823G5 Programming in Python

Aim

The aim of this lab is to provide students with hands-on experience in working with real-world datasets using Python and the `Pandas` library. By the end of this lab, students should be able to:

1. Load, inspect, and understand the structure of a dataset.
2. Perform data cleaning and preprocessing tasks, including handling missing values and duplicates.
3. Conduct exploratory data analysis (EDA) to gain insights into the dataset's characteristics.
4. Manipulate and transform data using `Pandas`, including creating new columns and applying filters.
5. Visualize data using `Matplotlib` and `Seaborn` for better understanding and interpretation.
6. Apply principles of tidy data to enhance data organization and readability.
7. Practice merging datasets to consolidate information and gain a comprehensive view of the data.

Exercises

Download the **`Lab9and10Exercises`** ZIP file for solutions and refer the **`company_data.csv`** file. For this lab we will use a Jupyter Notebook to solve the challenges as this oner of the preferred ways of working in Data Science and displays `DataFrame`s with rich formatting. To use a Jupyter notebook, it is best to launch Jupyter Lab. Alternatively, some IDEs such as VSCode can open and edit and display Jupyter notebooks.

Challenge 1

1) **Load and Inspect the Data**
   Load the "`company_data.csv`" file into a `Pandas DataFrame`.
   Display the first 10 rows of the dataset.
2) **Data Overview**
   Provide basic statistics for the numerical columns (e.g., count, mean, min, max).
   Display information about the data types and missing values.
3) **Data Cleaning**
   Check for and handle any missing values in the dataset.
   Identify and remove duplicate records if any.
4) **Exploratory Data Analysis (EDA)**
   Visualize the distribution of the "salary" column using a histogram.
   Create a boxplot to show the distribution of "rating" for different departments.

Challenge 2

1) **Data Manipulation with Pandas**
   Create a new column "`total_cost`" by multiplying the "`quantity`" and "`price`" columns.
   Add a column "`is_expensive`" with a boolean value indicating if the product is expensive (price > 200).
2) **Filtering and Subset Selection**
   Filter the dataset to include only records where the "`department`" is "`IT`" and "`rating`" is greater than 4.0.
   Select and display records for products with a quantity greater than 5.
3) **Data Visualization with Matplotlib and Seaborn**
   Create a scatter plot to visualize the relationship between "`quantity`" and "`price`".
   Use a bar plot to display the average salary for each department.
4) **Tidy Data and Merging**
   Transform the dataset into tidy format, keeping columns "`employee_id`," "`name`," "`department`," and "`salary`" as identifier variables.
   Merge the dataset with itself on the "`product_id`" column and display the result.

Dr. Benjamin Evans
B.D.Evans@sussex.ac.uk