**Bansilal Ramnath Agarwal Charitable Trust’s**

**VISHWAKARMA INSTITUTE OF INFORMATION TECHNOLOGY,**

**PUNE-48**

**Department of Information Technology**

**ITUA40201: DATA SCIENCE AND ANALYTICS**

**Assignment-2**

**Shreyas Shripad Kulkarni**

**BTech A Division**

**Roll No.: 431048**

**PRN: 22010443**

**AIM:** Predictive Modelling Exercise

**OBJECTIVE:** Build a predictive model using machine learning algorithms and evaluate its performance.

**TASKS:**

1. Select a dataset (e.g., customer churn, housing prices).
2. Perform data preprocessing, including handling missing values and categorical variables.
3. Split the dataset into training and testing sets.
4. Train a predictive model (e.g., regression, classification) using suitable algorithms (e.g., linear regression, logistic regression, decision trees).
5. Evaluate the model & performance using appropriate metrics (e.g., accuracy, mean squared error). Fine-tune the model by adjusting parameters or using ensemble methods. Compare and interpret the results.

**THEORY:**

In this assignment I have used the [housing price prediction dataset.](https://www.kaggle.com/datasets/yasserh/housing-prices-dataset)

We will be first load the dataset. After cleaning the dataset and preprocessing the dataset we will form a correlation matrix.

After encoding the dataset we will train the predictive model after splitting the dataset into train and testing datasets.

Once the model is trained, we will evaluate the model by calculating the training and testing score & training, testing accuracy along with R-Squared, Mean Average Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

We will use the regularization (Ridge regression) to fine tune our model.

* Correlation Matrix:A correlation matrix is a table or matrix that displays the correlation coefficients between many variables.A correlation matrix provides a concise overview of the relationships between variables, with values ranging from -1 to 1 indicating the strength and direction of linear correlations.
* Encoding: Encoding is essential in machine learning as it transforms data into a format suitable for algorithms, such as converting categorical variables into numerical representations.
* Training Accuracy: Training accuracy assesses how well a model fits the training data.
* Testing Accuracy: Testing accuracy measures its performance on unseen data, crucial for evaluating generalization.
* R-Squared: R-squared quantifies the goodness of fit in regression models, representing the proportion of variance in the dependent variable explained by independent variables.
* Regularization: Regularization is a technique used in machine learning and statistical modelling to prevent overfitting and improve the generalization performance of a model.

