TABULAR DATA AND DATAFRAMES

INTRODUCTION

When working with large datasets, it's essential to understand and manipulate the data efficiently. Tabular data is a common format where data is organized in rows and columns. In data science, we often need to analyze, transform, and visualize this data. Python offers powerful libraries to handle such tasks.

LIBRARIES FOR TABULAR DATA

Pandas

- **Description**: Pandas is a library that provides data structures and functions needed to manipulate structured data seamlessly. It allows you to work with Dataframes, which are analogous to relational tables.
- **Key Features**:
  - Named columns and indices.
  - Various operations on rows, columns, and dataframes.
  - Easy data manipulation and transformation.

Numpy

- **Description**: Numpy is a library for numerical computing. It deals with arrays (tensors), which are multi-dimensional data structures with values of the same type.
- **Key Features**:
  - Efficient mathematical operations.
  - Lesser overhead compared to dataframes.
  - Often used for numerical and scientific computing.

Other Useful Libraries

- **Matplotlib**: A library for data visualization and plotting graphs.
- **SciPy**: Provides additional scientific functions and builds on Numpy. Useful for tasks in probability, statistics, and more.

IMPORTING LIBRARIES

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from scipy import ... # specify exact sub-packages needed
```

PANDAS BASIC CONCEPTS

1. Series

A Series in Pandas is a sequence of values like a list or numpy array, but with an index. The index can be a simple integer or a more complex structure like a date interval.

2. DataFrame

A DataFrame is a collection of series with the same index, forming a 2D table.

Important DataFrame Operations

- **Column Selection**: df['A'] selects a single column. df[['B', 'A']] selects multiple columns.
- **Row Filtering**: df[df['A'] > 5] filters rows based on a condition.

3. Grouping and Aggregating

Grouping is useful for summarizing data, similar to pivot tables in Excel.

4. Loading Data

Pandas can load data from various sources like CSV files.