

```
from google.colab import drive
drive.mount('/content/drive')
```

Mounted at /content/drive

```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
```

```
df = pd.read_csv('///content/drive/MyDrive/project 2/kidney_disease.csv')
```

```
df.head()
```

|   | id | age  | bp   | sg    | al  | su  | rbc    | pc       | pcc        | ba         | ... | p |
|---|----|------|------|-------|-----|-----|--------|----------|------------|------------|-----|---|
| 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 | ... |   |

5 rows × 26 columns

```
df.dtypes
```

|                       | 0       |
|-----------------------|---------|
| <b>id</b>             | int64   |
| <b>age</b>            | float64 |
| <b>bp</b>             | float64 |
| <b>sg</b>             | float64 |
| <b>al</b>             | float64 |
| <b>su</b>             | float64 |
| <b>rbc</b>            | object  |
| <b>pc</b>             | object  |
| <b>pcc</b>            | object  |
| <b>ba</b>             | object  |
| <b>bgr</b>            | float64 |
| <b>bu</b>             | float64 |
| <b>sc</b>             | float64 |
| <b>sod</b>            | float64 |
| <b>pot</b>            | float64 |
| <b>hemo</b>           | float64 |
| <b>pcv</b>            | object  |
| <b>wc</b>             | object  |
| <b>rc</b>             | object  |
| <b>htn</b>            | object  |
| <b>dm</b>             | object  |
| <b>cad</b>            | object  |
| <b>appet</b>          | object  |
| <b>pe</b>             | object  |
| <b>ane</b>            | object  |
| <b>classification</b> | object  |

**dtype:** object

df.columns

```
Index(['id', 'age', 'bp', 'sg', 'al', 'su', 'rbc', 'pc', 'pcc', 'ba', 'bgr',
      'bu', 'sc', 'sod', 'pot', 'hemo', 'pcv', 'wc', 'rc', 'htn', 'dm',
      'cad',
      'appet', 'pe', 'ane', 'classification'],
      dtype='object')
```

```
df.isnull().sum()
```

|                       | <b>0</b> |
|-----------------------|----------|
| <b>id</b>             | 0        |
| <b>age</b>            | 9        |
| <b>bp</b>             | 12       |
| <b>sg</b>             | 47       |
| <b>al</b>             | 46       |
| <b>su</b>             | 49       |
| <b>rbc</b>            | 152      |
| <b>pc</b>             | 65       |
| <b>pcc</b>            | 4        |
| <b>ba</b>             | 4        |
| <b>bgr</b>            | 44       |
| <b>bu</b>             | 19       |
| <b>sc</b>             | 17       |
| <b>sod</b>            | 87       |
| <b>pot</b>            | 88       |
| <b>hemo</b>           | 52       |
| <b>pcv</b>            | 70       |
| <b>wc</b>             | 105      |
| <b>rc</b>             | 130      |
| <b>htn</b>            | 2        |
| <b>dm</b>             | 2        |
| <b>cad</b>            | 2        |
| <b>appet</b>          | 1        |
| <b>pe</b>             | 1        |
| <b>ane</b>            | 1        |
| <b>classification</b> | 0        |

**dtype:** int64

```
missing_percentage = df.isnull().mean() * 100  
threshold = 30
```

```

from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
for column in df.select_dtypes(include=['object']).columns:
    df[column] = label_encoder.fit_transform(df[column])
df.head()

```

|          | id | age  | bp   | sg    | al  | su  | rbc | pc | pcc | ba | ... | pcv | wc | rc | htn | dm |
|----------|----|------|------|-------|-----|-----|-----|----|-----|----|-----|-----|----|----|-----|----|
| <b>0</b> | 0  | 48.0 | 80.0 | 1.020 | 1.0 | 0.0 | 2   | 1  | 0   | 0  | ... | 32  | 72 | 34 | 1   | 4  |
| <b>1</b> | 1  | 7.0  | 50.0 | 1.020 | 4.0 | 0.0 | 2   | 1  | 0   | 0  | ... | 26  | 56 | 49 | 0   | 3  |
| <b>2</b> | 2  | 62.0 | 80.0 | 1.010 | 2.0 | 3.0 | 1   | 1  | 0   | 0  | ... | 19  | 70 | 49 | 0   | 4  |
| <b>3</b> | 3  | 48.0 | 70.0 | 1.005 | 4.0 | 0.0 | 1   | 0  | 1   | 0  | ... | 20  | 62 | 19 | 1   | 3  |
| <b>4</b> | 4  | 51.0 | 80.0 | 1.010 | 2.0 | 0.0 | 1   | 1  | 0   | 0  | ... | 23  | 68 | 27 | 0   | 3  |

