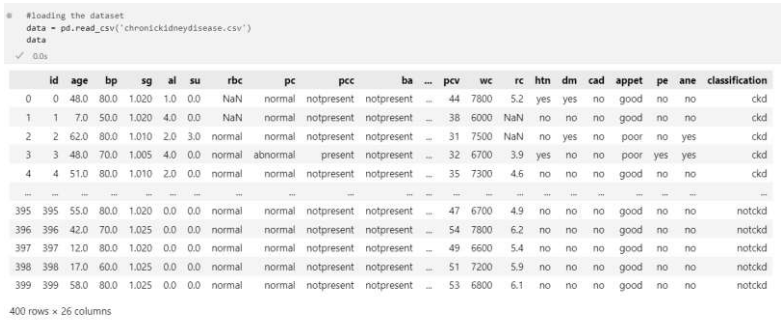
Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	xxxxxx
Project Title	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.

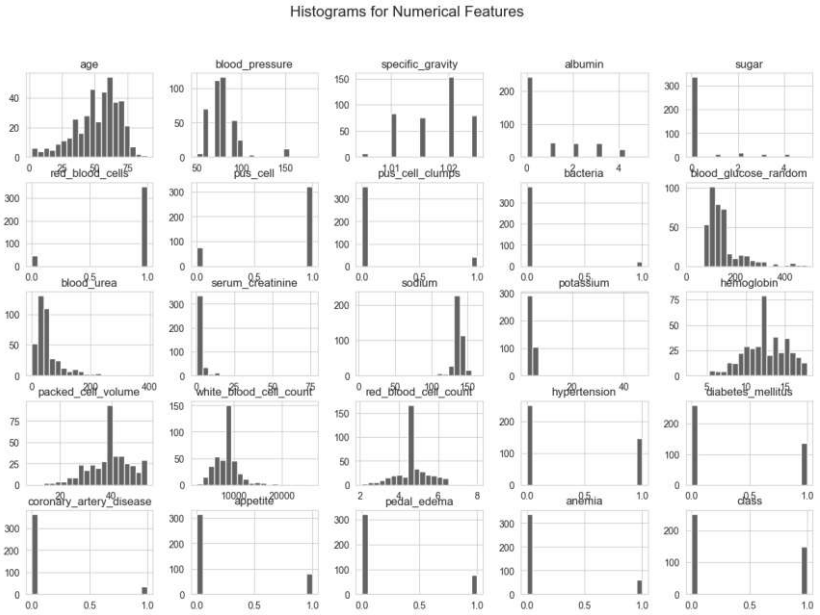
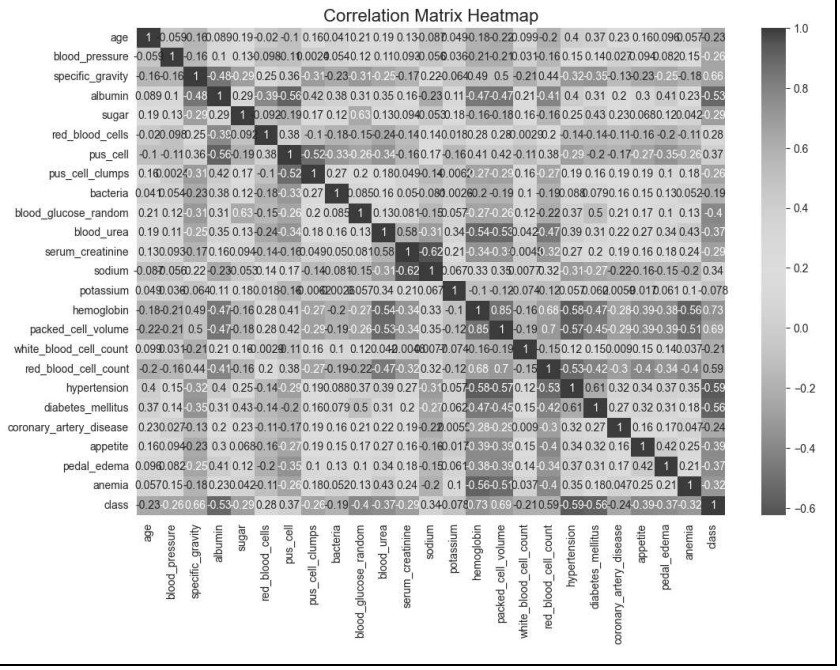
Section	Description
Data Overview	

