**PREDICTING HOUSE PRICE USING MACHINE LEARNING**

**PROBLEM DEFINITION:**

We all have experienced a time when we have to look up for a new house to buy. But then the journey begins with a lot of frauds, negotiating deals ,researching the local areas and so on. The machine learning model is given the test data but without the price of the properties in order to predict the price for them given the various features for the properties. The predicted price is then compared to the actual price in test data. House price prediction can help the developer determine the selling price of a house and can help customer to arrange the right time to purchase a house. There are three factors that influence price of a house which include physical conditions, concepts and location.

**DESIGN THINKING:**

The competition goal is to predict sale prices for homes in Ames, Iowa. You’re given a training and testing data set in csv format as well as a data dictionary.

**Training**: Our training data consists of 1,460 examples of houses with 79 features describing every aspect of the house. We are given sale prices (labels) for each house. The training data is what we will use to “teach” our models.

**Testing**: The test data set consists of 1,459 examples with the same number of features as the training data. Our test data set excludes the sale price because this is what we are trying to predict. Once our models have been built we will run the best one the test data and submit it to the Kaggle leaderboard.

You can familiarize yourself with the data on the competition page.

**Task**: Machine learning tasks are usually split into three categories; supervised, unsupervised and reinforcement. For this competition, our task is supervised learning.

Supervised learning uses examples and labels to find patterns in data:It’s easy to recognise the type of machine learning task in front of you from the data you have and your objective. We’ve been given housing data consisting of features and labels, and we’re tasked with predicting the labels for houses outside of our training data.

**Tools**:I used Python and Jupyter notebooks for the competition. Jupyter notebooks are popular among data scientist because they are easy to follow and show your working steps.

Please be aware this code is not for production purposes, it doesn’t follow software engineering best practices. I’ve sacrificed that somewhat for explainability.

**Libraries**: These are frameworks in python to handle commonly required tasks. I Implore any budding data scientists to familiarise themselves with these libraries:

Pandas — For handling structured data

Scikit Learn — For machine learning

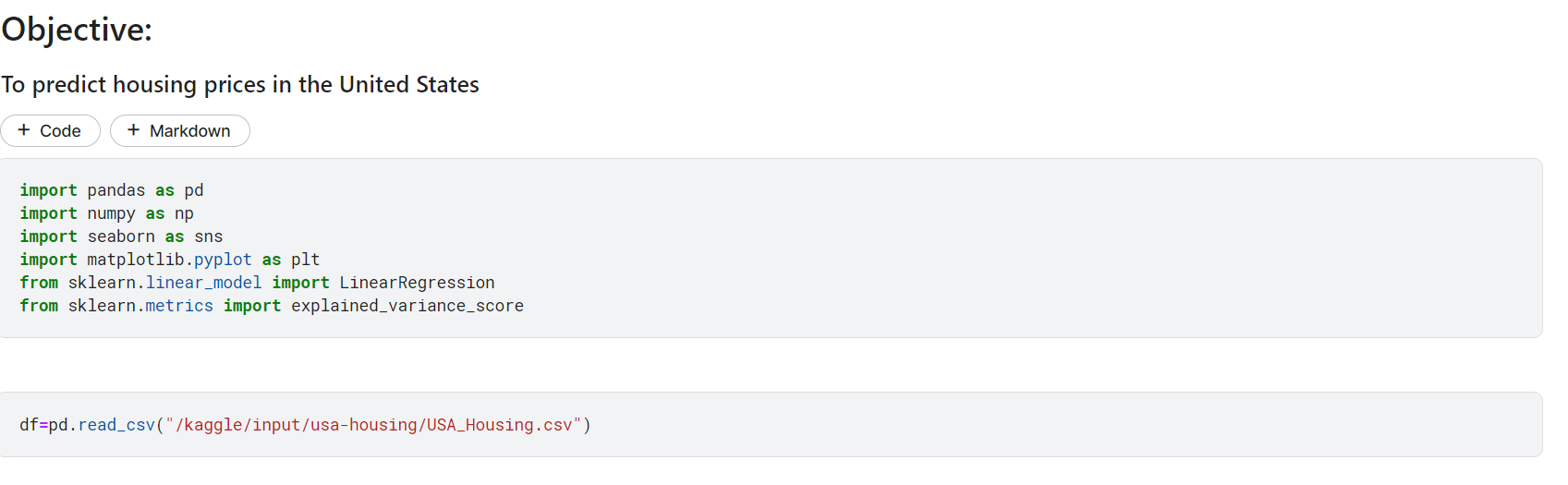
NumPy — For linear algebra and mathematics

