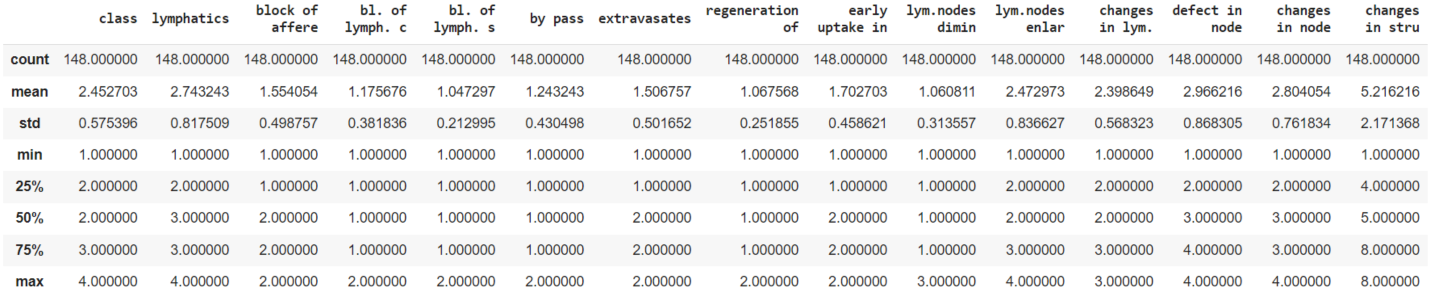
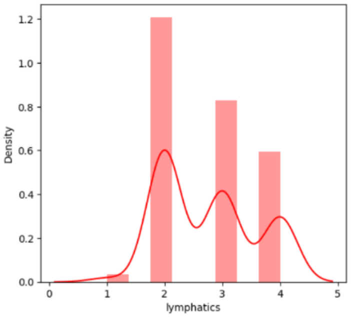
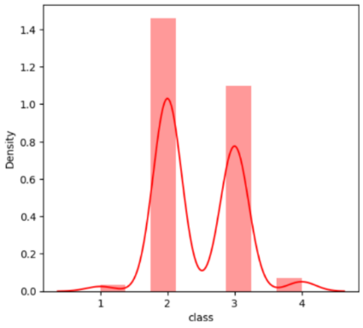
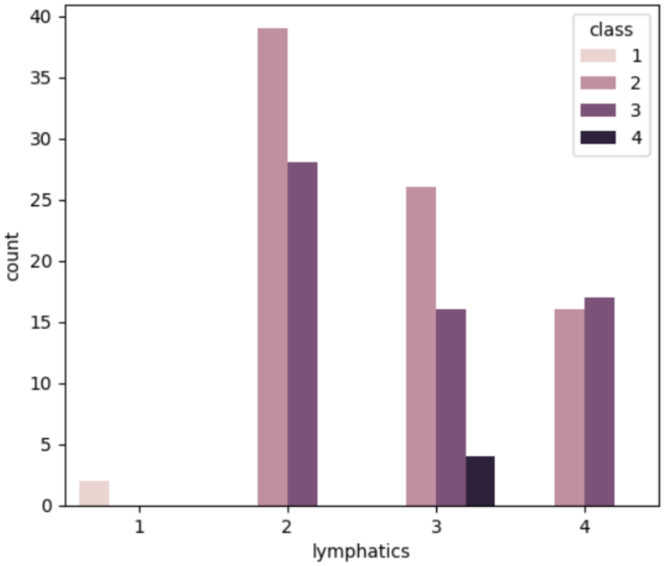
Data Collection and Preprocessing Phase

Date	12 JULY 2024
ID	740036
Project Title	Lymphography Classification using ML
Maximum Marks	6 Marks

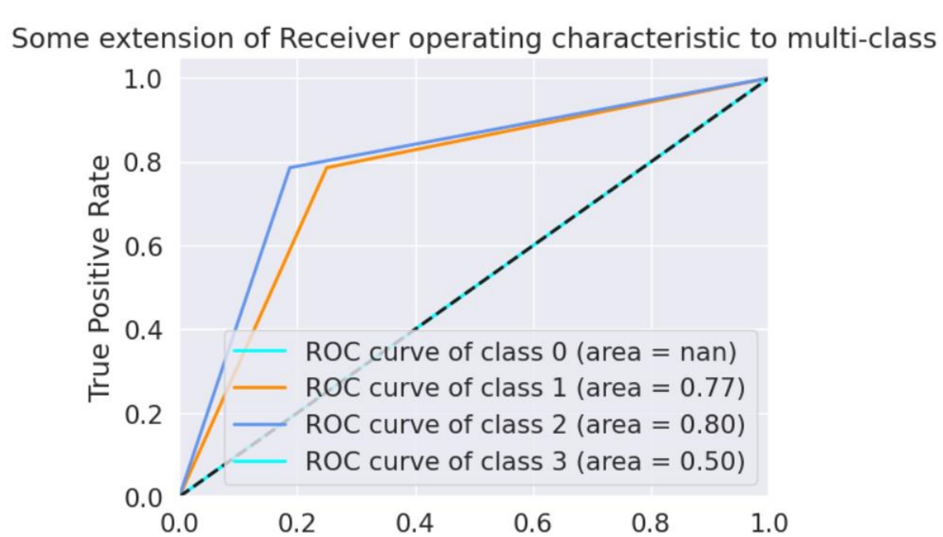
Data Exploration and Preprocessing Report

