**ARTIFITIAL INTELLIGENCE LAB**



**PROJECT REPORT**

**REALTY PRICE PREDICTION**

**GROUP MEMBERS:**

|  |  |
| --- | --- |
| BASIL ASLAM | (02-134191-002) |
| ALI FAWAD | (02-134191-094) |
| SHAHZAIN SHAFIQUE | (02-134191-051) |

Table of Contents

[**ABSTRACT** 3](#_Toc94549584)

[**INTRODUCTION** 3](#_Toc94549585)

[**AIM AND IMPORTANCE** 4](#_Toc94549586)

[**NEED AND MOTIVATION** 4](#_Toc94549588)

[**DATASET** 5](#_Toc94549589)

[**DATA EXPLORATION** 6](#_Toc94549590)

[**DATA VISUALIZATION** 6](#_Toc94549591)

[**DATA SELECTION** 7](#_Toc94549594)

[**DATA TRANSFORMATION** 9](#_Toc94549595)

[**LANGUAGE AND MODELS TESTED** 13](#_Toc94549596)

[**PYTHON** 13](#_Toc94549597)

[**LIBRARIES USED FOR THIS PROJECT INCLUDE** 13](#_Toc94549598)

[**MODELS TESTED** 13](#_Toc94549599)

[**DECISION TREE MODEL** 13](#_Toc94549600)

[**LASSO MODEL** 14](#_Toc94549602)

[**LINEAR REGRESSION MODEL** 15](#_Toc94549603)

[**RESULTS AND DISCUSSIONS** 16](#_Toc94549604)

[**BEST SUITED MODEL** 16](#_Toc94549605)

[**RMSE Bar Graph** 17](#_Toc94549608)

[**DEPLOYMENT APP** 17](#_Toc94549609)

[**CONCLUSION** 19](#_Toc94549611)

# **ABSTRACT**

House price forecasting is an important topic of real estate. The literature attempts to derive useful knowledge from historical data of property markets. Machine learning techniques are applied to analyze historical property transactions in Pakistan to discover useful models for house buyers and sellers. Revealed is the high discrepancy between house prices in the most expensive and most affordable suburbs in the city of Karachi. Moreover, experiments demonstrate that the Multiple Linear Regression that is based on mean squared error measurement is a competitive approach.

# **INTRODUCTION**

The real estate industry is a big one, with a lot of people involved, from regulators to private companies and investors. There is a considerable need among these stakeholders for a better grasp of the industry's operational mechanisms and driving causes.

Today, there is a wealth of data on pertinent statistics as well as other contextual elements, and it is natural to want to make sense of it all in order to better understand the sector. This has been done, for example, in Zillow's Zestimate and Kaggle's housing pricing competitions.

Non-traditional indicators have proven to be valuable forecasters of real estate trends in various circumstances. For example, in Seattle, apartments near speciality food stores like Whole Foods saw a larger increase in value than the national average.

This research can be seen as a step forward in the direction of more evidence-based decision-making for these stakeholders. Between 2017 and 2020, the initiative focuses on determining the value of residential properties in Karachi. Our project's goal was to develop a model that could anticipate changes in house prices in 2021 based on time and geography-dependent variables.

# **AIM AND IMPORTANCE**

# The goal is to forecast the most cost-effective property pricing for real estate buyers based on their budgets and goals. Future prices will be projected by examining recent market trends and price ranges, as well as forthcoming changes. The system works by combining the application with a website that accepts customer specifications. Customers can use this app to invest in real estate without having to deal with a real estate agent. It also reduces the risk associated with the transaction.

Purchasing or selling a home in today's market is stressful and costly. Because the customer must travel to different locations and pay a commission to the real estate agent. Furthermore, the customer/buyer has no idea whether the property will be profitable in the future.

These are the Parameters on which we will evaluate ourselves:

* Create an effective price prediction model
* Validate the model’s prediction accuracy
* Identify the important home price attributes which feed the model’s predictive power.

# **NEED AND MOTIVATION**

The general and standardized real estate characteristics are often listed separately from the asking price and general description. Because these characteristics are separately listed in a structured way, they can be easily compared across the whole range of potential houses. Because every house also has its own unique characteristics, such as a particular view or type of sink, house sellers can provide a summary of all the important features of the house in the description. All given real estate features can be considered by the potential buyers, but it is nearly impossible to provide an automated comparison on all variables due to the large diversity. This is also true in the other direction: house sellers have to make an estimation of the value based on its features in comparison to the current market price of similar houses. The diversity of features makes it challenging to estimate an adequate market price. Apart from providing a summary of the important features of the house, the house description is also a means of raising curiosity in the reader, or in other words to persuade the person. It is possible that there are certain word sequences in the natural language text that seduce potential buyers more than others. Therefore, there might be a relation between the language used in the description and the price of the property. This comparison does not focus primarily on the house characteristics, but on all words within the description.

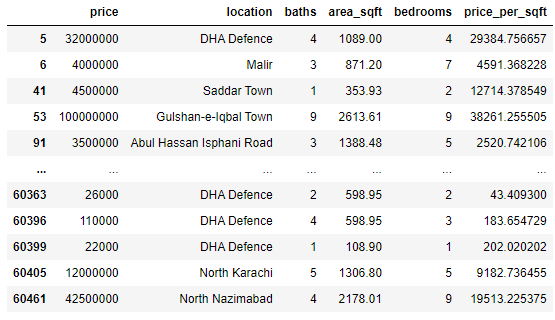
Having lived in Pakistan for so many years if there is one thing that I had been taking for granted, it’s that housing and rental prices continue to rise. Since the housing crisis of 2008, housing prices have recovered remarkably well, especially in major housing markets. However, in the 4th quarter of 2016, I was surprised to read that Rawalpindi housing prices had fallen the most in the last 4 years. In fact, median resale prices for condos and coops fell 6.3%, marking the first time there was a decline since Q1 of 2017. The decline has been partly attributed to political uncertainty domestically and abroad and the 2014 election. So, to maintain the transparency among customers and also the comparison can be made easy through this model. If customer finds the price of house at some given website higher than the price predicted by the model, so he can reject that house.

# **DATASET**

Here we have web scrapped the Data from zameen.com.pk website which is one of the leading real estate websites operating in PAKISTAN.

Our Data contains Karachi Houses only.

