**HOUSING: PRICE PREDICTION**

Submitted by:

Sampada Sambrani

Internship batch 34

e-mail:sampada.sambrani@gmail.com

**ACKNOWLEDGMENT**

I would like to express my deep and sincere gratitude to FLIP ROBO for giving me the opportunity to do this project. As a great bridge between academic and industry, this program educated me how to perform theoretical methodology in real life. I would like to express my sincere thankfulness to our assigned mentors for the continuous support of our queries, for their patience, enthusiasm, motivation and immense knowledge.

**INTRODUCTION**

The real estate sector is an important industry with many stakeholders ranging from regulatory bodies to private companies and investors. Among these stakeholders, there is a high demand for a better understanding of the industry operational mechanism and driving factors. Today there is a large amount of data available on relevant statistics as well as on additional contextual factors, and it is natural to try to make use of these in order to improve our understanding of the industry. Notably,

A US-based housing company named **Surprise Housing** has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia.

The main steps in this project are:

• Exploratory Data Analysis (EDA). By conducting explanatory data analysis, we obtain a better understanding of our data. This yields insights that can be helpful later when building a model, as well as insights that are independently interesting.

• Feature Selection In order to avoid overfitting issues.

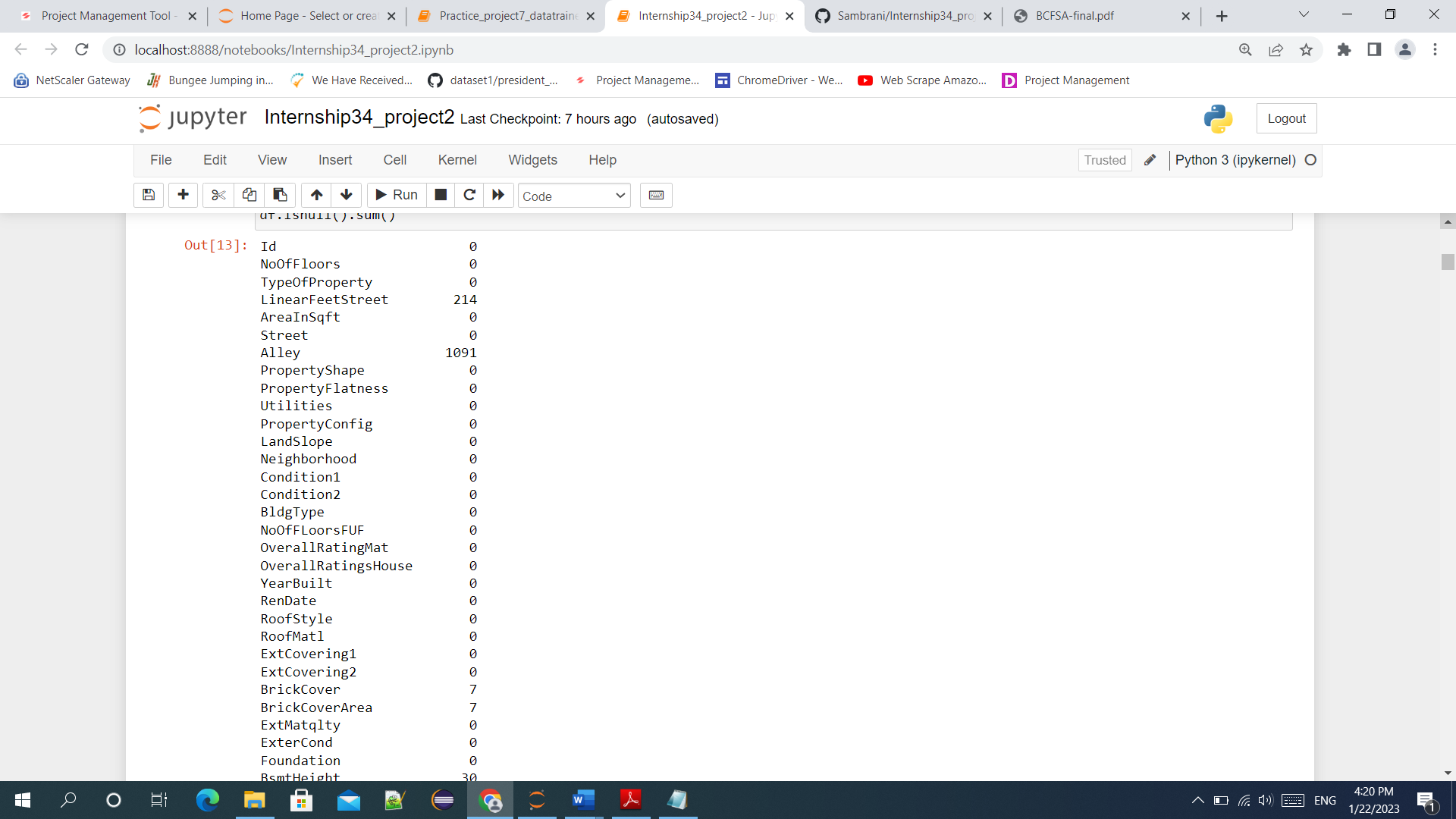
• Modeling We apply LinearRegression and check with the r2score.

