Relationship of Different Biochemical and Clinical Markers with Liver Cirrhosis Stages

Rahul Somabhai Chaudhary
Data Science
University of North Texas
Denton, TX, USA
RahulSomabhaiChaudhary@my.unt.ed
u

Introduction

Liver Cirrhosis is a chronic, progressive disease spread across the liver, making it failure to work properly over prolonged period of time. This condition, often resulting from alcohol abuse, hepatitis infection or fatty liver disease significantly impairing the liver function.

This study provides the aims to understands the effect of different components affecting the Liver resulting in Liver Cirrhosis such as Age, Cholesterol and many more by comparing data of 25000 individuals. In this study, the advance tools such as Tableau and languages like Python has been used.

Through this study, I am to identify the important trends and relations that offer insightful knowledge into complex interplay of factors affecting liver cirrhosis progression.

I. Dataset

A. *Collecting Dataset*

Dataset has been collected from Kaggle. It contains the information of 25000 individuals, with their data such as Age, Sex, Ascites, Hepatomegaly etc., this data provides a key information on understanding the level of liver cirrhosis of that individual.

Kaggel Dataset link: Liver Cirrhosis Stage Classification

B. *Dataset Information*

Dataset contains 25000 rows and 19 columns, each row represents different person while each column represents different characteristics affecting liver. 19 columns contains as follows:

1. Sex: Gender of the patients (Male/Female)
2. Age: Age of the patient in years
3. Ascites: It is condition of too much fluid in the abdomen. Here it represents the presence of Ascites or not (Y/N)
4. Hepatomegaly: Condition of enlarged liver, here it is represented by yes or no(Y/N)
5. Spiders: Presence of spider angiomas(Y/N)
6. Edema: It occurs when the fluid accumulates in the leg[1]. presence of edema N (no edema and no diuretic therapy for edema), S (edema present without diuretics, or edema resolved by diuretics), or Y (edema despite diuretic therapy)[2]
7. Stages: histologic stage of disease (1, 2, or 3)[2]
8. Status: status of the patient C (censored), CL (censored due to liver tx), or D (death)[2]
9. Drug: Type of drug used for treatment(D-penicillamine or placebo)
10. N_Days: Number of days between registration and the earlier of death, transplantation, or study analysis time in 1986[2]
11. Bilirubin: Serum bilirubin in[mg/dl]
12. Cholesterol: Blood cholesterol level of patient in [mg/dl]
13. Albumin: Albumin level in [gm/dl]
14. Copper: Urine copper level in [ug/day]
15. Alk_Phos: Alkaline phosphate level in [U/liter]
16. SGOT: Aspartate aminotransferase [U/ml]
17. Triglicerides: Triglycerides in [mg/dl]
18. Platelets: Platelets count per cubic [ml/1000]
19. Prothrombin: Prothrombin time in seconds

C. *Tools used for the dataset*

- Python: For data cleaning, Exploratory Data Analysis and visualization
- Tableau: For visualization

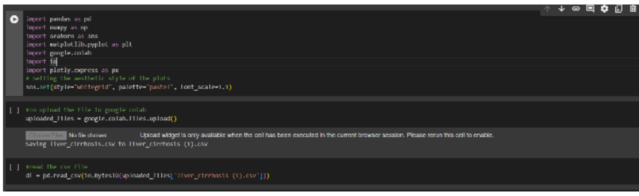
II. Data Cleaning And exploratory data analysis

To continue with the dataset and start our hypothesis, we need to do some cleaning such as removing null or missing values from our dataset and perform EDA to get knowledge of our dataset.

A. *Data Cleaning*

In this portion, we will perform the cleaning of data using python, we will start with

1. Importing necessary modules and the dataset



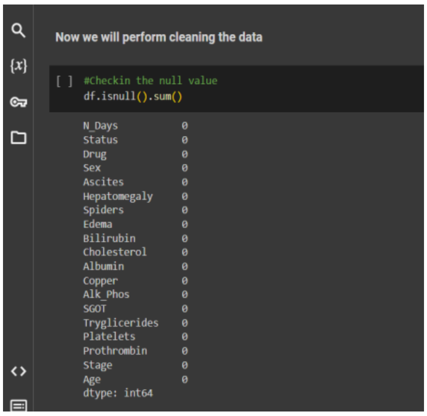
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import google.colab
import io
import plotly.express as px
# Setting the aesthetic style of the plots
sns.set(style="whitegrid", palette="pastel", font_scale=1.1)
```

```
#to upload the file to google colab
uploaded_files = google.colab.files.upload()
```

```
Choose Files No file chosen        Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving liver_cirrhosis.csv to liver_cirrhosis (1).csv
```

