Predicting House Prices Using Machine Learning.

# Test Heading

## **Test Heading 2**

### Test Heading

This Settings is fixed.

Test Normal

# Heading 1

## Heading 2

### Heading 3

**House Market Prices Predicting with Machine Learning**

Name: Talamarla Ethender Reddy

Supervised by: Helen Xiang.

Student ID: 19024775

Course: MSc Data Science and Analytics with Sandwich Placement.

Title: Final Project Report (FPR)

# Abstract

Everything in the world will go through the life cycle. Likewise, humans have their own life cycle as humans in childhood aimed to study hard and grow. Once every human reach adulthood wants to settle down. For this reason, it primarily focuses on houses, farming land, and others. Everyone has their own needs and requirements for their house, like location, bedrooms, and more. This necessity creates a real-estate world. Real estate is crucial not for single individuals but also corporates and the government as well.

Importantly, seeing this demand, there are massive spikes in the prices of the property. It relies on several factors for priciness. Sometimes due to greediness of the realtor can also cause these spikes. This evidence will find in "Chapter 2 analysis on data [Ref 2]". To control these spikes is extremely hard, but it is possible. Some strict regulations and acts should need to be brought. However, currently, End-user who wants to buy a home is losing so much. To help the end-user, this project was implemented. Through this project, a buyer was aware of the situation and price factors to negotiate better pricing.

In mathematical, anything to predict the need to have earlier data. For this project, data is based on Indian house prices and collected from Kaggle. Kaggle was the world's largest data science community owned by Google. Kaggle is a platform for data enthusiasts and data seekers will come together. Data seekers will publish their data on licensed and data enthusiasts will reveal their insights of data. For this project, data was collected from here [Ref data link]. This data was licensed under GPL (General Public License) v2[Ref data link]. Data/software can be free to use and distribute on this license without the owner's concern. More details have been discussed in "Chapter 3 under ethical issues".

Machine Learning is one technical advancement in human history. It is a mix of many studies. Machine learning has proven insightful in many business problems like insurance, banking sector, mobile and more. Here in this project while machine learning can also prove significantly effect on suggestions on property. Several fields like computer science, statistics, and mathematics are used and implemented to complete this project; see chapters 3 and 4 for details [ref 3 and 4].

After completing the detailed statistical analysis, researching a wide range of machine learning techniques, choosing the best possible techniques, and implementing a website, the end-user (Buyer) will input features like location, approvals, bedrooms, and more. Based on those, try to predict and visualise the high, and low prices can negotiate with the realtor. Proper tools are discussed in chapter 3, and Implementation results are displayed in chapter 4.

Table of Contents

# Chapter 1 - Introduction

## Introduction

In history, humankind has evolved many things. Once, humans lived from caves to currently living luxurious houses. Many things have changed over time. Throughout the time's humans begin living in a society. These necessities made huge demand on real-estate over the period.

Regularly people in their adulthood will tend to have their own house. Before buying the house, the buyer will check several factors within their budget. Rarely be able to find it.

Real estate was not only permitted houses. It also applies to business, farming and constructing sites as well. So, every need is interlinked with real estate.

## Current Issues

As stock market likewise real-estate is not consistent on price. Several factors are responsible for prices. Sometimes factors are discussed here. For any property, some fundamental factors will depend upon the price's factors.

### Urban/ Rural Areas

Location plays a significant role in property. Suppose the property is in urban more demand than the rural. Urban places generally will come from several places, and more industries situated near them. In urban more features will be present than the rural. Sometimes the opposite will also be actual rural areas will also more demand, for example, businesspeople and working people unsatisfied with the urban and working lifestyle. They want to find lonesome places if rural areas have near mountains, riversides or any nature scenarios that will bring more demand.

### Facilities

Another thing about the property is the facilities it provides. People live socially together. For any house before buying, it should have some features. Is water scarcity being not available? In India, some areas have water has scarce. Is this property being not one of the locations? Transportation is another primary factor need to look at and to education institutions nearby.

### Unnecessary spikes

Sometimes a property has huge price demand. Even regular features and situated on regular location amount will sky-rocket. What makes this difference. Frequently realtors will increase these spikes (proved in chapter 2 basic analysis on data). Owners assume huge price has because of some feelings about it.

## Project Specification

### About

Here project was developing for end-users to learn how much they can predict the price quotation using existing features. Through this project, end-users can understand nearby neighbourhoods and bidding prices and how much value this possesses based on the features that the house stands for. Expectations of this project user will project features does this house holds. Then using existing data and predictive modelling, the end-user will get a low price (the amount used can start bid from). High price (Maximum amount can be paid) exceeding this price means paying over price than was.

### Features need to specify

Every property is unique and obtains distinctive features, but commonly every property needs to have some features and based on that, the price will be dependent too.

|  |  |
| --- | --- |
| Factor | Description |
| Under Construction | Does this property on under construction? |
| RERA | Does this property government approval? |
| BHK No | How many bedrooms it has? |
| BHK OR RK | This property has bedroom or only one room and kitchen. |
| Square Feet | How much space it has? |
| Ready To Move | Does this property can move immediately? |
| Resale | Does this property have any resale value? |
| Location | Where property located on? |

Above mentioned features are necessary to predict the price. If the property is under construction means it has less price than ready to move. RERA approval is crucial for a property without these government facilities (like electricity, water, and home loans will not approve). This RERA approval factor is very crucial for the Resale feature. BHK or RK feature is necessary to estimate price.

Moreover, if BHK home and RK house near the same area do not have the same price, prices can fluctuate high and low based on Bedrooms and square feet factors if the property does not have resale value (means cannot sell this property of above or same value that currently have). People will think twice to buy.

### Study fields implemented

For project success along with the data must implement mathematical methods, computer science, and statistical without these studies, this project will never succeed.

Even simple data would be nothing if mathematics did not apply in it. To gain some knowledge about the data or transform mathematics needed. In this project, mathematics is core and fundamental. Much mathematical analysis has been researched and used that are discussed in chapter 2 research.

Another important field is computer science. Full implementation and prediction would not be possible if computer science is available. Computer evolution is technical miraculous for human history. It revolutionised technology. It transformed the human era. This project used program languages (Python, Html, CSS, JavaScript) and frameworks (NumPy, Pandas, matplotlib, sci-kit learn, TensorFlow, and more). Above mentioned languages and frameworks, everything has its essence. Those are mentioned in detail in chapter 4 implementation and appendices.

Statistics is another field that was used in this project. It is crucial to analyse and predict the values; without this implementation in this project, it will not be complete. Project results will never be able to achieve. Several implementations and visualisations were used (like the relationship between each feature, predicting high and low prices using t-distribution and more). Statistical methods and implementation are discussed in chapter 2 and chapter 4.

### Final model for end-user

It was finally binding all the fields and data to achieve the results for the project. For end-user usage, a website was constructed to predict the values with visualisations. This was discussed in detail in chapter 4 implementation.

## Aims and Objectives

Aim of thesis what is the difference between other ideas in this project and how the data stands for. This project's main aim was to combine all data, analyse and efficiently, smoothly understand by end-user.

### Aims

* Same thesis or similar our idea has implemented any over.
* Some background analyses on data. To know about the relationship between features
* Why price factor spike
* Does even prices will spike in rural not much crowdsourced
* In urban or rural how many bedrooms people will prefer
* If mediators rather than owners will influence any price of the property
* Which state/city is showing more interest in construction?
* Which state/city will spend more on real estate?

### Objectives

### 

* Finding the best regression technique.
* Collecting the best predicting regression techniques combining their results will give the best predictions.
* How the users can use these implementations
* What tools and precautions need to take care of for running smoothly?

## Feasibility

For any project success, certain supporting stages are needed to accomplish.

