

# **Predicting Housing and Apartment Rental Prices using Regression Analysis Algorithms**

*Arimuhanan, Karen B.  
Boteros, Angela Jane P.  
Forbes, Jake Sebastian S.  
Rodriguez, Reignoel D.*

## **Introduction**

In today's ever-evolving real estate market, accurately predicting housing and apartment rental prices is an invaluable tool for both buyers and sellers, as well as renters and landlords. By harnessing the power of advanced regression analysis algorithms, we can unlock insights into the factors that influence property prices, enabling us to make informed decisions and gain a competitive edge. The purpose of this research is to delve into the world of predictive analytics and explore how regression analysis algorithms can be employed to forecast housing and apartment rental prices. Regression analysis is a statistical technique that examines the relationships between a dependent variable (rental price) and multiple independent variables (such as location, property size, amenities, and market trends). By analyzing these variables and their impact on rental prices, we can develop robust predictive models that capture the intricacies of the real estate market.

Through this research, we aim to achieve several objectives. Firstly, we seek to identify the key determinants of housing and apartment rental prices, shedding light on the factors that play a pivotal role in shaping market dynamics. Additionally, we endeavor to evaluate the performance of different regression analysis algorithms, comparing their predictive capabilities and uncovering which algorithms yield the most accurate and reliable results. The findings of this study will have far-reaching implications for numerous stakeholders in the real estate industry. Buyers and renters will benefit from enhanced price transparency, enabling them to make more informed decisions when searching for their ideal home or apartment. Sellers and landlords, on the other hand, can optimize their pricing strategies based on empirical insights, resulting in better market positioning and potentially maximizing their returns.

## Definition of Terms

**Regression analysis.** A statistical technique that examines the relationships between a dependent variable and multiple independent variables to understand how independent variables affect the dependent variable.

**Market dynamics.** The factors and forces that influence the behavior and pricing of a market. The research aims to identify the key determinants of housing and apartment rental prices to contribute to understanding market dynamics.

**Price transparency.** Refers to the availability of accurate and reliable information regarding housing and apartment rental prices. Enhanced price transparency allows buyers and tenants to make more informed decisions during their property search.

**Pricing strategies.** The approaches and methods used by sellers and landlords to set the prices of housing and apartments. The research aims to provide empirical insights that can help optimize pricing strategies.

**Rental price forecasts.** Predictions or estimates of the prices or costs associated with renting a property.

**Risk mitigation.** The process of reducing or minimizing potential risks associated with underpricing or overpricing properties.

**Revenue losses.** Financial losses incurred due to incorrect pricing or valuation of rental properties.

**Financial risks.** Potential risks or uncertainties related to the financial aspects of owning or renting out properties.

**House rental system architecture.** The overall structure and components that make up a system designed for managing house rentals.

**Stakeholders.** Individuals or entities involved or affected by the rental system, such as property owners, tenants, and administrators.

## **Related Literature**

Prior to the revolutionary emergence of machine learning in predictive analysis, traditional statistical approaches for predicting the future prices of financial instruments like capital stocks were used. Technical analysts used other valuable assets such as house units. Unfortunately, typical statistical methods are inadequate for dealing with apartment renting. If there are too many explanatory variables to track, price prediction becomes challenging; consequently, machine learning is required to uncover the patterns and relationships existing in the data. Using the appropriate machine learning algorithms, it is feasible to convert people's experience (training data) into expertise (model). This enables realistic rental house pricing assessments. Years of real estate agent pricing apartment experience can be generalized into models that anticipate prices for previously unseen apartment sites. (Kok et al.)

Regression, at its essence, entails fitting a function as effectively as feasible through some data points. Multiple regression analysis is an excellent tool for forecasting the values of our target variable rental price  $y$ , which is dependent on other factors (or, in other words, independent variables) with numerical representations  $x$ . Some variables in regression analysis have a relationship termed a relationship in which changes in one variable cause changes in the other. Regression methods depict how changes in any given parameter affect the target value, and can be used to forecast a target value based on previously unknown parameters (Varma et al.).

House price forecasting is an important topic in the field of real estate. The literature aims to extract valuable insights from historical data in property markets. As a crucial aspect of real estate, particularly in markets like Mumbai, India, this study focuses on developing effective price prediction models using machine learning techniques and data collected from 99acres.com, a leading real estate website in India.

Data exploration involves analyzing the size, accuracy, patterns, and other attributes of the dataset. The data was collected from 99acres.com and specifically focuses on houses in Mumbai. The dataset was divided into a 9:1 ratio for training and testing purposes. Data visualization tools and techniques, such as charts, graphs, and maps, are used to represent information visually. In the context of

big data, data visualization plays a vital role in analyzing large amounts of information and identifying trends, outliers, and patterns. Python, along with libraries like Pandas, NumPy, Matplotlib, Seaborn, Scikit-Learn, and XGBoost, was employed for data analysis, modeling, and visualization in this project. Data selection is the process of determining the appropriate data type, source, and instruments for data collection. It precedes the actual data collection process and involves selecting relevant data based on research questions, existing literature, and accessibility to data sources. Log transformation is used to reduce skewness in highly skewed distributions. It helps make data patterns more interpretable and ensures the assumptions of inferential statistics are met. Log-transformed data allows for a comparison of geometric means.

Regression models, including linear regression, random forest regression, and XGBoost regression, were utilized for house price prediction in this study. The study found that linear regression exhibited the best performance among the models, making it suitable for deployment purposes. Random forest regression and XGBoost regression showed inferior performance and are not recommended for further deployment.

In conclusion, the aim of this study was successfully achieved by developing and evaluating house price prediction models using data from 99acres.com. The analysis revealed that the most effective model for predicting house prices in Mumbai was the linear regression model. The attribute that significantly influenced price prediction was the circle rate. With an RMSE score of 0.5025658262899986, the linear regression model demonstrated accurate performance. Recommendations for future deployment purposes suggest utilizing the linear regression model for house price predictions based on the findings of this study. (Udit and Uday Deo)

## **Methodology**

CRISP-DM is the methodology that the researchers employed in this study. This methodology will be the foundation for their study on predicting housing and apartment rental prices using regression analysis algorithms. Following this structured methodology, the researchers ensured a systematic and

organized approach to their research, allowing them to effectively address the various stages of the data mining process. The CRISP-DM framework provided a clear roadmap, guiding us the researchers through the stages of business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This methodology facilitated the researchers in achieving the objectives, including identifying key determinants of rental prices, evaluating the performance of regression analysis algorithms, and providing valuable insights for stakeholders in the real estate industry.