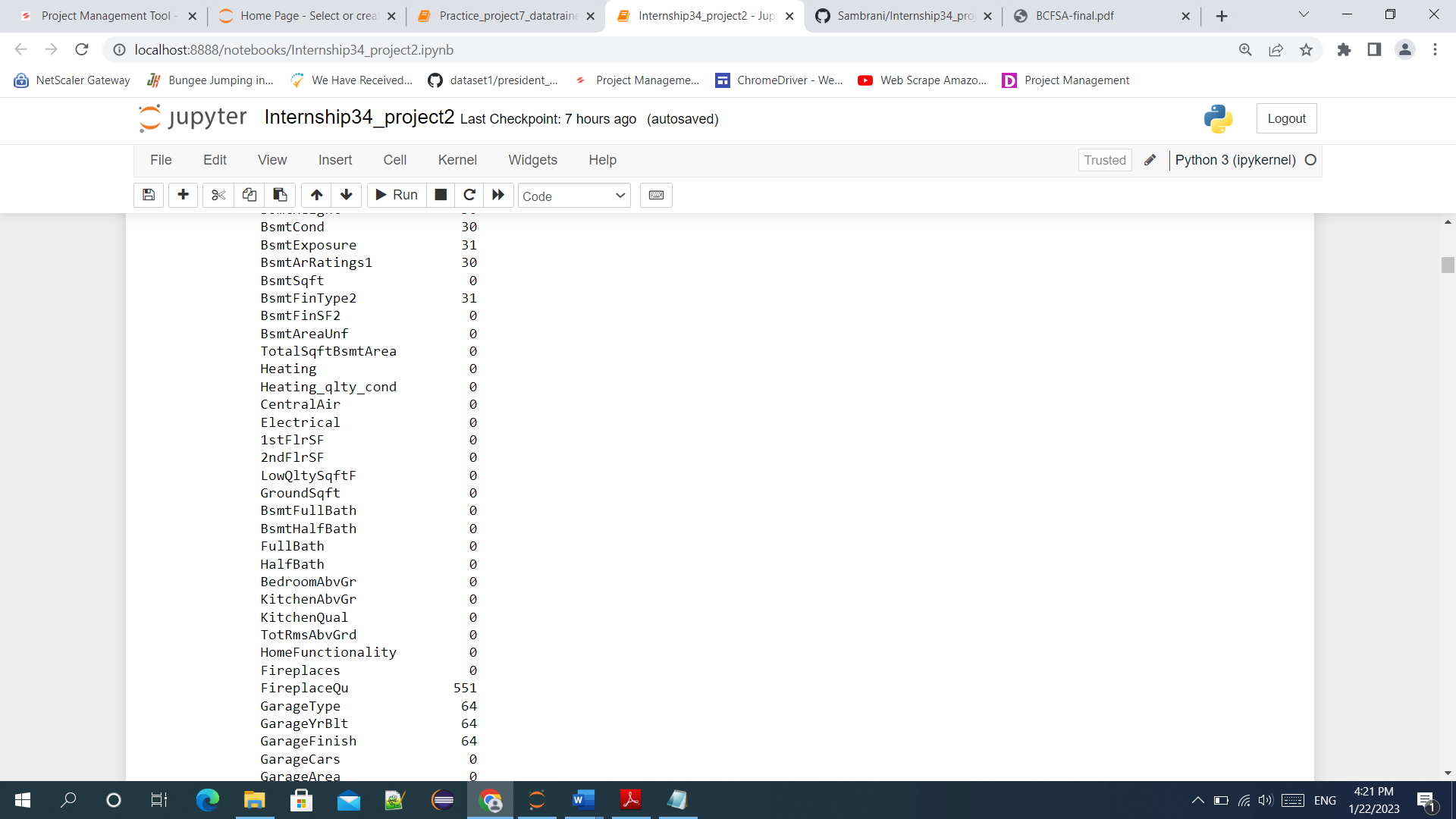
We are required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