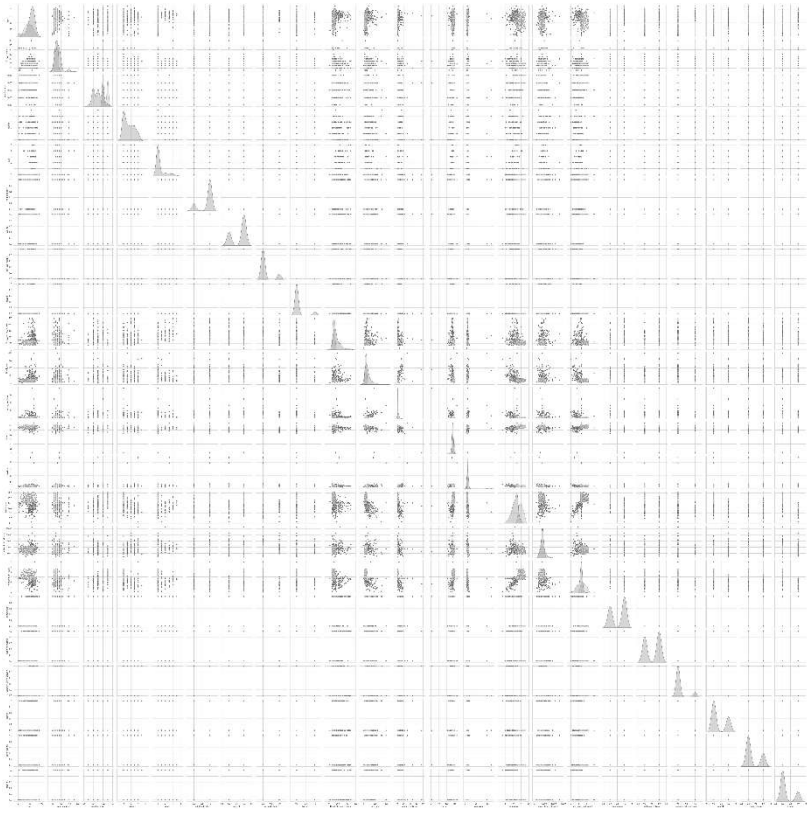
```
#loading the dataset
data = pd.read_csv('chronickidneydisease.csv')
data
```

0.0s

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	...	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	...	38	6000	NaN	no	no	no	good	no	no	ckd
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	...	31	7500	NaN	no	yes	no	poor	no	yes	ckd
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	...	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	...	35	7300	4.6	no	no	no	good	no	no	ckd
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
395	395	55.0	80.0	1.020	0.0	0.0	normal	normal	notpresent	notpresent	...	47	6700	4.9	no	no	no	good	no	no	notckd
396	396	42.0	70.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	...	54	7800	6.2	no	no	no	good	no	no	notckd
397	397	12.0	80.0	1.020	0.0	0.0	normal	normal	notpresent	notpresent	...	49	6600	5.4	no	no	no	good	no	no	notckd
398	398	17.0	60.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	...	51	7200	5.9	no	no	no	good	no	no	notckd
399	399	58.0	80.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	...	53	6800	6.1	no	no	no	good	no	no	notckd

400 rows × 26 columns

Univariate Analysis	
Bivariate Analysis	

Multivariate Analysis	
Outliers and Anomalies	-

Data Preprocessing Code Screenshots

Loading Data

```
#loading the dataset
data = pd.read_csv('chronickidneydisease.csv')
data
```

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	...	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	...	38	6000	NaN	no	no	no	good	no	no	ckd
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	...	31	7500	NaN	no	yes	no	poor	no	yes	ckd
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	...	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	...	35	7300	4.6	no	no	no	good	no	no	ckd
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
395	395	55.0	80.0	1.020	0.0	0.0	normal	normal	notpresent	notpresent	...	47	6700	4.9	no	no	no	good	no	no	notckd
396	396	42.0	70.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	...	54	7800	6.2	no	no	no	good	no	no	notckd
397	397	12.0	80.0	1.020	0.0	0.0	normal	normal	notpresent	notpresent	...	49	6600	5.4	no	no	no	good	no	no	notckd
398	398	17.0	60.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	...	51	7200	5.9	no	no	no	good	no	no	notckd
399	399	58.0	80.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	...	53	6800	6.1	no	no	no	good	no	no	notckd

400 rows × 26 columns

Handling Missing Data

```
data['blood_glucose_random'].fillna(data['blood_glucose_random'].mean(),inplace=True)
data['blood_pressure'].fillna(data['blood_glucose_random'].mean(),inplace=True)
data['blood_urea'].fillna(data['blood_urea'].mean(),inplace=True)
data['hemoglobin'].fillna(data['hemoglobin'].mean(),inplace=True)
data['packed_cell_volume'].fillna(data['packed_cell_volume'].mean(),inplace=True)
data['potassium'].fillna(data['potassium'].mean(),inplace=True)
data['red_blood_cell_count'].fillna(data['red_blood_cell_count'].mean(),inplace=True)
data['serum_creatinine'].fillna(data['serum_creatinine'].mean(),inplace=True)
data['sodium'].fillna(data['sodium'].mean(),inplace=True)
data['white_blood_cell_count'].fillna(data['white_blood_cell_count'].mean(),inplace=True)
```

