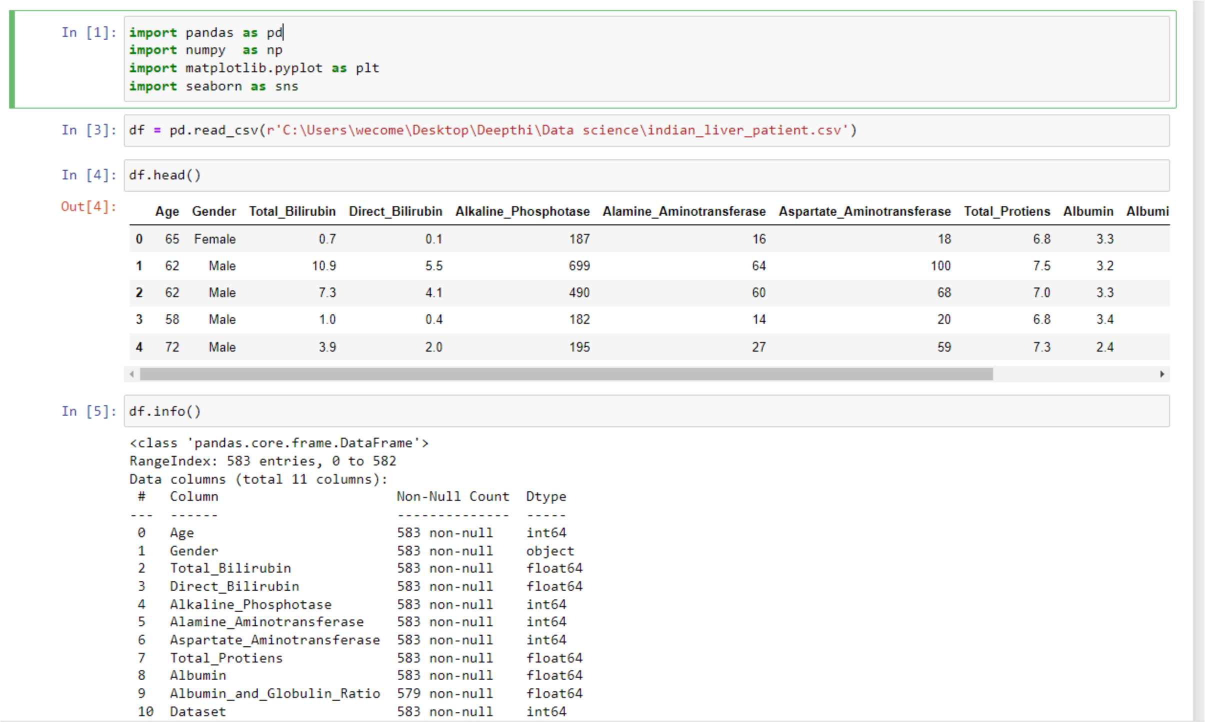
SPRINT-1 PROJECT DELIVERABLES

Data Preparation and Preprocessing

Reading the dataset:



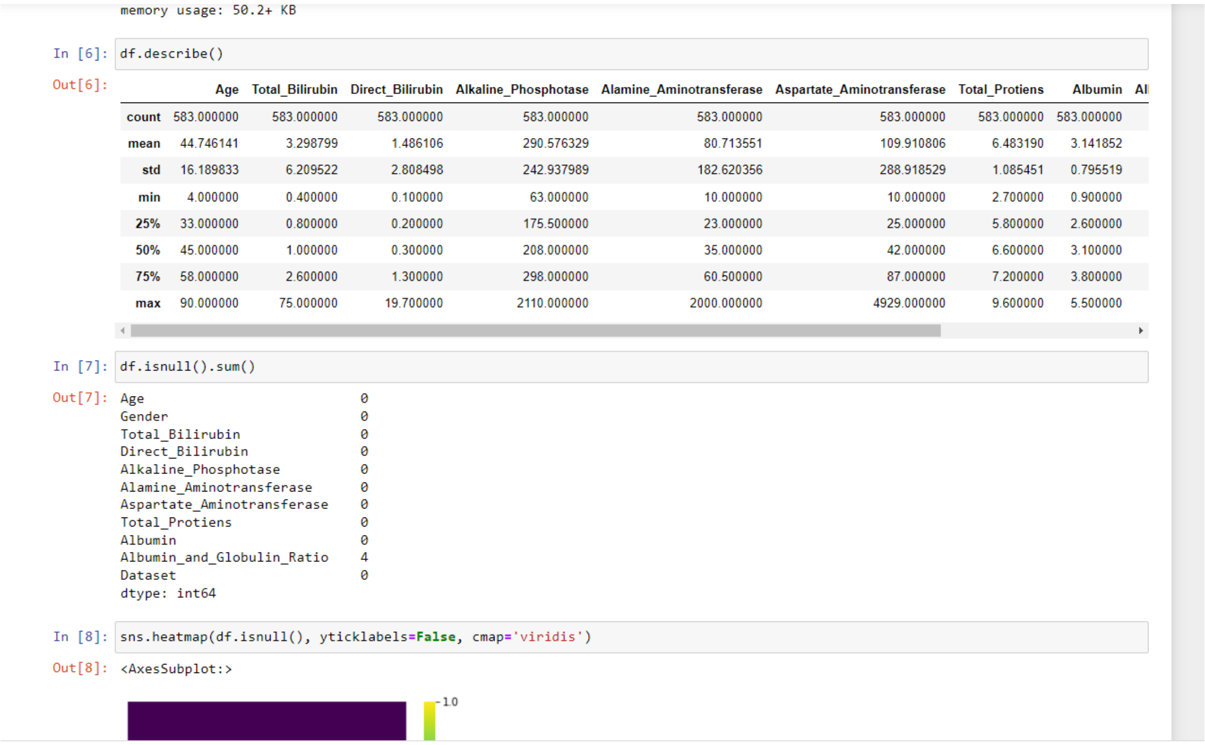
```
In [1]: import pandas as pd
        import numpy  as np
        import matplotlib.pyplot as plt
        import seaborn as sns

In [3]: df = pd.read_csv(r'C:\Users\wecome\Desktop\Deepthi\Data science\indian_liver_patient.csv')

In [4]: df.head()
```

Out[4]:

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumi
0	65	Female	0.7	0.1	187	16	18	6.8	3.3	
1	62	Male	10.9	5.5	699	64	100	7.5	3.2	
2	62	Male	7.3	4.1	490	60	68	7.0	3.3	
3	58	Male	1.0	0.4	182	14	20	6.8	3.4	
4	72	Male	3.9	2.0	195	27	59	7.3	2.4	

```
In [5]: df.info()

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 583 entries, 0 to 582
        Data columns (total 11 columns):
         #   Column                      Non-Null Count  Dtype
        ---  ------                      --------------  -----
         0   Age                         583 non-null    int64
         1   Gender                      583 non-null    object
         2   Total_Bilirubin             583 non-null    float64
         3   Direct_Bilirubin            583 non-null    float64
         4   Alkaline_Phosphotase        583 non-null    int64
         5   Alamine_Aminotransferase    583 non-null    int64
         6   Aspartate_Aminotransferase  583 non-null    int64
         7   Total_Protiens              583 non-null    float64
         8   Albumin                     583 non-null    float64
         9   Albumin_and_Globulin_Ratio  579 non-null    float64
         10  Dataset                     583 non-null    int64
```

Describe dataset:



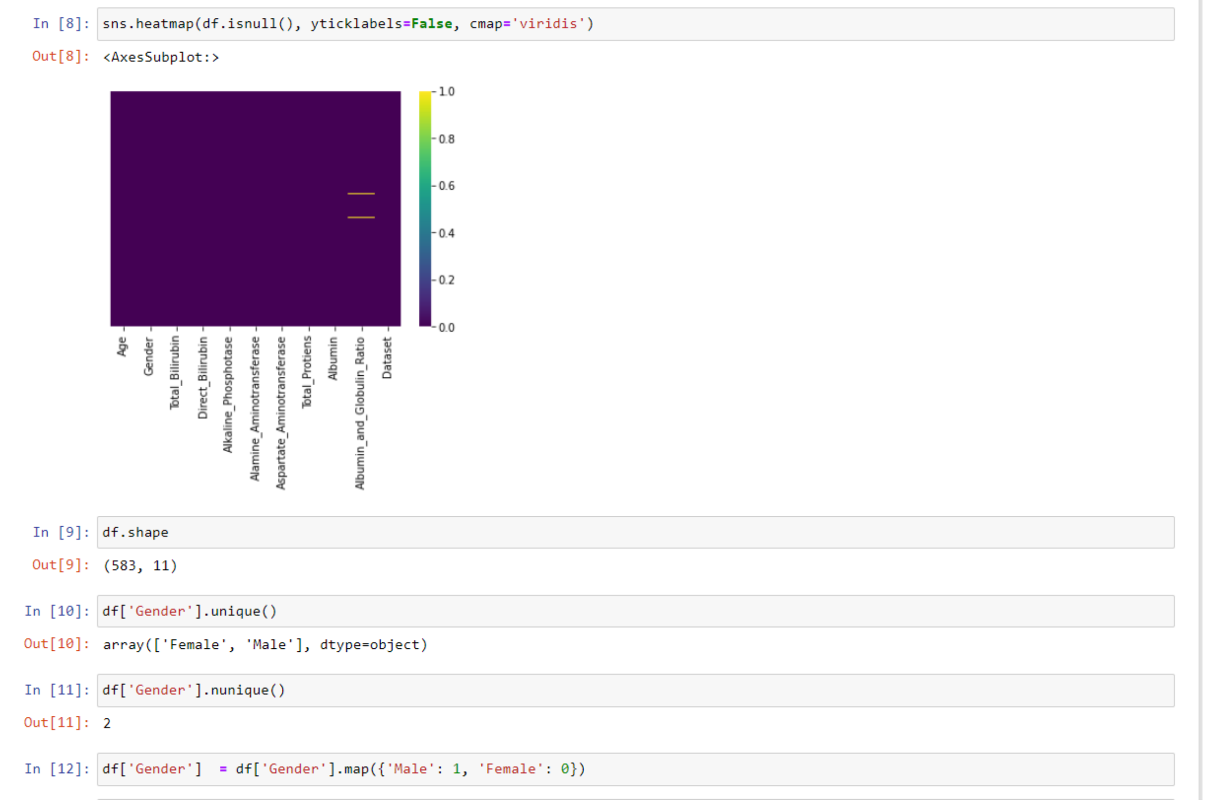
```
memory usage: 50.2+ KB

In [6]: df.describe()
```