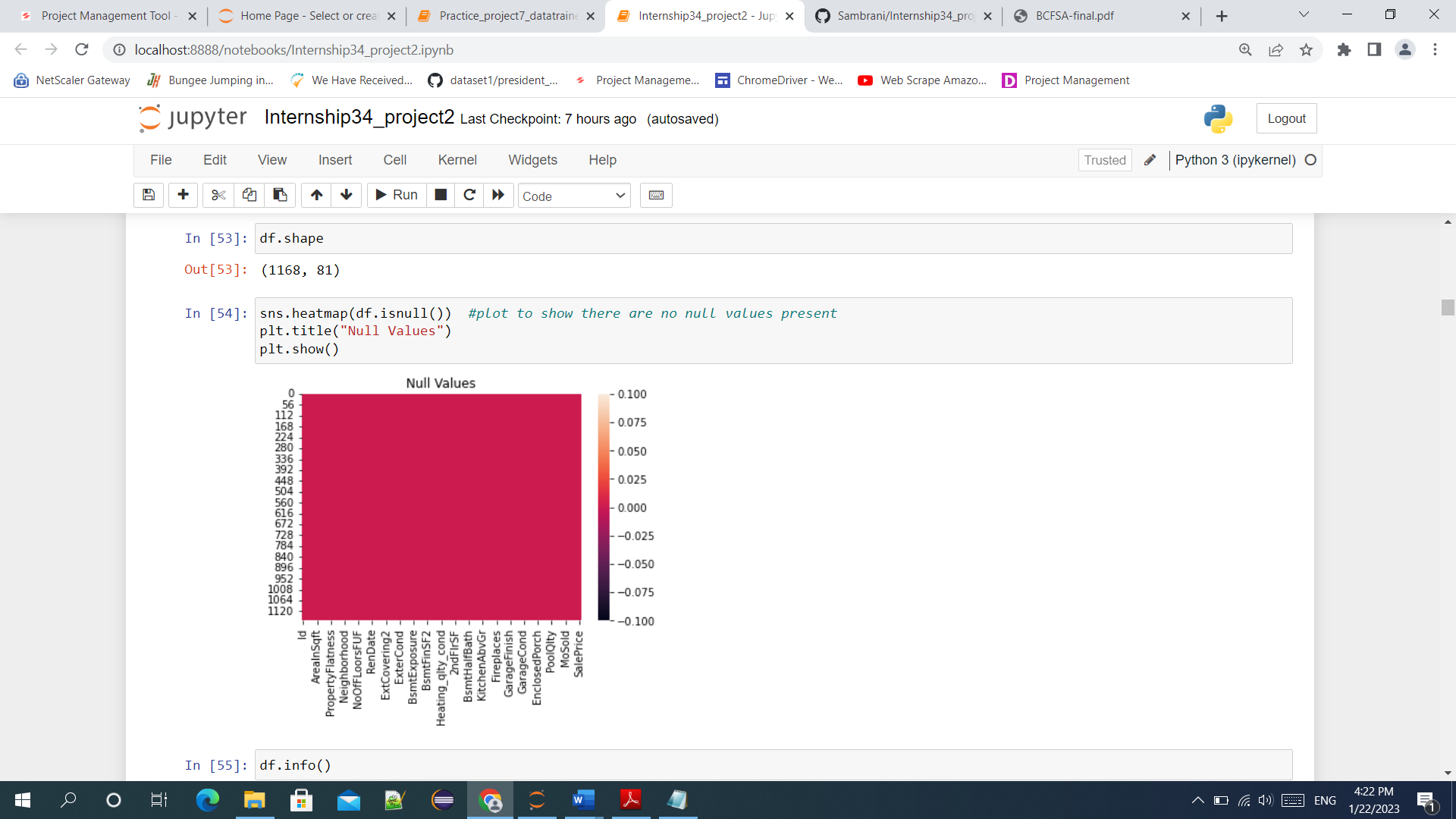
**Analytical Problem Framing**

* Mathematical/Analytical Modelling of the Problem

As our dataset had numerous null values , we first used fillna() to fill the null/nan values in all the columns , by using mean for numeric data columns and mode for categorical data columns as shown below



