# PYTHON – WORKSHEET 12 (DATA CLEANING)

**Instructions:** This is a data-cleaning worksheet. You have been given a dataset which includes the Melbourne house prices data. You have to use pandas library for data-cleaning. You have to attempt all the questions in the same order as given so that at the end you arrive at completely cleaned dataset with no missing values. Use jupyter notebook to solve the exercise.

**Questions:**

1. Determine the no. of rows and no. of columns in the dataset.
2. Display the datatypes of each column.
3. Determine the row indexes of rows having all null values (i.e. empty rows).
4. Determine the total number of empty rows and their percentage in whole dataset and remove them.
5. Determine the column names having missing values.
6. Determine the missing values and their percentage in each column.
7. Determine the row indexes having missing values.
8. Determine the missing values and their percentage in each row.
9. Determine the column names having more than 30% missing values and remove them from the dataset.
10. Determine the number and percentage of rows having more than 5 missing values. If the percentage of such rows is less than 20% then remove all these rows.
11. Display the description of data and percentage of missing values still left column wise.
12. Considering the 'Price' column as target variable which is to be predicted, how will you treat the missing values? Give a brief explanation and treat the missing values in 'Price' column.
13. Now check the percentage of missing values in each column again, if missing values exist, treat it with the most suitable method.
14. Now since you have completed your data cleaning, determine the number and percentage of columns removed.
15. Determine the number and percentage of rows removed in whole process. Also comment briefly that according to you the amount of data lost is justified or not.