Out[6]:

	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Al
count	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	
mean	44.746141	3.298799	1.486106	290.576329	80.713551	109.910806	6.483190	3.141852	
std	16.189833	6.209522	2.808498	242.937989	182.620356	288.918529	1.085451	0.795519	
min	4.000000	0.400000	0.100000	63.000000	10.000000	10.000000	2.700000	0.900000	
25%	33.000000	0.800000	0.200000	175.500000	23.000000	25.000000	5.800000	2.600000	
50%	45.000000	1.000000	0.300000	208.000000	35.000000	42.000000	6.600000	3.100000	
75%	58.000000	2.600000	1.300000	298.000000	60.500000	87.000000	7.200000	3.800000	
max	90.000000	75.000000	19.700000	2110.000000	2000.000000	4929.000000	9.600000	5.500000	

```
In [7]: df.isnull().sum()

Out[7]: Age                           0
        Gender                        0
        Total_Bilirubin               0
        Direct_Bilirubin              0
        Alkaline_Phosphotase          0
        Alamine_Aminotransferase      0
        Aspartate_Aminotransferase    0
        Total_Protiens                0
        Albumin                       0
        Albumin_and_Globulin_Ratio    4
        Dataset                       0
        dtype: int64

In [8]: sns.heatmap(df.isnull(), yticklabels=False, cmap='viridis')

Out[8]: <AxesSubplot:>
```