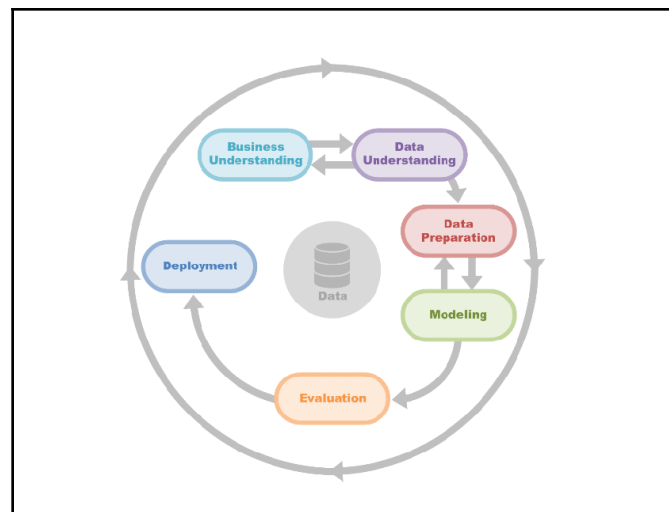


Figure 1.0 Crisp-DM

Source: CRISP-DM (data mining framework). | Download Scientific Diagram (researchgate.net)

## 1. Business Understanding

The first phase of CRISP-DM is the Business Understanding phase, which focuses on understanding the project objectives, requirements, and constraints in order to align the data mining process with business goals.

### 1.1 Assess the Situation

This entails gathering detailed information about the available resources, such as data sources, tools, and expertise, as well as understanding any constraints or limitations such as data quality and regulatory requirements.

### 1.1.1 Inventory of Resources

List the resources available to the project including:

**Personnel**, the involvement of skilled personnel plays a vital role in ensuring accurate and insightful results. This skilled personnel are the team called “*code alchemist*” which composed of four (4) members whose:

- Mr. Jake Sebastian S. Forbes, the developer.
- Ms. Reignoel D. Rodriguez, the data analyst.
- Ms. Angela Jane P. Boteros, the database specialist.
- Ms. Karen B. Arimuhanan, the dashboard designer.

**Data**, the availability and quality of data are crucial resources. This data are the variables that might determine the qualifications of the data, if it is ready for data analysis for predicting house/rental prices or not. Here are some data resources that researchers identified while analyzing the data:

- Prices
- Date
- Property Size (Floor Area)
- Amenities (Bedroom and Baths)
- Location (Municipality and Province)
- Agent (agent membership, agent name and verification of agent)
- Vendor (Lamudi or My Property)

**Computing resources**, adequate computing resources are essential to handle the computational requirements of the data analysis process.

- 8GB of RAM
- Windows 11

**Software**, Utilizing specialized software and tools for data analysis and regression modeling is crucial.

- Python Programming Language
- Anaconda Navigator 3
- Visual Studio Code for data exploration and developing the website
- PowerBI for data visualization
- Sqlyog for data warehousing
- Xampp for the database storage of the website

### **1.1.2 Requirements**

**User Friendly Interface.** The system should provide a user-friendly interface for easy interaction with the system. Users, such as landlords, tenants, or real estate professionals, should be able to input property information and retrieve predictions seamlessly. The interface should be intuitive, visually appealing, and provide clear and understandable results.

**Predicting Price Capability.** Providing price predictions for landlords, the integrated system empowers them to make data-driven decisions, optimize their rental income, and enhance their overall management of housing and rental properties.

**Accuracy and Reliability.** The system should ensure that the regression analysis algorithms used for prediction are accurate and reliable. The chosen algorithms should be capable of capturing complex relationships between the input features and the target variable, leading to accurate price estimations.

## **1.2 Desired Outputs of the Project**

The desired output of the project will be able to help the landlords with valuable insights and tools to make informed decisions about their rental properties.

- **Business Success Criteria**

In predicting housing and rental prices establishing a business success criteria is essential to evaluate the effectiveness and value of the predictive models.

**Predictive Model Accuracy** An important metric for success is how accurate the rental price forecasts are. Metrics like MAE and RMSE might be used as the success criteria. Predictions with lower error levels are more accurate.

**Scalability and Adaptability** prediction models should be able to handle various market conditions, property types, and geographic regions. Evaluation of the models performance across several geographies, as well as their capacity for handling newly acquired inputs and adapting to changing market dynamics.

**Risk Mitigation** Effective rental price prediction will help mitigate risks associated with underpricing or overpricing properties. Success criteria can include evaluating the models' ability to minimize revenue losses due to incorrect pricing, and reduce financial risks for property owners.

- **Data Mining Success Criteria**

**Model Performance,** Good model means the prediction has low prediction values, indicating accurate prediction of housing and rental prices.

**Feature Selections,** A good model should prioritize important features and provide insights into the relationships between these features and rental prices.

**Data Validity,** The model should be able to handle different amenities, locations, and time periods while maintaining its predictive accuracy.

To ensure the development of a robust model capable of providing accurate estimates of house and rental prices, researchers employ a comprehensive approach that involves web crawling through various websites to gather a significant volume of data. Instead of relying on a limited number of datasets, the researchers strive to collect an extensive dataset consisting of approximately fifteen thousand samples. This large and diverse dataset enables the researchers to capture a wide range of property characteristics, geographic locations, and rental prices, ensuring a more comprehensive representation of the housing and rental market.



- **Produce Project Plan**

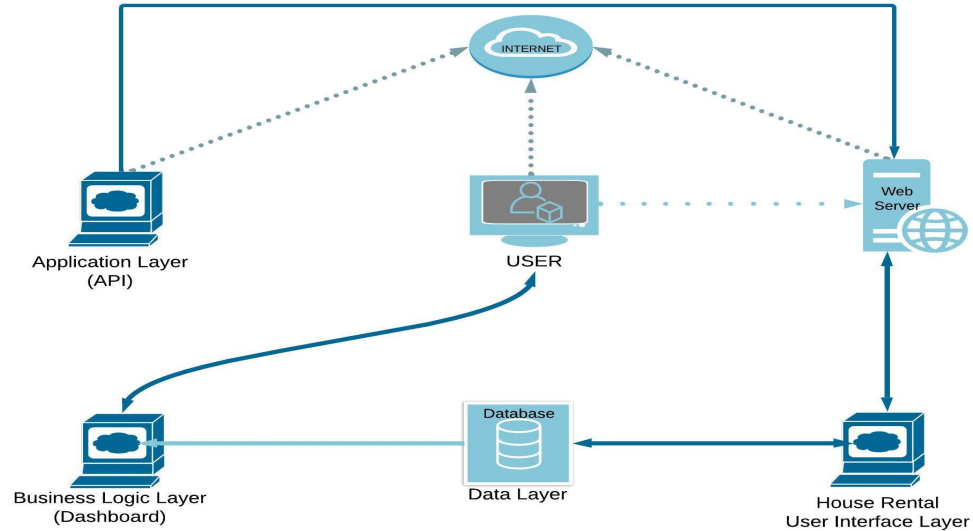


Figure 1.2 System Architecture of the system

Figure 1.2 is the house rental system architecture consists of several components that work together to facilitate the rental process. The stakeholders can interact with the system through various interfaces. Here's an overview of the components:

## User Interface Layer

**Web Application.** A user-friendly web application allows stakeholders to access the system using their web browsers. It provides interfaces for property owners to list their houses, tenants to search for available houses, and administrators to manage the system.

