# **📊 Data Cleaning Training Tasks**

### **General Submission Rules for All Tasks**

* Use a dataset from **Kaggle** (link is provided).
* Perform the required cleaning operations.
* Submit:  
  1. **Cleaned dataset (.csv)**
  2. **Python notebook (.ipynb)** with explanations.
  3. **Video (5–10 mins)** explaining steps & output. Upload video to **Google Drive** and **share the link** with your assigned buddy.

## **Task 1: Handling Missing Values**

* **Dataset:** Titanic - Machine Learning from Disaster
* **Task:** Identify missing values, decide whether to fill (mean/median/mode) or drop, and justify your choice.
* **Hint:** Use df.isnull().sum(), df.fillna(), df.dropna().

## **Task 2: Removing Duplicates**

* **Dataset:** Netflix Movies and TV Shows
* **Task:** Check for duplicate rows and remove them. Analyze how many records were lost after cleaning.
* **Hint:** Use df.duplicated(), df.drop\_duplicates().

## **Task 3: Handling Outliers**

* **Dataset:** House Prices - Advanced Regression Techniques
* **Task:** Detect outliers in numeric columns (e.g., SalePrice, LotArea) using IQR method and handle them.
* **Hint:** Use quantile(), IQR = Q3 - Q1, filter out values outside [Q1–1.5*IQR, Q3+1.5*IQR].

## **Task 4: String Cleaning & Standardization**

* **Dataset:** IMDb Movies Dataset
* **Task:** Clean text columns (e.g., remove extra spaces, standardize capitalization, correct typos).
* **Hint:** Use str.strip(), str.lower(), str.replace().

## **Task 5: Date & Time Cleaning**

* **Dataset:** Air Quality Data Set
* **Task:** Convert date/time columns to proper datetime format and extract useful features (year, month, weekday).
* **Hint:** Use pd.to\_datetime(), dt.year, dt.month, dt.dayofweek.

## **Task 6: Feature Engineering from Raw Data**

* **Dataset:** Uber Pickups in New York City
* **Task:** Create new features like hour\_of\_day, day\_of\_week from pickup datetime.
* **Hint:** Use .dt.hour, .dt.day\_name().

## **Task 7: Encoding Categorical Variables**

* **Dataset:** Students Performance Dataset
* **Task:** Convert categorical columns (gender, lunch type, test preparation course) into numeric using encoding.
* **Hint:** Use pd.get\_dummies() or LabelEncoder.

## **Task 8: Scaling & Normalization**

* **Dataset:** Mall Customers Dataset
* **Task:** Normalize income and spending score columns for better comparison.
* **Hint:** Use MinMaxScaler or (x - min) / (max - min).

## **Task 9: Data Type Conversion & Validation**

* **Dataset:** World Population Dataset
* **Task:** Ensure numeric columns are properly converted to integers/floats and check for invalid entries.
* **Hint:** Use df.dtypes, pd.to\_numeric(errors='coerce').

## **Task 10: Full Cleaning Project**

* **Dataset:** Zomato Restaurants Data
* **Task:** Perform end-to-end cleaning:  
  1. Handle missing values
  2. Remove duplicates
  3. Standardize text columns
  4. Correct data types
  5. Handle outliers
* **Hint:** Combine all techniques from previous tasks.