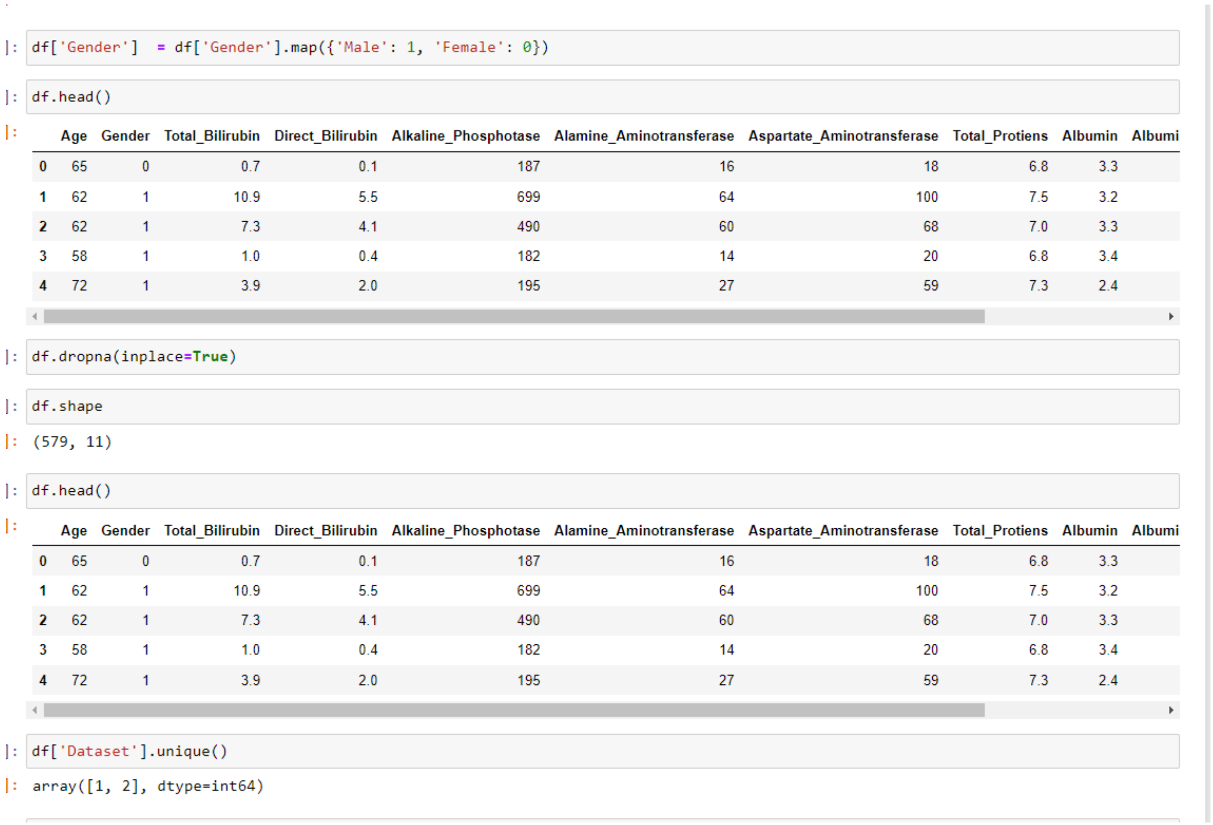
Heatmap for Visualizing Null Values:



```
In [8]: sns.heatmap(df.isnull(), yticklabels=False, cmap='viridis')
Out[8]: <AxesSubplot:>
```

```
In [9]: df.shape
Out[9]: (583, 11)

In [10]: df['Gender'].unique()
Out[10]: array(['Female', 'Male'], dtype=object)

In [11]: df['Gender'].nunique()
Out[11]: 2

In [12]: df['Gender']  = df['Gender'].map({'Male': 1, 'Female': 0})
```

Dropping Null Values from Dataset:



```
]: df['Gender']  = df['Gender'].map({'Male': 1, 'Female': 0})

]: df.head()
```

]:

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumi
0	65	0	0.7	0.1	187	16	18	6.8	3.3	
1	62	1	10.9	5.5	699	64	100	7.5	3.2	
2	62	1	7.3	4.1	490	60	68	7.0	3.3	
3	58	1	1.0	0.4	182	14	20	6.8	3.4	
4	72	1	3.9	2.0	195	27	59	7.3	2.4	

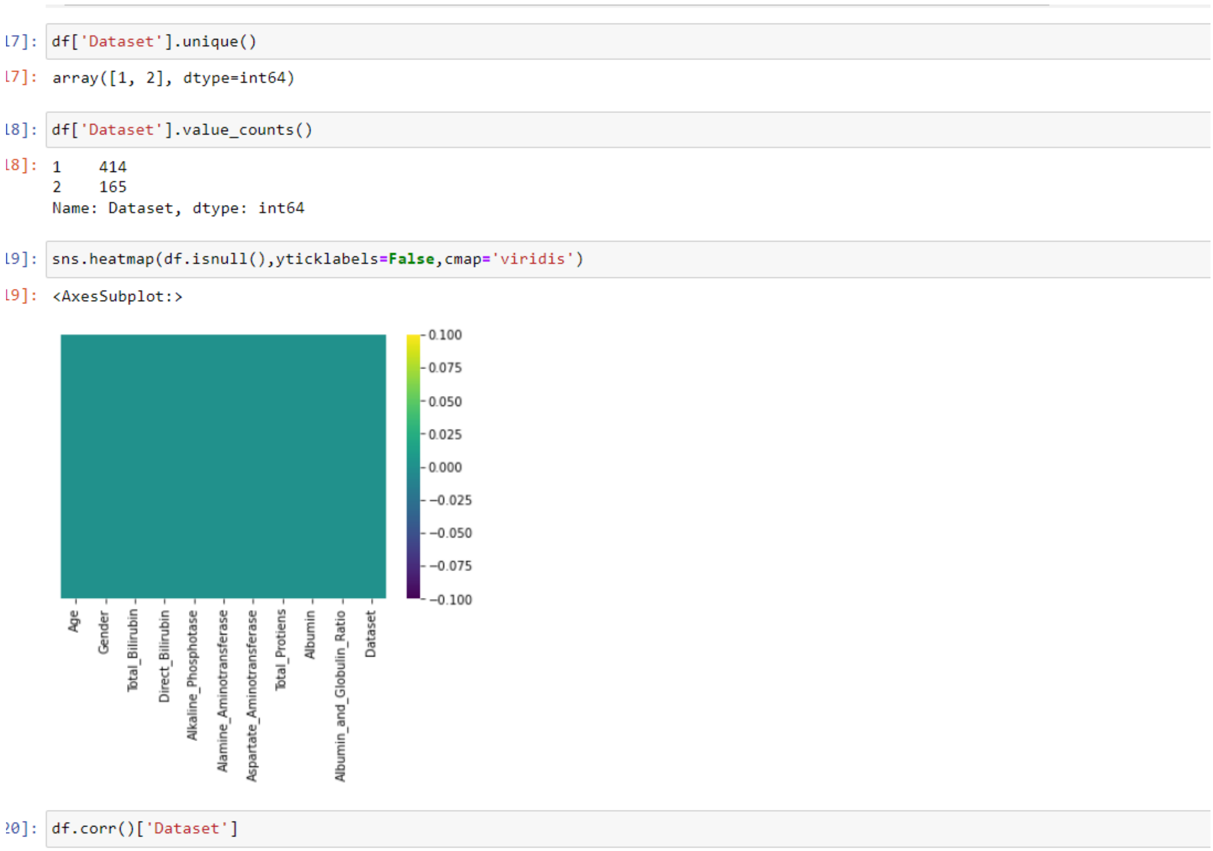
```
]: df.dropna(inplace=True)

]: df.shape
]: (579, 11)

]: df.head()
```

