Department of Computer Science and Engineering (Data Science)
St Joseph Engineering College, Vamanjoor, Academic Year:2025-26

HANDS ON WORKSEET-1

Subject: Principles of DataScience(22CDS54) Sem: V

Data Wrangling: It prepares raw-data for analysis.
This process involves discovering, cleaning and reformatting, restructuring/reshaping, enriching and validating data.

1. Discovering

(i) Read in the Crop_Production data from the csv file. Write code to list all column names that have trailing or leading spaces, check for inconsistent spellings or capitalization in categorical columns values like 'District'

Hints:

- *df.info/columns*
- *df[‘col’].unique*

(ii) Find the top 5 highest 'Yield' crops on average across the entire dataset.

Hints:

- *df.groupby:Splitting, Applying, and Combining.*
- *mean(),* sort_values(), head()

(iii) Find all crop names ('Crop') that appear fewer than 300 times in the entire dataset.

Hints:

- *value_count()*

2.Structuring

(i) Rename all column names that have trailing or leading spaces

Hints:

- *df.rename(columns={oldname,newname})*

(ii) What is the seasonal performance of each crop, in terms of both total Production and average Yield

Hints:

- *df.groupby a list of column names*
  - *agg(): apply one or more aggregation functions to one or more columns,*

(iii) What is the year-wise total production for each crop

Hints:

- *df.pivot(): transform a DataFrame from a "long" format to a "wide" format*
  Arguments: index, columns, and values.

(iv) Which crop, in which year, produced the highest average yield

Hints:

- *df.melt(): transform or reshape DataFrames from a "wide" format to a "long" format.*
  Arguments: id_vars, value_vars, var_name, value_name
  [use transformed data from pivot)

3. Cleaning

(i) Change inconsistent columns values like 'State' and 'District' from uppercase to lowercase and remove trailing or leading spaces

Hints:

- *df['col'].str.lower(), df['col'].str.strip()*

(ii) Instead of filling missing Production with 0 or a global mean, impute the NaN values by calculating Area * (Average_Yield). The Average_Yield should be the average for that specific District and Crop combination.

Hints:

- *groupby(['District', 'Crop']), transform('mean'), and fillna()).*

(iii) Find the wheat production of Odisha state

Hints:

- *filter by Multiple Conditions (AND - &)*

4. Enrichment

(i) Merge with External Data

Hints:

- *df.merge(data1,data2)*
  Check whether common column and its values are compatible or not before merge

(ii) Create new column {Pollutant ratios} derived from existing columns {SO2,NOx}

Hints:

- *df[new_col]= derived formula*
  Pollutant ratios= SO2/NOx

5. Validating

(i) The Yield column should equal Production / Area. Create a new column called Calculated_Yield using this formula. Then, find the sum of the difference between Calculated_Yield and Yield to see if the original data was accurate.

(ii) Check if any Area, Production, or Yield values are negative.

Check if Production > 0 and Area == 0

Check if any Crop_Year is in the future (e.g., > 2024).