will be displayed when a traverse through one column happens.

Note: Numbers for correlation coefficient and covariance might be the same as they appear on the other variable's UI figure.

Example for this occurrence: covariance for Age in ALB is -11.44. It appears to be the same in both UI figures for both summaries.

We will go into details for these numbers and figures later in the report.

Critically, in the midst of working on the data, we discovered that there were some missing and inconsistent values that we could not analyze, or else, all of our findings would be inaccurate and meaningless. So, to overcome this problem, we worked on an algorithm in MATLAB to find any missing or irregular values within each column and replace them with the mean value of that column.

Handling Inconsistent/Missing values:

Cleaning and correcting the data was accomplished through replacing the missing values with the **mean** of the column of numerical data. This would grant us having equal column vector size while comparing two columns of data (both of them would have 615 values, thus calculating the relational formulas would not be difficult).

After plotting the numbers, we noticed that the 'count for each category of people with liver conditions' bar graph and pie chart clearly showed how many individuals within the dataset were blood donors and how the rest of the categories barely differed from each other at a low frequency. This helps us know that the data gathered in this study is heavily skewed towards the blood donors category, and this makes sense because there are certainly more healthy people than those who have liver conditions in society.

The Age Distribution:

Moving on to the "Age" variable (a key factor in this dataset), we created and solved all of the necessary formulas for this numerical data in relation to all of the other biochemical markers. However, the essential part in all of this is figuring out what these numbers actually

represent and why they are significant, so plotting them onto different types of figures (scatter plots, histograms) helped us to identify the relationships. These figures reassured us that the central tendency of this data is consistent with the mean, median, and mode values (roughly 47) when knowing that the people range from 17-year-olds all the way to 77-year-olds. This also supports our interpolation and extrapolation numbers, as it signifies that most of the dataset contains relatively middle-aged people. The standard deviation and variance values were easily distinguishable from seeing the peaks of the density curve graph (located in the central tendency area) versus all of the gradually declining plots around 30 and 65 years old. The scatter plot visually represents a slightly right-skewed curve which explains the dataset containing more older people than younger. All of these figures are extremely helpful in finding unique characteristics of the data; for example, our boxplot distinctly shows the 1 outlier that deviates from our data (located above the upper whisker at around age 77). Another surprising feature we noticed was the correlation coefficient being slightly positive and negative in most (if not all) of the biochemical markers, which indicates how age could have a big influence on diagnosing liver diseases or conditions.

The Sex column of variables:

This section is indeed categorical, with the majority of samples being male we could predict that most of the people with liver conditions are potentially males. And aging factor also play a pivotal role in that. Other enzymes are also different for different sexes as we are delving into the enzyme column. So sex is also a factor in getting liver disease as its more seen in males.

The First Enzyme: ALB

This numeric value shows a unit of this certain biochemical in each person's body.

By looking at different figures now we specify exactly what is happening for this figure:

ALB Scatterplot: shows a normal distribution, although it has some outliers for certain variables with 82.2 units hitting the highest possible amount of this enzyme in a person's liver who is a blood donor male in his 50s. The lowest one however, for finding the lowest outlier, is again a male suspect blood donor in his early 70s scoring the lowest possible units of ALB.

Correlation and covariance are also provided below:

Index: -0.31

Age: -0.20

ALP: -0.14

ALT: 0.00

AST: -0.19

BIL: -0.22

CHE: 0.38

CHOL: 0.20

CREA: -0.00

GGT: -0.16

PROT: 0.55

With PROT and index having highest and lowest amount of relationship respectively with the covariances below:

Index: -317.50

Age: -11.44

ALP: -20.57

ALT: 0.24

AST: -36.96

BIL: -25.19

CHE: 4.79

CHOL: 1.33

CREA: -0.45

GGT: -49.11

PROT: 17.15

Covariance of index being the lowest and PROT being the highest.

