

# **Prediction and Analysis of Flat Price in Dhaka Using Advanced Regression Techniques**

**By**

**Fathe Muhammad Rafi**

**ID: 221-25-102**

This Report Presented in Partial Fulfillment of the Requirements for the  
Degree of Master of Science in Computer Science and Engineering

Supervised By

Abdus Sattar

Assistant Professor

Department of CSE

Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**JANUARY 2023**

## APPROVAL

This Project/Thesis titled “**Prediction and Analysis of Flat Price in Dhaka Using Advanced Regression Techniques**”, submitted by Fathe Muhammad Rafi ID No: 221-25-102 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 17-01-2023.

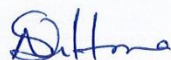
### BOARD OF EXAMINERS



**Chairman**

---

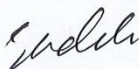
**Dr. Sheak Rashed Haider Noori, PhD**  
**Professor and Associate Head**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University



**Internal Examiner**

---

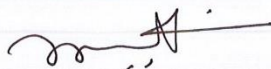
**Ms. Naznin Sultana**  
**Associate Professor**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University



**Internal Examiner**

---

**Mr. Md. Sadekur Rahman**  
**Assistant Professor**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University



**External Examiner**


---

**Dr. Mohammad Shorif Uddin, PhD**  
**Professor**  
Department of Computer Science and Engineering  
Jahangirnagar University

## DECLARATION

We hereby declare that, this thesis has been done by Fathe Muhammad Rafi under the supervision of, Abdus Sattar, Assistant Professor, Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for the award of any degree or diploma.

### Supervised by:



---

**Abdus Sattar**

**Assistant Professor**

Department of CSE

Daffodil International University

### Submitted by:



---

**Fathe Muhammad Rafi**

**ID: 221-25-102**

Department of CSE

Daffodil International University

## ACKNOWLEDGEMENT

First and Foremost, I'd like to express my thankfulness to the all-powerful Almighty as well as our gratitude to our parents. Without their support, we could not have finished our work.

I owe a great deal of gratitude to my esteemed supervisor, "**Mr. Abdus Sattar,**" an **Assistant Professor and Program Coordinator** in the Department of **Computer Science and Engineering** at Daffodil International University in Dhaka, Bangladesh. I finish the project thanks to his insightful knowledge and actionable recommendations. It was made possible to complete this project thanks to his unwavering tenacity, astute approach, constant comfort, steady and furious oversight, pragmatic judgment, significant direction, and reading a lot of substandard manuscripts and fixing them at every level.

I would like to express my sincere appreciation to "**Dr. Touhid Bhuiyan**", **Professor and Head**, Department of **Computer Science and Engineering**, Daffodil International University, as well as to other faculty members and the staff of the Computer Science and Engineering department at Daffodil International University, for their kind assistance in finishing my project.

I'm grateful to all of my classmates at Daffodil International University who participated in this discussion as we finished our course work.

## **ABSTRACT**

Prediction of flat costs could be a crucial space of realty. The literature tries to extract relevant info from historical real estate market data. So as to seek out models that are useful to flat patrons and sellers, machine learning techniques are wont to examine previous property transactions in Dhaka. It is clear that, Gulshan is the costliest area, Mirpur, Mohammadpur, Kallyanpur have similar price and Mohakhali is cheaper compare to other areas. To analysis this model I use different python's library for instance numpy, pandas, matplotlib, seaborn. Scipy and the like. Additionally, tests show that the Advanced Regression Techniques, which rely on mean squared error assessment, are a competitive strategy. To make this endeavor more effective, I gave it everything I had and got a good final prediction, which is R-square 0.87 and normal distribution for error count.

## TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE</b>
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
List of Figure	vii-viii
List of Tables	ix
<b>CHAPTER</b>	
<b>CHAPTER 1: INTRODUCTION</b>	<b>1-3</b>
1.1 Introduction	1
1.2 Motivation	2
1.3 Research Data	2
1.4 Expected Outcome	2-3
1.5 Layout of Report	3
<b>CHAPTER 2: BACKGROUND STUDIES</b>	<b>4-7</b>
2.1 Introduction	4
2.2 Related Works	4-7
2.3 Research summary	7
<b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>8-30</b>
3.1 Introduction	8
3.2 Implementation	8-10
3.3 Data	10-23
3.4 Feature Engineering	23-24
3.5 Python Libraries and Terms	25-30

**CHAPTER 4: EVALUATION AND EXPERIMENTAL RESULT 31-39**

4.1	Introduction	31
4.2	Algorithms	31-37
4.3	Model Performance and Result Analysis	37-39

**CHAPTER 5: CONCLUSION AND FUTURE WORK 40-41**

5.1	Conclusion	40
5.2	Future Work	40-41

**REFERENCES 42-45**

<b>LIST OF FIGURES</b>	<b>PAGE</b>
Figure 1.1: Different Locations in Dhaka City	1
Figure 3.2: Implementation Process	8
Figure 3.3.1.1: Dataset	10
Figure 3.3.1.2: Information of Dataset	11
Figure 3.3.1.3: Description of Dataset	11
Figure 3.3.3.1: Flat Price Based on Size of Flat	13
Figure 3.3.3.2: Flat Price in Different Location	13
Figure 3.3.3.3: Flat price per square feet in Different Location	14
Figure 3.3.3.4: Price with Number of Bed	14
Figure 3.3.3.5: Price with Number of Bath	15
Figure 3.3.3.6: Price with Number of Bed and Bath	15
Figure 3.3.3.7: Average Price of Bed Count in Different locations	16
Figure 3.3.3.8: Average Price of Bath Count in Different Locations	16
Figure 3.3.5: Correlation Heatmap	17
Figure 3.3.6.1: Heatmap of Missing Values	19
Figure 3.3.6.2: Heatmap without Missing Values	19
Figure 3.3.7.1: After Transforming Categorical to Numerical	21
Figure 3.3.7.2: Before Transformation of Price	21
Figure 3.3.7.3: After Transformation of Price	22
Figure 3.3.7.4: Before Transformation of Size	22
Figure 3.3.7.5: After Transformation of Size	23
Figure 3.5: Used Libraries	25



Figure 3.5.1: Numpy	26
Figure 3.5.2: Pandas	26
Figure 3.5.3: Matplotlib	27
Figure 3.5.4: SciPy	27
Figure 3.5.5: Scikit-Learn	28
Figure 3.5.6: TensorFlow	28
Figure 3.5.7: Seaborn	29
Figure 4.3.1: Trained Models	37
Figure 4.3.1: RMSE Analysis	37
Figure 4.3.3: RMSE of Different Models	38
Figure 4.3.4: Actual Result Vs Predicted Result	39
Figure 4.3.5: Distribution	39

## **LIST OF TABLES**

## **PAGE NO**

Table 1: Selected Features

24

Table 2: Model Results

38

Table 3: Final Prediction

38

# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

The difficulty of predicting flat prices can be readily overcome by using machine learning techniques. Considering changes in numerous variables, this can aid in predicting flat price fluctuations. As a country's property values evolve through time, so do house prices.

Additionally, data mining is the process of sorting through huge data sets to find patterns and develop links in order to use data analysis to solve problems. Data mining has also been increasingly used to forecast flat prices. Machine learning uses a variety of categorization algorithms to divide the data into a wide range of possible groups. For feature selection, classification, clustering, prediction, and rule framing, data mining algorithms are frequently utilized. As a result, I employed a variety of algorithms to predict future flat prices using a dataset of flat prices that I had acquired from various real estate companies and websites. The remainder of the paper discusses the project's literature review, dataset and system design, methodology, and the project's outcome, which is followed by a discussion and a conclusion.

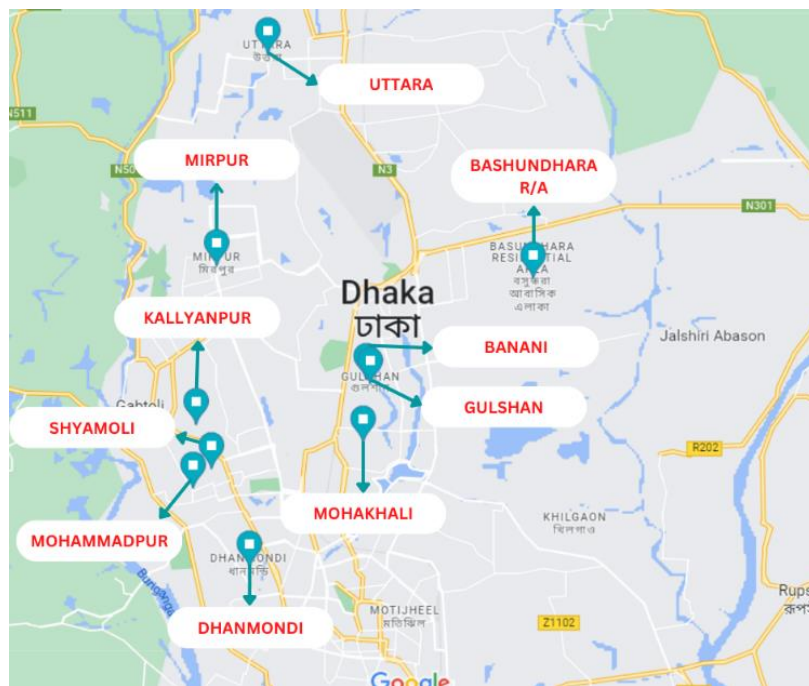


Figure 1.1: Different Locations in Dhaka City

## **1.2 Motivation**

After living in Dhaka for so many years and working in a real estate company for more than 1.5 years, I realized that, estimating flat price of different locations is quite complicated. The price of flat fluctuated significantly for the past few years. Different locations have different price range. For this reason, new customers often confused regarding flat price. Now-a-days, there are some fraud real estate agent who often gives false information to customer or buy flats with cheap price and sell them for higher price. By doing this kind of shameful work, they often cheat with customers.

By using our model, we can easily estimate the price of flat in different areas of Dhaka and get a primary knowledge of flat that how much it could be. So, we can easily help customers for doing this and assist them for making their decisions.

## **1.3 Research Data**

I have created an Excel sheet for storing the information of flats. I have gathered the data from different sources such as real estate Company, websites. Some data was confidential, which I had managed. After that, I assembled all the data and turned it into tabular form, then I had started to analyze the data. By using different python libraries, I had converted the data from qualitative to quantitative for analyzing. There had some useless feature that I had removed at the beginning of analyzing, such as pictures. All the data are primary data and collected by myself.

## **1.4 Expected Outcome**

Flat prices heavily rely on location, making accurate predictions of flats is quite challenging. By giving more attention to the noise of data, we can estimate the most probable estimating price. So, I had focused of different criteria such integration of multi-source data and appropriate machine learning techniques to generate more accurate and reliable output. I mainly focused on locations, bed, bath and size for predicting the price of flats and removed unnecessary features which had created an unbalanced situation.

Finally, the criteria on which I will evaluate this proposal are as follows:

1. Develop a reliable pricing prediction model.

2. Check the accuracy of the model's predictions.
3. Recognize the important flat rate elements that aid the model's prediction ability.
4. Make an effort to comprise the appropriate patterns using advance regression techniques.

## **1.5 Layout of Report**

First, the preamble first chapter discusses the goal of the study, which is what initially inspired the author to take on this particular issue statement. I will briefly discuss the goals of this research as well as the general context for the study in this part.

We looked at other publications from the computer science field that dealt with the same kinds of issues in the second part of the essay, which is the existing literature section (Chapter 2). Additionally, the goal of the background investigation was to identify any gaps in the earlier research that might have existed.

The data Methodology phase, which is covered in Chapter 3, contains specifics about the data gathering and conversion process that will be used in this project. The extraction of features and thorough investigation of the contributions made by each feature to the overall conclusion are also included in this step. We examine several data types in Chapter 4 of the book, "Implementation and Model Selection," determine which method is best for distinct patient data subsets, and then select the proper approach.

Additionally, the "Results and Study" chapter (4) included the proposed models as well as a comparison analysis of the prediction rate for each model. The study's findings are also summarized in a presentation.

## CHAPTER 2

### BACKGROUND STUDIES

#### 2.1 Introduction

This section will cover several earlier works that are connected to this subject. Real estate market is the costliest market for every country. So every year a bunch of analysis are done by researchers to find the appropriate patterns. Some of these work will be discussed in related work section.

#### 2.2 Related Work

As civilization progresses, the demand for housing will gradually increase. The cost of an apartment depends on many factors. The real estate industry has a huge impact on the economy as a whole. Buyers, real estate agents, and financial professionals all benefit from flat rate forecasts. Predicting actual home prices is important to prospective homeowners, developers, investors, and tax assessors. Accurate fixed-price forecasting algorithms fill information gaps, making decisions easier for both buyers and sellers. Due to their amazing benefits over conventional methods, machine learning algorithms have recently become frequently used for flat price prediction. [1]. They can make educated decisions on whether to buy a property and when is the best time to do so with the aid of a computer-based prediction system [2]. The house price index reflects changes in the value of homes' prices. For the projection of single-family home prices based on elements such house type, size, year of construction, amenities, and other elements that affect the demand and availability of homes, more exact methodologies are necessary. The authors examine composite pre-processing and feature extraction technique while just considering a few characteristics and datasets. A hybrid Gradient boost regression and Lasso model had previously been developed to predict the value of a single house. When a consumer wishes to oversee and identify relevant real estate for their investments, they typically contact a real estate agent. This strategy carries a significant level of risk because a bad prediction could result in clients losing their investment. Customers currently face high risk due to the manual estate valuation method that is applied [3]. Real estate is an important sector to boost the national economy and expand China's GDP. Real estate valuation and forecasting can help investors formulate sound investment strategies, enact appropriate government regulatory laws and long-term growth of the real estate

market [4]. For many buyers of homes, the distance to social and cultural hubs might be an important factor. Facilities like libraries, schools, and sports complexes are frequently visited by those customers who have assimilated into their daily lifestyle, since children need to attend to school and increase their cultural and physical education. The distance between the starting point and the destination is strongly correlated with the commute time. Greater convenience for all home members is provided by closer proximity, which raises the cost when comparing various solutions [5]. Models for predicting home prices aid consumers in making home purchases or investments. Consumers can determine whether house prices will rise or decline in the future, the size of the change in house prices, etc., as well as the rate of return on investment in real estate by using the house price prediction model. Consumers can use this model to evaluate if they should buy a local property in that year depending on whether they have an immediate need for one and whether the market is now seeing a price drop [6].