## Application Layer

***House Listing and Search.*** This module handles house listing functionality, allowing property owners to add details about their houses for rent, such as location, features, and rental terms. Tenants can search for houses based on their preferences.

***User Management.*** This module manages user profiles, authentication, and authorization. It ensures that stakeholders can securely access the system and their respective functionalities.

#### **Business Logic Layer**

***House Management.*** This component handles house-related operations, including house listing, availability management, and updating house details.

***User Administration.*** This component manages user accounts, authentication, and authorization, allowing administrators to oversee user-related activities.

***Dashboard.*** This component displays KPI's, such as the total number of tenants, number of pending complaints, and the total earning for the month.

#### **Data Layer**

***Database.*** The system utilizes a database to store and manage data, including house listings, house information, user profiles, and system configurations.

#### **Stakeholder Usage**

***End User.*** They can monitor the system, manage user accounts, handle disputes, generate reports, and ensure smooth operations. They can manage house listings and access financial reports. They can search for houses based on their preferences and view property details.

## **2. Data Understanding**

This phase involves a comprehensive exploration and familiarization with the available data. It ensures a comprehensive understanding of the data's characteristics, quality, and suitability for regression analysis algorithms. Researchers can make informed decisions regarding data preprocessing, feature selection, and model design to build accurate predictive models for housing and apartment rental price estimation.

## 2.1 Initial Data Report

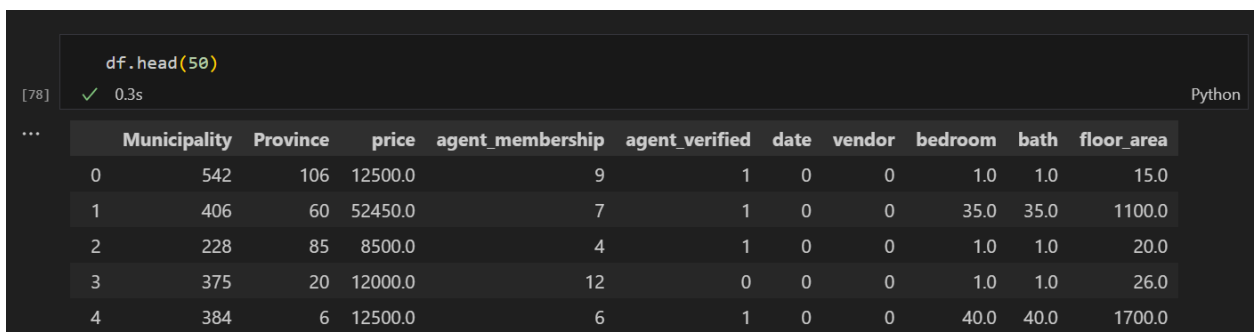
The data that has been collected is from the sites of my property and lamudi websites, these websites are online platforms or websites that specialize in real estate listings and property-related services. These websites serve as marketplaces where users can search for, buy, sell, or rent properties, including houses, apartments, land, and commercial spaces.

During the data crawling process, the researchers immediately noticed that the 'features' or amenities variable contained values for 'Bedroom,' 'Bath,' and 'Floor Area' combined within a single column. Recognizing the need for a more structured and granular representation of these variables, the researchers decided to create a new column specifically dedicated to capturing the individual values for Bedroom, Bath, and Floor Area. This approach aimed to improve data organization and facilitate further analysis by separating these important features into distinct columns. The data type of these new columns are string, so we converted it into an integer. So that the researchers will be able to accurately predict the prices based on these columns.

The researchers encountered a minor issue regarding the presence of symbols in the price variable. In order to ensure consistency and enable appropriate numerical calculations, the researchers made the decision to remove these symbols from the prices. This step aimed to convert the price values into a float data type, allowing for accurate mathematical operations and analysis. The researchers also saw the date that is string, so they converted it into datetime.

## 2.2 Describe Data

To check the datasets that have been collected, the researchers used the `dataFrame.head()` method to get the datasets.



```
df.head(50)
```

	Municipality	Province	price	agent_membership	agent_verified	date	vendor	bedroom	bath	floor_area
0	542	106	12500.0	9	1	0	0	1.0	1.0	15.0
1	406	60	52450.0	7	1	0	0	35.0	35.0	1100.0
2	228	85	8500.0	4	1	0	0	1.0	1.0	20.0
3	375	20	12000.0	12	0	0	0	1.0	1.0	26.0
4	384	6	12500.0	6	1	0	0	40.0	40.0	1700.0

As you can see, the figures are the datasets that have been collected by the researchers.

And to identify the number of rows and columns of the dataset, the researchers have used the `dataFrame.shape()` method to get the number of rows and columns. The result of this method is:

```
df.shape
[6] ✓ 0.1s Python
... (13855, 13)
```

As you can see, the researchers got 13,855 rows and 13 columns.

The researchers also checked the columns to see the variables and identify which columns are to be dropped or need to be refined. They used the `dataFrame.columns()`, to be able to see the columns.

```
df.columns
[74] ✓ 0.1s Python
... Index(['Municipality', 'Province', 'price', 'agent_membership',
         'agent_verified', 'date', 'vendor', 'bedroom', 'bath', 'floor_area'],
        dtype='object')
```

As you can see, the researchers got 10 columns.

The researchers also checked the data types of each column. They used the `dataFrame.dtypes` to get the data types of each column.

```
df.dtypes
✓ 0.1s Python
Unnamed: 0      int64
title           object
Municipality    object
Province        object
price           int64
agent           object
agent_membership int64
agent_verified  object
date            object
vendor          object
bedroom         int64
bath            int64
floor_area      int64
dtype: object
```

As you can see, there are so many object data types and int64.

The researchers saw that the price data type is int64, they wanted it to be float so that it is more accurate when predicting the price. And also, there are sentences like 'Contact Agent for price'. So the researchers changed it into a None value.

```
#Replacing the 'contact agent for price' into null values
def change_price(col):
    price_change = 'Contact agent for price';

    df[col][df[col] == price_change] = None;
```

✓ 0.1s Python

In this code, they replaced the string 'contact agent for price' into none values.

## 2.3 Verify Data Quality

The verification of data quality guarantees the integrity and reliability of the dataset, thus enabling accurate and insightful analysis.

### 2.3.1 Missing Data

Before checking the missing data, the researchers have checked the duplicate values first.

```
df.duplicated().sum()
```

✓ 0.0s Python

978

As you can see, they got 978 duplicate values. They removed the duplicated values and proceeded onto the next step.

