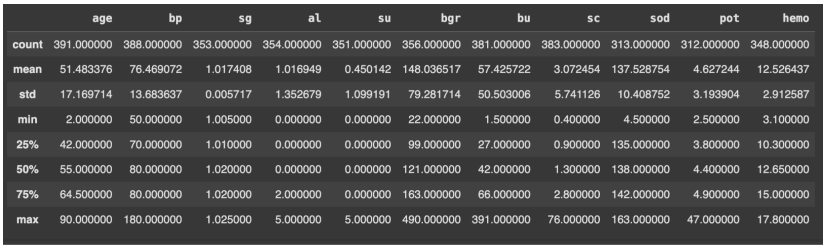
Data Collection and Preprocessing Phase

Date	18 June 2025
Team ID	SWTID1749641473
Project Title	Early Prediction for Chronic Kidney Disease Detection: A Progressive Approach to Health Management
Maximum Marks	6 Marks

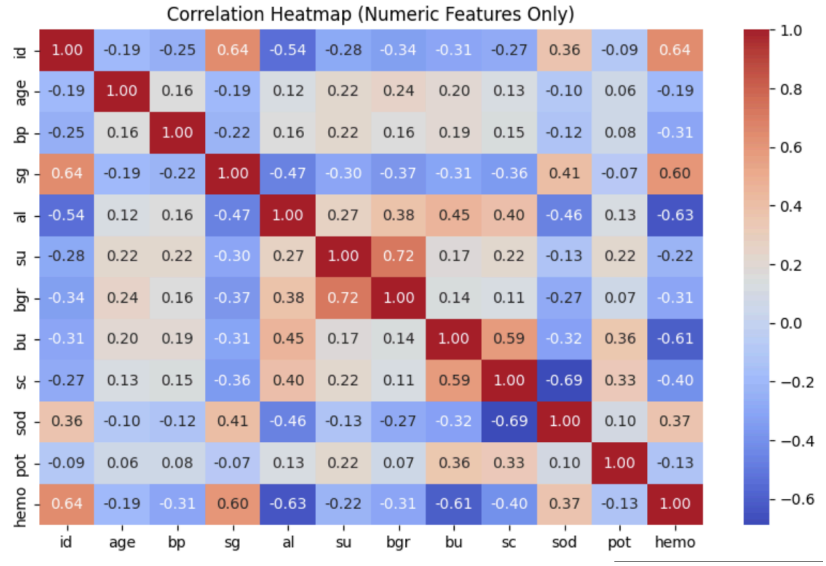
Data Exploration and Preprocessing Template

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: 400 rows × 26 columns Descriptive statistics: 