ALB Histogram and Density curve: shows that our pinnacle is at the range of ALB being from 40 to 42, with almost 100 people scoring this. Thus , the average units per person could be 41 units per molecule of blood in the liver. Density curse also illustrates that almost 60% of people have lower amounts of ALB enzyme in their body which makes the skewness of the curve to be left tilted and negative.

ALB boxplot: box plot shows most of the values fall lower than the median of our ALB units which is 41.9 (close to the mean) which means people tend to have lower units of ALB in their

body. Also, other outliers mentioned in the scatterplot are the furthest red crosses seen in the box plot.

The Second Enzyme: ALP

This numeric value shows a unit of this certain biochemical in each person's body.

By looking at different figures now we specify exactly what is happening for this figure:

ALP Scatterplot: shows a normal distribution, although it has some outliers for certain variables with 416.6 units hitting the highest possible amount of this enzyme in a person's liver who is a female diagnosed with cirrhosis in his 60s. The lowest one however, for finding the lowest outlier, is a male diagnosed with Cirrhosis in his mid 40s scoring the lowest possible units (11.3) of ALP.

Correlation and covariance are also provided below:

index: 0.02

Age: 0.17

ALB: -0.14

ALT: 0.17

AST: 0.06

BIL: 0.05

CHE: 0.03

CHOL: 0.12

CREA: 0.15

GGT: 0.44

PROT: -0.05

With GGT and PROT having highest and lowest amount of relationship respectively with the covariances below:

Index: 102.07

Age: 43.44

ALB: -20.57

ALT: 112.55

AST: 52.78

BIL: 24.52

CHE: 1.86

CHOL: 3.51

CREA: 190.82

GGT: 619.93

PROT: -7.42

Covariance of ALB being the lowest and GGT being the highest.

ALP Histogram and Density curve: shows that our pinnacle is at the range of ALP being from 60 to 70, with almost 140 people scoring this. Thus , the average units per person could be 65 units per molecule of blood in the liver. Density curse also illustrates that almost 65% of people have higher amounts of ALP enzyme in their body which makes the skewness of the curve to be right tilted and positive.

ALP boxplot: box plot shows most of the values fall higher than the median of our ALP units which is 66.7 (close to the mean) which means people tend to have higher units of ALP in their body. Also, other outliers mentioned in the scatterplot are the furthest red crosses seen in the box plot.

The Third Enzyme: ALT

This numeric value shows a unit of this certain biochemical in each person's body.

By looking at different figures now we specify exactly what is happening for this figure:

ALT Scatterplot: shows a normal distribution, although it has some outliers for certain variables with 325.3 units hitting the highest possible amount of this enzyme in a person's liver who is a female suspected to be a blood donor in his late 50s. The lowest one however, for finding the lowest outlier, is a male diagnosed with Cirrhosis in his mid 50s scoring the lowest possible units (0.9) of ALT.

Correlation and covariance are also provided below:

index -0.03

Age: -0.01

ALB: 0.00

ALP: 0.17

AST: 0.27

BIL: -0.04

CHE: 0.15

CHOL: 0.07

CREA: -0.04

GGT: 0.25

PROT: 0.09

With AST and BIL with CREA having highest and lowest amount of relationship respectively with the covariances below:

Index: -158.24

Age: -1.54

ALB: 0.24

ALP: 112.55

AST: 230.17

BIL: -19.26

CHE: 8.25

CHOL: 1.94

CREA: -54.48

GGT: 345.10

PROT: 12.96

Covariance of ALB being the lowest and GGT being the highest.

ALT Histogram and Density curve: shows that our pinnacle is at the range of ALT being from 10 to 20, with almost 220 people scoring this. Thus , the average units per person could be 15 units per molecule of blood in the liver. Density curse also illustrates that almost 65% of people have

higher amounts of ALT (same as ALP) enzyme in their body which makes the skewness of the curve to be right tilted and positive.

ALT boxplot: box plot shows all of the values fall higher than the median of our ALT units which is 23 (8 units more than the mean) which means all of the people tend to have higher units of ALT in their body. Also, other outliers mentioned in the scatterplot are the furthest red crosses seen in the box plot.