Housing is a key factor in determining family incomes and, as such, is one of the factors that explains household expenditure, which makes up around 60% of GDP in industrialized countries overall (55% of which is made up of consumption and 5% of which is made up of residential investment) [7]. Real Estate a person's first desire is property, which also serves as a barometer of their wealth and status in today's society. Real estate investments frequently seem profitable since property values do not decline suddenly. Numerous real estate investors, bankers, policymakers, and others will be impacted by changes in the value of the property. It seems like a seductive opportunity for investors to invest in real estate. Predicting the important estate price is hence a crucial economic indicator. With a total of 24.67 crore households, the Asian nation ranks second globally in terms of household numbers, according to the 2011 census [8]. The price is affected by a number of variables, including the interest rate, the cost of home ownership loans, the cost of building materials, and the minimum salary for employees. With so many different property designs and features available, potential buyers can occasionally feel overwhelmed when making a decision [6] [9] [13] [19]. The researchers faced a lot of drawbacks in this study. The biggest restriction is that we don't know anything about possible buyers or the sale's surroundings. Due to bidding wars and ego, factors like auctions can affect the price of a house. Since the data were collected over the course of a single year, little seasonality and economic considerations were taken into account [10].

In their study [11], Adding unquantifiable factors can help the model's prediction accuracy to some extent because they have a significant impact on house prices but weren't included in our research's model for predicting house prices. These unquantifiable factors, like policies, are those that have a significant impact on house prices. There may be some variation between the final forecast findings and the actual results due to the influence of unquantifiable factors. The housing market differs from every other market (Smith, 2011b) [14]. After completed their analysis he found a lot of problems and errors importantly in root mean square [15]. A key driver for predicting house prices is the correlation between housing costs and the economy. To assist buyers and sellers in their judgments, it is crucial to predict property prices objectively. To better and more accurately estimate home price trends, this project is being proposed. Various algorithms are used in this study to determine which produces the most exact and precise results [16]. Due to its generality, the Hybrid Regression approach is straightforward but significantly more effective than the other preceding methods [18]. The use of univariate and multivariate linear and polynomial regression, concluding that such methods, when applied individually, are fairly ineffective to complete the task. They suggested that a mixed model would produce better results than either one used alone, while affirming that high order polynomial regression tends to over fit the data while linear regression models tend to under fit it [19]. Unlike other research on the sale of houses, this one includes a component that affects the sale price rather than just a model for sale. They are unable to use feature engineering approaches like Principal Component Analysis and others since they could alter the outcome while making interpretation more difficult. They do not take ensemble methodologies for prediction into account for the same reason [20]. It can be said that using machine learning algorithms is a difficult process with many stages, and that these algorithms outperform conventional linear approaches. Since the value of the algorithms relies on the problem to be addressed and the type of data to be used, it is impossible to generalize that one approach is superior to another (tabular data, text, image, sound, etc.). All ML studies must also address the issue of data leakage and examine whether over fitting in the employed algorithms may lower the prediction accuracy [4] [8] [16] [21]. For the house price forecasts in the test data set using unseen data, they compared various performance metrics. It is clear that the RF performs better in this housing market than the other approaches when the  $R^2$ , RMSE, and MAPE values are taken into account. The complexity or nonlinearity of the housing markets in Boulder, Colorado, appears to be better captured by the RF technique than by other models [22]. According to this comparative



analysis, the Lasso regression model performed better than the Ridge regression model. Compared with the Ridge regression, the Lasso regression has a smaller mean square error and a higher adjusted R squared error. The Lasso regression had an adjusted R-squared value of 0.90. When 18 predictor variables were taken into account, the selling price of a home fell by 90%. A higher adjusted R-squared value indicates that the Lasso regression model is a more efficient model [23]. They take RMSE into account since the performance matrix is used to different datasets and these algorithms to find the best accurate model that predicts better outcomes [25]. Genuine domain is the slightest straightforward industry in our biological system and the genuine domain advertise is one of the foremost centered with respect to estimating and always fluctuating, individuals are cautious when they are attempting to purchase a modern house with their budgets and showcase strategies as well as venders when they are attempting to offer a house [26]. According to Zillow, "We have been unable to accurately forecast future home prices at various times in both directions by a significant amount" [27]. Method to handle several attribute types and merge them is a significant challenge in forecasting regression issues [28]. In arrange to expect the genuine bequest cost, Chernyshova et al. inquired about the interface between supply and request, which is the genuine domain cost shaped beneath the impact of social, financial, and fabric components [29]. They built a genuine domain record forecast demonstrate utilizing eleven machine learning models, compared the comes about of the eleven models, and came to the conclusion that support vector regressor performed better than other algorithms [30].

## **2.3 Research Summary**

Estimating flat price is one of the crucial part of real estate market. As price range vary from area to area, it is often difficult to predict a real price. Many fraud real estate agent takes this opportunity to cheat with customers. For this reason, this analysis has done to estimate an approximate price for the customer. For this research, data has been collected from various resources. Previous research has much more lacking, such as shortage of data that I tried to overcome in this research. I have applied different techniques for cleaning, merging the data. To estimate the probable price, I have used different advanced regression techniques. By doing this analysis, I found that, there has a huge gap between areas even though having same bed and bath facilities. In the future, it can be covered up the whole country with an implemented model to create an impact on real estate market.

## CHAPTER 3

### RESEARCH METHODOLOGY

#### 3.1 Introduction

For potential homebuyers, designers, financial specialists, appraisers, charge assessors, and other genuine bequest advertise members counting contract banks and guarantees, an exact projection of the house cost is vital. The foundation for the daily rise in housing demand is the advancement of civilization. The ability to estimate home values accurately has always piqued the interest of buyers, sellers, and bankers alike. A housing market is any market for real estate where prices are negotiated directly between sellers and purchasers or via the help of real estate brokers. As a result of the high need for housing globally, both individuals and businesses are lured to this sector. Demography, the economics, and politics are only a few of the variables that affect these demands. As a result, machine learning builds techniques and algorithms from data and utilize them to predict the results based on new data [16]. In this part, I am going to discuss Data, Features, Libraries and Implementation process.

#### 3.2 Implementation

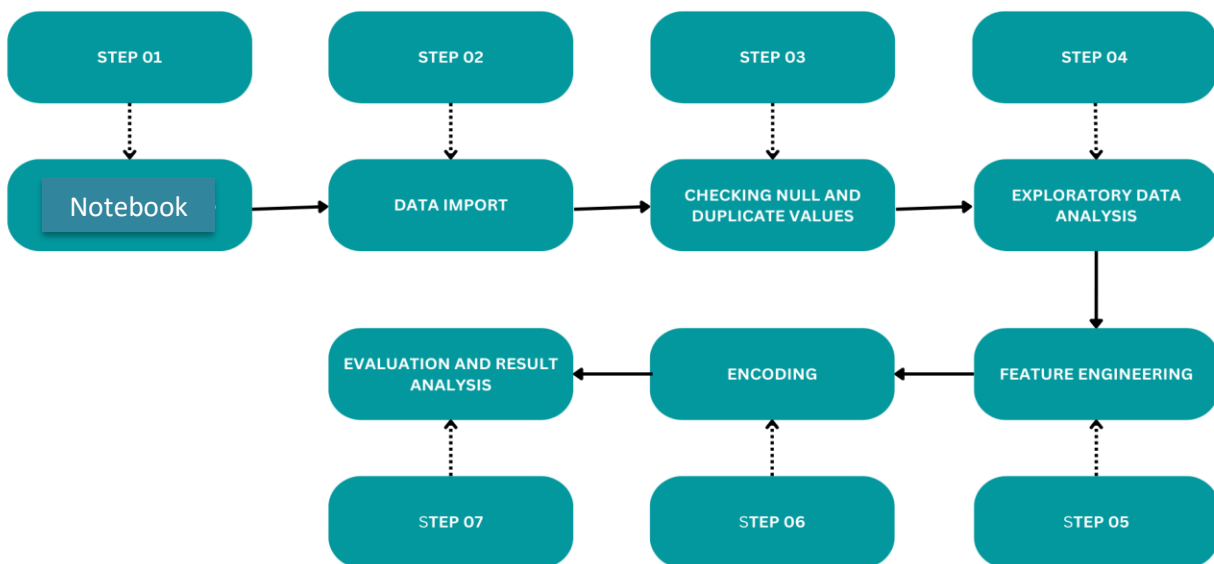


Figure 3.2: Implementation Process of Model

For analyzing, we have to use python libraries such as numpy, pandas, matplotlib, seaborn, scipy and the like. These libraries are used to organize and visualize data. There have a bunch of python libraries we can use to analyze data. Before working with data, we must have to import these libraries.

In the next step, we have to import all data that had collected before. There have several formats to save these data. But widely used saving format is comma separate value in short CSV. But we can import with other saving format. If there have any problem regarding column name, we can simplify in this step.

Then we have to deal with the data. In this step, null values have to check. If there have any null values, we must have to remove these value using python library. For this part, pandas library can be used, but we can also be used other techniques. Sometimes, there have other issues occurred that create an impact on result and one of the most common problems is duplicate values, moreover we can use libraries for removing duplicate values.

In the data exploratory step, we can visualize our data using libraries such as matplotlib, seaborn and the like. It is mandatory because we could not generate the idea about data without visualization. In numeric format everything is not clear, and we can't see correlation between attribute without visualization, it is also used to determine the pattern of data and how the data distributed. To know more about the connection between attributes, we use this step.

One of the most important step is feature engineering. Because all the features are not important, so we can't take all the features because it reduces effectiveness of the result. To generate more accurate final output, important features have to be determined.

After that, we must have to encode our data. Encoding means transferring qualitative data to quantitative data. This is just because existed algorithm can't work with text data because it is a mathematical analysis. For conversion, we can use libraries such one hot coding, standard coding and other pre-processing techniques. Besides, we can normalize our data into a constant scale in this stage.

Finally, after completion of all these step, we can start evaluation and result analyzing. To do this, we can use different algorithms based on the data and problem. There have several built in

algorithms such as supervised (classification, prediction), unsupervised (clustering, association) and reinforcement (control).

### 3.3 Dataset

In this part all information regarding will be discussed.

#### 3.3.1 Data Collection

This study uses primary data which had collected from several real estate companies and also scraped form the well-known real estate websites [5].

	Price_Crore	Location	Bed	Bath	Size_Sqft
0	1.90	Uttara	3.0	3.0	2,024
1	1.90	Uttara	3.0	3.0	2,004
2	3.90	Uttara	4.0	4.0	3,334
3	0.70	Uttara	3.0	3.0	1,289
4	2.55	Bashundhara	3.0	3.0	2,200
...	...	...	...	...	...
1370	0.52	Mirpur	3.0	3.0	1440
1371	0.45	Mirpur	3.0	3.0	1403
1372	0.45	Mirpur	2.0	2.0	850
1373	0.46	Mirpur	2.0	2.0	850
1374	0.57	Mirpur	3.0	3.0	1100

Figure 3.3.1.1: Dataset

Shape of this data set is 1375 rows and 5 columns.