After they check the duplicate values, they check the data if there is missing value using the `dataFrame.isna().sum()` method.

```
df.isna().sum()
```

✓ 0.0s Python

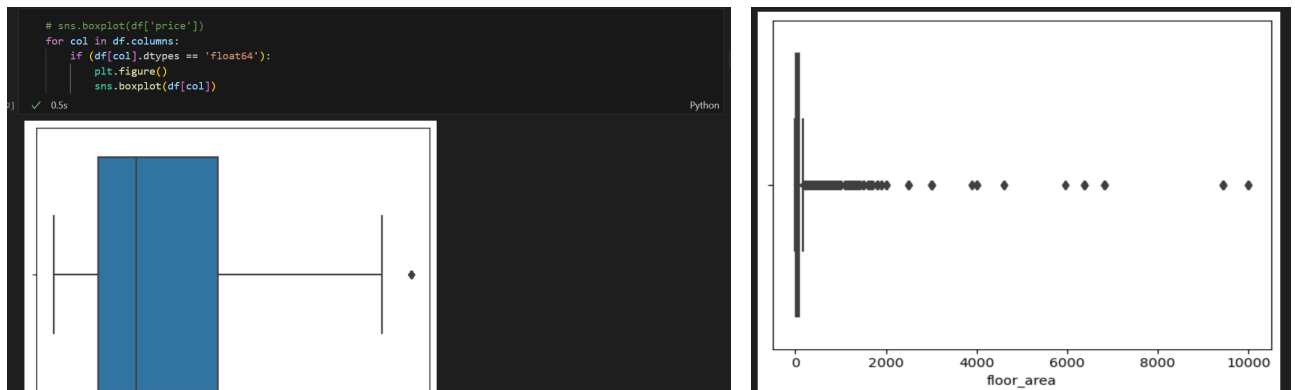
Municipality	0
Province	1988
price	0
agent_membership	0
agent_verified	0
date	0
vendor	0
bedroom	0
bath	0
floor_area	0
dtype: int64	

As you can see, there are missing values in the 'Province' variable. It has 1,988 missing values.

### 2.3.2 Outliers

In order to have a good model and analysis, the data should have no extreme values.

The first thing that the researchers did was to check the values if there are extreme values.

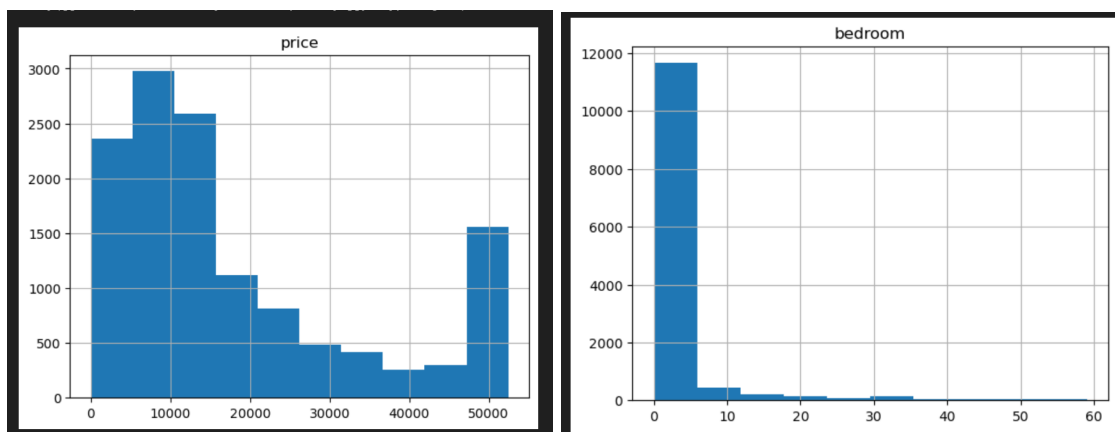


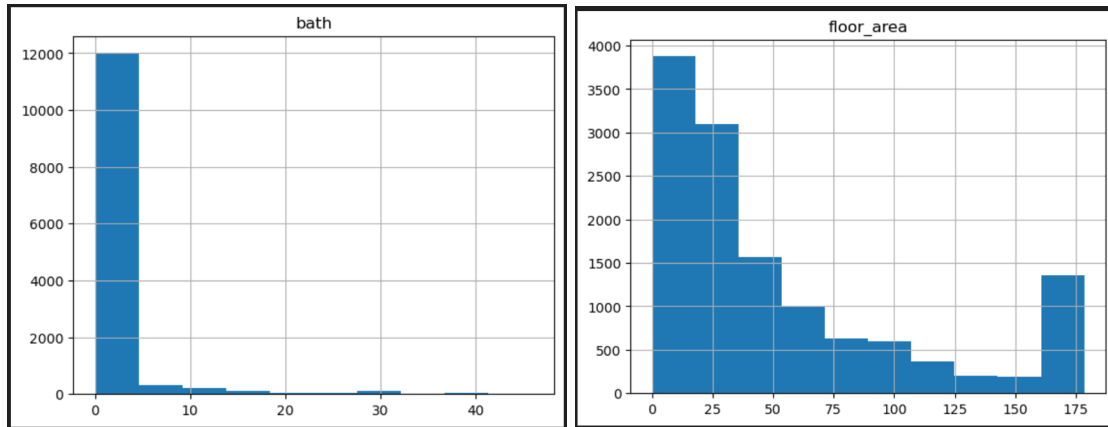
As you can see, these are the extreme values. They will remove that in the data cleaning process.

## 2.4 Initial Data Exploration

### 2.4.1 Distributions

The researchers used histogram to show if the feature or column is normally distributed or not. To show these, they used the `dataFrame.hist(column='')` method.

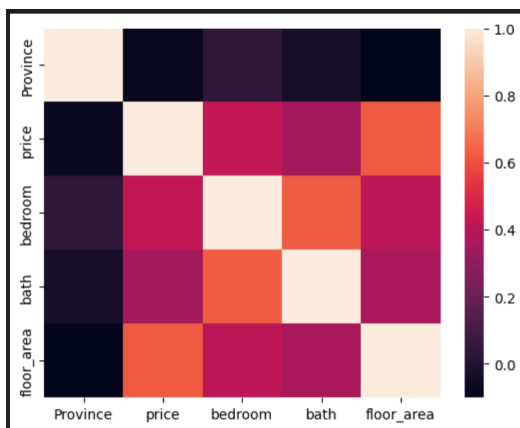




Based on the diagrams provided, it is evident that the features exhibit a left-skewed distribution. In a left-skewed distribution, the majority of the data points are concentrated towards the right side, while the tail of the distribution extends towards the left. This means that the data has a long tail on the left side, with fewer extreme values in that direction. Left-skewed distributions are also known as negatively skewed distributions.

## 2.4.2 Correlations

To identify if there is a relationship between variables, the researchers used the `sns.heatmap(dataFrame.corr)`.