Dataset looks as follows-



# **DATA EXPLORATION**

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, Python. Before it can conduct analysis on data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working.

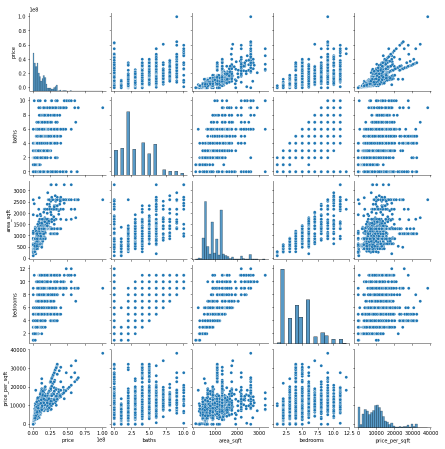
We divided the data 9:1 for Training and Testing purpose respectively.

# **DATA VISUALIZATION**

# Data visualization is the act of locating, evaluating, and comparing data in order to explain more clearly complicated ideas and make logical patterns simpler to spot. Many analytical tasks, such as data summaries, test data analysis, and model output analysis, require data visualization. Seeing good is one of the simplest ways to connect with others.

# Fortunately, Python includes a number of modules that may be used to extract data from data. Matplotlib, Seaborn, Bokeh, Altair, and others are among the most popular.

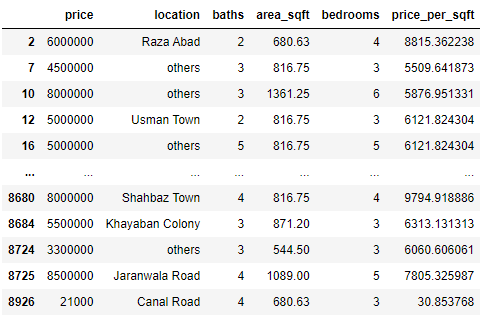
Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.



# **DATA SELECTION**

Data selection is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data. Data selection precedes the actual practice of data collection. This definition distinguishes data selection from selective data reporting (selectively excluding data that is not supportive of a research hypothesis) and interactive/active data selection (using collected data for monitoring activities/events, or conducting secondary data analyses). The process of selecting suitable data for a research project can impact data integrity.

The primary objective of data selection is the determination of appropriate data type, source, and instrument(s) that allow investigators to adequately answer research questions. This determination is often discipline-specific and is primarily driven by the nature of the investigation, existing literature, and accessibility to necessary data sources.



**Correlation Heatmap**

A correlation heatmap is a heatmap that shows a 2D correlation matrix between two discrete dimensions, using colored cells to represent data from usually a monochromatic scale. The values of the first dimension appear as the rows of the table while of the second dimension as a column. The color of the cell is proportional to the number of measurements that match the dimensional value. This makes correlation heatmap ideal for data analysis since it makes patterns easily readable and highlights the differences and variation in the same data. A correlation heatmap, like a regular heatmap, is assisted by a color bar making data easily readable and comprehensible.