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1375 entries, 0 to 1374
Data columns (total 5 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Price_Crore    1375 non-null   float64
1   Location        1375 non-null   object
2   Bed             1342 non-null   float64
3   Bath           1342 non-null   float64
4   Size_Sqft      1375 non-null   float64
dtypes: float64(4), object(1)

```

Figure 3.3.1.2: Information of Dataset

	Price_Crore	Bed	Bath	Size_Sqft
<b>count</b>	1375.000000	1375.000000	1375.000000	1375.000000
<b>mean</b>	1.520887	3.061818	3.189818	1707.101818
<b>std</b>	1.157353	0.603089	0.794567	738.662223
<b>min</b>	0.170000	1.000000	1.000000	250.000000
<b>25%</b>	0.780000	3.000000	3.000000	1250.000000
<b>50%</b>	1.200000	3.000000	3.000000	1550.000000
<b>75%</b>	1.875000	3.000000	4.000000	2050.000000
<b>max</b>	9.800000	7.000000	7.000000	8400.000000

Figure 3.3.1.3: Description of Dataset

### **3.3.2 Data Exploration**

Descriptive analytics, the first stage of data analysis, frequently comprises describing the data set's important characteristics, such as its volume, structure, correctness, early relationships in the data, and other attributes. Data analysts usually utilize visual analytics tools to perform this task, although Python, a more advanced statistical language, can also be employed. Before doing analysis on data collected from several sources of data and stored in data warehouses, an organization must determine how many instances are in a given dataset, what characteristics are included, how many missing values there are, and what general assumptions the data is inclined to support. The preliminary analysis of the data set may aid in answering queries of the research questions by acquainting researchers with the dataset they are working with. For training and testing purposes, I divided the data 8:2 each.

### **3.3.3 Data Visualization**

In visualization techniques, information and data are visually displayed. By integrating visual components like as bar chart, info graphics, and mappings, data visualization tools provide an easy method to identify and grasp patterns, anomalies, and relationships in the data. Data visualization technological tools become fundamental in the realm of big data for analyzing vast volumes of data or generating data-driven choices. Existing technologies, including as tableau, power BI, and others, allow us to view data.

In this visualization part I will add some pictures of my analysis. Where I have shown flat price based on flat size, flat price in different locations, per square feet price of flat in different locations, price with number of bed, price with number bath, average price of bed and bath.

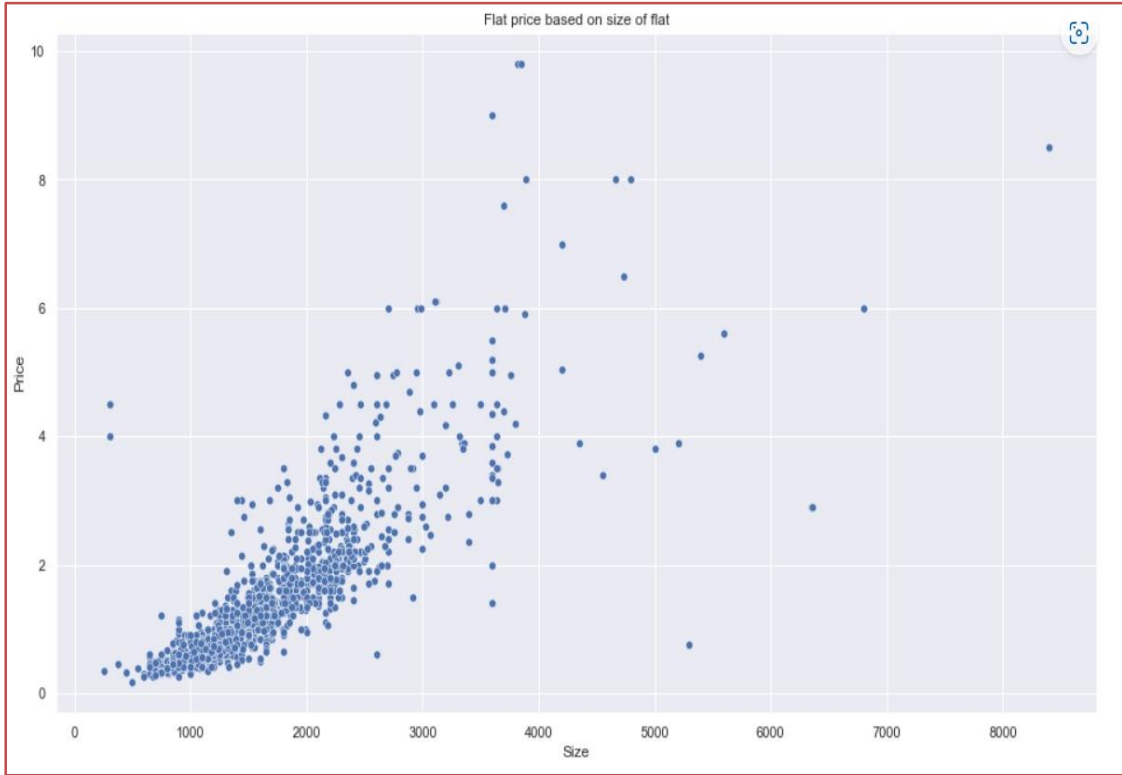


Figure 3.3.3.1: Flat Price Based on Size of Flat

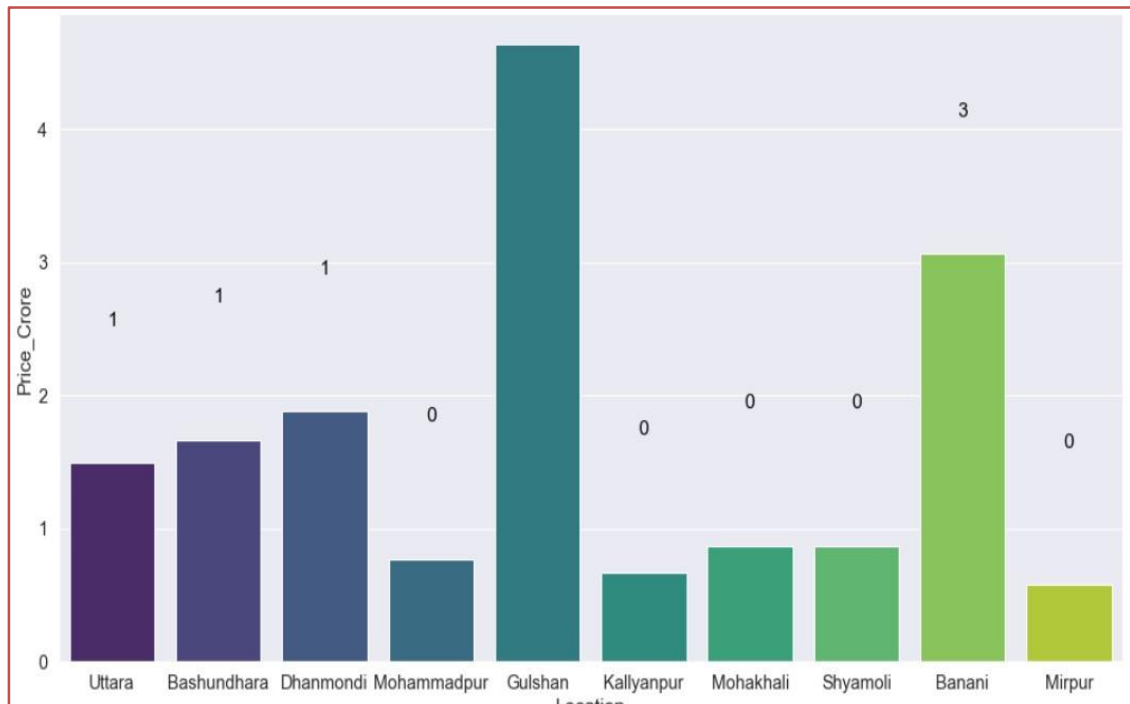


Figure 3.3.3.2: Flat Price in Different Location



Figure 3.3.3.3: Flat price per square feet in Different Location

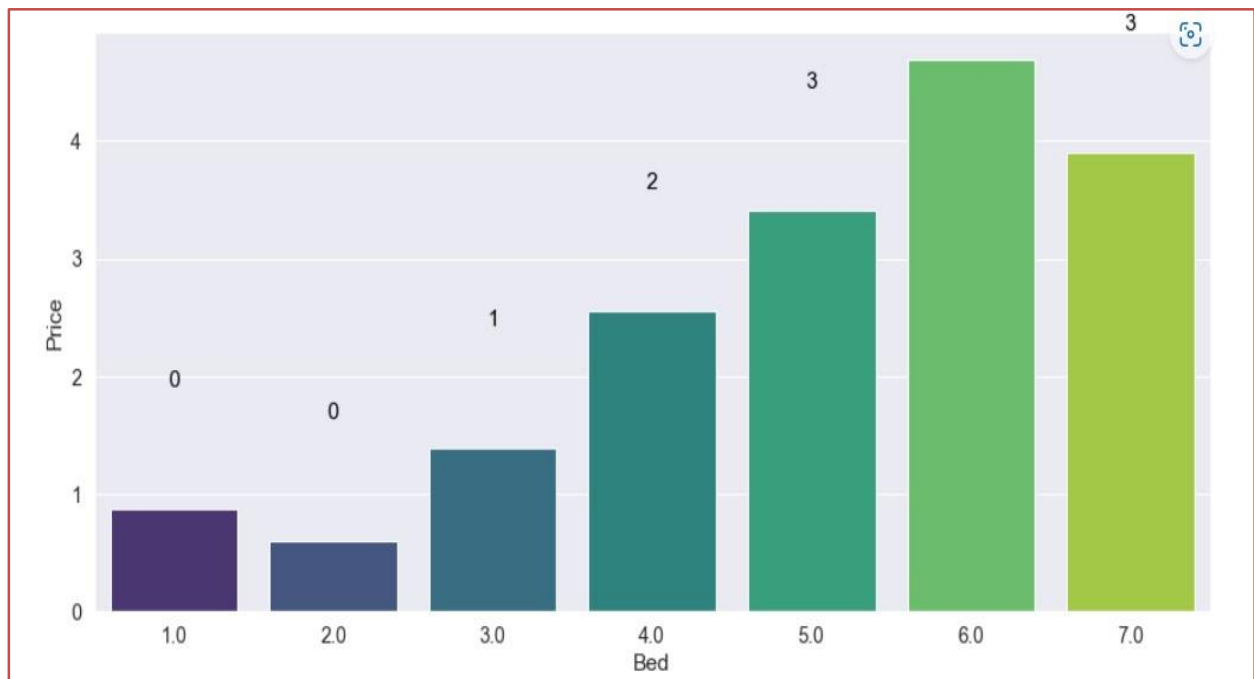


Figure 3.3.3.4: Price with Number of Bed



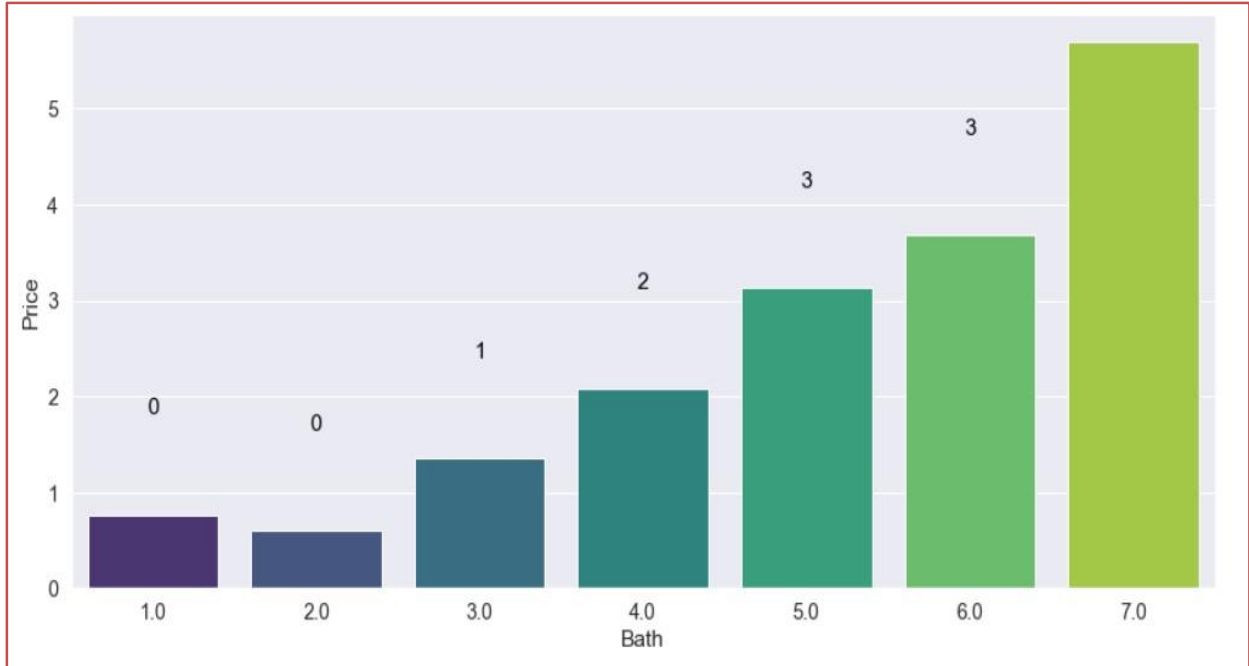


Figure 3.3.3.5: Price with Number of Bath



Figure 3.3.3.6: Price with Number of Bed and Bath

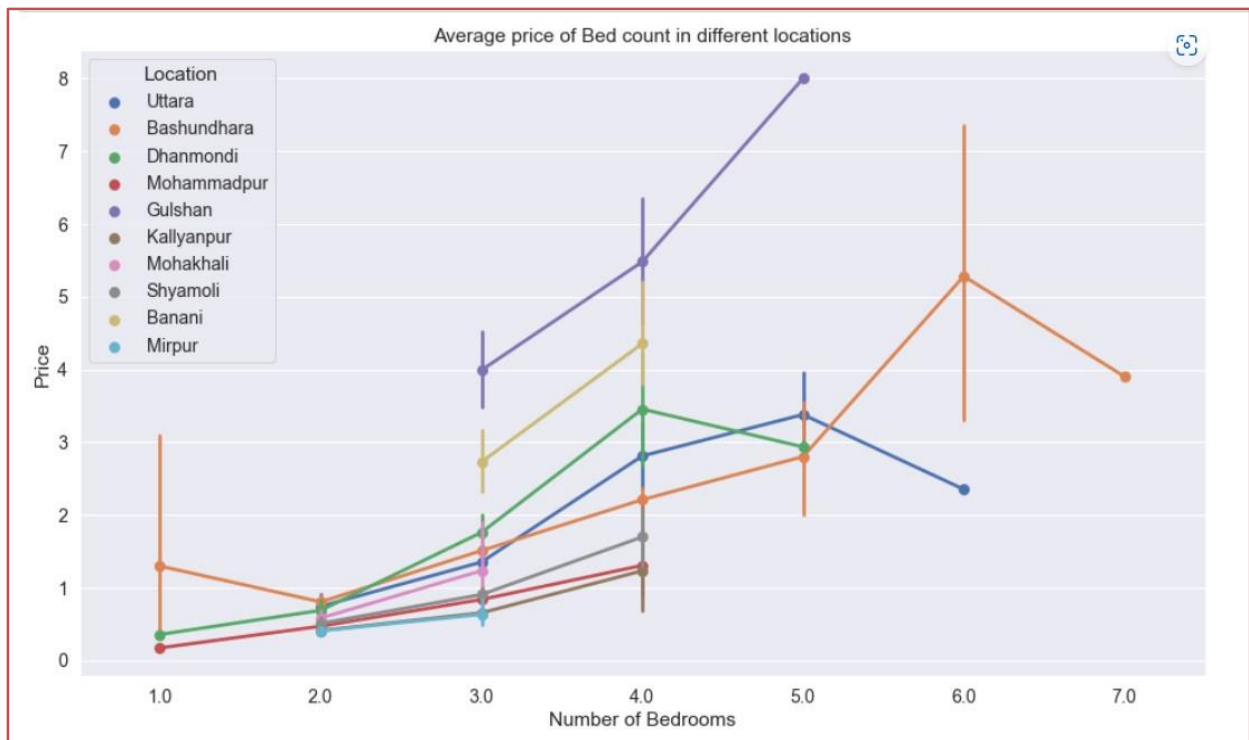


Figure 3.3.3.7: Average Price of Bed Count in Different locations

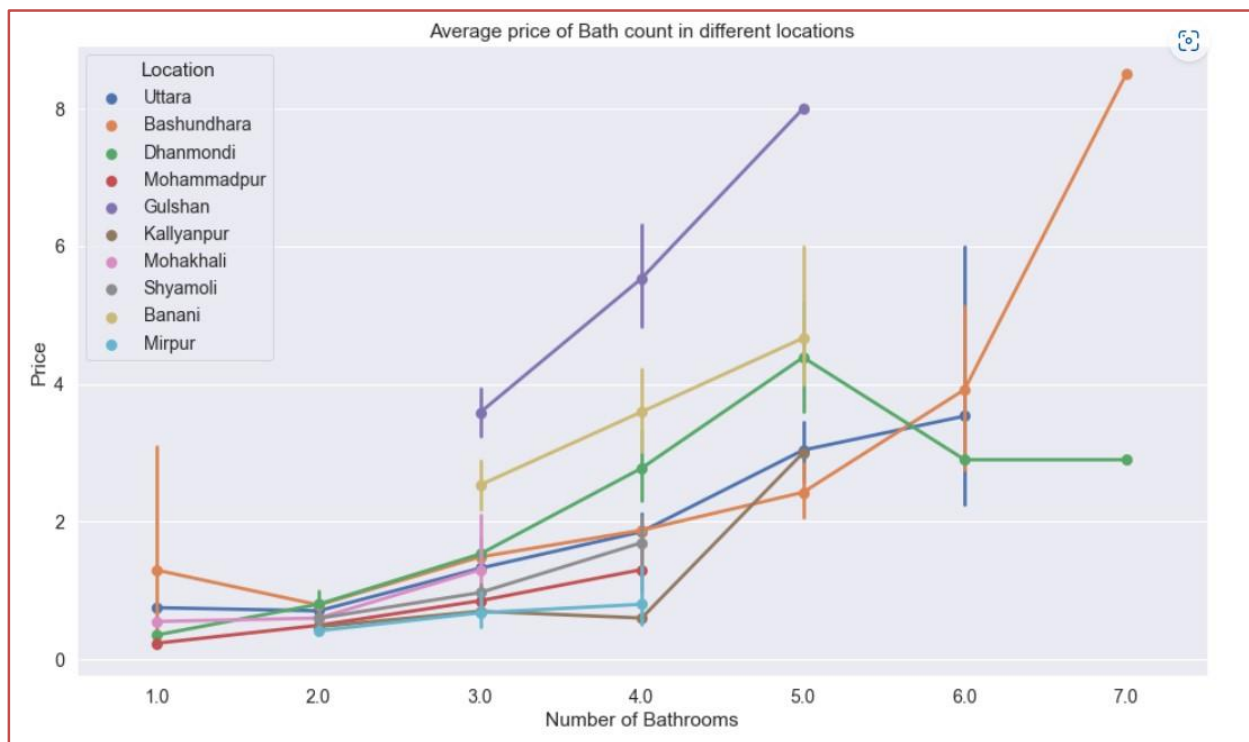


Figure 3.3.3.8: Average Price of Bath Count in Different Locations

### 3.3.4 Data Visualization Summary

After utilizing a graphical tool to analyze the information, it became clear that Gulshan apartments are costlier than those in other places. Prices for apartments with additional bedrooms and bathrooms are often higher. However, despite the fact that gulshan apartments have fewer bedrooms and bathrooms, the price per square foot in this neighborhood is higher than in any other places, and as a result, the price range immediately increased. However, Mohammadpur and Mohakhali have the most affordable prices for apartments. Additionally, locations next to Mohammedpur (Shyamoli, Kallyanpur) are likewise less expensive. Flats in Mirpur cost an average amount. There are more beds and baths in Bashundhara flats than in other places.

### 3.3.5 Data Selection

Data selection is the method of determining on the best source of data, type, and equipment to collect the data. The actual action of data collection begins before data selection. This concept distinguishes between interactive/active data selection (using collected data to monitor activities/events or conduct secondary data analysis) and selective data reporting (selectively removing data that is not supportive of a study premise). The approach used to choose acceptable data for a research project may impact data integrity.

The basic purpose of data selection is to choose the appropriate data form, origin, and instruments that will allow researchers to handle the problem of the study correctly. This option is frequently discipline-specific and is heavily impacted by the nature of the inquiry, the volume of preceding research, and the ease of access to relevant data sources.

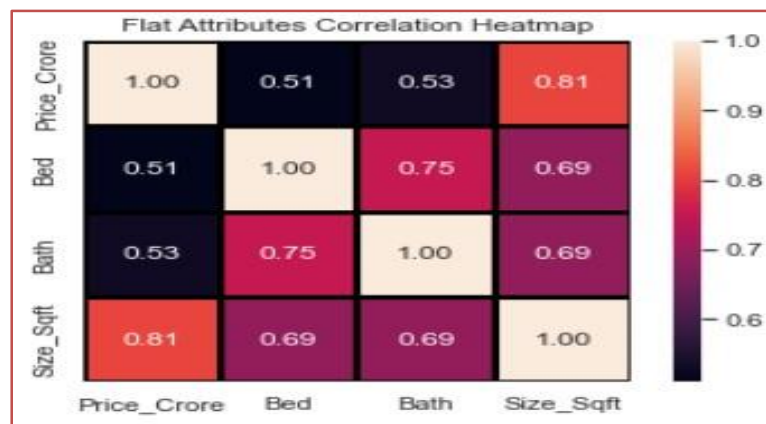


Figure 3.3.5: Correlation Heatmap