In the figure, the researchers can clearly see that there is no relationship between variables or features.

## 2.5 Data Quality Report

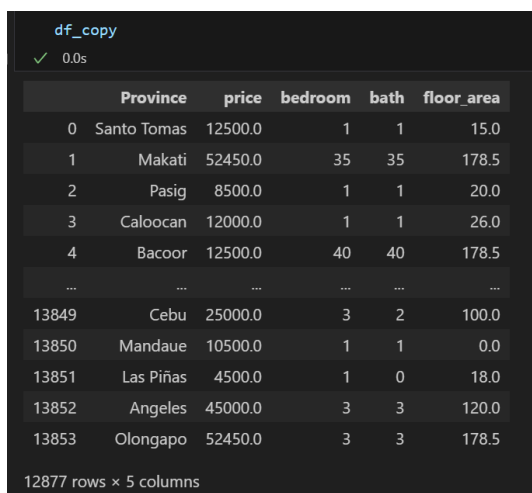
Upon examining the data, the researchers identified anomalies and outliers within the datasets. Recognizing that these irregularities could significantly impact the integrity and accuracy of subsequent analyses, the researchers diligently undertook data cleaning procedures to eliminate these outliers. The researchers also removed duplicated data, filled in missing values and removed extreme values. Researchers can say that the data is good and ready for the data preparation phase for selecting various predictive models to predict house and rental prices.

## 3. Data Preparation

This phase focuses on cleaning the data, handling missing values, encoding categorical variables, and creating new derived features. The data preparation phase is crucial in ensuring the data's quality, consistency, and suitability for regression analysis algorithms. Cleaning the data, selecting relevant features, encoding categorical variables, engineering new features, and appropriately scaling the data, researchers can create a well-prepared dataset ready for modeling and predicting housing and apartment rental prices.

### 3.1 Data Selection

The research selects the target variable based on our goal. Our goal is to be able to predict the house and rental prices. They choose the features affecting price. These features are the price (target variable), province, bedroom, bath and also the floor area of the house and apartment.



df\_copy  
✓ 0.0s

	Province	price	bedroom	bath	floor_area
0	Santo Tomas	12500.0	1	1	15.0
1	Makati	52450.0	35	35	178.5
2	Pasig	8500.0	1	1	20.0
3	Caloocan	12000.0	1	1	26.0
4	Bacoor	12500.0	40	40	178.5
...	...	...	...	...	...
13849	Cebu	25000.0	3	2	100.0
13850	Mandaue	10500.0	1	1	0.0
13851	Las Piñas	4500.0	1	0	18.0
13852	Angeles	45000.0	3	3	120.0
13853	Olongapo	52450.0	3	3	178.5

12877 rows x 5 columns



In these, the researchers copy the original datasets to preserve the data.

## 3.2 Data Cleaning

### 3.2.1 Drop Duplicates

The first thing that the researchers did in order to drop the duplicates is to check duplicate values by using `dataFrame.duplicated().sum()` method. With these methods, the researchers saw that there are 978 duplicate values in the datasets. And to drop this duplicates, the researchers used the `dataFrame.drop_duplicates()` method.

```
23] df.duplicated().sum()
✓ 0.0s Python
978

24] df = df.drop_duplicates()
✓ 0.0s Python
```

### 3.2.2 Fill-in Missing Values

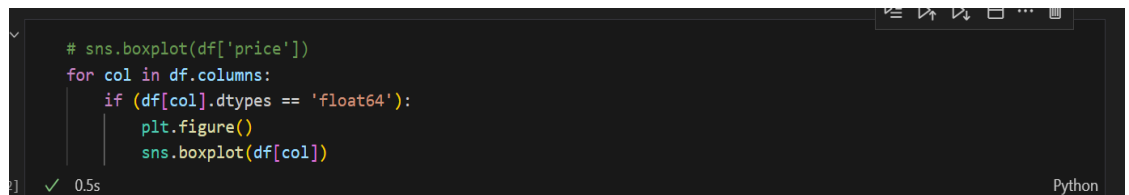
After removing duplicate values from the dataset, the researchers proceed to address the issue of missing data. The first step in filling in the missing values is to identify the features or variables that contain missing values. To accomplish this, the researchers utilize the `dataFrame.isna().sum()` function, which allows them to obtain a count of missing values for each feature in the dataset.

```
df.isna().sum()
✓ 0.0s
Municipality      0
Province         1988
price             0
agent_membership  0
agent_verified    0
date              0
vendor            0
bedroom           0
bath              0
floor_area        0
dtype: int64
```

In these figures, you can clearly see that there are 1,988 missing values in the 'Province' features. After finding the missing values, the researchers filled-in the data with 'NA', since this is a string value.

### 3.2.3 Remove Extreme Values

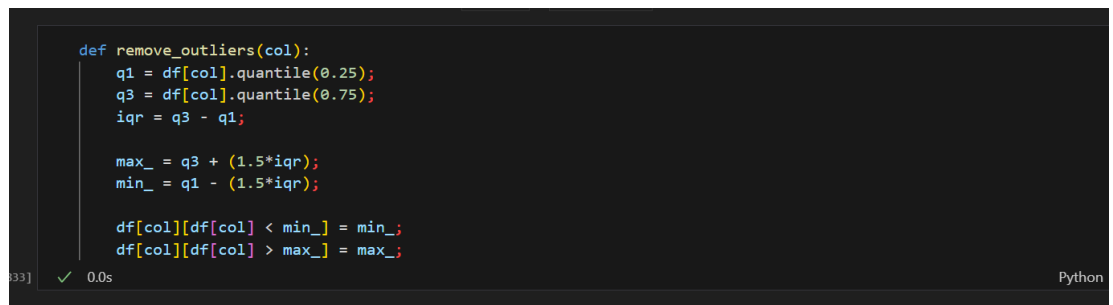
Once the researchers have filled in the missing values, the subsequent step in the data cleaning process is to identify and remove extreme values that could potentially interfere with the analysis. When dealing with outliers, it is important to first visualize the distribution and identify extreme values through a diagram or graph. To achieve this, the researchers employ the `sns.boxplot(dataFrame['col'])` method, where 'col' represents the column of interest in the dataset. By utilizing a boxplot, they can visualize the distribution of values within the specified column and identify any extreme values that lie outside the whiskers of the boxplot.



```
# sns.boxplot(df['price'])
for col in df.columns:
    if (df[col].dtypes == 'float64'):
        plt.figure()
        sns.boxplot(df[col])
```

The image shows a Jupyter Notebook cell with Python code. The code iterates through all columns of a DataFrame 'df'. For each column that is of type 'float64', it creates a new figure and generates a boxplot using 'sns.boxplot(df[col])'. The cell has a green checkmark and a runtime of 0.5s. The Python logo is visible in the bottom right corner.

