INSTITUTE OF AERONAUTICAL ENGINEERING
(Autonomous)
Dundigal - 500 043, Hyderabad, Telangana

COURSE CONTENT

DATA WRANGLING AND VISUALIZATION LABORATORY								
V Semester: CSE(DS)								
Course Code	**Category**	**Hours / Week**			**Credits**	**Maximum Marks**		
		L	**T**	**P**	**C**	**CIA**	**SEE**	**Total**
ACDC09	**Foundation**	0	0	3	1.5	30	70	100
Contact Classes: Nil	**Tutorial Classes: Nil**	**Practical Classes: 36**				**Total Classes:36**		
Prerequisites: Python Programming Laboratory								

I. COURSE OVERVIEW:

Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis. In this laboratory course the experiments are designed using Python programming language to demonstrate the methods of importing data from data files and world web, the methods of data cleaning and data conditioning and merging datasets, the methods of data exploration through summary tables and data visualization. Experiments on data visualization are devised to explore the relationships among the features of a dataset. As significant data is available on the world web, experiments are also designed to scrape data from world web. This laboratory course also provides some critical thinking experiments to enhance learning. The design and implementation skills gained in this laboratory course are quite useful for full-fledged for real-world data analysis needed for an entry-level data science engineer.

II. COURSE OBJECTIVES

The students will try to learn:

I. The Importing from and exporting data to Excel, CSV, JSON and XML files.
II. The working with data frames using Numpy and Pandas library.
III. The cleaning and exploration of the datasets.
IV. The visual representations of data patterns and draw conclusions

III. COURSE OUTCOMES:

At the end of the course students should be able to:

CO 1	**Demonstrate** the vector and the matrix operations using python libraries in python environment.
CO 2	**Experiment** with Read, write, and data manipulations of Excel, CSV, JSON, PDF and XML files in Python.
CO 3	**Apply** the data clean-up and formatting operations on the given dataset to prepare it ready for further exploration.
CO 4	**Analyze** the data statistically to find correlations and patterns in the data.
CO 5	**Analyze** the data using visual representations for understanding the statistical summaries and patterns.
CO 6	**Make** use of python libraries to fetch content from a web page.

IV. SYLLABUS

EXERCISES FOR DATA WRANGLING AND VISULZATION LABORATORY

Note: Students are encouraged to bring their own laptops for laboratory practice sessions.

1. Getting Started Exercises

1.1 Exercises on Arrays using NumPy Library

NumPy library gives an enormous range of fast and efficient ways of creating arrays and manipulating numerical data inside them. All of the elements in a NumPy array should be homogeneous. The different mathematical operations can be performed on arrays.

One way we can initialize NumPy arrays is from Python lists. Another way is to create a NumPy array use the function numpy.array().

Installing NumPy: pip install numpy

Write a program to create a 1-D and 2-D numpy array with with pseudo random numbers.

Hints

# Download a CSV file Brazilian-fire-dataset.csv containing some sample data and save it under the file path as dataset/Brazilian-fire-dataset.csv

# import the necessary libraries

# print the array

Try

1. Write a program to create a 3D NumPy array with ones.

2. Write a program to perform basic arithmetic operations (addition, subtraction, multiplication, and division) on NumPy arrays.

3. Write a program to access and print specific elements of a NumPy array.

4. Write a program to slice a 2D NumPy array to select rows and columns.

5. Write a program to reshape a 1D NumPy array into a 2D array.

1.2 Exercises on Matrix Operations using Scipy Library

Write a program to Create a 3x3 matrix and a 3x1 vector using NumPy, then solve a linear system of equations using SciPy's linear algebra functions.

This needs installation of SciPy package. Install using the command “pip install scipy”.

Hints

```
import numpy as np

from scipy.linalg import solve

# import the other necessary libraries

# print the matrix
```

Try

1. Write a program to Integrate a mathematical function numerically using SciPy's integration functions, such as quad.

2. Write a program to create square matrix B using sciPy's scipy.linalg.eig function to calculate the eigenvalues and eigenvectors of matrix B.

3. Write a program to create square matrix C to compute the inverse of matrix C.

1.3 Exercises on dataframes using Pandas Library

This needs installation of Pandas package. Install using the command “pip install pandas”.

