Data Collection and Preprocessing Phase

Date	20th June 2024
Team ID	SWTID1720080161
Project Title	Revolutionizing Liver Care : Predicting Liver Cirrhosis Using Advanced Machine Learning Techniques
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	

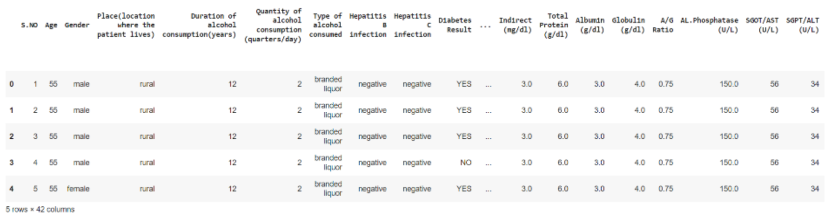
Dimensions : 950 x 42

Shape : (950,42)

```
df.shape
```

```
(950, 42)
```

Head:



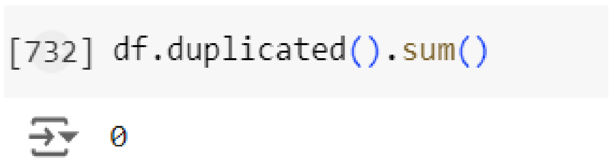
	S.NO	Age	Gender	Place(location where the patient lives)	Duration of alcohol consumption(years)	Quantity of alcohol consumption (quarters/day)	Type of alcohol consumed	Hepatitis B infection	Hepatitis C infection	Diabetes Result	...	Indirect (mg/dl)	Total Protein (g/dl)	Albumin (g/dl)	Globulin (g/dl)	A/G Ratio	AL.Phosphatase (U/L)	SGOT/AST (U/L)	SGPT/ALT (U/L)
0	1	55	male	rural	12	2	branded liquor	negative	negative	YES	...	3.0	6.0	3.0	4.0	0.75	150.0	56	34
1	2	55	male	rural	12	2	branded liquor	negative	negative	YES	...	3.0	6.0	3.0	4.0	0.75	150.0	56	34
2	3	55	male	rural	12	2	branded liquor	negative	negative	YES	...	3.0	6.0	3.0	4.0	0.75	150.0	56	34
3	4	55	male	rural	12	2	branded liquor	negative	negative	NO	...	3.0	6.0	3.0	4.0	0.75	150.0	56	34
4	5	55	female	rural	12	2	branded liquor	negative	negative	YES	...	3.0	6.0	3.0	4.0	0.75	150.0	56	34

5 rows × 42 columns

Overview of columns:

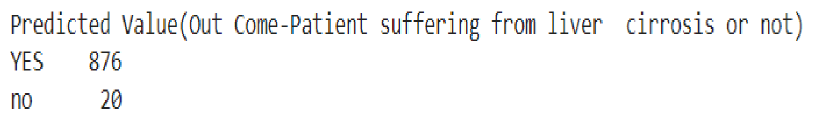
```
Data columns (total 42 columns):
 #   Column                                                           Non-Null Count  Dtype  
---  ------                                                           --------------  -----  
 0   S.NO                                                             1250 non-null   float64
 1   Age                                                              1250 non-null   float64
 2   Gender                                                           1250 non-null   object 
 3   Place(location where the patient lives)                          1116 non-null   object 
 4   Duration of alcohol consumption(years)                           1250 non-null   float64
 5   Quantity of alcohol consumption (quarters/day)                   1250 non-null   float64
 6   Type of alcohol consumed                                         1250 non-null   object 
 7   Hepatitis B infection                                            1250 non-null   object 
 8   Hepatitis C infection                                            1250 non-null   object 
 9   Diabetes Result                                                  1250 non-null   object 
 10  Blood pressure (mmhg)                                            1250 non-null   object 
 11  Obesity                                                          1250 non-null   object 
 12  Family history of cirrhosis/ hereditary                          1250 non-null   object 
 13  TCH                                                              591 non-null    float64
 14  TG                                                               591 non-null    object 
 15  LDL                                                              591 non-null    object 
 16  HDL                                                              582 non-null    float64
 17  Hemoglobin  (g/dl)                                               1250 non-null   float64
 18  PCV  (%)                                                         1220 non-null   float64
 19  RBC  (million cells/microliter)                                  698 non-null    float64
 20  MCV   (femtoliters/cell)                                         1241 non-null   float64
 21  MCH  (picograms/cell)                                            592 non-null    float64
 22  MCHC  (grams/deciliter)                                          578 non-null    float64
 23  Total Count                                                      1240 non-null   float64
 24  Polymorphs  (%)                                                  1250 non-null   float64
 25  Lymphocytes  (%)                                                 1250 non-null   float64
 26  Monocytes   (%)                                                  1241 non-null   float64
 27  Eosinophils   (%)                                                1242 non-null   float64
 28  Basophils  (%)                                                   1201 non-null   float64
 29  Platelet Count  (lakhs/mm)                                       1250 non-null   float64
 30  Total Bilirubin    (mg/dl)                                       1250 non-null   object 
 31  Direct    (mg/dl)                                                1250 non-null   float64
 32  Indirect     (mg/dl)                                             1195 non-null   float64
 33  Total Protein     (g/dl)                                         1189 non-null   float64
 34  Albumin   (g/dl)                                                 1241 non-null   float64
 35  Globulin  (g/dl)                                                 1221 non-null   float64
 36  A/G Ratio                                                        785 non-null    object 
 37  AL.Phosphatase      (U/L)                                        1240 non-null   float64
 38  SGOT/AST      (U/L)                                              1250 non-null   float64
 39  SGPT/ALT (U/L)                                                   1250 non-null   float64
 40  USG Abdomen (diffuse liver or  not)                              1250 non-null   object 
 41  Predicted Value(Out Come-Patient suffering from liver  cirrosis or not)  1196 non-null   object 
dtypes: float64(27), object(15)
```

Duplicate rows:



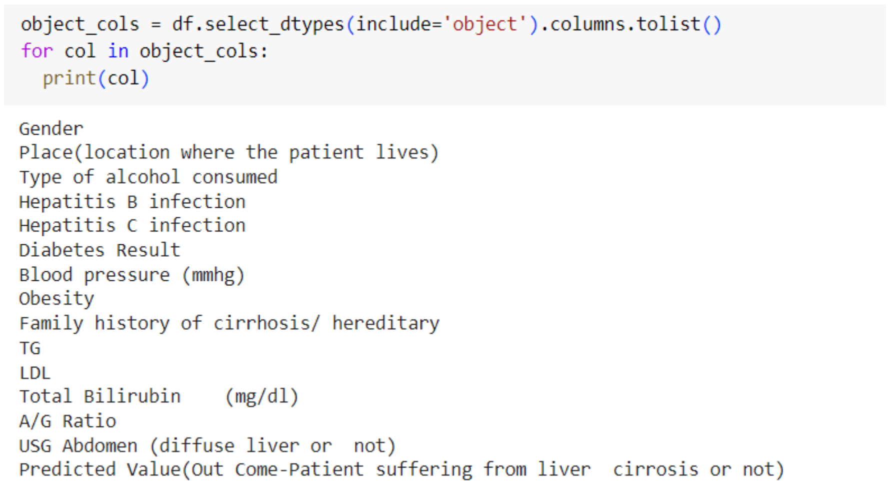
```
[732] df.duplicated().sum()

0
```

Target value to predict:



```
Predicted Value(Out Come-Patient suffering from liver  cirrosis or not)
YES    876
no      20
```

Object columns:



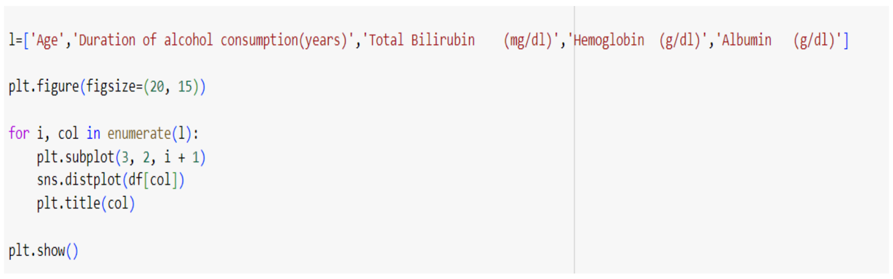
```
object_cols = df.select_dtypes(include='object').columns.tolist()
for col in object_cols:
  print(col)
```

```
Gender
Place(location where the patient lives)
Type of alcohol consumed
Hepatitis B infection
Hepatitis C infection
Diabetes Result
Blood pressure (mmhg)
Obesity
Family history of cirrhosis/ hereditary
TG
LDL
Total Bilirubin    (mg/dl)
A/G Ratio
USG Abdomen (diffuse liver or  not)
Predicted Value(Out Come-Patient suffering from liver  cirrosis or not)
```

Univariate Analysis

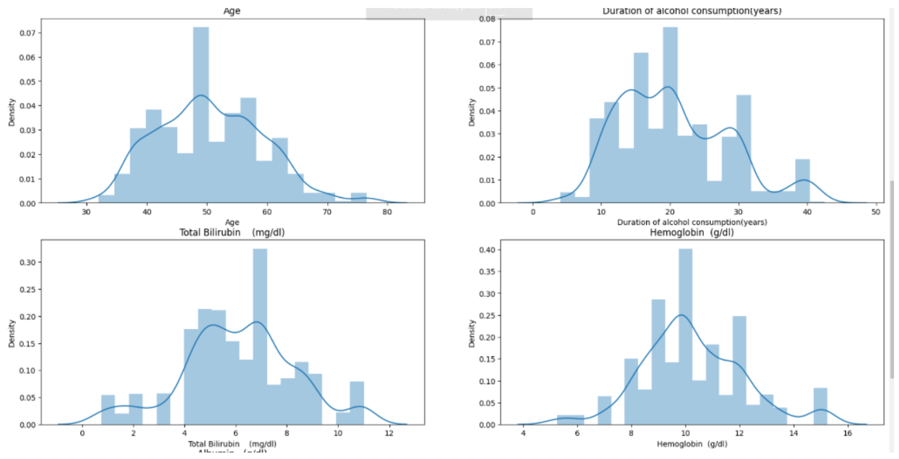
Exploration using Distplots:

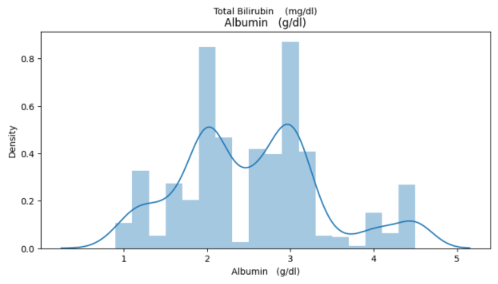
Code:



```
l=['Age','Duration of alcohol consumption(years)','Total Bilirubin    (mg/dl)','Hemoglobin  (g/dl)','Albumin   (g/dl)']

plt.figure(figsize=(20, 15))

for i, col in enumerate(l):
    plt.subplot(3, 2, i + 1)
    sns.distplot(df[col])
    plt.title(col)

plt.show()
```

Plots:





Inference:

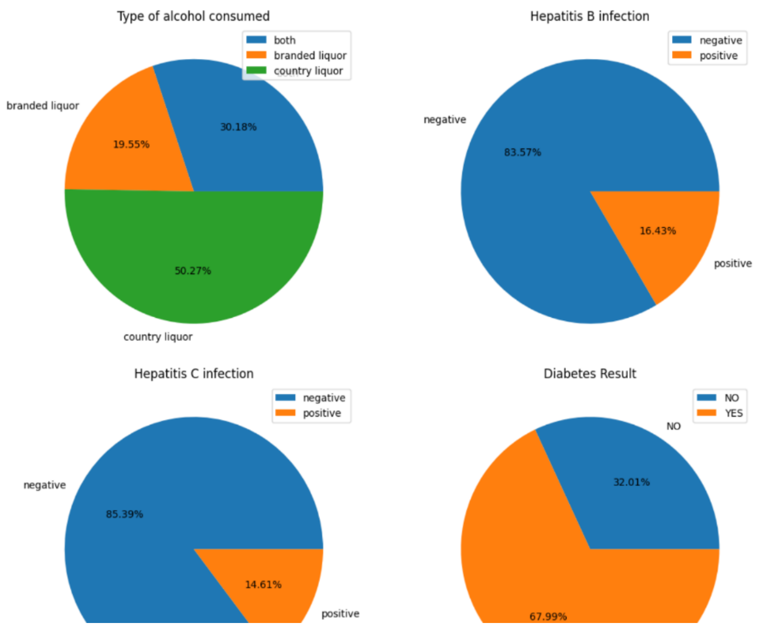
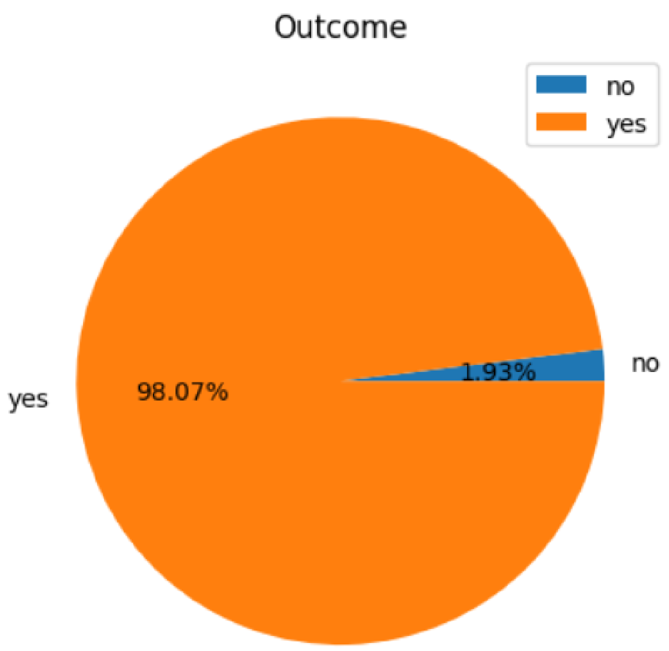
Inferences from Density Plots

Age Distribution:

- The majority of patients fall within the 40-60 age range.
- There is a noticeable peak around the age of 50, indicating a higher frequency of patients in their early 50s.