The researchers employ a process of plotting columns with outliers and utilize quantile percentiles to remove these outliers. This approach allows for a more robust analysis by eliminating extreme values that may skew the results and affect the accuracy of predictions.



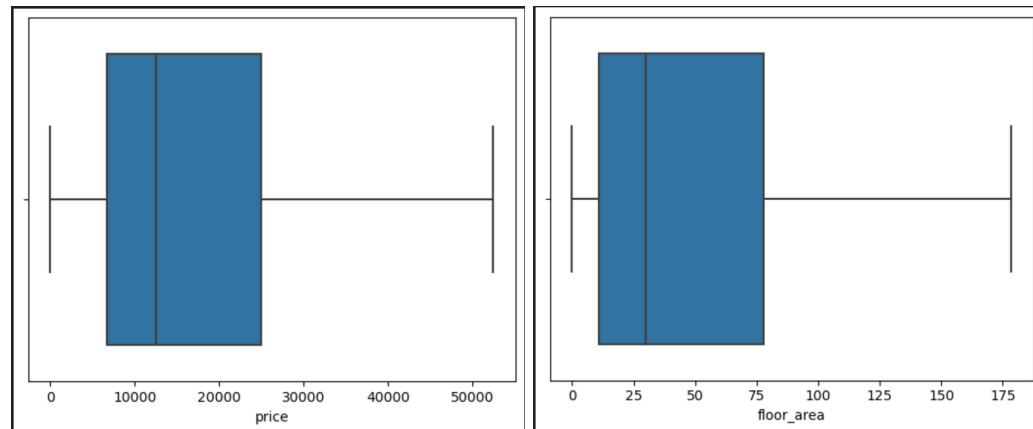
```
def remove_outliers(col):
    q1 = df[col].quantile(0.25);
    q3 = df[col].quantile(0.75);
    iqr = q3 - q1;

    max_ = q3 + (1.5*iqr);
    min_ = q1 - (1.5*iqr);

    df[col][df[col] < min_] = min_;
    df[col][df[col] > max_] = max_;
```

The image shows a Jupyter Notebook cell with a Python function 'remove\_outliers'. The function takes a column name 'col' as input. It calculates the first quartile (q1), third quartile (q3), and interquartile range (iqr). It then determines the maximum and minimum values that define the range of non-outliers (1.5 \* iqr from the quartiles). Finally, it updates the DataFrame 'df' by replacing values below 'min\_' with 'min\_' and values above 'max\_' with 'max\_'. The cell has a green checkmark and a runtime of 0.0s. The Python logo is visible in the bottom right corner.

After these, the outliers have already been removed.



Due to the nature of the price and floor variables having a float data type, the researchers decided to specifically focus on removing outliers from these two variables. Outliers are extreme values that deviate significantly from the majority of the data points and can potentially skew the analysis or model results.

#### 3.2.4 Convert Categorical values into numerical

After removing outliers from the dataset, the researchers proceed to convert categorical values into a numerical format, which is necessary for predicting house and rental prices. To accomplish this, the researchers employ the Label Encoding technique, a common method used for handling categorical variables. The researchers apply Label Encoding to the 'Province' feature, which initially consists of string values representing different provinces. Label Encoding assigns a unique numeric code to each distinct category within the 'Province' feature, effectively transforming it into a numerical representation.

```
from sklearn.preprocessing import LabelEncoder;

encoder = LabelEncoder();

df['Province'] = encoder.fit_transform(df['Province'])

]
```

df  
✓ 0.1s

	Province	price	bedroom	bath	floor_area
0	106	12500.0	1	1	15.0
1	60	52450.0	35	35	1100.0
2	85	8500.0	1	1	20.0
3	20	12000.0	1	1	26.0
4	6	12500.0	40	40	1700.0
...	...	...	...	...	...
13849	24	25000.0	3	2	100.0
13850	67	10500.0	1	1	0.0
13851	51	4500.0	1	0	18.0
13852	0	45000.0	3	3	120.0
13853	77	52450.0	3	3	217.0

### 3.2.5 Feature Scaling

In the feature scaling, the researchers will scale the data to standardize the range of independent variables or features of data. The sklearn.preprocessing MinMaxScaler, will be used to scale the data.

```
from sklearn.preprocessing import MinMaxScaler;

scaler = MinMaxScaler()
df_scaled = df.copy()
✓ 0.2s

for col in df.columns:
    if (col != 'price'):
        df_scaled[col] = scaler.fit_transform(df_scaled[col].values.reshape(-1,1))
✓ 0.2s
```

## 3.3 Feature Selection, Engineering & Extraction

The researchers focus on refining the dataset by selecting relevant features, creating new features, and extracting valuable information from the existing data. This phase aims to enhance the predictive power of the model and improve the accuracy of the price predictions.

### 3.3.1 Feature Selection

In the feature selection phase, the researchers carefully choose the features that are expected to have a significant impact on predicting the prices of houses and rentals. The selected features are considered to be highly relevant and informative for the prediction task.

df\_scaled ## final dataset

✓ 0.2s

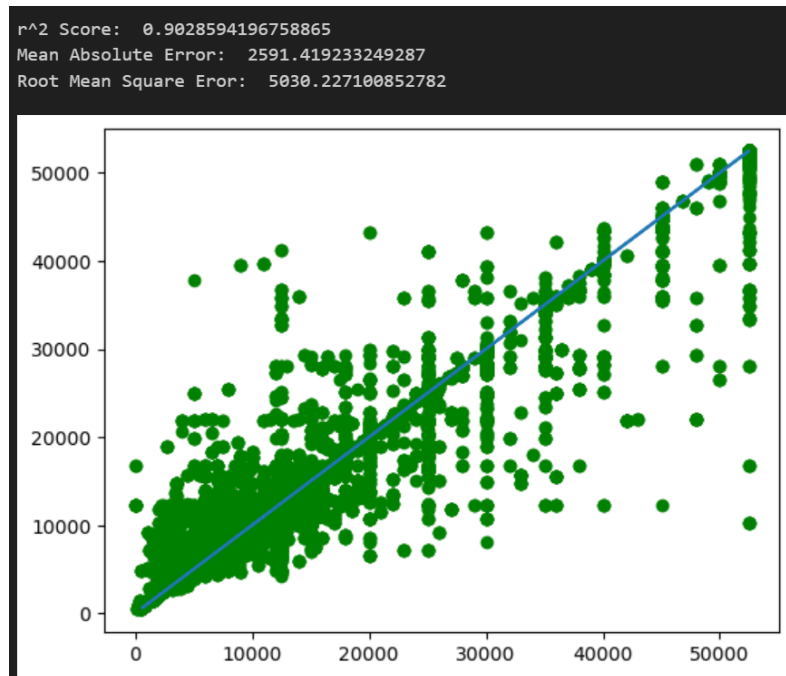
	Province	price	bedroom	bath	floor_area
0	0.809160	12500.0	0.003205	0.005	0.0015
1	0.458015	52450.0	0.112179	0.175	0.1100
2	0.648855	8500.0	0.003205	0.005	0.0020
3	0.152672	12000.0	0.003205	0.005	0.0026
4	0.045802	12500.0	0.128205	0.200	0.1700
...	...	...	...	...	...
13849	0.183206	25000.0	0.009615	0.010	0.0100
13850	0.511450	10500.0	0.003205	0.005	0.0000
13851	0.389313	4500.0	0.003205	0.000	0.0018
13852	0.000000	45000.0	0.009615	0.015	0.0120
13853	0.587786	52450.0	0.009615	0.015	0.0217

