Name: Aditya Warjale
Roll no: 31282

Assignment No - 1

- **TITLE**: Data Wrangling I

- **PROBLEM STATEMENT**:

1. Import all the required python libraries.
2. Load the dataset into pandas dataframe.
3. Data Processing - check for missing values in the data using pandas isnull(), describe() functions to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the dataframe.
4. Data formatting and Normalization - Summarize the types of variables by checking the data types (i.e. character, numeric, integer, factor and logical) of the variables in the dataset. If variables are not in the correct data type, apply proper type conversions.
5. Turn categorial variables into quantitative variables in python.

In addition to the code and output, explain every operation that you do in the above steps and explain everything that you do to import/read/scrap the dataset.

- **LEARNING OBJECTIVES**:

1. To understand the constraint of a dataset
2. To understand various libraries and methods used to analyse a dataset.
3. To be able to differntiate between different columns with repect to their datatype, value, missing values, etc.
4. To form relations to improve analytics of the data set.
5. To tackle empty value in a large data set.
6. To graphically display various outcomes.

- LEARNING OUTCOMES:

1. Load the .csv file into a dataframe for further analysis.
2. Process the dataset by finding out the missing data and filling them with mean value of the column.
3. Graphically display the categorical values.

- SOFTWARE AND HARDWARE REQUIREMENTS:

1. H/W – i5 10th Gen, OS - Windows
2. S/W – Jupyter Notebook / VS code - Python

- THEORY:

1. Libraries used:
   a) Panda – used to analyze the dataframe and perform various operation on the data.
   b) matplotlib.pyplot – stats based interface to plot data

2. Reading data:
   eg: – df = pd.read_csv ('data.csv')
   where, df → dataframe variable (2D Labeled data structure with columns of different data types)
   pd → pandas library is imported as 'pd'
   read_csv() → function used to read a .csv file.

3. Dataframe attributes:
   a) df.shape → gives the dimensions of the data set
   b) df.dtypes → gives the datatype of all the columns present in the dataset

4. Dataframe methods:
   a) df.head() → provides the first 5 rows from the dataset
   b) df.tail() → provides the last 5 rows from the dataset
   c) df.describe() → displays count, mean, std deviation, min, max, sum for numeric and float data type columns.
   d) df.info() → displays all the columns alongwith information about the existence of value & datatype

e) df.isnull() → used to find NaN values in the dataset

f) df.isnull().sum() → count of NaN values from each column

5. Replacing null value by using mean value of column.

```
mean = df['BuildingArea'].mean()
df['BuildingArea'].fillna(value = mean, inplace = True)
```

6. Changing datatype of any column.

```
df['Type'] = df['Type'].astype('category')
```

7. Categorial variables into quantitative variables.

```
df['Types codes'] = df,['Type'].cat.codes
```

→ assigns codes to the categorial variable in alphatical order.

8. Graphical Analysis.

```
eg:- plt.hist(df['Type'])
     plt.show()
```

- CONCLUSION:

Hence we learnt how to use pandas to extract data from a csv file and implemented various operations on the dataset.