### 1.5.1 Pre-processing

Once data is collected, specific procedures to perform for further processing. Pre-processing needs to check the data whether the data have any missing values. Suppose anything is missing is needed to fill the gaps. If any data is missing, it is crucial to take proper action, such as replacing average values or removing such fields based on much weightage they will bring to results. Another important thing is bringing data to the correct format (here, pandas data frame); otherwise, implementing mathematical solutions would not be possible.

### Analysis

Another crucial step is analysis. Here, once pre-processed data analysis should be done. How each feature relates to each other and how much they correlate with each other. The dimensionality reduction technique can remove negative impact and unnecessary features that do not support the final output. In this step, once the analysis is completed, able to know in which machine learning or neural networks will be used.

### Transformation

Before going ahead to the machine learning techniques, it is crucial that data needed to be transformed. In machine learning techniques, expect numbers. Categorical values or text values do not fit for machine learning approach. Though it needs to transform, and some other system expects only scaled data. To prepare this stage is extremely useful.

### Machine Learning

Several machine learning approaches are available; no rule was made for the particular use case—researching main techniques and evaluating performances once the best model was selected. To wrap up the model using some statistical methods. Now this project can move forward final stage.

### Neural Networks

Alternatively, neural networks are another approach for machine learning. Neural networks have gained the most popularity recently. Even complex problems are solved using neural networks (like face recognition, DNA matching problems, self-driving cars, and more). Neural networks replicate the human brain functionality. Fitting our project thesis can bring the best possible solution.

### End-User Application

Finally, combine all methods and data. Now end-user can use these predictions to display the predictions. The website was built to interact with the end-user. Application is convenient for a user to visualise predicting prices.

### Out-of-scope

Due to time factors, not highlighting similar features of houses in the same areas and prices they bid for.

# Chapter 2 - Research

## 2.1 Background research similar ideas

Similar like this project have been implemented if it is there any difference between this project will be able to know. After lot research surprisingly same idea of this project have not implemented yet. Some similarities have found. Those are discussed are here.

There are a lot of suggestions websites and blogs are present to guide the user what needs to be done and what need to care about. However, those are allowed to guidelines but not dwelling research on those areas. Some critical analysis on those guidelines.

## 2.2 Suggestions suggested by similar websites and assessment

### 2.2.1 Local Experts

Local experts will guide valuable information on locally and pros and cons on the location. But evaluating the price based on features and location may or may not be precise enough. Even though the taking advise from locals need to see current and earlier sold houses.

### 2.2.2 Search online property sites

Searching online is one of the useful. This will be able to find the house that are sold and prices are currently purposing. There are a lot of useful websites are present currently like “Magic bricks, housing.com, 99 -acres”.

These websites allow owners, builders, & realtors can post the property for rental, lease, selling and buying purpose. Here buyer and seller can visit the website, contact, and negotiate the prices. If mutual agreement has happened deal will be completed. These websites act as a platform for properties buying. These websites take a little commission for bring users to posted owners.

However, these platforms will never suggest the buyer features based prices. It can only show the listings of properties. It will never show earlier histories of selling houses.

### 2.2.3 Online price calculators

Price calculators will help only to calculate the construction estimation based on square feet and house how buyer wants to build. This tool can also very much helpful to buyer to estimate and how much the construction materials will be used to build.

Nevertheless, here user need to have a property and thinking about built. But these will useless when buyer wants to buy constructed property or not interested to construct at once. Moreover, these tools will never help to negotiate the prices and find the situations, features and prices.

## 2.3 Research on languages and frameworks

### 2.3.1 Research on language

Several languages are present in computer science. Particularly choosing Python is complex, and several considerations have been taken care of. Python was built by “Guido van Rossum in the late 1980”. This language is relatively easy to learn and easy to adopt small- and large-scale apps. Even though it is object-oriented programming, it does not need to follow all the guidelines like Java and C. New learner will quickly adapt to Python, unlike the other languages.

Even though Python is open source, it has strong community support. This builds individual developers' good confidence in making programs on contradictory like other languages python wide range of applications. Almost it covers every application. Most importantly, python language is widely known for machine learning and neural networks; using extensive python libraries, it supports vision, text, databases, and more readily available frameworks and is easy to implement.

Suppose Python did not support some underlying technology. Python can inherit another programming language to itself. For example, the “Jython” framework can work with java libraries in Python. Similarly, “Cython” use c libraries in Python. Python most dynamic and extensible programming language.

### 2.3.2 Research on machine learning and visualisations

If any machine tries to work without an explicit program is called machine learning. Even though the machine learning term was present in 1959, but it has got popular recently. The 21st century was a data era because technology has more advanced than any decade, and much data is present. A data is nothing when did not understand. So, machine learning was brought into the picture it also artificial intelligence. Using some advanced mathematical methods and data machines will be able to predict the future.

Machine learning has solved many business problems like medicinal fields (How much drug will be effective), Email filtering (Lot of spam emails will be filtered), Speech and text recognition (Lot of IoT devices were built and auto-text prediction in mails) and more.

**Pandas**

Pandas were built on top of NumPy. This framework was flexible and expressive data structures. Through this, a multi-dimensional array can easily manipulate, sort, and index. Complex data manipulation is crucial for any application. Because for mathematically achieving a goal, everyone knows several implementations and manipulations need to be done on inputs. If this does not make fast and reliable, then predictions would not be possible at any time.

**Scikit-learn**

Scikit-learn is one of the most important libraries for machine learning. Machine learning is a mix of various study fields. It possesses various techniques and methods to work with. So, sci-kit is an open-source library and implemented most of the machine-learning techniques and methods. With this library can be able to perform data transformations, predictive model building and more.

**Matplotlib and Seaborn**

This library is for visualisations. A picture reveals more than a thousand words. In the statistical world, charts will talk more about the data. In this project, all analysis is covered with these libraries.

### 2.3.3 Research on Neural Networks

After decades of experimentation and workings of how the brain works, many practical applications have been made in that neural networks are one. Neural networks will mimic brain functionality in computer science. This solution also helps to predict the same as machine learning. Neural networks have solved many businesses use cases like self-driving cars, image recognition, voice recognition, and more. In this project, deep neural networks will be significantly fit for predictions.

**TensorFlow**

TensorFlow is a free and open-source library. It supplies a range of solutions and almost several types of neural networks (Deep Neural Networks, Recurrent Neural Networks, Convolutional Neural Networks and more). This library can also supply GPU support; this is very crucial for complex computational power.

### 2.3.4 Research on web applications

The total aim was that end-user should be using the application smoothly and efficiently. Nowadays, to reach an end-user, there are two ways. The first proposal is to design a mobile or system application. Choosing the first choice will require much time and different architecture. One design is not sufficient for all Operating systems. Each operating system has its language to fulfil the application. The second proposal was designing a web application that is the simplest and common application across all channels. Almost every operating system has a web browser in it.

**Web frameworks and web languages**

**Frontend Development**

Every web application has 2 phases, frontend which users see, feel, and use and the backend connects the main logic with UI workings integrating databases and workings of internal logic. For this project, frontend HTML5, CSS, and JavaScript were used.

HML5 was markup language uses for structuring the website. This language is fundamental for a web application. CSS uses for styling the application. User experience matters regardless of how much application is practical. To smooth and efficient workings, JavaScript and CSS were used.