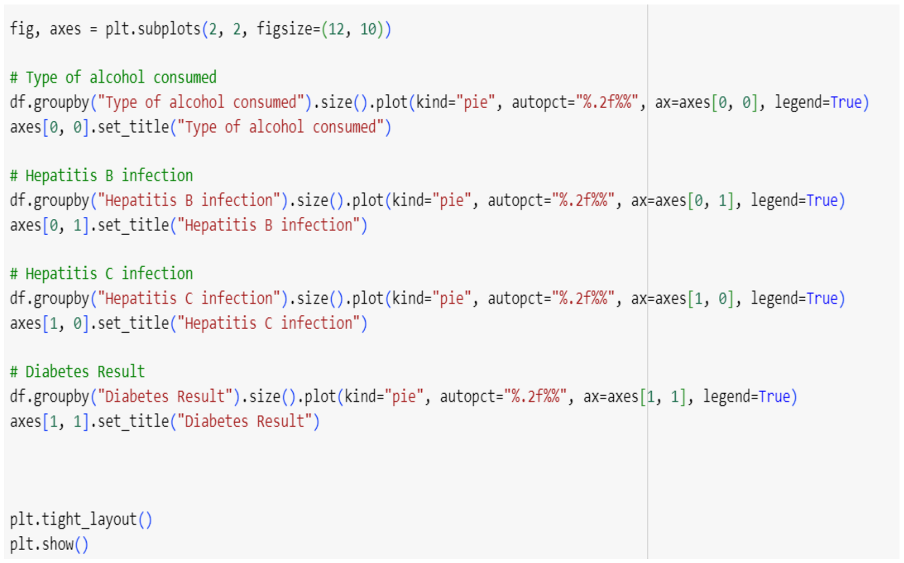
2. **Duration of Alcohol Consumption**:
    - The duration of alcohol consumption varies widely among patients.
    - A significant proportion of patients have been consuming alcohol for around 15-25 years, with a peak at approximately 20 years.
3. **Total Bilirubin**:
    - The total bilirubin levels show a wide distribution, with a peak around 6 mg/dl.
    - There are some patients with very high bilirubin levels, indicating possible liver dysfunction.
4. **Hemoglobin Levels**:
    - Hemoglobin levels are generally distributed around a mean of approximately 10 g/dl.
    - The distribution shows a peak around 10-12 g/dl, suggesting that most patients have moderate to normal hemoglobin levels.

Representing all the important catogorical columns in pie chart

	Pie charts:  

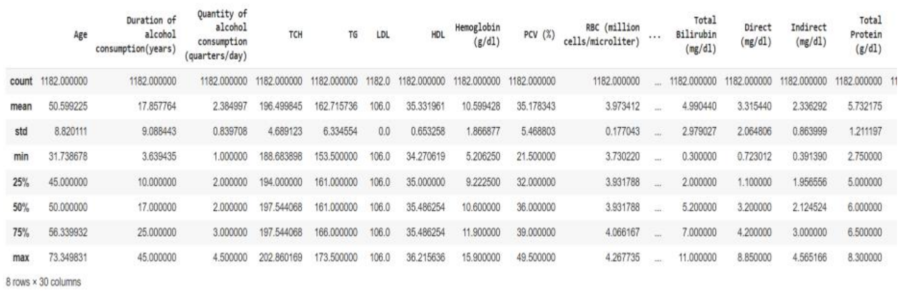
Type of alcohol consumed
both
branded liquor
country liquor
30.18%
19.55%
50.27%
Hepatitis B infection
negative
positive
83.57%
16.43%
Hepatitis C infection
negative
positive
85.39%
14.61%
Diabetes Result
NO
YES
32.01%
67.99%
Outcome
no
yes
98.07%
1.93%

Code:



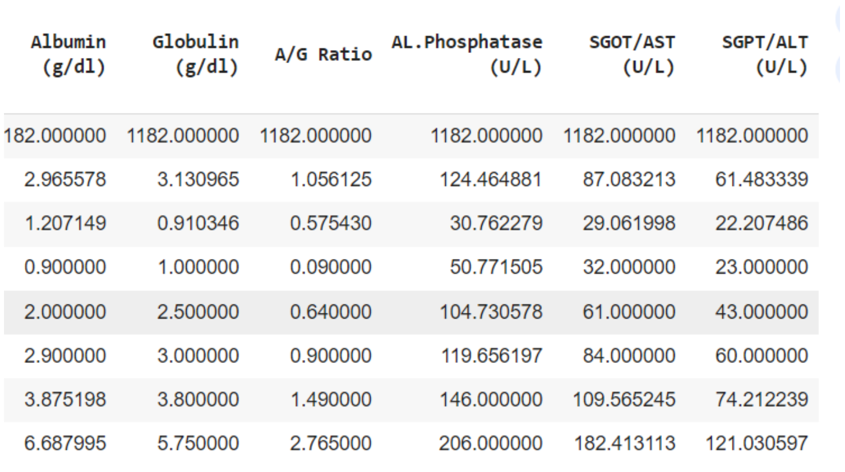
```
fig, axes = plt.subplots(2, 2, figsize=(12, 10))

# Type of alcohol consumed
df.groupby("Type of alcohol consumed").size().plot(kind="pie", autopct="%.2f%%", ax=axes[0, 0], legend=True)
axes[0, 0].set_title("Type of alcohol consumed")

# Hepatitis B infection
df.groupby("Hepatitis B infection").size().plot(kind="pie", autopct="%.2f%%", ax=axes[0, 1], legend=True)
axes[0, 1].set_title("Hepatitis B infection")

# Hepatitis C infection
df.groupby("Hepatitis C infection").size().plot(kind="pie", autopct="%.2f%%", ax=axes[1, 0], legend=True)
axes[1, 0].set_title("Hepatitis C infection")

# Diabetes Result
df.groupby("Diabetes Result").size().plot(kind="pie", autopct="%.2f%%", ax=axes[1, 1], legend=True)
axes[1, 1].set_title("Diabetes Result")



plt.tight_layout()
plt.show()
```

Statistical analysis for individual variables:



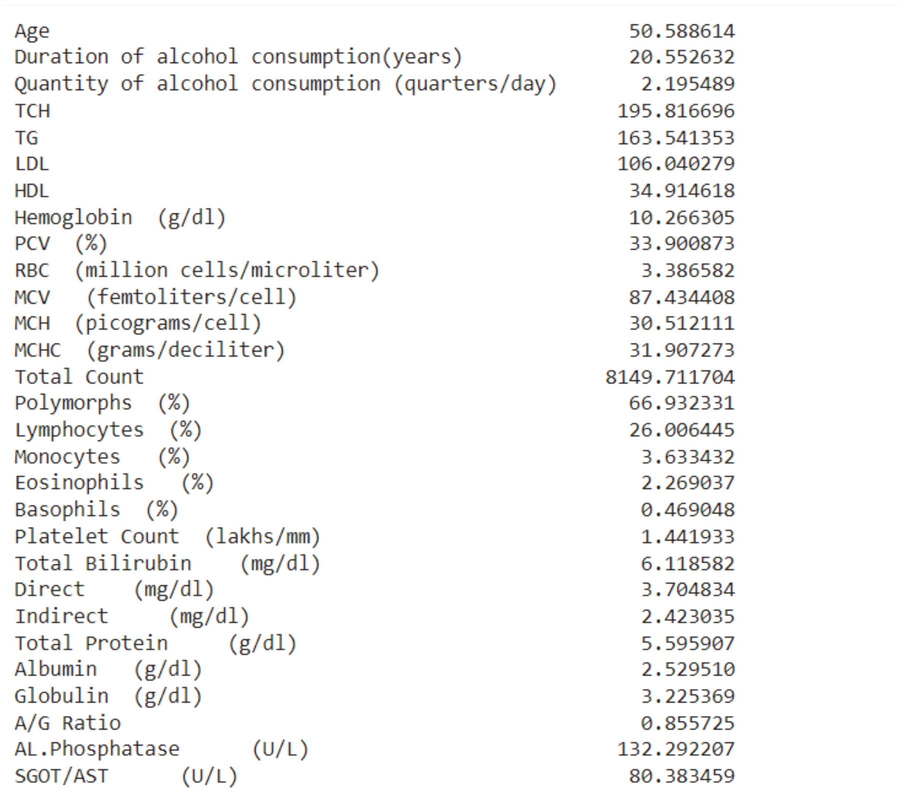
	Age	Duration of alcohol consumption(years)	Quantity of alcohol consumption (quarters/day)	TCH	TG	LDL	HDL	Hemoglobin (g/dl)	PCV (%)	RBC (million cells/microliter)	...	Total Bilirubin (mg/dl)	Direct (mg/dl)	Indirect (mg/dl)	Total Protein (g/dl)
count	1182.000000	1182.000000	1182.000000	1182.000000	1182.000000	1182.0	1182.000000	1182.000000	1182.000000	1182.000000	...	1182.000000	1182.000000	1182.000000	1182.000000
mean	50.599225	17.857764	2.384997	196.499845	162.715736	106.0	35.331961	10.599428	35.178343	3.973412	...	4.990440	3.315440	2.336292	5.732175
std	8.820111	9.088443	0.839708	4.689123	6.334554	0.0	0.653258	1.866877	5.468803	0.177043	...	2.979027	2.064806	0.863999	1.211197
min	31.738678	3.639435	1.000000	188.683898	153.500000	106.0	34.270619	5.206250	21.500000	3.730220	...	0.300000	0.723012	0.391390	2.750000
25%	45.000000	10.000000	2.000000	194.000000	161.000000	106.0	35.000000	9.222500	32.000000	3.931788	...	2.000000	1.100000	1.956556	5.000000
50%	50.000000	17.000000	2.000000	197.544068	161.000000	106.0	35.486254	10.600000	36.000000	3.931788	...	5.200000	3.200000	2.124524	6.000000
75%	56.339932	25.000000	3.000000	197.544068	166.000000	106.0	35.486254	11.900000	39.000000	4.066167	...	7.000000	4.200000	3.000000	6.500000
max	73.349831	45.000000	4.500000	202.860169	173.500000	106.0	36.215636	15.900000	49.500000	4.267735	...	11.000000	8.850000	4.565166	8.300000

8 rows × 30 columns



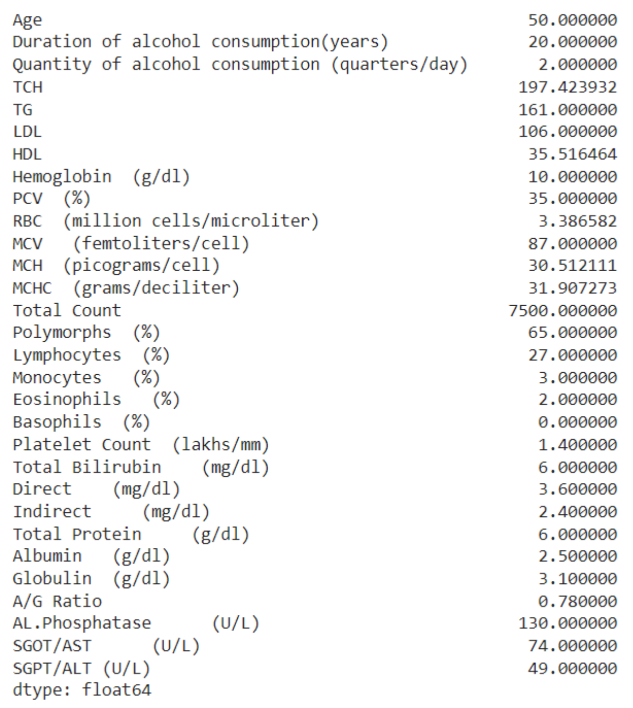
Albumin (g/dl)	Globulin (g/dl)	A/G Ratio	AL.Phosphatase (U/L)	SGOT/AST (U/L)	SGPT/ALT (U/L)
182.000000	1182.000000	1182.000000	1182.000000	1182.000000	1182.000000
2.965578	3.130965	1.056125	124.464881	87.083213	61.483339
1.207149	0.910346	0.575430	30.762279	29.061998	22.207486
0.900000	1.000000	0.090000	50.771505	32.000000	23.000000
2.000000	2.500000	0.640000	104.730578	61.000000	43.000000
2.900000	3.000000	0.900000	119.656197	84.000000	60.000000
3.875198	3.800000	1.490000	146.000000	109.565245	74.212239
6.687995	5.750000	2.765000	206.000000	182.413113	121.030597

