Mini-Project: Simple Data Manipulation

Project Overview

Objective: Analyze a real-world dataset to derive insights that can inform decision-making.

Dataset Example: The UCI Machine Learning Repository has a dataset titled **"Wine Quality"**, which contains various physicochemical tests and quality ratings for different wines. This dataset is perfect for exploring relationships between attributes and quality ratings, making it a suitable choice for our analysis.

Expected Outcome: A comprehensive report summarizing your findings, including visualizations and key statistics.

Steps for Your Mini-Project

Step1: Load the Dataset

1. **Loading Data**:

o Download the dataset from UCI Wine Quality Dataset.

o Use pd.read_csv() to load the dataset into a Pandas DataFrame.

Code Template:

```
46import pandas as pd
df = pd.read_csv('winequality-red.csv', sep=';') # Note the separator is semicolon
print(df.head())
```

Step 2: Explore the Data

2. **Exploratory Data Analysis (EDA)**:

o Use df.info(), df.describe(), and df.isnull().sum() to understand the structure and identify missing values.

o Investigate the distribution of the quality ratings.

Code Template:

```
print(df.info())
print(df.describe())
print(df.isnull().sum())
```

Step 3: Data Cleaning

3. **Data Cleaning**:

o Check for duplicates and handle any missing values appropriately.

o You may want to drop rows with missing values or fill them with appropriate statistics.

Code Template:

```
df.drop_duplicates(inplace=True) # Remove duplicate rows
df.fillna(df.mean(), inplace=True) # Fill missing values with column means
```

Step 4: Analyze and Manipulate Data

4. **Data Manipulation**:

o Use groupby() to summarize data based on quality.

o Calculate the average values of the physicochemical properties for each quality rating.

Code Template:

```
quality_summary = df.groupby('quality').mean()
print(quality_summary)
```

Step 5: Summarize Findings

6. **Report Writing**:

o Prepare a report summarizing your key findings.

o Include visualizations and statistics that highlight important insights, such as which physicochemical properties are most correlated with wine quality.

Report Structure:

o **Introduction**: Describe the dataset and objectives.

o **Data Exploration**: Summarize findings from the EDA.

o **Data Cleaning**: Discuss any cleaning steps taken.

o **Analysis**: Present insights from data manipulation.

o **Visualizations**: Include relevant charts and explain their significance.

o **Conclusion**: Summarize key takeaways.

Step 6: Present Your Findings

7. **Presentation**:

o Prepare to present your findings to the class using Jupyter Notebook or PowerPoint.

o Highlight key insights and explain your visualizations and their implications.

Example Analysis Questions

To guide your analysis, consider the following questions:

· What is the distribution of wine quality ratings in the dataset?

· Which physicochemical properties are most strongly correlated with wine quality?

· Are there any noticeable trends or patterns in the data based on the quality ratings?

· How does the alcohol content affect the quality of wine?