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>Data Overview</td><td>
<u>Dimension:</u>
148 rows × 19 columns
<u>Descriptive statistics:</u>

<table>
<tr><th></th><th>class</th><th>lymphatics</th><th>block of affere</th><th>bl. of lymph. c</th><th>bl. of lymph. s</th><th>by pass</th><th>extravasates</th><th>regeneration of</th><th>early uptake in</th><th>lym.nodes dimin</th><th>lym.nodes enlar</th><th>changes in lym.</th><th>defect in node</th><th>changes in node</th><th>changes in stru</th></tr>
<tr><td>count</td><td>148.000000</td><td>148.000000</td><td>148.000000</td><td>148.000000</td><td>148.000000</td><td>148.000000</td><td>148.000000</td><td>148.000000</td><td>148.000000</td><td>148.000000</td><td>148.000000</td><td>148.000000</td><td>148.000000</td><td>148.000000</td><td>148.000000</td></tr>
<tr><td>mean</td><td>2.452703</td><td>2.743243</td><td>1.554054</td><td>1.175676</td><td>1.047297</td><td>1.243243</td><td>1.506757</td><td>1.067568</td><td>1.702703</td><td>1.060811</td><td>2.472973</td><td>2.398649</td><td>2.966216</td><td>2.804054</td><td>5.216216</td></tr>
<tr><td>std</td><td>0.575396</td><td>0.817509</td><td>0.498757</td><td>0.381836</td><td>0.212995</td><td>0.430498</td><td>0.501652</td><td>0.251855</td><td>0.458621</td><td>0.313557</td><td>0.836627</td><td>0.568323</td><td>0.868305</td><td>0.761834</td><td>2.171368</td></tr>
<tr><td>min</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td></tr>
<tr><td>25%</td><td>2.000000</td><td>2.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>2.000000</td><td>2.000000</td><td>2.000000</td><td>2.000000</td><td>4.000000</td></tr>
<tr><td>50%</td><td>2.000000</td><td>3.000000</td><td>2.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>2.000000</td><td>1.000000</td><td>2.000000</td><td>1.000000</td><td>2.000000</td><td>2.000000</td><td>3.000000</td><td>3.000000</td><td>5.000000</td></tr>
<tr><td>75%</td><td>3.000000</td><td>3.000000</td><td>2.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>2.000000</td><td>1.000000</td><td>2.000000</td><td>1.000000</td><td>3.000000</td><td>3.000000</td><td>4.000000</td><td>3.000000</td><td>8.000000</td></tr>
<tr><td>max</td><td>4.000000</td><td>4.000000</td><td>2.000000</td><td>2.000000</td><td>2.000000</td><td>2.000000</td><td>2.000000</td><td>2.000000</td><td>2.000000</td><td>3.000000</td><td>4.000000</td><td>3.000000</td><td>4.000000</td><td>4.000000</td><td>8.000000</td></tr>
</table>
</td></tr>
<tr><td></td><td></td></tr>
</table>