Mean of all numerical columns:

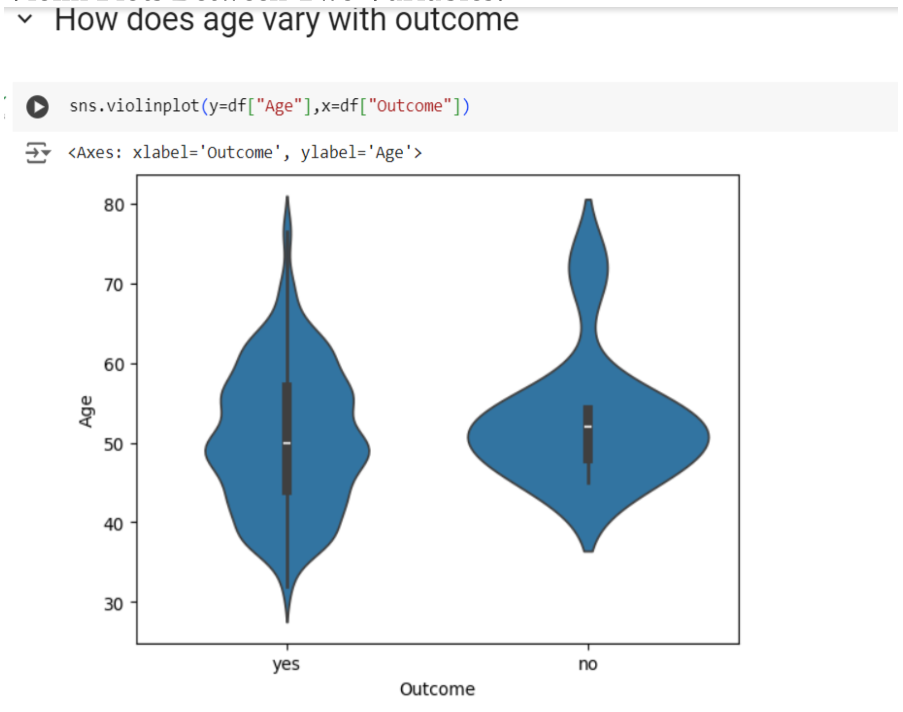


```
Age                                             50.588614
Duration of alcohol consumption(years)          20.552632
Quantity of alcohol consumption (quarters/day)   2.195489
TCH                                            195.816696
TG                                             163.541353
LDL                                            106.040279
HDL                                             34.914618
Hemoglobin  (g/dl)                              10.266305
PCV  (%)                                        33.900873
RBC  (million cells/microliter)                  3.386582
MCV   (femtoliters/cell)                        87.434408
MCH  (picograms/cell)                           30.512111
MCHC  (grams/deciliter)                         31.907273
Total Count                                   8149.711704
Polymorphs  (%)                                 66.932331
Lymphocytes  (%)                                26.006445
Monocytes   (%)                                  3.633432
Eosinophils   (%)                                2.269037
Basophils  (%)                                   0.469048
Platelet Count  (lakhs/mm)                       1.441933
Total Bilirubin    (mg/dl)                       6.118582
Direct    (mg/dl)                                3.704834
Indirect     (mg/dl)                             2.423035
Total Protein     (g/dl)                         5.595907
Albumin   (g/dl)                                 2.529510
Globulin  (g/dl)                                 3.225369
A/G Ratio                                        0.855725
AL.Phosphatase      (U/L)                      132.292207
SGOT/AST      (U/L)                             80.383459
```

Median:



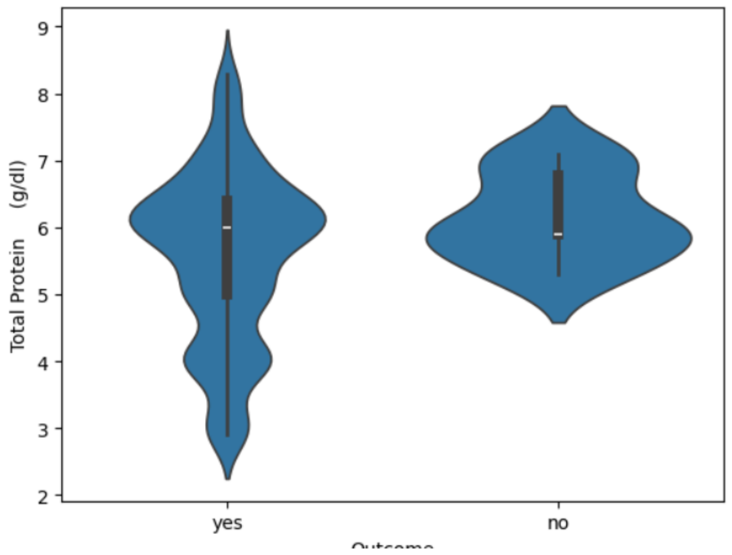
```
Age                                              50.000000
Duration of alcohol consumption(years)           20.000000
Quantity of alcohol consumption (quarters/day)    2.000000
TCH                                             197.423932
TG                                              161.000000
LDL                                             106.000000
HDL                                              35.516464
Hemoglobin  (g/dl)                               10.000000
PCV  (%)                                         35.000000
RBC  (million cells/microliter)                   3.386582
MCV   (femtoliters/cell)                         87.000000
MCH  (picograms/cell)                            30.512111
MCHC  (grams/deciliter)                          31.907273
Total Count                                    7500.000000
Polymorphs  (%)                                  65.000000
Lymphocytes  (%)                                 27.000000
Monocytes   (%)                                   3.000000
Eosinophils   (%)                                 2.000000
Basophils  (%)                                    0.000000
Platelet Count  (lakhs/mm)                        1.400000
Total Bilirubin    (mg/dl)                        6.000000
Direct    (mg/dl)                                 3.600000
Indirect     (mg/dl)                              2.400000
Total Protein     (g/dl)                          6.000000
Albumin   (g/dl)                                  2.500000
Globulin  (g/dl)                                  3.100000
A/G Ratio                                         0.780000
AL.Phosphatase      (U/L)                       130.000000
SGOT/AST      (U/L)                              74.000000
SGPT/ALT (U/L)                                   49.000000
dtype: float64
```

Bivariate Analysis	**Violin Plots Between Two Variables:** How does age vary with outcome  `sns.violinplot(y=df["Age"],x=df["Outcome"])` `<Axes: xlabel='Outcome', ylabel='Age'>` **Inference:** Inferences from Violin Plot The violin plot shows the age distribution for patients with and without liver cirrhosis. • **Patients with Liver Cirrhosis (Yes)**: ◦ Broader age distribution with multiple peaks. ◦ Concentration around 50-60 years. • **Patients without Liver Cirrhosis (No)**: ◦ More uniform age distribution. ◦ Noticeable peak around 50 years. **Conclusion** Liver cirrhosis affects a wider range of ages, especially 50-60 years, while the age distribution for patients without cirrhosis is more consistent.

How does protein influence outcome

```
sns.violinplot(y=df["Total Protein     (g/dl)"],x=df["Outcome"])
```

```
<Axes: xlabel='Outcome', ylabel='Total Protein     (g/dl)'>
```



Inference:

Total Protein Distribution:

Patients with liver cirrhosis ("yes") have a wider distribution of total protein levels ranging from approximately 3 g/dl to 9 g/dl.

Patients without liver cirrhosis ("no") have a slightly narrower distribution, with total protein levels ranging from approximately 4.5 g/dl to 8 g/dl.

Median Total Protein Levels:

The median total protein level in patients with liver cirrhosis is slightly higher than in those without liver cirrhosis, as indicated by the white dot in the center of each violin.els.

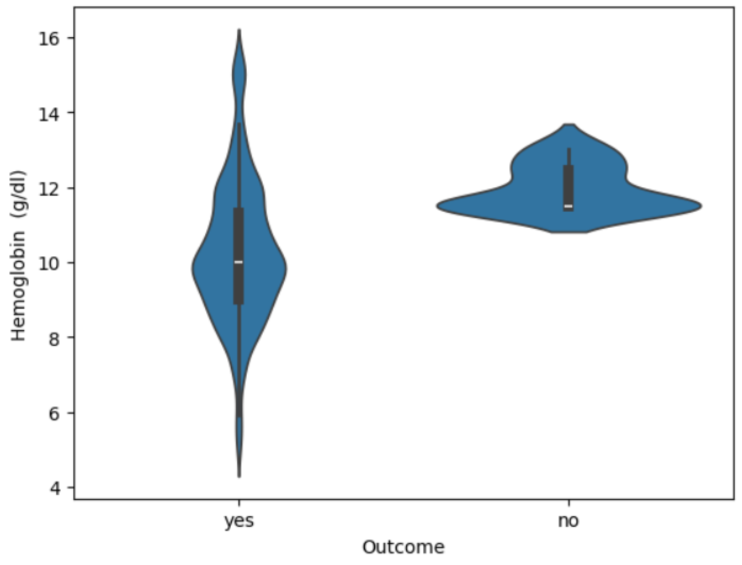
Double-click (or enter) to edit

PROTEIN LEVEL HAS CONSIDERABLE EFFECT ON OUTCOME

How does haemoglobin affect the outcome

```
sns.violinplot(y=df["Hemoglobin  (g/dl)"],x=df["Outcome"])
```

```
<Axes: xlabel='Outcome', ylabel='Hemoglobin  (g/dl)'>
```



Inference:

Distribution:

Cirrhosis ("yes"): Hemoglobin levels range broadly from approximately 4 g/dl to 16 g/dl.

No cirrhosis ("no"): Hemoglobin levels are more concentrated, ranging from about 11 g/dl to 14 g/dl. Median Levels:

Cirrhosis: The median hemoglobin level is around 10 g/dl.

No Cirrhosis: The median hemoglobin level is also around 11.5 g/dl.

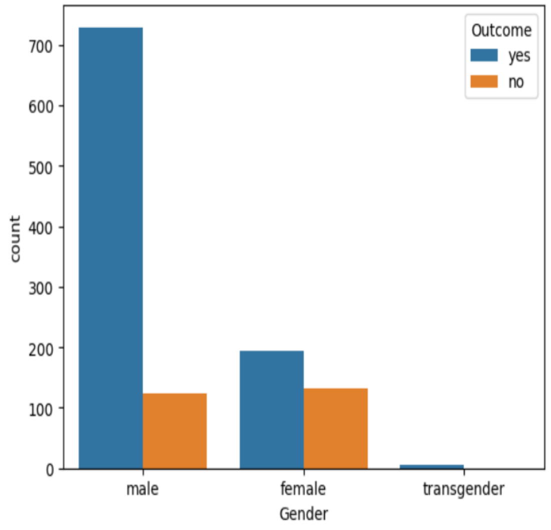
Comparison:

- Liver disease is associated with a wider range of hemoglobin levels.
- No liver disease shows more consistent hemoglobin levels centered around 11.5 g/dl.