```
#read the csv file
df = pd.read_csv(io.BytesIO(uploaded_files['liver_cirrhosis (1).csv']))
```

2. Checking for any null values



Now we will perform cleaning the data

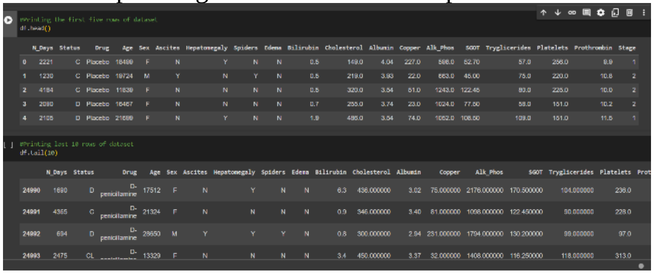
```
#Checkin the null value
df.isnull().sum()
```

```
N_Days          0
Status          0
Drug            0
Sex             0
Ascites         0
Hepatomegaly    0
Spiders         0
Edema           0
Bilirubin       0
Cholesterol     0
Albumin         0
Copper          0
Alk_Phos        0
SGOT            0
Tryglicerides   0
Platelets       0
Prothrombin     0
Stage           0
Age             0
dtype: int64
```

As the data doesn't contains any null values we can move forward and perform EDA

B. *Performing Exploratory Data Analysis*

Now to understand the dataset better and to find the relationship among the attributes we will perform EDA.



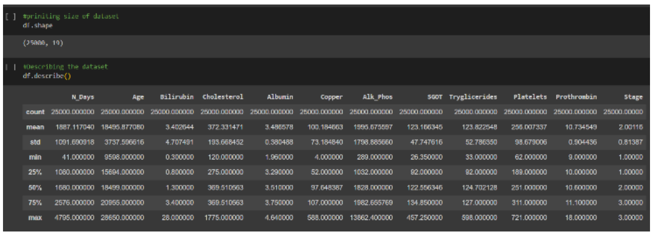
```
#Printing the first five rows of dataset
df.head()
```

	N_Days	Status	Drug	Age	Sex	Ascites	Hepatomegaly	Spiders	Edema	Bilirubin	Cholesterol	Albumin	Copper	Alk_Phos	SGOT	Tryglicerides	Platelets	Prothrombin	Stage
0	2221	C	Placebo	18499	F	N	Y	N	N	0.5	149.0	4.04	227.0	598.0	52.70	57.0	256.0	9.9	1
1	1230	C	Placebo	19724	M	Y	N	Y	N	0.5	219.0	3.93	22.0	663.0	45.00	75.0	220.0	10.8	2
2	4184	C	Placebo	11839	F	N	N	N	N	0.5	320.0	3.54	51.0	1243.0	122.45	80.0	225.0	10.0	2
3	2090	D	Placebo	16467	F	N	N	N	N	0.7	255.0	3.74	23.0	1024.0	77.50	58.0	151.0	10.2	2
4	2105	D	Placebo	21699	F	N	Y	N	N	1.9	486.0	3.54	74.0	1052.0	108.50	109.0	151.0	11.5	1

```
#Printing last 10 rows of dataset
df.tail(10)
```

	N_Days	Status	Drug	Age	Sex	Ascites	Hepatomegaly	Spiders	Edema	Bilirubin	Cholesterol	Albumin	Copper	Alk_Phos	SGOT	Tryglicerides	Platelets	Prot
24990	1690	D	D-penicillamine	17512	F	N	Y	N	N	6.3	436.000000	3.02	75.000000	2176.000000	170.500000	104.000000	236.0	
24991	4365	C	D-penicillamine	21324	F	N	N	N	N	0.9	346.000000	3.40	81.000000	1098.000000	122.450000	90.000000	228.0	
24992	694	D	D-penicillamine	28650	M	Y	Y	Y	N	0.8	300.000000	2.94	231.000000	1794.000000	130.200000	99.000000	97.0	
24993	2475	CL	D-penicillamine	13329	F	N	N	N	N	3.4	450.000000	3.37	32.000000	1408.000000	116.250000	118.000000	313.0	

Here by function head and tail we can print some of the first and last rows of dataset.



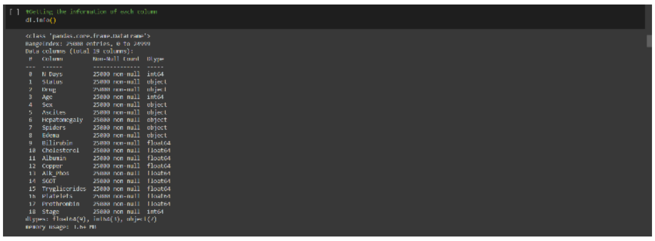
```
#priniting size of dataset
df.shape
```