**Backend Development**

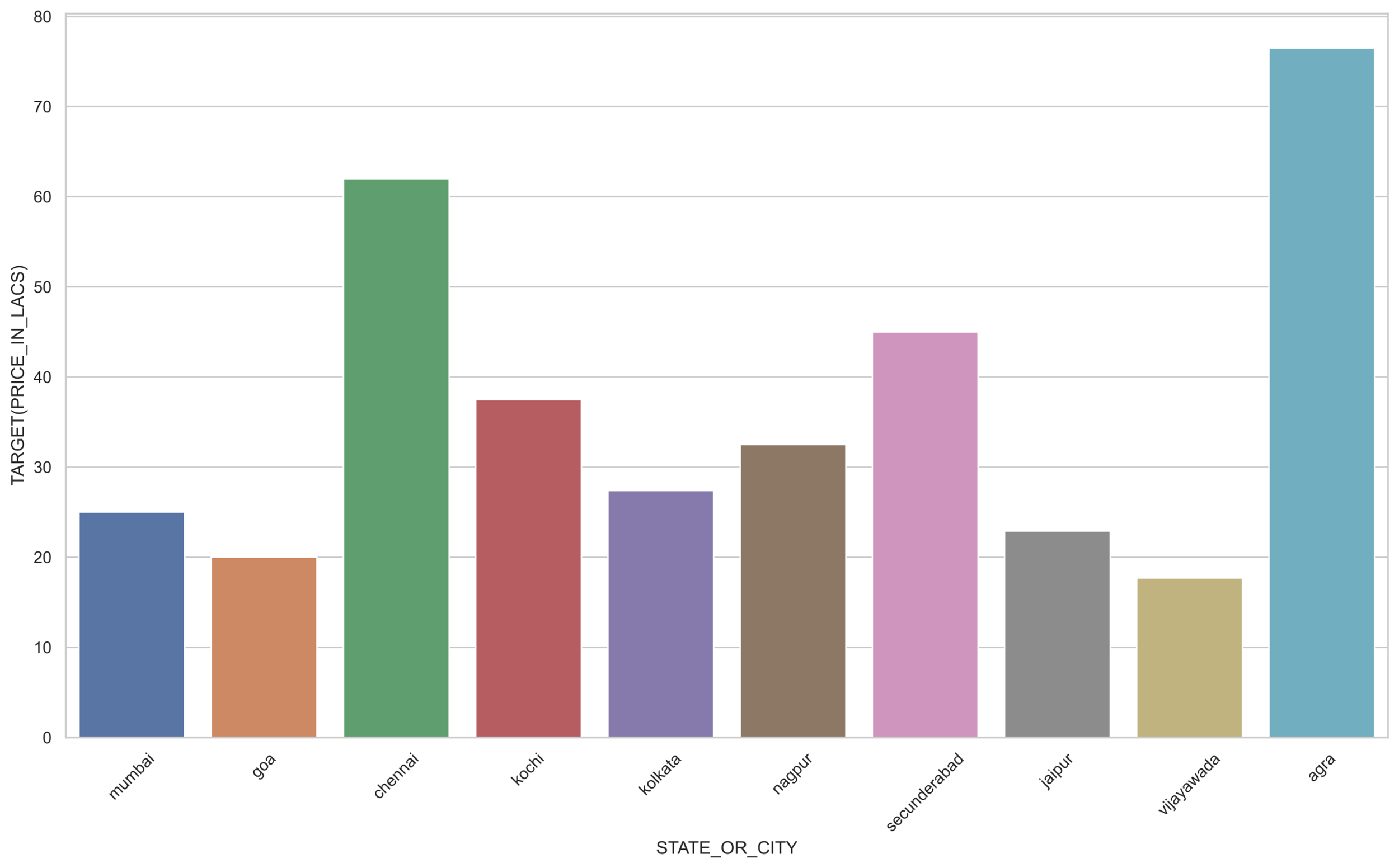
Even though how much pretty and usage in frontend without connecting backend. Web applications were useless. Backend will stay like a backbone for a web application. For this project, Django was a perfect choice.

Django is a web development framework developed using python. It is easy, reliant and trusted by many other web developers, and it also has strong community support.  2.4 Background research on house data

## 2.4 Background research on house data

### 2.4.1 Which cities will spend a high average on housing

To Identify the above question. Two main tasks are needed. Identifying main cities in India and average spends on the housing market.

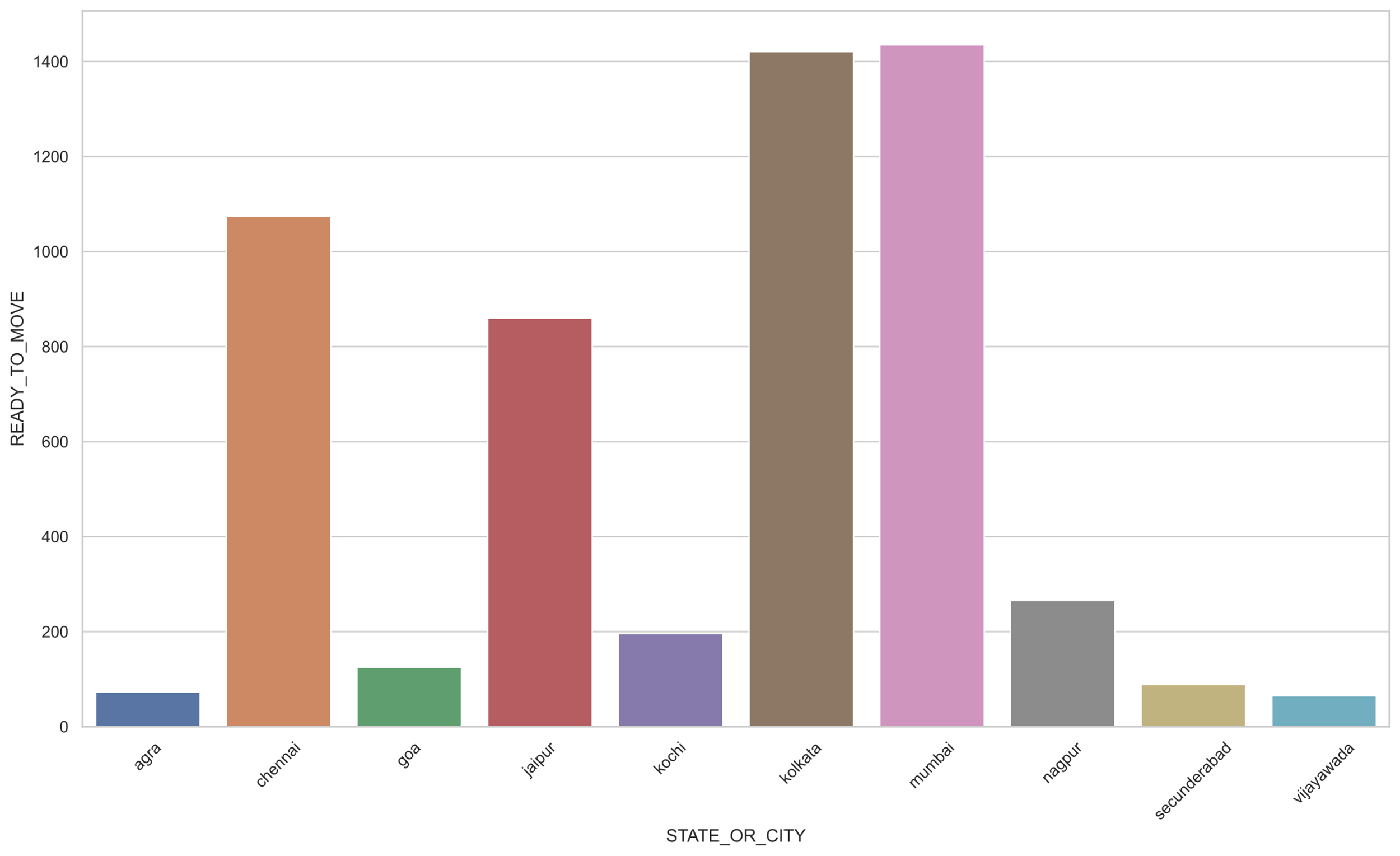


Ref: Please refer appendices (1)

India is the seventh-largest country in the world, Rather than Identifying the most spending housing areas. Another way is identifying main cities average spending will reveal which city spending most on real estate. [[appendix 7.a](#_7.a._Identifying_top)]

### 2.4.2 Which city was constructing real estate more rapidly

Several factors needed to develop urban areas. Real estate plays a significant role in that. People across different parts of India will come to cities, finding for work & live here.