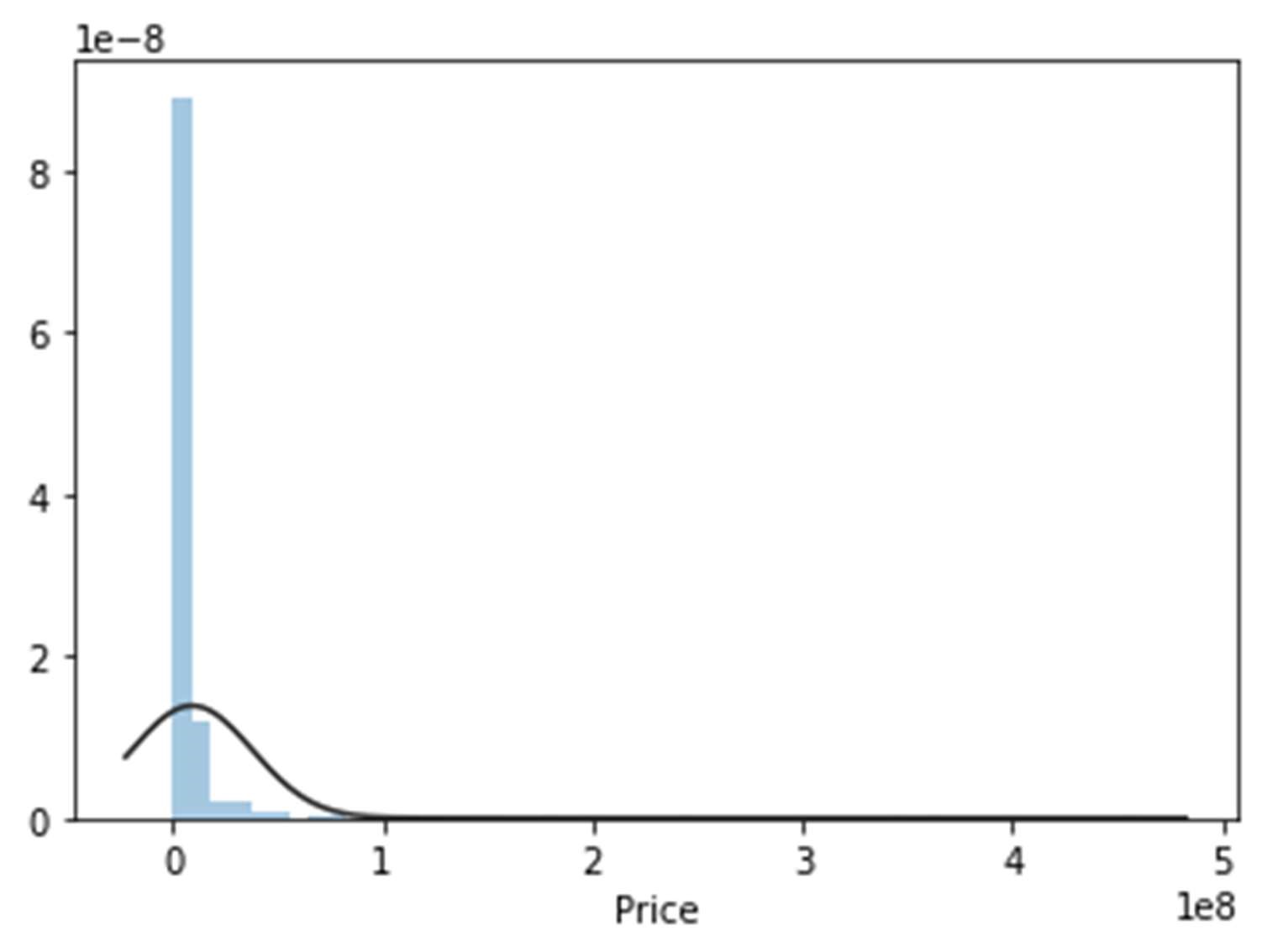


# **DATA TRANSFORMATION**

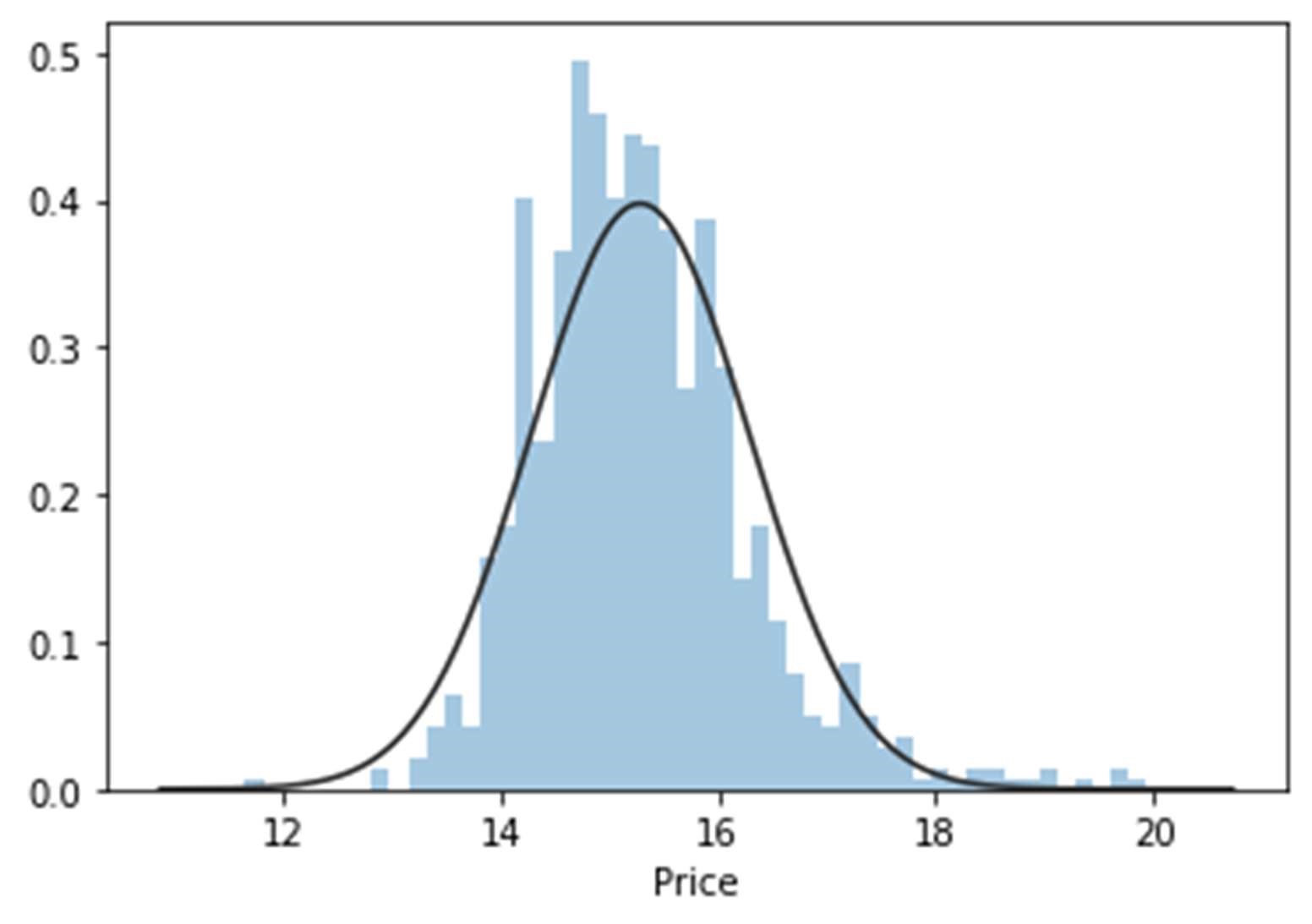
The log transformation can be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics.

It is hard to discern a pattern in the upper panel whereas the strong relationship is shown clearly in the lower panel. The comparison of the means of log-transformed data is actually a comparison of geometric means. This occurs because, as shown below, the anti-log of the arithmetic mean of log-transformed values is the geometric mean.