✓ 0.0s Python

```
data['age'].fillna(data['age'].mode()[0],inplace=True)
data['hypertension'].fillna(data['hypertension'].mode()[0],inplace=True)
data['pus_cell_clumps'].fillna(data['pus_cell_clumps'].mode()[0],inplace=True)
data['appetite'].fillna(data['appetite'].mode()[0],inplace=True)
data['albumin'].fillna(data['albumin'].mode()[0],inplace=True)
data['pus_cell'].fillna(data['pus_cell'].mode()[0],inplace=True)
data['red_blood_cells'].fillna(data['red_blood_cells'].mode()[0],inplace=True)
data['bacteria'].fillna(data['bacteria'].mode()[0],inplace=True)
data['anemia'].fillna(data['anemia'].mode()[0],inplace=True)
data['sugar'].fillna(data['sugar'].mode()[0],inplace=True)
data['diabetes_mellitus'].fillna(data['diabetes_mellitus'].mode()[0],inplace=True)
data['pedal_edema'].fillna(data['pedal_edema'].mode()[0],inplace=True)
data[ 'specific_gravity'].fillna(data[ 'specific_gravity'].mode()[0],inplace=True)
```

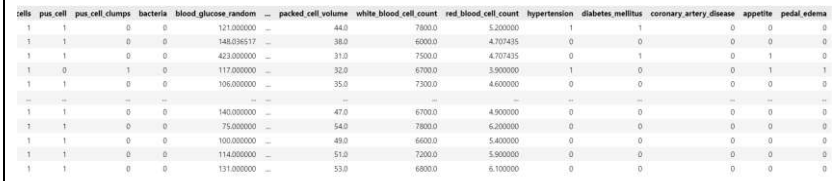
✓ 0.0s Python

Data Transformation

```
for i in catcols: #looping through all categorical columns
    print("Label Encoding of:", i)
    LEi = LabelEncoder() # creating an object of Label Encoder
    print(c(data[i]))
    data[i] = LEi.fit_transform(data[i])
    print(c(data[i]))
    print("*"*100)
```

[42] ✓ 0.0s

```
Label Encoding of: red_blood_cells
Counter({'normal': 351, 'abnormal': 47})
Counter({1: 351, 0: 47})
****************************************************************************************************
Label Encoding of: appetite
Counter({'good': 316, 'poor': 82})
Counter({0: 316, 1: 82})
****************************************************************************************************
Label Encoding of: class
Counter({'ckd': 250, 'notckd': 148})
Counter({0: 250, 1: 148})
****************************************************************************************************
Label Encoding of: pedal_edema
Counter({'no': 322, 'yes': 76})
Counter({0: 322, 1: 76})
****************************************************************************************************
Label Encoding of: anemia
Counter({'no': 338, 'yes': 60})
Counter({0: 338, 1: 60})
****************************************************************************************************
Label Encoding of: pus_cell
Counter({'normal': 322, 'abnormal': 76})
Counter({1: 322, 0: 76})
****************************************************************************************************
Label Encoding of: coronary_artery_disease
...
Label Encoding of: bacteria
Counter({'notpresent': 376, 'present': 22})
Counter({0: 376, 1: 22})
****************************************************************************************************
```

Feature Engineering	
Save Processed Data	

:ells	pus_cell	pus_cell_clumps	bacteria	blood_glucose_random	...	packed_cell_volume	white_blood_cell_count	red_blood_cell_count	hypertension	diabetes_mellitus	coronary_artery_disease	appetite	pedal_edema
1	1	0	0	121.000000	...	44.0	7800.0	5.200000	1	1	0	0	0
1	1	0	0	148.036517	...	38.0	6000.0	4.707435	0	0	0	0	0
1	1	0	0	423.000000	...	31.0	7500.0	4.707435	0	1	0	1	0
1	0	1	0	117.000000	...	32.0	6700.0	3.900000	1	0	0	1	1
1	1	0	0	106.000000	...	35.0	7300.0	4.600000	0	0	0	0	0
..	..	..	..	..	..	..	..	..	..	..	..	..	..
1	1	0	0	140.000000	...	47.0	6700.0	4.900000	0	0	0	0	0
1	1	0	0	75.000000	...	54.0	7800.0	6.200000	0	0	0	0	0
1	1	0	0	100.000000	...	49.0	6600.0	5.400000	0	0	0	0	0
1	1	0	0	114.000000	...	51.0	7200.0	5.900000	0	0	0	0	0
1	1	0	0	131.000000	...	53.0	6800.0	6.100000	0	0	0	0	0