CSE446: Data Management & Visualization

Experiment 1: Variable Types and Graphical Data Exploration

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns


# Load dataset

data = pd.read_csv('data/sample_dataset.csv')


# Display variable types

print(data.dtypes)


# Visualize data

sns.pairplot(data)

plt.title("Pairplot for Data Exploration")

plt.show()
```

Experiment 2: Identifying Outliers and Missing Data

```
import pandas as pd

import numpy as np


# Load dataset

data = pd.read_csv('data/sample_dataset.csv')


# Identify missing data

missing_data = data.isnull().sum()

print("Missing Data:\n", missing_data)
```

```
# Handle missing data (e.g., fill with mean)
data_filled = data.fillna(data.mean())

# Detect outliers using IQR
Q1 = data.quantile(0.25)
Q3 = data.quantile(0.75)
IQR = Q3 - Q1
outliers = (data < (Q1 - 1.5 * IQR)) | (data > (Q3 + 1.5 * IQR))
print("Outliers:\n", outliers)

# Visualize outliers
sns.boxplot(data=data)
plt.title("Boxplot to Visualize Outliers")
plt.show()
```

Experiment 3: Descriptive Statistics and Scatter Plots

```
# Descriptive statistics
print("Descriptive Statistics:\n", data.describe())

# Scatter plot
plt.scatter(data['Variable1'], data['Variable2'])
plt.title("Scatter Plot of Variable1 vs Variable2")
plt.xlabel("Variable1")
plt.ylabel("Variable2")
plt.show()
```

Experiment 4: Grouping and Recoding Variables

```
# Group data by a categorical variable
grouped_data = data.groupby('Category').mean()
print("Grouped Data:\n", grouped_data)
```

```
# Recode string variables
data['Category_Recode'] = data['Category'].replace({'Type1': 1, 'Type2': 2, 'Type3': 3})
print("Recode:\n", data[['Category', 'Category_Recode']])
```