Write a program to create a series using pandas library as:

given_list = [2, 4, 5, 6, 9]

Hints

```
import pandas as pd

# import the other necessary libraries

# print the series
```

Try

1. Write a Pandas program to create the today’s date.

2. Write a Pandas program to select columns by data type of a given dataframe.

Sample Output:

```
Original DataFrame
name date_of_birth age
0 Alberto Franco 17/05/2002 18.5
1 Gino Mcneill 16/02/1999 21.2
2 Ryan Parkes 25/09/1998 22.5
3 Eesha Hinton 11/05/2002 22.0
4 Syed Wharton 15/09/1997 23.0

Select numerical columns
age
0 18.5
```

```
1 21.2
2 22.5
3 22.0
4 23.0

Select string columns
name date_of_birth
0 Alberto Franco 17/05/2002
1 Gino Mcneill 16/02/1999
2 Ryan Parkes 25/09/1998
3 Eesha Hinton 11/05/2002
4 Syed Wharton 15/09/1997
```

3. Write a Pandas program to display the dimensions or shape of the World alcohol consumption dataset “https://docs.google.com/spreadsheets/d/e/2PACX-1vRdo6mReKHWZ996tjlkkftal2gwTh7Yjf4HuqpojAfQ5B5f6H2u6iL3xdTK2liTV_HIzVYjzPQCU7ub/pubhtml“. Also extract the column names from the dataset.

2. Exercises on data importing from data files-I

2.1 Working with Excel files

1. Write a Python program to read an xlsx file into a dataframe using a library in python from the link https://www.dol.gov/sites/dolgov/files/ETA/naws/pdfs/NAWS_A2E197-xls.zip.

Hints

```
# Download a excel file

# import the necessary libraries

# read the excel file
```

Try

1. Write a Python program to read specific columns from an xlsx file and put it into a dataframe.

2. Write a Python program to write a dataframe into the Excel file using a python library in python.

2.2 Working with CSV files

Write a Python program that reads data from a CSV file that may use either a comma or a semicolon as the delimiter. Parse the data and print it.

Hints

```
# import the necessary libraries

# Define potential delimiters to try

# read the csv file
```

Note: Download a CSV file "sales_data.csv" containing some sample data and save it under the file path as dataset/ sales_data.csv from the https://www.kaggle.com/datasets/manovirat/groceries-sales-data/download?datasetVersionNumber=3.

Download the "sales_data.csv" file or create a CSV file with similar characteristics. Ensure that the file contains at least three different sets of headers for different time periods.

Hints

```
# import the necessary libraries

# Define potential headers to try

# read the csv file
```

Write a Python program to read the CSV file into a dataframe using a library in python.

Hints

```
# Download a CSV file Brazilian-fire-dataset.csv containing some sample data and save it under the file path as dataset/Brazilian-fire-dataset.csv

# import the necessary libraries

# read the csv file
```

Create a CSV file containing international characters (e.g., accented letters or non-ASCII characters) and save it with UTF-8 encoding. Write a Python program to read and process the data from this file using UTF-8 encoding.

Hints

```
# import the necessary libraries

# Define potential encoding parameters to try

# read the csv file
```

Notes: Download a CSV file " international_data.csv" containing some sample data and save it under the file path as dataset/ international_data.csv

Try

1. Write a program to read a given CSV file as a dictionary.

2. Create a text file that contains data with a custom delimiter of your choice (e.g., # or |). Write a Python program to read the data from the file, split it using the custom delimiter, and display the parsed data.

3. Write a Python program that reads data from a CSV file that may contain mixed delimiters (e.g., some rows using commas, and some rows using semicolons). Parse the data and separate it into two lists: one for comma-delimited rows and one for semicolon-delimited rows.

4. Download the "https://data.montgomerycountymd.gov/api/views/v76h-r7br/rows.json?accessType=DOWNLOAD " file. Ensure that the file contains at least three different sets of keys for different records and dynamically handles the different keys.

5. Create another CSV file containing international characters and save it with ISO-8859-1 (Latin-1) encoding. Write a Python program to read and process the data from this file using ISO-8859-1 encoding.

6. Create a CSV file with mixed encoding errors, such as characters that are not valid in UTF-8. Write a Python program that reads the file and handles encoding errors gracefully by skipping problematic lines or replacing invalid characters.

2.3 Working with PDF files

Write a Python program to write a dataframe into the CSV file using a python library in python.

Hints

```
# import the necessary libraries

# Define potential delimiters to try

# Write a dataframe into a CSV file
```

Try

1. Write a Python program to display a pdf file into a dataframe using a library in python.

Note: This needs installation of PYPDF2 package. Install using the command “pip install PyPDF2”.

2. Write a Python program to write a dataframe into the pdf file using a python library in python.

Note: This needs installation of pdfkit package. Install using the command “pip install pdfkit”.

3. Exercises on data importing from data files-II

3.1 Working with xml files

Write a Python program to read the xml file into a dataframe in python.

Hints

```
# Download a xml file containing some sample data and save it under the file path as dataset/filename.xml

# import the necessary libraries

# read the xml file
```

Note: This needs use of built xml.etree.ElementTree module in python.