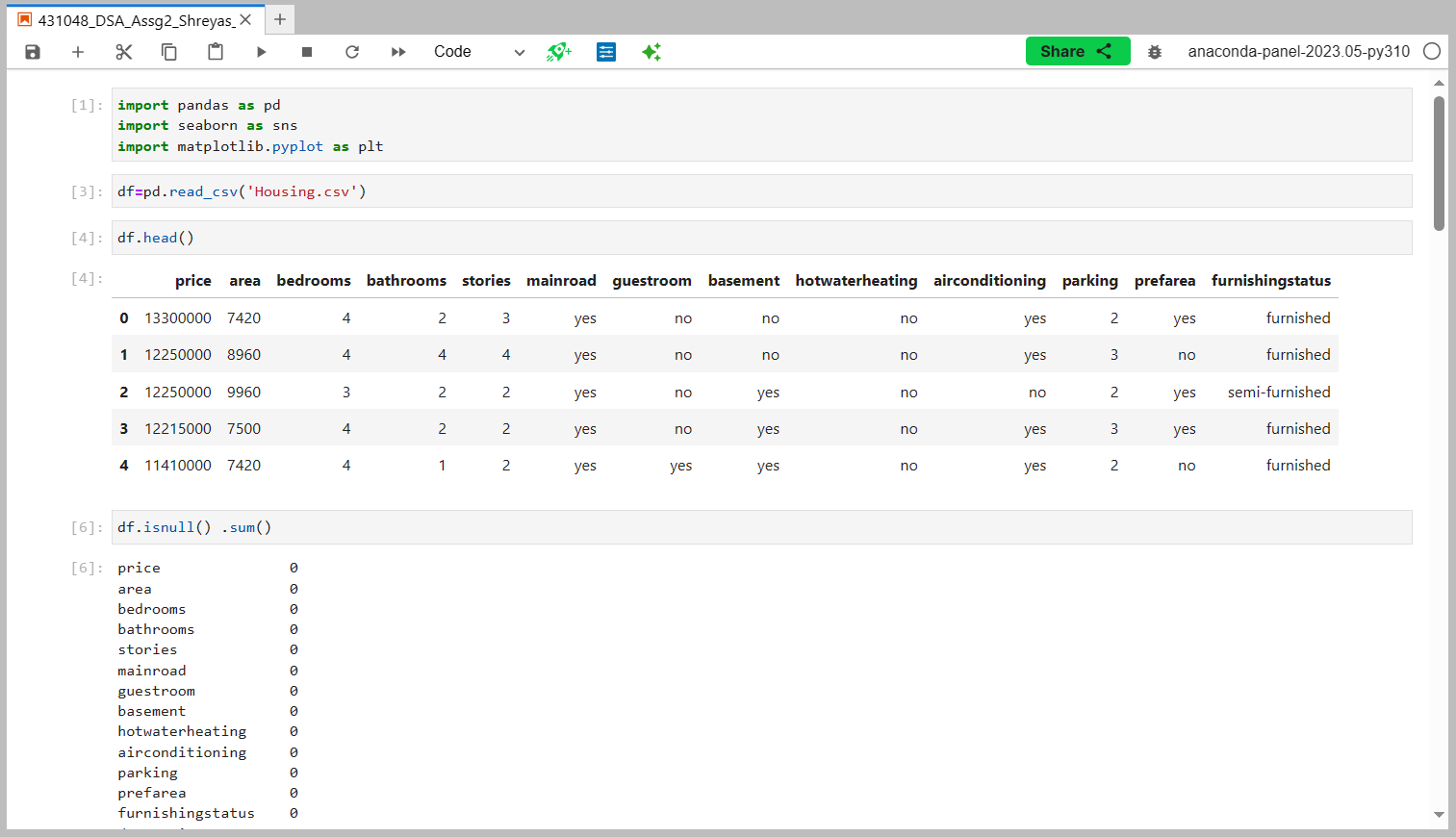
There are 2 types of regularization:

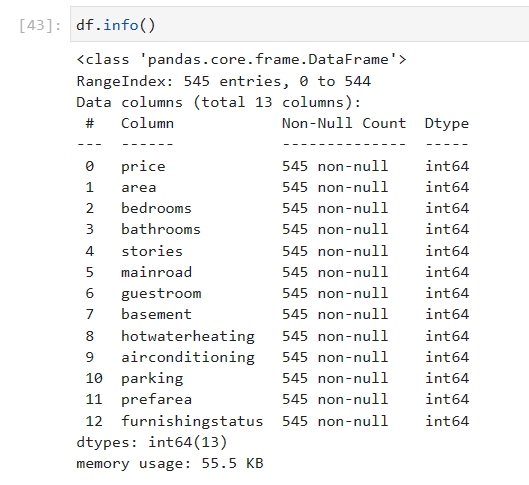
* Lasso Regression (L1 Regularization): Lasso Regression is also a type of regularization linear model. It also adds a penalty term to the cost function. *The main aim of Lasso Regression is to reduce the features and hence can be used for Feature Selection.*
* Ridge Regression (L2 Regularization): Ridge regression is a type of regularized regression model. This means it is a variation of the standard linear regression model that includes a regularized term in the cost function. The purpose of this is to prevent Overfitting.