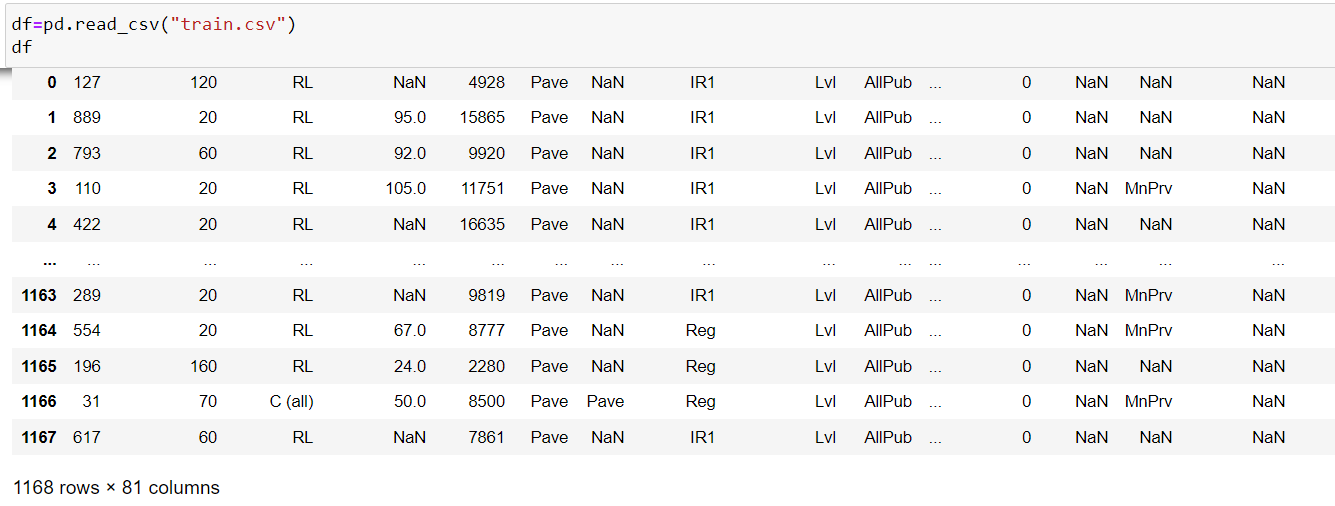


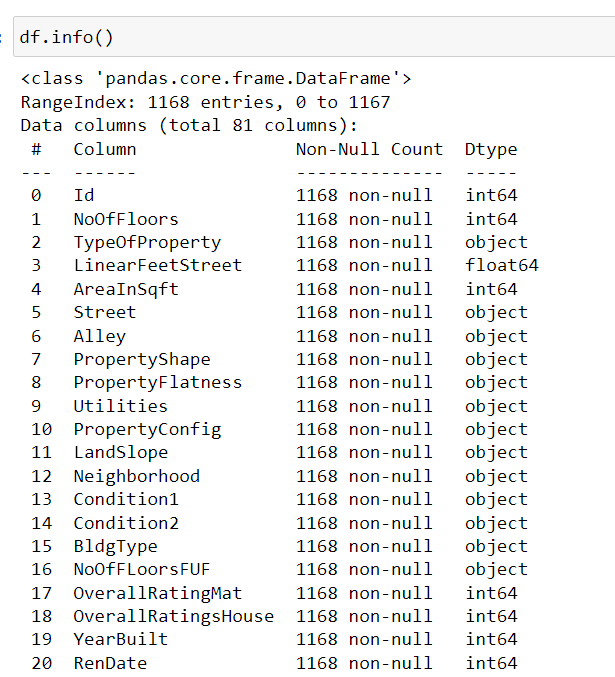


* The statistical summary was obtained by using describe (), which gives some idea of the percentile, mean ,count,standard deviation , min and max values , so that we can get idea whether our data is skewed ,our data is highly spreaded and some knowledge of outliers are present or not / may be present.
* The correlation w.r.t the target variable was checked, to get the features which are positively and negatively correlated with the target variable.We have checked for multicollinearity exist or not using the corr().The acceptable range is <+/-0.7
* The skewness was checked for the feature variables only ,using the skew(). If high skewness is present then we using various transformation techniques to reduce the skewness, and even after transformation if not reduced then we can drop the columns. The acceptable range for skewness is +/-0.5
* Data Sources and their formats

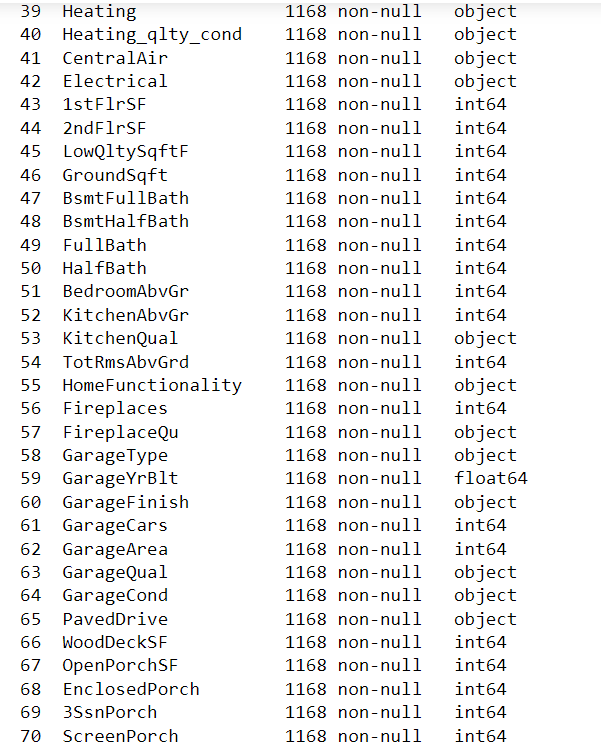
A US-based housing company named **Surprise Housing** has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file.

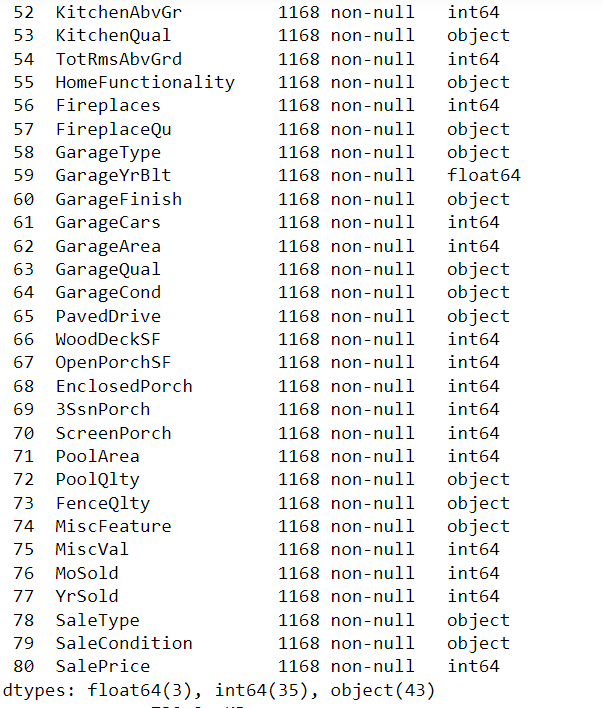
* Two datasets are being provided (test.csv, train.csv). We have to train on train.csv dataset and predict on test.csv file







