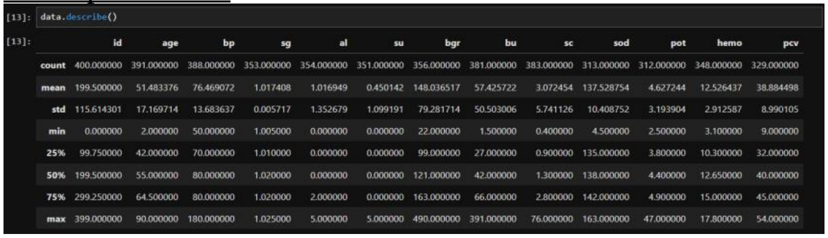
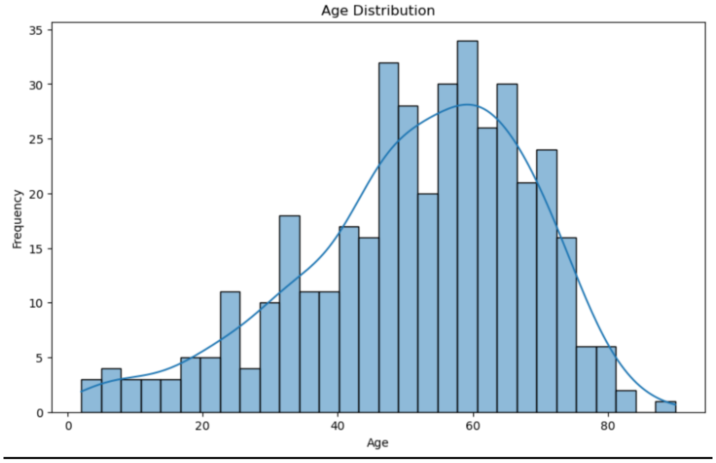
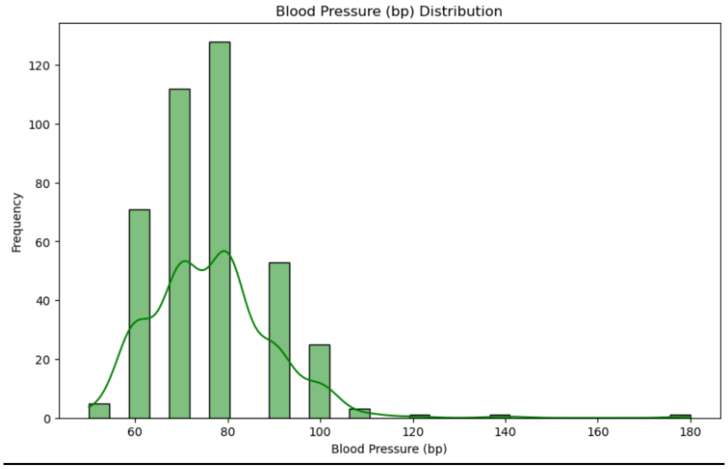
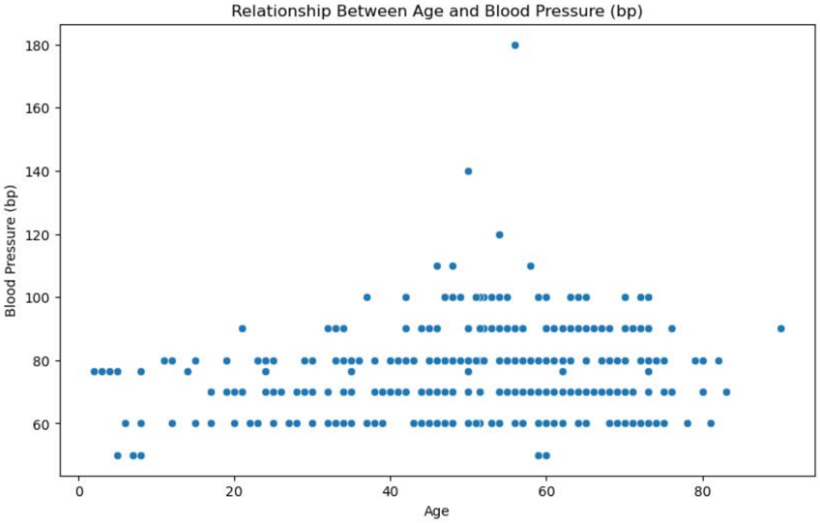
Data Collection and Preprocessing Phase