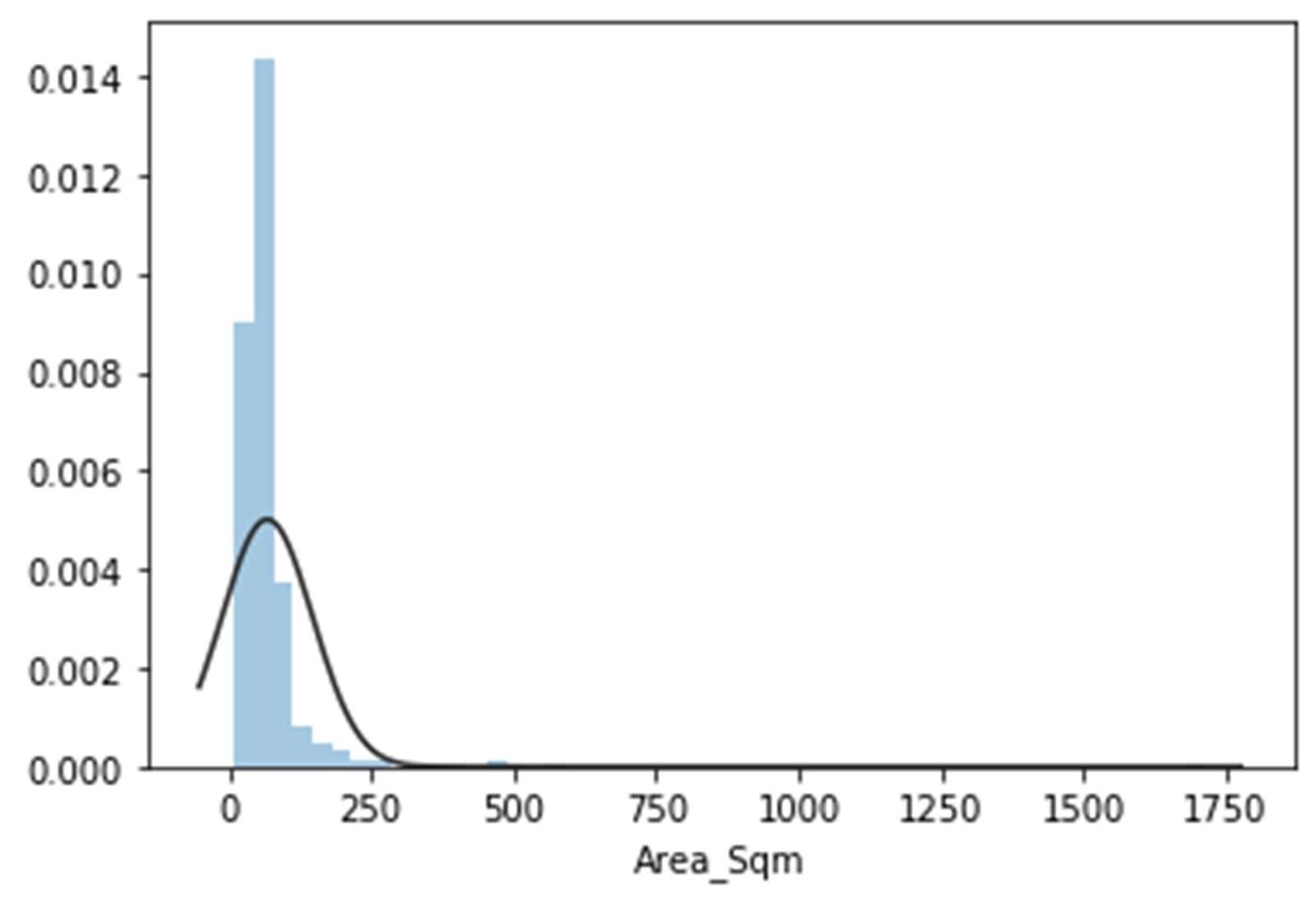
**Skewed Price**



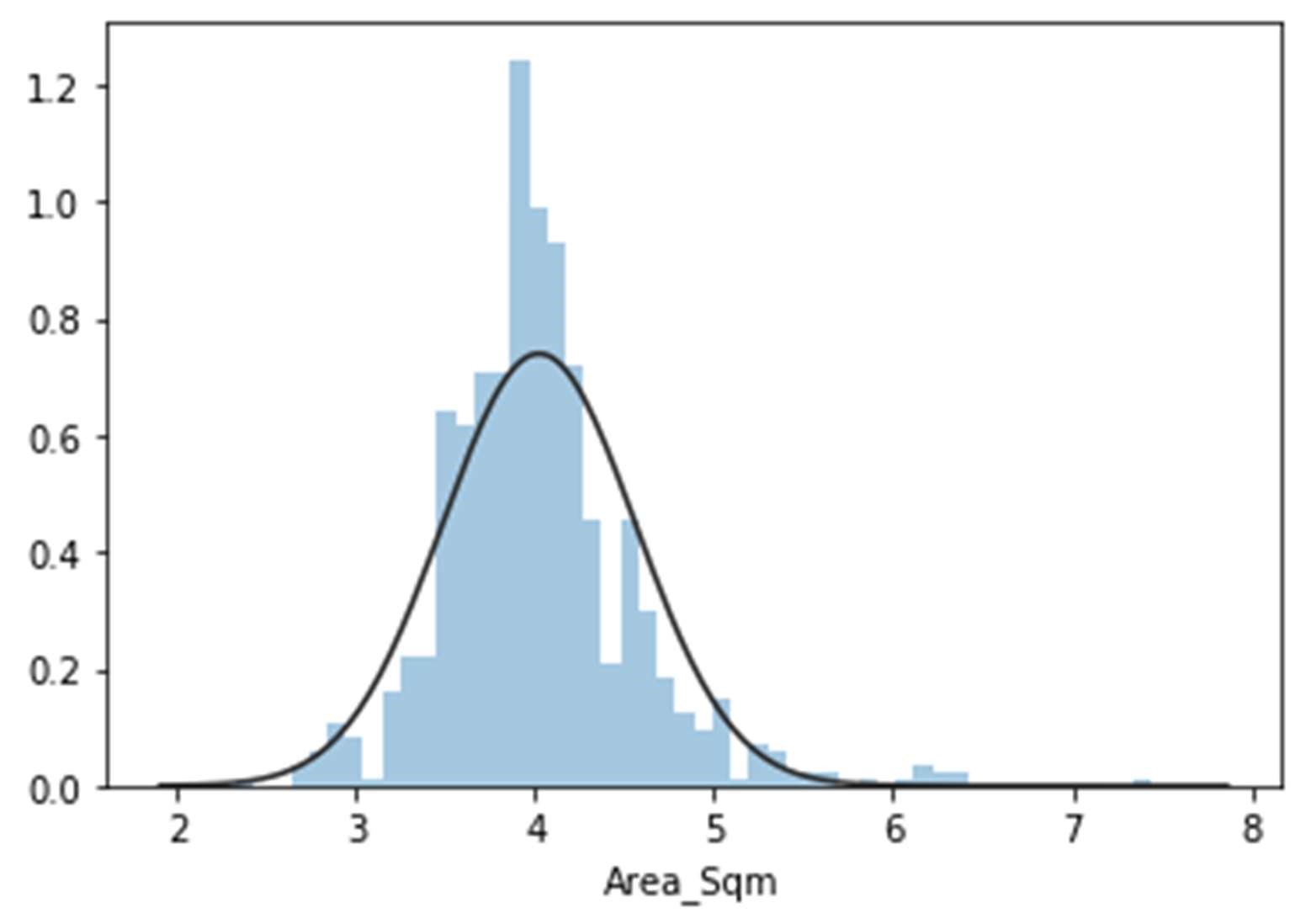
**Normal Price**



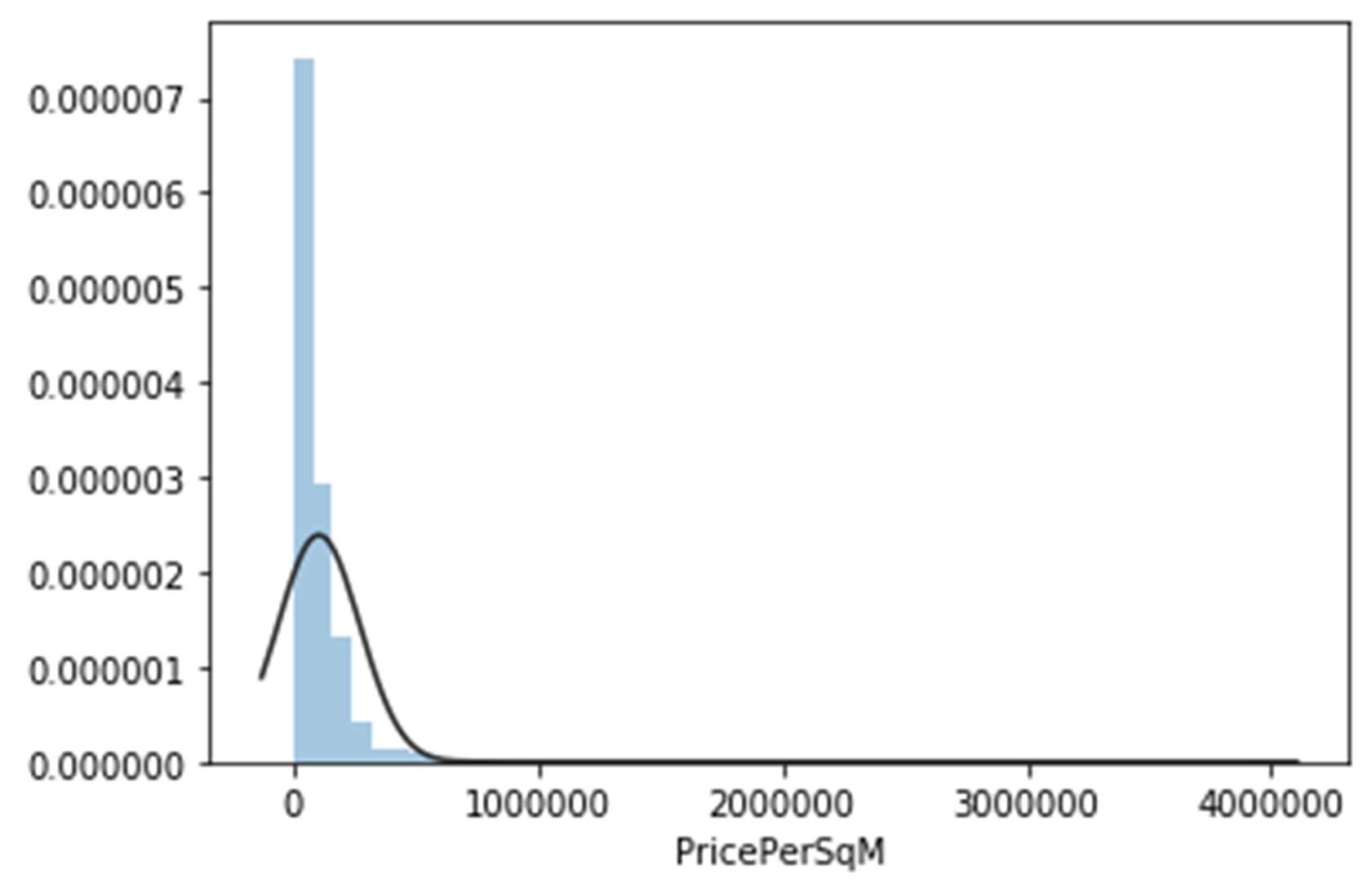
**Skewed Area**



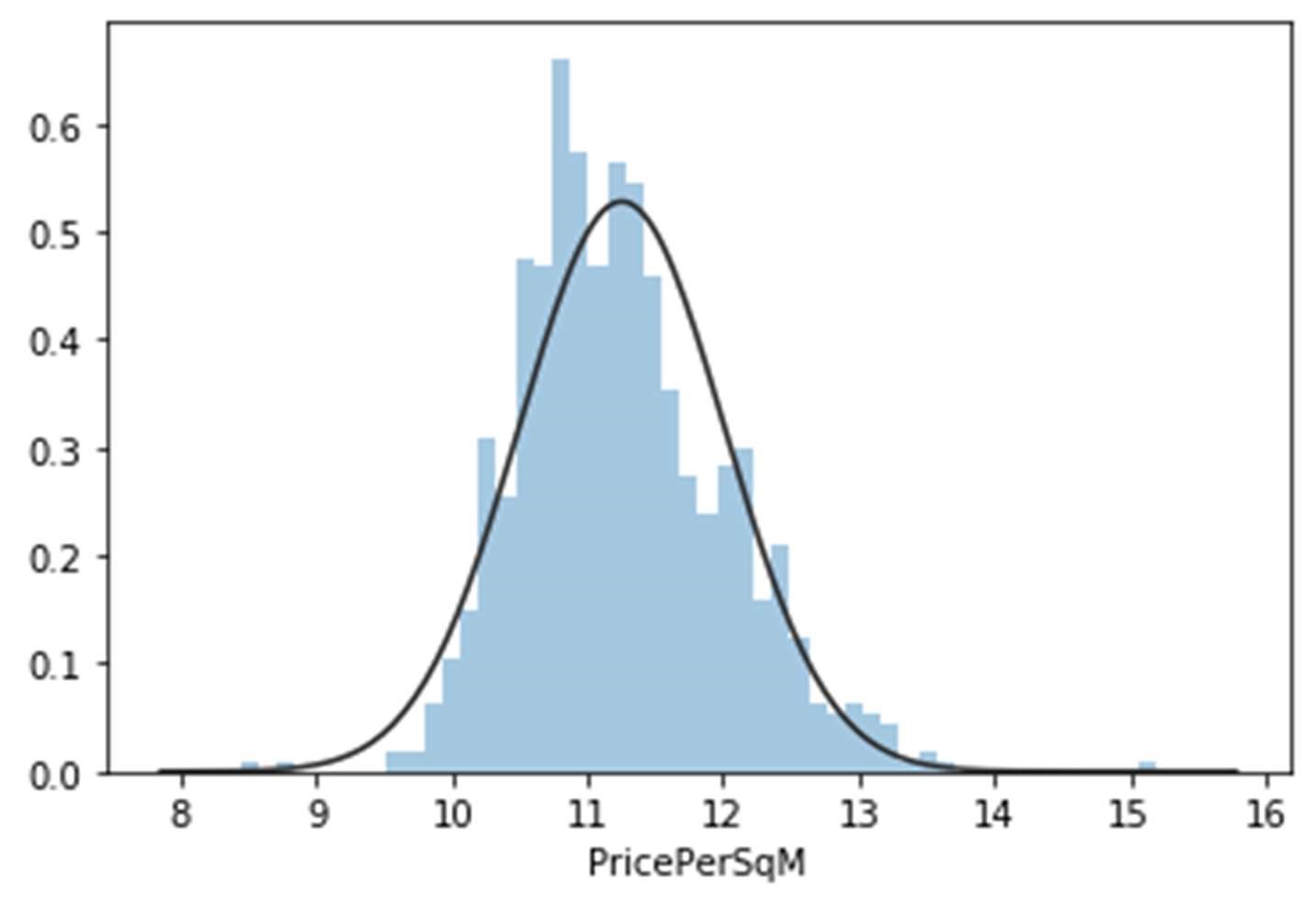
**Normal Area**



**Skewed Price/Sq.**



**Normal Price/Sq.**



# **LANGUAGE AND MODELS TESTED**

## **PYTHON**

Python is widely used in scientific and numeric computing:

* SciPy is a collection of packages for mathematics, science, and engineering.
* Pandas is a data analysis and modelling library.
* NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed.