5 rows × 26 columns

```

for column in df.select_dtypes(include=['float64']).columns:
    mean_value = df[column].mean()
    df[column] = df[column].fillna(mean_value)

```

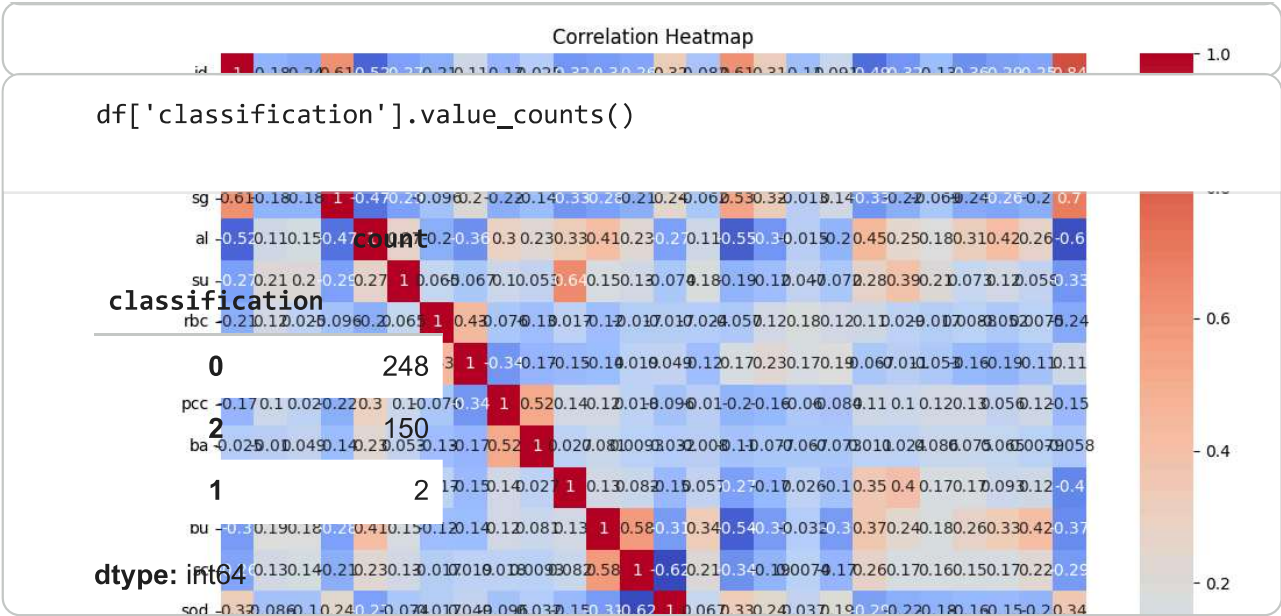
```
df.isnull().sum()
```

|                       | 0 |
|-----------------------|---|
| <b>id</b>             | 0 |
| <b>age</b>            | 0 |
| <b>bp</b>             | 0 |
| <b>sg</b>             | 0 |
| <b>al</b>             | 0 |
| <b>su</b>             | 0 |
| <b>rbc</b>            | 0 |
| <b>pc</b>             | 0 |
| <b>pcc</b>            | 0 |
| <b>ba</b>             | 0 |
| <b>bgr</b>            | 0 |
| <b>bu</b>             | 0 |
| <b>sc</b>             | 0 |
| <b>sod</b>            | 0 |
| <b>pot</b>            | 0 |
| <b>hemo</b>           | 0 |
| <b>pcv</b>            | 0 |
| <b>wc</b>             | 0 |
| <b>rc</b>             | 0 |
| <b>htn</b>            | 0 |
| <b>dm</b>             | 0 |
| <b>cad</b>            | 0 |
| <b>appet</b>          | 0 |
| <b>pe</b>             | 0 |
| <b>ane</b>            | 0 |
| <b>classification</b> | 0 |