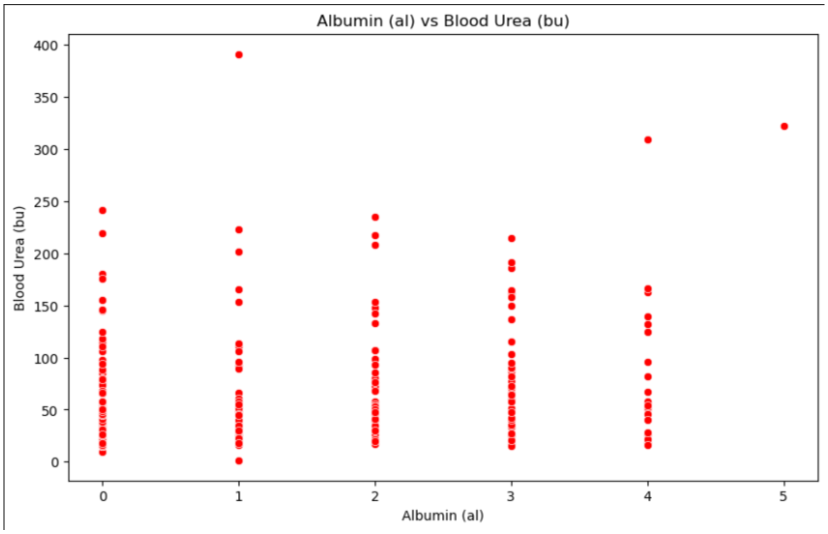
Date	15 March 2024
Team ID	SWTID1721205662
Project Title	Early Prediction of Chronic Kidney Disease Using Machine Learning
Maximum Marks	6 Marks