* The Software Carpentry Course teaches basic skills for scientific computing, running boot camps and providing open-access teaching materials.

## **LIBRARIES USED FOR THIS PROJECT INCLUDE**

* Pandas
* NumPy
* Matplotlib
* Seaborn
* sklearn

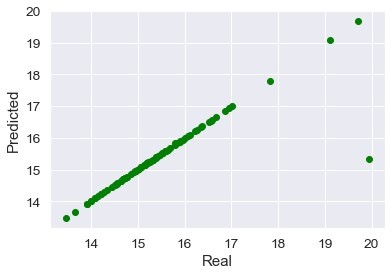
## **MODELS TESTED**

### **DECISION TREE MODEL**

### A decision tree is a flowchart-like tree structure in which an internal node represents a feature (or attribute), a branch represents a decision rule, and each leaf node indicates the result. The root node is the topmost node in a decision tree. It learns to partition based on the value of an attribute. Recursive partitioning is a method of partitioning the tree in a recursive manner. This flowchart-like structure assists you in making decisions. It's a flowchart diagram-style depiction that closely resembles human thinking. As a result, decision trees are simple to comprehend and interpret.

For classification and regression, Decision Trees (DTs) are a non-parametric supervised learning method. The goal is to learn simple decision rules from data attributes to develop a model that predicts the value of a target variable. A tree is an approximation to a piecewise constant.