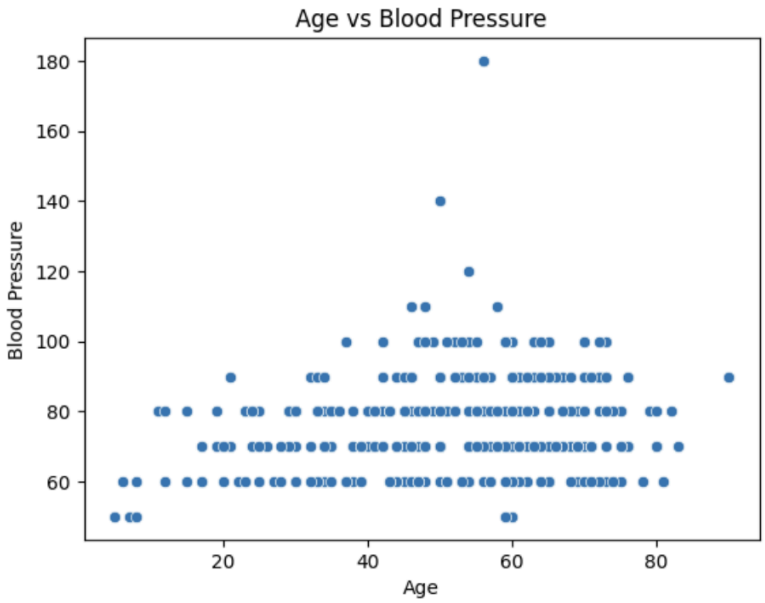
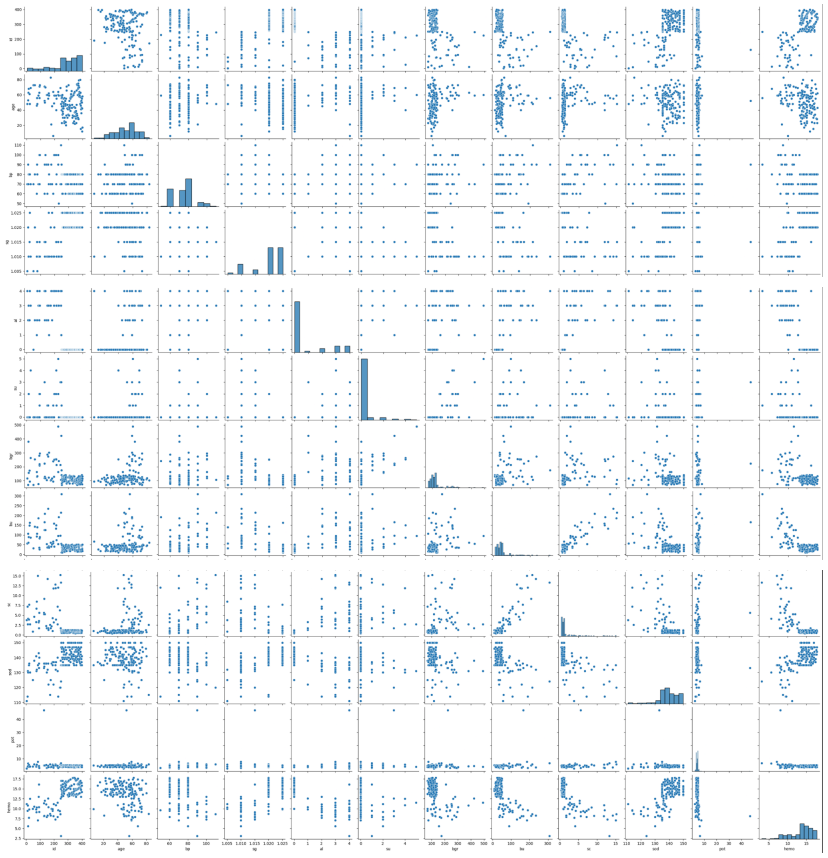
	age	bp	sg	al	su	bgr	bu	sc	sod	pot	hemo
count	391.000000	388.000000	353.000000	354.000000	351.000000	356.000000	381.000000	383.000000	313.000000	312.000000	348.000000
mean	51.483376	76.469072	1.017408	1.016949	0.450142	148.036517	57.425722	3.072454	137.528754	4.627244	12.526437
std	17.169714	13.683637	0.005717	1.352679	1.099191	79.281714	50.503006	5.741126	10.408752	3.193904	2.912587
min	2.000000	50.000000	1.005000	0.000000	0.000000	22.000000	1.500000	0.400000	4.500000	2.500000	3.100000
25%	42.000000	70.000000	1.010000	0.000000	0.000000	99.000000	27.000000	0.900000	135.000000	3.800000	10.300000
50%	55.000000	80.000000	1.020000	0.000000	0.000000	121.000000	42.000000	1.300000	138.000000	4.400000	12.650000
75%	64.500000	80.000000	1.020000	2.000000	0.000000	163.000000	66.000000	2.800000	142.000000	4.900000	15.000000
max	90.000000	180.000000	1.025000	5.000000	5.000000	490.000000	391.000000	76.000000	163.000000	47.000000	17.800000

