**House Prices in India**

Student’s Name:

Professor’s Name:

Institute of Affiliation:

Course:

Date:

**1.0** **Introduction**

1.1 Introduction

Real Estate business in India is booming due to the increase in population (over 1.3 billions) and urbanization. This has led to creation of more employment and rise in the nation's GDP. The aim of the research is to analyze the variables affecting the housing price in India by creating a house price prediction model.

An understanding of the house prices is vital for the companies in real estate industry as it helps them make strategic decisions and better position themselves in the market. House prices analysis also helps companies develop valuable insights into the dynamics of demand-supply, market trends, and pricing strategies. Critical studying of the house prices will enable the developers to align their business plans with the recent market dynamics thus capitalizing on the profits.

The impact of house pricing has significant relevance to India's business landscape. The real estate industry accounts for a sizable portion of India's Gross Domestic Product (GDP), providing employment opportunities while influencing markets negatively or positively given any potential price fluctuations.

When property values escalate rapidly over time within an Indian regional landscape due to favorable conditions like infrastructure investments; adding demand to available properties- homeowners' net worth rises consequently leading to surging consumer spending across diverse areas like retail outlets—amongst other sectors, including increasing construction scopes directly linked with profitable land ownership structures.

Notably, banks are susceptible to a "housing crash" that could result from the adverse fluctuations in property prices; hence they often provide loans and mortgages based on house values. Housing demand massively influences infrastructure development, encouraging investment activity given favorable price trends.

Moreover, stable or appreciating house prices tend to attract domestic and international investors' attention, impacting investor sentiment positively for business expansions with a vested interest in desirable property holdings. It's noteworthy that government policies such as affordable housing initiatives directly impact house pricing levels alongside overall economic development goals with positive impacts on business activities like real estate investment opportunities while enhancing living standards for citizens around an area of interest.

In essence, understanding India's varied business dynamics requires thorough knowledge application of factors influencing the nation's entrepreneurship spirit fulfilled through efficient management guided by varied factors like market pricing trends. Stakeholders like policymakers, investors both small and multinational corporations alike can benefit from understanding this dynamic industry through monitoring and assessing different aspects that shape the nation's economic landscape amidst market turbulence always bound to happen in any sector growth efforts globally.

The purpose of the research is to identify the factors or rather the variables that play crucial role in determining the prices of houses in India. To achieve this, a predictive model has to be constructed and then its accuracy validated. This will enable the ever growing population of India to figure out the type of house they want to acquire with their respective price tags.

1.2 Objectives and Scope of the Research

Objectives

* To identify the input variables that determine the house prices in India.
* To create and develop a house prices prediction model.
* To validate the accuracy of the model's prediction.

Scope

The research will cover a sample of house prices dataset whereby, the input variables will be analyzed, and several models developed and their accuracy validated using machine learning algorithms such as ridge regression and random forests.

**2.0 Literature Review**

2.1 Business

The Indian government through the ministry of Housing and Urban Affairs initiated programs such as Pradhan Mantri Awas Yojana (PMAY) in 2015 to address the housing issues such as affordability. This came as a result of rapid increase in India's population. The government also developed policies that regulate real estate industry. The booming house development sector in India ensured that each social class accessed the houses both in urban and rural at affordable prices.

The Real Estate Act, 2016 (RERA) was brought to limelight to offer protection to the interests of homebuyers and also to hold real estate developers accountable. Any form complain and arising disputes were managed and solved by RERA.

According to Knight Frank India Real Estate Report, 2021, regional variations such as amount jobs, infrastructures, demand-supply dynamics, and economic growth significantly affected the house pricing. House prices in cities such as Delhi and Mumbai were relatively higher compared to house prices in rural areas which were comparatively lower.

**3.0 Methodology**

3.1 Data

The data was obtained from Kaggle website. It was stored in a comma separated value (csv) file by the name House prices India. The dataset is made up of 23 columns and 14620 rows. The price column is used as the target feature while the other columns as input variables. The data has no Nan values hence clean and ready for data analysis.

The India house price dataset plays a very important role in the house price prediction analysis. This is because the dataset has historical data on various aspects such as latitudes and longitudes, nearness to schools and airports, prices, number of bedrooms and bathrooms, etc. The analysis of this dataset enables creation of predictive models that can be used to estimate/forecast the future prices of houses in India based on the available data. The following methods in which the India house price dataset relates to the house price prediction research: Training data, feature selection, model creation and development, validation and testing, and evaluation and improvement of the model.

3.2 Tools/Analytics

The analysis of the house prices in India was conducted on Jupyter notebook using Python language. Machine learning techniques such as Exploratory Data Analysis (EDA), ridge regression, and random forests were used in this research to visualize the data and predict prices.

Application of ridge regression in this research enabled the model to generalize better hence making it easier for a homebuyer(s) to understand how houses are priced depending on the features they have.

Random forests technique was applied to average the results of several decisions trees thus reducing errors when predicting the house prices.