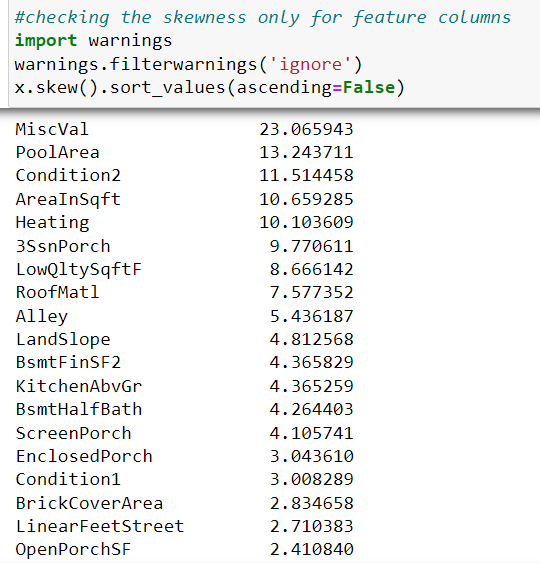


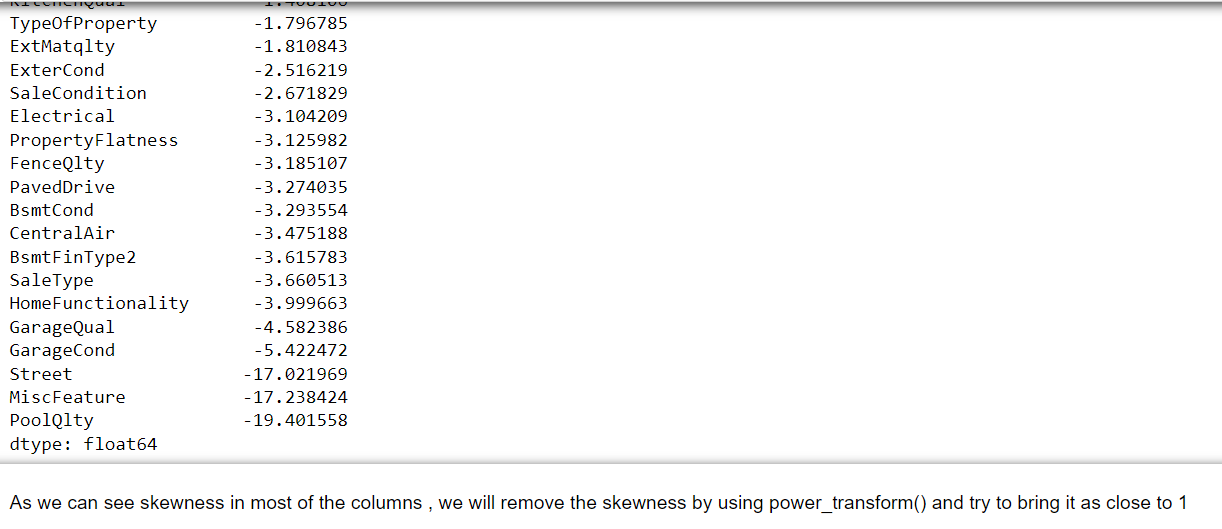


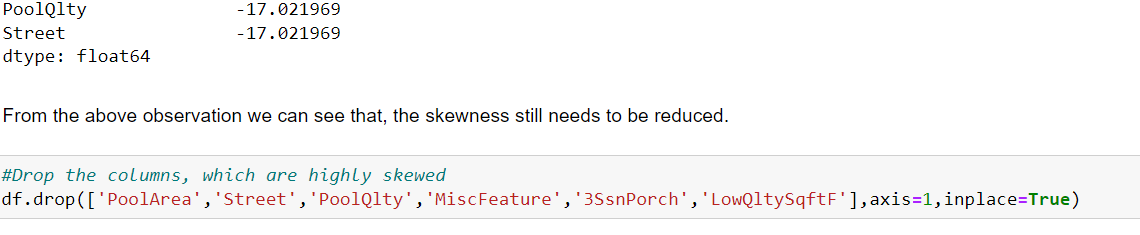
* Data Preprocessing Done

For data cleaning we have firstly filled all the null values with respective mean,median or mode depending on the type of data.

After checking with the skewness and applying power transformation to reduce the skewness , even if the columns had high skewness, we have to drop that columns ( very high negative and positive skewed columns).







* Data Inputs- Logic- Output Relationships

The data inputs and Output relationship in our project was w.r.t the data set provided .

* (test.csv, train.csv). We have to train on train.csv dataset and predict on test.csv file.
* From the input (feature variables) we need to determine:

Which variables are important to predict the price of variable?

How do these variables describe the price of the house?