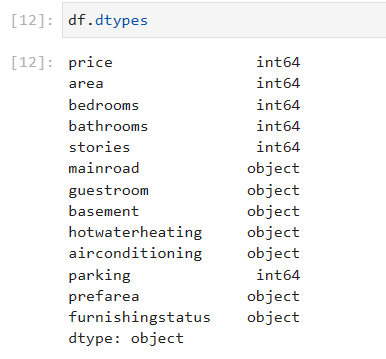
**IMPLEMENTATION:**

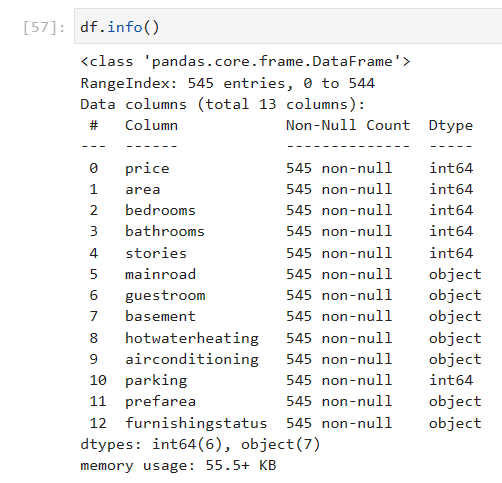
In this assignment we will be using the standard dataset (Housing).

Step 1: Load Dataset.