The Forth Enzyme: AST

This numeric value shows a unit of this certain biochemical in each person's body.
By looking at different figures now we specify exactly what is happening for this figure:

AST Scatterplot: shows a normal distribution, although it has some outliers for certain variables with 324 units hitting the highest possible amount of this enzyme in a person’s liver who is a male diagnosed with Hepatitis in his mid 50s. The lowest one however, for finding the lowest outlier, is a suspected male donor in his late 40s scoring the lowest possible units (10.6) of AST.
Correlation and covariance are also provided below:

index: 0.33

Age: 0.09

ALB: -0.19

ALP: 0.06

ALT: 0.27

BIL: 0.31

CHE: -0.21

CHOL: -0.21

CREA: -0.02

GGT: 0.49

PROT: 0.04

With GGT and CHE plus CHOL having highest and lowest amount of relationship respectively with the covariances below:

Index: 1955.69

Age: 29.50

ALB: -36.96

ALP: 52.78

ALT: 230.17

BIL: 203.26

CHE: -15.22

CHOL: -7.71

CREA: -35.21

GGT: 888.58

PROT: 7.13

Covariance of ALB being the lowest and Index being the highest.

AST Histogram and Density curve: shows that our pinnacle is at the range of AST being from 20 to 30, with almost 300 people scoring this. Thus , the average units per person could be 25 units per molecule of blood in the liver. Density curse also illustrates that almost 70% of people have higher amounts of AST enzyme in their body which makes the skewness of the curve to be right tilted and positive.

AST boxplot: box plot shows most of the values fall higher than the median of our AST units which is 25.9 (close to the mean) which means people tend to have higher units of AST in their body. Also, other outliers mentioned in the scatterplot are the furthest red crosses seen in the box plot.

The Fifth Enzyme: BIL

This numeric value shows a unit of this certain biochemical in each person's body.
By looking at different figures now we specify exactly what is happening for this figure:

BIL Scatterplot: shows a normal distribution, although it has some outliers for certain variables with 254 units hitting the highest possible amount of this enzyme in a person’s liver who is a male diagnosed with Cirrhosis in his mid 40s. The lowest one however, for finding the lowest outlier, is a suspected male donor in his late 40s scoring the lowest possible units (0.8) of BIL.
Correlation and covariance are also provided below:

index: 0.18

Age: 0.03

ALB: -0.22

ALP: 0.05

ALT: -0.04

AST: 0.31

CHE: -0.33

CHOL: -0.16

CREA: 0.03

GGT: 0.22

PROT: -0.04

With AST and CHE having highest and lowest amount of relationship respectively with the covariances below:

Index: 634.29

Age: 6.43

ALB: -25.19

ALP: 24.52

ALT: -19.26

AST: 203.26

CHE: -14.46

CHOL: -3.45

CREA: 30.56

GGT: 233.38

PROT: -4.39

Covariance of GGT being the lowest and Index being the highest.

BIL Histogram and Density curve: shows that our pinnacle is at the range of BIL being from 0 to 10, with almost 420 people scoring this. Thus , the average units per person could be 5 units per molecule of blood in the liver. Density curse also illustrates that almost 80% of people have higher amounts of BIL enzyme in their body which makes the skewness of the curve to be right tilted and positive.

BIL boxplot: box plot shows most of the values fall higher than the median of our BIL units which is 7.3 (close to the mean) which means people tend to have higher units of BIL in their body. Also, other outliers mentioned in the scatterplot are the furthest red crosses seen in the box plot.

The Sixth Enzyme: CHE

This numeric value shows a unit of this certain biochemical in each person's body.
By looking at different figures now we specify exactly what is happening for this figure:

CHE Scatterplot: shows a normal distribution, although it has some outliers for certain variables with 16.41 units hitting the highest possible amount of this enzyme in a person’s liver who is a male diagnosed with Hepatitis in his mid 40s. The lowest one however, for finding the lowest outlier, is a male diagnosed with Cirrhosis in his mid 50s scoring the lowest possible units (1.42) of CHE.

