Practical Experiment: Data Cleaning, Preprocessing, and Analysis in Azure

Objective:

To implement a Python script to read data, perform data cleaning and preprocessing steps, and measure variance and range using Microsoft Azure.

Steps:

1. **Set up an Azure Environment**:
    - Create an Azure account if you don't have one.
    - Create a new resource group in Azure.
    - Create an Azure Machine Learning workspace.
2. **Create a Compute Instance**:
    - Go to the Azure Machine Learning workspace.
    - Under the "Compute" section, create a new Compute Instance (e.g., Standard_DS11_v2).
3. **Connect to the Compute Instance**:
    - Once the Compute Instance is running, open the Jupyter Notebook interface from the Azure Machine Learning workspace.
4. **Install Necessary Packages**:
    - In the Jupyter Notebook, create a new notebook and install the necessary packages (Pandas, NumPy, etc.) if not already installed.

python

```
!pip install pandas numpy
```

5. **Download the Dataset**:
    - Use the `wget` command to download the dataset and save it as `adult.csv`.

python

```
!wget https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data -O adult.csv
```

6. **Implement the Data Cleaning and Preprocessing Script**:
    - Write the Python script to read data, perform data cleaning and preprocessing steps, and measure variance and range.

python

```
import pandas as pd
import numpy as np
```

```
# Step 1: Read 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('adult.csv', header=None, names=columns,
na_values=' ?')

# Display the first few rows of the dataset
print("Original Data:\n", data.head())

# Step 2: Handle Missing Values
# Fill missing values with the most frequent value in each column
data = data.apply(lambda x: x.fillna(x.value_counts().index[0]))

# Display the dataset after handling missing values
print("\nData After Handling Missing Values:\n", data.head())

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

# Display the dataset after removing duplicates
print("\nData After Removing Duplicates:\n", data.head())

# Step 4: Transform Data
# Apply a transformation to the 'capital_gain' column (e.g., log
transformation)
data['capital_gain'] = data['capital_gain'].apply(lambda x: np.log(x
+ 1))

# Display the dataset after transformation
print("\nData After Transformation:\n", data.head())

# Step 5: Replace Values
# Replace specific values in the 'income' column
data['income'] = data['income'].replace({' <=50K': 'Low', ' >50K':
'High'})

# Display the dataset after replacing values
print("\nData After Replacing Values:\n", data.head())

# Step 6: Detect and Filter Outliers
# Calculate the Z-scores to detect outliers
data['age_zscore'] = (data['age'] - data['age'].mean()) /
data['age'].std()

# Filter out rows where the Z-score is greater than 3 or less than -
3
data = data[(data['age_zscore'] < 3) & (data['age_zscore'] > -3)]

# Display the dataset after filtering outliers
print("\nData After Filtering Outliers:\n", data.head())

# Drop the Z-score column for final cleaned data
data.drop(columns=['age_zscore'], inplace=True)
```

```
# Display the final cleaned dataset
print("\nFinal Cleaned Data:\n", data.head())

# Measure Variance and Range
def calculate_variance_and_range(column):
    variance = data[column].var()
    data_range = data[column].max() - data[column].min()
    return variance, data_range

columns_to_analyze = ['age', 'capital_gain', 'capital_loss']
for col in columns_to_analyze:
    variance, data_range = calculate_variance_and_range(col)
    print(f"\nVariance and Range for '{col}':\nVariance:
{variance}\nRange: {data_range}")

# Save the cleaned data to a new CSV file
data.to_csv('cleaned_adult.csv', index=False)
print("\nCleaned data saved to 'cleaned_adult.csv'")
```

7. **Run the Script on Azure**:
    - Execute the script in the Jupyter Notebook on your Azure Compute Instance.

Explanation:

1. **Set up Azure Environment**: Create and configure an Azure resource group, machine learning workspace, and compute instance.
2. **Connect to the Compute Instance**: Use the Jupyter Notebook interface from the Azure Machine Learning workspace.
3. **Install Necessary Packages**: Install Python and the necessary libraries using pip.
4. **Download the Dataset**: Use the `wget` command to download the dataset and save it as `adult.csv`.
5. **Implement the Data Cleaning and Preprocessing Script**: The provided script reads the dataset, handles missing values, removes duplicates, transforms data, replaces values, detects outliers, and measures variance and range.
6. **Run the Script on Azure**: Execute the script on your Azure environment to perform data cleaning, preprocessing, and analysis.