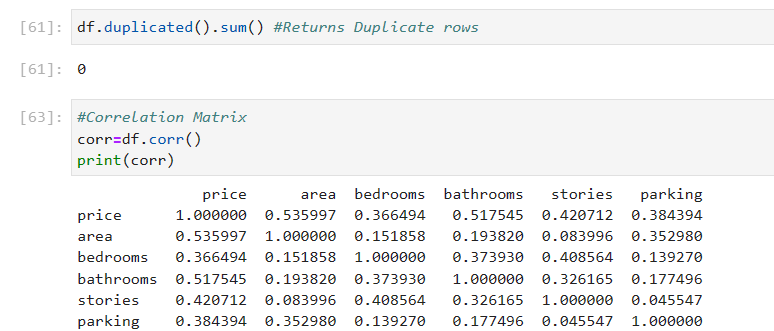




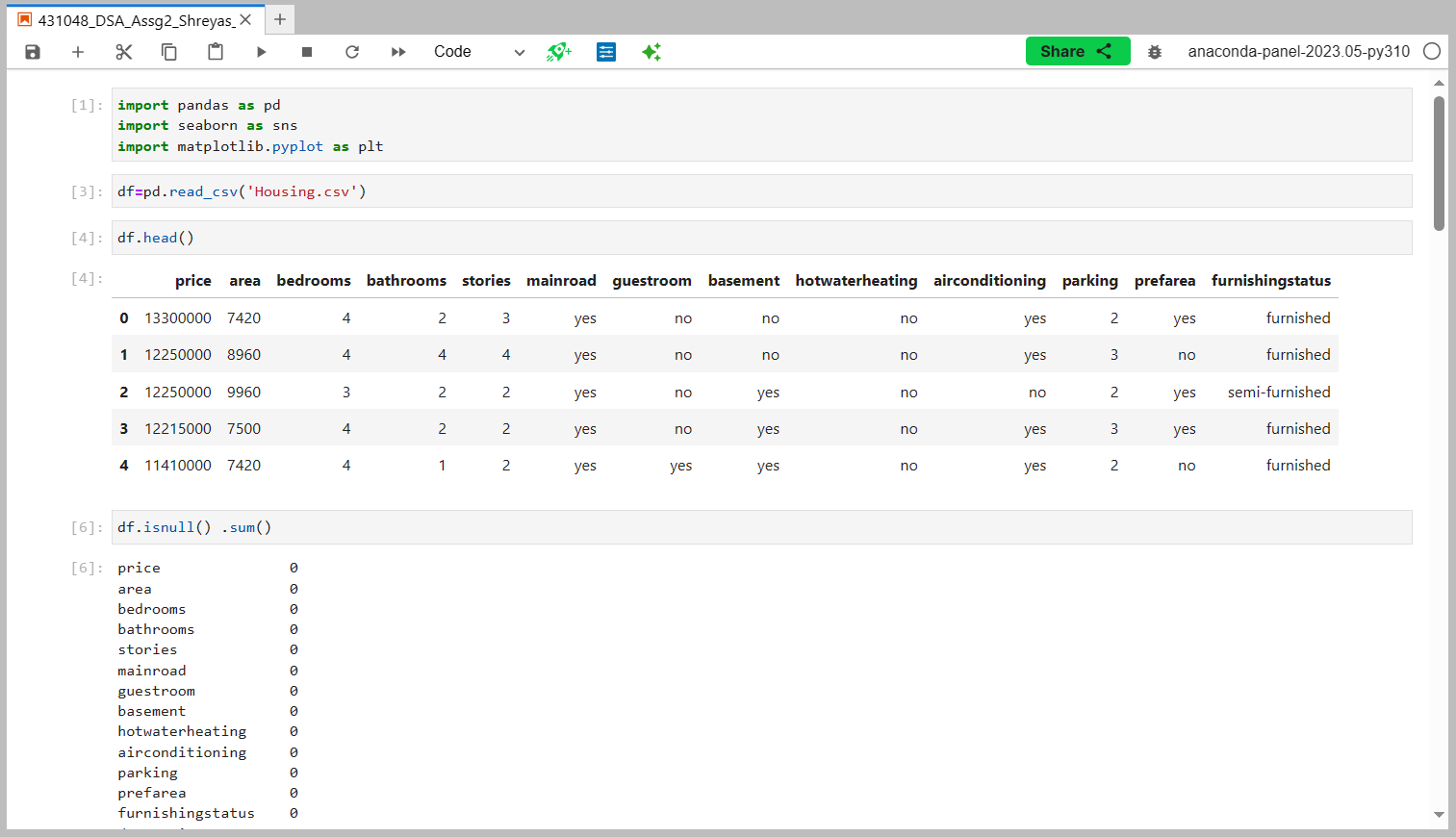




Step 2: Clean Dataset

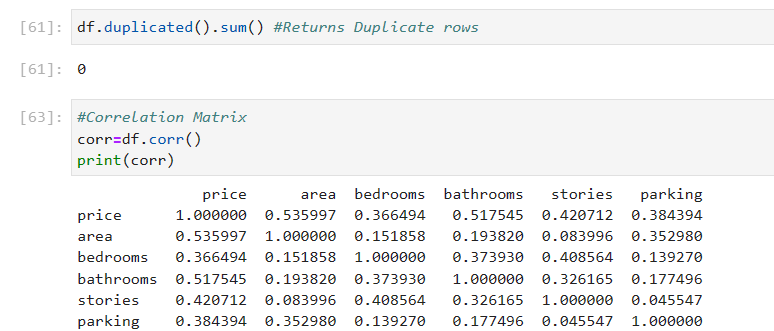


df.isnull() .sum() is the count of missing values in the corresponding column

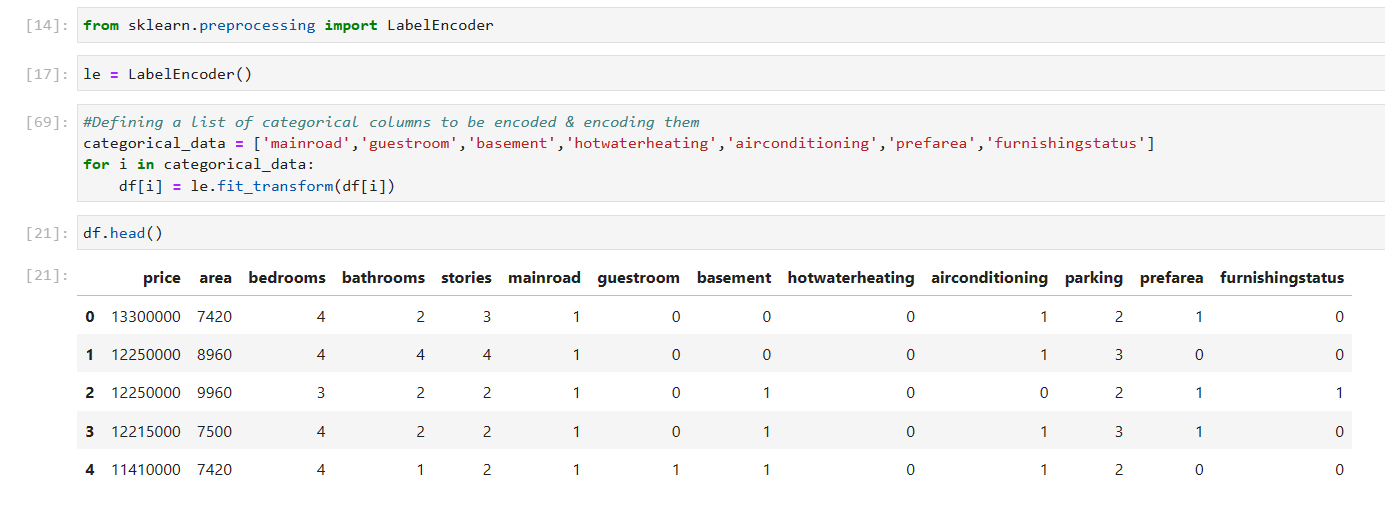


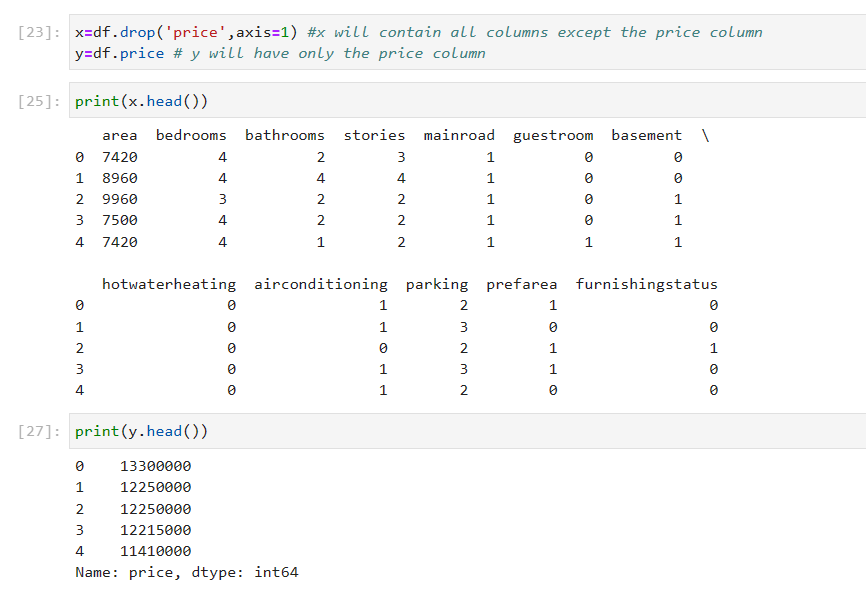
