TP Worksheet: Data Wrangling

Ismail JAMIAI

December 17, 2024

Introduction

This TP worksheet contains five mini-projects designed to help you practice data wrangling techniques using the pandas library in Python. Each project includes a description, a link to the dataset, and a set of tasks to complete.

Exercise 1: Titanic Survivor Analysis with Visualization

Description: You have a dataset of Titanic passengers. Your task is to create a pandas DataFrame from the data, then calculate the number of survivors and non-survivors by passenger class. Next, visualize this data using a bar chart.

Tasks:

1. Load the Titanic dataset from the provided URL.
2. Calculate the number of survivors and non-survivors for each passenger class.
3. Create a bar chart to visualize the number of survivors and non-survivors by class.

Dataset: `https://raw.githubusercontent.com/chrisalbon/sim_data/master/titanic.csv`

Exercise 2: Handling Missing Values and Age Analysis

Description: You have a DataFrame containing data on Titanic passengers, but some age values are missing. Your task is to replace these missing values with the mean age. Then, analyze the age distribution by creating a histogram.

Tasks:

1. Load the Titanic dataset.
2. Replace missing age values with the mean age.
3. Create a histogram to visualize the age distribution of passengers.

Dataset: `https://raw.githubusercontent.com/chrisalbon/sim_data/master/titanic.csv`

Exercise 3: Renaming Columns and Adding a New Feature

Description: You have a DataFrame with unclear column names. Your task is to rename the columns to make them more descriptive. Then, add a new column that indicates whether a passenger is a child (age < 18) or an adult.

Tasks:

1. Load the Titanic dataset.
2. Rename the columns to make them more descriptive.
3. Add a new column to indicate if the passenger is a child or an adult.

Dataset: `https://raw.githubusercontent.com/chrisalbon/sim_data/master/titanic.csv`

Exercise 4: Grouping by Class and Calculating Detailed Statistics

Description: You have a DataFrame containing information about Titanic passengers, grouped by class (1st, 2nd, 3rd). Your task is to calculate the mean, median, minimum, and maximum age for each class. Then, display these statistics in a DataFrame.

Tasks:

1. Load the Titanic dataset.
2. Group the data by passenger class.
3. Calculate the mean, median, minimum, and maximum age for each class.
4. Display the calculated statistics in a DataFrame.

Dataset: `https://raw.githubusercontent.com/chrisalbon/sim_data/master/titanic.csv`

Exercise 5: Merging DataFrames and Sales Analysis

Description: You have two DataFrames, one containing information about employees and the other containing their sales. Your task is to merge these two DataFrames on a common column (e.g., employee ID). Then, calculate the total sales by employee and display the top 5 sellers.

Tasks:

1. Create two DataFrames: one for employees and one for sales.
2. Merge the two DataFrames on the employee ID column.
3. Calculate the total sales by employee.
4. Display the top 5 sellers based on total sales.

Dataset: You can use fictional data as in the example or real data if available.