Univariate Analysis	 
Bivariate Analysis	

Multivariate Analysis

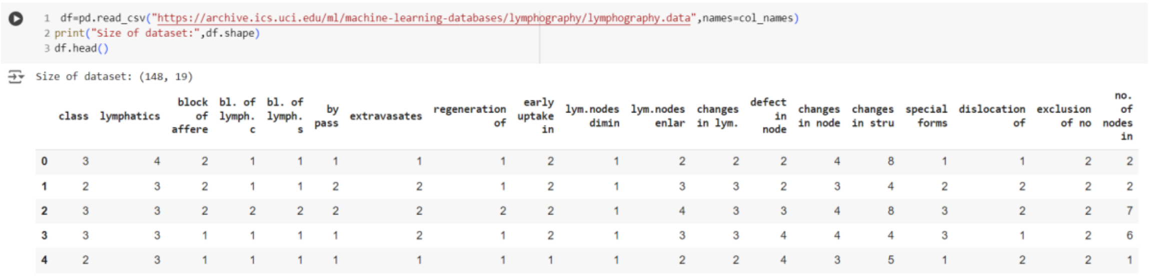


Outliers and Anomalies

-

Data Preprocessing Code Screenshots

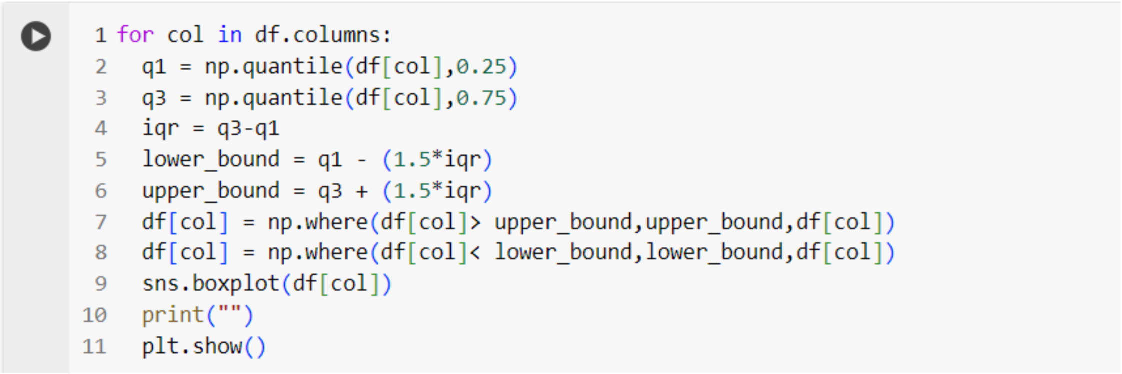
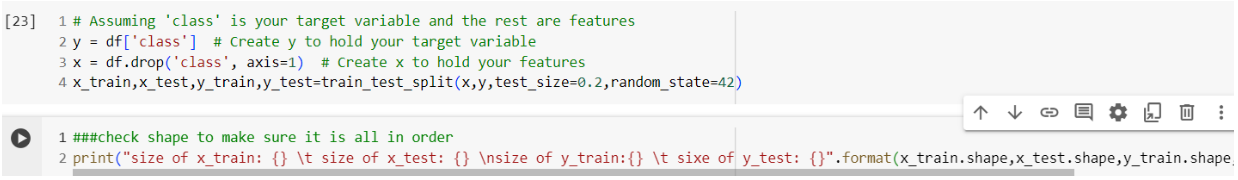
Loading Data



```
df=pd.read_csv("https://archive.ics.uci.edu/ml/machine-learning-databases/lymphography/lymphography.data",names=col_names)
print("Size of dataset:",df.shape)
df.head()
```

```
Size of dataset: (148, 19)
```

	class	lymphatics	block of affere	bl. of lymph. c	bl. of lymph. s	by pass	extravasates	regeneration of	early uptake in	lym.nodes dimin	lym.nodes enlar	changes in lym.	defect in node	changes in node	changes in stru	special forms	dislocation of	exclusion of no	no. of nodes in
0	3	4	2	1	1	1	1	1	2	1	2	2	2	4	8	1	1	2	2
1	2	3	2	1	1	2	2	1	2	1	3	3	2	3	4	2	2	2	2
2	3	3	2	2	2	2	2	2	2	1	4	3	3	4	8	3	2	2	7
3	3	3	1	1	1	1	2	1	2	1	3	3	4	4	4	3	1	2	6
4	2	3	1	1	1	1	1	1	1	1	2	2	4	3	5	1	2	2	1

<table>
<tr><td>Handling Outliers</td><td>
<pre>
for col in df.columns:
  q1 = np.quantile(df[col],0.25)
  q3 = np.quantile(df[col],0.75)
  iqr = q3-q1
  lower_bound = q1 - (1.5*iqr)
  upper_bound = q3 + (1.5*iqr)
  df[col] = np.where(df[col]> upper_bound,upper_bound,df[col])
  df[col] = np.where(df[col]< lower_bound,lower_bound,df[col])
  sns.boxplot(df[col])
  print("")
  plt.show()
</pre></td></tr>
<tr><td>Training and Testing</td><td>
<pre>
[23]
Assuming 'class' is your target variable and the rest are features
y = df['class'] # Create y to hold your target variable
x = df.drop('class', axis=1) # Create x to hold your features
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=42)
</pre>
<pre>
###check shape to make sure it is all in order
print("size of x_train: {} \t size of x_test: {} \nsize of y_train:{} \t sixe of y_test: {}".format(x_train.shape,x_test.shape,y_train.shape
</pre></td></tr>
<tr><td>Feature Engineering</td><td>Attached the codes in final submission.</td></tr>
<tr><td>Save Processed Data</td><td>-</td></tr>
</table>