No missing values or duplicate values which means that our dataset is clean

Step 3: Form the Correlation matrix



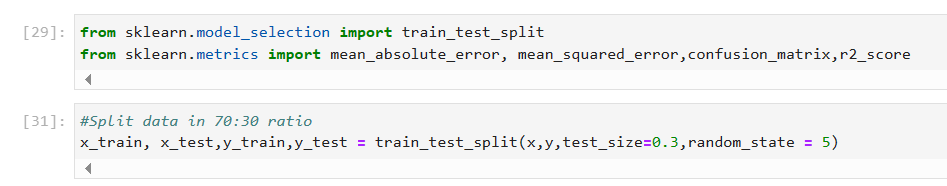
Step 4 : Encoding Categorical Data



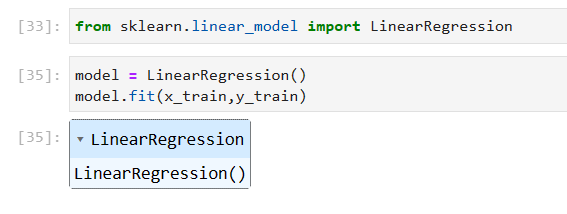


Step 5: Split the dataset

We will split the data in 7:3 Ratio i.e 20% of the dataset will be training data & 30% data will be test data.