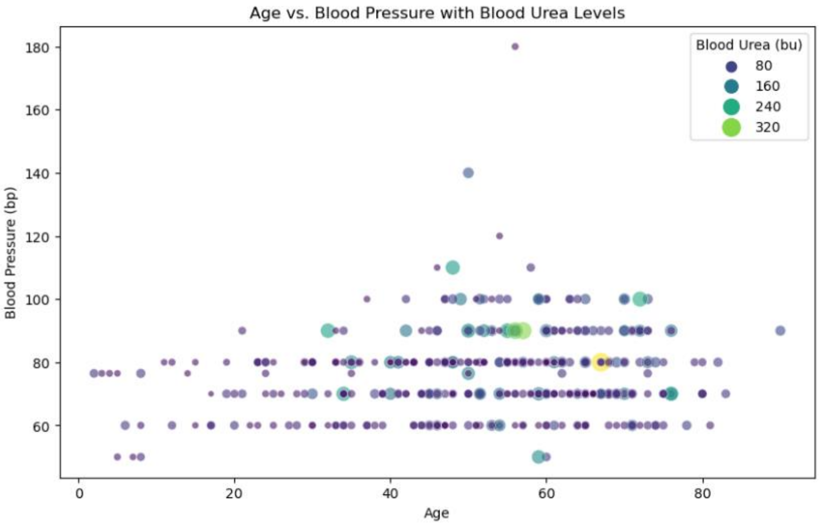
Data Exploration and Preprocessing

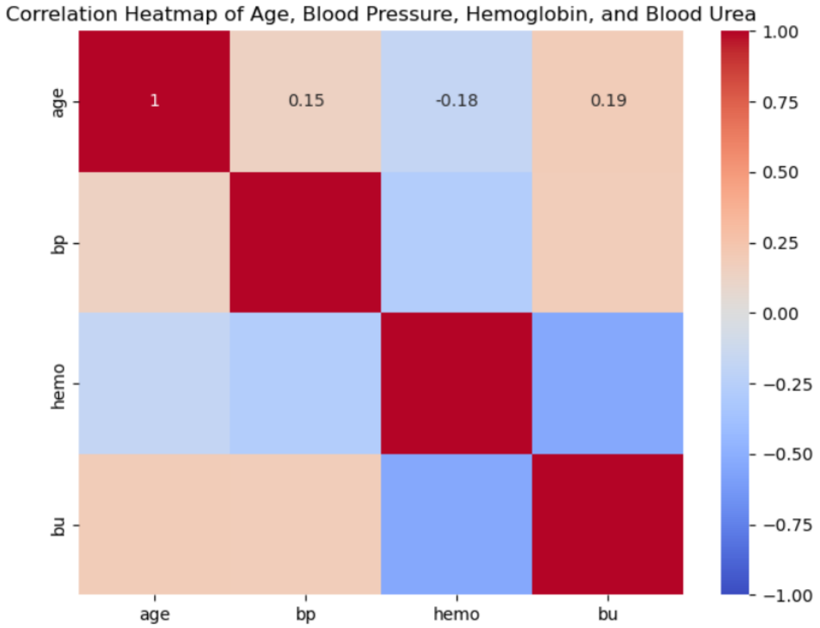
Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong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

<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>Data Overview</td><td>Dimension:
400 rows x 26 columns
Descriptive Stats:

[13]: data.describe()
[13]:
<table>
<tr><th></th><th>id</th><th>age</th><th>bp</th><th>sg</th><th>al</th><th>su</th><th>bgr</th><th>bu</th><th>sc</th><th>sod</th><th>pot</th><th>hemo</th><th>pcv</th></tr>
<tr><td>count</td><td>400.000000</td><td>391.000000</td><td>388.000000</td><td>353.000000</td><td>354.000000</td><td>351.000000</td><td>356.000000</td><td>381.000000</td><td>383.000000</td><td>313.000000</td><td>312.000000</td><td>348.000000</td><td>329.000000</td></tr>
<tr><td>mean</td><td>199.500000</td><td>51.483376</td><td>76.469072</td><td>1.017408</td><td>1.016949</td><td>0.450142</td><td>148.036517</td><td>57.425722</td><td>3.072454</td><td>137.528754</td><td>4.627244</td><td>12.526437</td><td>38.884498</td></tr>
<tr><td>std</td><td>115.614301</td><td>17.169714</td><td>13.683637</td><td>0.005717</td><td>1.352679</td><td>1.099191</td><td>79.281714</td><td>50.503006</td><td>5.741126</td><td>10.408752</td><td>3.193904</td><td>2.912587</td><td>8.990105</td></tr>
<tr><td>min</td><td>0.000000</td><td>2.000000</td><td>50.000000</td><td>1.005000</td><td>0.000000</td><td>0.000000</td><td>22.000000</td><td>1.500000</td><td>0.400000</td><td>4.500000</td><td>2.500000</td><td>3.100000</td><td>9.000000</td></tr>
<tr><td>25%</td><td>99.750000</td><td>42.000000</td><td>70.000000</td><td>1.010000</td><td>0.000000</td><td>0.000000</td><td>99.000000</td><td>27.000000</td><td>0.900000</td><td>135.000000</td><td>3.800000</td><td>10.300000</td><td>32.000000</td></tr>
<tr><td>50%</td><td>199.500000</td><td>55.000000</td><td>80.000000</td><td>1.020000</td><td>0.000000</td><td>0.000000</td><td>121.000000</td><td>42.000000</td><td>1.300000</td><td>138.000000</td><td>4.400000</td><td>12.650000</td><td>40.000000</td></tr>
<tr><td>75%</td><td>299.250000</td><td>64.500000</td><td>80.000000</td><td>1.020000</td><td>2.000000</td><td>0.000000</td><td>163.000000</td><td>66.000000</td><td>2.800000</td><td>142.000000</td><td>4.900000</td><td>15.000000</td><td>45.000000</td></tr>
<tr><td>max</td><td>399.000000</td><td>90.000000</td><td>180.000000</td><td>1.025000</td><td>5.000000</td><td>5.000000</td><td>490.000000</td><td>391.000000</td><td>76.000000</td><td>163.000000</td><td>47.000000</td><td>17.800000</td><td>54.000000</td></tr>
</table>
</td></tr>
<tr><td>Univariate Analysis</td><td></td></tr>
</table>

