KALINDI COLLEGE
UNIVERSITY OF DELHI

Indian Liver Patient Records

NAME	COLLEGE RNO	UNIVERSITY RNO
Anshika Varshney	21570004	21033570008
Prachi Aggarwal	21570015	21033570042

Unique Paper Code	32347507
Course	Bsc(H) Computer Science
Semester	5th Semester
Subject	Data Analysis and Visualization
Submitted to	Dr. Rajani Mam

INTRODUCTION

The Indian Liver Patient Records dataset is a collection of medical data related to liver patients in India. This dataset contains various attributes such as age, gender, total proteins, albumin, bilirubin levels, and more. The primary goal of this dataset is to facilitate research and analysis in the field of hepatology (study of liver diseases) and healthcare.

Researchers and data analysts use this dataset to explore patterns, correlations, and factors related to liver diseases among patients. It provides an opportunity to study the relationships between different clinical features and the diagnosis of liver diseases. Moreover, it allows for the development of predictive models to identify potential risk factors or indicators associated with liver ailments.

Analysing this dataset helps in understanding the demographics of liver patients, detecting potential correlations between different medical attributes, and possibly developing tools for early diagnosis or risk assessment in liver diseases. The insights drawn from this dataset can contribute to advancements in medical research, patient care, and the development of better diagnostic and treatment approaches for liver-related conditions.

The dataset for Indian liver patient records is taken from the source Kaggle:

Indian Liver Patient Records (kaggle.com)

QUERIES MADE WITH THE HELP OF INDIAN LIVER PATIENT RECORDS DATASET

Query1:Loading the Dataset: Loading the dataset from the given path and displaying the first few rows to understand its structure.

Display the first few rows of the dataset

liver_data = pd.read_csv('../input/indian-liver-patient-records/indian_liver_patient.csv')

liver_data.head()

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	A
0	65	Female	0.7	0.1	187	16	
1	62	Male	10.9	5.5	699	64	
2	62	Male	7.3	4.1	490	60	
3	58	Male	1.0	0.4	182	14	
4	72	Male	3.9	2.0	195	27	

Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	Dataset
18	6.8	3.3	0.90	1
100	7.5	3.2	0.74	1
68	7.0	3.3	0.89	1
20	6.8	3.4	1.00	1
59	7.3	2.4	0.40	1

Query2:Summary Statistics: Generating summary statistics for numerical columns like count, mean, standard deviation, etc., providing an overview of the data distribution.

Get summary statistics of numerical columns

summary_stats = liver_data.describe()

summary_stats

	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase
count	583.000000	583.000000	583.000000	583.000000	583.000000
mean	44.746141	3.298799	1.486106	290.576329	80.713551
std	16.189833	6.209522	2.808498	242.937989	182.620356
min	4.000000	0.400000	0.100000	63.000000	10.000000
25%	33.000000	0.800000	0.200000	175.500000	23.000000
50%	45.000000	1.000000	0.300000	208.000000	35.000000
75%	58.000000	2.600000	1.300000	298.000000	60.500000
max	90.000000	75.000000	19.700000	2110.000000	2000.000000

Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	Dataset
583.000000	583.000000	583.000000	579.000000	583.000000
109.910806	6.483190	3.141852	0.947064	1.286449
288.918529	1.085451	0.795519	0.319592	0.452490
10.000000	2.700000	0.900000	0.300000	1.000000
25.000000	5.800000	2.600000	0.700000	1.000000
42.000000	6.600000	3.100000	0.930000	1.000000
87.000000	7.200000	3.800000	1.100000	2.000000
4929.000000	9.600000	5.500000	2.800000	2.000000