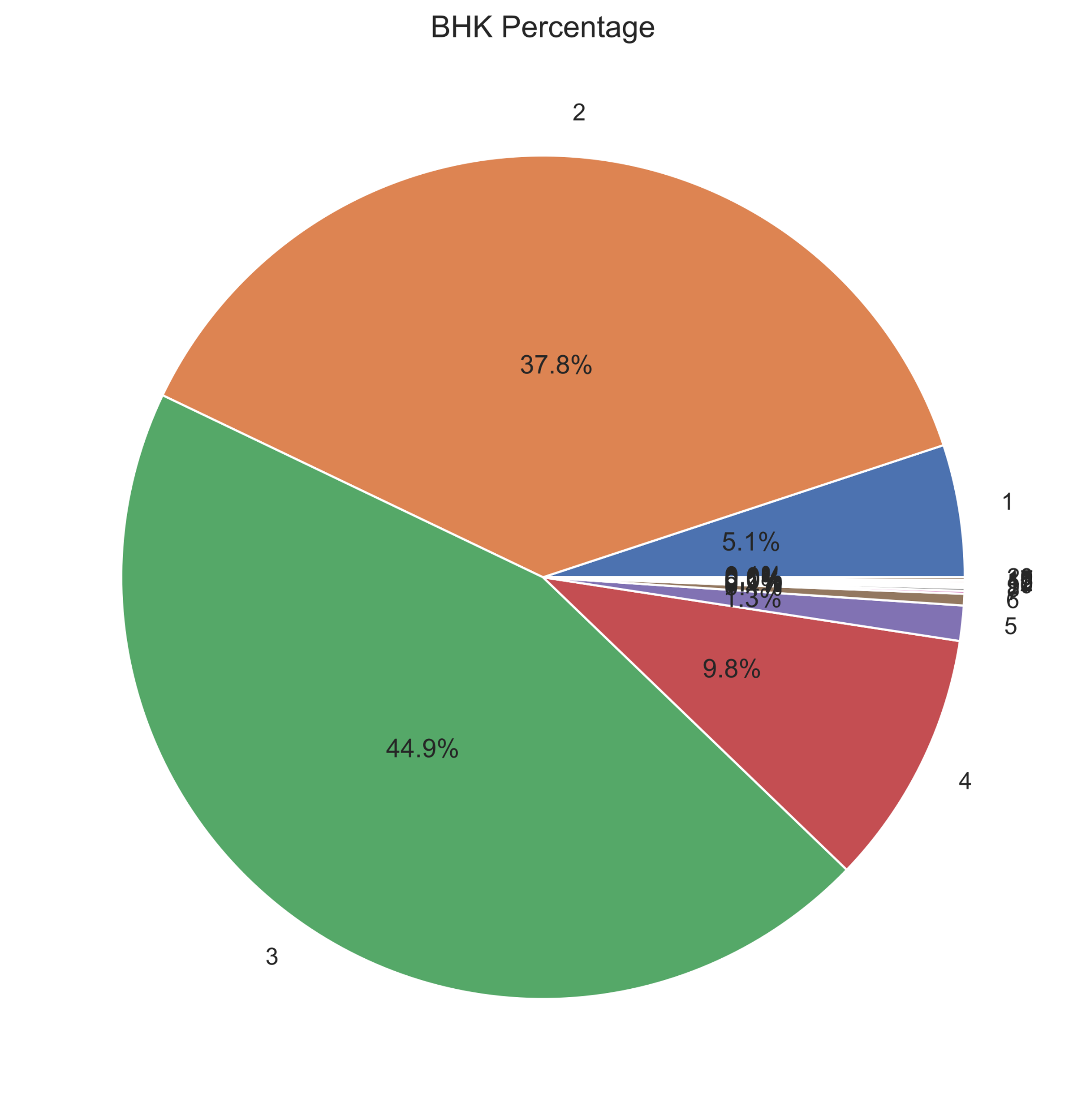


Ref: Please refer appendices (2)

Mumbai, Kolkata, Chennai, & Jaipur was rapidly developing on real estate. Mumbai, Kolkata, Chennai are coastal areas and most developed crowdsourced areas as well. So that factors have influencing real estate development. The Prime minister has announced a high-speed railway and Delhi to Mumbai express highway connecting through Jaipur. These factors have influenced real estate to grow in Jaipur. [[appendix 7.b](#_7.b._Which_City/State)]