* If any of the columns are least important in determining the house prices, we can drop from the dataset
* The output data is our predictions made for the house price based on the input data(feature columns), as test.csv consist of only feature columns.
* Upon training the data for feature and target columns in train.csv , based on these assumptions , the machine has predicted the output(house price) on the test.csv.
* State the set of assumptions (if any) related to the problem under consideration

Only assumption made here is , based on the high skewness values for certain column, we have dropped the columns

* Hardware and Software Requirements and Tools Used

**Hardware specification**

Processor: intel CORE i3 (10th gen) minimum

RAM: 2GB and above

Hard disc capacity: Minimum of 100GB

Display type: Standard VGA

**Software specification**

Operating system: Windows 10

Front end : Jupyter framework(anaconda)

Programming tool: Anaconda

Internet browser: Google chrome

**Libraries**

import numpy as np – for numeric algebra

import pandas as pd- for data representation

import sklearn

import matplotlib.pyplot as plt- for data visualization

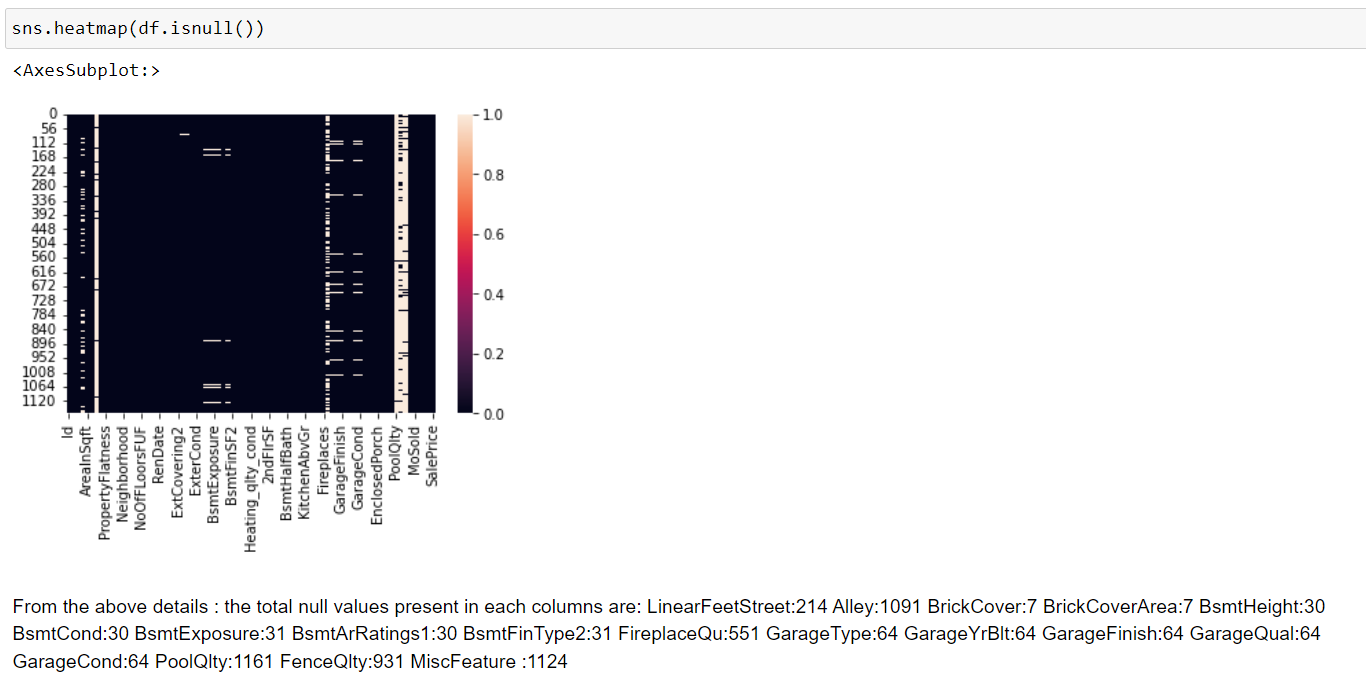
import seaborn as sns

from sklearn.model\_selection import train\_test\_split

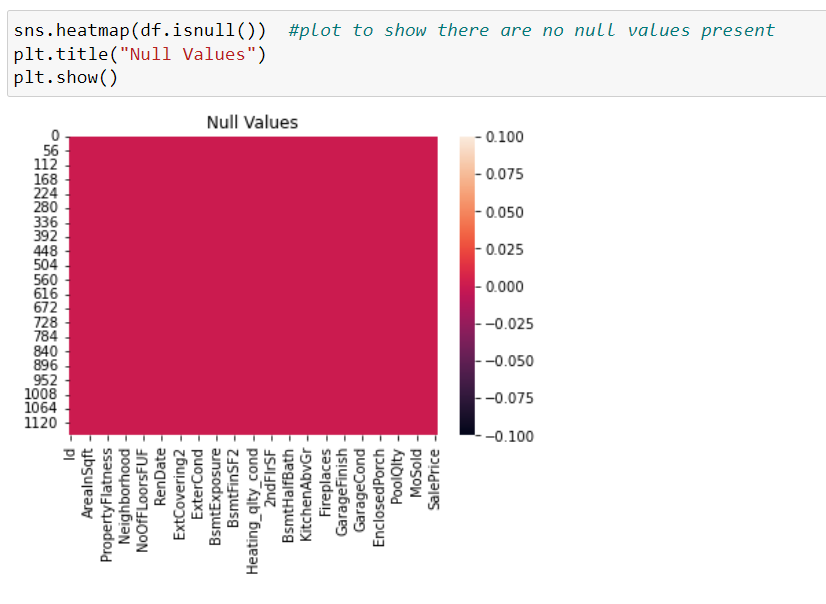
* Scikit-learn is one of the most popular ML libraries for classical ML algorithms. It is built on top of two basic Python libraries, viz., NumPy and SciPy. Scikit-learn supports most of the supervised and unsupervised learning algorithms. Scikit-learn can also be used for data-mining and data-analysis, which makes it a great tool who is starting out with ML.
* SciPy is a very popular library among Machine Learning enthusiasts as it contains different modules for optimization, linear algebra, integration and statistics.
* NumPy is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is very useful for fundamental scientific computations in Machine Learning. It is particularly useful for linear algebra
* Pandas is a popular Python library for data analysis.
* Matplotlib is a very popular Python library for data visualization

