

```

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings

warnings.filterwarnings("ignore")

# 1. Load the Dataset
df = pd.read_csv('kidney_disease.csv')

# تنظيف البيانات الأولية (تحويل الأعمدة النصية إلى أرقام إذا لزم الأمر)
df['pcv'] = pd.to_numeric(df['pcv'], errors='coerce')
df['wc'] = pd.to_numeric(df['wc'], errors='coerce')
df['rc'] = pd.to_numeric(df['rc'], errors='coerce')

# --- STEP 1: Handle Missing Values ---
# تقسيم الأعمدة إلى نوعين: عددية وفئوية
numerical_cols = df.select_dtypes(include=['float64', 'int64']).columns
categorical_cols = df.select_dtypes(include=['object']).columns

# ملء القيم المفقودة للأعمدة العددية باستخدام المتوسط (Mean)
for col in numerical_cols:
    df[col] = df[col].fillna(df[col].mean())

# ملء القيم المفقودة للأعمدة الفئوية باستخدام المنوال (Mode)
for col in categorical_cols:
    df[col] = df[col].fillna(df[col].mode()[0])

print("Missing values handled successfully.")

# --- STEP 2: Handle Outliers ---
# للتعامل مع القيم الشاذة في الأعمدة العددية (المدى الربيعي) سنستخدم طريقة IQR
for col in numerical_cols:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    # استبدال القيم الشاذة بالحدود العليا والدنيا (Capping)
    df[col] = np.where(df[col] > upper_bound, upper_bound,
                      np.where(df[col] < lower_bound, lower_bound, df[col]))

print("Outliers handled successfully using IQR capping.")

# --- STEP 3: Data Visualization ---
plt.figure(figsize=(15, 10))

# أ. توزيع الفئات (Target Distribution)
plt.subplot(2, 2, 1)
sns.countplot(x='classification', data=df, palette='viridis')
plt.title('Classification Distribution (CKD vs NotCKD)')

# ب. العلاقة بين العمر وضغط الدم
plt.subplot(2, 2, 2)
sns.scatterplot(x='age', y='bp', hue='classification', data=df)
plt.title('Age vs Blood Pressure')

# ج. توزيع الهيموجلوبين (Distribution of Hemoglobin)
plt.subplot(2, 2, 3)
sns.histplot(df['hemo'], kde=True, color='red')
plt.title('Distribution of Hemoglobin')

# د. مصفوفة الارتباط (Correlation Heatmap)
plt.subplot(2, 2, 4)
sns.heatmap(df[numerical_cols].corr(), annot=False, cmap='coolwarm')
plt.title('Numerical Features Correlation')

plt.tight_layout()
plt.show()

# حفظ الملف النهائي (اختياري)
# df.to_csv('cleaned_kidney_data.csv', index=False)

```

Missing values handled successfully.
Outliers handled successfully using IQR capping.