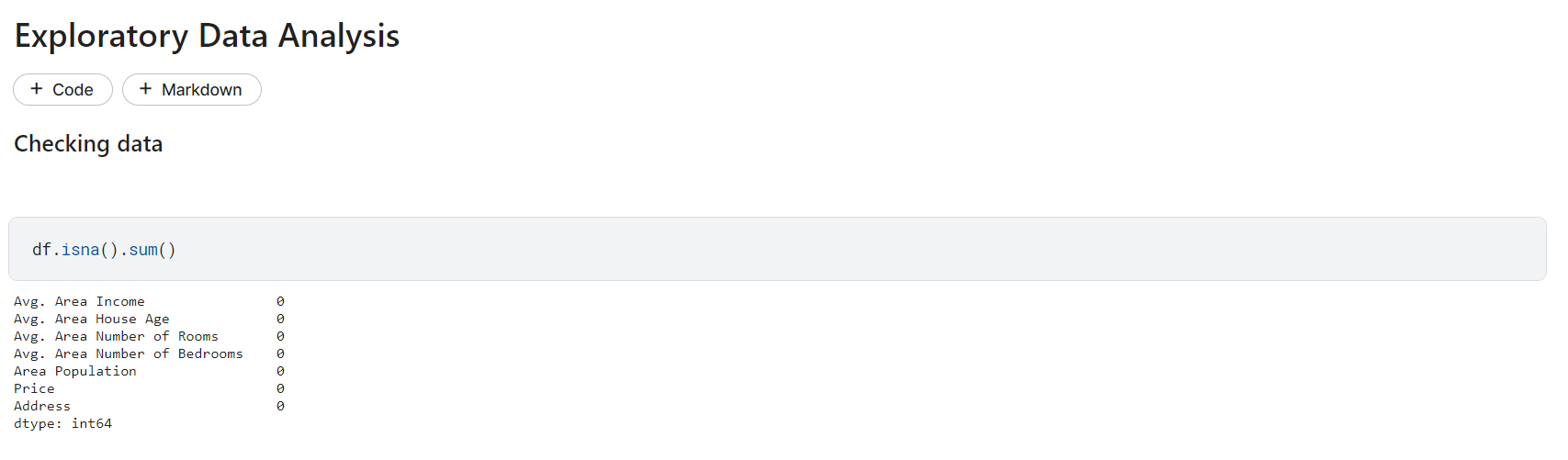
Seaborn — For data visualization

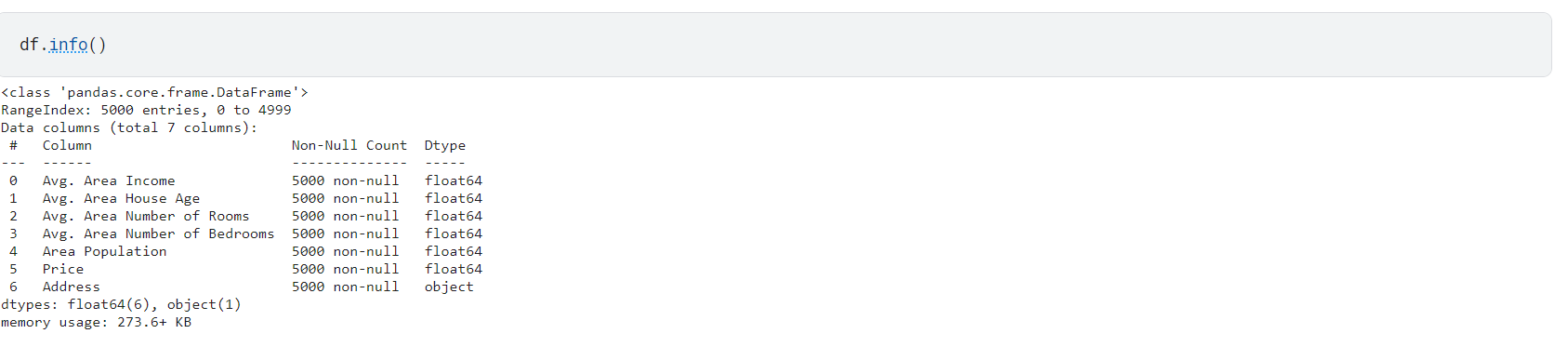
**Project Pipeline:**

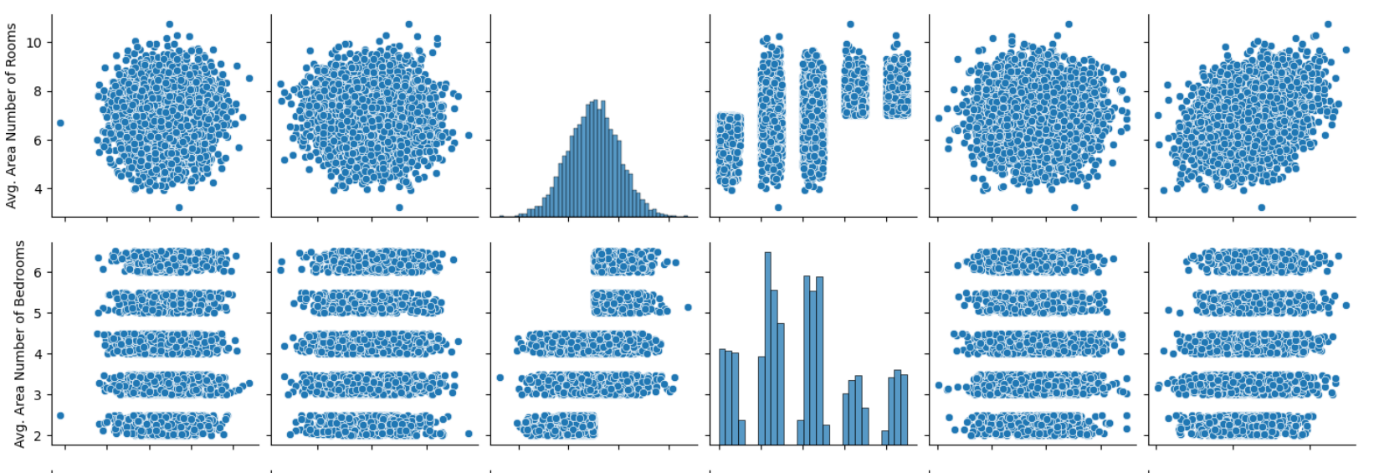
Generally speaking, machine learning projects follow the same process. Data ingestion, data cleaning, exploratory data analysis, feature engineering and finally machine learning.

The pipeline is not linear and you might find you have to jump back and forth between different stages. It’s important I mention this because tutorials often make you believe the process is much cleaner than in reality. So please keep this in mind, your first machine learning project might be a mess.