Correlation and covariance are also provided below:

index: -0.27

Age: -0.08

ALB: 0.38

ALP: 0.03

ALT: 0.15

AST: -0.21

BIL: -0.33

CHOL: 0.42

CREA: -0.01

GGT: -0.11

PROT: 0.29

With CHOL and BIL having highest and lowest amount of relationship respectively with the covariances below:

Index: -106.03

Age: -1.67

ALB: 4.79

ALP: 1.86

ALT: 8.25

AST: -15.22

BIL: -14.46

CHOL: 1.04

CREA: -1.22

GGT: -13.30

PROT: 3.49

Covariance of Index being the lowest and ALT being the highest.

CHE Histogram and Density curve: shows that our pinnacle is at the range of CHE being from 7 to 9, with almost 120 people scoring this. Thus , the average units per person could be 6 units per molecule of blood in the liver. Density curse also illustrates that almost 55% of people have higher amounts of CHE enzyme in their body which makes the skewness of the curve to be right tilted and positive.

CHE boxplot: box plot shows most of the values fall higher than the median of our CHE units which is 8.26 (close to the mean) which means half of the people to have higher units of CHE in their body while the other half have it at a lower amount. Also, other outliers mentioned in the scatterplot are the furthest red crosses seen in the box plot.

The Seventh Enzyme: CHOL

This numeric value shows a unit of this certain biochemical in each person's body.

By looking at different figures now we specify exactly what is happening for this figure:

CHOL Scatterplot: shows a normal distribution, although it has some outliers for certain

variables with 9.67 units hitting the highest possible amount of this enzyme in a person's liver who is a male diagnosed with Cirrhosis in his early 50s. The lowest one however, for finding the lowest outlier, is a male diagnosed with Cirrhosis in his mid 50s scoring the lowest possible units (1.43) of CHOL.

Correlation and covariance are also provided below:

index: -0.09

Age: 0.12

ALB: 0.20

ALP: 0.12

ALT: 0.07

AST: -0.21

BIL: -0.16

CHE: 0.42

CREA: -0.05

GGT: -0.01

PROT: 0.21

With CHE and AST having highest and lowest amount of relationship respectively with the covariances below:

Index: -17.11

Age: 1.40

ALB: 1.33

ALP: 3.51

ALT: 1.94

AST: -7.71

BIL: -3.45

CHE: 1.04

CREA: -2.66

GGT: -0.42

PROT: 1.25

Covariance of Index being the lowest and ALP being the highest.

CHOL Histogram and Density curve: shows that our pinnacle is at the range of CHOL being from 5 to 5.5, with almost 120 people scoring this. Thus , the average units per person could be 5.25 units per molecule of blood in the liver. Density curse also illustrates that almost 55% of people have higher amounts of CHOL enzyme in their body which makes the skewness of the curve to be right tilted and positive.

CHOL boxplot: box plot shows most of the values fall higher than the median of our BIL units which is 5.31 (close to the mean) which means people tend to have higher units of CHOL in their body. Also, other outliers mentioned in the scatterplot are the furthest red crosses seen in the box plot.

The Eighth Enzyme: CREA

This numeric value shows a unit of this certain biochemical in each person's body.

By looking at different figures now we specify exactly what is happening for this figure:

CREA Scatterplot: shows a normal distribution, although it has some outliers for certain variables with an astonishing 1079.1 units hitting the highest possible amount of this enzyme in a person’s liver who is a male diagnosed with Cirrhosis in his mid 40s. The lowest one however, for finding the lowest outlier, is a male blood donor in his early 50s scoring the lowest possible units (8) of CREA.

Correlation and covariance are also provided below:

index: -0.03

Age: -0.02

ALB: -0.00

ALP: 0.15

ALT: -0.04

AST: -0.02

BIL: 0.03

CHE: -0.01

CHOL: -0.05

GGT: 0.12

PROT: -0.03

With GGT and CHOL having highest and lowest amount of relationship respectively with the covariances below:

Index: -229.74

Age: -11.15

ALB: -0.45

ALP: 190.82

ALT: -54.48

AST: -35.21

BIL: 30.56

CHE: -1.22

CHOL: -2.66

GGT: 329.10

PROT: -8.51

Covariance of index being the lowest and GGT being the highest.

