Data Collection and Preprocessing Phase
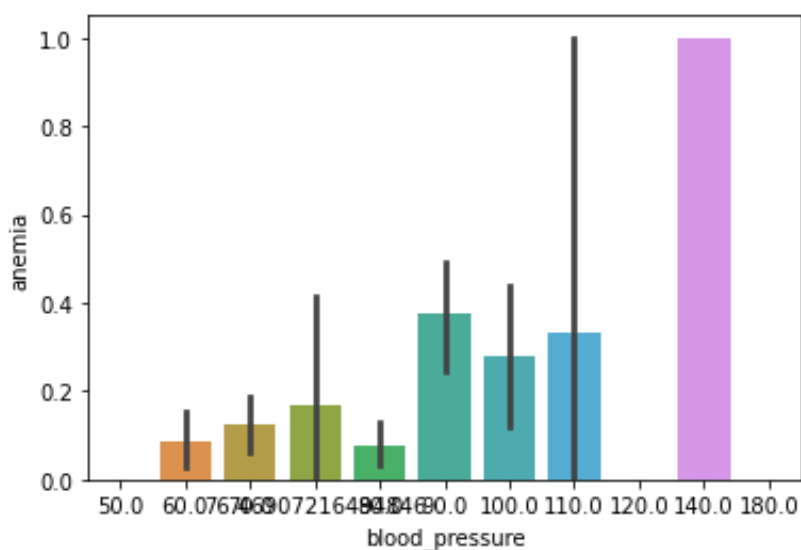
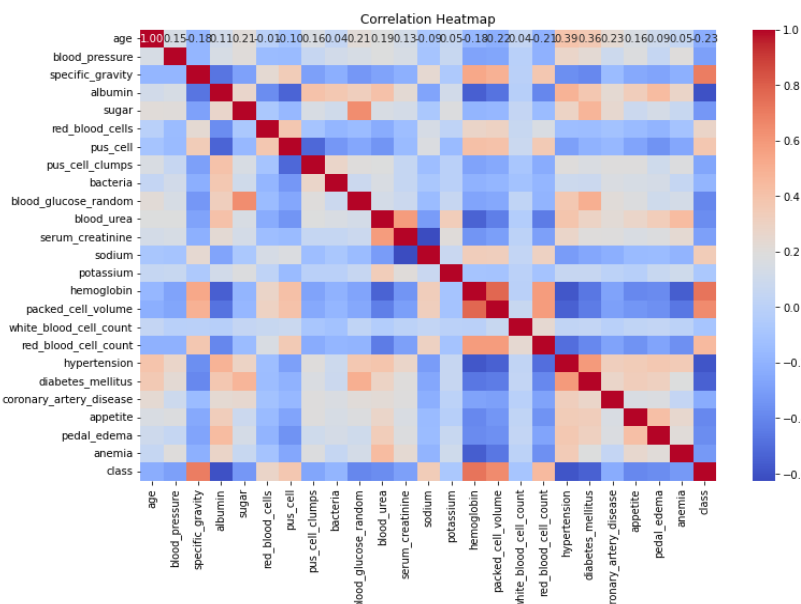| Date	15 July 2024
Team ID	739846
Project Title	Early Prediction Of Chronic Kidney Disease
Maximum Marks	6 Marks

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 399rows x 26columns
Univariate Analysis	

Bivariate Analysis	
Multivariate Analysis	
Outliers and Anomalies	-
Data Preprocessing Code Screenshots	
| Loading Data | ```
# Load dataset
filename = "C:\\Users\\soumy\\OneDrive\\Desktop\\mini p\\data.csv"
data = pd.read_csv(filename)
``` |

| Handling Missing Data | ```python
# Handle missing values
data = data.apply(lambda x: x.fillna(x.mean()) if x.dtype.kind in 'biufc' else x.fillna(x.mode().iloc[0]))
``` |
| --- | --- |
| Data Transformation | ```python
Clean categorical variables with incorrect values
data['class'] = data['class'].replace("ckd\t", "ckd")
data['coronary_artery_disease'] = data['coronary_artery_disease'].replace('\tno', 'no')
data['diabetes_mellitus'] = data['diabetes_mellitus'].replace(to_replace={'\tno': 'no', '\tyes': 'yes','yes':'yes'})
```

```python
# Encode categorical variables
cat_cols = data.select_dtypes(include=['object']).columns
label_encoders = {}
for col in cat_cols:
    le = LabelEncoder()
    data[col] = le.fit_transform(data[col])
    label_encoders[col] = le
``` |
| Feature Engineering | Attached code in final submission. |
| Save Processed Data | - |