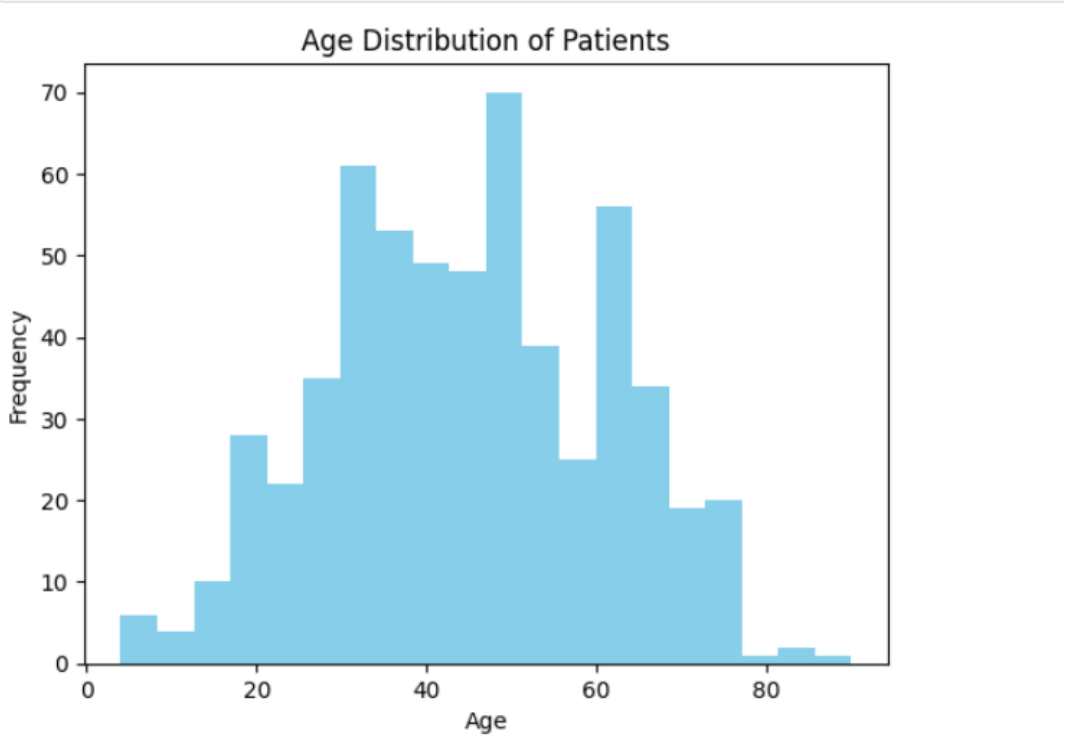
Query3:Count of Patients by Gender: Determining the number of patients based on gender to observe the gender distribution in the dataset.

```
# Count patients by gender

patient_count_by_gender = liver_data['Gender'].value_counts()

patient_count_by_gender
```

```
Gender
Male      441
Female    142
Name: count, dtype: int64
```

Query4:Age Distribution of Patients: Creating histograms to visualize the age distribution among all patients and specifically among patients diagnosed with liver disease.

```
# Plot age distribution of patients

import matplotlib.pyplot as plt

plt.hist(liver_data['Age'], bins=20, color='skyblue')

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.title('Age Distribution of Patients')

plt.show()
```



Query5: Average Total Proteins: Calculating the average total protein levels for patients with and without liver disease to observe any significant differences.

```
# Calculate average total proteins for patients with and without liver disease

avg_total_proteins_with_disease=liver_data[liver_data['Dataset'] == 1]['Total_Protiens'].mean()

avg_total_proteins_without_disease= liver_data[liver_data['Dataset'] == 2]['Total_Protiens'].mean()


print("AverageTotal Proteins for Patients with Liver Disease:", avg_total_proteins_with_disease)

print("Average Total Proteins for Patients without Liver Disease:", avg_total_proteins_without_disease)
```

```
Average Total Proteins for Patients with Liver Disease: 6.459134615384615
Average Total Proteins for Patients without Liver Disease: 6.543113772455088
```