### 2.4.3 How many bedrooms most people prefer

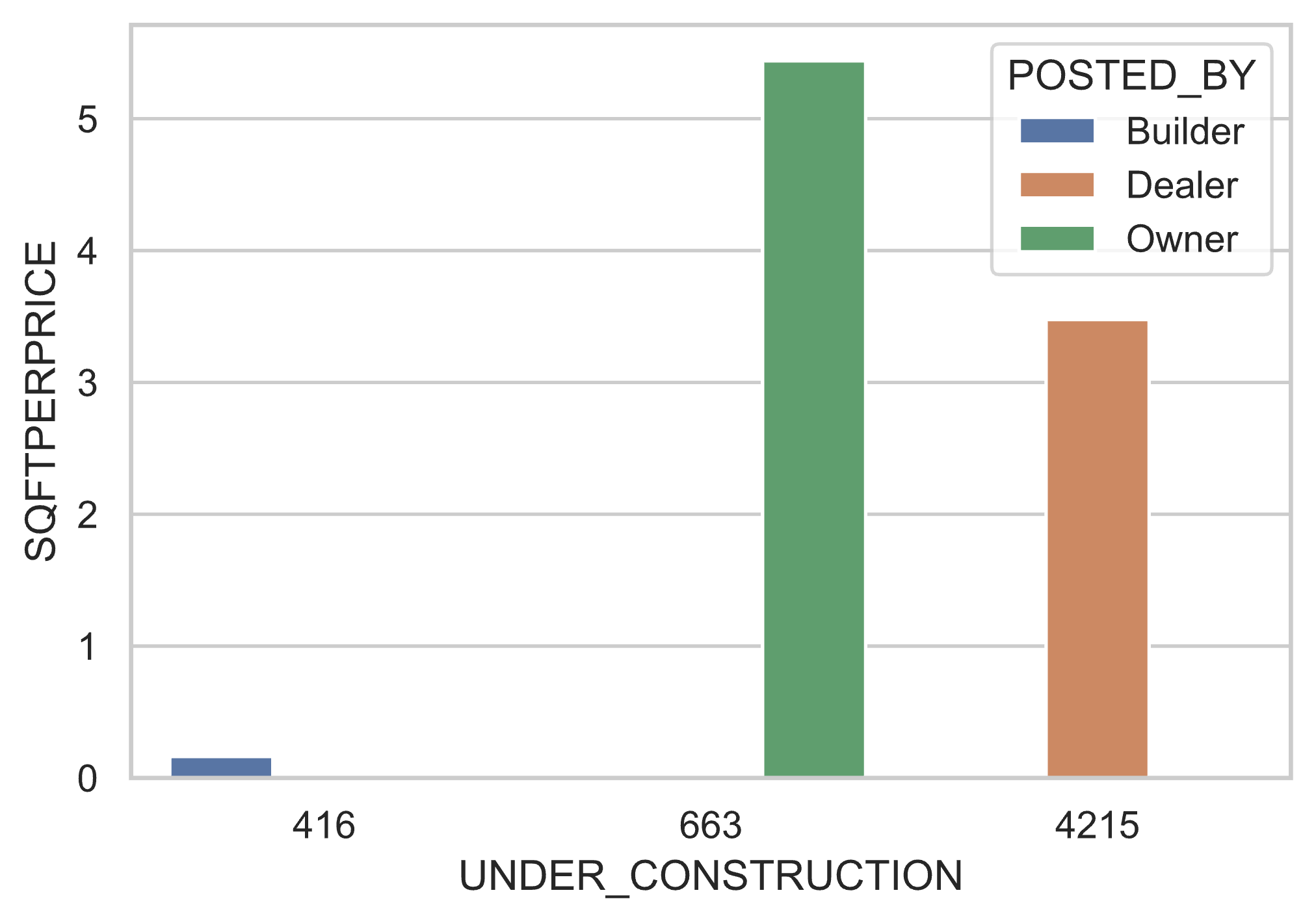
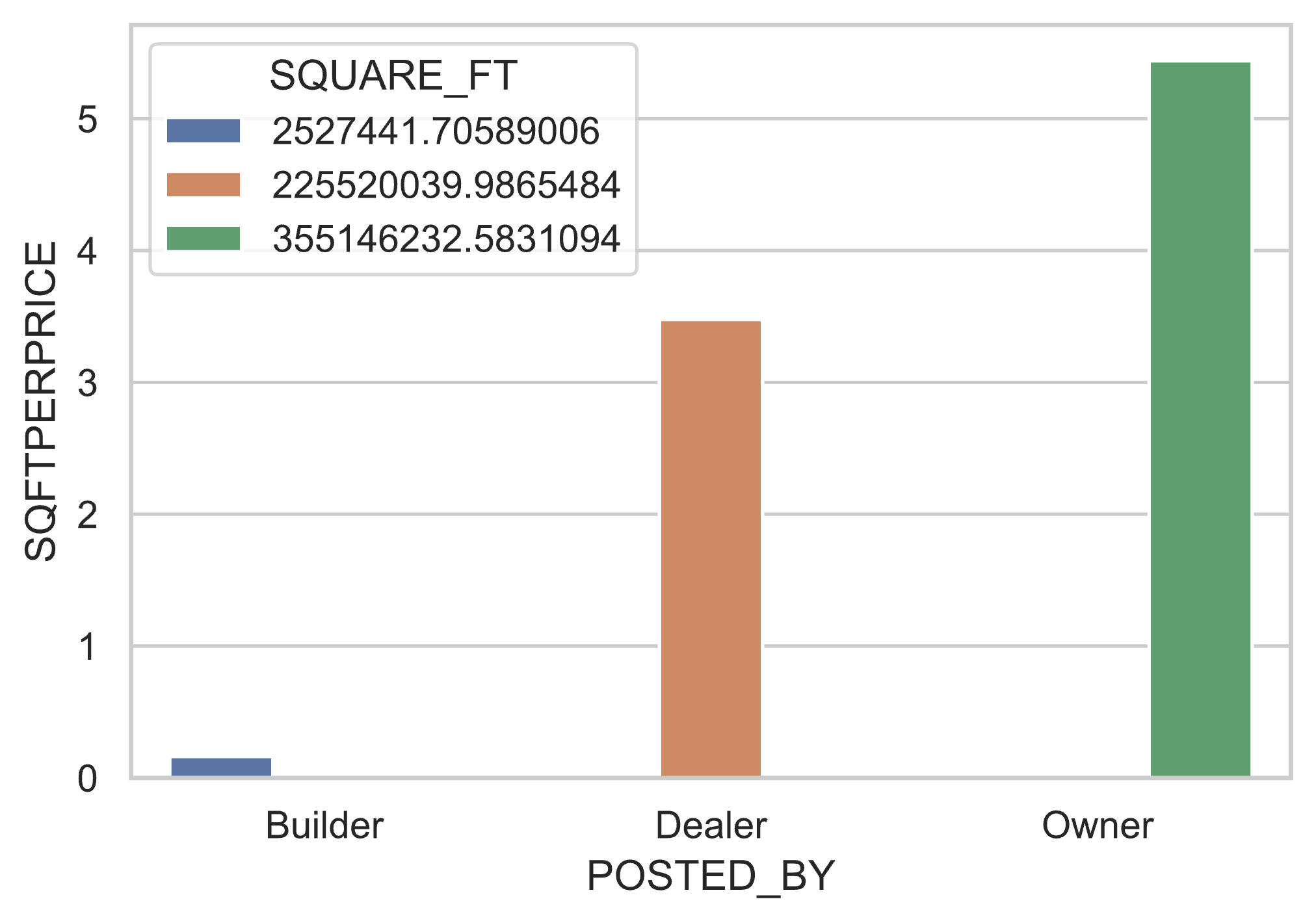
Everyone will think before buying a home. How spacious it is and how many bedrooms are there, and many features are cross-checked before buying.



Ref: Please refer appendices (3)

Applying analysis on BHK Distribution percentage. Three bedrooms have 44.9%, two bedrooms have 37.8%, Four bedrooms has 9.8%, and One bedroom 5.1%. Most Middle-income person can mostly afford 2 or 3 bedrooms. So, most housing categories mainly focus on two or three bedrooms. Upper and upper-middle-income groups can afford four and above. [[appendix 7.c](#_7.c._Which_bedrooms)]

### 2.4.4 Did who posted will influence the price



Considering the data, the Owner has posted more amount than the dealer or builder. On contradictory Owner has posted more SQFT than the builder or dealer. This concludes that the Owner owns more land and sells, coming to dealer/builder, dividing the land, building more houses, and selling at the profit margin. [[appendix 7.d](#_7.d._Did_price)]

## Chapter 3 – Methodology

## 3.1 Factors responsible for the house price

House prices are not independent it is dependent on several factors. Some factors are discussed below.

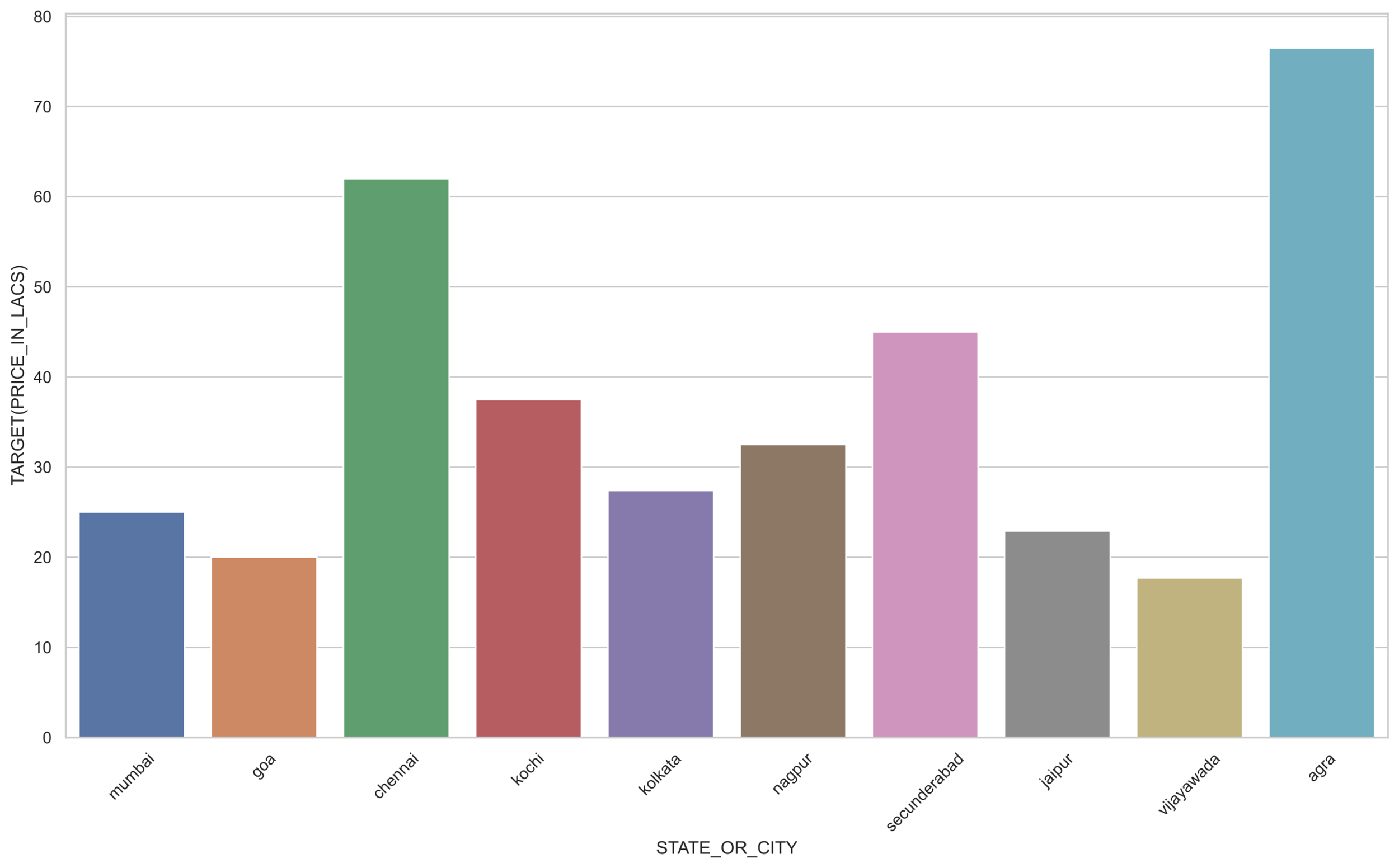
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Economic Change** | **Supply Demand** | **Interest Rates** | **Location** | **Laws** |
| Country economic status | More demand and Supply | Banks offers lower interest rates | Schools | Government laws that offers to buy/build houses. |
| Hospitals |
| Currency trends | Transportation | Tax exemptions |
| Neighbourhood |