**Real Vs Predicted**

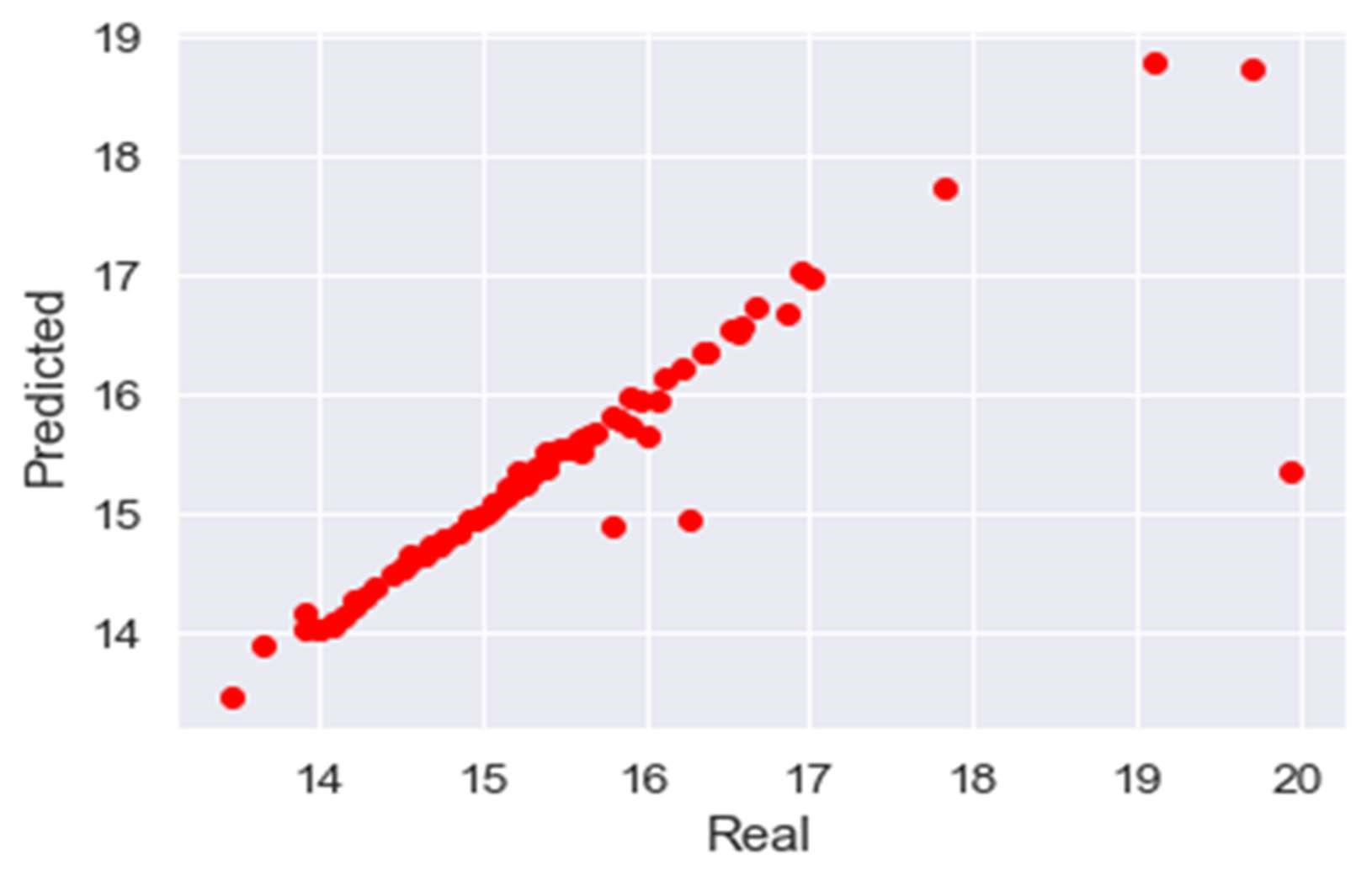


### **LASSO MODEL**

Lasso regression is a type of regularization. For a more accurate forecast, it is preferred over regression approaches. Shrinkage is used in this model. Data values are shrunk towards a central point known as the mean in shrinkage. Simple, sparse models are encouraged by the lasso approach (i.e. models with fewer parameters). This type of regression is ideal for models with a lot of multicollinearity or when you wish to automate elements of the model selection process, such as variable selection and parameter removal.

Lasso Regression is a sort of regularized linear regression with an L1 penalty that is widely used. This causes the coefficients for input variables that don't contribute much to the prediction task to diminish. This penalty allows some coefficient values to be set to zero, thereby removing input variables from the model and allowing for automatic feature selection.

**Real Vs Predicted**



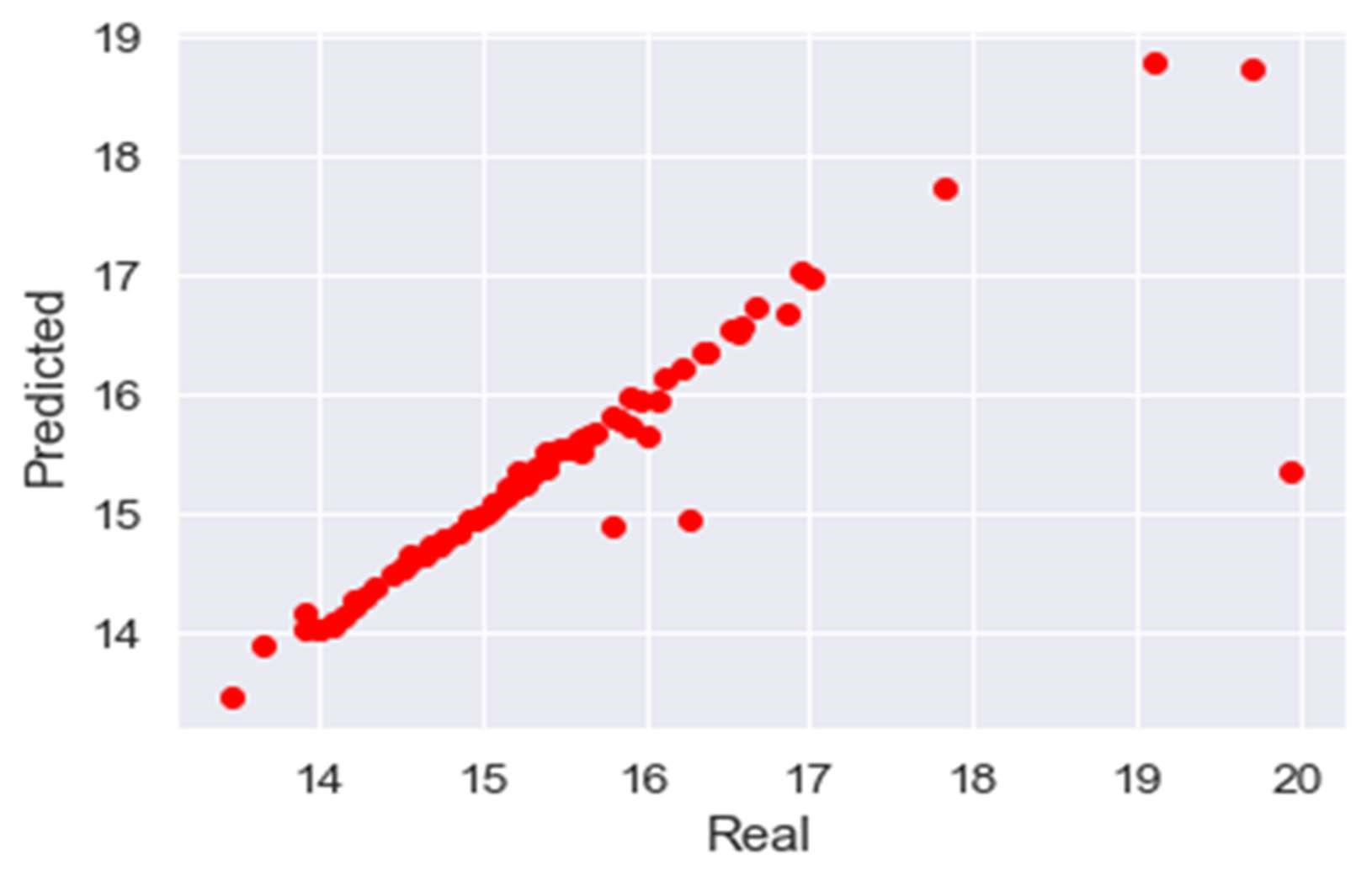
### **LINEAR REGRESSION MODEL**

By fitting a linear equation to observed data, linear regression seeks to model the relationship between two variables. One variable is regarded as an explanatory variable, while the other is regarded as a dependent variable. A modeler might, for example, use a linear regression model to match people's weights to their heights.

A modeler should first evaluate whether or not there is a link between the variables of interest before attempting to fit a linear model to observed data. This does not necessarily indicate that one variable causes the other (for example, greater SAT scores do not always imply higher college grades), but rather that the two factors have a substantial relationship. When determining the strength of a relationship between two variables, a scatterplot can be useful. If the suggested explanatory and dependent variables appear to have no relationship (i.e., the scatterplot shows no increasing or decreasing trends), then fitting a linear regression model to the data is unlikely to yield a useful model.