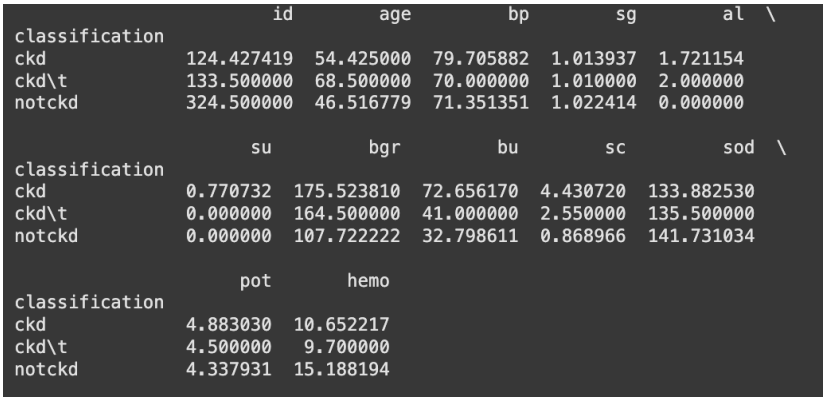
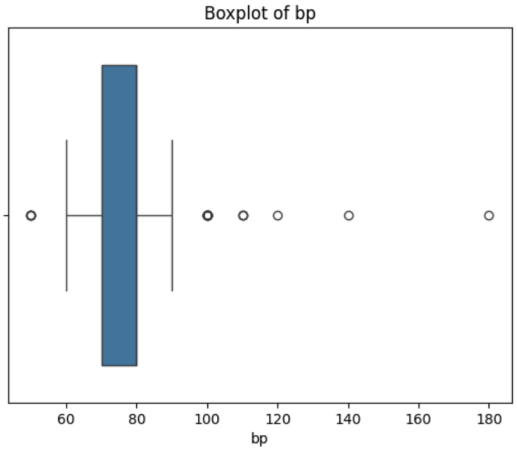
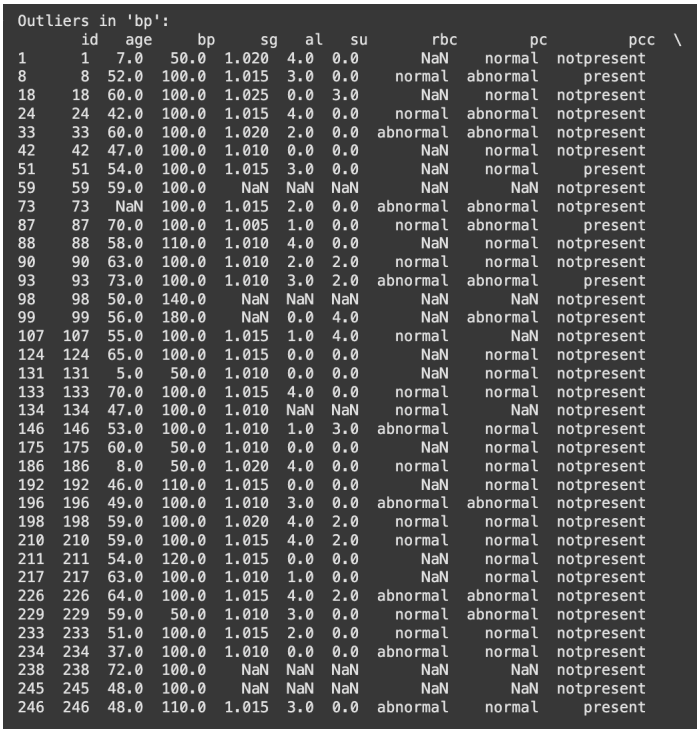
```
First 5 rows:
   id   age    bp     sg   al   su     rbc        pc         pcc          ba  \
0   0  48.0  80.0  1.020  1.0  0.0     NaN    normal  notpresent  notpresent
1   1   7.0  50.0  1.020  4.0  0.0     NaN    normal  notpresent  notpresent
2   2  62.0  80.0  1.010  2.0  3.0  normal    normal  notpresent  notpresent
3   3  48.0  70.0  1.005  4.0  0.0  normal  abnormal     present  notpresent
4   4  51.0  80.0  1.010  2.0  0.0  normal    normal  notpresent  notpresent

   ...  pcv    wc   rc  htn   dm  cad appet   pe  ane classification
0  ...   44  7800  5.2  yes  yes   no  good   no   no            ckd
1  ...   38  6000  NaN   no   no   no  good   no   no            ckd
2  ...   31  7500  NaN   no  yes   no  poor   no  yes            ckd
3  ...   32  6700  3.9  yes   no   no  poor  yes  yes            ckd
4  ...   35  7300  4.6   no   no   no  good   no   no            ckd

[5 rows x 26 columns]
```