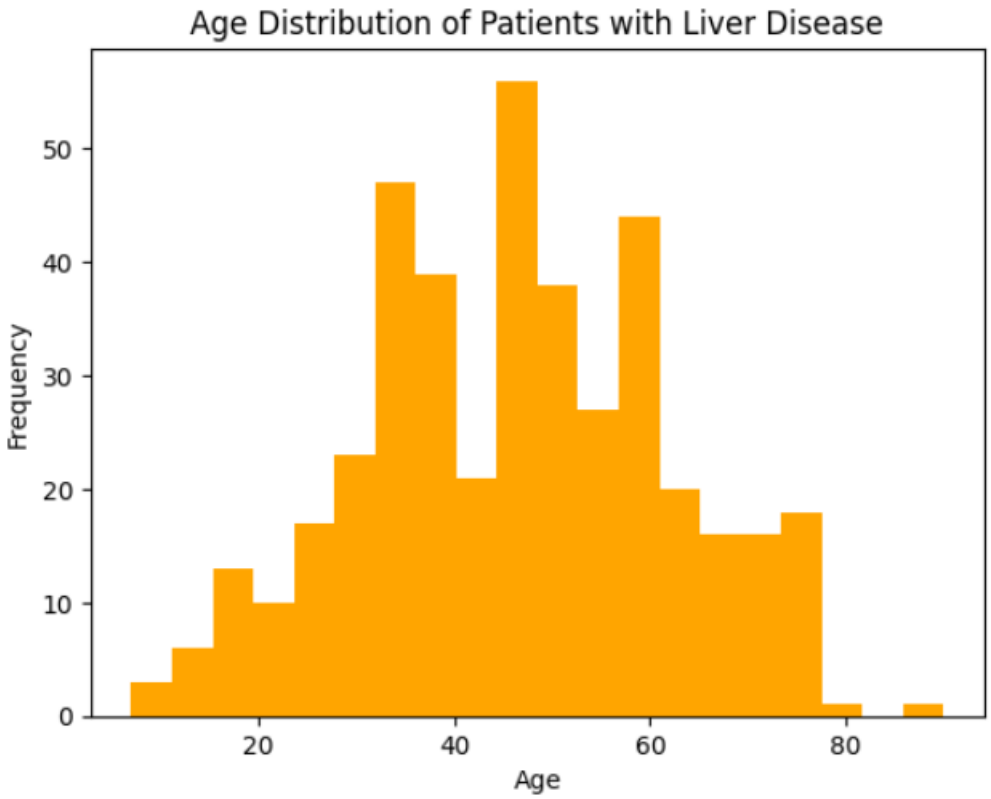
Query6: Gender-wise Distribution of Liver Disease: Understanding the count of patients diagnosed with liver disease categorized by gender.

```
# Gender-wise count of patients with liver disease

gender_liver_disease_count = liver_data[liver_data['Dataset'] == 1]['Gender'].value_counts()

print("Gender-wise distribution of patients with liver disease:")

print(gender_liver_disease_count)
```

```
Gender-wise distribution of patients with liver disease:
Gender
Male      324
Female     92
Name: count, dtype: int64
```

Query7: Age distribution of patients with liver disease: Creates a histogram to visualize the age distribution among patients diagnosed with liver disease.