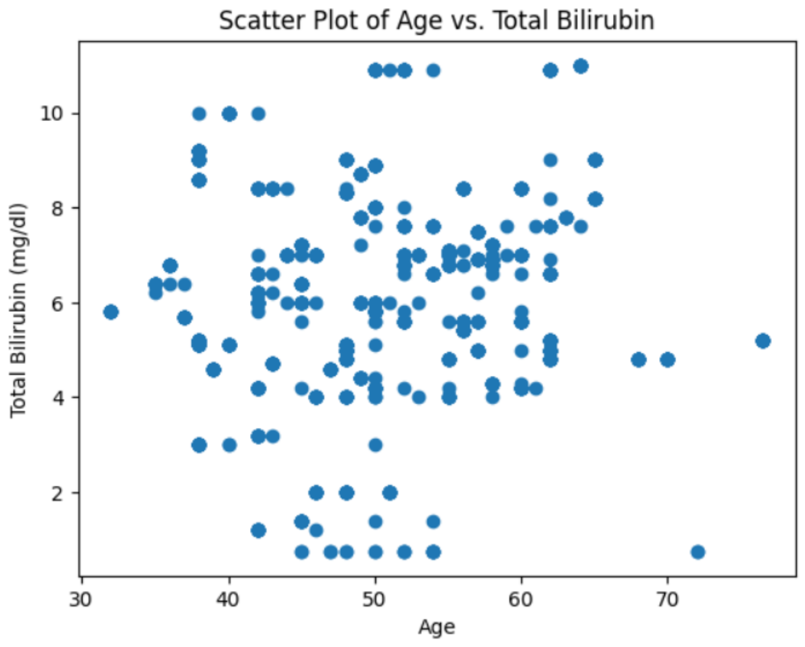
How does the distribution of gender affect the outcome

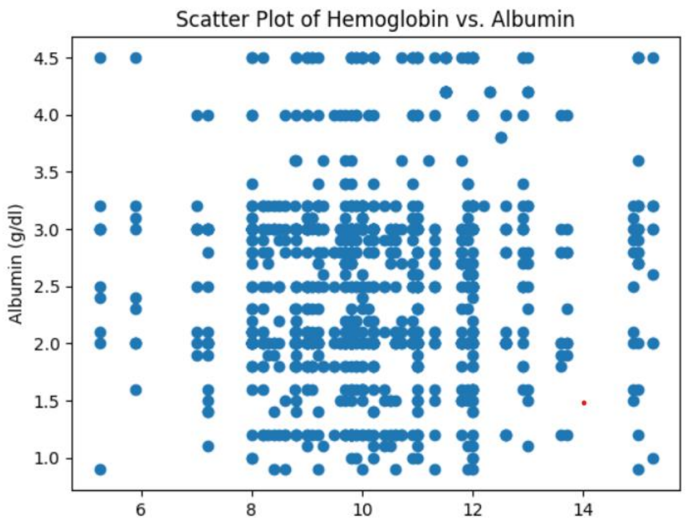
```
sns.countplot(data=df,x="Gender",hue="Outcome")
```

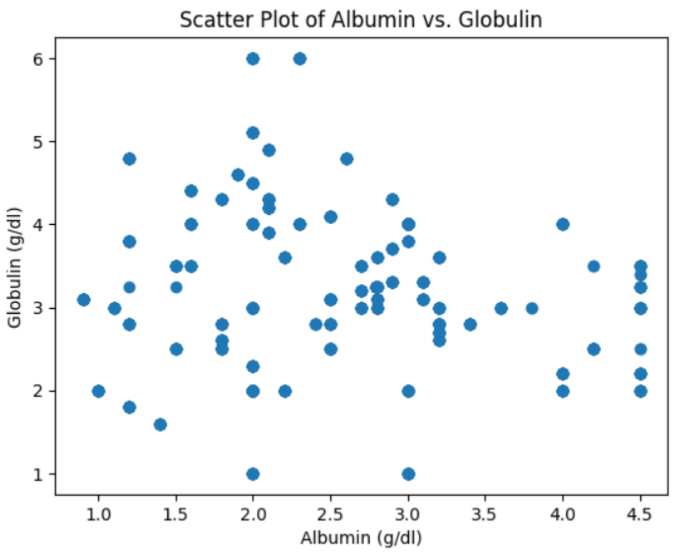
```
<Axes: xlabel='Gender', ylabel='count'>
```



Scatter Plots:



	Inference: No Clear Trend: • There doesn't appear to be a clear linear relationship between age and Total Bilirubin levels. • Total Bilirubin levels are spread across the age range without a consistent pattern.  **Inference:** A large cluster of data points is concentrated around Haemoglobin levels of 8 to 12 g/dl and Albumin levels of 1.3 to 3 g/dl. This suggests that most individuals in the dataset have Haemoglobin levels within this range.



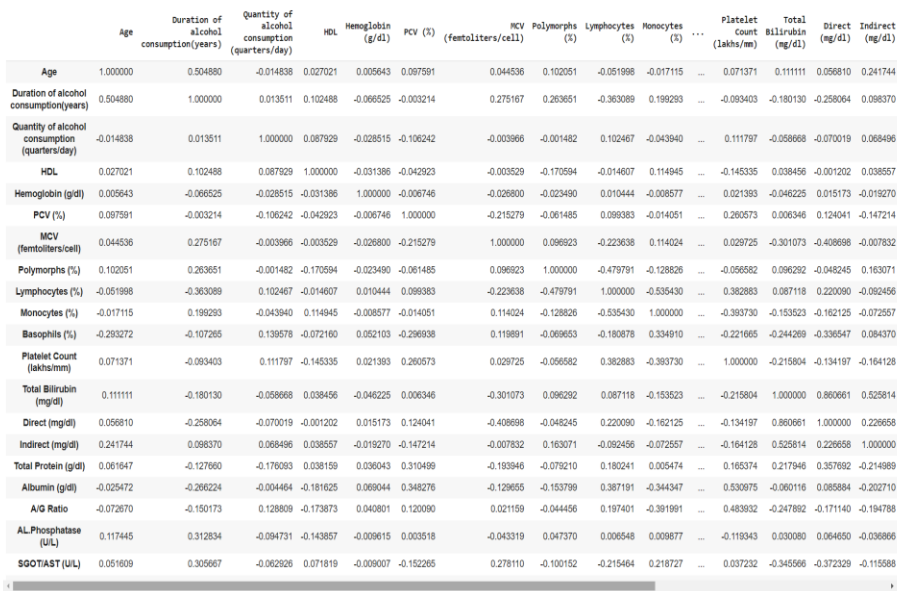
Inference:

Inferences from the Scatter Plot of Albumin vs. Globulin:

1. **No Strong Correlation**:
   - The scatter plot indicates no strong linear relationship between albumin and globulin levels. The data points are widely scattered, suggesting that variations in albumin levels do not directly predict changes in globulin levels.
2. **Range of Values**:
   - Most albumin levels fall within the range of 2.0 to 4.0 g/dl, while globulin levels range from 2.0 to 5.0 g/dl. There are some outliers with higher globulin levels up to 6.0 g/dl and albumin levels up to 4.5 g/dl, indicating diverse liver function profiles among the patients.

Multivariate Analysis

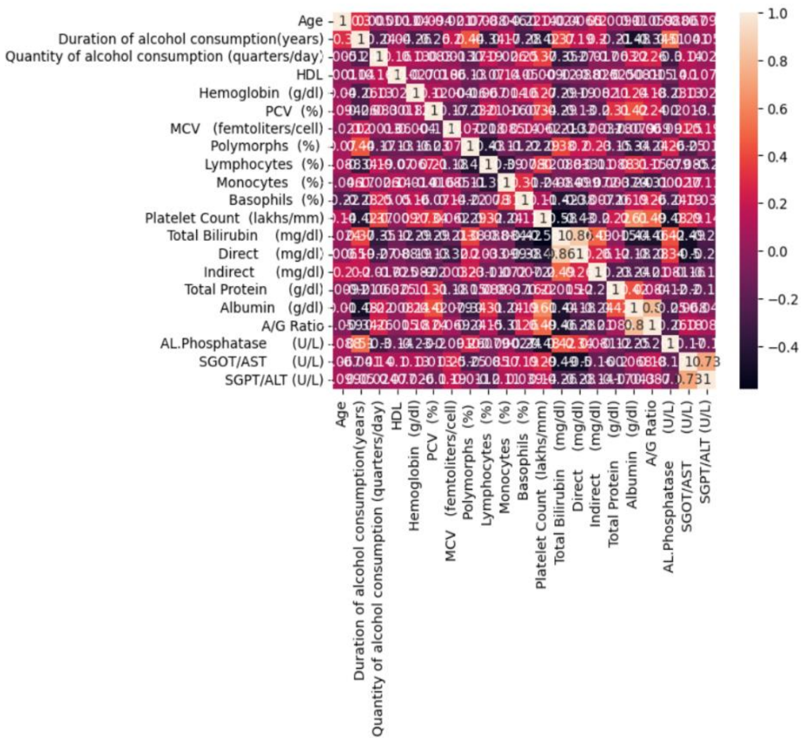
Correlation Matrix:



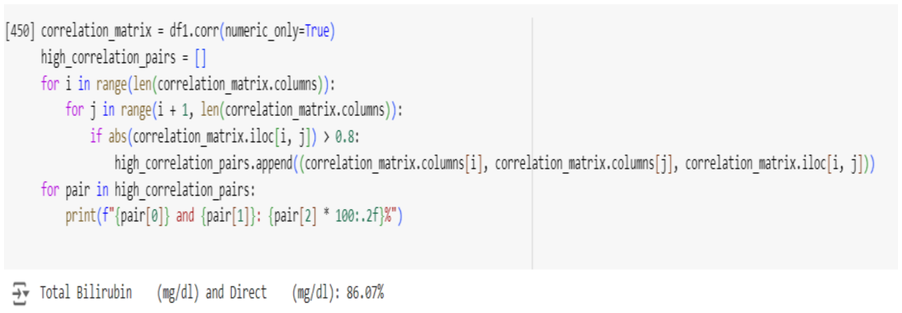
	Age	Duration of alcohol consumption(years)	Quantity of alcohol consumption (quarters/day)	HDL	Hemoglobin (g/dl)	PCV (%)	MCV (femtoliters/cell)	Polymorphs (%)	Lymphocytes (%)	Monocytes (%)	...	Platelet Count (lakhs/mm)	Total Bilirubin (mg/dl)	Direct (mg/dl)	Indirect (mg/dl)
Age	1.000000	0.504880	-0.014838	0.027021	0.005643	0.097591	0.044536	0.102051	-0.051998	-0.017115	...	0.071371	0.111111	0.056810	0.241744
Duration of alcohol consumption(years)	0.504880	1.000000	0.013511	0.102488	-0.066525	-0.003214	0.275167	0.263651	-0.363089	0.199293	...	-0.093403	-0.180130	-0.258064	0.098370
Quantity of alcohol consumption (quarters/day)	-0.014838	0.013511	1.000000	0.087929	-0.028515	-0.106242	-0.003966	-0.001482	0.102467	-0.043940	...	0.111797	-0.058668	-0.070019	0.068496
HDL	0.027021	0.102488	0.087929	1.000000	-0.031386	-0.042923	-0.003529	-0.170594	-0.014607	0.114945	...	-0.145335	0.038456	-0.001202	0.038557
Hemoglobin (g/dl)	0.005643	-0.066525	-0.028515	-0.031386	1.000000	-0.006746	-0.026800	-0.023490	0.010444	-0.008577	...	0.021393	-0.046225	0.015173	-0.019270
PCV (%)	0.097591	-0.003214	-0.106242	-0.042923	-0.006746	1.000000	-0.215279	-0.061485	0.099383	-0.014051	...	0.260573	0.006346	0.124041	-0.147214
MCV (femtoliters/cell)	0.044536	0.275167	-0.003966	-0.003529	-0.026800	-0.215279	1.000000	0.096923	-0.223638	0.114024	...	0.029725	-0.301073	-0.408698	-0.007832
Polymorphs (%)	0.102051	0.263651	-0.001482	-0.170594	-0.023490	-0.061485	0.096923	1.000000	-0.479791	-0.128826	...	-0.056582	0.096292	-0.048245	0.163071
Lymphocytes (%)	-0.051998	-0.363089	0.102467	-0.014607	0.010444	0.099383	-0.223638	-0.479791	1.000000	-0.535430	...	0.382883	0.087118	0.220090	-0.092456
Monocytes (%)	-0.017115	0.199293	-0.043940	0.114945	-0.008577	-0.014051	0.114024	-0.128826	-0.535430	1.000000	...	-0.393730	-0.153523	-0.162125	-0.072557
Basophils (%)	-0.293272	-0.107265	0.139578	-0.072160	0.052103	-0.296938	0.119891	-0.069653	-0.180878	0.334910	...	-0.221665	-0.244269	-0.336547	0.084370
Platelet Count (lakhs/mm)	0.071371	-0.093403	0.111797	-0.145335	0.021393	0.260573	0.029725	-0.056582	0.382883	-0.393730	...	1.000000	-0.215804	-0.134197	-0.164128
Total Bilirubin (mg/dl)	0.111111	-0.180130	-0.058668	0.038456	-0.046225	0.006346	-0.301073	0.096292	0.087118	-0.153523	...	-0.215804	1.000000	0.860661	0.525814
Direct (mg/dl)	0.056810	-0.258064	-0.070019	-0.001202	0.015173	0.124041	-0.408698	-0.048245	0.220090	-0.162125	...	-0.134197	0.860661	1.000000	0.226658
Indirect (mg/dl)	0.241744	0.098370	0.068496	0.038557	-0.019270	-0.147214	-0.007832	0.163071	-0.092456	-0.072557	...	-0.164128	0.525814	0.226658	1.000000
Total Protein (g/dl)	0.061647	-0.127660	-0.176093	0.038159	0.036043	0.310499	-0.193946	-0.079210	0.180241	0.005474	...	0.165374	0.217946	0.357692	-0.214989
Albumin (g/dl)	-0.025472	-0.266224	-0.004464	-0.181625	0.069044	0.348276	-0.129655	-0.153799	0.387191	-0.344347	...	0.530975	-0.060116	0.085884	-0.202710
A/G Ratio	-0.072670	-0.150173	0.128809	-0.173873	0.040801	0.120090	0.021159	-0.044456	0.197401	-0.391991	...	0.483932	-0.247892	-0.171140	-0.194788
AL.Phosphatase (U/L)	0.117445	0.312834	-0.094731	-0.143857	-0.009615	0.003518	-0.043319	0.047370	0.006548	0.009877	...	-0.119343	0.030080	0.064650	-0.036866
SGOT/AST (U/L)	0.051609	0.305667	-0.062926	0.071819	-0.009007	-0.152265	0.278110	-0.100152	-0.215464	0.218727	...	0.037232	-0.345566	-0.372329	-0.115588