```
(25000, 19)
```

```
#Describing the dataset
df.describe()
```

	N_Days	Age	Bilirubin	Cholesterol	Albumin	Copper	Alk_Phos	SGOT	Tryglicerides	Platelets	Prothrombin	Stage
count	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.00000
mean	1887.117040	18495.877080	3.402644	372.331471	3.486578	100.184663	1995.675597	123.166345	123.822548	256.007337	10.734549	2.00116
std	1091.690918	3737.596616	4.707491	193.668452	0.380488	73.184840	1798.885660	47.747616	52.786350	98.679006	0.904436	0.81387
min	41.000000	9598.000000	0.300000	120.000000	1.960000	4.000000	289.000000	26.350000	33.000000	62.000000	9.000000	1.00000
25%	1080.000000	15694.000000	0.800000	275.000000	3.290000	52.000000	1032.000000	92.000000	92.000000	189.000000	10.000000	1.00000
50%	1680.000000	18499.000000	1.300000	369.510563	3.510000	97.648387	1828.000000	122.556346	124.702128	251.000000	10.600000	2.00000
75%	2576.000000	20955.000000	3.400000	369.510563	3.750000	107.000000	1982.655769	134.850000	127.000000	311.000000	11.100000	3.00000
max	4795.000000	28650.000000	28.000000	1775.000000	4.640000	588.000000	13862.400000	457.250000	598.000000	721.000000	18.000000	3.00000

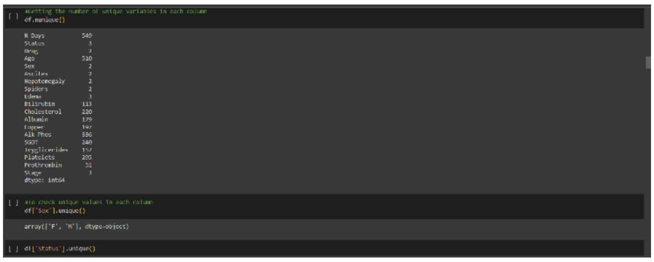
The function shape provides the size of dataset while the description provides the statistical calculation of dataset.



```
#Getting the information of each column
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 19 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   N_Days         25000 non-null  int64  
 1   Status         25000 non-null  object 
 2   Drug           25000 non-null  object 
 3   Age            25000 non-null  int64  
 4   Sex            25000 non-null  object 
 5   Ascites        25000 non-null  object 
 6   Hepatomegaly   25000 non-null  object 
 7   Spiders        25000 non-null  object 
 8   Edema          25000 non-null  object 
 9   Bilirubin      25000 non-null  float64
 10  Cholesterol    25000 non-null  float64
 11  Albumin        25000 non-null  float64
 12  Copper         25000 non-null  float64
 13  Alk_Phos       25000 non-null  float64
 14  SGOT           25000 non-null  float64
 15  Tryglicerides  25000 non-null  float64
 16  Platelets      25000 non-null  float64
 17  Prothrombin    25000 non-null  float64
 18  Stage          25000 non-null  int64  
dtypes: float64(9), int64(3), object(7)
memory usage: 3.6+ MB
```

Method info provides information such as type and non-null values of columns.



```
#Getting the number of unique variables in each column
df.nunique()
```

```
N_Days           540
Status             3
Drug               2
Age              510
Sex                2
Ascites            2
Hepatomegaly       2
Spiders            2
Edema              3
Bilirubin        113
Cholesterol      220
Albumin          179
Copper           197
Alk_Phos         336
SGOT             240
Tryglicerides    157
Platelets        295
Prothrombin       51
Stage              3
dtype: int64
```

```
#To check unique values in each column
df['Sex'].unique()
```

```
array(['F', 'M'], dtype=object)
```

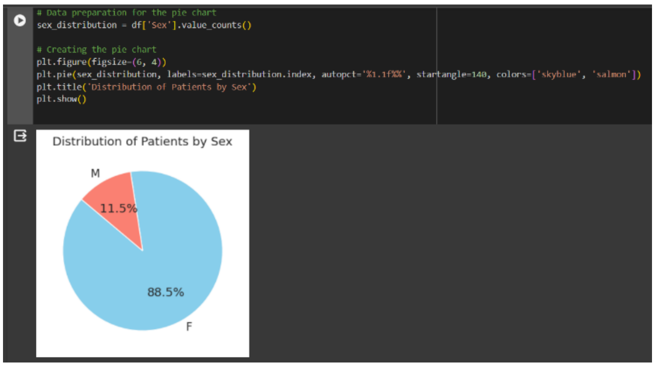
```
df['Status'].unique()
```

Unique provides the number of unique values in columns.

C. *Finding the relations among attributes*

We will implement the relationship among the elements using python,

1. What is the ratio of male to female in dataset?

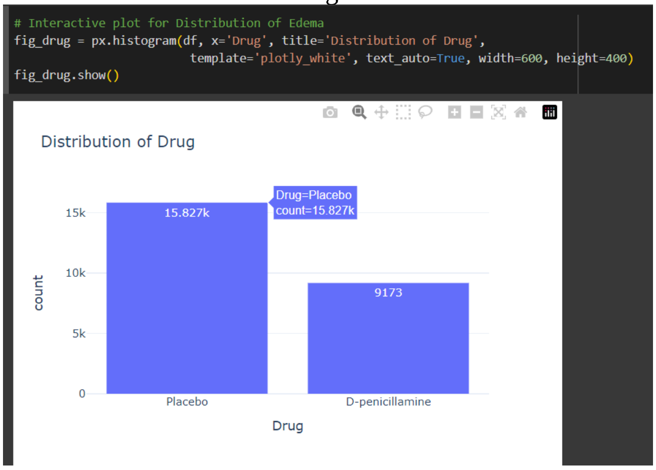