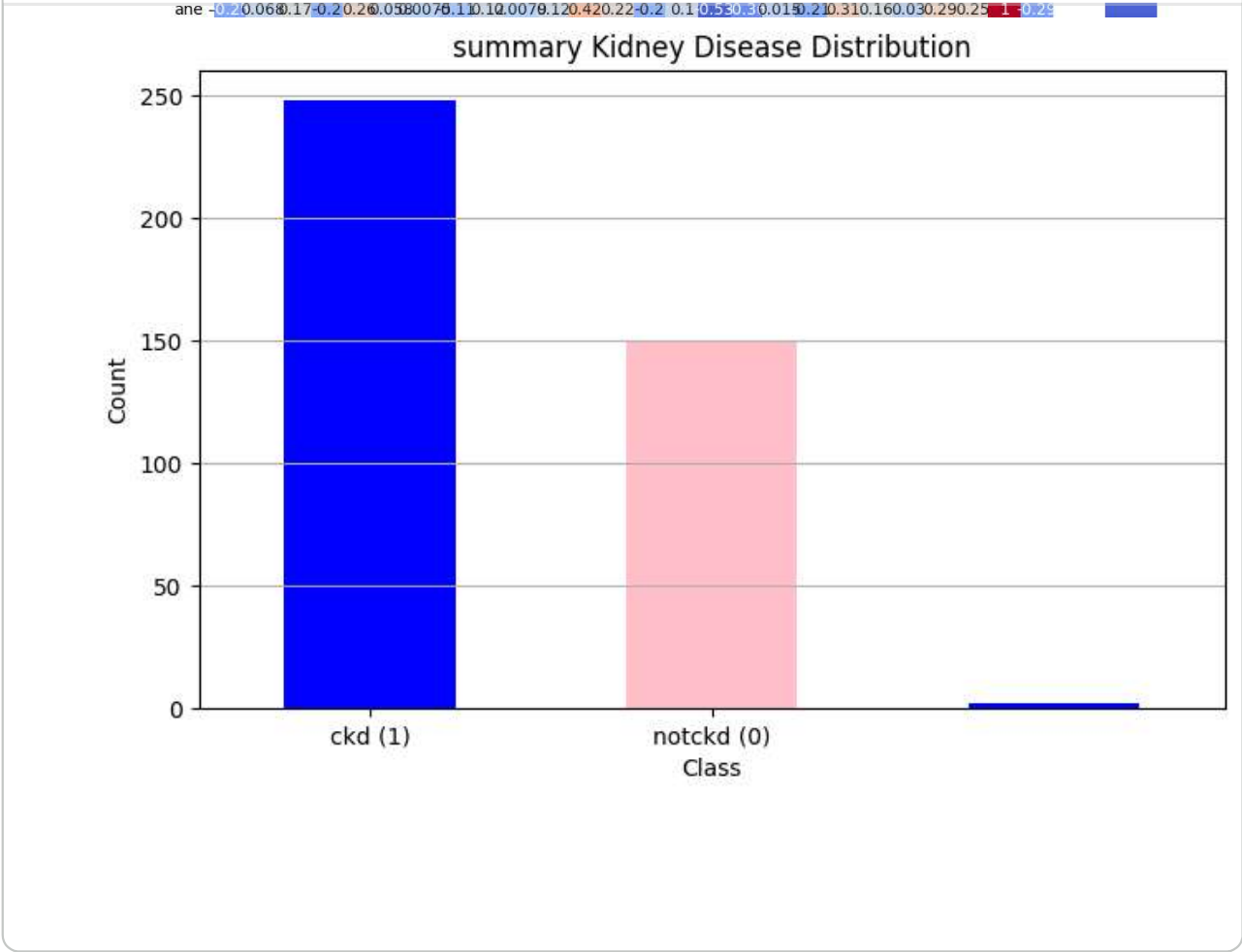
**dtype:** int64

```
plt.figure(figsize=(12, 12))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```

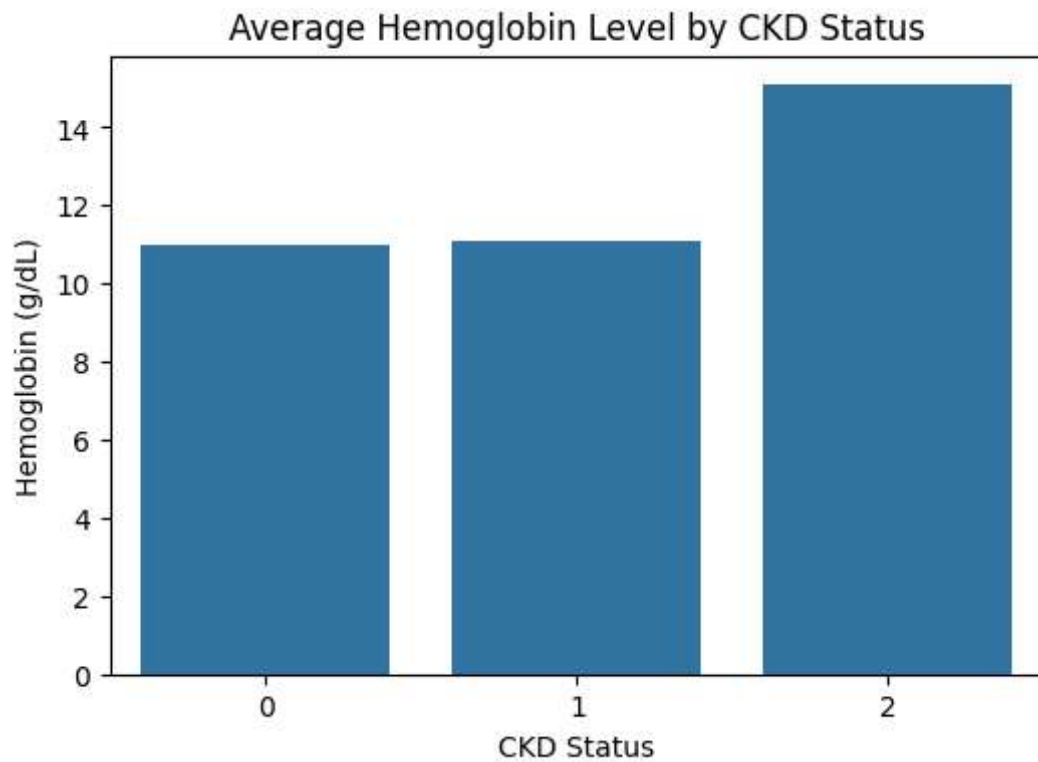




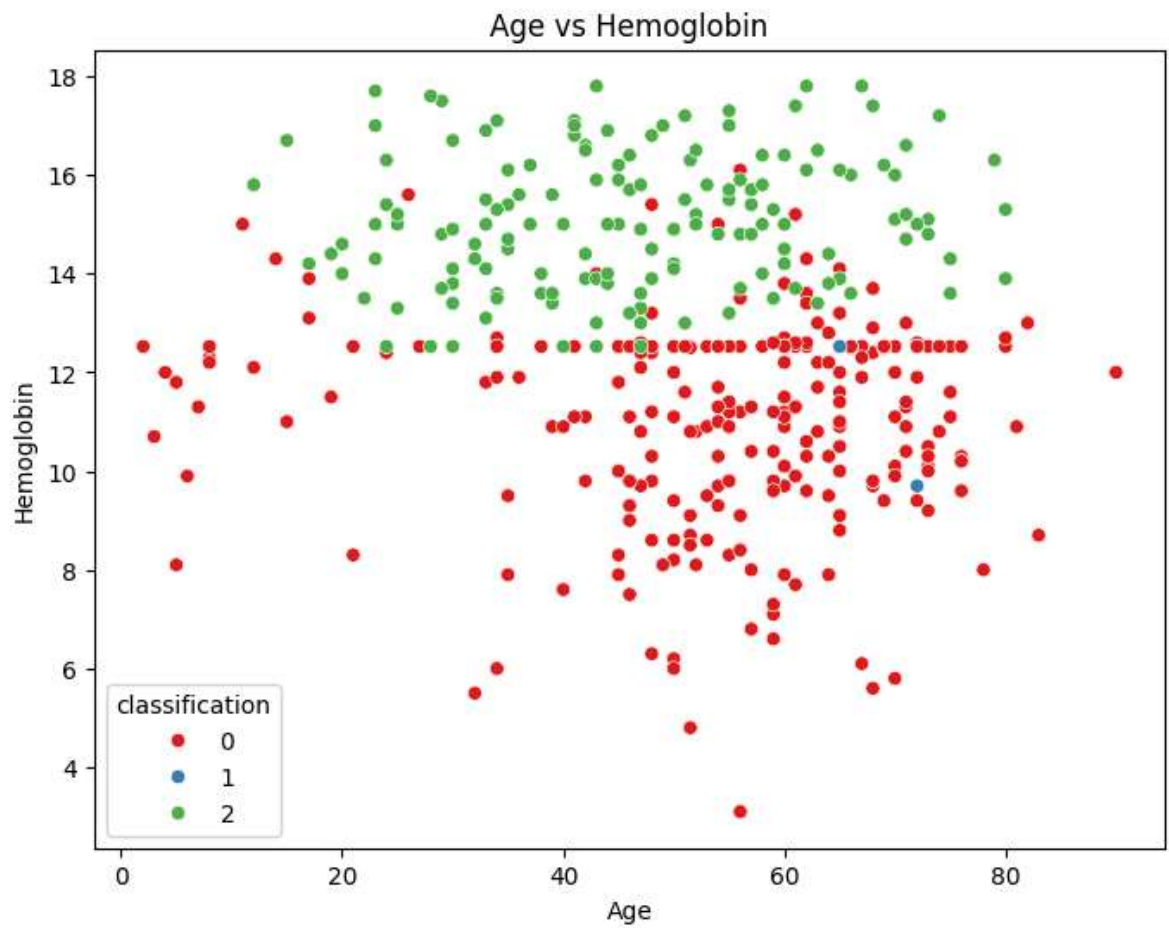
```
plt.figure(figsize=(8, 5))
df['classification'].value_counts().plot(kind='bar', color=['blue',
'pink'])
plt.xticks([0, 1], ['ckd (1)', 'notckd (0)'], rotation=0)
plt.title('summary Kidney Disease Distribution')
plt.xlabel('Class')
plt.ylabel('Count')
plt.grid(axis='y')
plt.show()
```