Heatmap:

```
<Axes: >
```



Columns having high correlation:



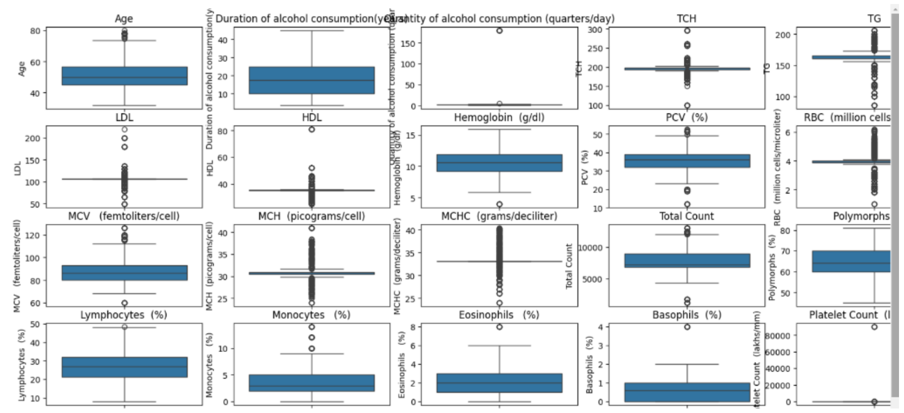
```
[450] correlation_matrix = df1.corr(numeric_only=True)
      high_correlation_pairs = []
      for i in range(len(correlation_matrix.columns)):
          for j in range(i + 1, len(correlation_matrix.columns)):
              if abs(correlation_matrix.iloc[i, j]) > 0.8:
                  high_correlation_pairs.append((correlation_matrix.columns[i], correlation_matrix.columns[j], correlation_matrix.iloc[i, j]))
      for pair in high_correlation_pairs:
          print(f"{pair[0]} and {pair[1]}: {pair[2] * 100:.2f}%")
```

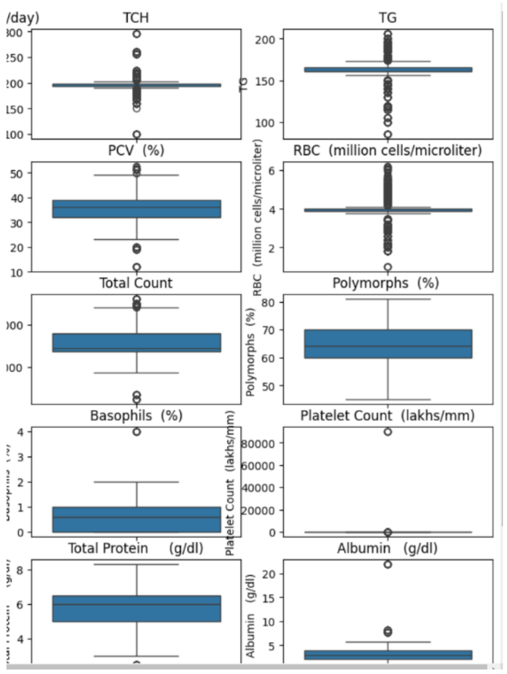
```
Total Bilirubin    (mg/dl) and Direct    (mg/dl): 86.07%
```

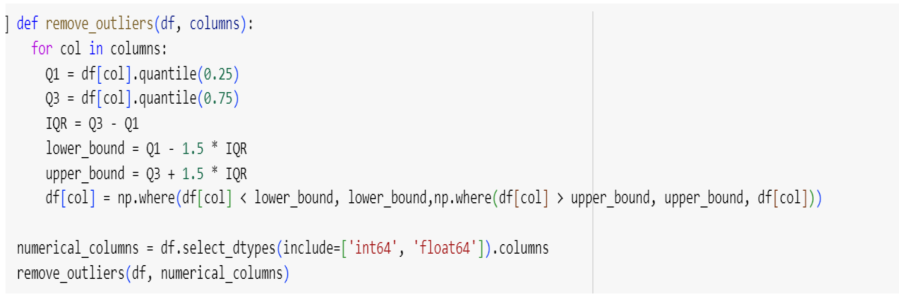
Outliers and Anomalies

Identification using boxplot:

```
c=0
plt.figure(figsize=(20,15))
for i in df.columns:
  if(type(df[i][0])!=str):
    plt.subplot(7,5,c+1)
    # Attempt to convert the column to numeric, handling errors by coercing them to NaN
    sns.boxplot(df[i].apply(pd.to_numeric, errors='coerce'))
    plt.title(i)
    c=c+1
plt.show()
```

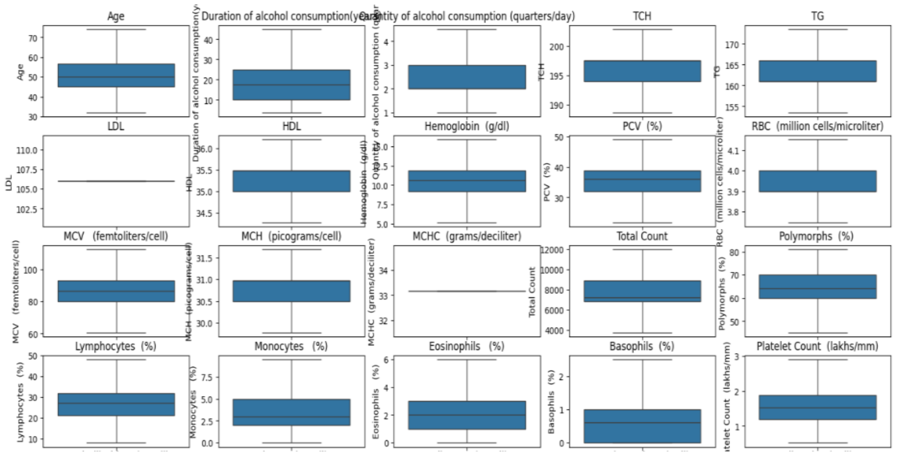


Removal  **using IQR:**



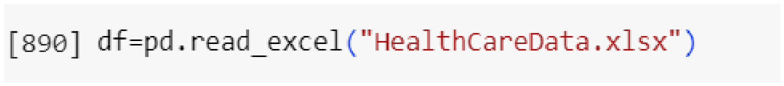
```
] def remove_outliers(df, columns):
    for col in columns:
      Q1 = df[col].quantile(0.25)
      Q3 = df[col].quantile(0.75)
      IQR = Q3 - Q1
      lower_bound = Q1 - 1.5 * IQR
      upper_bound = Q3 + 1.5 * IQR
      df[col] = np.where(df[col] < lower_bound, lower_bound,np.where(df[col] > upper_bound, upper_bound, df[col]))

  numerical_columns = df.select_dtypes(include=['int64', 'float64']).columns
  remove_outliers(df, numerical_columns)
```

After removing:



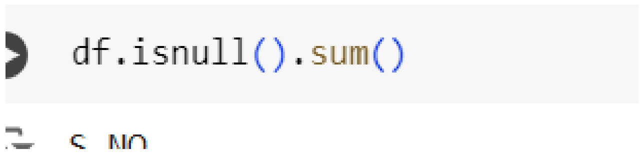
Data Preprocessing Code Screenshots

Loading Data



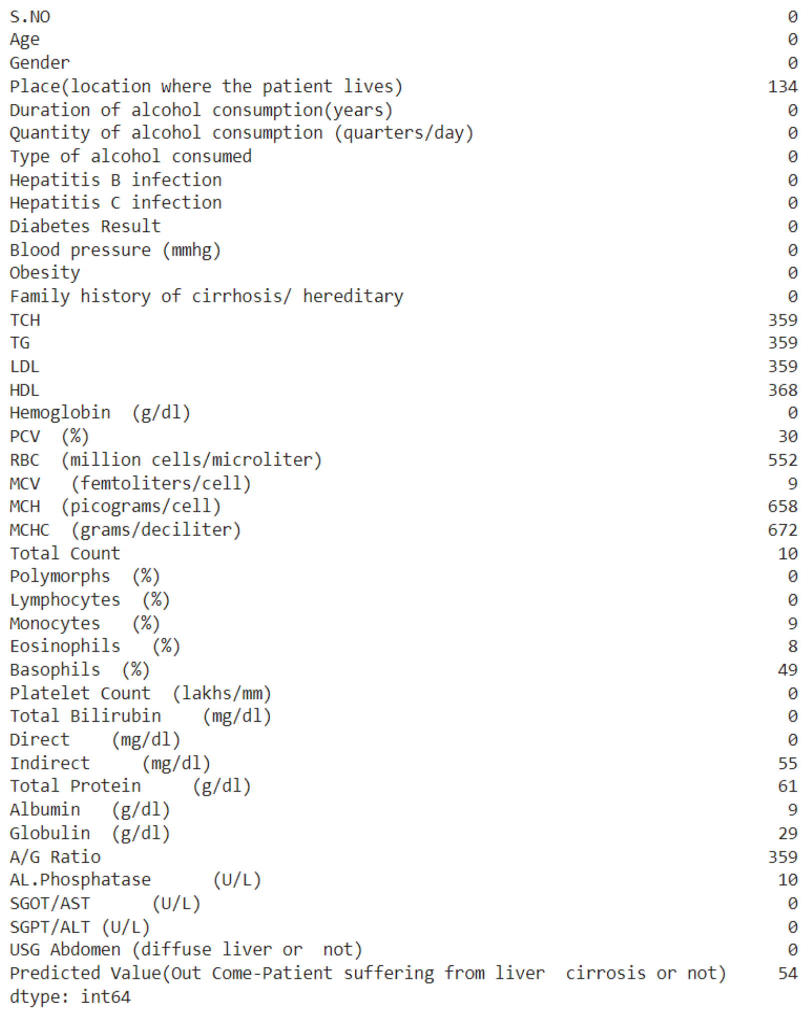
```
[890] df=pd.read_excel("HealthCareData.xlsx")
```

Handling Missing Data



```
df.isnull().sum()

S NO
```

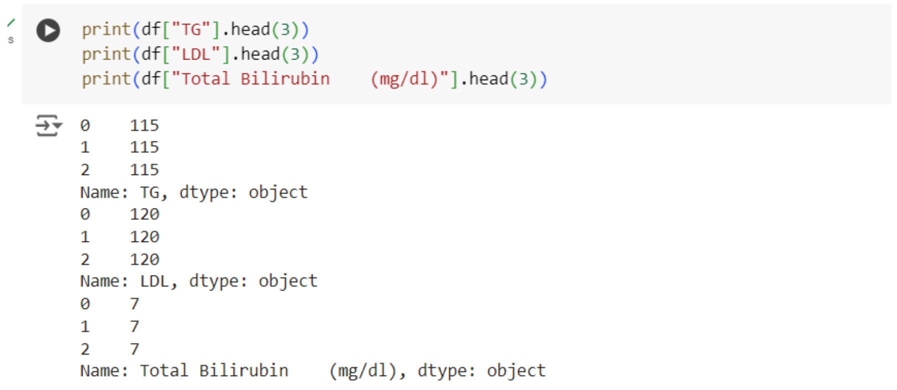
Missing values in Data:



```
S.NO                                                                        0
Age                                                                         0
Gender                                                                      0
Place(location where the patient lives)                                   134
Duration of alcohol consumption(years)                                      0
Quantity of alcohol consumption (quarters/day)                              0
Type of alcohol consumed                                                    0
Hepatitis B infection                                                       0
Hepatitis C infection                                                       0
Diabetes Result                                                             0
Blood pressure (mmhg)                                                       0
Obesity                                                                     0
Family history of cirrhosis/ hereditary                                     0
TCH                                                                       359
TG                                                                        359
LDL                                                                       359
HDL                                                                       368
Hemoglobin  (g/dl)                                                          0
PCV  (%)                                                                   30
RBC  (million cells/microliter)                                           552
MCV   (femtoliters/cell)                                                    9
MCH  (picograms/cell)                                                     658
MCHC  (grams/deciliter)                                                   672
Total Count                                                                10
Polymorphs  (%)                                                             0
Lymphocytes  (%)                                                            0
Monocytes   (%)                                                             9
Eosinophils   (%)                                                           8
Basophils  (%)                                                             49
Platelet Count  (lakhs/mm)                                                  0
Total Bilirubin    (mg/dl)                                                  0
Direct    (mg/dl)                                                           0
Indirect     (mg/dl)                                                       55
Total Protein     (g/dl)                                                   61
Albumin   (g/dl)                                                            9
Globulin  (g/dl)                                                           29
A/G Ratio                                                                 359
AL.Phosphatase      (U/L)                                                  10
SGOT/AST      (U/L)                                                         0
SGPT/ALT (U/L)                                                              0
USG Abdomen (diffuse liver or  not)                                         0
Predicted Value(Out Come-Patient suffering from liver  cirrosis or not)    54
dtype: int64
```