### 3.3.6 Data Preprocessing

A step known as data preprocessing in the machine learning and assessment process takes raw data and turns it into a format that computers and machine learning algorithms can comprehend and analyze. Text, images, videos, and other sorts of raw, real-world data are inefficient. It is frequently deficient and lacks a regular, consistent design, and it may be erroneous and irregular. Machines like to process information that is cleanly arranged; they interpret data as 1s and 0s. Structured data, such as full numbers and percentages, may therefore be computed quickly. However, unstructured material, such as text and photographs, must be cleaned and processed before analysis. Gathering information from several sources, you may receive data in a variety of formats. While the end goal of the method is to reformat your data for machines, you must begin with similarly prepared data. If your project investigates family income from many countries, you must translate each amount into a single currency. Outliers may have a substantial influence on data analysis results. If you averaged the test results for the whole class, one student's 0%, for example, might dramatically affect the numbers. Check for any blank text or survey questions, as well as any missing data fields. This might be due to incomplete or erroneous data. To solve the issue of missing data, you must first clean the data. The process of adding missing data, filling gaps, and removing erroneous or irrelevant information from a data collection is known as data cleaning. The most important preprocessing step in ensuring that your data is fit for use for your downstream objectives is data purification. The data cleansing will resolve any discrepancy that your data quality check uncovered. Depending on the type of data you're working with, you might need to apply one of a number of cleansers to your files. There are many different techniques and technologies for preprocessing data, including the following:

- Sampling is the method of choosing selection of data from a vast population of data.
- Transformation alters unprocessed data to produce a single input.
- Using denoising, data may be made noise-free.
- Imputation, which generates data with statistical significance in place of missing values.
- Organizing data through normalization makes it easier to access.
- Feature extraction is the process of determining a meaningful subset of features that is significant in a given context.

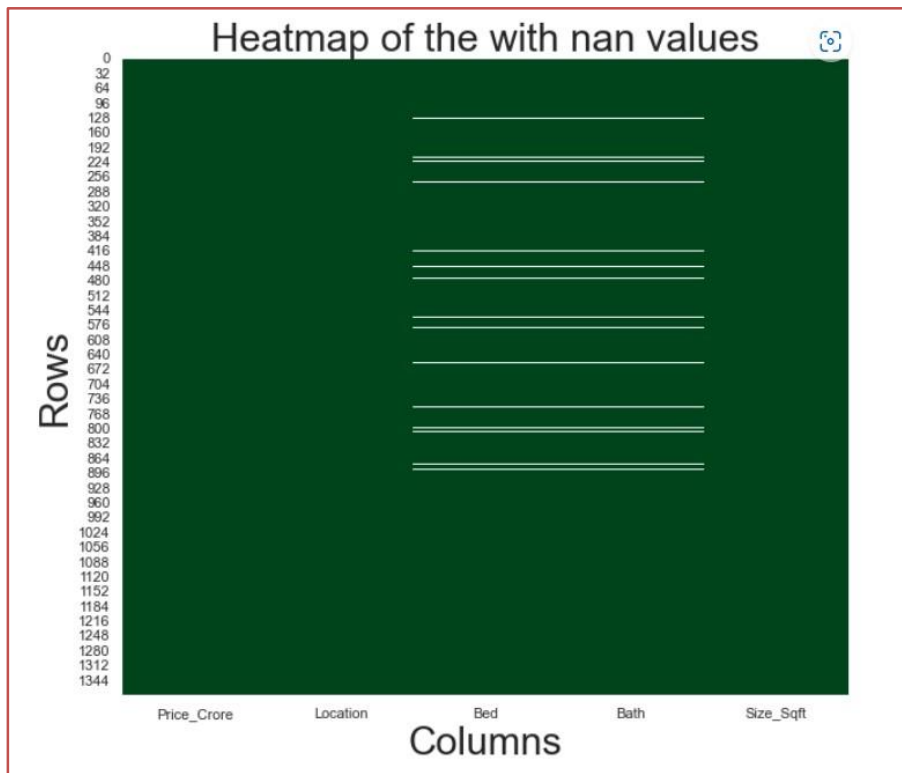


Figure 3.3.6.1: Heatmap of Missing Values

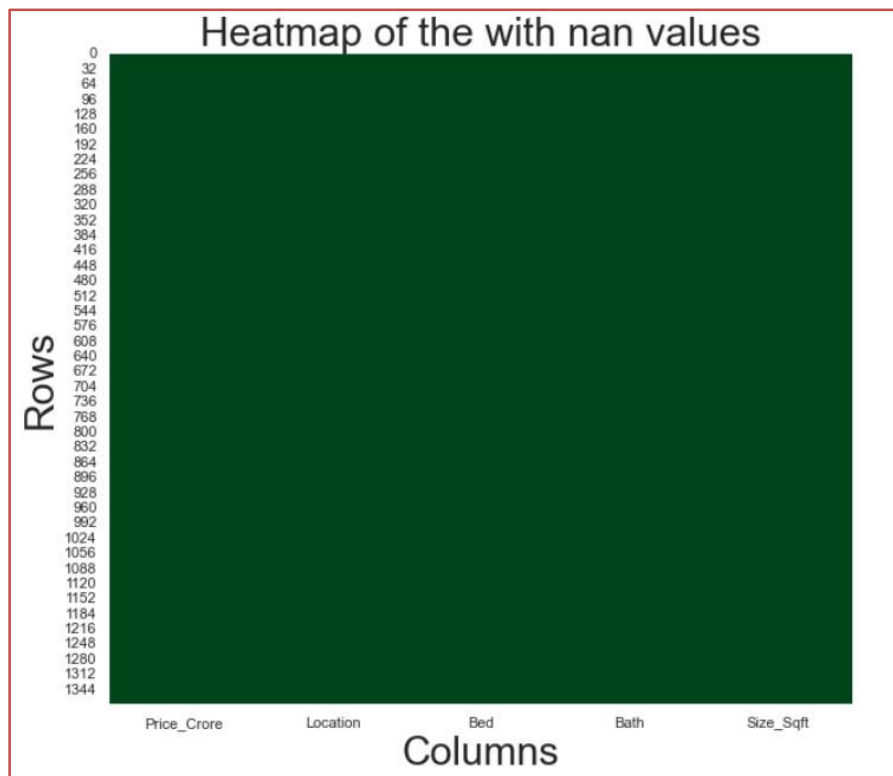


Figure 3.3.6.2: Heatmap without Missing Values

### **3.3.7 Data Transformation**

Data transformation is a technique for converting data to a machine-readable format. Data stored in various formats such as integer, float, text, image, and the like. So, these data must be converted into a machine-readable format. There are some libraries and preprocessing techniques for doing this task. Most usable techniques are one-hot coding and level encoding. In this project, level encoding was used to convert the data. Another transformation of the data is converting all the data into one scale to increase accuracy. For doing this normalization technique, a standard scaler is used. The process of converting data from a source format into a format that may be utilized for a number of purposes is known as data transformation. It takes place during the ETL (Extract, Load, Transform) procedure, when the data has to be identified, removed from its current location, and transferred into a single repository. By resolving problems like missing values and inconsistencies, this raw data from the Data Transformation Process in Data Mining must be cleaned and made ready for transformation.

Data smoothing is a method for reducing noise in a dataset. Noise is the term used to describe the distorted and nonsensical data included in a dataset. To draw attention to the unique qualities of the data, smoothing procedures are applied. After the noise has been eliminated, the technique can identify even the smallest changes in the data to find unique patterns. The act of compiling data from numerous sources and storing it in a certain, standardized manner makes it possible to access it quickly; this is known as data aggregation. Data is gathered, saved, examined, and then presented as a report or summary. It is advantageous in the Data Transformation Process in machine learning when working with discrete data since discretization uses decision tree-based algorithms to create concise, accurate results. The concept of the Data Transformation Process in Data Mining hierarchies is used in this procedure to transform low-level characteristics of data into high-level data properties. It is useful to switch from a lower to a higher conceptual level in order to better understand the facts. If a dataset contains information of age data, for instance, it may be expressed as (8, 40). At a higher conceptual level, it is converted into a category value (young, old). Here, the data is altered to fit inside a predetermined range. When characteristics are distributed over several ranges or sizes, data modeling and mining may be challenging. Normalization makes it easier to use data mining techniques and to extract data more quickly, both of which are crucial steps in the data transformation process used in data mining.

	Price_Crore	Location	Bed	Bath	Size_Sqft
0	1.90	Uttara	3.0	3.0	2,024
1	1.90	Uttara	3.0	3.0	2,004
2	3.90	Uttara	4.0	4.0	3,334
3	0.70	Uttara	3.0	3.0	1,289
4	2.55	Bashundhara	3.0	3.0	2,200
5	0.95	Uttara	3.0	3.0	1,450
6	0.62	Dhanmondi	3.0	2.0	930
7	1.75	Mohammadpur	3.0	3.0	1,758
8	2.50	Bashundhara	3.0	3.0	2,075
9	6.10	Gulshan	4.0	4.0	3,120

	Price_Crore	Location	Bed	Bath	Size_Sqft