	 
Bivariate Analysis	

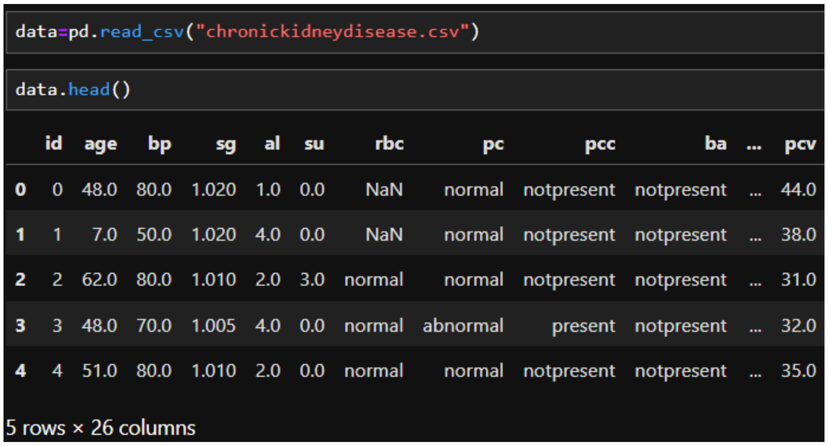
<table>
<tr>
<td></td>
<td>


</td>
</tr>
<tr>
<td>Multivariate Analysis</td>
<td>




</td>
</tr>
</table>

<table>
<tr><td>Outliers and Anomalies</td><td>-</td></tr>
<tr><td colspan="2"><b>Data Preprocessing Code Screenshots</b></td></tr>
<tr><td>Loading Data</td><td>