```
# Data preparation for the pie chart
sex_distribution = df['Sex'].value_counts()

# Creating the pie chart
plt.figure(figsize=(6, 4))
plt.pie(sex_distribution, labels=sex_distribution.index, autopct='%1.1f%%', startangle=140, colors=['skyblue', 'salmon'])
plt.title('Distribution of Patients by Sex')
plt.show()
```

The ratio of male to female is 11.5:88.5 respectively.

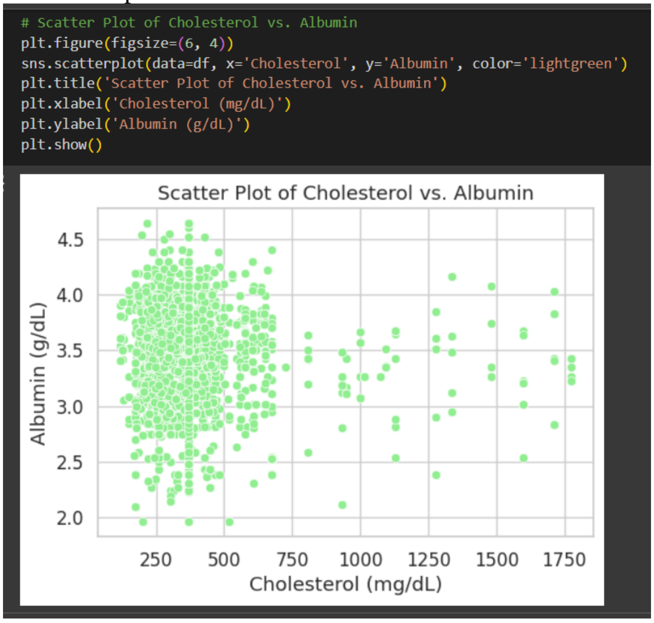
2. What is the distribution of drugs?



```
# Interactive plot for Distribution of Edema
fig_drug = px.histogram(df, x='Drug', title='Distribution of Drug',
                        template='plotly_white', text_auto=True, width=600, height=400)
fig_drug.show()
```

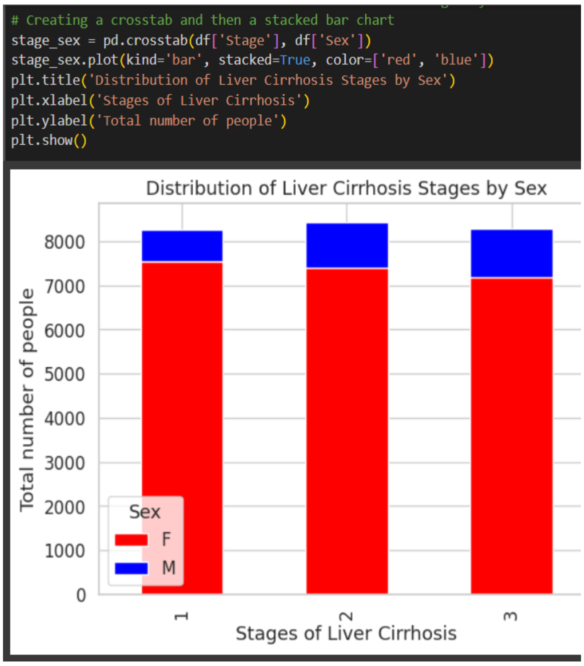
From the figure, we can say that the 15,827 have given placebo while the rest of the 9,173 have been treated with D-penicillamine.

3. Relationship between cholesterol and albumin.



```
# Scatter Plot of Cholesterol vs. Albumin
plt.figure(figsize=(6, 4))
sns.scatterplot(data=df, x='Cholesterol', y='Albumin', color='lightgreen')
plt.title('Scatter Plot of Cholesterol vs. Albumin')
plt.xlabel('Cholesterol (mg/dL)')
plt.ylabel('Albumin (g/dL)')
plt.show()
```

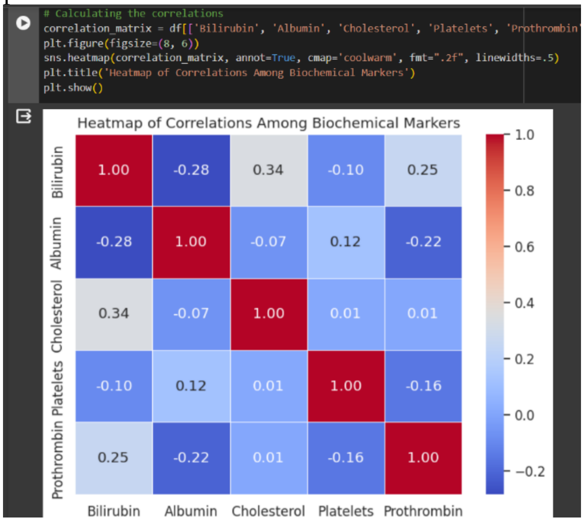
4. Relationship between male and female individual regarding their liver cirrhosis stages.