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

**Real Vs Predicted**



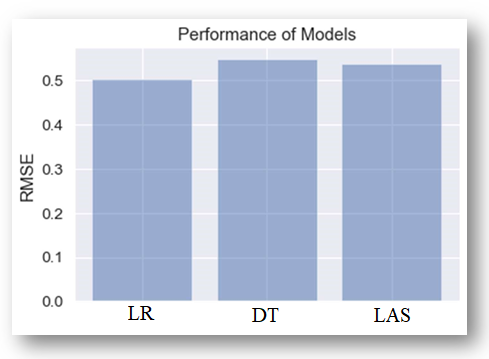
# **RESULTS AND DISCUSSIONS**

# **BEST SUITED MODEL**

# So, our study showed that Decision Tree Model displayed the best performance for this Dataset and can be used for deploying purposes.

# Linear regression model and lasso model are far behind, so can’t be recommended for further deployment purposes.

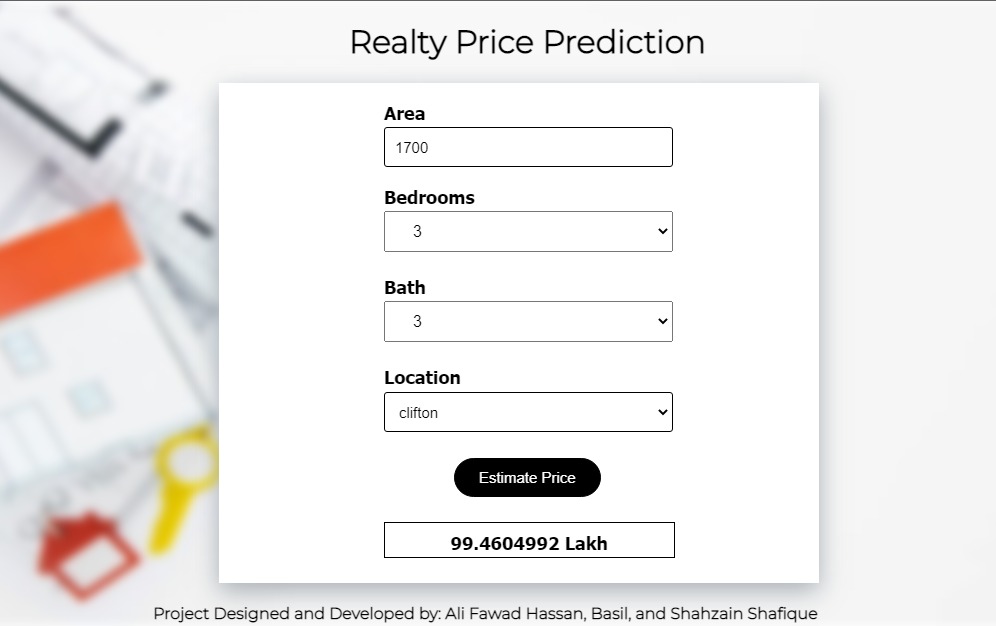
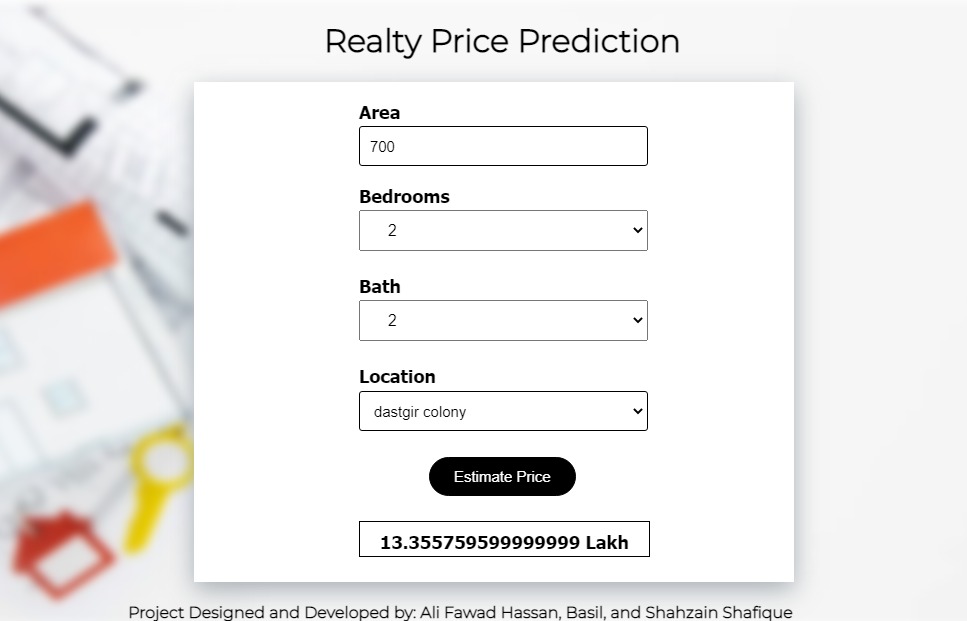
## **RMSE Bar Graph**



## **DEPLOYMENT APP**

## The Model is deployed through Python Web App Flask in collaboration with HTML CSS and JAVASCRIPT.

The Front End designed for the price prediction of houses in Karachi is shown below with some examples:

****

# **CONCLUSION**

In today's real estate environment, storing and extracting such large amounts of data for one's personal needs has become difficult. Furthermore, the information gathered should be useful. The Data Mining Algorithm is used to its full potential by the system. The system makes the most efficient use of such information. By raising the accuracy of estate selection and lowering the risk of investing in an estate, the Data mining algorithm helps to satisfy clients. There are other elements that might be added to improve the system's acceptance. One of the primary future goals is to expand the estate database to include other cities, allowing users to look at more properties and make more informed decisions.

There are a variety of alternative models that can be used to make predictions. The data provided as input to such a model should be compatible with the tool and the operators engaged. Additionally, a larger number of data sets can be used to improve the model's accuracy. The major goal of using a different model should be to reduce calculation time and make the entire process easier to complete.

So, our Aim is achieved as we have successfully ticked all our parameters as mentioned in our Aim Column. It is seen that circle rate is the most effective attribute in predicting the house price and that the Decision Tree is the most effective model for our Dataset with RMSE score of 0.9025658262899986.