

Data Preprocessing and Visualization - Project No. 1

Student: Mohammed Alhawshabi

Instructor: Dr. Hajar

This project demonstrates basic data preprocessing techniques using Python. The tasks include handling missing values, detecting and removing outliers using the IQR method, and performing data visualization to better understand the dataset.

```
# Import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Load datasets
heart = pd.read_csv("heart.csv")
kidney = pd.read_csv("kidney_disease.csv")

# -----
# 1. Handle Missing Values
# -----

print("Missing values in Heart dataset:")
print(heart.isnull().sum())

print("Missing values in Kidney dataset:")
print(kidney.isnull().sum())

# Fill numerical missing values with mean
heart = heart.fillna(heart.mean(numeric_only=True))
kidney = kidney.fillna(kidney.mean(numeric_only=True))

# Fill categorical missing values with mode
for column in kidney.select_dtypes(include='object').columns:
    kidney[column].fillna(kidney[column].mode()[0], inplace=True)

# -----
# 2. Handle Outliers (IQR Method)
# -----

def remove_outliers(df):
    Q1 = df.quantile(0.25)
    Q3 = df.quantile(0.75)

    IQR = Q3 - Q1

    df_clean = df[~((df < (Q1 - 1.5 * IQR)) |
                    (df > (Q3 + 1.5 * IQR))).any(axis=1)]

    return df_clean

heart_clean = remove_outliers(heart.select_dtypes(include=np.number))

# -----
# 3. Data Visualization
# -----

# Histogram
heart_clean.hist(figsize=(10,8))
plt.title("Heart Dataset Distribution")
plt.show()
```

```
# Boxplot
plt.figure(figsize=(10,6))
sns.boxplot(data=heart_clean)
plt.title("Outlier Visualization")
plt.show()

# Correlation Heatmap
plt.figure(figsize=(10,6))
sns.heatmap(heart_clean.corr(), annot=True)
plt.title("Correlation Heatmap")
plt.show()
```