Univariate Analysis	Mean age: 51.48337595907928 Median age: 55.0 Mode age: 60.0 Standard Deviation: 17.16971408926224
Bivariate Analysis	Correlation matrix: (see table below) 

Correlation matrix:

	id	age	bp	sg	al	su	bgr	bu	sc	sod	pot	hemo
id	1.000000	-0.185308	-0.245744	0.642156	-0.541993	-0.283416	-0.338673	-0.307175	-0.268683	0.364251	-0.092347	0.640298
age	-0.185308	1.000000	0.159480	-0.191096	0.122091	0.220866	0.244992	0.196985	0.132531	-0.100046	0.058377	-0.192928
bp	-0.245744	0.159480	1.000000	-0.218836	0.160689	0.222576	0.160193	0.188517	0.146222	-0.116422	0.075151	-0.306540
sg	0.642156	-0.191096	-0.218836	1.000000	-0.469760	-0.296234	-0.374710	-0.314295	-0.361473	0.412190	-0.072787	0.602582
al	-0.541993	0.122091	0.160689	-0.469760	1.000000	0.269305	0.379464	0.453528	0.399198	-0.459896	0.129038	-0.634632
su	-0.283416	0.220866	0.222576	-0.296234	0.269305	1.000000	0.717827	0.168583	0.223244	-0.131776	0.219450	-0.224775
bgr	-0.338673	0.244992	0.160193	-0.374710	0.379464	0.717827	1.000000	0.143322	0.114875	-0.267848	0.066966	-0.306189
bu	-0.307175	0.196985	0.188517	-0.314295	0.453528	0.168583	0.143322	1.000000	0.586368	-0.323054	0.357049	-0.610360
sc	-0.268683	0.132531	0.146222	-0.361473	0.399198	0.223244	0.114875	0.586368	1.000000	-0.690158	0.326107	-0.401670
sod	0.364251	-0.100046	-0.116422	0.412190	-0.459896	-0.131776	-0.267848	-0.323054	-0.690158	1.000000	0.097887	0.365183
pot	-0.092347	0.058377	0.075151	-0.072787	0.129038	0.219450	0.066966	0.357049	0.326107	0.097887	1.000000	-0.133746
hemo	0.640298	-0.192928	-0.306540	0.602582	-0.634632	-0.224775	-0.306189	-0.610360	-0.401670	0.365183	-0.133746	1.000000

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

	
Outliers and Anomalies	 

classification	id	age	bp	sg	al
ckd	124.427419	54.425000	79.705882	1.013937	1.721154
ckd\t	133.500000	68.500000	70.000000	1.010000	2.000000
notckd	324.500000	46.516779	71.351351	1.022414	0.000000

classification	su	bgr	bu	sc	sod
ckd	0.770732	175.523810	72.656170	4.430720	133.882530
ckd\t	0.000000	164.500000	41.000000	2.550000	135.500000
notckd	0.000000	107.722222	32.798611	0.868966	141.731034