**Model/s Development and Evaluation**

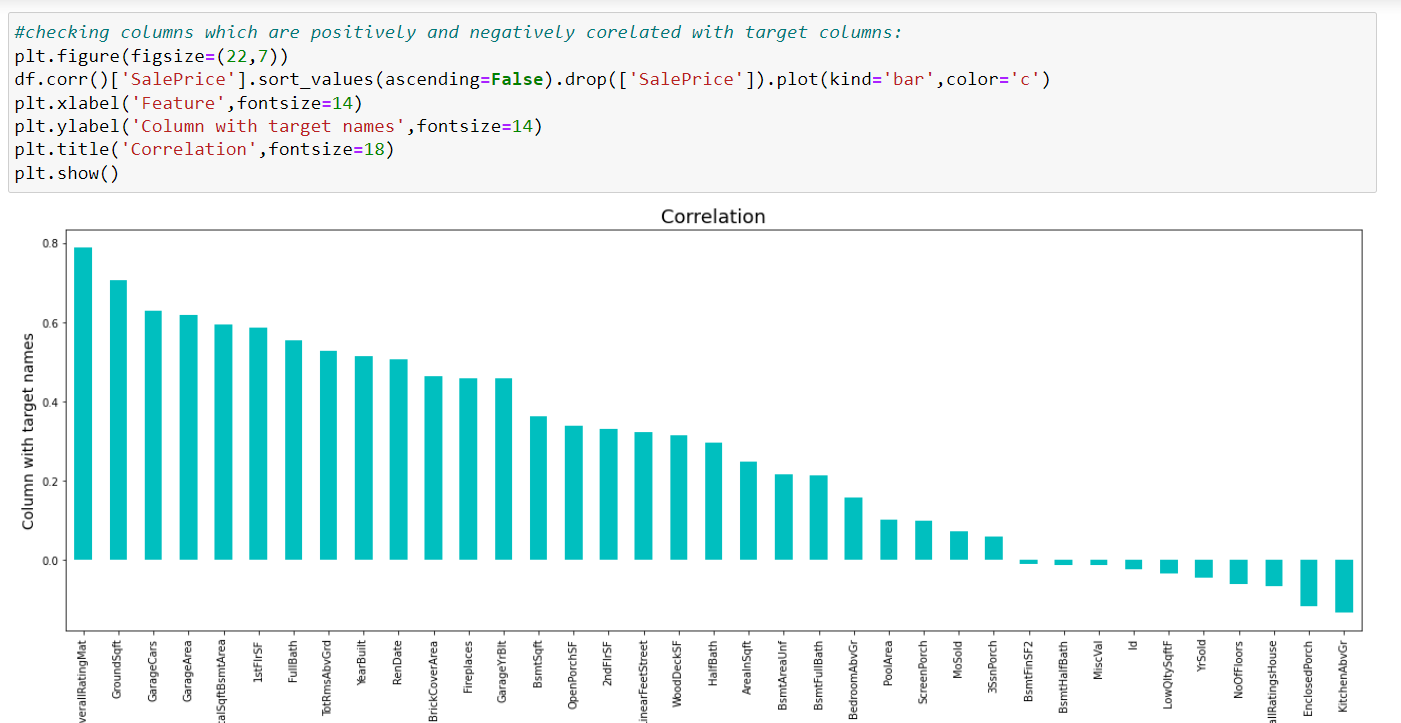
* Our dataset had some null values, as shown in the fig below, these null values were filled by fillna().