]:

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumi
0	65	0	0.7	0.1	187	16	18	6.8	3.3	
1	62	1	10.9	5.5	699	64	100	7.5	3.2	
2	62	1	7.3	4.1	490	60	68	7.0	3.3	
3	58	1	1.0	0.4	182	14	20	6.8	3.4	
4	72	1	3.9	2.0	195	27	59	7.3	2.4	

```
]: df['Dataset'].unique()
]: array([1, 2], dtype=int64)
```

Heatmap to check if there is any Null Value:

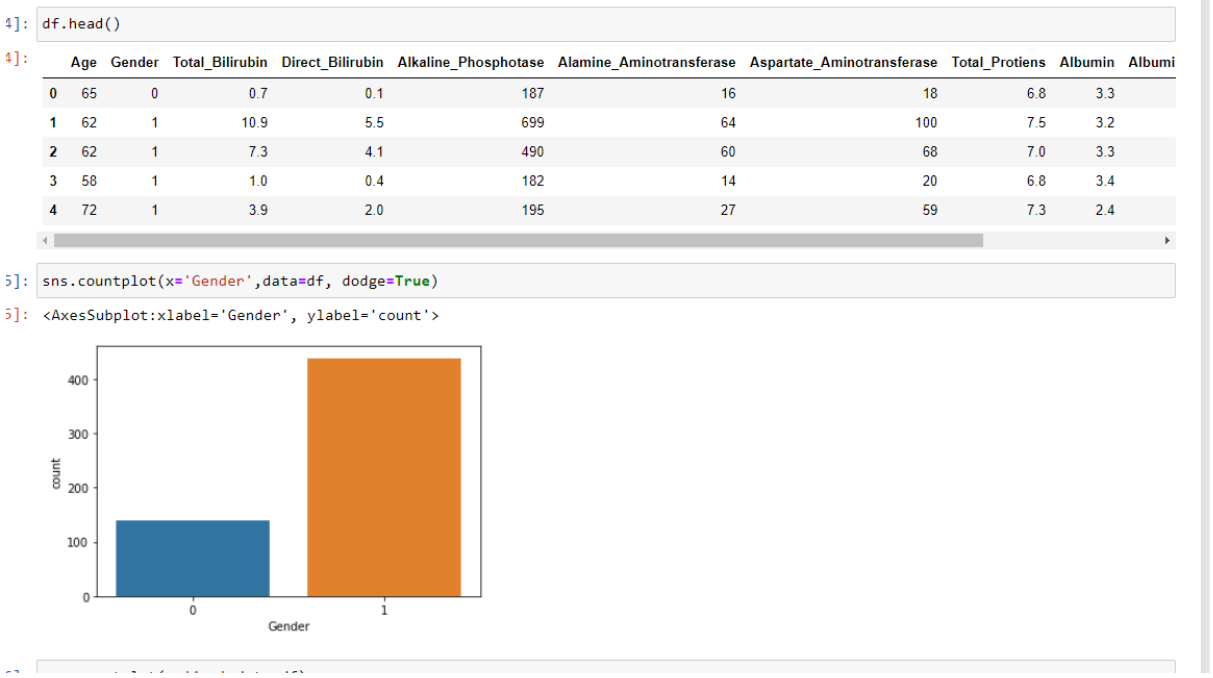


```
l7]: df['Dataset'].unique()
l7]: array([1, 2], dtype=int64)

l8]: df['Dataset'].value_counts()
l8]: 1    414
     2    165
     Name: Dataset, dtype: int64

l9]: sns.heatmap(df.isnull(),yticklabels=False,cmap='viridis')
l9]: <AxesSubplot:>
```

```
20]: df.corr()['Dataset']
```