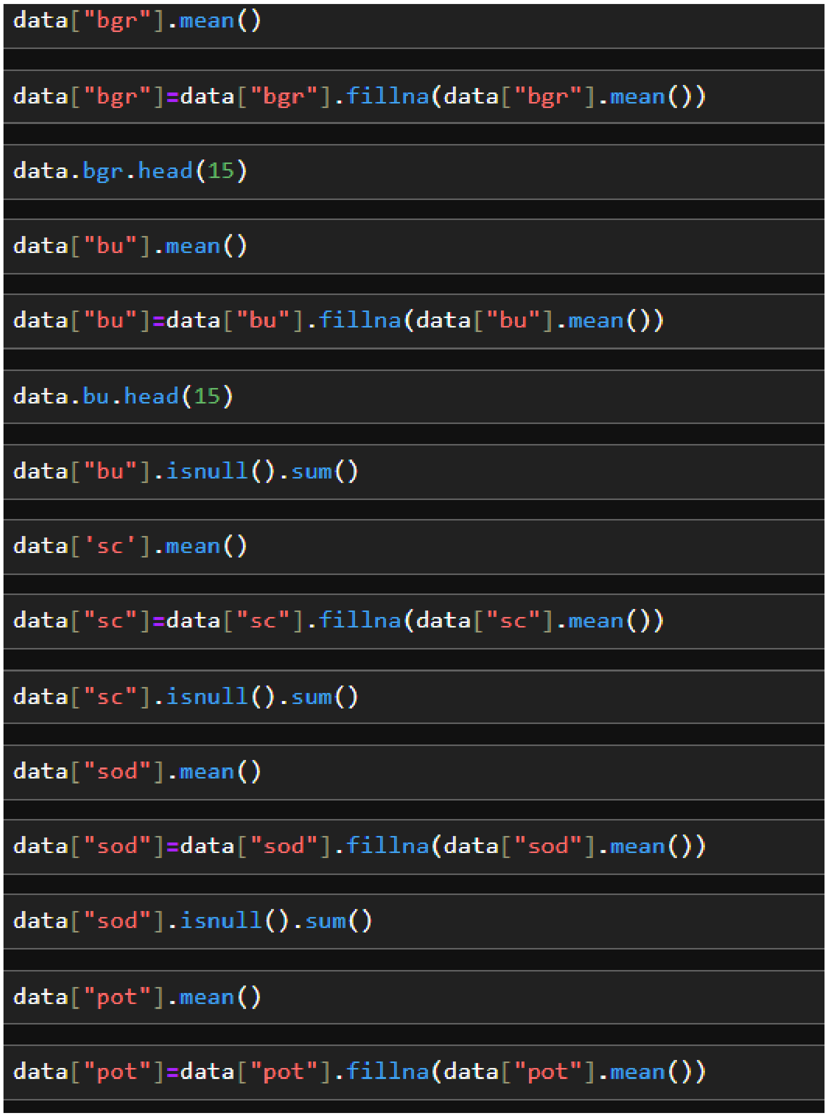
```
data=pd.read_csv("chronickidneydisease.csv")
```

```
data.head()
```

```
   id   age    bp     sg   al   su     rbc        pc         pcc          ba  ...  pcv
0   0  48.0  80.0  1.020  1.0  0.0     NaN    normal  notpresent  notpresent  ...  44.0
1   1   7.0  50.0  1.020  4.0  0.0     NaN    normal  notpresent  notpresent  ...  38.0
2   2  62.0  80.0  1.010  2.0  3.0  normal    normal  notpresent  notpresent  ...  31.0
3   3  48.0  70.0  1.005  4.0  0.0  normal  abnormal     present  notpresent  ...  32.0
4   4  51.0  80.0  1.010  2.0  0.0  normal    normal  notpresent  notpresent  ...  35.0

5 rows × 26 columns
```

</td></tr>
<tr><td>Handling Missing Data</td><td>



```
data["bgr"].mean()
```

```
data["bgr"]=data["bgr"].fillna(data["bgr"].mean())
```

```
data.bgr.head(15)
```

```
data["bu"].mean()
```

```
data["bu"]=data["bu"].fillna(data["bu"].mean())
```

```
data.bu.head(15)
```

```
data["bu"].isnull().sum()
```

```
data['sc'].mean()
```

```
data["sc"]=data["sc"].fillna(data["sc"].mean())
```

```
data["sc"].isnull().sum()
```

```
data["sod"].mean()
```

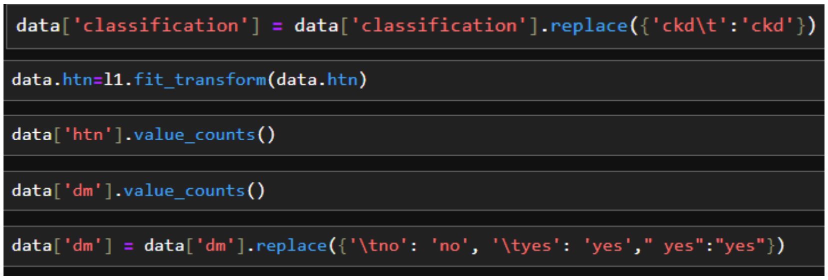
```
data["sod"]=data["sod"].fillna(data["sod"].mean())
```

```
data["sod"].isnull().sum()
```

```
data["pot"].mean()
```

```
data["pot"]=data["pot"].fillna(data["pot"].mean())
```

</td></tr>
</table>

Data Transformation	
	`data['classification'] = data['classification'].replace({'ckd\t':'ckd'})`
	`data.htn=l1.fit_transform(data.htn)`
	`data['htn'].value_counts()`
	`data['dm'].value_counts()`
	`data['dm'] = data['dm'].replace({'\tno': 'no', '\tyes': 'yes'," yes":"yes"})`

Feature Engineering	Attached the codes in final submission.
Save Processed Data	-