CREA Histogram and Density curve: shows that our pinnacle is at the range of CREA being from 60 to 80, with almost 300 people scoring this. Thus , the average units per person could be 70 units per molecule of blood in the liver. Density curse also illustrates that almost 70% of people have higher amounts of CREA enzyme in their body which makes the skewness of the curve to be right tilted and positive.

CREA boxplot: box plot shows most of the values fall either above or beneath the median of our CREA units which is 77 (close to the mean) which means people tend to have higher or lower units of CREA in their body. Also, other outliers mentioned in the scatterplot are the furthest red crosses seen in the box plot.

The Ninth Enzyme: GGT

This numeric value shows a unit of this certain biochemical in each person's body.

By looking at different figures now we specify exactly what is happening for this figure:

GGT Scatterplot: shows a normal distribution, although it has some outliers for certain variables with 650.9 units hitting the highest possible amount of this enzyme in a person's liver who is a female diagnosed with Cirrhosis in his early 60s. The lowest one however, for finding the lowest outlier, is a female blood donor in his early 30s scoring the lowest possible units (4.5) of GGT.

Correlation and covariance are also provided below:

index: 0.25

Age: 0.15

ALB: -0.16

ALP: 0.44

ALT: 0.25

AST: 0.49

BIL: 0.22

CHE: -0.11

CHOL: -0.01

CREA: 0.12

PROT: -0.01

With ALP and ALB having highest and lowest amount of relationship respectively with the covariances below:

Index: 2406.49

Age: 84.14

ALB: -49.11

ALP: 619.93

ALT: 345.10

AST: 888.58

BIL: 233.38

CHE: -13.30

CHOL: -0.42

CREA: 329.10

PROT: -3.46

Covariance of ALB being the lowest and Index being the highest.

GGT Histogram and Density curve: shows that our pinnacle is at the range of GGT being from 0 to 20, with almost 250 people scoring this. Thus , the average units per person could be 10 units per molecule of blood in the liver. Density curse also illustrates that almost 60% of people have

higher amounts of GGT enzyme in their body which makes the skewness of the curve to be right tilted and positive.

GGT boxplot: box plot shows all of the values fall higher than the median of our GGT units which is 23.3 (half the value of the mean) which means people tend to have higher units of GGT in their body. Also, other outliers mentioned in the scatterplot are the furthest red crosses seen in the box plot.

The Tenth Enzyme: PROT

This numeric value shows a unit of this certain biochemical in each person's body.
By looking at different figures now we specify exactly what is happening for this figure:

PROT Scatterplot: shows a normal distribution, although it has some outliers for certain variables with 90 units hitting the highest possible amount of this enzyme in a person's liver who is a male diagnosed with Hepatitis in his late 20s. The lowest one however, for finding the lowest outlier, is a suspected male donor in his late 40s scoring the lowest possible units (44.8) of PROT.

Correlation and covariance are also provided below:

index: -0.11

Age: -0.15

ALB: 0.55

ALP: -0.05

ALT: 0.09

AST: 0.04

BIL: -0.04

CHE: 0.29

CHOL: 0.21

CREA: -0.03

GGT: -0.01

With ALB and Age having highest and lowest amount of relationship respectively with the covariances below:

Index: -109.09

Age: -8.34

ALB: 17.15

ALP: -7.42

ALT: 12.96

AST: 7.13

BIL: -4.39

CHE: 3.49

CHOL: 1.25

CREA: -8.51

GGT: -3.46

Covariance of Index being the lowest and ALB being the highest.

PROT Histogram and Density curve: shows that our pinnacle is at the range of PROT being from 70 to 72, with almost 115 people scoring this. Thus , the average units per person could be 71 units per molecule of blood in the liver. Density curse also illustrates that almost 55% of people have lower amounts of PROT enzyme in their body which makes the skewness of the curve to be left tilted and positive.