EDA : Exploratory Data Analysis

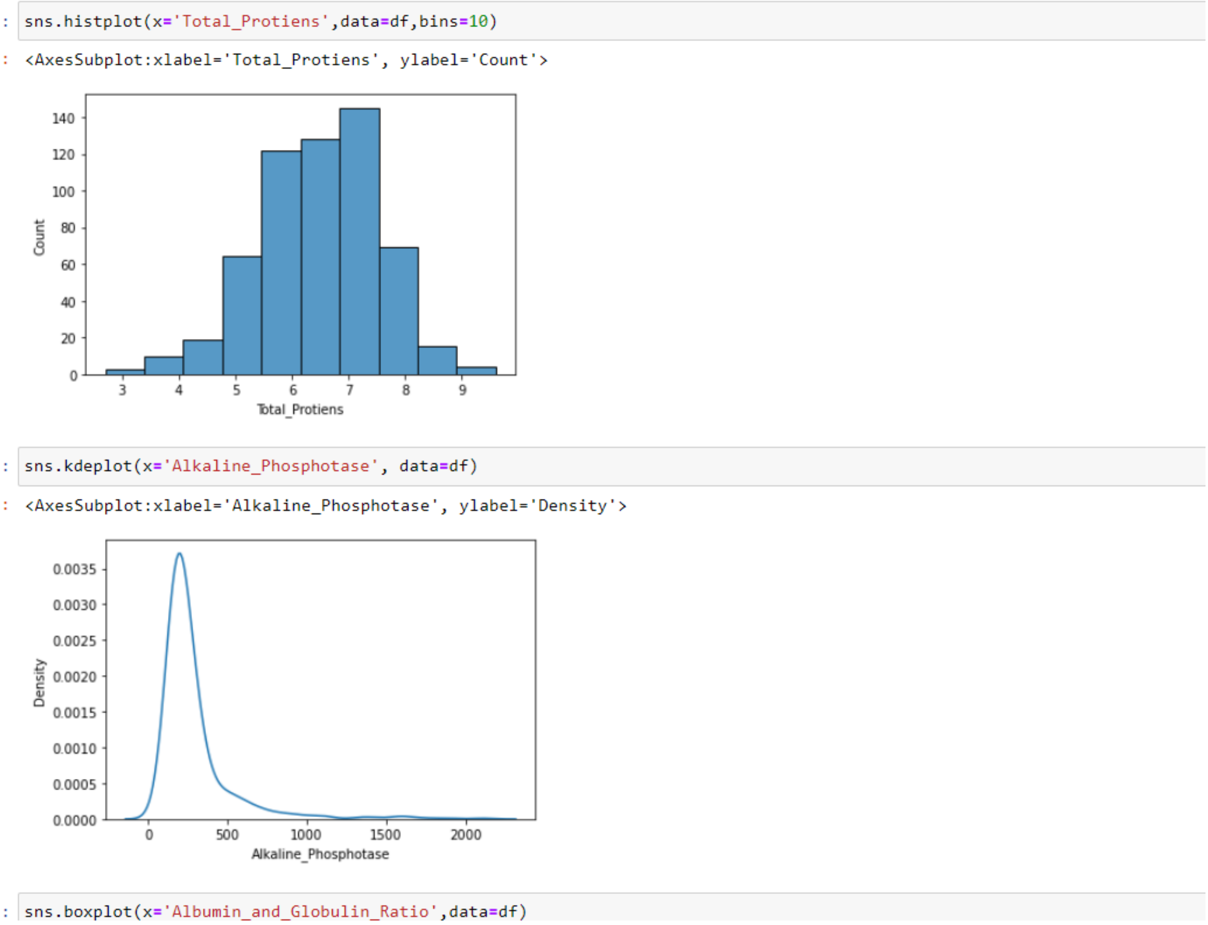
Uni – variate Analysis:



```
4]: df.head()
4]:
```