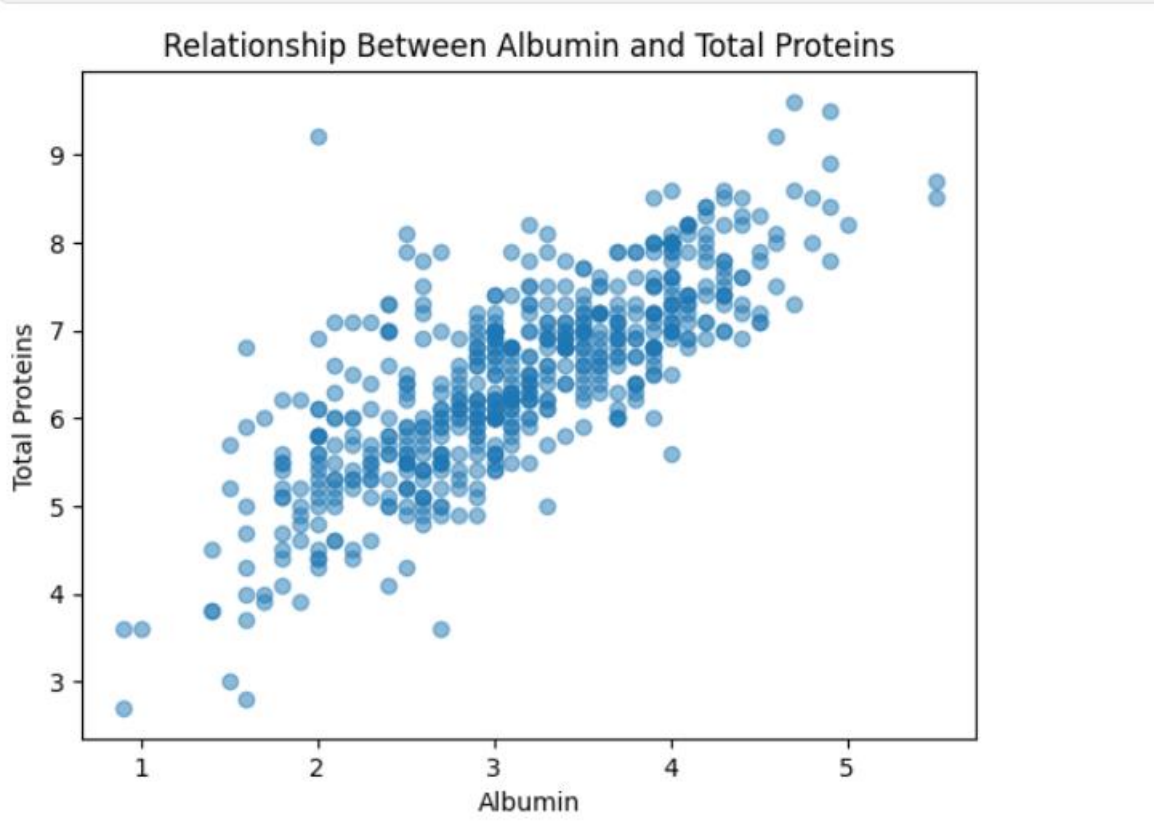
```
# Age distribution of patients with liver disease
age_liver_disease = liver_data[liver_data['Dataset'] == 1]['Age']
plt.hist(age_liver_disease, bins=20, color='orange')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.title('Age Distribution of Patients with Liver Disease')
plt.show()
```



Query8: Relationship between Albumin and Total Proteins: Displaying a scatter plot to observe the relationship between albumin and total proteins.

```
# Scatter plot of Albumin vs Total Proteins
```

```
plt.scatter(liver_data['Albumin'], liver_data['Total_Protiens'], alpha=0.5)
plt.xlabel('Albumin')
plt.ylabel('Total Proteins')
plt.title('Relationship Between Albumin and Total Proteins')
plt.show()
```