Try

1. Write a Python program that reads data from a xml file that may use either a comma or a semicolon as the delimiter. Parse the data and print it.

3. Write a Python program to read the html file: https://www.baseball-reference.com/players/r/riverma01.shtml into a dataframe in python.

3.2 Working with json files

Write a Python program to write a dataframe into the json file using a python library in python.

Hints

```
# import the necessary libraries

# Define potential encoding parameters to try

# read the file
```

Try

1. Write a Python program that reads data from a json file that may use either a comma or a semicolon as the delimiter. Parse the data and print it.

2. Write a Python program to read the json file into a dataframe in python.

3. Write a Python program to convert Python dictionary object (sort by key) to JSON data.

3.3 Working with html files

Write a Python program to write a dataframe into the xml file using a python library in python.

Hints

```
# import the necessary libraries

# Define potential delimiters to try

# Write a dataframe into a CSV file
```

Try

1. Write a Python program that reads data from a html file that may use either a comma or a semicolon as the delimiter. Parse the data and print it.

2. Write a Python program to write a dataframe into the html file using a python library in python.

4. Exercises on Data Frames-I

4.1 Create data structures and convert into a dataframe

Write a program to create a Python dictionary with keys name, score, attempts and qualify as column names and values as lists containing exam data. Display a dataframe.

exam data = {‘name’: [‘Anastasia’, ‘Dima’, ‘Katherine’, ‘James’, ‘Emily’, ‘Michael’, ‘Matthew’, ‘Laura’, ‘Kevin’,‘Jonas’], ‘score’: [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19], ‘attempts’: [1, 3, 2, 3, 2, 3, 1, 1, 2, 1], ‘qualify’: [‘yes’, ‘no’, ‘yes’, ‘no’, ‘no’, ‘yes’, ‘yes’, ‘no’, ‘no’, ‘yes’]}
labels = [‘a’, ‘b’, ‘c’, ‘d’, ‘e’, ‘f’, ‘g’, ‘h’, ‘i’, ‘j’]

Hints

```
# import the necessary libraries

# Create a Python dictionary with keys as column names and lists as values

# Create a dataframe from the dictionary

# Display the dataframe
```

Note: To create a Pandas dataframe from a dictionary, you can use constructor from the Pandas library. print() can be used to display the data structure.

Try

1. Write a program to create a Python 2D array as input to your dataframe. Display the dataframe.

2. Write a program to create a Python series as input to your dataframe. Display the dataframe.

3. Write a program to create a Python dataframe as input to your dataframe. Display the dataframe.

4. Write a program to create a Python list as input to your dataframe. Display the dataframe.

5. Write a program to create a Python tuple as input to your dataframe. Display the dataframe.

4.2 Perform sorting operations on python dataframe

Write a program to sort a dataframe based on the column ‘category’ from the dataset shopping_trends.csv available on https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset.

Hints

```
# import the necessary libraries

# Create a dataframe from the given csv file

# Sort and display the dataframe based on the column
```

Try

1. Write a program to rearrange the columns from the given order in the dataframe.

2. Write a program to obtain alternate rows from a dataframe.

3. Write a program to obtain alternate rows from a dataframe.

4. Write a program to obtain a new dataframe to the existing dataframe, To apply function to every cell of dataframe.

4.3 Perform manipulation operations on python dataframe

Write a program to filter the rows from the above dataframe.

Hints

```
# import the necessary libraries

# Create a dataframe from the given dataset

# Display the filtered rows
```

Note: Compare the original rows with filtered rows.

Try

1. Write a program to add a new column 'Customer Type' on the above dataframe.

2. Write a program to remove a column 'Customer Type' on the above dataframe.

4.4 Perform grouping and aggregation operations on python dataframe

Write a program for performing aggregation by grouping by a column 'Frequency of Purchases' and calculate sum and mean of 'Review Rating' column from the above dataframe.

Hints

```
# import the necessary libraries

# Create a dataframe from the given dataset

# Perform grouping by a column

# Display the aggregated sum
```

Note: Compare the original columns with grouped columns.

Try

1. Write a program for performing aggregation by grouping by a column 'Frequency of Purchases' and calculate count, min and max of 'Review Rating' column from the above dataframe.

2. Write a program for performing aggregation by grouping by a column 'Subscription Status' and calculate sum and mean of 'Purchase Amount (USD)' column from the above dataframe.

3. Write a program for performing aggregation by grouping by a column 'Subscription Status' and calculate count, min and max of 'Purchase Amount (USD)' column from the above dataframe.