12877 rows × 5 columns

The researchers select the ‘Province’, ‘price’, ‘bedroom’, ‘bath’, and ‘floor\_area’ features to accurately predict the house and rental price estimation. After removing the other features, the accuracy and efficiency of the predictive model has improved. The selected features are expected to capture various aspects that directly or indirectly affect the prices of houses/rentals. These features could include quantitative variables such as the number of bedrooms and baths, the floor area, or the location's proximity to amenities.

### 3.3.2 Feature Engineering & Extraction

The researchers focus on refining the dataset by selecting relevant features, creating new features, and extracting valuable information from the existing data. The feature engineering & extraction aims to enhance the predictive power of the model and improve the accuracy of the price predictions. The highest accuracy of the feature engineering in our study is 91.24% with  $r^2$  score of 90.28%, RSME is 5030.22, and MAE is 2591.41 prediction error. This is the Principal Component Analysis, where it is a linear dimensionality reduction technique that can be utilized for extracting information.



#### 4. Modeling

The researchers develop and apply appropriate regression analysis algorithms to build predictive models. The researchers select the regression algorithms that are most suitable for their price prediction task. This involves considering various factors such as the nature of the dataset, the complexity of the problem, and the specific goals of the study. Once the regression algorithms are chosen, the researchers proceed with training the models using their prepared dataset. They input the relevant features or variables, such as location, size, amenities, and other factors that influence rental price. The models then learn from this data to establish the underlying patterns and associations. After training and refining the models, the researchers evaluate their effectiveness and performance. They assess the models' predictive power by measuring metrics. This evaluation enables them to understand how well the models can estimate rental prices and identify the key determinants that significantly influence the pricing.

##### 4.1 Splitting Training and Test Dataset

After the data cleaning and preparation stages, the researchers move on to the data modeling phase. To build a robust and reliable model, The Researchers begin by splitting the

available data into training and test sets. The researchers allocate 85% of the data for training purposes, which involves fitting the model to learn the underlying patterns and relationships in the datasets. The model can learn from a significant amount of data thanks to this greater fraction of the data, which improves its capacity for precise prediction. The test datasets, which are used to assess the effectiveness of the trained model, are reserved for the remaining 15% of the data. Keeping this portion of the data separate, the researchers can assess how well the model generalizes to unseen data and estimate its predictive capabilities.

The researchers set a specific value of random state which is 42 to ensure reproducibility and consistency. It can replicate the results and compare different models or techniques effectively.

```
X_train, X_test, y_train, y_test = train_test_split(X,y,random_state=42, train_size = .85);
```

## 4.2 Choosing an Algorithm

In the initial stage of the analysis, the researchers focused on understanding the dataset and identifying the target variable. Since the target variable in this study is a numerical value representing the price, the researchers determined that a regression analysis algorithm would be suitable for predicting housing and apartment rental prices.

Once the target variable was identified, the researchers proceeded to explore and compare various regression analysis algorithms to select the one that would yield the highest accuracy scores. This involved testing and evaluating multiple algorithms to determine their predictive performance. The researchers aimed to find an algorithm that could effectively capture the relationships between the input features and the target variable, leading to accurate price predictions. They assessed the performance of different algorithms by considering evaluation metrics such as mean squared error, mean absolute error, or R-squared.

Through this comparative analysis, the researchers were able to identify the predictive algorithm that demonstrated the highest accuracy scores. This algorithm was considered the most suitable for predicting housing and apartment rental prices based on the dataset at hand.

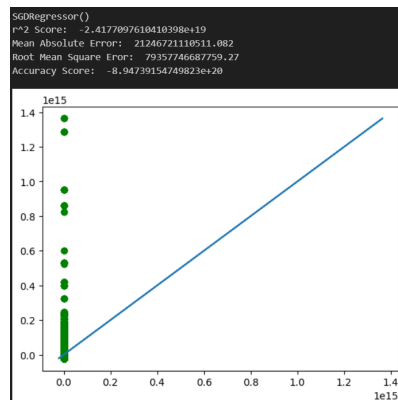
### 4.3 Fitting the Model

The researchers fit the model to find who has the highest accuracy score to predict the price of houses/rental. The researchers used the `model.fit(train_datasets, test_datasets)`

```
model.fit(X_train, y_train);
```

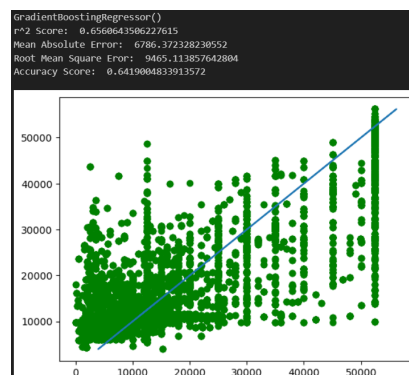
#### 4.3.2 Stochastic Gradient Descent Regression

In the `SGDRegressor`, the researchers got a very low accuracy score since the results are negative. Who got a -8.947% accuracy score. Indicating that this model is not good for predicting the price.



#### 4.3.4 Gradient Boosting Regression

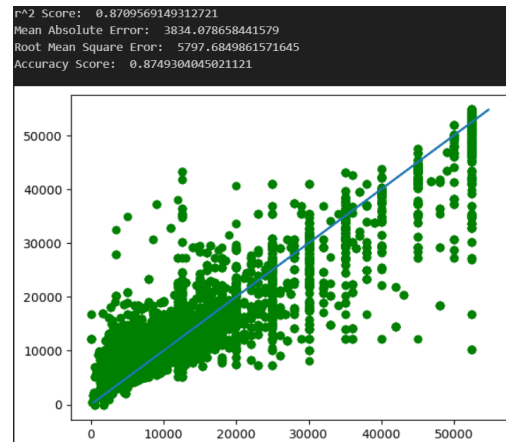
The Gradient Boosting Regression also has a low accuracy score. Who got a 64.19% accuracy score. Indicating that this model is not good for predicting the price.





