Introduction to Pandas Week 1 Project: Beginner Level

Welcome to Pandas Week 1!

This week, we will focus on exploring, cleaning, and analysing datasets using Pandas. By the end of this assignment, you will have hands-on experience with essential pandas functionalities, including data loading, cleaning, manipulation, and simple analysis using real-world datasets.

PART A - Learning Resources

To help you get started, here are some learning resources to guide you through using pandas:[](https://www.youtube.com/watch?v=I3FBJdiExcg)

PART B - Hands-On Assignment

Dataset

For this assignment, you will work with the [Titanic Dataset](https://www.kaggle.com/datasets/brendan45774/test-file) available on Kaggle. This dataset contains information about passengers aboard the Titanic, such as age, sex, ticket class, fare, and whether they survived the disaster.

Assignment Questions:

Data Loading and Cleaning:

1. Loading the Dataset

How would you load the Titanic dataset into a pandas DataFrame?

How can you display the first few rows to inspect the data?

2. Missing Values:

How many missing values are present in each column of the dataset?

What method can you use to check for this?

3. Handling Missing Data:

If the `Age` column contains missing values, how would you handle them?

How would you handle missing values in the `Embarked` column?

Would you drop the rows entirely, or would you fill them with a default value?

4. Removing Duplicates:

How can you check for duplicate rows in the dataset?

How can you remove duplicate rows in the dataset?

5. Data Types:

How do you check for the data types for each columns?

Which columns in the dataset might have incorrect data types, and how would you ensure all columns have appropriate data types (e.g., converting `PassengerId` to a categorical type)?

Data Exploration:

1. Summary Statistics:

What methods can you use to calculate basic summary statistics (e.g., mean, median, min, max) for numeric columns like `Age`, `Fare`, etc.?

2. Survival by Class:

How can you group passengers by `Pclass` and `Survived` to determine how many passengers in each class survived?

3. Average Age and Survival:

How would you calculate the average age of passengers who survived compared to those who did not?

4. Gender Distribution:

How can you determine the distribution of genders (`Sex`) among passengers?

Data Analysis:

1. Survival Rates by Age Groups:

How would you calculate the survival rate for passengers in different age groups (e.g., <18, 18–40, 40–60, 60+)?

What method would you use to bin the ages into these categories?

2. Survival Rates by Gender and Class:

How can you use the `groupby()` function with `Pclass` and `Sex` to determine survival rates based on gender and passenger class?

Data Visualisation:

1. Bar Chart of Survival Rates:

How would you create a bar chart to compare survival rates for different age groups and genders?

PART C - Submission Instructions

You are expected to submit your assignment by the end of the week. Submission will be done on Twitter or LinkedIn.

Twitter Submission:

- Take a screenshot of your code/outputs.

- Tag:

1. The Official page: @TDataInitiative

2. The project coordinator: @The\_Jonathaan

3. Use the hashtag: #TDI\_Pandas

LinkedIn Submission:

- Upload your screenshots of code and results.

- Tag:

1. The TDI Official page: @TheDataInitiative

2. The project coordinator: @JonathaanCoordinator

GitHub Submission Instructions:

For this assignment, you will be required to submit your work via GitHub. Please follow the instructions below to submit your project:

* 1. Create a New Repository:
     + Log in to your GitHub account and create a new repository named TDI\_DE.
     + Make sure the repository is public so that it can be reviewed.
  2. Project Structure:
     + In your repository, structure your project folder like this:

Titanic\_Data\_Analysis/

├── README.md

├── data/

│ └── titanic.csv # (Optional - Kaggle link can be referenced here)

├── notebooks/

│ └── titanic\_analysis.ipynb # Jupyter notebook for analysis

│ └── visuals/

└── bar\_chart.png # Place any generated visualizations here

3. The README.md should briefly describe the project and instructions for running the code.

4. Add your Jupyter notebook that contains the code for loading, cleaning, exploring, and visualising the Titanic dataset.

Push the Code:

* + - Add and commit your files to the repository.
    - Push the repository to GitHub

PART D - Correction Class

Correction classes will be held every \*\*Saturday from 4 pm to 6 pm Nigerian Time\*\* on the TDI official Discord or a Google Meet link will be shared before the class.

Good luck with your assignment! We hope you find it both challenging and rewarding. If you have any questions, feel free to reach out to the mentors or the community on the TDI platform.