Cleaning Numerical columns:

We can see TG LDL and Bilirubin are object type but they have numeric values

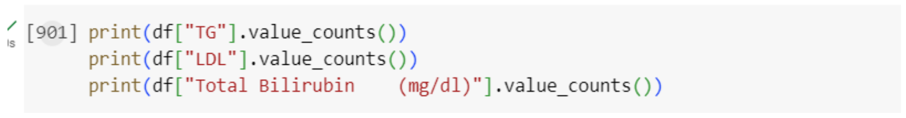


```
print(df["TG"].head(3))
print(df["LDL"].head(3))
print(df["Total Bilirubin    (mg/dl)"].head(3))
```

```
0    115
1    115
2    115
Name: TG, dtype: object
0    120
1    120
2    120
Name: LDL, dtype: object
0    7
1    7
2    7
Name: Total Bilirubin    (mg/dl), dtype: object
```

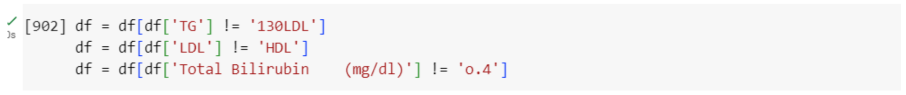
By using value_counts() we can notice that:

- TG contains a row - 130LDL
- LD contains a row - HDL
- Bilirubin contains a row - o.4



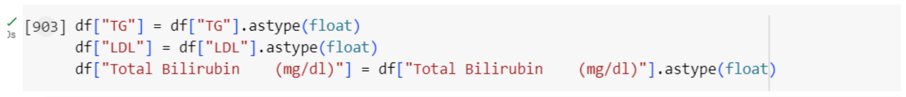
```
[901] print(df["TG"].value_counts())
      print(df["LDL"].value_counts())
      print(df["Total Bilirubin    (mg/dl)"].value_counts())
```

Dropping those rows



```
[902] df = df[df['TG'] != '130LDL']
      df = df[df['LDL'] != 'HDL']
      df = df[df['Total Bilirubin    (mg/dl)'] != 'o.4']
```

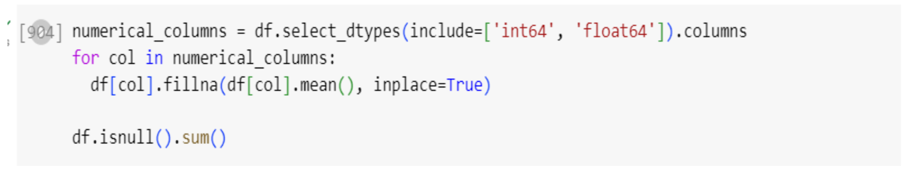
Converting into float



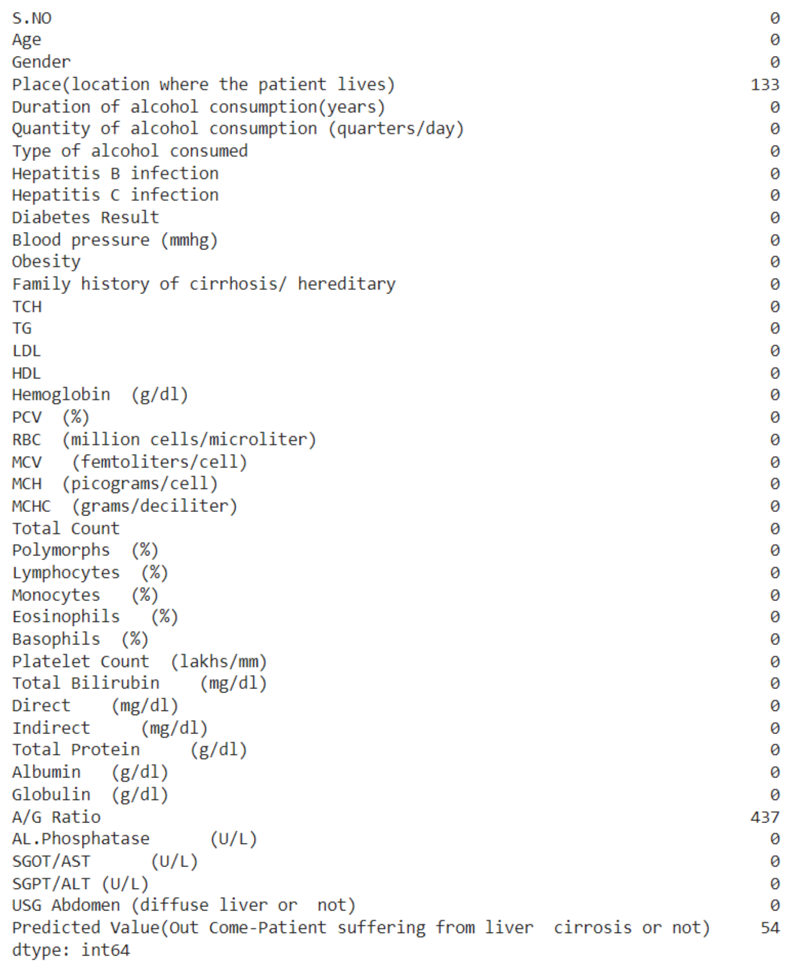
```
[903] df["TG"] = df["TG"].astype(float)
      df["LDL"] = df["LDL"].astype(float)
      df["Total Bilirubin    (mg/dl)"] = df["Total Bilirubin    (mg/dl)"].astype(float)
```

Filling numeric columns with mean:

Filling all numerical columns with their mean

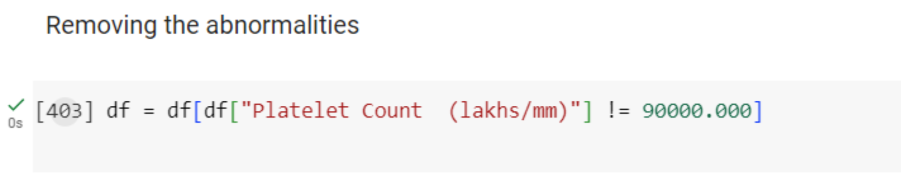


```
[904] numerical_columns = df.select_dtypes(include=['int64', 'float64']).columns
      for col in numerical_columns:
        df[col].fillna(df[col].mean(), inplace=True)

      df.isnull().sum()
```



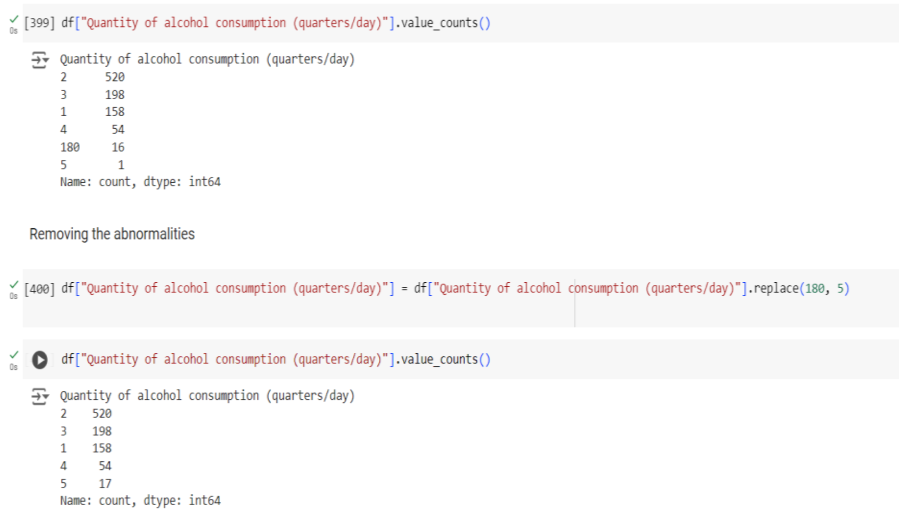
```
S.NO                                                                    0
Age                                                                     0
Gender                                                                  0
Place(location where the patient lives)                               133
Duration of alcohol consumption(years)                                  0
Quantity of alcohol consumption (quarters/day)                          0
Type of alcohol consumed                                                0
Hepatitis B infection                                                   0
Hepatitis C infection                                                   0
Diabetes Result                                                         0
Blood pressure (mmhg)                                                   0
Obesity                                                                 0
Family history of cirrhosis/ hereditary                                 0
TCH                                                                     0
TG                                                                      0
LDL                                                                     0
HDL                                                                     0
Hemoglobin  (g/dl)                                                      0
PCV  (%)                                                                0
RBC  (million cells/microliter)                                         0
MCV   (femtoliters/cell)                                                0
MCH  (picograms/cell)                                                   0
MCHC  (grams/deciliter)                                                 0
Total Count                                                             0
Polymorphs  (%)                                                         0
Lymphocytes  (%)                                                        0
Monocytes   (%)                                                         0
Eosinophils   (%)                                                       0
Basophils  (%)                                                          0
Platelet Count  (lakhs/mm)                                              0
Total Bilirubin    (mg/dl)                                              0
Direct    (mg/dl)                                                       0
Indirect     (mg/dl)                                                    0
Total Protein     (g/dl)                                                0
Albumin   (g/dl)                                                        0
Globulin  (g/dl)                                                        0
A/G Ratio                                                             437
AL.Phosphatase      (U/L)                                               0
SGOT/AST      (U/L)                                                     0
SGPT/ALT (U/L)                                                          0
USG Abdomen (diffuse liver or  not)                                     0
Predicted Value(Out Come-Patient suffering from liver  cirrosis or not)    54
dtype: int64
```

Cleaning Abnormalities found in data:



Removing the abnormalities

```
[403] df = df[df["Platelet Count  (lakhs/mm)"] != 90000.000]
```

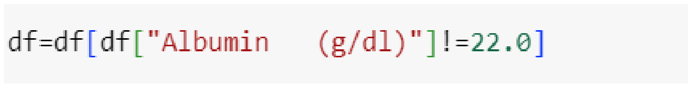


```
[399] df["Quantity of alcohol consumption (quarters/day)"].value_counts()

Quantity of alcohol consumption (quarters/day)
2      520
3      198
1      158
4       54
180     16
5        1
Name: count, dtype: int64
```

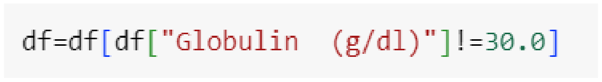
Removing the abnormalities

```
[400] df["Quantity of alcohol consumption (quarters/day)"] = df["Quantity of alcohol consumption (quarters/day)"].replace(180, 5)
```

```
df["Quantity of alcohol consumption (quarters/day)"].value_counts()

Quantity of alcohol consumption (quarters/day)
2    520
3    198
1    158
4     54
5     17
Name: count, dtype: int64
```

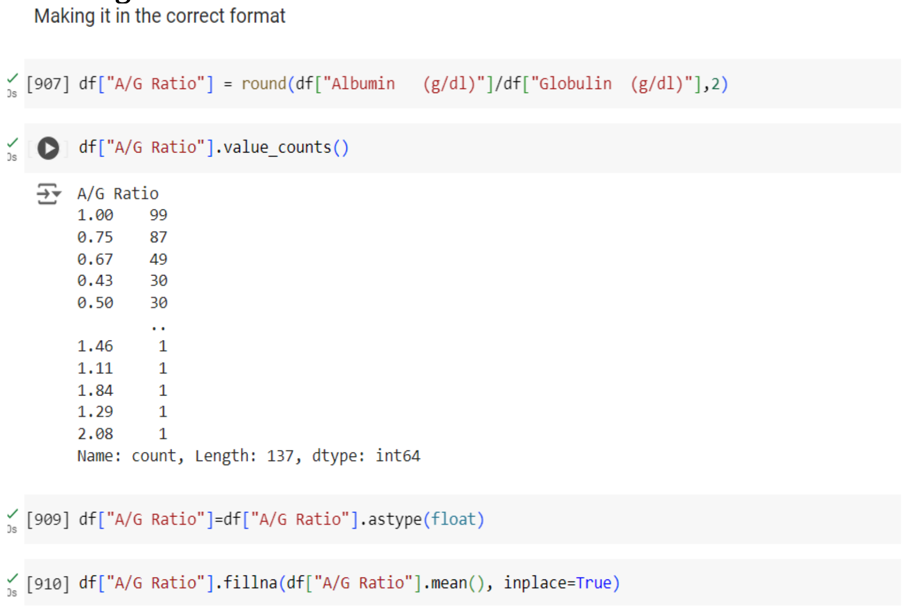


```
df=df[df["Albumin   (g/dl)"]!=22.0]
```



```
df=df[df["Globulin  (g/dl)"]!=30.0]
```