In India, people will trust that investing in land is the safest investment overall the period. Even inflation/deflation, people believed betting their savings on real estate will be more beneficial than other investing options. If year by year, investing in real estate steady increase across the country. Even on fundamental analysis on cities property sales were high. Ref [1]

Chart, line chart

Description automatically generated

Ref1



Ref: Please refer appendices (1)

What factors people thinking investing in properties will better option. Due to the continuous increase in prices every year, Mortgage will give regular incomes; government tax exemptions have imposed property and building-related (Section 80C, Section 24, Section 80EE https://www.bajajfinserv.in/tax-benefits-on-home-loan). Even regular income people will get benefits under these laws. Additionally, state governments even pay half of the payment for constructing (under scheme Awas yojana).

Buying individual property locations was mainly considered. Property must be reachable to hospitals, workgroups, good neighbourhoods, schools, transportation. The above factors were crucial for day-to-day life for a living. This factor makes central urban properties make demand and more sharp prices.

Sometimes the government will break the previously imposed taxes to encourage investment in real estate. Like on covid times, the central government eases RERA regulations fees. Because in India government impose 18% tax on the total price of the house. Due to high taxes and non-movement of the country economically, these conditions make the real estate nearly broke for balancing this government has exempted for some time. This boost real estate almost immediately.

## 3.2 Choice of methods

The predictive model can be developed by using data mining techniques. Data mining techniques are helpful to understand feature importance and analyzing the data. It is crucial for investigating rightful machine learning and neural networks techniques is vital for better predictions. There are several machine learning and hyper-parameters are available. Each has its unique possibilities to predict even in complex data patterns. It is very crucial for finding the correct technique and parameters for a better solution. Even in high scaling of traffic should be able to predict concisely. In machine learning and neural networks, there are several papers, articles and journals are published. The main aim of this thesis was careful to understand the algorithm and correctly fit in the problem.

### 3.2.1 Feature Importance

In data, all features do not correspond to dependent factors even it impacts negatively. "Irrelevant or partially relevant features can negatively impact model performance (Eleanor Roosevelt,2018)".

On implementing the feature selection on the house, data will get to know which feature supports predictions and which features do not help. For selecting features, we use the "sklearn SelectKBest" algorithm. This algorithm internally uses the chi-square test to find the feature importance.

Formula



c = degree of freedom

O = Observed value(s)

E = Expected value(s)

Degree of freedom refers to independent factor’s maximum vary overall. Expected refers to an average of values.

### 3.2.2 Linear Regression

Linear regression is primarily applied for predictive analysis. It is the most straightforward regression technique and very basic. Several advanced regression techniques are available, but all are developed on top of linear regression.



 = predicted value

 = weights

X = independent features

Linear regression will predict the output based on adding little weights to independent features and scales error using distances (using Euclidean distances or similar.) and correct it regularly (Uses Gradient Descent) to find an optimal weight for predicting.

House prediction data is more complex than usual it needs more sophisticated techniques to fit well. Some techniques are seen and discussed in great detail and how well fit in our problem.

House Predictions

Step 0:







Residuals



If all errors are more than convergence level. Then using gradient descent algorithm optimise the weights.



Repeat until Convergence level



**Justification of use**

Using this basic regression, able to know how well it will fit in the regression technique. How much it can improve when tuning hyper-parameters.

### 3.2.3 Logistic Regression

This regression can also be called a binary classifier. It usually compares results and searches which class it belongs to. If the result is greater than 50%, it concludes the result. That is why the price should not be float. It should be a natural number. This regression can impose a penalty for weights; this dramatically and quickly finds optimal weights.

Formula





**Chart, histogram

Description automatically generated**

Ref: by hands-on scikit-learn

**Justification of use**

Implementing in a project will be able to know. Applying “L1, L2, and elastic-net” penalty to the learning rate will perform any better than linear regression.

### 3.2.4 Support Vector Regression

Support vector was developed at AT&T Bell laboratories. This technique will also work classification and regression tasks. Support vectors are well suited for complex datasets. Our data does not represent a straight line to predict the values. Above mentioned techniques try to draw a straight line on the optimal solution. In this technique, while applying kernel, can a draw a line to mostly perfectly separated. See the below picture of how it can represent the data.

Chart, scatter chart

Description automatically generated

Source: hands on sci-kit learn

Formula





 : training sample.

 : target value.

 : inner product plus intercept.

: free parameter that serves as threshold.



This technique can also be used to find outlier and outlier detections. Sometimes in our project, the property can be sold amount or more features on less price unusually. This type of residuals cannot be able to figure out with our model. Using this technique with kernel prediction can be sensitive or hard on future decisions.

How kernel can be able to separate data, and other techniques were unable to. The answer is when opting for kernel trick in support vector machine. Hyperplane will be the same dimension as features were for previous techniques. In SVM, the kernel multiplies the features and tries to divide the data with hyperplane into multi-dimensions; this can achieve more accurate results than the previous techniques. Even curve data can be able to split into groups.

**Justification of use**

Applying kernel to data will predict with fewer residuals than the other regression techniques. This technique can also be non-sensitive to outliers (which means unusual transaction prices on a property cannot change the model prediction rate).

### 3.2.5 Random Forest

In statistics, always population was better than the single and The law of large numbers declares that many trials results will lead to average number. Similarly, ensemble techniques also follow a similar theory. Here many regression techniques will combine and predict the value collectively.

Using the above technique, Random Forest was created by "TinKam Ho in 1995". Random forest was the ensemble technique of the Decision tree. Here random forest multiplies random subsets of features and increases features like tree branches. It will increase predictions. This regressor will also help find each feature importance.

Random Forest is a combination of bagging and pasting. It trains predictors with different random subsets of the training set. If sampling is performed with replacement is called Bagging, or without replacement is called pasting. Random forest is a combination of both. It will try both and bring the best suitable option.

In machine learning for regression or classification tasks, Support vector machines and Random Forest techniques are more powerful than any other. It works like magic and predicts accurately will not much change on further on new data. It will optimise well on changes.

Bagging: 

Pasting: 

Note: Bagging and Pasting formula is same only difference is pasting will replace training subsets but bagging doesn’t.

