Week 1 Lesson Plans: Data Handling with Pandas & NumPy

Goal: Build a strong foundation in data manipulation.

Day 1 – Mon 22 Sept | NumPy Refresher: Arrays, Indexing, Operations

⏱️ **2 hours**

- **(5 min) Intro** → Course overview, why NumPy is essential.
- **(15 min) Quick reminder** → Difference between Python lists vs NumPy arrays.
- **(30 min) Arrays & Creation** → `np.array`, `np.arange`, `np.linspace`, `np.zeros`, `np.ones`, `np.random`.
- **(20 min) Indexing & Slicing** → accessing elements, rows/columns, slices.
- **(25 min) Reshaping & Operations** → `.reshape()`, transpose, math ops (sum, mean, std, matrix ops).
- **(20 min) Mini Project** → Build 5×5 matrix (1–25), extract rows/cols, compute stats.
- **(5 min) Reflection** → Why is NumPy faster than lists?

Day 2 – Tue 23 Sept | Pandas Series & DataFrames

⏱️ **2 hours**

- **(5 min) Reminder** → NumPy arrays → Pandas built on NumPy.
- **(15 min) Theory** → What is Pandas, Series, DataFrame.
- **(30 min) Series** → create, index, slicing, operations.
- **(40 min) DataFrames** → create from dict/CSV, `.head()`, `.tail()`, `.info()`, `.describe()`.
- **(20 min) Indexing** → `loc` vs `iloc`.
- **(5 min) Reflection** → Why DataFrames are more powerful than arrays?

Day 3 – Wed 24 Sept | Import & Export Data (CSV, Excel, JSON)

⏱️ **2 hours**

- **(5 min) Reminder** → Yesterday: building DataFrames manually → today: importing real data.
- **(20 min) Theory** → file formats (CSV, Excel, JSON) + importance.
- **(30 min) Importing** → `pd.read_csv`, `pd.read_excel`, `pd.read_json`.
- **(25 min) Exporting** → `.to_csv`, `.to_excel`, `.to_json`.
- **(30 min) Hands-on** → Load Titanic dataset (`sns.load_dataset("titanic")`), inspect, save first 100 rows to CSV.
- **(10 min) Reflection** → Which format is most common in real-world data & why?

Day 4 – Thu 25 Sept | Data Cleaning (Missing Values, Duplicates, Types)

⏱️ **2 hours**

- **(5 min) Reminder** → Yesterday we imported messy datasets → today we clean them.
- **(20 min) Theory** → importance of cleaning, bad data = bad models.
- **(30 min) Handling missing values** → `isnull()`, `dropna()`, `fillna()`.
- **(20 min) Handling duplicates** → `duplicated()`, `drop_duplicates()`.
- **(25 min) Data types** → `df.dtypes`, `astype()`.
- **(15 min) Mini Challenge** → Clean Titanic dataset: fill missing ages, drop duplicates.
- **(5 min) Reflection** → Why cleaning is 80% of data science work?

Day 5 – Fri 26 Sept | Data Wrangling (Filtering, GroupBy, Merge, Pivot)

⏱️ **2 hours**

- **(5 min) Reminder** → Yesterday we cleaned → now we transform data.
- **(20 min) Theory** → what is wrangling, why needed for insights.
- **(30 min) Filtering** → conditions, multiple conditions (`df[df['Age'] > 30]`).
- **(30 min) GroupBy & Aggregations** → `.groupby('col').mean()`.
- **(20 min) Merge & Join** → `pd.merge`, `pd.concat`.
- **(10 min) Pivot tables** → `pd.pivot_table()`.
- **(10 min) Mini Challenge** → On Titanic dataset: group by sex, compute survival rates; pivot table by class & sex.

- **(5 min) Reflection** → How wrangling prepares data for visualization & ML.