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumi
0	65	0	0.7	0.1	187	16	18	6.8	3.3	
1	62	1	10.9	5.5	699	64	100	7.5	3.2	
2	62	1	7.3	4.1	490	60	68	7.0	3.3	
3	58	1	1.0	0.4	182	14	20	6.8	3.4	
4	72	1	3.9	2.0	195	27	59	7.3	2.4	

```
5]: sns.countplot(x='Gender',data=df, dodge=True)
5]: <AxesSubplot:xlabel='Gender', ylabel='count'>
```



```
: sns.histplot(x='Total_Protiens',data=df,bins=10)
```

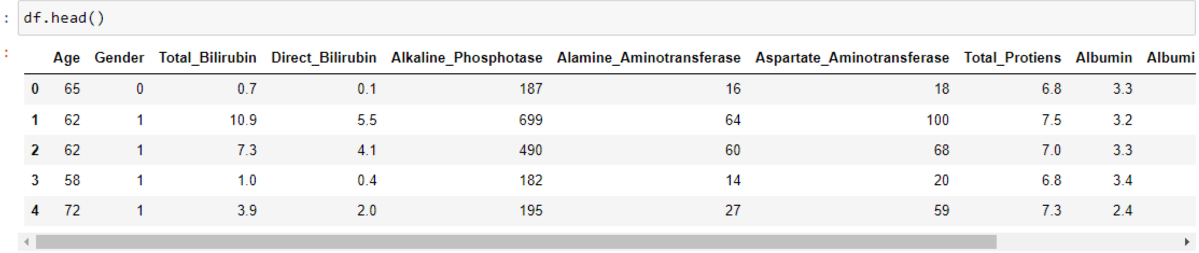
```
: <AxesSubplot:xlabel='Total_Protiens', ylabel='Count'>
```

```
: sns.kdeplot(x='Alkaline_Phosphotase', data=df)
```

```
: <AxesSubplot:xlabel='Alkaline_Phosphotase', ylabel='Density'>
```

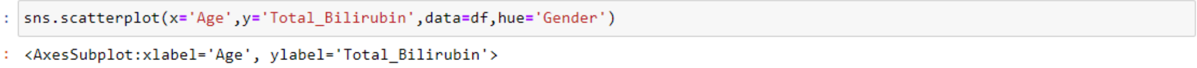
```
: sns.boxplot(x='Albumin_and_Globulin_Ratio',data=df)
```

Bi – variate Analysis:



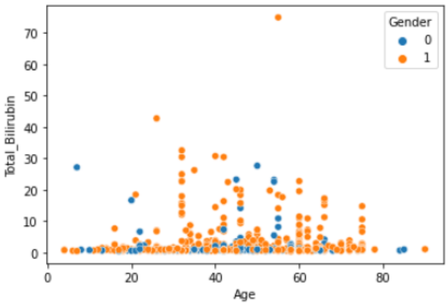
```
df.head()
```

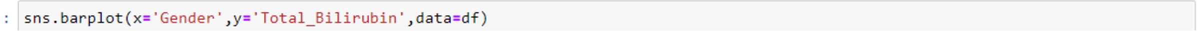
	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumi
0	65	0	0.7	0.1	187	16	18	6.8	3.3	
1	62	1	10.9	5.5	699	64	100	7.5	3.2	
2	62	1	7.3	4.1	490	60	68	7.0	3.3	
3	58	1	1.0	0.4	182	14	20	6.8	3.4	
4	72	1	3.9	2.0	195	27	59	7.3	2.4	



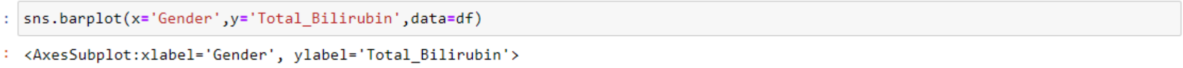
```
sns.scatterplot(x='Age',y='Total_Bilirubin',data=df,hue='Gender')
```

```
<AxesSubplot:xlabel='Age', ylabel='Total_Bilirubin'>
```



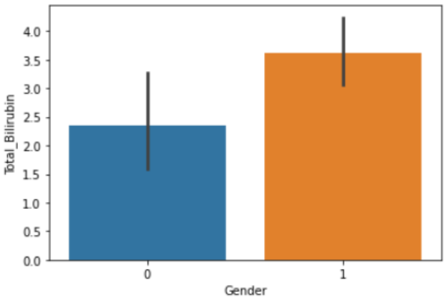


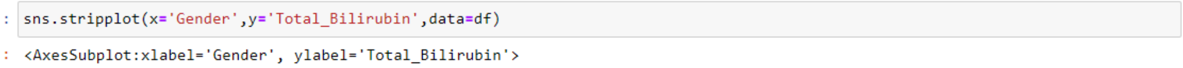
```
sns.barplot(x='Gender',y='Total_Bilirubin',data=df)
```



```
sns.barplot(x='Gender',y='Total_Bilirubin',data=df)
```

