Assignment 4

Data Analysis and Visualization Using Pandas

1. Data Set Selection

I'll download the Iris dataset from the UCI Machine Learning Repository.

2. Data Loading

python code

```
import pandas as pd
```

```
# Load the Iris dataset into a Pandas DataFrame
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
column_names = ['sepal_length', 'sepal_width', 'petal_length', 'petal_width', 'class']
df = pd.read_csv(url, names=column_names)
```

3. Data Exploration

Structure and Features

Let's start by understanding the structure and features of the dataset.

Python code

```
# Display the first few rows of the dataset
print(df.head())

# Check the dimensions of the dataset
print(f"Dataset dimensions: {df.shape}")

# Check the data types and presence of missing values
print(df.info())

# Statistical summary of numerical columns
print(df.describe())
```

Insights:

- The dataset contains 150 instances and 5 columns.
- There are no missing values, and all columns are numerical except for the 'class' column, which is categorical.
- Summary statistics (mean, min, max, quartiles) provide insights into the range and distribution of each numerical feature.

4. Data Cleaning

Since the Iris dataset is clean and well-structured, typically no cleaning steps are necessary. However, if there were missing values or duplicates, we would handle them here.

5. Data Visualization

Example Visualizations:

Python code

```
import matplotlib.pyplot as plt
import seaborn as sns

# Pairplot to visualize pairwise relationships in the dataset
sns.pairplot(df, hue='class', height=2.5)
plt.suptitle("Pairplot of Iris Dataset")
plt.show()

# Boxplot for each feature to visualize the distribution
plt.figure(figsize=(10, 6))
sns.boxplot(data=df.drop(columns='class'), orient='h')
plt.title("Boxplot of Features in Iris Dataset")
plt.show()

# Histogram of each feature grouped by class
plt.figure(figsize=(10, 6))
for i, feature in enumerate(df.columns[:-1]):
    plt.subplot(2, 2, i + 1)
    sns.histplot(data=df, x=feature, hue='class', kde=True)
plt.suptitle("Histograms of Iris Dataset Features")
plt.tight_layout()
plt.show()
```

Insights:

- **Pairplot**: It shows pairwise relationships between features colored by class ('setosa', 'versicolor', 'virginica'). Insights include how features correlate and how well-separated classes are.
- **Boxplot**: It gives a visual summary of the distribution of each feature, highlighting potential outliers and the overall spread of the data.
- **Histograms**: These show the distribution of each feature, providing insights into the range and frequency of values within each class.

6. Analysis and Insights

- **Pairplot**: We observe that the Iris setosa species is well-separated from the other two species across various feature combinations, indicating distinct feature distributions.

- **Boxplot**: Petal length and width show noticeable differences across different Iris species, especially 'setosa' which tends to have smaller dimensions compared to 'versicolor' and 'virginica'.
- **Histograms**: They confirm the distribution patterns seen in the pairplot and provide a closer look at the density of values within each feature for each class.