5. Exercises on Data Frames-II

5.1 Perform merging and joining operations on dataframe

Write a program to merge two dataframe based on the common column ‘id’ using the two dictionaries from the following data:

data1 = {'id': [1, 2, 3], 'name': ['Alice', 'Bob', 'Carol'], 'age': [25, 30, 35]}

data2 = {'id': [1, 2, 4], 'occupation': ['Software Engineer', 'Data Scientist', 'Manager'], 'salary': [100000, 120000, 150000]}

Hints

```
# import the necessary libraries

# Create two dataframe from the given dictionary

# Display the merged dataframe
```

Note: Perform outer merge.

Try

1. Write a program to merge two dataframe based on the common column ‘id’ using the two dictionaries to perform inner merge.

2. Write a program to perform left merge two dataframe based on the multiple common column ‘id’ and ‘name’ using the two dictionaries from the following data:

data1 = {'id': [1, 2, 3], 'name': ['Alice', 'Bob', 'Carol'], 'age': [25, 30, 35]}

data2 = {'id': [1, 2, 4], 'name': ['Alice', 'Bob', 'Carol'], 'occupation': ['Software Engineer', 'Data Scientist', 'Manager'], 'salary': [100000, 120000, 150000]}

3. Write a program to perform right merge two dataframe based on the multiple common column ‘id’ and ‘occupation’ using the two dictionaries from the above data.

5.2 Perform reshaping operations on dataframe

Write a program to use reshape operations like pivot data from the dataset fatalities_isr_pse_conflict_2000_to_2023.csv from https://www.kaggle.com/datasets/willianoliveiragibin/fatalities-in-the-israeli-palestinian.

Hints

```
# import the necessary libraries

# Create one dataframe from the given dataset

# Display the pivot data
```

Try

1. Write a program to use reshape operations like stack data on the above dataframe.

2. Write a program to use reshape operations like unstack data on the above dataframe.

3. Write a program to use reshape operations like melt data on 'name' column for value column 'event_location' from the above dataframe.

6. Exercises on Data Exploration

6.1 Explore data with conditioning

Write a program to access and print the first five rows of the dataframe from the dataset heart_attack_prediction_dataset file of https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset.

Hints

# import the necessary libraries

# Access the first five rows of the above dataframe

Hints

Head() method can be used to access and print the first five rows of a dataframe. print() can be used to display the data structure.

Write a program to select the rows where the 'Heart Rate' is normal for an adult on the above dataframe.

# import the necessary libraries

# select rows where the 'Heart Rate' is normal for an adult on the above dataframe.

Hints

The condition dataframe ['Heart Rate '] >= 60 & dataframe ['Heart Rate '] >= 100 can be applied on dataframe.

1.

2. Write a program to where the 'Heart Rate' is low for an adult on the above dataframe.

3. Write a program to where the 'Heart Rate' is high for an adult on the above dataframe.

4. Write a program to delete the 'Age' column from the dataframe.

5. Write a Pandas program to display a data type information about above specified dataframe.

6.2 Summary statistics of the data

Write a program to display the statistical summary of a dataframe.

Hints

df.describe ( ) can be used to display statistical summary of a dataframe.

1. Write a program to count the number of rows and columns of a dataframe.

2. Write a program to display the statistical summary by data type of above dataframe.

3. Write a Pandas program to display a summary of the basic information about above specified dataframe and its data.

4. Write a program to display the statistical summary for specific groups 'Sex' column of above dataframe.

5. Write a program to generate specific percentiles of above dataframe.

6. Write a program to use mean(), sum(), min(), max(), and std() for specific column-wise calculations on above dataframe.

7. Write a program to use mean(), sum(), min(), max(), and std() for specific column-wise calculations i.e. on 'Heart Rate', 'Income', 'Exercise Hours Per Week', 'BMI' and 'Triglycerides' columns on above dataframe.

6.3 Identification of outliers in the data

Write a program to identify outliers using z-score method on above dataframe.

Hints

stats.zscore ( ) function can be used to identify outliers of a dataframe.

1. Write a program to identify outliers using IQR (Interquartile Range) Method on above dataframe.

2. Write a program to identify outliers using Tukey's Method on above dataframe.

7. Exercises on Data Preprocessing – Handling missing values

7.1 Handling missing values - Removing

Write a program to create a dataframe 'data' containing the data from the following dataset "https://raw.githubusercontent.com/selva86/datasets/master/Cars93_miss.csv".

Hints

```
# Identify missing values in the dataframe using '.isnull()' method
```

Write a program and consider given a dataframe 'data' with missing values, use the .dropna() method to remove all rows containing missing values. Display the resulting dataframe. Also verify the same by printing the summary reports. Also print the rest rows and columns summary.

Hints

```
#use the ‘.dropna()’ method to remove all rows

Write code
```

Try

1. Write a program and consider the above dataframe with missing values, calculate the percentage of missing values in each column and create a summary table showing the columns with the highest percentage of missing values.