0	1.90	9	3.0	3.0	6.794187
1	1.90	9	3.0	3.0	6.786311
2	3.90	9	4.0	4.0	7.186941
3	0.70	9	3.0	3.0	6.433938
4	2.55	1	3.0	3.0	6.860219
5	0.95	9	3.0	3.0	6.528386
6	0.62	2	3.0	2.0	6.170202
7	1.75	7	3.0	3.0	6.682223
8	2.50	1	3.0	3.0	6.813912
9	6.10	3	4.0	4.0	7.135080

Figure 3.3.7.1: After Transforming Categorical to Numerical

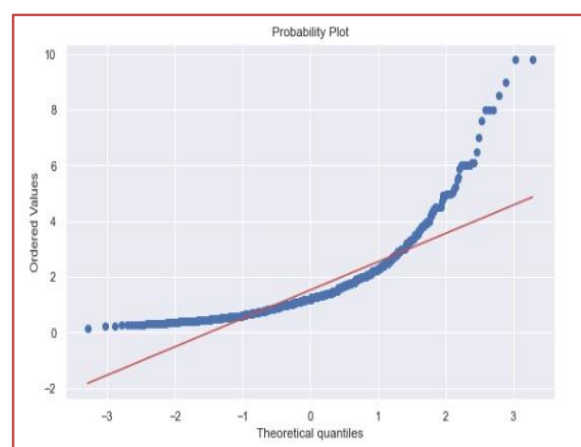
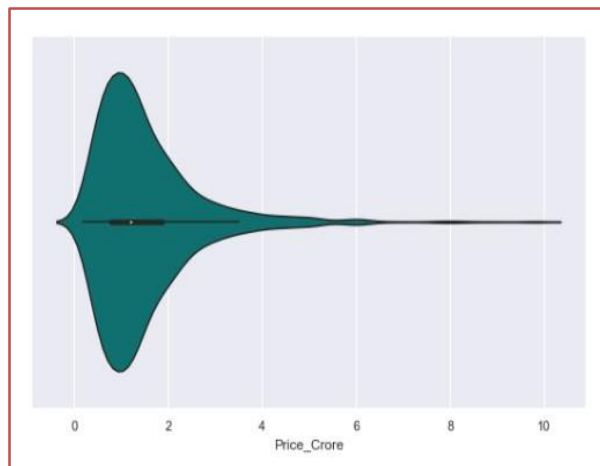
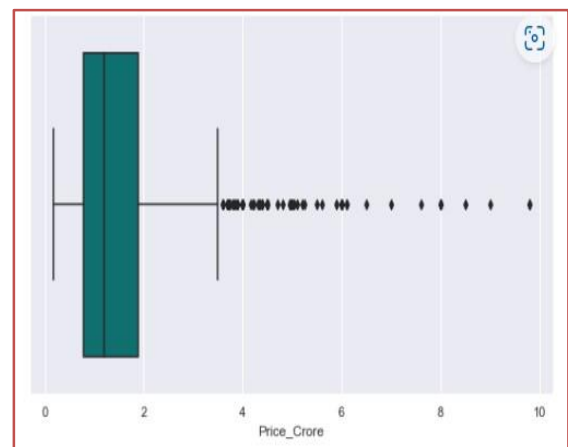
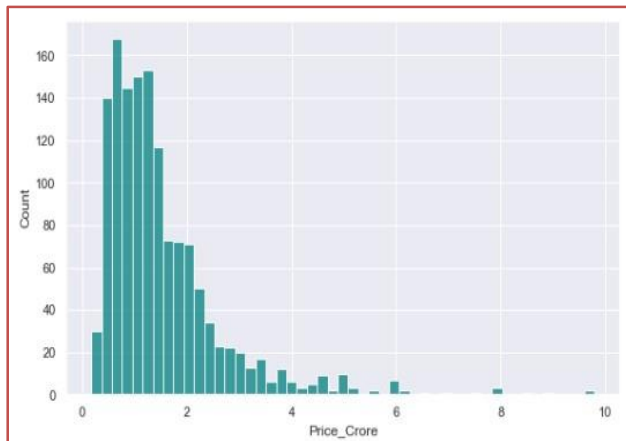


Figure 3.3.7.2: Before Transformation of Price

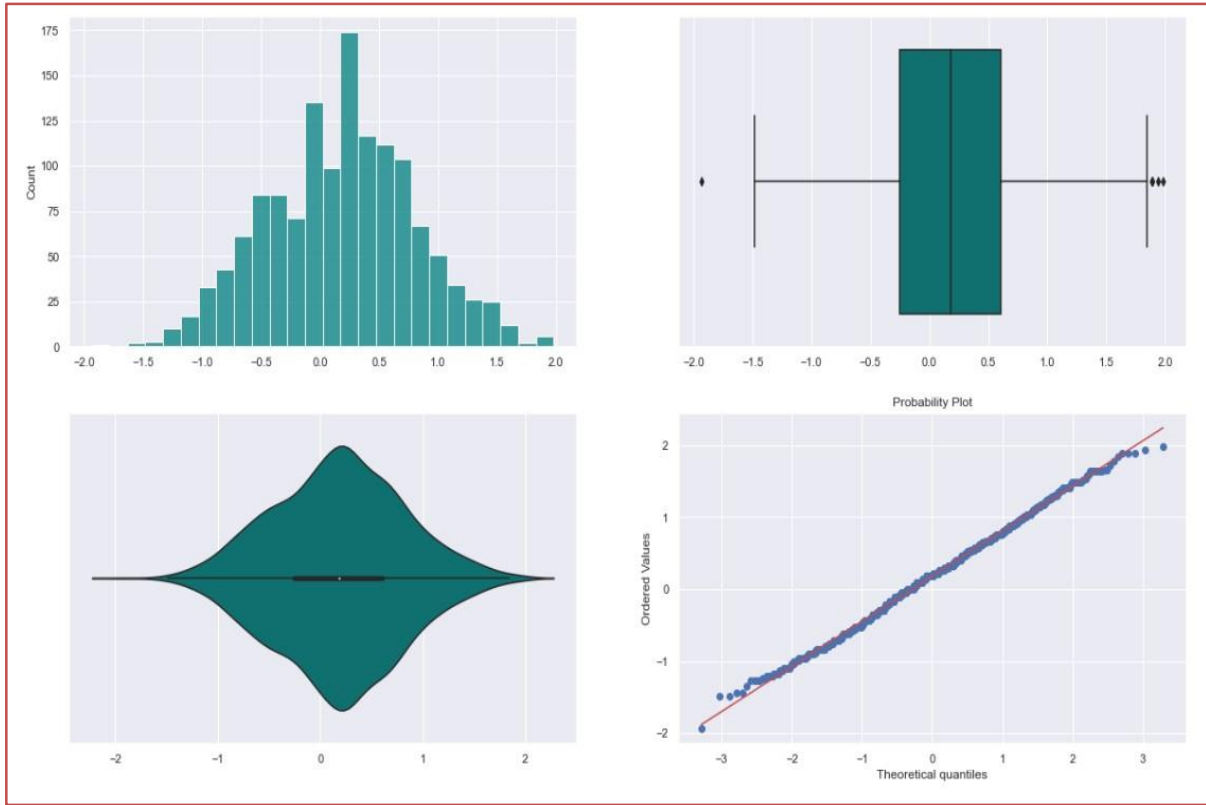


Figure 3.3.7.3: After Transformation of Price

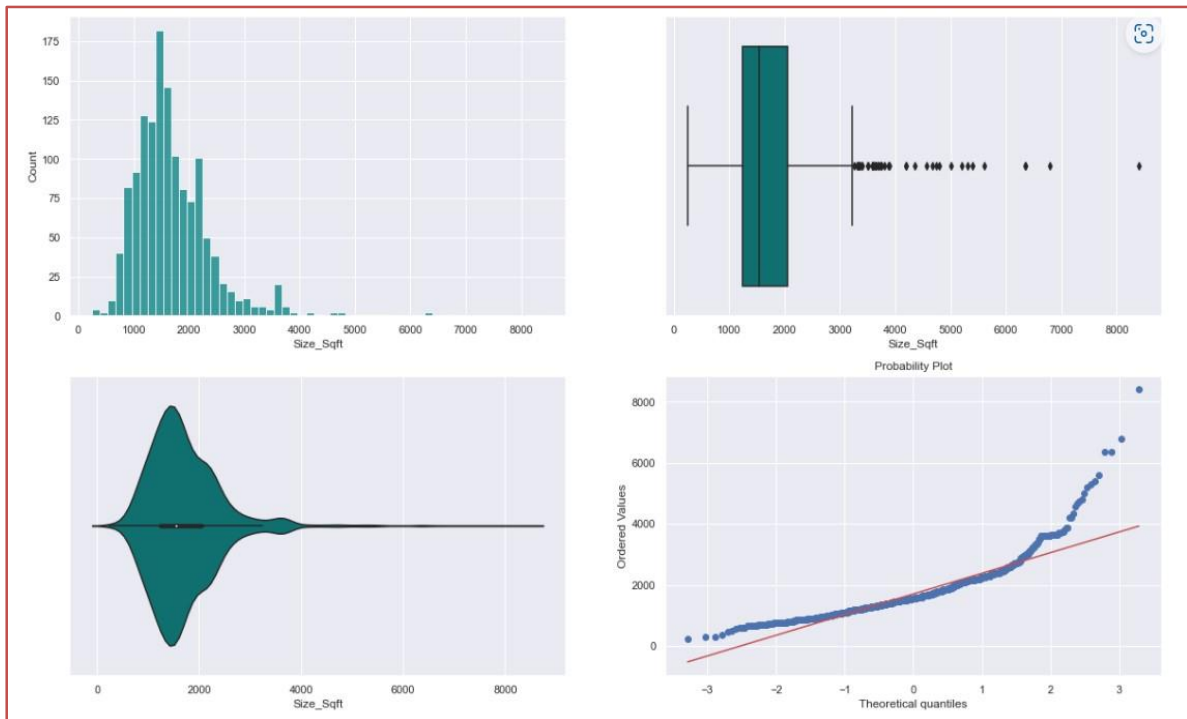


Figure 3.3.7.4: Before Transformation of Size



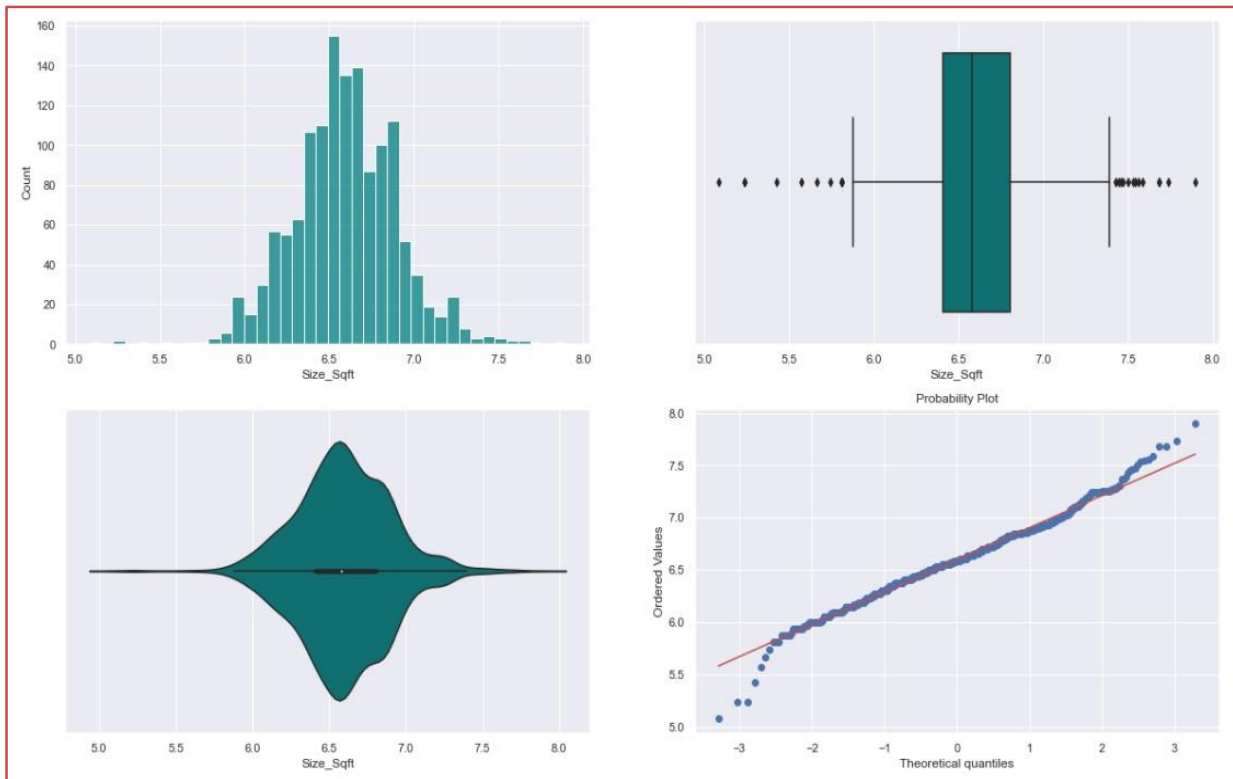


Figure 3.3.7.5: After Transformation of Size

### 3.4 Feature Engineering

The process of adding attribute-value pairs to a dataset that is stored in a table is known as feature engineering. Features or descriptive properties are other names for attribute-value pairs. When pre-processing data for supervised learning algorithms in machine learning, feature engineering is crucial. Data must be kept in a single table with rows containing training samples and columns listing attribute-value pairs for supervised learning algorithms. The improvement of supervised learning output accuracy is a key objective of feature engineering. An input that a machine learning model utilizes to generate precise predictions is referred to as a feature. A machine learning recommendation engine may utilize the characteristics "subject," "word count," "reading level," and "time-to-read" if a website sells books to determine what material a visitor would be interested in reading next.

One of the most crucial aspects of machine learning is feature engineering, however due to the amount of human involvement needed, the process is sometimes referred to as an art.

The data scientist or machine learning engineer must possess extensive subject expertise. This implies that they need to have a thorough grasp of the business issues that each model is intended to solve, as well as the technical know-how necessary to prepare the data for training. Good soft skills are also necessary for feature engineering, which calls for ML engineers and data scientists. When deciding which variables to utilize, they often need to consult other domain specialists. The time-consuming procedure of pre-processing the data may make the difference between a machine learning model that produces correct predictions and one that does not.

For doing this project manual feature analysis had done. When data had been collected, many unnecessary features were in the dataset for instance image data, builder name, details location, status and the like. After analysis, the data I had decided that, I should remove thesis features. Besides, there had other features of the flats such as elevator facility, electricity facility, parking facilities, and the like. Moreover, I hadn't taken these features for analysis, because I found that, almost all the flat have parking and electricity facilities. So I didn't take these features. But after Excel sheet analysis, I found that flat having price more than 50-60 lacks have elevator facility. As I am doing analysis upon price based on locations. So I removed this feature also to generate more accurate model. Finally, after analysis, 5 feature were finalized.

Table 1: Selected Features

Location	Location of flats.
Size	Flats size in square feet.
Bed	Number of bed.
Bath	Number of bath
Price	Price of flats.

### 3.5 Python Libraries and Terms

A number of Python libraries were utilized to conduct this investigation. A collection of interconnected modules is referred to as a "Python library." It contains groups of code that may be used repeatedly in several projects. It streamlines and makes Python programming more usable for programmers. Python libraries play a significant role in the fields of data science, machine learning, data visualization, and other related fields. To prevent having to modify existing code in our product, we use libraries. Nevertheless, how it functions. Actually, the library files in the MS Windows environment have a DLL extension (Dynamic Load Libraries). When we link a library with our program and run it, the linker will automatically search for that library. It extracts those library's functions, then suitably interprets the program using them. That is how our software uses a library's methods. We'll examine how we include libraries into our Python scripts in more detail. The Python Standard Library includes Python's exact syntax, semantics, and tokens. It has built-in modules that provide users access to key system functionalities such as I/O and a few more core modules. The C programming language is used to develop the majority of Python libraries. With over 200 core modules, the Python standard library is rather large. Python is a high-level programming language thanks to all of these factors working together. It is crucial to use the Python Standard Library. It is necessary in order for programmers to access Python's features. In addition to this, though, Python also has a number of additional libraries that facilitate programming.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
sns.set()

from scipy.stats import probplot, boxcox
from scipy.special import inv_boxcox
import pylab

from sklearn.preprocessing import StandardScaler
from sklearn import preprocessing
```