```
# Creating a crosstab and then a stacked bar chart
stage_sex = pd.crosstab(df['Stage'], df['Sex'])
stage_sex.plot(kind='bar', stacked=True, color=['red', 'blue'])
plt.title('Distribution of Liver Cirrhosis Stages by Sex')
plt.xlabel('Stages of Liver Cirrhosis')
plt.ylabel('Total number of people')
plt.show()
```

Distribution of Liver Cirrhosis Stages by Sex
Total number of people
8000
7000
6000
5000
4000
3000
2000
1000
0
Sex
F
M
1
2
3
Stages of Liver Cirrhosis

5. Correlation matrix showing relationship between bilirubin, albumin, cholesterol, platelets, prothrombin.



```
# Calculating the correlations
correlation_matrix = df[['Bilirubin', 'Albumin', 'Cholesterol', 'Platelets', 'Prothrombin'
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=.5)
plt.title('Heatmap of Correlations Among Biochemical Markers')
plt.show()
```

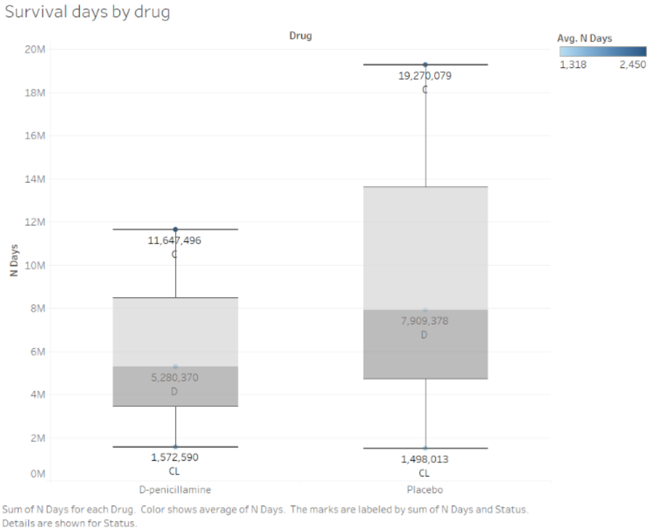
III. Hypothesis Testing Using Tableau

From the above visualizations, we can conclude that the attributes are related to each other, so with the use of tableau, we will now focusing on proving the hypothesis. Below are the mentioned hypothesis that we are going to focus upon.

1. Patients receiving different drugs have different survival rates.
2. The presence of ascites, hepatomegaly or edema is more frequent at higher stages of liver cirrhosis.
3. Cholesterol levels are not significantly different between patients with and without edema.
4. There is no significant relationship between age and prothrombin time.

A. *Patients receiving different drugs have different survival rates.*

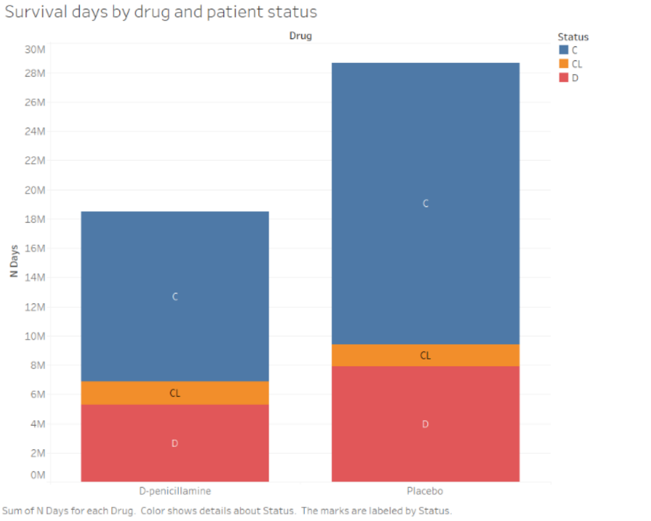
Visualization 1:
Here, I have used the box-and-whisker plot for visualization. Box plot is the way of visualizing data based on five-number summary, first quartile (Q1), median (Q2), third quartile (Q3), minimum and maximum.



Sum of N Days for each Drug. Color shows average of N Days. The marks are labeled by sum of N Days and Status. Details are shown for Status.

Result: From this plot we can easily estimate that the difference in days of recovery from placebo is way higher than those from D-penicillamine, hence our hypothesis can be valid from this visualization.

Visualization 2:

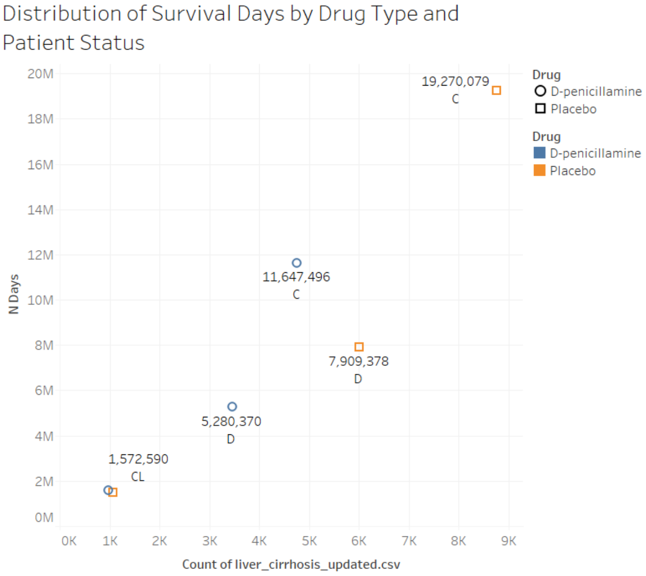


Sum of N Days for each Drug. Color shows details about Status. The marks are labeled by Status.

Here, the visualization used is a bar chart.
Bar chart uses the length or height of bars to represent the quantity of data [3].

Result: From the bars we can conclude that, for all the status whether it is C, CL, or D the placebo has been used more as a recovery drug than those of D-penicillamine, proving the validation of our hypothesis.

Visualization 3:



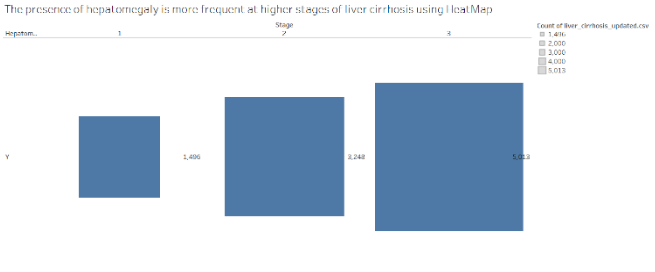
Total number of patients vs. sum of N Days. Color shows details about Status. Shape shows details about Drug. The marks are labeled by sum of N Days and Status.

The visualization used is called scatter plot.
Scatter plot uses dots to represent the data. The dots on the cartesian plain representing data are placed on x and y axis.[4]

Result: From the visualization, we can conclude that the placebo has higher number of survival day respect to the patient count compared to those of D-Penicillamine. Hence the hypothesis is valid.

B. The presence of ascites, hepatomegaly, or edema is more frequent at higher stages of liver cirrhosis.

Visualization 1:



Total number of patients broken down by Stage vs. Hepatomegaly. Size shows count of total number of patients. The marks are labeled by count of total number of patients. The view is filtered on Hepatomegaly, which keeps Y.

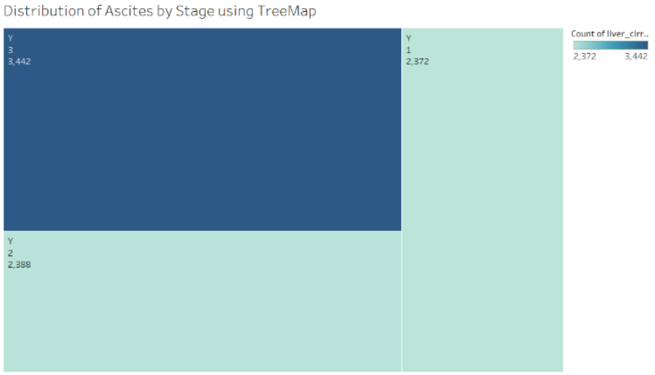
Visualization used to prove this hypothesis is called heat map. A graphical representation of data in the form of color coding to represents the different values of that dataset.[5]

Result: From the visualization, after filtering the data by the patient having hepatomegaly, we can conclude that as the liver cirrhosis stages increases, the number of patient with hepatomegaly increases as well, making our hypothesis valid.

Visualization 2:

The visualization used is tree map for validating the hypothesis.

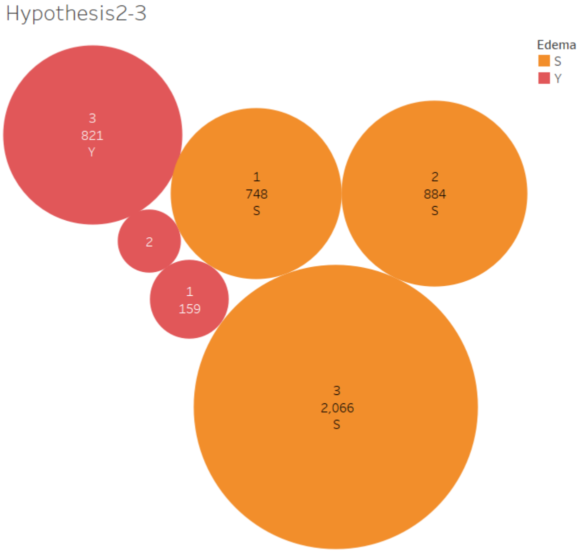
Tree map represents the data into hierarchical and rectangular form, each rectangle representing the data and size of rectangle is based on size of data.



Ascites, Stage and count of total number of patients. Color shows count of total number of patients. Size shows count of total number of patients. The marks are labeled by Ascites, Stage and count of total number of patients. The view is filtered on Ascites, which keeps Y.

Result: From the visualization, after sorting the patients with the ascites as present, we can see that the number of patients are higher in higher stages of liver cirrhosis and the number decreases as the stages of liver cirrhosis decreases, making our hypothesis valid.

Visualization 3:



Stage, count of total number of patients and Edema. Color shows details about Edema. Size shows count of total number of patients. The marks are labeled by Stage, count of total number of patients and Edema. The view is filtered on Edema, which excludes N and Null.

Here, packed bubble is used for visualization.

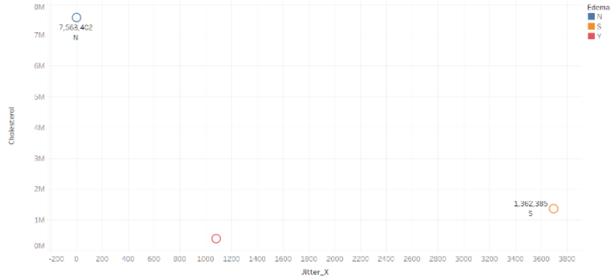
In packed bubble chart, data are displayed in the form of clusters of circles, with bigger the data represented by the larger circle and vice versa.[6]

