Assignment: Data Sources in Artificial Intelligence (AI)

Part 1: Data Basics

1. **Question**: Define data and explain its importance in Artificial Intelligence (AI).
    - **Hint**: Discuss the role of data in training AI models, making predictions, and driving decision-making.

Part 2: Types of Data

2. **Question**: Differentiate between Quantitative Data and Qualitative Data. Provide two examples of each.
    - **Hint**: Quantitative Data involves numerical values, while Qualitative Data involves descriptive attributes.
3. **Question**: Explain Structured Data, Unstructured Data, and Semi-Structured Data with examples.
    - **Hint**: Structured Data is organized in rows and columns, Unstructured Data lacks a predefined structure, and Semi-Structured Data has elements of both.

Part 3: Big Data: Volume, Variety, Velocity

4. **Question**: Describe the three Vs of Big Data: Volume, Variety, and Velocity. Provide an example for each.
    - **Hint**: Volume refers to the large amount of data, Variety refers to different types of data, and Velocity refers to the speed of data generation and processing.

Part 4: Database and Other Tools

5. **Question**: List and explain three tools used for managing and analyzing data in AI.
    - **Hint**: Include tools such as SQL databases, NoSQL databases, Hadoop, Spark, Pandas, and NumPy.

Part 5: Data Process

6. **Question**: Outline the steps involved in the data process, from data collection to data visualization.
    - **Hint**: Include data collection, data cleaning, data transformation, data analysis, and data visualization.

Part 6: How Much Data Do You Need for AI?

7. **Question**: Discuss how the amount of data needed for AI varies depending on the complexity of the problem and the algorithm used.
    - **Hint**: Compare simple linear regression models with deep learning models.

Part 7: Data Sources

8. **Question**: Differentiate between Primary Data Source and Secondary Data Source with examples.
    - **Hint**: Primary Data is collected firsthand, while Secondary Data is gathered from existing sources.
9. **Question**: Provide examples of Qualitative Data and Quantitative Data in the context of a customer feedback system.
    - **Hint**: Qualitative Data could be customer reviews, while Quantitative Data could be ratings.
10. **Question**: Explain the importance of Historical and Real-Time Data in AI applications. Provide examples.
    - **Hint**: Historical Data is used for trend analysis and forecasting, while Real-Time Data is used for immediate decision-making.
11. **Question**: Describe Internal Data and External Data with examples.
    - **Hint**: Internal Data is generated within an organization, while External Data is collected from outside sources.

Practical Exercise

12. **Exercise**: Implement a Python script to read data from a CSV file, perform data cleaning (handling missing values, removing duplicates), and visualize the data using line plots and scatter plots.
    - **Hint**: Use the Pandas library for data manipulation and Matplotlib/Seaborn for data visualization.

python

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

# Step 1: Read Data
url = 'https://archive.ics.uci.edu/ml/machine-learning-
databases/adult/adult.data'
columns = ['age', 'workclass', 'fnlwgt', 'education', 'education_num',
'marital_status', 'occupation',
           'relationship', 'race', 'sex', 'capital_gain',
'capital_loss', 'hours_per_week', 'native_country', 'income']
data = pd.read_csv(url, header=None, names=columns, na_values=' ?')

# Step 2: Handle Missing Values
data = data.apply(lambda x: x.fillna(x.value_counts().index[0]))
```

```
# Step 3: Remove Duplicates
data.drop_duplicates(inplace=True)

# Step 4: Visualize Data
for column in data.select_dtypes(include=['float64', 'int64']).columns:
    plt.figure(figsize=(10, 6))
    plt.plot(data[column])
    plt.title(f'Line Plot for {column}')
    plt.xlabel('Index')
    plt.ylabel(column)
    plt.grid(True)
    plt.show()

for column in data.select_dtypes(include=['float64', 'int64']).columns:
    plt.figure(figsize=(10, 6))
    plt.scatter(data.index, data[column])
    plt.title(f'Scatter Plot for {column}')
    plt.xlabel('Index')
    plt.ylabel(column)
    plt.grid(True)
    plt.show()
```