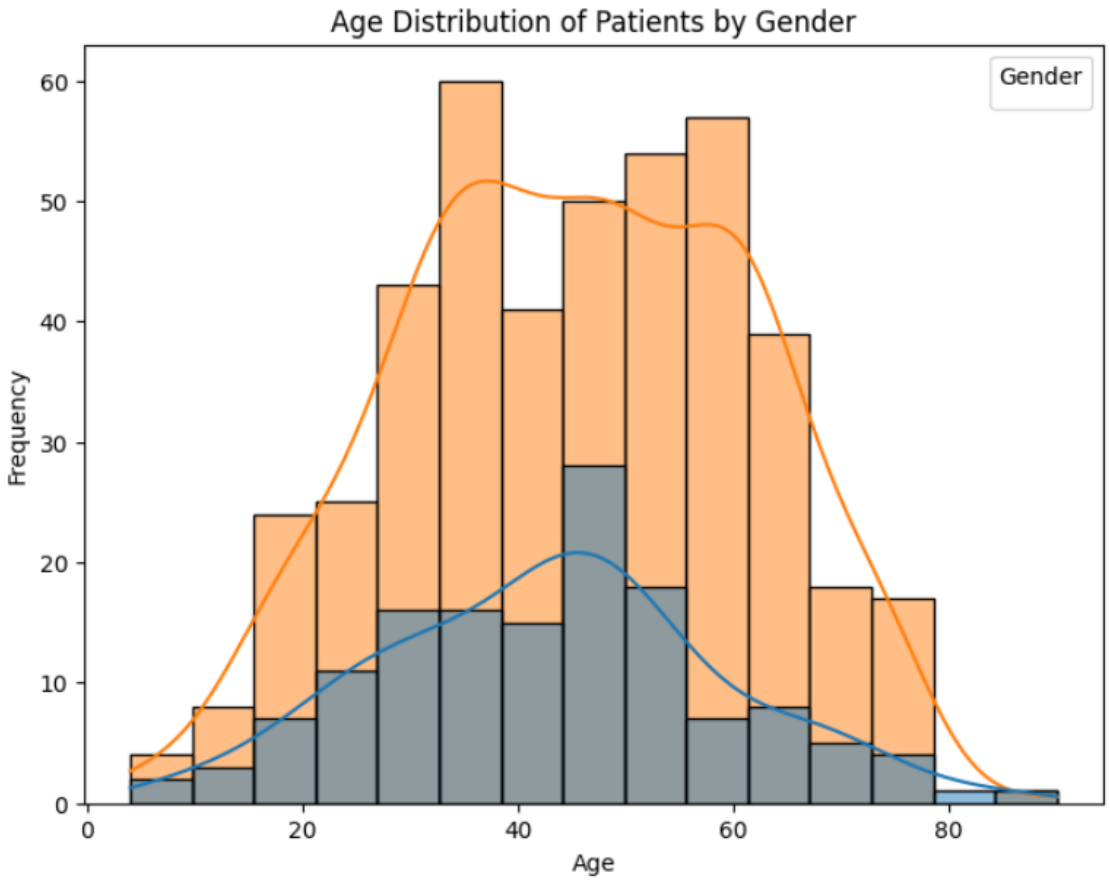


Query9:Age Distribution of Patients by Gender: Visualizing the age distribution of patients categorized by gender.

```
# Age distribution of patients by gender
plt.figure(figsize=(8, 6))
sns.histplot(data=liver_data, x='Age', hue='Gender', kde=True)
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.title('Age Distribution of Patients by Gender')
```

plt.legend(title='Gender')

plt.show()



Query10: Percentage of Patients with Liver Disease: Calculating the percentage of patients diagnosed with liver disease from the entire dataset.

Calculate the percentage of patients with liver disease

percentage_with_disease = (liver_data['Dataset'].sum() / len(liver_data)) * 100

print("Percentage of patients with liver disease:", percentage_with_disease)

```
Percentage of patients with liver disease: 128.64493996569468
```

CONCLUSION

The insights derived from the diverse analyses conducted through Python functions are expansive, delving into various facets of the dataset. These analyses encompass a wide spectrum, from grasping the nuances of demographic distributions to unraveling potential correlations between attributes such as protein levels or bilirubin. This exploration extends further, aiming to elucidate plausible connections to the diagnosis of liver disease. Through intricate computations and systematic visualizations orchestrated by the code, the dataset undergoes a thorough investigation, offering a comprehensive array of visual representations and detailed numerical synopses. This multifaceted approach significantly amplifies the understanding of the dataset's complexities and sheds light on potential interrelationships among its variables.

Now, concerning individuals at a heightened risk of liver disease, multiple factors contribute significantly to increased susceptibility. Foremost among these factors is chronic alcohol consumption, recognized as a leading cause of liver-related complications. Additionally, viral infections like hepatitis B and C pose substantial risks, often leading to chronic liver inflammation and potential long-term damage. Lifestyle elements such as obesity and diabetes are also contributors, impacting liver health due to associated metabolic dysfunctions. Exposure to specific toxins or medications can exacerbate risks, posing threats to liver function over time.

Moreover, genetic predispositions or a personal history of liver conditions may elevate the likelihood of developing liver diseases, underscoring the role of hereditary factors in susceptibility. Recognizing and comprehending these risk factors assumes paramount importance in identifying vulnerable populations. It enables the formulation and implementation of targeted preventive measures and early intervention strategies. These initiatives aim to curtail the onset or progression of liver ailments by addressing the underlying causes or predisposing conditions.

Understanding the complexity of these risk factors within diverse populations aids in tailoring interventions that consider individual susceptibilities. By emphasizing preventive healthcare measures and advocating for lifestyle modifications, medical professionals and public health practitioners can strive towards reducing the incidence and severity of liver diseases among at-risk populations. Ultimately, this multifaceted understanding serves as a cornerstone for more effective healthcare strategies, promoting better liver health and well-being on a broader scale.