Cleaning A/G Ratio:



Making it in the correct format

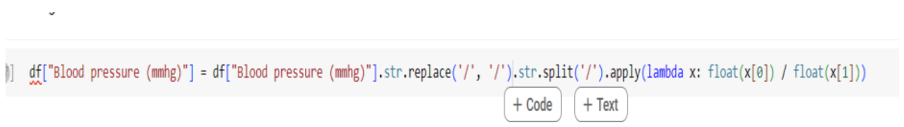
```
[907] df["A/G Ratio"] = round(df["Albumin   (g/dl)"]/df["Globulin  (g/dl)"],2)
```

```
df["A/G Ratio"].value_counts()

A/G Ratio
1.00    99
0.75    87
0.67    49
0.43    30
0.50    30
        ..
1.46     1
1.11     1
1.84     1
1.29     1
2.08     1
Name: count, Length: 137, dtype: int64
```

```
[909] df["A/G Ratio"]=df["A/G Ratio"].astype(float)
```

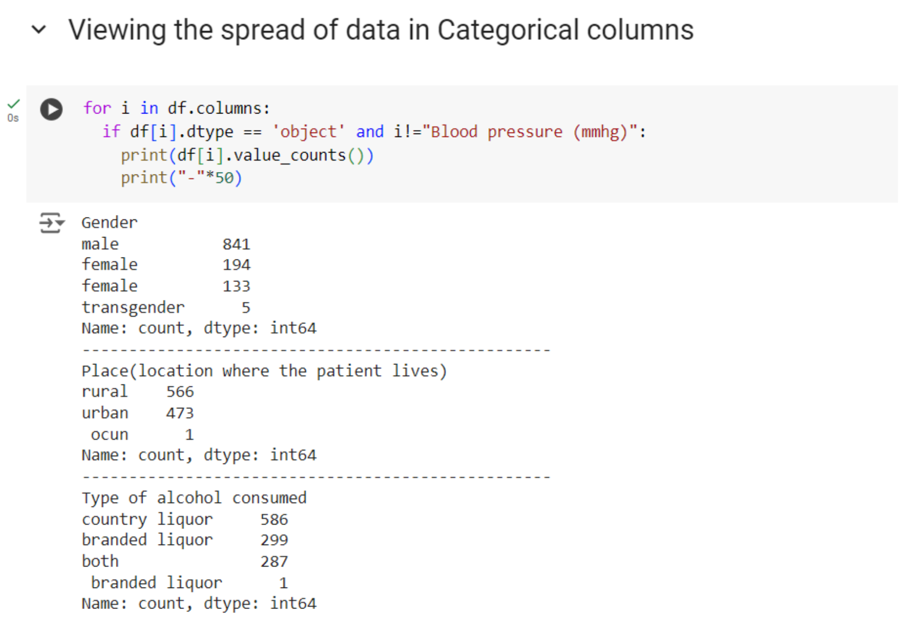
```
[910] df["A/G Ratio"].fillna(df["A/G Ratio"].mean(), inplace=True)
```

Cleaning And Transforming Blood Pressure:



```
df["Blood pressure (mmhg)"] = df["Blood pressure (mmhg)"].str.replace('/', '/').str.split('/').apply(lambda x: float(x[0]) / float(x[1]))
```

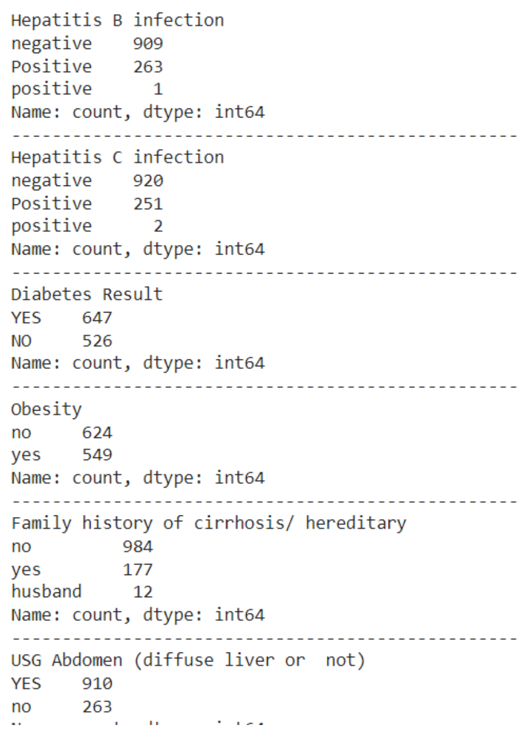
Cleansing Categorical Columns:



Viewing the spread of data in Categorical columns

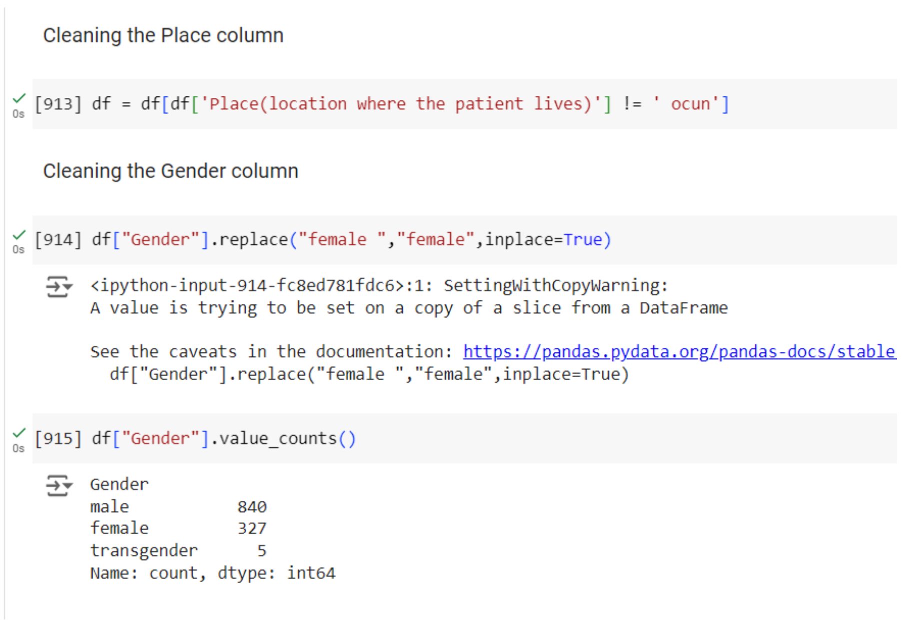
```
for i in df.columns:
  if df[i].dtype == 'object' and i!="Blood pressure (mmhg)":
    print(df[i].value_counts())
    print("-"*50)
```

```
Gender
male           841
female         194
female         133
transgender      5
Name: count, dtype: int64
--------------------------------------------------
Place(location where the patient lives)
rural     566
urban     473
 ocun       1
Name: count, dtype: int64
--------------------------------------------------
Type of alcohol consumed
country liquor     586
branded liquor     299
both               287
 branded liquor      1
Name: count, dtype: int64
```



```
Hepatitis B infection
negative    909
Positive    263
positive      1
Name: count, dtype: int64
--------------------------------------------------
Hepatitis C infection
negative    920
Positive    251
positive      2
Name: count, dtype: int64
--------------------------------------------------
Diabetes Result
YES    647
NO     526
Name: count, dtype: int64
--------------------------------------------------
Obesity
no     624
yes    549
Name: count, dtype: int64
--------------------------------------------------
Family history of cirrhosis/ hereditary
no          984
yes         177
husband      12
Name: count, dtype: int64
--------------------------------------------------
USG Abdomen (diffuse liver or  not)
YES    910
no     263
```

Removing all the abnormalities:


Cleaning the Place column

```
[913] df = df[df['Place(location where the patient lives)'] != ' ocun']
```

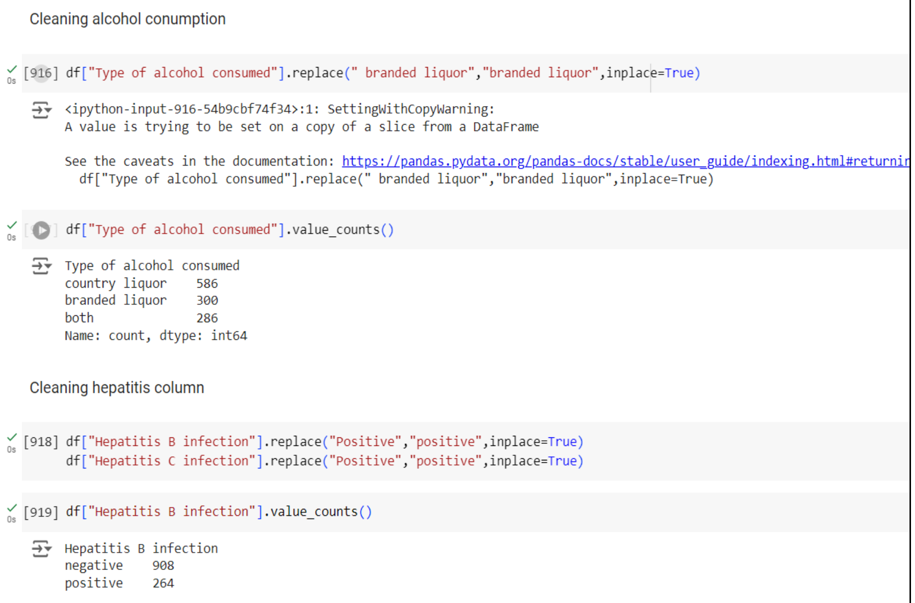
Cleaning the Gender column

```
[914] df["Gender"].replace("female ","female",inplace=True)
```

```
<ipython-input-914-fc8ed781fdc6>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable
  df["Gender"].replace("female ","female",inplace=True)
```

```
[915] df["Gender"].value_counts()
```

```
Gender
male           840
female         327
transgender      5
Name: count, dtype: int64
```


Cleaning alcohol conumption

```
[916] df["Type of alcohol consumed"].replace(" branded liquor","branded liquor",inplace=True)
```

```
<ipython-input-916-54b9cbf74f34>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returnin
  df["Type of alcohol consumed"].replace(" branded liquor","branded liquor",inplace=True)
```

```
df["Type of alcohol consumed"].value_counts()
```

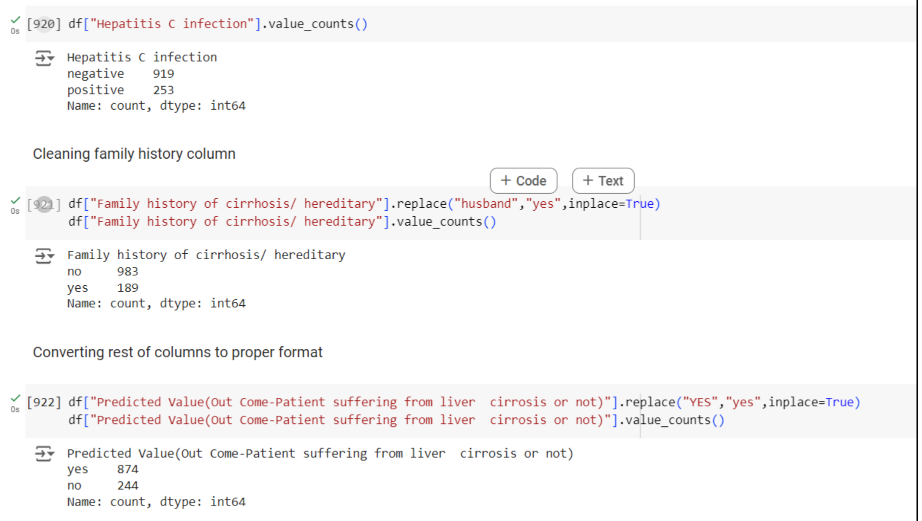
```
Type of alcohol consumed
country liquor    586
branded liquor    300
both              286
Name: count, dtype: int64
```

Cleaning hepatitis column

```
[918] df["Hepatitis B infection"].replace("Positive","positive",inplace=True)
      df["Hepatitis C infection"].replace("Positive","positive",inplace=True)
```

```
[919] df["Hepatitis B infection"].value_counts()
```

```
Hepatitis B infection
negative    908
positive    264
```



```
[920] df["Hepatitis C infection"].value_counts()

Hepatitis C infection
negative    919
positive    253
Name: count, dtype: int64
```