Result: After sorting the edema level to Y and S, we can conclude that the rise in number of edema patient is higher in patients with higher stages of liver cirrhosis, concluding that the hypothesis is valid.

C. Serum cholesterol levels are not significantly different between patients with and without edema.

Visualization 1:
The visualization used is scatter plot.

Relationshiup between cholesterol of patient with edema

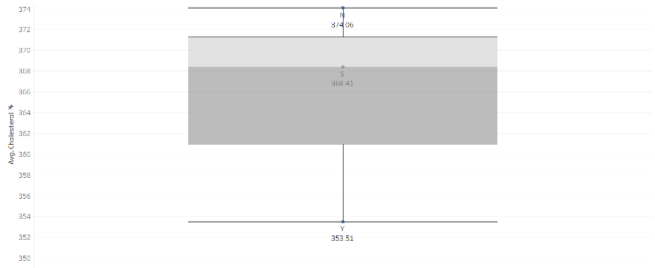


Sum of Jitter_X(Adjust position of Y and N of edema using Random() value generator) vs. sum of Cholesterol. Color shows details about Edema. The marks are labeled by sum of Cholesterol and Edema.

Result: From the visualization, we can clearly see there is a large difference in level of cholesterol with respect to patient with all stages of edema, hence we can conclude that the hypothesis is invalid.

Visualization 2:

Average choleterol level alongside edema

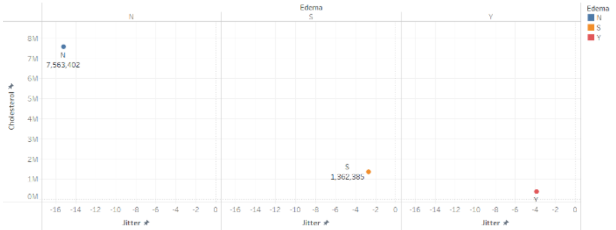


Average of Cholesterol. The marks are labeled by Edema and average of Cholesterol. Details are shown for Edema.

For this visualization, I have used box-and-whisker plot.
Result: For the average cholesterol level of patient with edema is 353.51 while the patient without edema it is 374.06, implying to the significant difference, from the visualization we can nullify our hypothesis.

Visualization 3:

Cholesterol level alongside edema with Jitter function



Sum of Jitter(Random number generator) vs. sum of Cholesterol broken down by Edema. Color shows details about Edema. The marks are labeled by Edema and sum of Cholesterol.

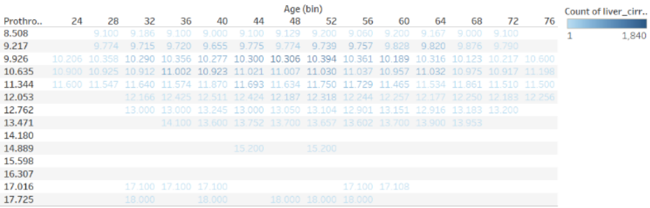
Here, visualization used is called Scatter plot with Jitter. Jitter is a function that will adjust a position of 'Y' and 'N' of edema using Random () function. Here we will use an amplifier with a value of 0.1.

Result: From the visualization, it can clearly conclude that there is a significant difference between cholesterol levels of patients with and without edema. Hence our hypothesis is invalid.

D. *There is no significant relationship between age and prothrombin time.*

Visualization 1:
Visualization used to check the validity of hypothesis is heat map.

HeatMap representing the relationship between age and prothrombin



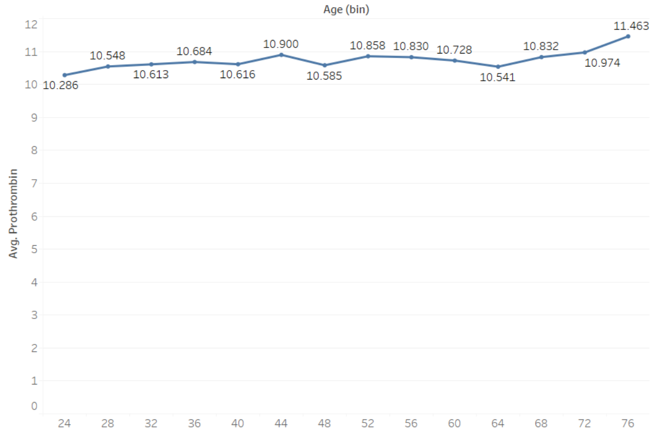
	Age (bin)													
Prothro..	24	28	32	36	40	44	48	52	56	60	64	68	72	76
8.508		9.100	9.186	9.100	9.000	9.100	9.129	9.200	9.060	9.200	9.167	9.000	9.100	
9.217		9.774	9.715	9.720	9.655	9.775	9.774	9.739	9.757	9.828	9.820	9.876	9.790	
9.926	10.206	10.358	10.290	10.356	10.277	10.300	10.306	10.394	10.361	10.189	10.316	10.123	10.217	10.600