1. Sample with replacement, n training examples from X, Y; call these 
2. Train a Regression tree  on 

Chart, radar chart

Description automatically generated

Source: ref random forest Wikipedia

**Justification for use**

A random forest can improve prediction more on our project while multiplying more subset features and deriving hyper-planes on multiple dimensionalities.

### 3.2.6 Gradient Boosting Regressor

Gradient Boosting Regressor is another ensemble technique. This is similar technique of Random Forest only difference it will add tree one at a time and fit for correct predictions. This type model is called boosting in machine learning.

Gradient Boosting Regressor is best known for listed below.

* Tree constraints: depth of trees and count of the trees.
* Weighted updates: learning rate will limit each tree depending on how much it contributes for predictions.
* Random Sampling: It tries to random subsets of features as trees and try to predict the output.

Formula



*  : Expected value
* : Loss function
*  : actual output
* : predicted output

**Justification for use**

This model can be extremely useful in our project. Even though Random Forest is similar. Random forest will never be able to supply which subset of features will supporting the final output. By, implementing this model can prove beneficial to project.

### 3.2.7 Voting ensemble method

All regression is good in their way. Nevertheless, in this project include all possibilities to predict. The voting technique is another technique in ensemble methods. It will combine multiple regression or classification techniques and bring majority votes which purposed by most techniques.

**Justification of use**

To make solid predictions in all possibilities. This project will choose the best regression techniques to combine and predict the final output price and make a confidence interval.

### 3.3 Data Collection

Crucial analysis on this project is prior sold houses data. Complete this project; Further investigation performed finally found data on Kaggle. Kaggle is a data science community platform. Data was accurate and precise. Moreover, the data is open source. This data was licensed under GPLv2. For more details of this license, please refer to the Ethical and commercial context section.

### 3.4 Project Design

**Use case diagram**

The use case diagram illustrates what respective fields will need to provide by the end-user, and based on that, machine learning models will fulfil the house's predictions to the end-user.

Diagram

Description automatically generated

**System Diagram**

Here in system diagram illustrates two use case scenarios.

Diagram

Description automatically generated

**Use case scenario 1**

Earlier data was collected from sold houses. Then data were cleaned and processed and transferred to the voting regressor. A voting regressor is a collection of best-chosen models and predicts the price combining all models. One prediction was completed sent further for statistical analysis from their final visualisations is shown.

**Use case scenario 2:**

User fed the data sent for data cleaning from there to voting regressor. Here all models are pre-trained and ready to predict accurately. Once the voting model is predicted, sent statistical analysis for 95% confidence interval. Once results are collected are shown to the end-user with visualisations.

### 3.5 Validation

Any machine learning project is all about choosing the best model, not only prediction accuracy and efficiency. Suppose the model predicts the accuracy initially. When the model is deployed on production, it will fail catastrophically due to heavy variation of trained data. Efficiency is also crucial as prediction accuracy. For selecting the best model, the below steps are followed.

**Step 1**: Select a model and evaluate the score.

**Step 2**: If the model performed better, tune the hyperparameters. Every step of tuning must take care of the model should not underfitting or overfitting.

**Step 3**: If the model not performing better and efficiently, reject the model and repeat step 1.

**Step 4**: If the model is tuned and predicting better and efficiently, select the model.

**Step 5**: After selecting the best models, combine and predict the output.

**Step 6**:  Based on the predicted value, apply Student’s T distribution and make a 95% confidence interval.

**Step 7**: Show the results to the end-user.

### 3.6 Ethical Issues

**Legal**

Data was used in this project was licensed under GPLv2 (General Public License). Under this license, data/software is free to use, change, and distribute without any author concerns.

All tools and frameworks which are used in this project were open sources. No legal issues will ever arise on Institute or whoever using it.

**Dataset:**

Kaggle.com. 2021. House Price Prediction Challenge. [online] Available at: <https://www.kaggle.com/anmolkumar/house-price-prediction-challenge> [Accessed 1 July 2021].

**License:**

Gnu.org. 2021. GNU General Public License v2.0 - GNU Project - Free Software Foundation. [online] Available at: <http://www.gnu.org/licenses/old-licenses/gpl-2.0.en.html> [Accessed 1 July 2021].

**Social**

No persons were involved in this project, or any other practices were made to hurt the feelings or beliefs of persons. This project is intended to help people who are needy to analyse real estate.

**Ethical**

Strictly this project does not expose any gender or any other beliefs. This project aim was to help buy the best house in society according to the buyer budget.

Strictly this project does not expose/ hurt any gender, religion, or any other beliefs.

**Professional**

Under any circumstances supporting confidentiality, anonymity will not be uneased.

All issues and concerns were following under the guidelines of the university. No extra practices were made to disrupt the above issues.

### 3.7 Consideration of commercial and economic context

**Considering similar ideas**

The core idea of this project has never been implemented ever yet. This project was constructed keeping in mind other policies and never overruled. Nearly nowhere this project disobeyed until our notice. If any, please bring to our notice that respective action will be taken shortly.

**Considering business**

This project will respect real estate or any other personals. This project will not discriminate against others at any cost. This project was intended to help buyers in price negotiations but never intended to hurt others or aggravate others. If we do so, please bring to our notice that respective action will be taken shortly.

## Chapter 4 – Quality and Results

### 4.1 Code Editors

**Jupyter**

Jupyter notebook is a python codebook. Which means we can interpret the code and explain its functionality. It uses like a record book. This editor will be helpful for data science and others. In terms of data science, while researching and analysing, understanding each step is necessary. Jupyter notebook will be more reliable in that matter.

**PyCharm**

PyCharm's community editor is beneficial for developing an end-to-end model. JetBrains corporation was created, incredible code editors. Even google was distributing to developers for their android development. Most of the developers were recommended for development. There are two distributions available, but in this project, the community version will be used.

Mainly PyCharm will be used for developing a web application for our project. In this editor, not only python also languages like HTML, CSS, JavaScript is supported to develop. PyCharm is also supporting version controlling. This feature enables fewer bugs and reliable source management.

### 4.2 Others Important tools and techniques

This section discussed essential and distinct features, which helped this project build and cannot be combined with other areas.

**Virtual Environment**

In python there another critical feature is the virtual environment. Every project in python has used a distinct set of libraries and different requirements of versions. When combining all together, it will be problematic for projects. A virtual environment is simple and more reliant. Under this feature, this project was extracted and import. Any environment (operating system) and architecture can be used (like cloud, servers, and personal computers).

**Job Library**

Unlike other projects in python, the machine learning project is distinct. The first model was pre-trained in machine learning and made necessary changes; then, it is ready for deployment. Every time web application restarted or republishing machine-learning model cannot be trained. The “JobLib library” used once the machine learning model is pre-trained can be saved to the filesystem and loaded through this library.

In this project as well, several models are saved and used in the end-to-end model.

### 4.3 Data pre-processing

**Encoding**

Machine learning is a mathematical field. Even data should be numerical values only; otherwise, ml techniques will not accept. Generally, any format of data will transform numerical values and later be sent to machine learning techniques.

A picture containing text

Description automatically generated

Table

Description automatically generated

Here in this project, “POSTED\_BY and BHK\_OR\_RK”, features represent categorical values. These values need to transform numerical values.

Table

Description automatically generated

In machine learning, there is a possibility available called Label Encoder. This Label Encoder gives each categorical class a number. Through this, it will be able to transform object types into numerical values. Here in our project, “POSTED\_BY and BHK\_OR\_RK” will transform using Label Encoder. Using Label Encoder later can also use as reverse transformations mean transformed numerical values will transform to categorical classes.

**Splitting**

Even machine learning is a model is built. Some inquiries will arise.

* Even the model is built. How efficient is a model?
* How to test model prediction rate and how well it performs in production.