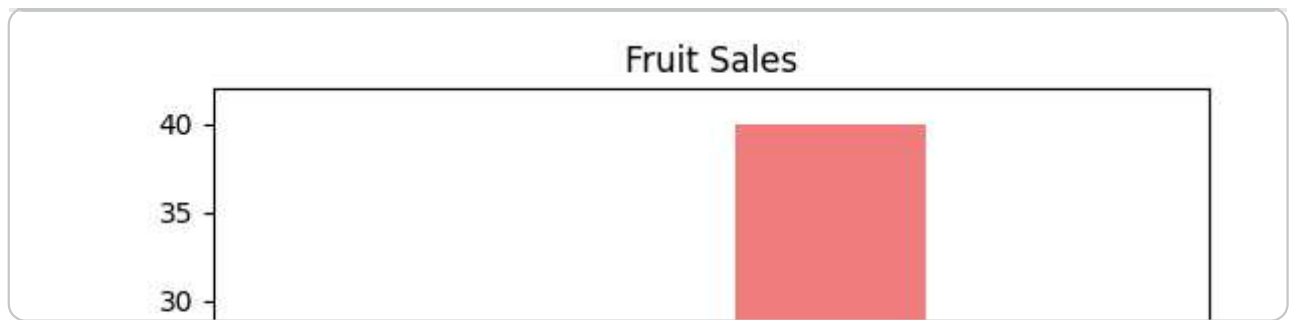
```
import seaborn as sns
plt.figure(figsize=(6,4))
sns.barplot(x='classification', y='hemo', data=df, estimator='mean',
ci=None)
plt.title("Average Hemoglobin Level by CKD Status")
plt.xlabel("CKD Status")
plt.ylabel("Hemoglobin (g/dL)")
plt.show()
```



```
plt.figure(figsize=(8, 6))
sns.scatterplot(x='age', y='hemo', data=df, hue='classification',
palette='Set1')
plt.title('Age vs Hemoglobin')
plt.xlabel('Age')
plt.ylabel('Hemoglobin')
plt.show()
```



```
fruits = ['apple', 'Bananas', 'Cherries', 'Dates']  
sales = [25, 18, 40, 20]  
plt.bar(fruits, sales, color='lightcoral')  
plt.title('Fruit Sales')  
plt.xlabel('Fruit Type')  
plt.ylabel('Quantity Sold')  
plt.show()
```



```
sales_B = [20, 25, 30]
x = np.arange(len(years))
width = 0.35
plt.bar(x - width/2, sales_A, width, label='Product A',
color='orange')
plt.bar(x + width/2, sales_B, width, label='Product B', color='green')
plt.title('Yearly Sales Comparison')
plt.xlabel('Year')
plt.ylabel('Sales')
plt.xticks(x, years)
plt.legend()
plt.show()
```