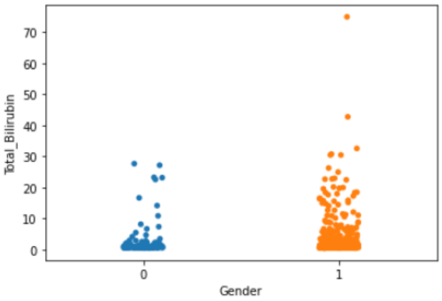
```
<AxesSubplot:xlabel='Gender', ylabel='Total_Bilirubin'>
```



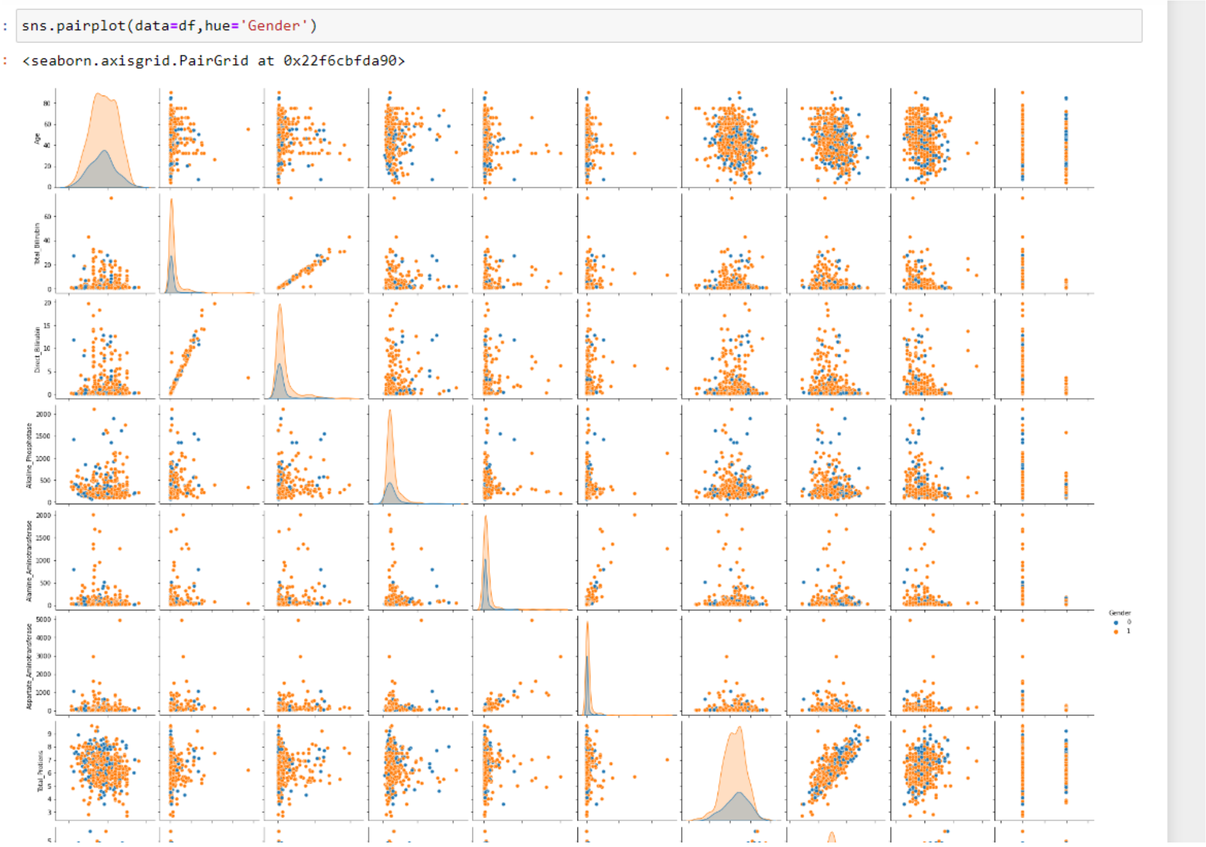


```
sns.stripplot(x='Gender',y='Total_Bilirubin',data=df)
```

```
<AxesSubplot:xlabel='Gender', ylabel='Total_Bilirubin'>
```



Multi – variate Analysis:



```
: sns.pairplot(data=df,hue='Gender')
```

```
: <seaborn.axisgrid.PairGrid at 0x22f6cbfda90>
```