Figure 3.5: Used Libraries

### 3.5.1 Numpy

The abbreviation "Numerical Python" is Numpy. It is the one that is most frequently used. This well-liked machine learning program supports large matrices and multi-dimensional data. For rapid computations, it features built-in mathematical functions. The Array method represents one of this library's key features.



Figure 3.5.1: Numpy

### 3.5.2 Pandas

The Pandas library is vital to data scientists. A collection of analytical tools and custom high-level data structures are available in this free machine learning toolkit. Data administration, cleansing, and analysis are all made simpler by it. Pandas supports a variety of operations, including sorting, re-indexing, repetition, composition, data transformation, infographics, aggregation, and others.

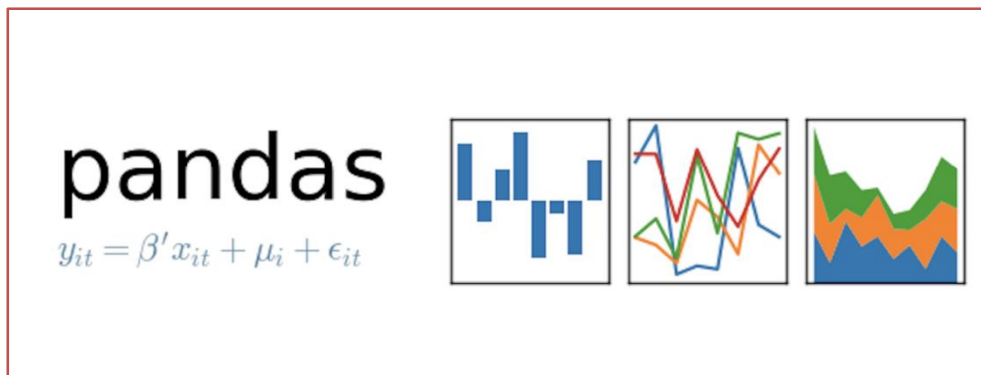


Figure 3.5.2: Pandas

### 3.5.3 Matplotlib

This library manages plotting numerical data. It is employed in data analysis as a result. Additionally, it generates incredibly complex visuals like graphs, scatter diagrams, graphs, and pie charts using an open-source framework.



Figure 3.5.3: Matplotlib

### 3.5.4 SciPy

SciPy is short for "Scientific Python." For advanced scientific calculations, it is an open-source library. This library is based on a Numpy extension. To conduct complicated calculations, it collaborates with Numpy. Numpy supports indexing and sorting of array data, whereas SciPy stores the numerical data code. Engineers and application developers also often utilize it.



Figure 3.5.4: SciPy

### 3.5.5 Scikit-Learn

This well-known Python tool is used to manage challenging data. Scikit-learn is indeed an open-source package that makes machine learning easier. It supports a range of supervised and unsupervised techniques, such as linear regression, classification, and clustering. This package works in tandem with both SciPy and Numpy.



Figure 3.5.5: Scikit-learn

### 3.5.6 TensorFlow

This library was created in collaboration between Google and the Brain Team. It is an open-source library for high-level computing. It also features in machine learning and deep learning methods. It has a considerable number of tensor operations. In order to address difficult physics and mathematical problems, researchers also utilize this Python package.



Figure 3.5.6: TensorFlow

### 3.5.7 Seaborn

For plotting statistical visualizations, Python's Seaborn visualization library is amazing. It provides wonderful default color combinations and styles to improve the aesthetic appeal of statistics charts. The Pandas data structures are closely tied to it and it is built on top of the Matplotlib toolbox. With Seaborn, analysis of the data and comprehension will be centered on visualization. It provides dataset-oriented APIs that let us transition between various visual representations of the exact variables for a deeper understanding of the dataset.



Figure 3.5.7: Seaborn

### 3.5.8 Root Mean Square

The arithmetic mean of the squares of a collection of values, sometimes referred to as the mean square in statistics, is defined as the root-mean-square (RMS). RMS is a deviation from the extended means that has an exponent of 2. It is sometimes referred to as a quadratic mean. A changing function built on an integral of the squares of the values existing at any given period in a cycle is one more explanation of root-mean-square.

The arithmetic mean squared or the squares of the function that defines the continuous waveform, to put it another way, is the RMS of a set of integers.

Used the Root Mean Square formula below before determine the RMS value of a set of data values.

The RMS is provided by: for a set of n values including  $x_1, x_2, x_3, \dots, X_n$

xrms =

$$\sqrt{\frac{(x_1^2 + x_2^2 + \dots + x_n^2)}{N}} \dots \dots \dots (1)$$

A continuous function f(t), defined for the range T1 ≤ t ≤ T2, has the following formula:

frms =

$$\sqrt{\frac{1}{T_2 - T_1} \int_{T_1}^{T_2} [f(t)^2 dt]} \dots \dots \dots (2)$$

The RMS of a function is always the same as the RMS of a periodic function. The RMS value of the continuous function can be approximated by taking the RMS of a string of evenly spaced items. Furthermore, the RMS value of different waveforms may be determined without the need of mathematics.

### 3.5.9 Root Mean Square Error

The Root Mean Square Error, or RMSE, is a popular measure of how effectively an estimator or mode predicts disparities between statistics (population values and samples). The RMSE describes the sample standard deviation of the variances between predicted and actual values. Each of these disparities is known as a residual when calculated across the data sample that was employed to estimate; when calculated outside the sample, they are known as prediction errors. The RMSE aggregates the magnitude of forecasted errors made multiple times into a single metric of predictive power.

To calculate RMSE in connection to an estimated variable in a projected model, xmodel, the square root of the mean squared error is utilized.

RMSE =

$$\sqrt{\frac{\sum_{i=1}^n (X_{obs,i} - X_{model,i})^2}{n}} \dots \dots \dots (3)$$

At time I Xobs stands for "observed values" while Xmodel stands for "modeled values."



## CHAPTER 4

### EVALUATION AND EXPERIMENTAL RESULTS

#### 4.1 Introduction

This analysis is evaluated by 10 advance regression techniques. In machine learning, every analysis is one kind of prediction. Prediction means estimating the most probable value. After analyzing past data, necessary patterns can drown to generate future outcomes. Supervised algorithm had used in this model because dataset had clear level and Prediction algorithms were used just because this dataset is a combination of numerical and categorical data. All of this analysis are statistical analysis. Because without statistical analysis, this can't be done. All the build in algorithm are made by mathematical terms, and they all have specific requirements. When these requirements are done, these algorithms utilize those requirements and evaluate the result and give an estimation upon requirements.

#### 4.2 Algorithms

There are 10 algorithms used in this analysis and They are:

- Lasso
- Bayesian Ridge
- Ridge
- XGBoost
- KNN
- ANN
- Random Forest
- Support Vector
- Lightgbm
- CatBoost
- Gradient Boost

### 4.2.1 Lasso

The acronym "LASSO" stands for Least Absolute Shrinkage and Selection Operator. It is a statistical formula for regularizing data models and choosing features. It is chosen over regression approaches for a more precise forecast. In this model, shrinkage is utilized. Shrinkage is the term for when data values decrease as they approach the mean. The lasso method promotes simple, sparse models (i.e. models with fewer parameters). This particular type of regression is perfectly suited when a model shows a high level of collinearity or when you want to automate some processes in the model selection process, such as variable selection and parameter removal.