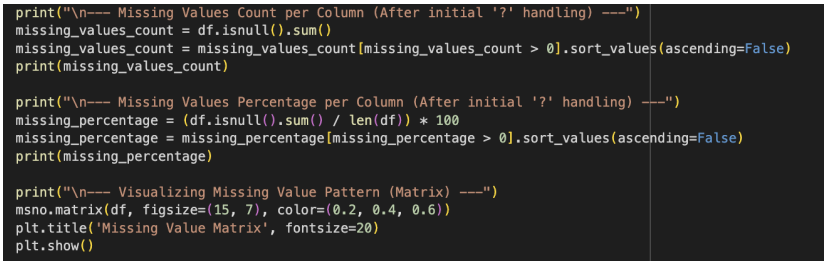
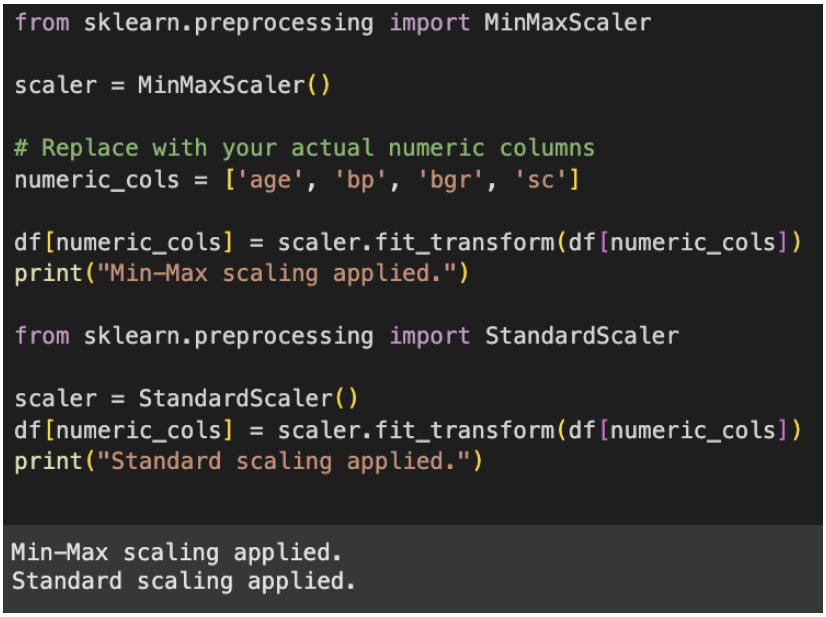
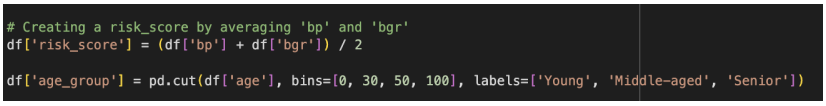
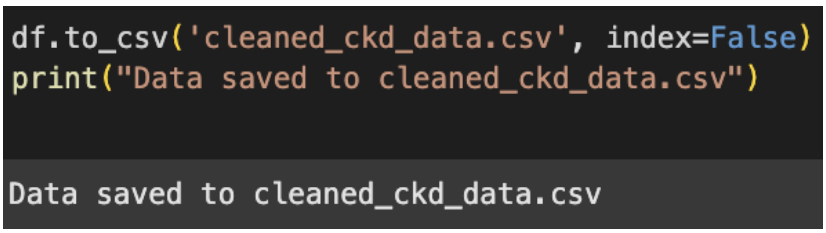
classification	pot	hemo
ckd	4.883030	10.652217
ckd\t	4.500000	9.700000
notckd	4.337931	15.188194

Outliers in 'bp':