The above questions are basic and necessary for evaluating a model. For these, we need to have test data to evaluate a model. Using total on model training will not make the model efficient sometimes. It leads to overfitting and worsens on unseen attributes. To overcome data needs to split two categories. First, training data will be used to train the model. Second, test data will be used to evaluate the model.



Graphical user interface, text

Description automatically generated

Using the “sklearn from train\_test\_split” method, data is divided into train and test data. “test\_size” can specify the test data length. Every time running this method splitting data will be randomly done. Means one record can be a training set, or a test set cannot be predicted. If we want to divide consistently, “random\_state” needs to be set.

### 4.4 Machine Learning Implementation

**Linear Regression**

Linear regression (As discussed in chapter 3) was implemented on house prices predictions. Although linear regression was straightforward, it will not perform better on complex and distinct datasets. It also not performs better on outliers. So linear regression expects scaling on data.

**Scaling**

Data can be more varied; this can affect predictions. So, data needed to be scaled. Frequently two types of scaling are using in machine learning. First Standard-Scaler this method transforms data between 0 and 1. Second, the MinMaxScaler method will transform the data between –1 to 1. In this project, standard scaler and most suitable for the house price dataset.

Standardization



Mean



Standard Deviation



**Implementation**Graphical user interface, text, application, email

Description automatically generated

After scaling, the data is consigned to linear regression. The above code shows that the Linear regression method was imported from the sklearn library after creating a linear regression object without any hyper-parameters. The model was and evaluated with test data. So, it performed worse than ever. Linear regression model able to predict 9% accurate total data of the time.

So, what steps can be taken to improve models? As discussed in an earlier section, validation (3.52) step tuning hyper-parameters can show improvements.

**Text

Description automatically generated**

**Hyper-Parameters Finetune**

As shown above, “GridsearchCv” is one of the best techniques for hyper-parameter tuning; as seen in the code, what parameters need to be tuned must be specified in list format. “GridSearchCV” method to work expects a model on which parameters need to be tuned. Then “GridSearchCV” will train the model using the “K-Fold cross-validation method” one parameter sequence at a time. Finally, after cross confirming, the results of a combination of parameters will be given. The final model will be build based on the “GridSearchCV” suggested parameters.

**Model Result**

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Even though hyper-parameters fine-tuned results were not satisfying.

**What can be done**

Even linear regression is failed after hyper-parameters tuning. However, it can be improved. So, multiplying the data attributes as features can be improved the linear regression; this method is called polynomial features. Multiplying features will create multi-dimensions then it is easy to divide hyper-planes to predict. (Appendices no)

**Support Vector Machine**

Support vector regression (discussed in chapter 3) is implemented on the house prices dataset. Support vector regression expects data not to be scaled. It makes soft margins on hyperplanes.

Graphical user interface, text, application, email

Description automatically generated

At the first implementation of support vector regression, the predicted scores were not satisfied. Before coming to conclusions, support vector regression is widely known for its kernels. Here above code was implemented without any kernel. As per the validation section (chapter 3), tuning hyper-parameters can be improved the model efficiency.

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text

Description automatically generated

The above code shows significant improvements to the model. The House prices dataset will be fitting incredibly on kernel linear. So, this method can be acceptable to the final end-to-end model. Upon carefully seeing these scores, it looks like the model is overfitting means too sensitive to unknown values. So, in this case, controlling “Tol and C” values will eliminate the overfitting of the model.

Graphical user interface, text, application

Description automatically generated

The above code was to find the best “Tol and C” value, not to overfit the model. Implementation of the above code adjusting “Tol and C” will not significantly affect the model. So, it looks like the model is showing more variance and less bias. To conclude this, finally controlling degrees of freedom will be helpful.

Graphical user interface, text

Description automatically generated

**Results**

On looking at the code controlling degrees of freedom also will not help the model not to overfit. So, this model cannot accept for the final end-to-end model.

**What can be done**

Even though the support vector machine is overfitting, there are two ways to overcome it. First, supplying more data to the model can be helpful. Secondly, using a more complex model will be helpful for better prediction and efficiency.

**Decision Tree Regression**

Decision trees are known for simplicity and efficiency. As the name suggests, this algorithm works like trees here. Leaves are features, and branches are conjunction of features; it can predict smartly through this.

Graphical user interface, text, application, email

Description automatically generated

**Implementation**

Basic implementation of the technique supplies satisfactory results, most probably acceptable to the end-to-end model. The first model has shown 88% correct predictions of the time. However, the model can be more perfected. Decision trees can specify hyper-parameter “Depth”, which means to predict how much deep the branches can be allowed to travel. Another hyper-parameter is “max-features”, how many max features combination the predicted value will be.

**Hyper-parameter fine tuning**

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

On performing hyper-parameters tuning on “Depth and Max-feature”. Shows performance improvements and without compromising model efficiency. The results show “Depth is 2” and “Max-Feature is auto”. The model performs best on the housing dataset. Alternatively, showing “max-feature is log2” and “depth is 2 to 6” has the same results as above.

**Result**

This model performs effectively on the house price datasets and is non-sensitive to unknown records. This model can be used as the end-to-end model.

**What can be done**

Currently, the model is working at best and changing anything can lead to overfitting the model.

**Random Forest Regression**

Random forests are already specified in the earlier chapter (Chapter 3). Now here, implementing on house price data.

**Implementation**

Graphical user interface, text

Description automatically generated with medium confidence

Here implementing the random forest performs elegantly on the housing dataset. It shows a total of 94% accurately without losing efficiency. Some hyper-parameters are concluded at the basis of model initiation. “n-estimators” are before predicting the final output estimate various times and give best possibilities. “max\_leaf\_nodes” while predicting take maximum specified leaf nodes into accountability. Furthermore, “n\_jobs” is commonly used in the sklearn library to work the model using all cores of processors for faster processing.

**Hyper-parameter fine tuning**

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

Random forest performs better at the first model itself. However, hyperparameter tuning reveals the efficiency and peak performance, even though the hyperparameters are the same as decision tree regression. Bagging and Pasting differences will any impact the model.

As shown, results implementing these hyper-parameters on random forests negatively affecting on predictions and efficiency.

**Results**

Concluding, not using any added hyper-parameters will make the model performs better than ever.

**Voting Regression**

Voting Regression is one of the ensemble methods. Ensemble methods is already discussed in earlier chapter (Chapter 3). Unlike others regression methods voting regression will not do anything. But in this project, it plays very crucial role. Voting regression will predicts the final output combining multiple regressors. Voting regression will work in theory of “law of large numbers”.

**Implementation**

Graphical user interface, text, application, email

Description automatically generated

Here voting regressor has initialised and combination of regressions methods. Above four regression methods were finalised. As shown in the picture which regression is initialised and what parameters mentioned.

Voting regression reveals compelling results and this prove averagely 94% accurate of all the time and more stable on unseen data.

References :

* <https://www.globalpropertyguide.com/Asia/India/Price-History>
* Feature selection: <https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e>
* Feature importance : <https://www-sciencedirect-com.ezproxy.herts.ac.uk/science/article/pii/S1532046418301400?via%3Dihub>
* Features importance : <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-213>
* Feature Importance algorithm: select k best use chiquare selection feature: <https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223>
* Support Vector Machine: <https://shuzhanfan.github.io/2018/05/understanding-mathematics-behind-support-vector-machines/>
* Support Vector Machine : <https://en.wikipedia.org/wiki/Support-vector_machine>
* Random Forest : <https://en.wikipedia.org/wiki/Random_forest>
* Research : https://www.houselogic.com/sell/how-to-sell-step-by-step/home-market-analysis/
* Research Tool: <https://www.bricknbolt.com/cost-estimator>
* Gradient Boosting Regressor: <https://machinelearningmastery.com/gradient-boosting-machine-ensemble-in-python/>
* Gradient Boosting Regressor: https://en.wikipedia.org/wiki/Gradient\_boosting