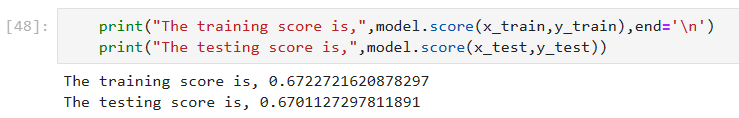


Step 6: Train the predictive model

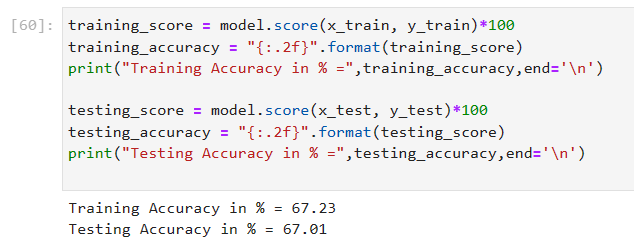


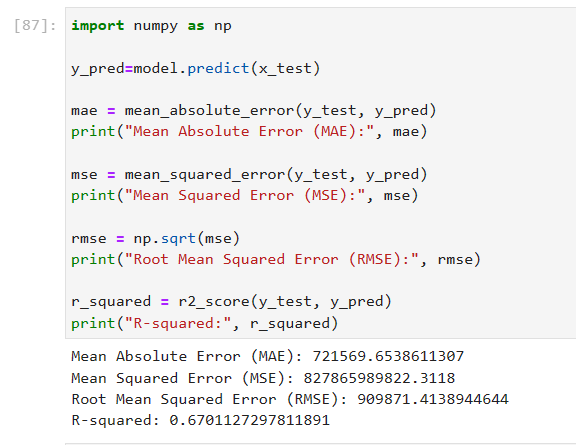
Step 7: Evaluating the model

Training & Testing score



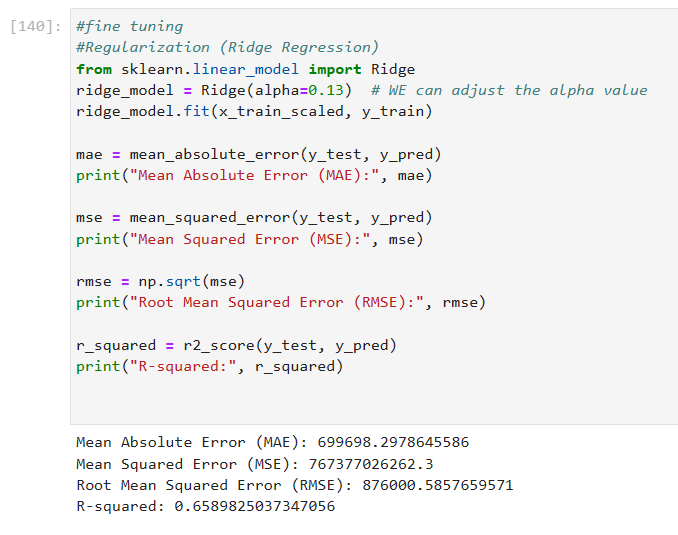
Training & Testing Accuracy





Step 8: Fine Tuning

We will use the regularization (Ridge regression) to fine tune our model.



We can clearly see that the MAE,MSE & RMS has been reduced which indicates the model is finely tuned and is fine that the previous model

**CONCLUSION**: We have learnt, understood and performed predictive modelling on housing prediction dataset.

[GitHub Repository](https://github.com/Shreyas100100/DATA-SCIENCE-AND-ANALYTICS)