PROT boxplot: box plot shows most of the values fall lower than the median of our PROT units which is 72.2 (close to the mean) which means people tend to have lower units of PROT in their body. Also, other outliers mentioned in the scatterplot are the furthest red crosses seen in the box plot.

Conclusion

Looking back at this entire process and especially the analysis of this dataset, it is needless to say that we discovered many new connections in the factors surrounding the Hepatitis C virus. Although the data we worked with wasn't perfect by any means (it contained many missing values), it still gave us knowledge on the general trends and patterns associated with liver conditions, as well as specific molecular level component takeaways (how each of the biochemical markers react to our age and gender). We initially predicted that there would be signs of a positive correlation between HCV and certain numerical/categorical data, and we're glad that we could design scripts in MATLAB to support our claims both numerically and visually.

References

Gilat, A. (2014). *MATLAB: An Introduction with Applications* (5th ed.). Wiley.

Kennedy, G. (2024). *Explore Data with MATLAB Plots*. MathWorks. https://matlabacademy.mathworks.com/details/explore-data-with-matlab-plots/otmledp

Lichtinghagen, R., Klawonn, F., & Hoffmann, G. (2020). *HCV data*. UCI Machine Learning Repository. https://doi.org/10.24432/C5D612

Mayo Clinic. (2023). *Hepatitis C*. Mayo Foundation for Medical Education and Research (MFMER). https://www.mayoclinic.org/diseases-conditions/hepatitis-c/symptoms-causes/syc-20354278

Moler, C., & Little, J. (2000). *MATLAB Help Center: bioma.data.DataMatrix*. MathWorks. https://www.mathworks.com/help/bioinfo/ref/bioma.data.datamatrix.html?searchHighlight=for+name+%3D+column_names%282%3Aend%29+++++colName+%3D+name%7B1%7D%3B+++++column_data+%3D+data_set.%28colName%29%3B&s_tid=srchtitle_support_results_1_for+name+%253D+column_names%25282%253Aend%2529+++++colName+%253D+name%257B1%257D%253B+++++column_data+%253D+data_set.%2528colName%2529%253B

Moler, C., & Little, J. (2000). *MATLAB Help Center: categories*. MathWorks. https://www.mathworks.com/help/matlab/ref/categorical.categories.html

Moler, C., & Little, J. (2000). *MATLAB Help Center: corrcoef*. MathWorks. https://www.mathworks.com/help/matlab/ref/corrcoef.html?searchHighlight=corrcoef&s_tid=srchtitle_support_results_1_corrcoef

Moler, C., & Little, J. (2000). *MATLAB Help Center: Covariance*. MathWorks. https://www.mathworks.com/help/mbc/mbccommandline/mbcmodel.responsefeatures.covariance.html?searchHighlight=covariance&s_tid=srchtitle_support_results_1_covariance&status=SUCCESS

Moler, C., & Little, J. (2000). *MATLAB Help Center: Extrapolation for Interpolant Fit Types*. MathWorks. https://www.mathworks.com/help/curvefit/extrapolation-methods.html?searchHighlight=extrapolationmethod&s_tid=srchtitle_support_results_1_extrapolationmethod

Moler, C., & Little, J. (2000). *MATLAB Help Center: figure*. MathWorks. https://www.mathworks.com/help/matlab/ref/figure.html?searchHighlight=figure&s_tid=srchtitle_support_results_1_figure

Moler, C., & Little, J. (2000). *MATLAB Help Center: fprintf problem*. MathWorks. https://www.mathworks.com/support/search.html/answers/39530-fprintf-problem.html?fq%5B%5D=asset_type_name:answer&fq%5B%5D=category:matlab/startup-and-shutdown&page=1

Moler, C., & Little, J. (2000). *MATLAB Help Center: gca*. MathWorks.https://www.mathworks.com/help/matlab/ref/gca.html?s_tid=doc_ta