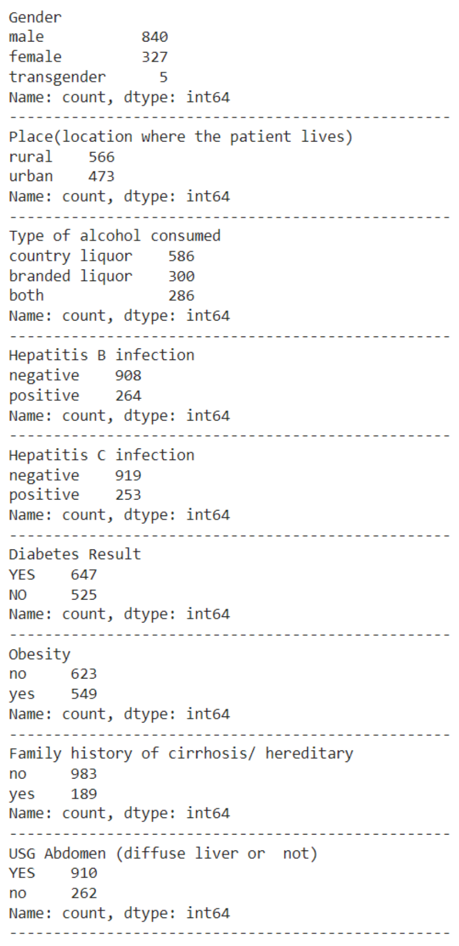
Cleaning family history column

```
[921] df["Family history of cirrhosis/ hereditary"].replace("husband","yes",inplace=True)
      df["Family history of cirrhosis/ hereditary"].value_counts()

Family history of cirrhosis/ hereditary
no     983
yes    189
Name: count, dtype: int64
```

Converting rest of columns to proper format

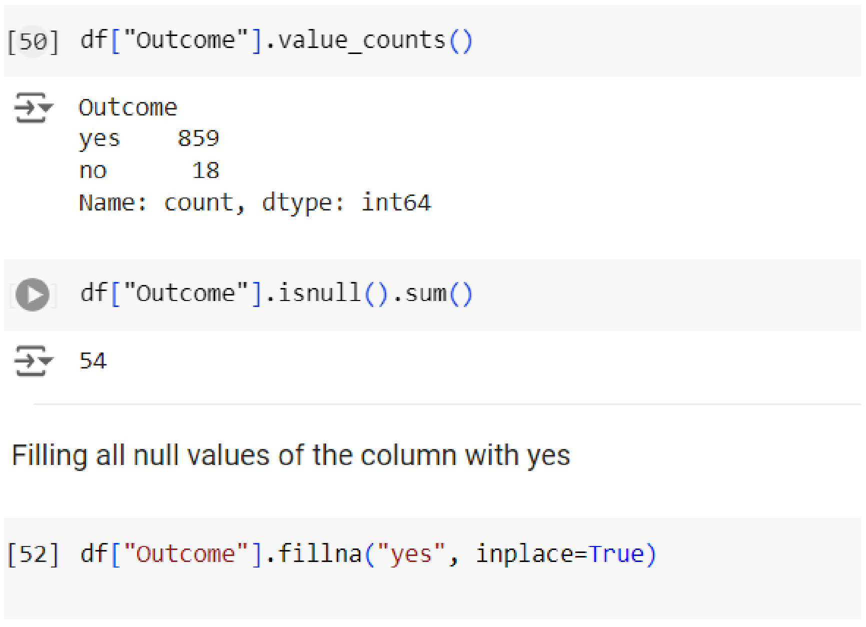
```
[922] df["Predicted Value(Out Come-Patient suffering from liver  cirrosis or not)"].replace("YES","yes",inplace=True)
      df["Predicted Value(Out Come-Patient suffering from liver  cirrosis or not)"].value_counts()

Predicted Value(Out Come-Patient suffering from liver  cirrosis or not)
yes    874
no     244
Name: count, dtype: int64
```

After cleaning:



```
Gender
male           840
female         327
transgender      5
Name: count, dtype: int64
--------------------------------------------------
Place(location where the patient lives)
rural    566
urban    473
Name: count, dtype: int64
--------------------------------------------------
Type of alcohol consumed
country liquor    586
branded liquor    300
both              286
Name: count, dtype: int64
--------------------------------------------------
Hepatitis B infection
negative    908
positive    264
Name: count, dtype: int64
--------------------------------------------------
Hepatitis C infection
negative    919
positive    253
Name: count, dtype: int64
--------------------------------------------------
Diabetes Result
YES    647
NO     525
Name: count, dtype: int64
--------------------------------------------------
Obesity
no     623
yes    549
Name: count, dtype: int64
--------------------------------------------------
Family history of cirrhosis/ hereditary
no     983
yes    189
Name: count, dtype: int64
--------------------------------------------------
USG Abdomen (diffuse liver or  not)
YES    910
no     262
Name: count, dtype: int64
--------------------------------------------------
```

Cleaning the outcome:



```
[50] df["Outcome"].value_counts()
```

```
Outcome
yes    859
no      18
Name: count, dtype: int64
```

```
df["Outcome"].isnull().sum()
```

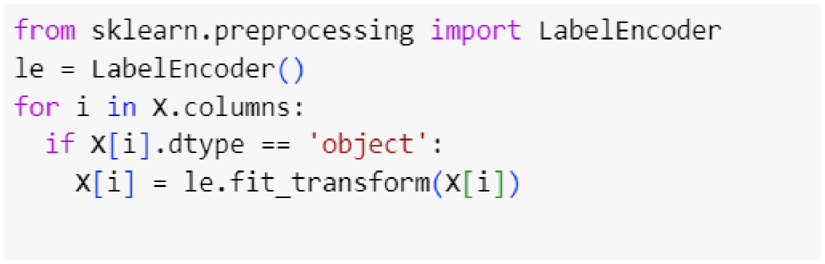
```
54
```

Filling all null values of the column with yes

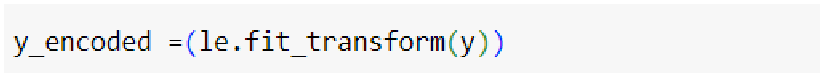
```
[52] df["Outcome"].fillna("yes", inplace=True)
```

Data Transformation

Encoding all the categorical columns:

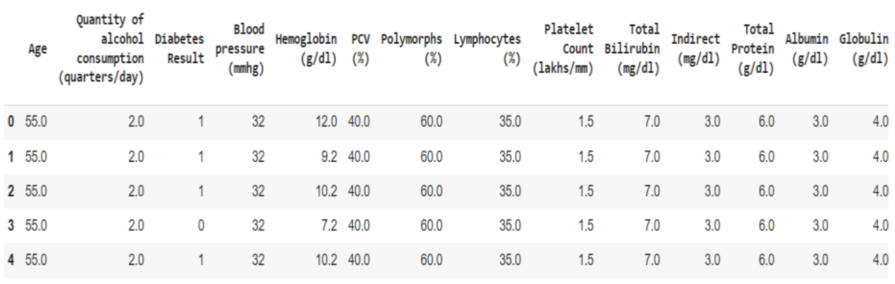


```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
for i in X.columns:
  if X[i].dtype == 'object':
    X[i] = le.fit_transform(X[i])
```



```
y_encoded =(le.fit_transform(y))
```

Encoded Data:



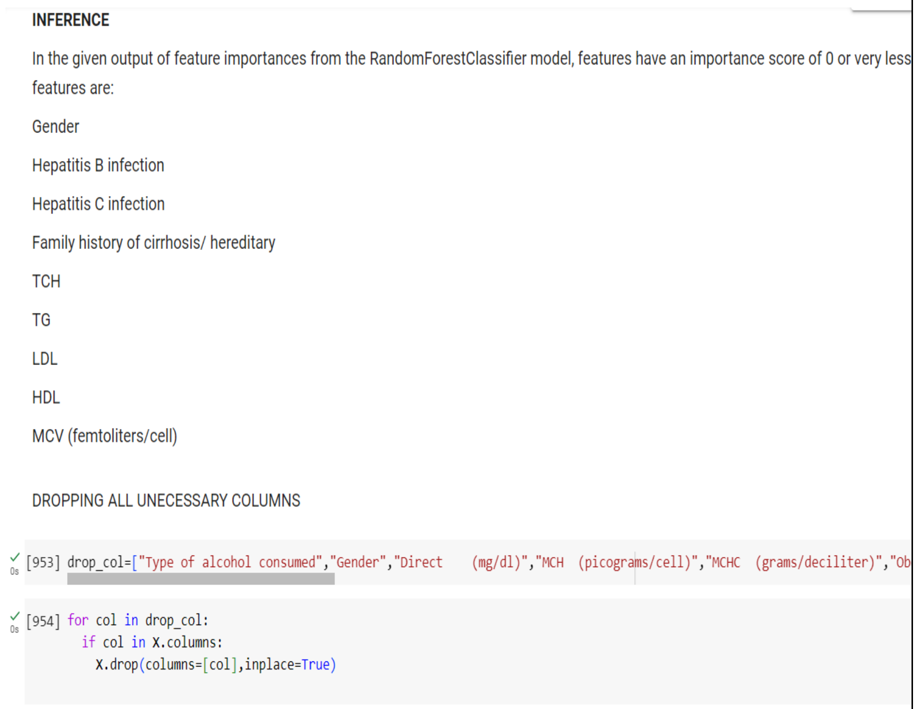
	Age	Quantity of alcohol consumption (quarters/day)	Diabetes Result	Blood pressure (mmhg)	Hemoglobin (g/dl)	PCV (%)	Polymorphs (%)	Lymphocytes (%)	Platelet Count (lakhs/mm)	Total Bilirubin (mg/dl)	Indirect (mg/dl)	Total Protein (g/dl)	Albumin (g/dl)	Globulin (g/dl)
0	55.0	2.0	1	32	12.0	40.0	60.0	35.0	1.5	7.0	3.0	6.0	3.0	4.0
1	55.0	2.0	1	32	9.2	40.0	60.0	35.0	1.5	7.0	3.0	6.0	3.0	4.0
2	55.0	2.0	1	32	10.2	40.0	60.0	35.0	1.5	7.0	3.0	6.0	3.0	4.0
3	55.0	2.0	0	32	7.2	40.0	60.0	35.0	1.5	7.0	3.0	6.0	3.0	4.0
4	55.0	2.0	1	32	10.2	40.0	60.0	35.0	1.5	7.0	3.0	6.0	3.0	4.0

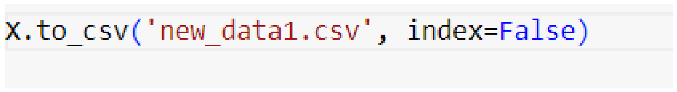
Feature Engineering

Feature Importance:

```python
from sklearn.ensemble import RandomForestClassifier

model = RandomForestClassifier(n_estimators=100)
model.fit(X, y)
importances = model.feature_importances_

# Print feature importances
for feature, importance in zip(X.columns, importances):
    print(f"{feature}: {importance:.4f}")
```

```
Age: 0.0006
Gender: 0.0000
Duration of alcohol consumption(years): 0.1940
Quantity of alcohol consumption (quarters/day): 0.0206
Type of alcohol consumed: 0.0000
Hepatitis B infection: 0.0000
Hepatitis C infection: 0.0000
Diabetes Result: 0.0044
Blood pressure (mmhg): 0.0001
Obesity: 0.0000
Family history of cirrhosis/ hereditary: 0.0001
TCH: 0.0001
TG: 0.0001
LDL: 0.0002
HDL: 0.0003
Hemoglobin  (g/dl): 0.0011
PCV  (%): 0.0007
RBC  (million cells/microliter): 0.0282
MCV   (femtoliters/cell): 0.0007
MCH  (picograms/cell): 0.0194
MCHC  (grams/deciliter): 0.0534
Total Count: 0.0010
Polymorphs  (%) : 0.0104
Lymphocytes  (%): 0.0058
Monocytes   (%): 0.0025
Eosinophils   (%): 0.0000
Basophils  (%): 0.0074
Platelet Count  (lakhs/mm): 0.0203
Total Bilirubin    (mg/dl): 0.1604
Direct    (mg/dl): 0.1125
Indirect     (mg/dl): 0.0092
Total Protein     (g/dl): 0.0024
Albumin   (g/dl): 0.0800
Globulin  (g/dl): 0.0003
A/G Ratio: 0.0518
AL.Phosphatase      (U/L): 0.0204
SGOT/AST      (U/L): 0.0199
SGPT/ALT (U/L): 0.0114
USG Abdomen (diffuse liver or  not): 0.1605
```

	Removing Unecessary Features:
	
	INFERENCE
	In the given output of feature importances from the RandomForestClassifier model, features have an importance score of 0 or very less
	features are:
	Gender
	Hepatitis B infection
	Hepatitis C infection
	Family history of cirrhosis/ hereditary
	TCH
	TG
	LDL
	HDL
	MCV (femtoliters/cell)
	DROPPING ALL UNECESSARY COLUMNS
	`[953] drop_col=["Type of alcohol consumed","Gender","Direct    (mg/dl)","MCH  (picograms/cell)","MCHC  (grams/deciliter)","Ob`
	`[954] for col in drop_col:`
	`        if col in X.columns:`
	`          X.drop(columns=[col],inplace=True)`

Save Processed Data	 `X.to_csv('new_data1.csv', index=False)`