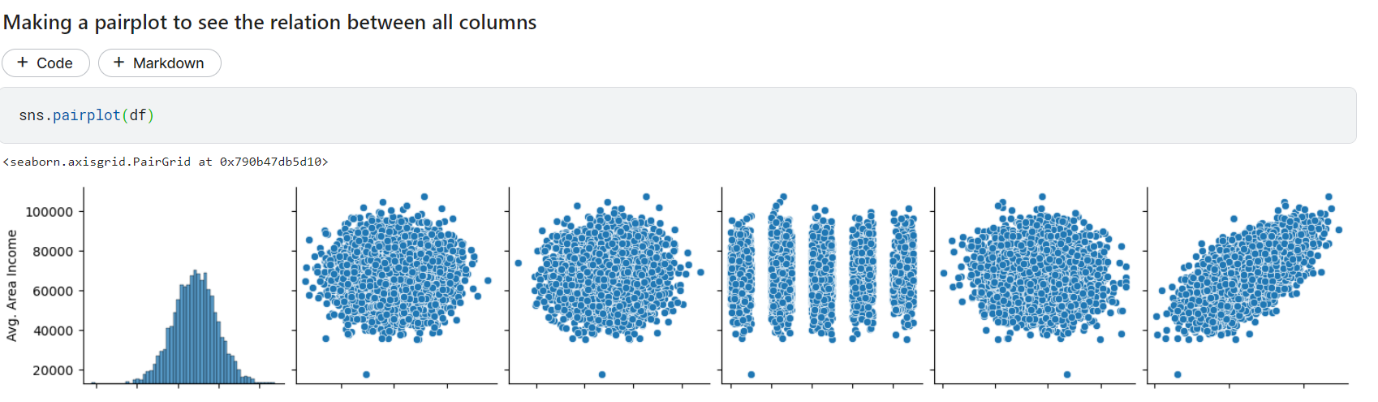


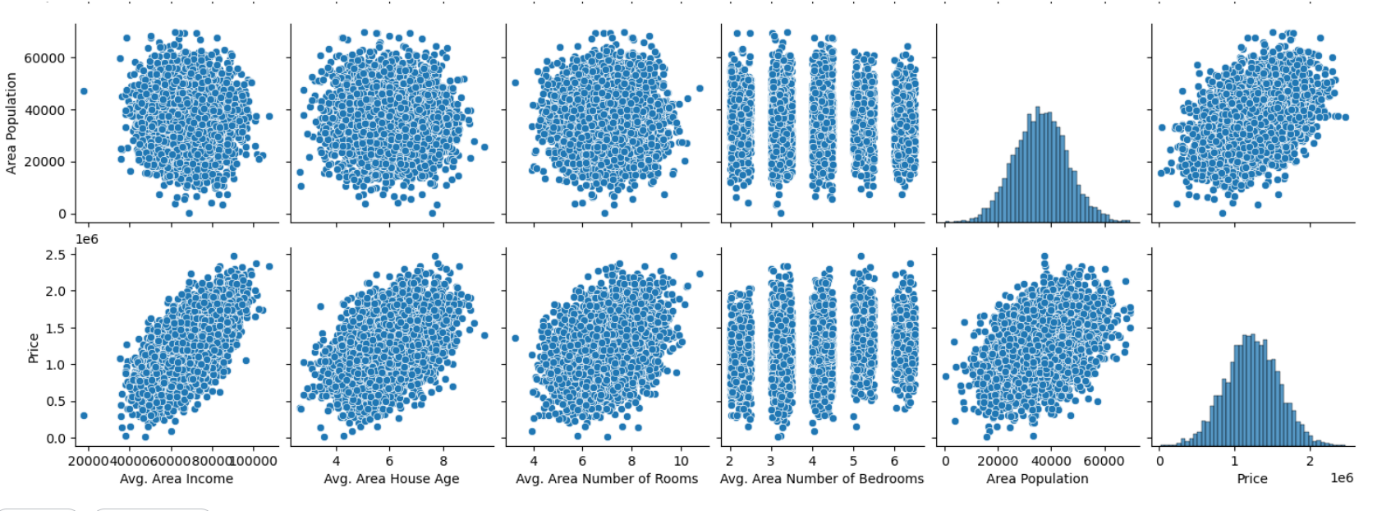


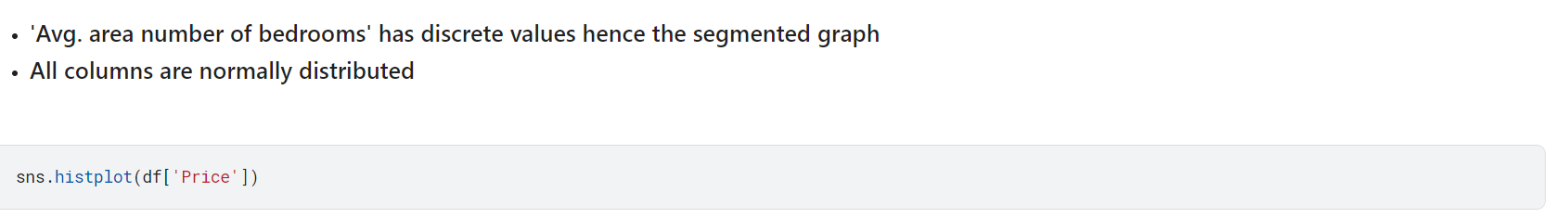


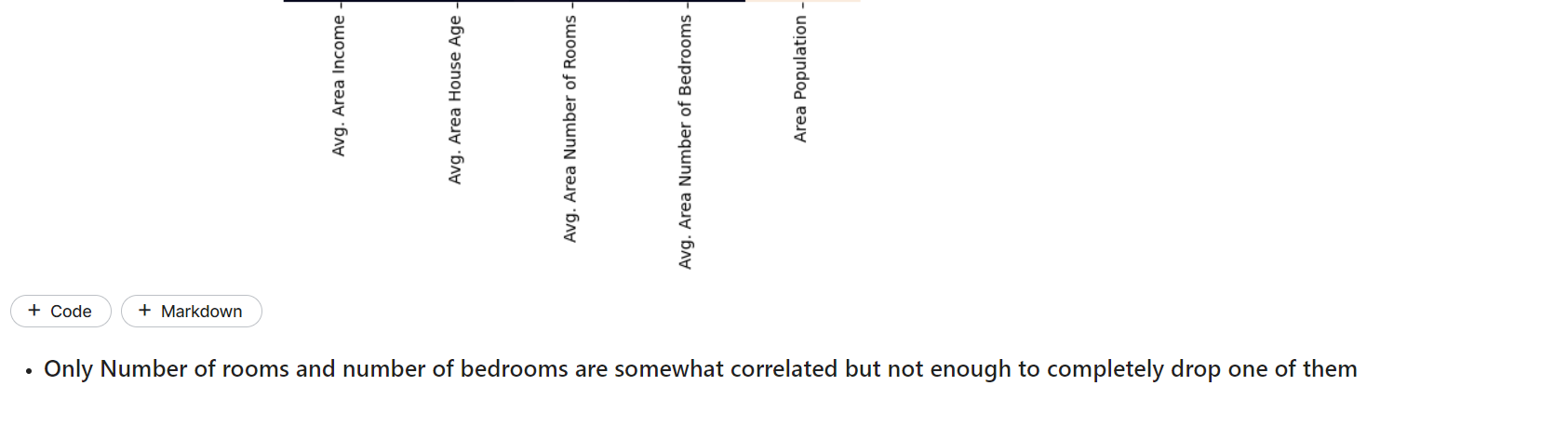
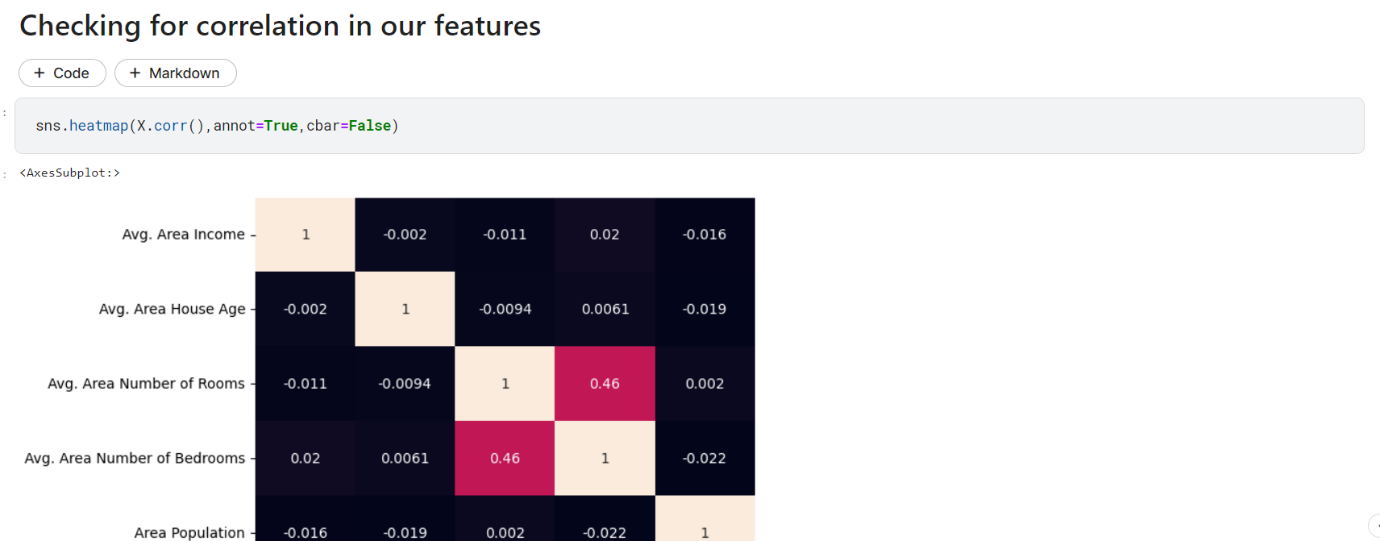