2. Write a program and consider the above dataframe, calculate and display the number of missing values in each column individually. Sort the columns by the number of missing values in descending order.

3. Write a program and consider the above dataframe, calculate and display the number of missing values in each row individually. Sort the rows by the number of missing values in ascending order.

4. Write a program and consider above dataframe with missing values, remove all columns containing missing values. Display the resulting dataframe.

5. Write a program and consider above dataframe, remove rows that have more than a specified number of missing values. Display the resulting dataframe.

6. Write a program to create a summary report that includes information about the number of rows and columns removed due to missing values after removal operations.

7.2 Handling missing values - Imputation

Write a program to impute the missing values in the numerical column of the given dataframe ‘data’ with the mean of the non-missing values.

Hints

```
# Sample Dataframe with missing values

# Impute missing values with the mean
```

Try

1. Write a program and consider given dataframe with missing values in a numerical column, impute the missing values in that column with the median of the non-missing values.

2. Write a program and consider given dataframe with missing values in a numerical column, impute the missing values in that column with the median of the non-missing values.

3. Write a program and consider given dataframe with missing values, impute missing values in a specific column with a constant value of your choice (e.g., 0).

7.3 Handling missing values - Interpolation

Write a program and consider the given dataframe with missing values in a numerical column, use linear interpolation to impute the missing values in that column.

Hints

```
# import libraries

# Perform linear interpolation on the 'Value' column using ‘.interpolate()’ method
```

Try

1. Write a program to define a custom interpolation function that takes into account domain-specific knowledge or specific data characteristics to impute missing values in a dataframe.

2. Write a program and consider above dataframe with time-ordered data and a numerical column containing missing values, use interpolation techniques that consider the time steps between observations to impute missing values more accurately.

3. Write a program and consider given dataframe with time-ordered data and a numerical column containing missing values, impute the missing values in that column. Ensure that missing values are filled with the previously available value in the column. Perform Forward Fill using forward fill (.ffill()) method

4. Write a program and consider given a dataframe with multiple columns, use forward fill to impute missing values in specific columns of your choice while keeping other columns unaffected.

5. Write a program to create a summary report that includes information about the number of missing values before and after forward fill and backward fill operations, as well as any specific patterns or trends observed.

8. Exercises on Time Series Data

8.1 Time series data preparation

Write a program to load a time series dataset daily-minimum-temperatures-in-me.csv into a pandas dataframe from the https://www.kaggle.com/datasets/shenba/time-series-datasets. Preprocess the data by setting the datetime column as the index and ensuring that it's in the correct datetime format.

Hints

```
#import libraries

# preprocess the data by setting the datetime column as the index in correct format using ‘.to_datetime()’ method
```

Try

1. Write a program and load a time series dataset with irregular time intervals (timestamps) into a pandas dataframe. Resample the data to have a regular time interval (e.g., daily, hourly) and fill any missing data with appropriate values (e.g., forward fill, backward fill, interpolation) on the following data:

data = {'Timestamp': ['2022-01-01 08:00:00', '2022-01-01 10:30:00', '2022-01-02 09:15:00'], 'Value': [10, 15, 20]}

2. Write a program and load above time series dataset that includes time zone information. Convert the timestamps to a common time zone (e.g., UTC) in pandas.

3. Write a program and load a time series dataset that exhibits seasonality (e.g., monthly sales data). Create additional columns to represent the year, quarter, month, day of the week, or any other seasonal components to facilitate seasonal analysis on the following data:

data = {'Date': ['2022-01-15', '2022-02-20', '2022-03-10', '2022-04-05', '2022-05-18'], 'Sales': [1000, 1200, 800, 1100, 1500]}

8.2 Time series aggregation

Write a program and load a daily time series dataset and aggregate it to monthly intervals. Calculate summary statistics such as the monthly mean, sum, minimum, and maximum values for the following data:

data = {'Date': pd.date_range(start='2022-01-01', end='2022-12-31', freq='D'), 'Value': [10, 15, 20, 18, 25, 30] * 61}   # Six values repeated for each day in 2022

Hints

```
# Create a dataframe from the daily time series data

# Aggregate to monthly intervals and calculate summary statistics using ‘.agg() method
```

Try

1. Write a program and load an hourly time series dataset and aggregate it to daily intervals. Calculate the daily total, average, or maximum values for the following data:

data = {'Timestamp': pd.date_range(start='2022-01-01', end='2022-01-03', freq='H'), 'Value': [10, 15, 20, 18, 25, 30, 22, 28, 35, 40, 32, 38, 45, 50, 42, 48, 55, 60]}

2. Write a program and load a weekly time series dataset and aggregate it to monthly intervals. Calculate monthly summaries, such as the total sales for each month.