	id	age	bp	sg	al	su	rbc	pc	pcc
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent
8	8	52.0	100.0	1.015	3.0	0.0	normal	abnormal	present
18	18	60.0	100.0	1.025	0.0	3.0	NaN	normal	notpresent
24	24	42.0	100.0	1.015	4.0	0.0	normal	abnormal	notpresent
33	33	60.0	100.0	1.020	2.0	0.0	abnormal	abnormal	notpresent
42	42	47.0	100.0	1.010	0.0	0.0	NaN	normal	notpresent
51	51	54.0	100.0	1.015	3.0	0.0	NaN	normal	present
59	59	59.0	100.0	NaN	NaN	NaN	NaN	NaN	notpresent
73	73	NaN	100.0	1.015	2.0	0.0	abnormal	abnormal	notpresent
87	87	70.0	100.0	1.005	1.0	0.0	normal	abnormal	present
88	88	58.0	110.0	1.010	4.0	0.0	NaN	normal	notpresent
90	90	63.0	100.0	1.010	2.0	2.0	normal	normal	notpresent
93	93	73.0	100.0	1.010	3.0	2.0	abnormal	abnormal	present
98	98	50.0	140.0	NaN	NaN	NaN	NaN	NaN	notpresent
99	99	56.0	180.0	NaN	0.0	4.0	NaN	abnormal	notpresent
107	107	55.0	100.0	1.015	1.0	4.0	normal	NaN	notpresent
124	124	65.0	100.0	1.015	0.0	0.0	NaN	normal	notpresent
131	131	5.0	50.0	1.010	0.0	0.0	NaN	normal	notpresent
133	133	70.0	100.0	1.015	4.0	0.0	normal	normal	notpresent
134	134	47.0	100.0	1.010	NaN	NaN	normal	NaN	notpresent
146	146	53.0	100.0	1.010	1.0	3.0	abnormal	normal	notpresent
175	175	60.0	50.0	1.010	0.0	0.0	NaN	normal	notpresent
186	186	8.0	50.0	1.020	4.0	0.0	normal	normal	notpresent
192	192	46.0	110.0	1.015	0.0	0.0	NaN	normal	notpresent
196	196	49.0	100.0	1.010	3.0	0.0	abnormal	abnormal	notpresent
198	198	59.0	100.0	1.020	4.0	2.0	normal	normal	notpresent
210	210	59.0	100.0	1.015	4.0	2.0	normal	normal	notpresent
211	211	54.0	120.0	1.015	0.0	0.0	NaN	normal	notpresent
217	217	63.0	100.0	1.010	1.0	0.0	NaN	normal	notpresent
226	226	64.0	100.0	1.015	4.0	2.0	abnormal	abnormal	notpresent
229	229	59.0	50.0	1.010	3.0	0.0	normal	abnormal	notpresent
233	233	51.0	100.0	1.015	2.0	0.0	normal	normal	notpresent
234	234	37.0	100.0	1.010	0.0	0.0	abnormal	normal	notpresent
238	238	72.0	100.0	NaN	NaN	NaN	NaN	NaN	notpresent
245	245	48.0	100.0	NaN	NaN	NaN	NaN	NaN	notpresent
246	246	48.0	110.0	1.015	3.0	0.0	abnormal	normal	present

```
              ba  ...  pcv      wc   rc  htn     dm  cad appet   pe  ane  \
1    notpresent  ...   38    6000  NaN   no     no   no  good   no   no
8    notpresent  ...   33    9600  4.0  yes    yes   no  good   no  yes
18   notpresent  ...   37   11400  4.3  yes    yes  yes  good   no   no
24      present  ...   39    8300  4.6  yes     no   no  poor   no   no
33   notpresent  ...   29     NaN  NaN  yes     no   no  poor   no   no
42   notpresent  ...   33    9200  4.5  yes     no   no  good   no  yes
51   notpresent  ...   33     NaN  NaN  yes    yes   no  poor  yes   no
59   notpresent  ...  NaN     NaN  NaN  yes    yes   no  good   no  yes
73   notpresent  ...   14    6300  NaN  yes     no   no  good  yes  yes
87   notpresent  ...   32    5800    5  yes    yes   no  poor   no   no
88   notpresent  ...  NaN   13200  4.7  yes  \tyes   no  good   no   no
90      present  ...   40    9800  4.2  yes     no  yes  good   no   no
93   notpresent  ...   30    7000  3.2  yes    yes  yes  poor   no   no
98   notpresent  ...   18    5800  2.3  yes    yes   no  poor   no  yes
99   notpresent  ...   32   10400  4.2  yes    yes   no  poor  yes   no
107  notpresent  ...   34   13600  4.4  yes    yes   no  good   no   no
124  notpresent  ...   28    5500  3.6  yes     no   no  good   no   no
131  notpresent  ...   36   12400  NaN   no     no   no  good   no   no
133  notpresent  ...   37  \t8400  8.0  yes     no   no  good   no   no
134  notpresent  ...   33   10200  3.8   no    yes   no  good   no   no
146  notpresent  ...  NaN     NaN  NaN   no    yes   no  good   no   no
175  notpresent  ...  NaN    4200  3.4  yes     no   no  good   no   no
186  notpresent  ...  NaN     NaN  NaN   no     no   no  good  yes   no
192  notpresent  ...  NaN     NaN  NaN   no     no   no  good   no   no
196  notpresent  ...   24    9600  3.5  yes    yes   no  poor  yes  yes
198  notpresent  ...   30   26400  3.9  yes    yes   no  poor  yes   no
210  notpresent  ...   20    9800  3.9  yes    yes  yes  good   no  yes
211  notpresent  ...  NaN     NaN  NaN   no     no   no  good   no   no
217  notpresent  ...   36   10500  4.3   no    yes   no  good   no   no
226     present  ...   26    7500  3.4  yes    yes   no  good  yes   no
229  notpresent  ...   31   15700  3.8   no    yes   no  good  yes   no
233     present  ...  NaN     NaN  NaN   no     no   no  poor   no   no
234  notpresent  ...   44    4100  5.2  yes     no   no  good   no   no
238  notpresent  ...   28     NaN  NaN  yes    yes   no  good   no  yes
245  notpresent  ...   19    7200  2.6  yes     no  yes  poor   no   no
246  notpresent  ...   26    5000  2.5  yes     no  yes  good   no  yes
```

```
    classification
1              ckd
8              ckd
18             ckd
24             ckd
33             ckd
42             ckd
51             ckd
59             ckd
73             ckd
87             ckd
88             ckd
90             ckd
93             ckd
98             ckd
99             ckd
107            ckd
124            ckd
131            ckd
133            ckd
134            ckd
146            ckd
175            ckd
186            ckd
192            ckd
196            ckd
198            ckd
210            ckd
211            ckd
217            ckd
226            ckd
229            ckd
233            ckd
234            ckd
238            ckd
245            ckd
246            ckd

[36 rows x 26 columns]

Original size: 400 rows
After removing outliers in 'bp': 352 rows
```