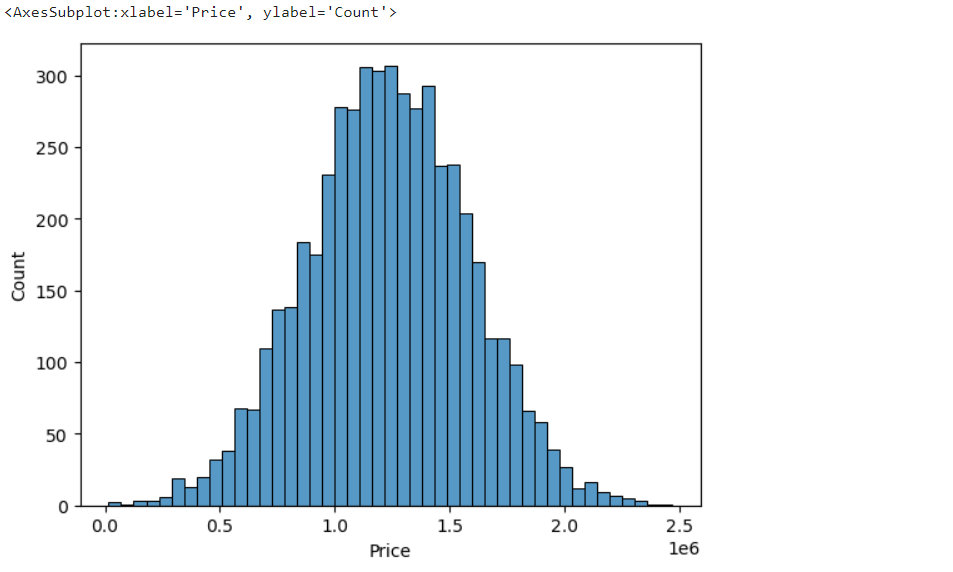


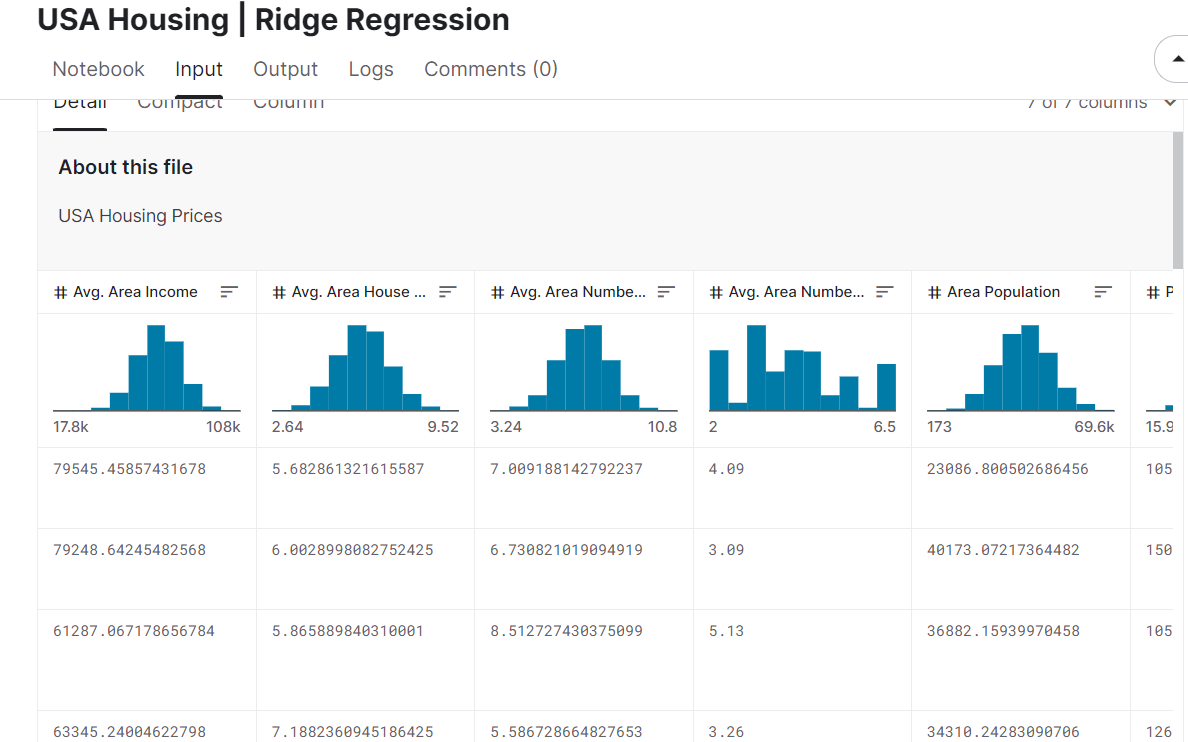


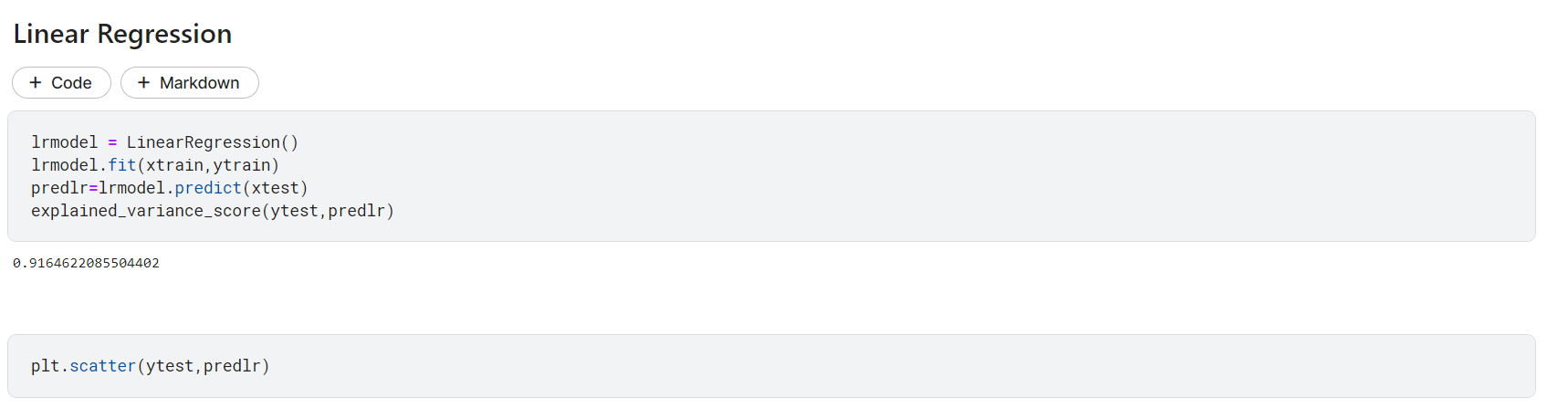


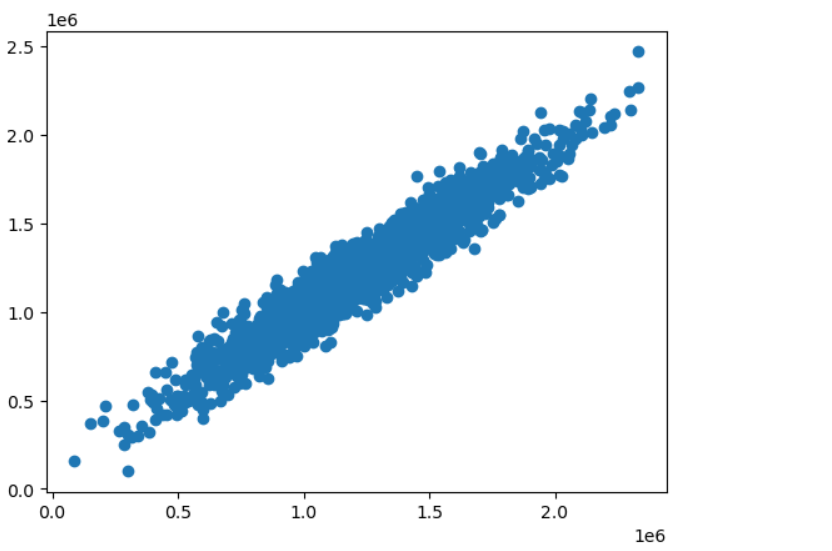


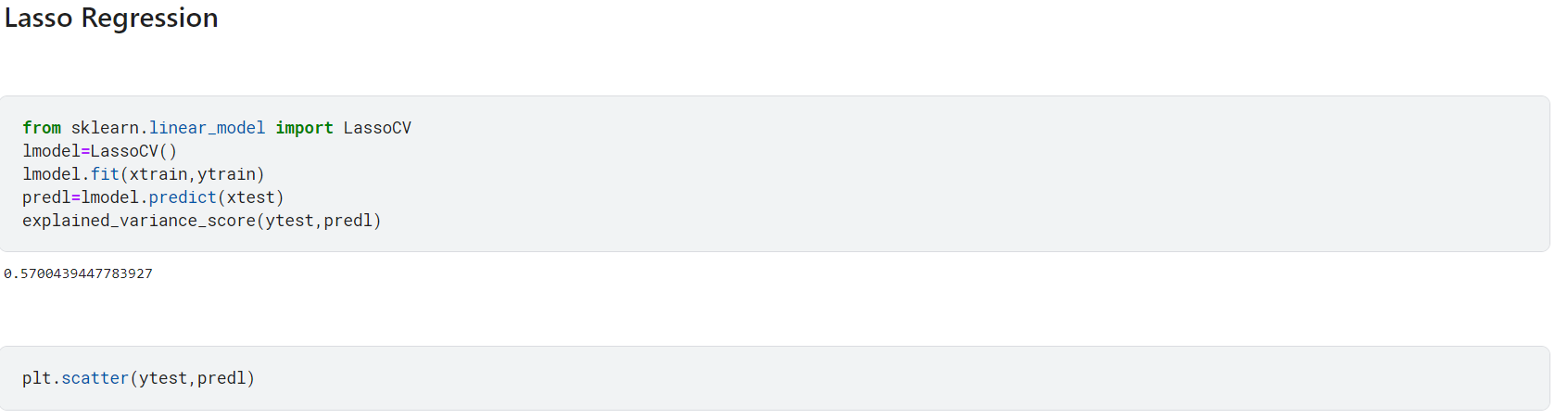


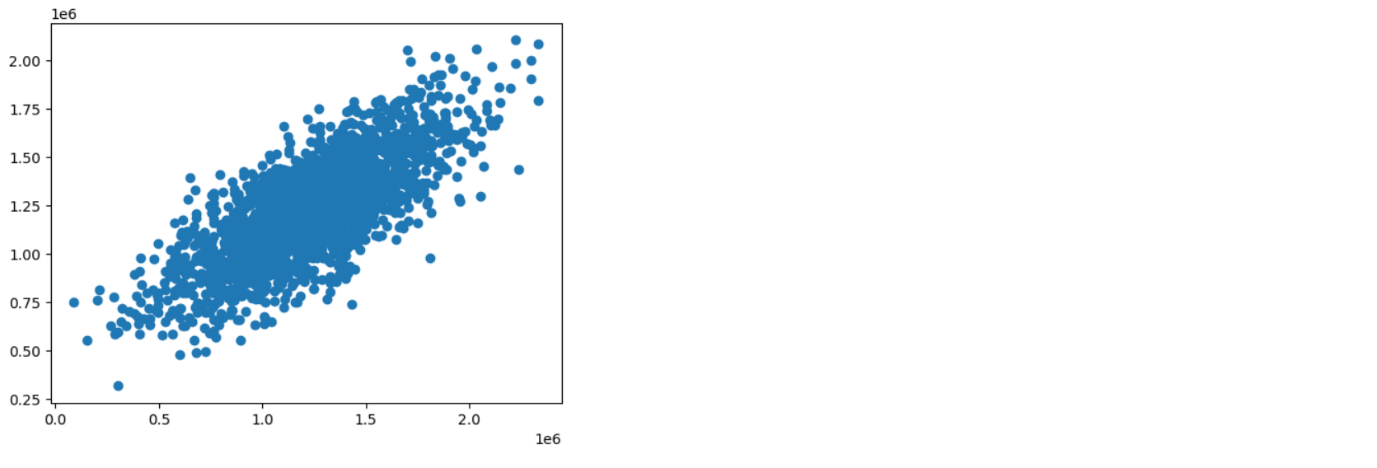


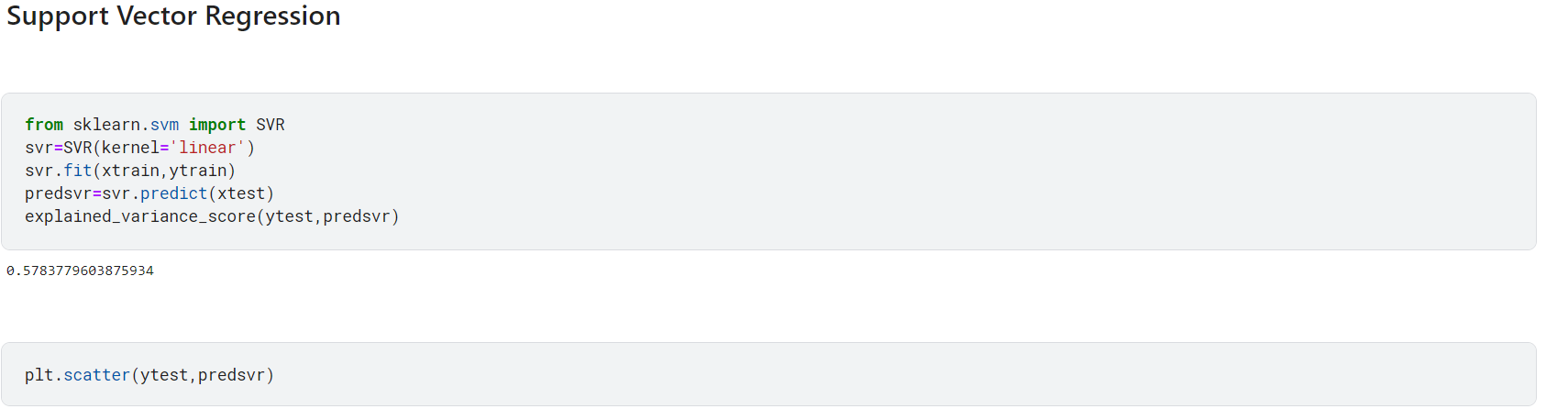


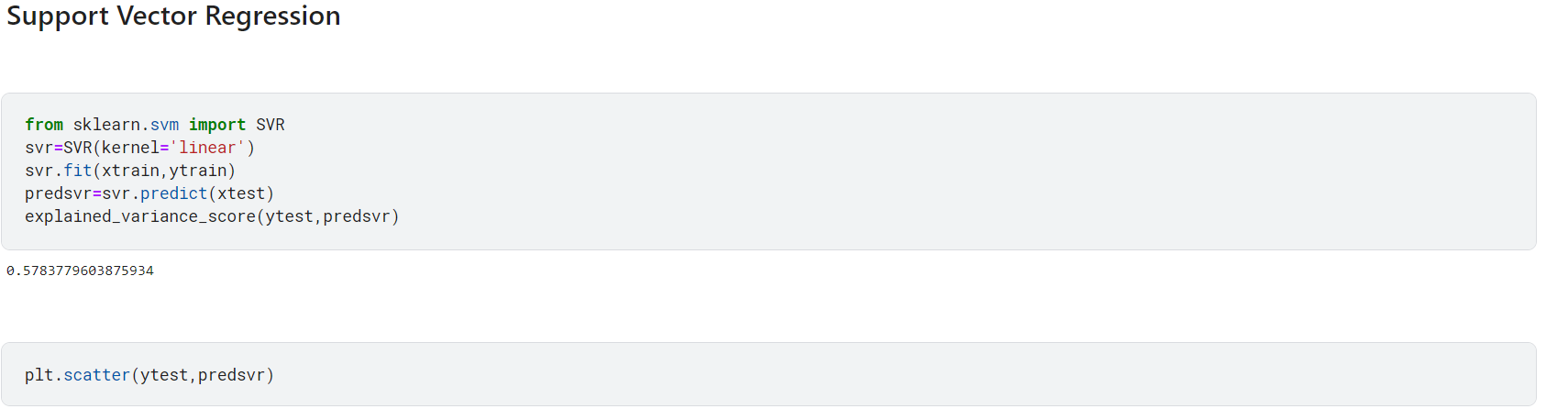