3. Write a program and load a time series dataset as following with seasonal patterns (e.g., daily temperature readings). Aggregate the data to a higher level of granularity (e.g., monthly) while preserving important seasonal information.

data = {'Date': pd.date_range(start='2022-01-01', end='2022-12-31', freq='D'), 'Temperature': [30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52] * 30 # Daily temperature values (seasonal pattern)}

9. Exercises on working with Data Tables

9.1 Exercises on joining data tables

Write a program to join two data tables along rows from the dataset as used two times ‘https://github.com/17rsuraj/data-curious/blob/master/TowardsDataScience/Dummy_course_data.xlsx’.

Hints

```
import matplotlib library
```

```
# Use ‘join( )’ function to add two data tables
```

Try

1. Write a program to join two data tables along columns from the dataset using ‘merge( ) function.

2. Write a program to join two data tables along rows from the dataset using ‘join( ) function.

3. Write a program to join two data tables along columns from the given dataset using ‘merge( )’ function..

9.2 Exercises on pivot tables

Write a program to create a Pivot table with multiple indexes from the data set of “https://github.com/ven-27/datasets/blob/master/titanic.csv” file.

Hints

```
# Import necessary libraries

# Create a dataframe

# Use ‘pivot_table( )’ function to create a pivot table
```

Try

1. Write a program to create a Pivot table on the given dataset and find survival rate by gender.

2. Write a program to create a Pivot table on the given dataset and find survival rate by gender, age wise of various classes.

3. Write a program to create a Pivot table on the given dataset and calculate number of women and men were in a particular cabin class.

9.3 Exercises on contingency tables

Write a program to make a contingency table between the ‘Age’ and ‘Survived’ columns on the above dataset.

Hints

```
# Import necessary libraries

# Use the ‘crosstab( )’ function to create contingency table
```

Try

1. Write a program to make a contingency table between the ‘Pclass’ and ‘Sex’ columns on the above dataset.

2. Write a program to make a contingency table between the ‘PassengerId’ + ‘Sex’ and ‘Survived’ columns on the above dataset.

9.4 Exercises on grouping data

Write a program to split the above dataset into groups based on passenger class and also check the type of GroupBy ( ) object.

Hints

```
# use the ‘requests’ and the ‘Beautiful Soup’ libraries for web scraping and data extraction
```

Try

1. Write a program to split the above dataframe into groups based on survived and passenger class.

2. Write a program to split the above dataframe into groups based on single column and multiple columns. Find the size of the grouped data.

10. Exercises on Web Scraping

10.1 Scraping a list of items from a website

Write a program to scrape a list of articles from a news website https://ywepaper.thehindu.com/reader.

Hints

```
# use the ‘requests’ and the ‘Beautiful Soup’ libraries for web scraping and data extraction
```

Try

1. Write a program to scrape a List of Products from an E-commerce Website https://www.amazon.in/?&ext_vrnc=hi&tag=googhydrabk1-21&ref=pd_sl_7hz2t19t5c_e&adgrpid=58355126069&hvpone=&hvptwo=&hvadid=610644601173&hvpos=&hvnetw=g&hvrand=2271425446877080510&hvqmt=e&hvdev=c&hvdvcmdl=&hvlocint=&hvlocphy=9062186&hvtargid=kwd-10573980&hydadcr=14453_2316415.

2. Write a program to scrape a List of Job Postings from a Job Search Website https://www.naukri.com/engineering-jobs?src=discovery_trendingWdgt_homepage_srch.

3. Write a program to scrape a List of Books from an Online Bookstore https://www.bookswagon.com/.

10.2 Scraping data from a table

Write a program to scrape data from a table on a world health organization website:
https://rho.emro.who.int/Indicator/TermID/46.

Hints

```
# use the ‘requests’ and the ‘Beautiful Soup’ libraries for web scraping and data extraction
```

Try

1. Write a program to scrape a Table with Headers on an above world health organization website.

2. Write a program to scrape a Table with Pagination on an above world health organization website.

10.3 Scraping images from a website

Write a program to locate the all images URLs in the webpage's HTML to do scraping from the above website.

Hints

```
# use the ‘requests’ and the ‘Beautiful Soup’ libraries for web scraping and data extraction
```

Try

1. Write a program to scrape Images from a Gallery on the world health organization website https://www.emro.who.int/media/multimedia/posters.html.

2. Write a program to scrape Images from a Search Results Page as https://www.emro.who.int/search/en/index.htm?q=images.

10.4 Scraping data with pagination

Write a program to scrape data with pagination from a website involves navigating through multiple pages to collect all the desired information.

