TITLE	Data Wrangling I
PROBLEM STATEMENT/ DEFINITION	Perform the following operations using Python on any open-source dataset (e.g., data.csv) 1. Import all the required Python Libraries. 2. Locate an open-source data from the web (e.g. https://www.kaggle.com). Provide a clear description of the data and its source (i.e., URL of the web site). 3. Load the Dataset into pandas' data frame. 4. Data Preprocessing: check for missing values in the data using pandas isnull (), describe() function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame. 5. Data Formatting and Data Normalization: Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set. If variables are not in the correct data type, apply proper type conversions. 6. Turn categorical variables into quantitative variables in Python. In addition to the codes and outputs, explain every operation that you do in the above steps and explain everything that you do to import/read/scrape the data set.
OBJECTIVE	1. To do pre-processing on the given dataset. 2. To provide initial statistics 3. Data formatting and type conversions on the dataset columns.
S/W PACKAGES AND HARDWARE APPARATUS USED	S/W- Jupyter Notebook/ Weka/ Python OS-LINUX 64 bit OS H/W: Core 2 DUO/i3/i5/i7 64-bit processor
REFERENCES	1. CHIRAG SHAH, "A HANDS-ON INTRODUCTION TO DATA SCIENCE",ISBN 978-1-108-47244-9 2. Wes McKinney and the Pandas Development Team, "Pandas: powerful Python data analysis toolkit" 3. https://pandas.pydata.org/
STEPS	**Refer to student activity flow chart if found necessary by subject teacher and relevant to the subject manual. Describe steps only.**
INSTRUCTIONS FOR WRITING JOURNAL	1. Title 2. Problem statement 3. Learning objective 4. Learning outcome 5. Theory (includes methods, libraries and functions, 6. Analysis (as per assignment), 7. Conclusion.

1. Title: Data Wrangling I

2. Problem statement:

Perform the following operations using Python on any open-source dataset (e.g., data.csv)
1. Import all the required Python Libraries.
2. Locate an open-source data from the web (e.g. https://www.kaggle.com). Provide a clear description of the data and its source (i.e., URL of the web site).
3. Load the Dataset into pandas' data frame.
4. Data Preprocessing: check for missing values in the data using pandas isnull (), describe() function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame.
5. Data Formatting and Data Normalization: Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set. If variables are not in the correct data type, apply proper type conversions.
6. Turn categorical variables into quantitative variables in Python.

In addition to the codes and outputs, explain every operation that you do in the above steps and explain everything that you do to import/read/scrape the data set.

3. Learning objective:

1. To do pre-processing on the given dataset.
2. To provide initial statistics
3. Data formatting and type conversions on the dataset columns.

4. Learning outcome:

After performing this assignment students will be able to:

- Do preprocessing on the dataset.
- Provide initial statistics of all columns in the dataset.
- Apply data formatting and type conversions on the dataset columns.

5. Theory:

Libraries: Pythhon libraries Pandas

Pandas: **Pandas** is a Python library. **Pandas** is used to analyze data. Pandas allows importing data from various file formats such as comma-separated values, JSON, SQL database tables or queries, and Microsoft Excel.

Methods and functions:

Pandas.read_csv(): to load csv file into Pandas dataframe.

Shape(): Pandas shape function provides the shape of the dataset in terms of number of rows and columns.

dtypes: Pandas dtype attribute provides datatype of each column in the dataset.

Isnull(): to check presence of null values in the dataset.

describe(): The describe() method is **used for calculating some statistical data like percentile, mean and std of the numerical values of the Series or DataFrame**. It analyzes both numeric and object series and also the DataFrame column sets of mixed data types.

Converting column from object type to numerical data type: **pandas.to_numeric() function is used to convert object data type to numeric data type**. This function will try to change non-numeric objects (such as strings) into integers or floating-point numbers as appropriate.
Converting object data type to datetime data type:
Pandas to_datetime()method is used to convert string with date and time to datetime type.

Conversion of categorical feature/column to numeric column:
1. **Label Encoder:** It is used to transform non-numerical labels to numerical labels (or nominal categorical variables). Numerical labels are always between 0 and n_classes-1. This approach is more flexible because it allows encoding as many category columns as you would like and choose how to label the columns using a prefix. Proper naming will make the rest of the analysis just a little bit easier.
2. **Dummy Coding:** Dummy coding is a commonly used method for converting a categorical input variable into continuous variable. ‘Dummy’, as the name suggests is a duplicate variable which represents one level of a categorical variable. Presence of a level is represent by 1 and absence is represented by 0. For every level present, one dummy variable will be created.
3. **One-Hot Encoder:** Though label encoding is straight but it has the disadvantage that the numeric values can be misinterpreted by algorithms as having some sort of hierarchy/order in them. This ordering issue is addressed in another common alternative approach called ‘One-Hot Encoding’. In this strategy, each category value is converted into a new column and assigned a 1 or 0 (notation for true/false) value to the column.

6. Analysis:

Observed missing values in the dataset and deal with them. Observed descriptive statistics of numerical columns in the dataset and the categorical features in the dataset and do conversion.

7. Conclusion:

Using python libraries Pandas dataset shape, datatypes are observed. Dataset is cleaned after finding missing values, descriptive cs are observed, columns data type is changed as per the requirement, categorical columns are converted to numeric type.