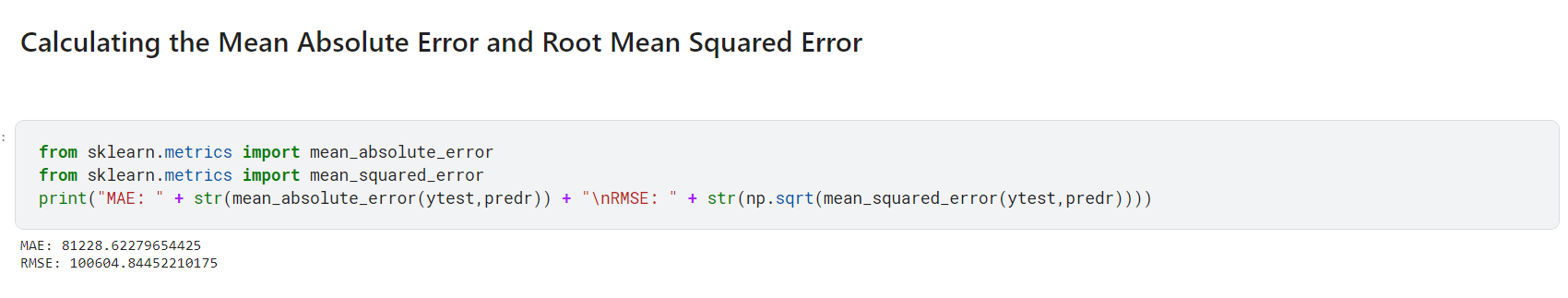


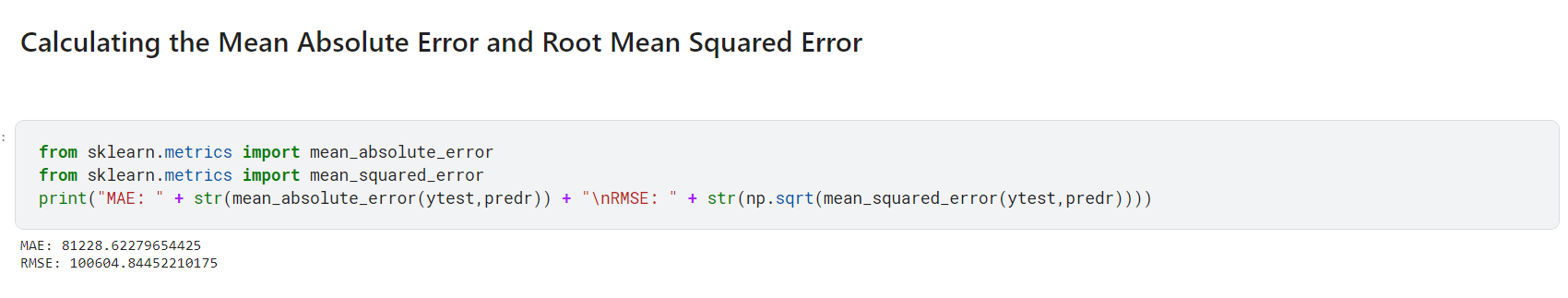




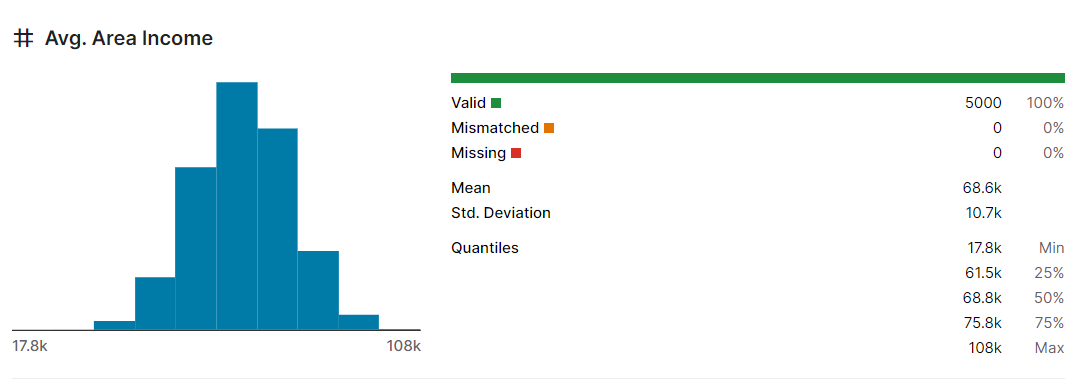


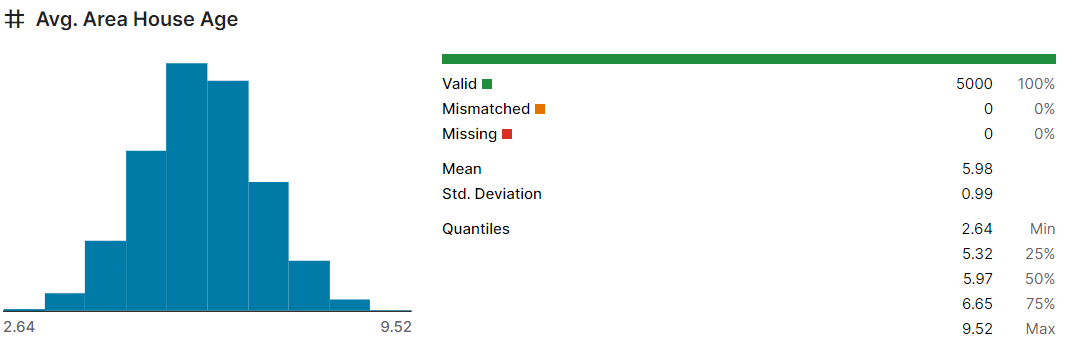


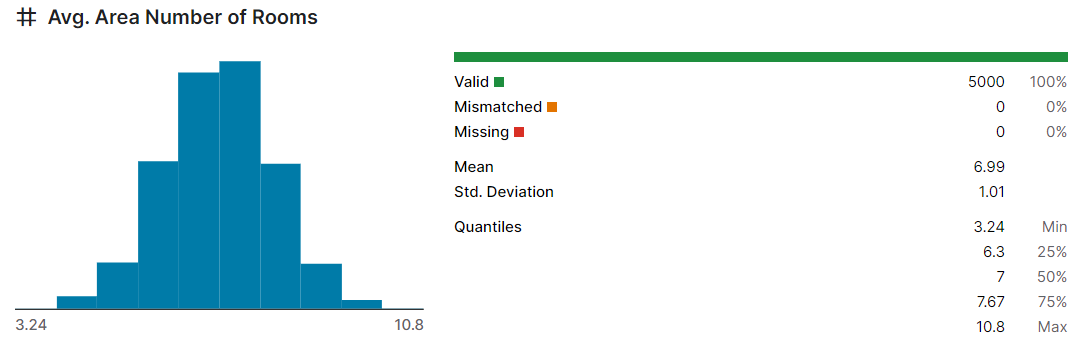


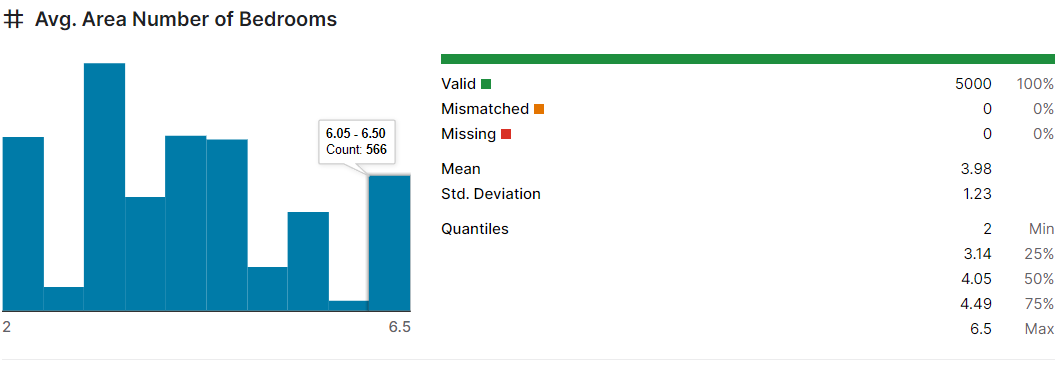


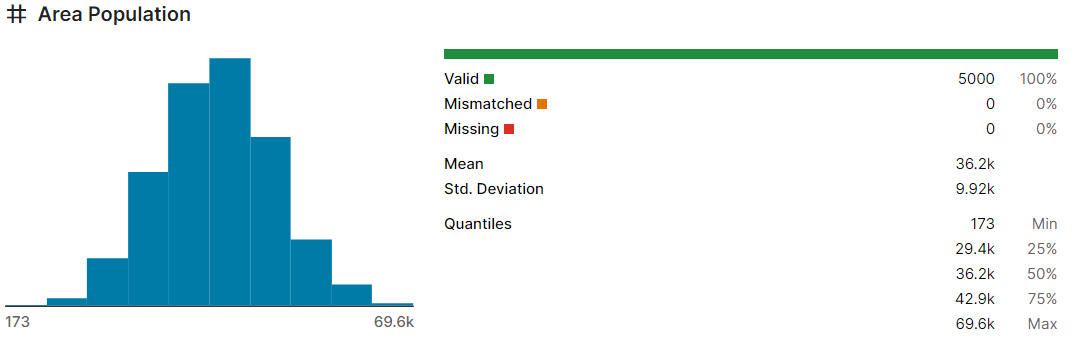
**EVALUATIONS:**

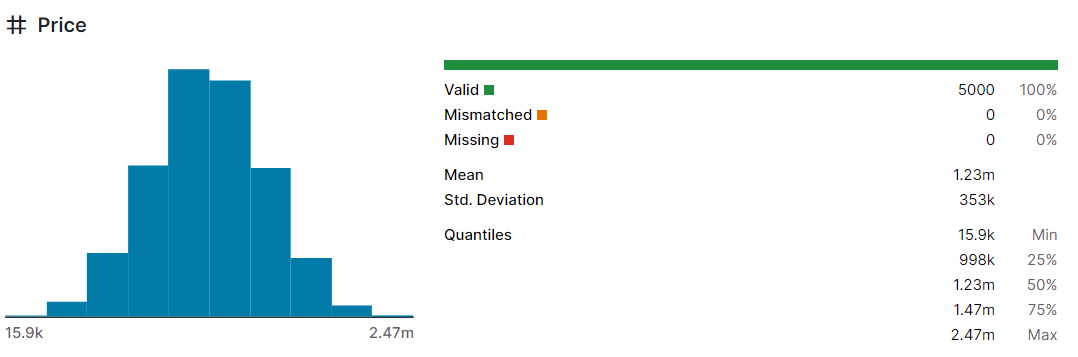