Software-based Lasso solutions are available for quadratic programming problems. The algorithm's goal is to minimize:

$$\sum_{i=1}^n (y_i - \sum_j x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

Which is the same as minimizing the sum of squares with constraint  $\sum |\beta_j| \leq s$  ( $\Sigma$  = summation notation). A regression model that is simpler to understand is produced when some of the  $s$  are minimized to absolutely zero.

A tuning parameter,  $\lambda$  controls the strength of the L1 penalty.  $\lambda$  is basically the amount of shrinkage:

When  $\lambda = 0$ , no parameters are eliminated. The estimate is equal to the one found with linear regression.

As  $\lambda$  increases, more and more coefficients are set to zero and eliminated (theoretically, when  $\lambda = \infty$ , all coefficients are eliminated).

As  $\lambda$  increases, bias increases.

As  $\lambda$  decreases, variance increases.

If an intercept is included in the model, it is usually left unchanged.

### 4.2.2 Ridge

Ridge regression is a model-tuning approach used while evaluating multicollinear data. This method is used to achieve L2 regularization. When there is an issue with multicollinearity, least-squares are unbiased, and variances are large, resulting in predicted values that are far from the actual values.

Ridge regression imposes a certain type of constraint on the parameters ( $\beta$ 's):  $\hat{\beta}_{ridge}$ . The ridge is selected to reduce the penalized sum of squares:

$$\sum_{i=1}^n (y_i - \sum_{j=1}^p x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2 \dots \dots \dots (1)$$

$\hat{\beta}_{ls}$  is an unbiased estimator of  $\beta$ ;  $\hat{\beta}_{ridge}$  is a biased estimator of  $\beta$ .

The definition of the effective degrees of freedom connected to  $\beta_1, \beta_2, \dots, \beta_p$  is:

$$df(\lambda) = tr(X(X'X + \lambda I)^{-1} X') = \sum_{j=1}^p \frac{d_j^2}{d_j^2 + \lambda} \dots \dots \dots (2)$$

Calculating the variance-covariance matrix is simple due to the linear nature of the ridge estimator.

$$var(\hat{\beta}_{ridge}) = \sigma^2 (X'X + \lambda I)^{-1} X'X (X'X + \lambda I)^{-1} \dots \dots \dots (3)$$

### 4.2.3 XGBoost

The XGBoost approach is useful for generating supervised regression models. Knowing about its (XGBoost) objective function and base learners lets one to determine the veracity of this proposition. The aim function includes a gradient descent and a regularization term. It describes the difference between real and expected values, i.e. how near or far the model outputs are to the actual values. Reg:logistics and reg:linear are the most often used loss functions in XGBoost for binary classification and regression tasks, respectively. One method used in ensemble learning is called XGBoost. In ensemble learning, many models—often referred to as base learners—are trained and combined to provide a single prediction. XGBoost anticipates having base learners who are consistently terrible at the rest, such that bad forecasts will cancel out when all of the forecasts are summed together and better forecasts will add up to create final positive forecasts.

Regression problems result in continuous or real results. Two popular regression methods are decision trees and linear regression. The root-mean-square error (RMSE) and mean-square error are two examples of the metrics used in regression (MSE). Each of the key participants in our XGBoost models has a vital task to complete.

RMSE is the square root of the mean squared error (MSE).

MAE: Since it is not mathematically sound, it is used less frequently than other metrics even though it represents the absolute sum of actual and predicted differences.

$$RMSLE(y, y^{\wedge}) = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log(1 + \hat{y}_i) - \log(1 + y_i))^2} \dots\dots\dots(1)$$

where  $\hat{y}_i$  is the predicted value of the target for instance  $i$ , and  $y_i$  is the actual value of the target for instance  $i$ .

The asymmetry arises because

$$\log(1 + \hat{y}_i) - \log(1 + y_i) = \log\left(\frac{1 + \hat{y}_i}{1 + y_i}\right) \dots\dots\dots(2)$$

#### 4.2.4 KNN

Classification and regression problems may be solved with the k-nearest neighbors (KNN) algorithm, a simple and approachable supervised machine learning method. The KNN method assumes that related items could be identified close by. In other words, linked things are close to one another. KNN uses a mathematical notion that many of us learned as children: calculating the distance between points on a graph—to represent the idea of similarity, also known as distance, proximity, or closeness. There are several ways to calculate distance, and depending on the situation, one approach may be preferable to another. The straight-line distance, or Euclidean distance, is a popular and well-known alternative, nevertheless.

Euclidean:

$$\sqrt{\sum_{i=1}^k (x_i - y_i)^2} \dots\dots\dots(1)$$

Manhattan:

$$\sum_{i=1}^k |x_i - y_i| \dots\dots\dots(2)$$

### 4.2.5 Random Forest

Random Forest is an ensemble method that can handle both regression and classification problems. It does so with the use of many decision trees using a technique known as Bootstrap and Aggregation, or bagging. As opposed to relying just on one decision tree, the basic idea behind this is to combine a number of decision trees to get the ultimate result. Random Forest employs several different decision trees as its main learning models. To build sample datasets for each model, row and feature sampling is conducted at random from the dataset. Bootstrap is the name of it.

$$MSE = \frac{1}{N} \sum_{i=1}^N (f_i - y_i)^2$$

### 4.2.6 Support Vector Machine

Data analysis for classification and regression analysis utilizing supervised machine learning models and associated learning techniques is known as support vector regression. The Support Vector Machine, or SVM, concept is the basis of SVR. It is one of the popular Machine Learning models that may be used to solve classification problems or categorize data when it is impossible to utilize linear deviation. Although we frequently use the term "regression issues," the most accurate term is "categorization." The SVM approach aims to find a hyperplane in an N-dimensional space that unambiguously classifies the data points. The size of the hyperplane depends on the quantity of features. Essentially, if there are just two input features, the hyperplane is a line. If there are three input characteristics, the hyperplane transforms into a 2-D plane. Anything with more than three traits is difficult to imagine.

Linear line:

$$y^{\wedge} = WT x + b \dots\dots\dots(1)$$

Loss after training:

$$RMSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \dots\dots\dots(2)$$

### 4.2.7 Light Gradient Boosted Machine

A decision tree-based gradient boosting framework called LightGBM helps models run more efficiently while using less memory.

It uses two cutting-edge techniques: Gradient-based One Side Sampling and Exclusive Feature Bundling (EFB), which solve its shortcomings, replace the histogram-based core approach used in all GBDT (Gradient Boosting Decision Tree) systems. The two GOSS and EFB methods, which are described here, are the characteristics of the LightGBM Algorithm. They work together to make the model operate and offer it an edge over other GBDT frameworks. For gradient-based one-side sampling, use the LightGBM approach.

In the calculation of information gain, various data instances play a variety of functions. The information gain will be greater for the examples with bigger gradients (i.e., under-trained cases). To maintain the accuracy of information gain estimate, GOSS only randomly discards cases with minor gradients (those that are below a certain threshold or those are in the top percentiles). With the same goal sampling rate, this method may provide a gain estimate that is more accurate than uniformly random sampling, particularly when the information gain value has a wide range.

$$IG(B, V) = En(B) - \sum_{v \in \text{Values}(V)} |B_v| BEn(B_v) \dots \dots \dots (1)$$

$$En(B) = \sum d = 1D - pd \log_2 pd \dots \dots \dots (2)$$

### 4.2.8 Gradient Boost and CatBoost

An ensemble machine learning approach called gradient boosting is often employed to address classification and regression issues. It is simple to use, handles heterogeneous data effectively, and even handles relatively tiny data. In essence, it makes a strong learner out of a group of several poor ones.

Yandex created the open-source boosting package known as CatBoost or Categorical Boosting. In addition to regression and classification, CatBoost may also be used for ranking, recommendation mechanisms, forecasting, and even personal assistants.

This gradient boosting method is improved by CatBoost.

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (a_i - t_i)^2 w_i}{\sum_{i=1}^N w_i}} \dots\dots\dots(1)$$

$$p_i = \frac{1}{1 + e^{-a_i}} \dots\dots\dots(2)$$

### 4.3 Model Performance and Result Analysis

```
models = {
    'ridge' : Ridge(),
    'xgboost' : XGBRegressor(),
    'catboost' : CatBoostRegressor(verbose=0),
    'lightgbm' : LGBMRegressor(),
    'gradient boosting' : GradientBoostingRegressor(),
    'lasso' : Lasso(),
    'random forest' : RandomForestRegressor(),
    'bayesian ridge' : BayesianRidge(),
    'support vector' : SVR(),
    'knn' : KNeighborsRegressor(n_neighbors = 4)
}
```

Figure 4.3.1: Trained Models

```
Epoch 1/5
35/35 [=====] - 1s 8ms/step - loss: 0.1690 - rmse: 0.4111 - val_loss: 0.0871 - val_rmse: 0.2951
Epoch 2/5
35/35 [=====] - 0s 3ms/step - loss: 0.1155 - rmse: 0.3399 - val_loss: 0.1023 - val_rmse: 0.3198
Epoch 3/5
35/35 [=====] - 0s 3ms/step - loss: 0.1079 - rmse: 0.3285 - val_loss: 0.0805 - val_rmse: 0.2837
Epoch 4/5
35/35 [=====] - 0s 3ms/step - loss: 0.1091 - rmse: 0.3303 - val_loss: 0.0920 - val_rmse: 0.3034
Epoch 5/5
35/35 [=====] - 0s 2ms/step - loss: 0.1185 - rmse: 0.3443 - val_loss: 0.0809 - val_rmse: 0.2845
```

Figure 4.3.2: RMSE Analysis

After folding 10 times, results are showing below:

Table 2: Model Results

Name	RMSE
Lasso	0.618
Bayesian Ridge	0.3182
Ridge	0.3184
XGBoost	0.289
KNN	0.285
ANN	0.284
Random Forest	0.279
Support Vector	0.276
Lightgbm	0.271
Catboost	0.268
Gradient Boost	0.263

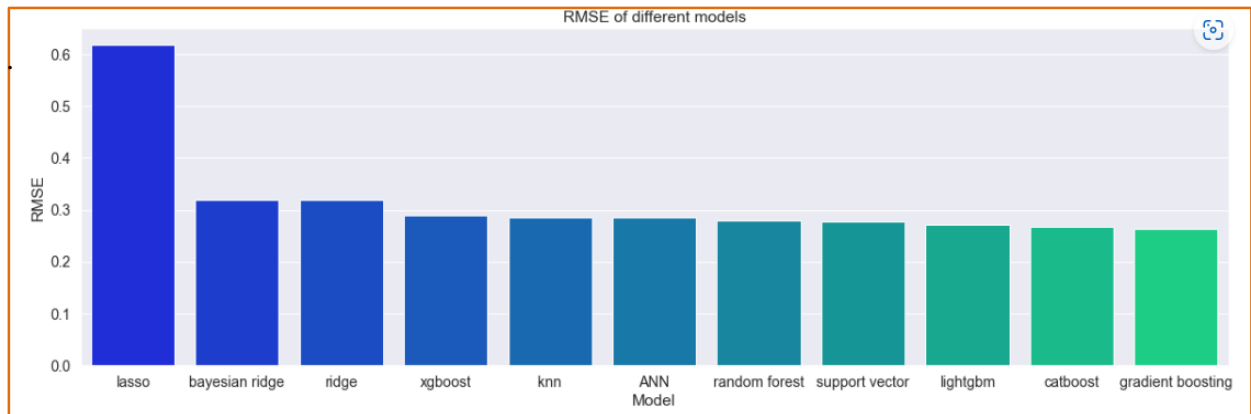


Figure 4.3.3: RMSE of Different Models

Table 3: Final Prediction

RMSE	0.391
R-Square	0.87

$$R - square = 1 - \frac{\sum(Y_{actual} - Y_{predicted})^2}{\sum(Y_{actual} - Y_{mean})^2}$$