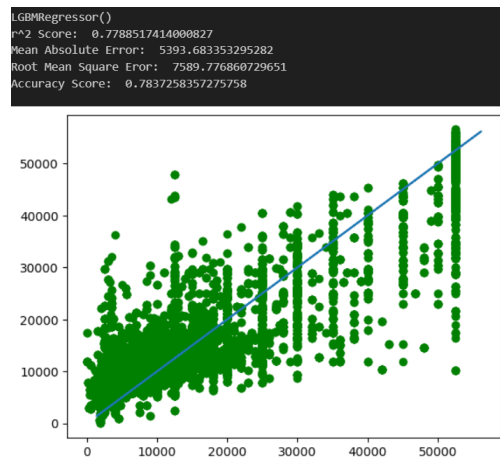
#### 4.3.7 XGBoost Regression

In this model, the researchers got an 87.49% accuracy score. Indicating that this is an acceptable model and have the ability to predict the house/rental prices.



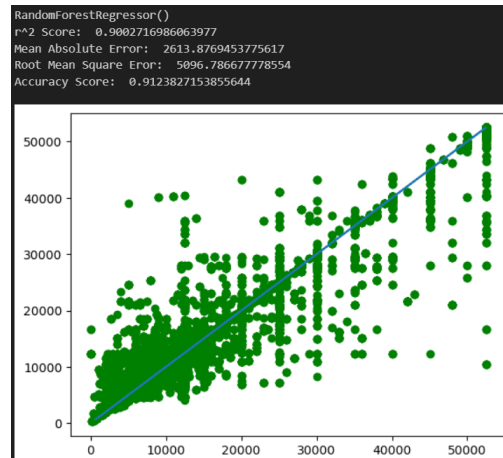
#### 4.3.8 Light GBM Regression

In these models, researchers almost got an acceptable model. It has a 78.37% accuracy score. The researchers think that this model will not be able to give a good prediction of house/rental prices.



#### 4.3.8 Random Forest Regression

In this model, researchers got a 91.23% accuracy score. Which determines that this is a good model and that can be used for predicting accurately the house/rental prices.



After fitting different regression algorithms and evaluating their performance, the researchers identified the algorithm that yielded the highest accuracy scores for predicting housing and apartment rental prices. This algorithm demonstrated superior predictive capabilities in capturing the relationships between the input features and the target variable.

#### 4.4 Hyperparameter Tuning

The researchers did not employ the hyperparameter tuning as the model is already good for predicting the house/rental prices.

### 5. Results and Evaluation

After applying various regression analysis algorithms and building predictive models, the researchers proceeded to evaluate the performance and effectiveness of these models. During the evaluation stage, the researchers assessed the quality and accuracy of the developed models in predicting rental prices. Comparing the predicted rental prices with the actual rental prices from the dataset, the researchers gained insight into the models' predictive capabilities. The evaluation stage allowed the researchers to analyze the significance and contribution of various factors or determinants in influencing rental prices. They examined the coefficients or weights assigned to different features in the regression models to identify which variables had the most substantial impact on rental prices. This analysis helped in understanding the key drivers and factors affecting the pricing dynamics in the real estate market.

## 5.1 Model Accuracy Reports

Model	r <sup>2</sup> Score	MAE	RMSE	Accuracy Score
BayesianRidge	0.131769058	11761.13731	15038.5123	0.138380531
SGDRegressor	-2.42E+19	2.12467E+13	7.93577E+13	-8.95E+20
ElasticNet	0.131785061	11762.37228	15038.37371	0.138417933
GraddientBoostingRegressor	0.656064351	6786.372328	9465.113858	0.641900483
LinearRegressor	0.131729248	11758.42697	15038.85707	0.13837332
KernelRidge	-0.26410057	13291.62499	18145.88104	-0.200903055
XGBRegressor	0.870956915	3834.078658	5797.684986	0.874930405
LGBMRegressor	0.778851741	5393.683353	7589.776861	0.783725836
RandomForestRegressor	0.900271699	2613.876945	5096.786678	0.912382715

Table 1.1 Tabular Results of the evaluated model

Table 1.1 shows the tabular results of the evaluated model, as you can see the potential models that the researchers can use in predicting the house/rental prices is the XGB Regression and the RandomForest Regression Algorithms. Who got a 87.49% accuracy score for XGB Regression and 91.23% for Random Forest Regression.

## 5.2 Evaluation

To choose the best model for predicting housing and apartment rental prices, the researchers employed several evaluation metrics to assess the prediction error and accuracy. Two commonly used metrics in this study were the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The average size of the forecast mistakes is determined by the RMSE. The average of the squared differences between the expected and actual values is computed. A lower RMSE number reflects higher prediction accuracy since it reflects a better fit between the predicted and actual prices. The MAE measures the average magnitude of the absolute differences between the predicted and actual values. It provides a more straightforward interpretation of the prediction error. Again, a lower MAE value indicates better prediction accuracy, with smaller deviations between the predicted and actual prices. The closer these metrics are to zero, the better the model performs in capturing the underlying patterns and relationships in the data, resulting in more accurate predictions.

The accuracy of the model's predictions is also evaluated by the researchers through cross-validation. In cross-validation, the data are divided into several subsets, the model is trained on a subset of the data, and its performance is assessed on the remaining subset of the data. Through this approach, the model's capacity for generalization is estimated, and its prediction accuracy is more thoroughly assessed.

### **5.3 Choosing a Model**

The best model for forecasting housing and apartment rental costs, according to the researchers' evaluation of numerous regression algorithms, is the Random Forest Regression algorithm. This model was chosen since it outperformed all other examined models in terms of accuracy and had a low prediction error. The choice of this algorithm guarantees that the researchers have a trustworthy and effective instrument for supporting real estate market stakeholders.

## **6. Deployment**

The researchers create an API that provides real-time predictions based on the trained regression models. The deployment stage involves ensuring the scalability, reliability, and efficiency of the deployed system. This phase focuses on creating a usable and reliable system that enables stakeholders to access accurate rental price predictions. Effectively deploying the models, the researchers can provide valuable insights to the real estate industry and assist in decision-making processes related to rental pricing.

To enhance the usability and accessibility of the chosen Random Forest Regression model for predicting housing and apartment rental prices, the researchers integrated it with an Application Programming Interface (API) within the system. Integrating the model with an API, users can interact with the model and make predictions conveniently through a user-friendly interface. The API serves as an intermediary that allows the system to communicate with the model and provide seamless access to its prediction capabilities.

### Works Cited

- Deo, Udit. "HOUSE PRICE PREDICTION USING REGRESSION TECHNIQUES." *researchgate.net*, [https://www.researchgate.net/publication/349477129\\_House\\_Price\\_Prediction](https://www.researchgate.net/publication/349477129_House_Price_Prediction).
- Karanko, Lauri. "Rental Price Prediction on Greater Helsinki Apartments." *Rental Price Prediction on Greater Helsinki Apartments*, 2022, p. 51.
- Varma, Sarma, Doshi, Nair. "House Price Prediction Using Machine Learning and Neural Networks." *Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018, DOI: 10.1109/ICICCT.2018.8473231.