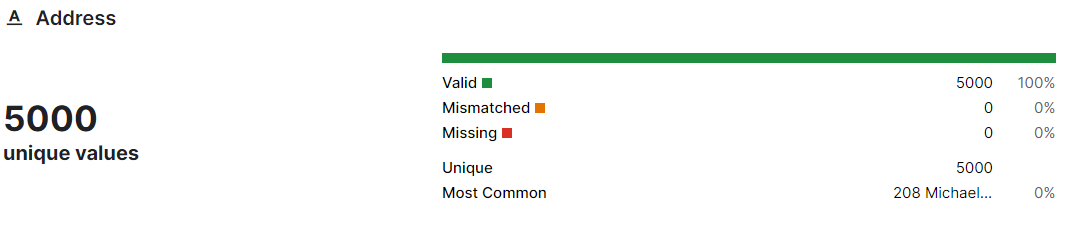












**MODEL AND ACCURACY:**

As we have train the model to determine the continuous values, so we will be using these regression models.

* svm – support vector machine
* random forest regression
* catboost classifier

**SVM - SUPPORT VECTOR MACHINE:**

It can used for both regression and classification problems or models.It finds hyperplane in the n-dimensional plane

**RANDOM FROEST REGRESSION:**

It is an ensemble technique that uses multiple of decisions trees and can be used for both regression and classification tasks.

**CATBOOST CLASSIFIER:**

It is a machine learning algorithm implemented by Yandex and is open-source. It is simple to interface with deep learning frameworks such as Apple’s Core ML and Google’s TenserFlow.

**CONCLUSIONS:**

Clearly ,**svm** **model** is giving better accuracy as the mean absolute error is least among all other regression models i.e, 0.18 approx.To get much better results ensemble learning techniques like **Bagging** and **Boosting** can used.