Data Preprocessing Code Screenshots	
Loading Data	
Handling Missing Data	
Data Transformation	
Feature Engineering	
Save Processed Data	

```
dataset_path = 'chronickidneydisease.csv'
try:
    df = pd.read_csv(dataset_path)
    print(f"\nDataset '{dataset_path}' loaded successfully.")
    print(f"Initial dataset shape: {df.shape}")
except FileNotFoundError:
    print(f"ERROR: Dataset '{dataset_path}' not found.")
    print("Please ensure you have downloaded 'chronickidneydisease.csv' and placed it in the 'CKD_Prediction_App/dataset/' folder.")
    exit()
except Exception as e:
    print(f"An error occurred while reading the dataset: {e}")
    exit()
```

```
print("\n--- Missing Values Count per Column (After initial '?' handling) ---")
missing_values_count = df.isnull().sum()
missing_values_count = missing_values_count[missing_values_count > 0].sort_values(ascending=False)
print(missing_values_count)

print("\n--- Missing Values Percentage per Column (After initial '?' handling) ---")
missing_percentage = (df.isnull().sum() / len(df)) * 100
missing_percentage = missing_percentage[missing_percentage > 0].sort_values(ascending=False)
print(missing_percentage)

print("\n--- Visualizing Missing Value Pattern (Matrix) ---")
msno.matrix(df, figsize=(15, 7), color=(0.2, 0.4, 0.6))
plt.title('Missing Value Matrix', fontsize=20)
plt.show()
```

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

# Replace with your actual numeric columns
numeric_cols = ['age', 'bp', 'bgr', 'sc']

df[numeric_cols] = scaler.fit_transform(df[numeric_cols])
print("Min-Max scaling applied.")

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
df[numeric_cols] = scaler.fit_transform(df[numeric_cols])
print("Standard scaling applied.")
```

```
Min-Max scaling applied.
Standard scaling applied.
```

```
# Creating a risk_score by averaging 'bp' and 'bgr'
df['risk_score'] = (df['bp'] + df['bgr']) / 2

df['age_group'] = pd.cut(df['age'], bins=[0, 30, 50, 100], labels=['Young', 'Middle-aged', 'Senior'])
```

```
df.to_csv('cleaned_ckd_data.csv', index=False)
print("Data saved to cleaned_ckd_data.csv")
```

```
Data saved to cleaned_ckd_data.csv
```