Hints

```
# Use the ‘requests’ and the ‘Beautiful Soup’ libraries for web scraping and data extraction

# Set the number of pages to scrape

# Print the scraped data
```

Try

1. Write a program to scrape Data with AJAX Pagination.

2. Write a program to scrape Images from a Search Results Page as https://www.emro.who.int/search/en/index.htm?q=images.

3. Write a program to implement error handling to gracefully handle HTTP errors, such as 404 or 503, and exceptions that may occur during scraping, such as timeouts.

10.5 Scraping with user-agent

Write a program to scrape a webpage with a custom User-Agent as:

headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/58.0.3029.110 Safari/537.36'}

Hints

```
# use the ‘requests’ library to set a User-Agent header to simulate a web browser request and avoid being blocked
```

Try

1. Write a program to modify the program to use a different User-Agent string. You can find various User-Agent strings online, or you can generate your own to mimic different browsers or devices.

2. Write a program to Extend the program to scrape a specific page on the website with the custom User-Agent header.

3. Write a program to modify the program to scrape multiple pages on the website with the custom User-Agent header. You can use a loop to iterate through different URLs.

4. Write a program to create a list of User-Agent strings and randomly select one for each request. This exercise simulates a more dynamic scraping behavior.

11. Exercises on Visualization-I

11.1 Exercises on `line plot`

Write a program to plot a line plot on the following dataset https://gist.githubusercontent.com/seankross/a412dfbd88b3db70b74b/raw/5f23f993cd87c283ce766e7ac6b329ee7cc2e1d1/mtcars.csv.

Hints

```
# Use the Matplotlib library

# The method plt.plot(x, y) can be use to plot the line plot
```

Try

1. Write a program to line plot customizing Line Styles and Markers**.**

2. Write a program to line plot adding Grid Lines.

3. Write a program to line plot adding Logarithmic Scale on the Y-axis.

11.2 Exercises on `scatter plot`

Write a program to plot scatter plot to visualize the relationship between two variables on the following dataset "https://github.com/CoreyMSchafer/code_snippets/blob/master/Python/Matplotlib/07-ScatterPlots/2019-05-31-data.csv".

Hints

```
# Use the Matplotlib library

# The method plt.scatter(x, y) can be use to plot the line plot
```

Try

1. Write a program to plot a scatter Plot with Color and Marker Customization.

2. Write a program to plot a scatter Plot with Size and Transparency.

3. Write a program calculate correlations between variables to quantify the strength and direction of a relationship between the variables.

11.3 Exercises on bar chart

Write a program to plot bar plot to visualize the relationship between two variables on the following dataset "https://github.com/CoreyMSchafer/code_snippets/blob/master/Python/Matplotlib/07-ScatterPlots/2019-05-31-data.csv".

Hints

```
# Use the Matplotlib library

# Create a basic vertical bar chart using method 'plt.bar(categories, values)' with parameters
```

Try

1. Write a program to plot a vertical bar plot on the above data.

2. Write a program to plot a grouped bar plot on the above data.

3. Write a program to plot a bar plot on the above data using seaborn library.

11.4 Stacked bar plot

Write a program to plot Stacked bar plot to visualize the relationship between two variables on the following dataset "https://github.com/CoreyMSchafer/code_snippets/blob/master/Python/Matplotlib/07-ScatterPlots/2019-05-31-data.csv".

Hints

```
# Use the Matplotlib library

# Create a basic stacked bar chart using method 'plt.bar(categories, values1)' and 'plt.bar(categories, values2, bottom=values1)' parameters
```

Try

1. Write a program to plot a horizontal stacked bar plot on the above data using Matplotlib library.

2. Write a program to plot a horizontal stacked bar plot on the above data using seaborn library.

12. Exercises on Visualization-II

12.1 Exercises on histogram

Write a program to plot a Histogram for visualizing the distribution of a dataset using Matplotlib library on the following dataset https://gist.githubusercontent.com/seankross/a412dfbd88b3db70b74b/raw/5f23f993cd87c283ce766e7ac6b329ee7cc2e1d1/mtcars.csv.

Hints

```
# Use the Matplotlib library

# Create a histogram with bin parameters in ‘plt.hist(data, bins)’ method
```

Try

1. Write a program to plot a histogram with different Customization.

2. Write a program to plot a Stacked Histograms on the above data points.

12.2 Exercises on density plots

Write a program to plot a density plot on given the dataset ‘car_crashes’ in which is the most common speed due to which most of the car crashes happened.

Hints

```
# Use the Matplotlib & seaborn library

# Create a histogram with bin parameters in ‘plot.density()’ method
```

Try

1. Write a program to plot a density plot on given the dataset ‘tips’ and calculate what was the most common tip given by a customer.