10.635	10.900	10.925	10.912	11.002	10.923	11.021	11.007	11.030	11.037	10.957	11.032	10.975	10.917	11.198
11.344	11.600	11.547	11.640	11.574	11.870	11.693	11.634	11.750	11.729	11.465	11.534	11.861	11.510	11.500
12.053			12.166	12.425	12.511	12.424	12.187	12.318	12.244	12.257	12.177	12.250	12.183	12.256
12.762			13.000	13.000	13.245	13.000	13.050	13.104	12.901	13.151	12.916	13.183	13.200	
13.471				14.100	13.600	13.752	13.700	13.657	13.602	13.700	13.900	13.953		
14.180														
14.889						15.200		15.200						
15.598														
16.307														
17.016			17.100	17.100	17.100				17.100	17.108				
17.725			18.000		18.000		18.000	18.000	18.000					

Average of Prothrombin broken down by Age (bin) vs. Prothrombin (bin). Color shows count of total patients.

Result: From the visualization, the average prothrombin time increases with increase in age group, this proves that our hypothesis is invalid.

Visualization 2:

LineGraph of prothrombin with respect to age



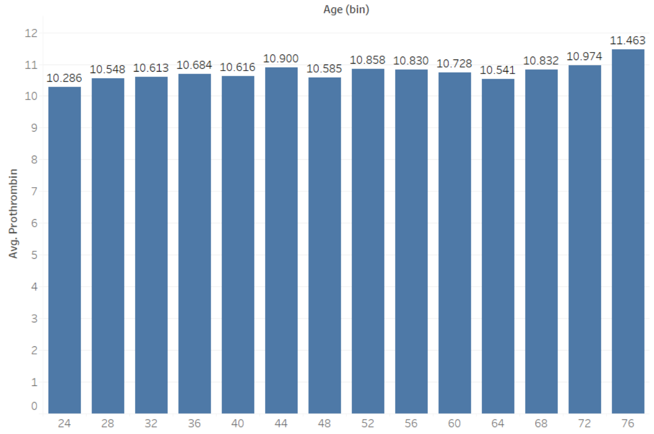
The trend of average of Prothrombin for Age (bin). The marks are labeled by average of Prothrombin.

Visualization used is called line chart.
Line chart is representation of data with points and joining the points to observe the trends using lines.[7]

Result: From the visualization, the lowest prothrombin time is 10.286 by the age group of 20-24 (lowest age group), while the highest prothrombin levels are 11.463 of the age group 73-76 (highest age group), implicating the rise of prothrombin level with increase in age, hence our hypothesis is invalid.

Visualization 3:

Histogram of prothrombin time to that of age



Average of Prothrombin for each Age (bin). The marks are labeled by average of Prothrombin.

Visualization used is histogram.
Histogram is the distribution of data by making bins in the form of bars.[8]

Result:
From the visualization, it can be clearly observed that the relationship between age and prothrombin exists, making our hypothesis invalid.

CONCLUSION

From the above results we can conclude that,

- Patient receiving different drugs have different survival rate.
- The ascites, hepatomegaly, or edema levels are directly proportional to liver cirrhosis stages.
- Cholesterol levels are significantly different between the patient with edema and without edema.
- Age and prothrombin time are related, with increase in age the prothrombin time increases.

REFERENCES

[1] Mayo Clinic
[2] Liver Cirrhosis Stage Classification Dataset by Aadarsh Velu
[3] Bar Chart wikipedia-https://en.wikipedia.org/wiki/Bar_chart
[4] A complete guide to scatter plot- Atlassian
[5] Heat map- Optimizely.
[6] Build a packed bubble chart by tableau.
[7] A complete guide to line chart- Atlassian
[8] Histogram Unveiled- Atlassian