Moler, C., & Little, J. (2000). *MATLAB Help Center: Interpolation*. MathWorks. https://www.mathworks.com/help/curvefit/interpolation.html?searchHighlight=interapolation&s_tid=srchtitle_support_results_1_interapolation

Moler, C., & Little, J. (2000). *MATLAB Help Center: isstring*. MathWorks. https://www.mathworks.com/help/matlab/ref/isstring.html?searchHighlight=if+iscell%28column_data%29+%7C%7C+ischar%28column_data%29+%7C%7C+iscategorical%28column_data%29&s_tid=srchtitle_support_results_1_if+iscell%2528column_data%2529+%257C%257C+ischar%2528column_data%2529+%257C%257C+iscategorical%2528column_data%2529

Moler, C., & Little, J. (2000). *MATLAB Help Center: polyfit*. MathWorks. https://www.mathworks.com/help/matlab/ref/polyfit.html?searchHighlight=polyfit&s_tid=srchtitle_support_results_1_polyfit

Moler, C., & Little, J. (2000). *MATLAB Help Center: rmmissing*. MathWorks. https://www.mathworks.com/help/matlab/ref/rmmissing.html?s_tid=doc_ta

Moler, C., & Little, J. (2000). *MATLAB Help Center: Set Data Properties*. MathWorks. https://www.mathworks.com/help/stateflow/ug/set-data-properties-1.html?searchHighlight=column_names+%3D+data_set.Properties.VariableNames%3B&s_tid=srchtitle_support_results_1_column_names+%253D+data_set.Properties.VariableNames%253B

Moler, C., & Little, J. (2000). *MATLAB Help Center: set*. MathWorks. https://www.mathworks.com/help/matlab/ref/set.html?s_tid=doc_ta

Moler, C., & Little, J. (2000). *MATLAB Help Center: skewness*. MathWorks. https://www.mathworks.com/help/stats/skewness.html?searchHighlight=skewness&s_tid=srchtitle_support_results_1_skewness

Moler, C., & Little, J. (2000). *MATLAB Help Center: sprintf*. MathWorks. https://www.mathworks.com/help/matlab/ref/string.sprintf.html?searchHighlight=sprintf&s_tid=srchtitle_support_results_1_sprintf

Moler, C., & Little, J. (2000). *MATLAB Help Center: strcat*. MathWorks. https://www.mathworks.com/help/matlab/ref/strcat.html?searchHighlight=strcat&s_tid=srchtitle_support_results_1_strcat

Moler, C., & Little, J. (2000). MATLAB Help Center: strcmp. MathWorks.https://www.mathworks.com/help/matlab/ref/strcmp.html?searchHighlight=strcmp&s_tid=srchtitle_support_results_1_strcmp

Moler, C., & Little, J. (2000). *MATLAB Help Center: summary*. MathWorks. https://www.mathworks.com/help/matlab/ref/double.summary.html?searchHighlight=summary+values&s_tid=srchtitle_support_results_1_summary+values

Moler, C., & Little, J. (2000). *MATLAB Help Center: uifigure*. MathWorks. https://www.mathworks.com/help/matlab/ref/uifigure.html?searchHighlight=uifigure&s_tid=srchtitle_support_results_1_uifigure

Moler, C., & Little, J. (2000). *MATLAB Help Center: uitextarea*. MathWorks. https://www.mathworks.com/help/matlab/ref/uitextarea.html?searchHighlight=uitextarea&s_tid=srchtitle_support_results_1_uitextarea

Moler, C., & Little, J. (2000). *MATLAB Help Center: xticks*. MathWorks. https://www.mathworks.com/help/matlab/ref/xticks.html?s_tid=doc_ta

Placek, B. (2024). *Clean and Prepare Data for Analysis*. MathWorks. https://matlabacademy.mathworks.com/details/clean-and-prepare-data-for-analysis/otmlpda

World Health Organization. (2024). *Hepatitis C*. https://www.who.int/news-room/fact-sheets/detail/hepatitis-c