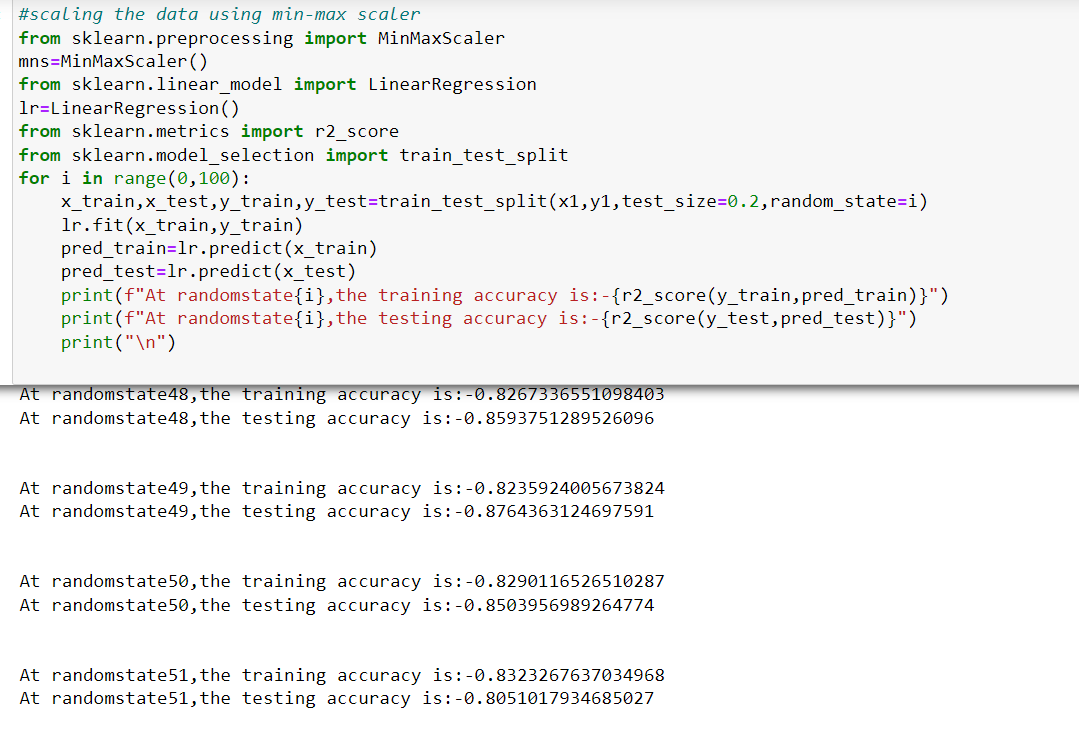
* After the null/nan values were replaced with mean/mode depending on the columns datatype, the fig below shows that there are no null values present.

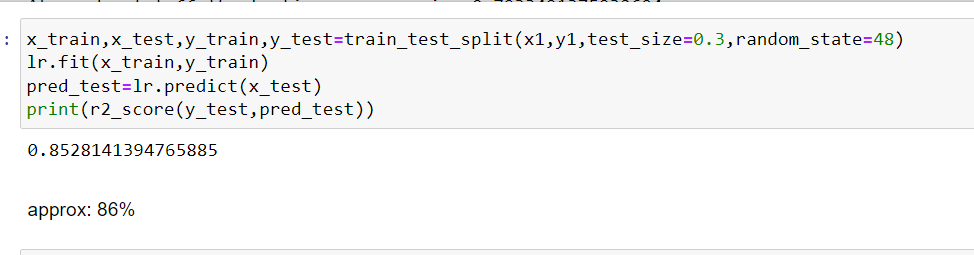


* The correlation w.r.t the target variable was found using corr(), the below fig shows which columns are positively and negatively correlated, which can show us that which features are affecting the house price.

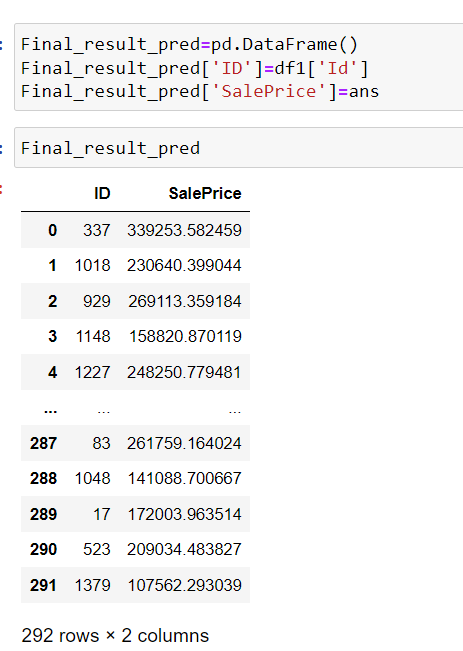


* Since the target variable has continuous numeric values , its see that the problem is based on Linear Regression . The below figure shows the test results and predicted results.





* The predictions to be made on the test.csv dataset to determine the house prices.



**CONCLUSION**

* Key Findings and Conclusions of the Study

Our Dataset contained numerical as well as categorical variables.

You need to find important features which affect the price positively or negatively

* Learning Outcomes of the Study in respect of Data Science

According to the results, the Linear Regression Model

obtained the most remarkable accuracy. The outcome of training the given file was seen by the results obtained at the test file , where the predictions made were approx. 85%

.

* Limitations of this work and Scope for Future Work

People will be able to utilize this program in the future to

acquire the most accurate pricing of a home.

This application may be converted into a Flutter application to get support for Android and iOS devices, allowing it to be used

everywhere. It can also be used as an external or internal

service for apps that display property for rent.

Users may apply this methodology to various fields, such as tuition

costs in a specific location, swimming pool rates, and data science-type models.