Figure 4.3.4: Actual Result Vs Predicted Result

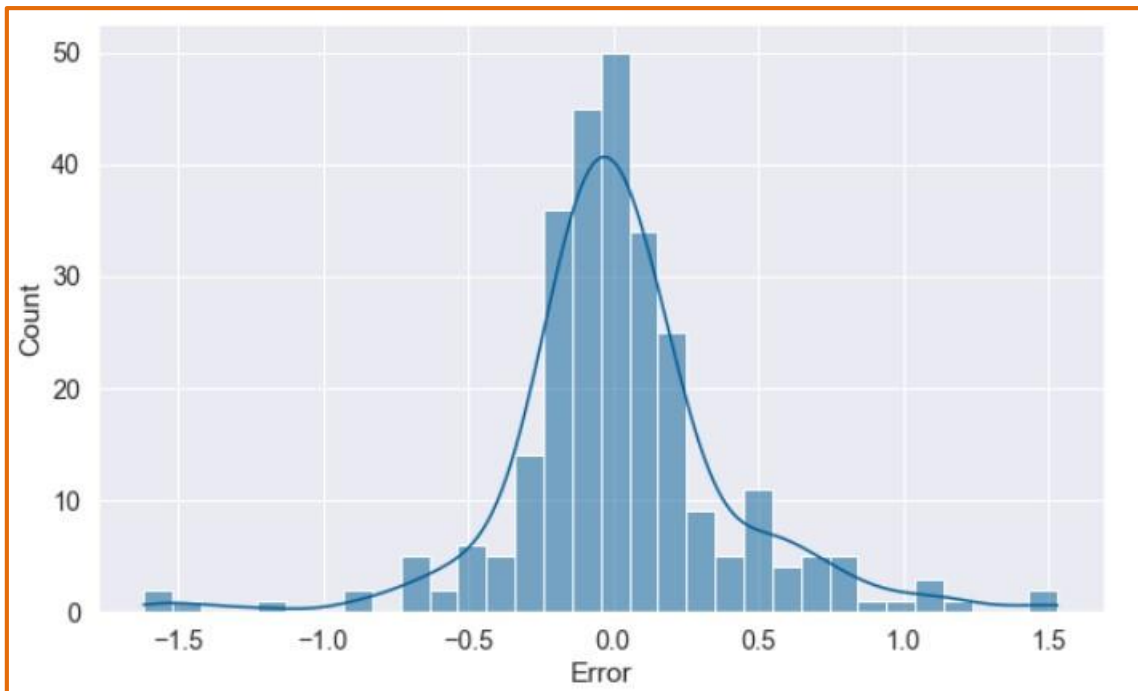


Figure 4.3.5: Distribution

## CHAPTER 5

### CONCLUSION AND FUTURE WORK

#### 5.1 Conclusion

In this work, we looked at 10 machine learning algorithms that were used to create housing price prediction models for forecasting the cost of apartments in the Bangladeshi metropolis, Dhaka. The first scientific housing dataset and pure analysis for the Dhaka housing market are things I contributed to this study by gathering flat data. Advanced Regression methods, Support Vector Regression, Boosting, and Deep Learning Network are a few examples of machine learning algorithms that can predict values that are extremely near to the listed price. R-Square 0.87 of the findings demonstrate that Lasso outperforms the other machine learning techniques. In this work, we analyze the effectiveness of multiple machine learning regression models to discover the optimal model for a more accurate forecast of home prices. To the best of our knowledge, no machine learning or other home forecasting methods are currently in use. I am confident that machine learning flat price prediction models will aid people involved in the real estate industry and prospective purchasers in making wise decisions on the purchase of a home. Future research on the real estate market, stock price forecasting, and forecasting oil and petroleum prices may all be built on this work. Future house price predictions might be improved upon by combining this textual tabular dataset with visual elements of the homes, such as pictures of the inside and outside. And finally, other macroeconomic factors that affect the housing market, such as the price of gold, the stock price index, property taxes, and the appraised value of a property, can have an impact on home prices. Taking these factors into account can help develop house price prediction models that can accurately predict home prices.

#### 5.2 Future Work

In sum, this analysis had done with the flat information of Dhaka city. In future, I want to develop a model with the data of every district of Bangladesh. I have done this analysis with 5 features that is another point I will development in future. By analysis with the image data more reliable prediction can be done in future. As, real estate market is one of the costliest market of a country

which rapidly create an influence against GDP. So, if we development this field more widely it will create new era of a country. So, my heart and soul try will be increased this market widely collect more data and build fast and secure model to estimate prices.

## REFERENCES

- [1] J. J. Jui, M. M. Imran Molla, B. S. Bari, M. Rashid, and M. J. Hasan, "Flat price prediction using linear and random forest regression based on machine learning techniques," *Embracing Industry 4.0*, pp. 205–217, 2020.
- [2] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, "House price prediction using Random Forest Machine Learning Technique," *Procedia Computer Science*, vol. 199, pp. 806–813, 2022.
- [3] Jangaraj, Avanija & Sunitha, Gurram & Madhavi, Reddy & Kora, Padmavathi & Hitesh, R & Associate, Sai. (2021). Prediction of House Price Using XGBoost Regression Algorithm. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*. 12. 2151-2155.
- [4] G. Liu, "Research on prediction and analysis of real estate market based on the multiple linear regression model," *Scientific Programming*, vol. 2022, pp. 1–8, 2022.
- [5] Q. Zhang, "Housing price prediction based on multiple linear regression," *Scientific Programming*, vol. 2021, pp. 1–9, 2021.
- [6] N. Chen, "House price prediction model of Zhaoqing City based on correlation analysis and multiple linear regression analysis," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–18, 2022.
- [7] J. M. Pagés and L. Á. Maza, "ANALYSIS OF HOUSE PRICES IN SPAIN," *BANCO DE ESPAÑA*, vol. 1579-8666, 2003.
- [8] Thakur, Amey & Satish, Mega. Bangalore House Price Prediction," *International Research Journal of Engineering and Technology (IRJET)*", 8. 193-196, 2021.
- [9] R. E. Febrita, W. F. Mahmudy, and A. P. Wibawa, "High Dimensional Data Clustering using Self-Organized Map," *Knowledge Engineering and Data Science (KEDS)*, vol. 2, no. 1, pp. 31-40, 2019.
- [10] K. Case and R. Shiller, "Is There a Bubble in the Housing Market?," *Brookings Papers on Economic Activity*, vol. 2003, no. 2, pp. 299-362, 2003.
- [11] N. Chen, "House price prediction model of Zhaoqing City based on correlation analysis and multiple linear regression analysis," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–18, 2022.
- [12] C. Zhou, "House price prediction using polynomial regression with particle swarm optimization," *Journal of Physics: Conference Series*, vol. 1802, no. 3, p. 032034, 2021.
- [13] C. Vasquez and V. Chellamuthu, "House price prediction with statistical analysis in support Vector Machine Learning for regression estimation," *Curiosity: Interdisciplinary Journal of Research and Innovation*, vol. 2, 2021.
- [14] M. B. Abdul Hamid, M. H. B. Mohd Izhar, M. W. B. Ismail, T. K. Wing, and T. T. Joon, "Advanced Regression Techniques," 2020.
- [15] Agarwal, Umang & Gupta, Smriti & Goyal, Madhav. HOUSE PRICE PREDICTION USING LINEAR REGRESSION IN ML. 10.13140/RG.2.2.11175.62887, 2022.
- [16] Konwar, Robart & Kakati, Angshuman & Das, Bhagyashree & Shah, Borah & Muchahari, Monoj. House Price Prediction Using Machine Learning. *The Journal of Philosophy Psychology and Scientific Methods*. 9. 2455-6211, 2021.

- [17] J. chougale, A. Shinde, N. Deshmukh, D. Sawant, and V. Latke, "House price prediction using machine learning and image processing," *Journal of University of Shanghai for Science and Technology*, vol. 23, no. 06, pp. 961–965, 2021.
- [18] Q. Truong, M. Nguyen, H. Dang, and B. Mei, "Housing price prediction via Improved Machine Learning Techniques," *Procedia Computer Science*, vol. 174, pp. 433–442, 2020.
- [19] M. I. G. U. E. L. A. N. G. E. L. MANRIQUE, O. B. Sierra, D. Otero Gomez, H. Laniado, R. Mateus C, and D. A. Millan, "Housing-price prediction in Colombia using machine learning," 2021.
- [20] Özdemir, Ozancan. (2022). House Price Prediction Using Machine Learning: A Case in Iowa. 10.13140/RG.2.2.19846.86086.
- [21] R.-T. Mora-Garcia, M.-F. Cespedes-Lopez, and V. R. Perez-Sanchez, "Housing price prediction using machine learning algorithms in COVID-19 times," *Land*, vol. 11, no. 11, p. 2100, 2022.
- [22] Pai, P.-F. and Wang, W.-C. (2020). Using machine learning models and actual transaction data for predicting real estate prices. *Applied Sciences*, 10(17):5832.
- [23] S. J. Xin and K. Khalid, "Modelling House Price Using Ridge Regression and Lasso Regression," *International Journal of Engineering & Technology*, vol. 7, pp. 498–501, 2018.
- [24] S. Mysore, "Prediction of house prices using machine learning," *International Journal for Research in Applied Science and Engineering Technology*, vol. 10, no. 6, pp. 1780–1785, 2022.
- [25] P. Mali, S. Patil, P. Gujar, and P. M. Tiwari, "PREDICTION OF HOUSE SALES PRICES USING MACHINE LEARNING ALGORITHM," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 3, no. 3, 2021.
- [26] House prices - advanced regression techniques. Kaggle. (n.d.). <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>. Retrieved 3 Dec 2021.
- [27] Y. Yang, "Research on house price prediction based on multi-dimensional data fusion," *International Journal of Advanced Network, Monitoring and Controls*, vol. 5, no. 1, pp. 1–8, 2020.
- [28] G. Liu, "Research on prediction and analysis of real estate market based on the multiple linear regression model," *Scientific Programming*, vol. 2022, pp. 1–8, 2022.
- [29] M. Imran, U. Zaman, M. Waqar, and A. Zaman, "Using Machine Learning Algorithms for Housing Price Prediction: The Case of Islamabad Housing Data," *Soft Computing and Machine Intelligence*, vol. 1, no. 1, 2021.
- [30] Bajari P, Benkard CL & Krainer J, "House prices and consumer welfare", *Journal of Urban Economics*, 58(3), 2010, pp.474–487.
- [31] Amri S & Tularam GA, "Performance of Multiple Linear Regression and Nonlinear Neural Networks and Fuzzy Logic Techniques in Modelling House Prices", *Journal of Mathematics and Statistics*, 8(4), 2012, pp.419–434.
- [32] Mak S, Choy L & Ho W, "Quantile Regression Estimates of Hong Kong Real Estate Prices", *Urban Studies*, 47(11), 2010, pp.2461– 2472.
- [33] Limsombunchai V, Gan C, & Lee M, "House Price Prediction : Hedonic Price Model vs Artificial Neural Network", *American Journal of Applied Sciences*, 1(3), 2004, pp.193–201.

- [34] Graham MH, "Confronting multicollinearity in ecological multiple regression", *Ecology*, 84(11), 2003, pp.2809-2815.
- [35] Kraha A, Turner H, Nimon K, Zientek & Henson RK, "Interpreting multiple regression in the face of multicollinearity", *Frontiers in Psychology*, 3, 2012, pp.1–10.
- [36] Bin Shafi MA, Bin Rusiman MS and Che Yusof NSH, "Determinants Status of Patient After Receiving Treatment at Intensive Care Unit: A Case Study in Johor Bahru", *I4CT 2014 - 1st International Conference on Computer, Communications, and Control Technology*, 6914150, 2014, pp.80 – 82.
- [37] Pasha GR & Shah MA, "Application of Ridge regression to multi-collinear data", *Journal of Research Science*, 15(1), 2004, pp.97–106.
- [38] Meinshausen N & Bühlmann P, "High-dimensional graphs and variable selection with the Lasso", *The annals of statistics*, 34(3), 2006, pp.1436–1462.
- [39] Calhoun CA, "Property Valuation Methods and Data in the United States", *Housing Finance International*, 16(2), 2001, pp.12–23.
- [40] Khalid K, Mohamed I and Abdullah NA, "An Additive Outlier Detection Procedure in Random Coefficient Autoregressive Models", *AIP Conference Proceedings*, 1682, 2015, 050017.
- [41] Mohamed I, Khalid K And Yahya MS, "Combined Estimating Function for Random Coefficient Models with Correlated Errors", *Communications In Statistics—Theory And Methods*, 45(4), 2016, pp.967-975.
- [42] Rusiman MS, Hau OC, Abdullah AW, Sufahani SF, Azmi NA, "An Analysis of Time Series for the Prediction of Barramundi (Ikan Siakap) Price in Malaysia", *Far East Journal of Mathematical Sciences*, 102(9), 2017, pp.2081-2093.
- [43] Hastie T, Tibshirani R & Friedman J, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 1st ed. New York: Springer, 2001.
- [44] Wakefield J, *Bayesian and Frequentist Regression Methods*, 1st ed. New York: Springer Science and Business Media, 2013.
- [45] Chai T & Draxler RR, "Root mean square error (RMSE) or mean absolute error (MAE)—Arguments against avoiding RMSE in the literature", *Geoscientific Model Development*, 7(3), 2014, pp.1247–1250.
- [46] Rusiman MS, Nasibov E and Adnan R, "The Optimal Fuzzy C-regression Models (OFCRM) in Miles per Gallon of Cars Prediction", *Proceedings – 2011 IEEE Student Conference on Research and Development, SCOReD 2011*, 6148760, 2011, pp.333-338.

[47] Shafi MA and Rusiman MS, "The Use of Fuzzy Linear Regression Models for Tumor Size in Colorectal Cancer in Hospital of Malay-sia", *Applied Mathematical Sciences* 9 (56), 2015, pp.2749-2759.

[48] Kutner MH, Nachtsheim CJ & Neter J, *Applied Linear Regression Models* 4th ed., New York: McGraw-Hill Higher Education, 2003.

## Prediction and Analysis of Flat Price in Dhaka

---

### ORIGINALITY REPORT

---

15%

SIMILARITY INDEX

7%

INTERNET SOURCES

2%

PUBLICATIONS

13%

STUDENT PAPERS

---

### PRIMARY SOURCES

---

1

[www.geeksforgeeks.org](http://www.geeksforgeeks.org)

Internet Source

2%

2

Submitted to Daffodil International University

Student Paper

2%

3

Submitted to Islamic University of Lebanon

Student Paper

1%

4

Submitted to Coventry University

Student Paper

1%

5

Submitted to Liverpool John Moores University

Student Paper

1%

6

Submitted to Visvesvaraya Technological University, Belagavi

Student Paper

1%

7

Submitted to Pace University

Student Paper

1%

8

Submitted to Dr. S. P. Mukherjee International Institute of Information Technology (IIIT-NR)

Student Paper

1%

---