2. Write a program to plot a density plot on given the dataset ‘tips’ and calculate what was the most common tip given by a customer using plot.kde( ) function.

12.3 Exercises on box plots

Write a program to plot a box plot on given the dataset 'https://github.com/ven-27/datasets/blob/master/titanic.csv' in which whether the person has survived or not during the sink of titanic and different details of the person.

Hints

```
# Use the Matplotlib & seaborn library

# Create a box plot with bin parameters in 'plt.boxplot( )' method
```

Try

1. Write a program to plot a box plot of distribution of survived based on the age of the passenger on given the dataset.

2. Write a program to box plot of distribution of survived based on the Class of the passenger on given the dataset.

12.4 Exercises on pair plots

Write a program to generate pair plots for pairwise relationship of variables on the 'penguins' dataset from seaborn module.

Hints

```
# Use the seaborn library

# Create a pair plot using 'seaborn.pairplot( )' method
```

Try

1. Write a program to generate pair plots for pairwise relationship of variables on the above dataset to assigning a hue variable as 'species'.

2. Write a program to generate pair plots for pairwise relationship of variables on the above dataset to assigning the kind parameter as 'kde' to draw KDEs using kdeplot( ) funtion.

3. Write a program to generate pair plots for pairwise relationship of variables on the above dataset to assigning a vars value as x_vars and y_vars to select the variables to plot.

13. Final Notes

Dealing with the raw data which has no longer been accessible and cannot be utilized should be processed to use efficiently. Then here comes data-wrangling which helps to turn non-resourceful (raw) data into valuable data which in turn returns valuable information. Each step in the process of wrangling the data is to the best possible analysis. The outcome is expected to generate a robust and reliable analysis of the data.

The problems in this tutorial are certainly NOT challenging. Check out these sites:

- The Springboard- (https://www.springboard.com/blog/data-science/data-wrangling/)
- Educative - https://www.educative.io/courses/data-wrangling-with-python/coding-challenges-on-standardization
- iMocha - https://www.imocha.io/tests/data-wrangling-with-python-test
- Kaggle - https://www.kaggle.com/code/prakharrathi25/data-wrangling
- Github - https://moderndive.github.io/moderndive_labs/static/PS/PS03_data_wrangling.html
- Rpubs - https://rpubs.com/Rokshana_Data_knowledge/891691
- Towardsdatascience - https://towardsdatascience.com/conquer-the-python-coding-round-in-data-science-interviews-5e27c4513be3
- Into the minds - https://www.intotheminds.com/blog/en/the-11-challenges-of-data-preparation-and-data-wrangling/
- Posit - https://posit.co/blog/wrangling-unruly-data/
- Futurelearn - https://www.futurelearn.com/info/courses/data-analytics-python-data-wrangling-and-ingestion/0/steps/186670

Student must have any one of the following certification:

1. Coursera - Data Wrangling, Analysis and AB Testing with SQL
2. Coursera - Data Wrangling with Python Specialization
3. Coursera - Practical Data Wrangling with Pandas
4. Coursera - Fundamental Tools of Data Wrangling
5. GeeksforGeeks – Data Wrangling in Python
6. Stanford – Data Wrangling
7. Udemy – Data Wrangling with Python
8. Pluralsight - Data Wrangling with Python 3
9. Harvard University - Data Science: Wrangling
10. Udacity - Advanced Data Wrangling

V. TEXT BOOKS:

1. Eric Jacqueline Kazil & Katharine Jarmul, "Data Wrangling with Python", O'Reilly Media, Inc,2016.

VI. REFERENCE BOOKS:

1. Dr. Tirthajyoti Sarkar, Shubhadeep, "Data Wrangling with Python: Creating actionable data from raw sources", Packt Publishing Ltd,2019.
2. Wes McKinney, "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython", O'Reilly, 2nd Edition,2018.
3. Jake VanderPlas, "Python Data Science Handbook: Essential Tools for Working with Data", O'Reilly, 2017.
4. Y. Daniel Liang, "Introduction to Programming using Python", Pearson,2012.

VII. ELECTRONICS RESOURCES

1. https://www.dataquest.io/blog/sci-kit-learn-tutorial/
2. https://www.ibm.com/support/knowledgecenter/en/SS3RA7_sub/modeler_tutorial_ddita/modeler_tutorial_ddita-gentopic1.html
3. https://archive.ics.uci.edu/ml/datasets.php
4. https://www.edx.org/course/analyzing-data-with-python
5. http://math.ecnu.edu.cn/~lfzhou/seminar/[Joel_Grus]_Data_Science_from_Scratch_First_Princ.pdf
6. https://www.programmer-books.com/introducing-data-science-pdf/

VIII. MATERIALS ONLINE (Include full stack)

1. Course template
2. Lab Manual