**4.0 Analysis**

Download the dataset

At the beginning, we opendatasets library.

pip install opendatasets --upgrade

Install and import some python libraries that we will use.

import opendatasets as od

import os

from zipfile import ZipFile

import numpy as np # linear algebra

import pandas as pd # data processing

import seaborn as sns; sns.set(style="ticks", color\_codes=True)

import matplotlib.pyplot as plt

Now we import the csv file with the dataset to colab, then use pandas library to read, manipulate, and analyze it.

House\_price\_df = pd.read\_csv('House Price India.csv')

Data Exploration

Let's print the first few rows (about 10) of the data using '*data.head()*' library.

print(House\_price\_df.head(10))

House\_price\_df.columns

From the above dataset, one can observe the following as the input variables: date, number of bedrooms, area of the house, number of schools nearby, lot area, number of floors, condition of the house, waterfront present, number of views, living area, number of bathrooms excluding basement, lot area renovation, Area of the basement, built year, renovation year, postal code, Latitude, Longitude, living area renovation, grade of the house, and the distance from the airport.

Let's view the last 5 rows of the dataset

House\_price\_df.tail(5)

Exploratory Data Analysis

Remove the 'id' column

House\_price\_df.drop('id', axis=1, inplace=True)

The 'id' column was removed because it is an index column which is only used to identify each row or rather used to identify different types of houses and their pricing in the dataset. Therefore, it is not necessarily important when analyzing the data.

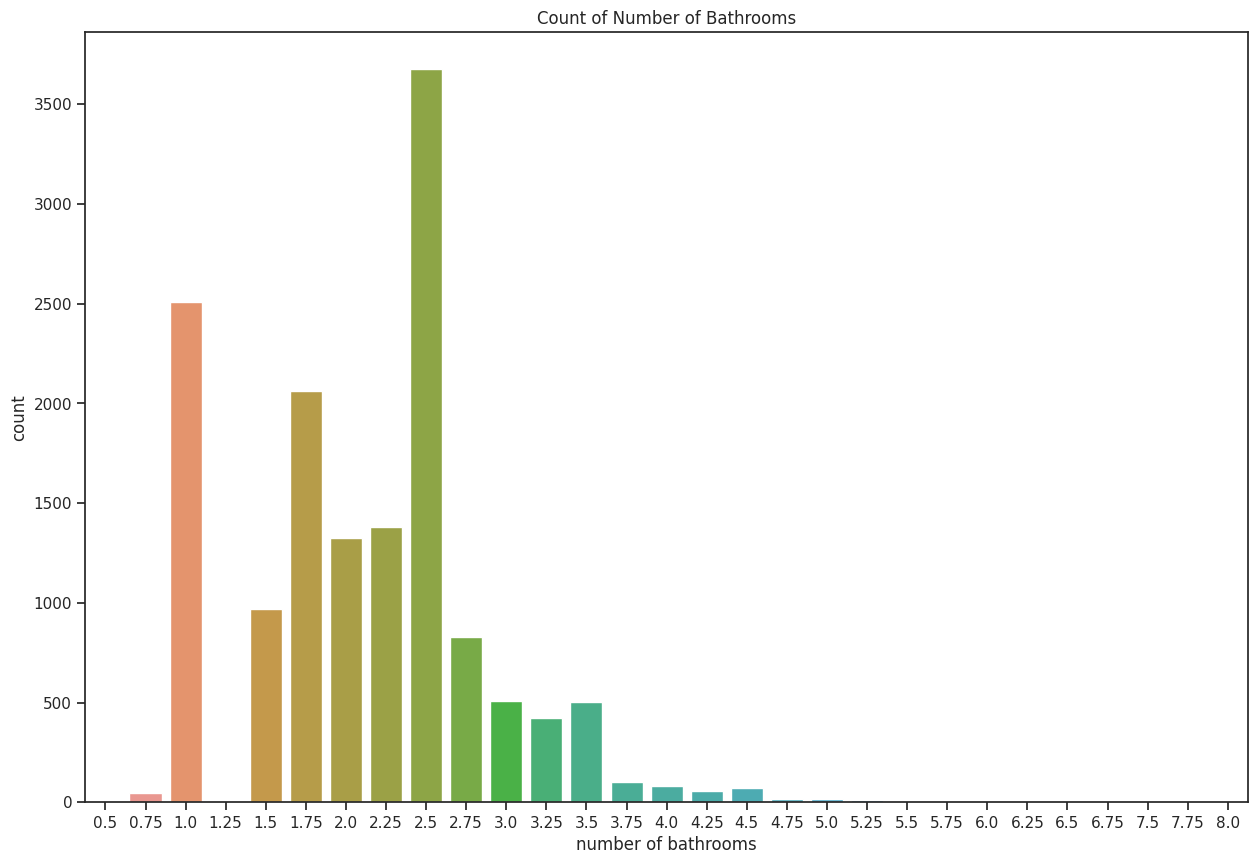
Correlation between Input Variables and Target feature (Price)



The following can be inferred from the above heatmap:

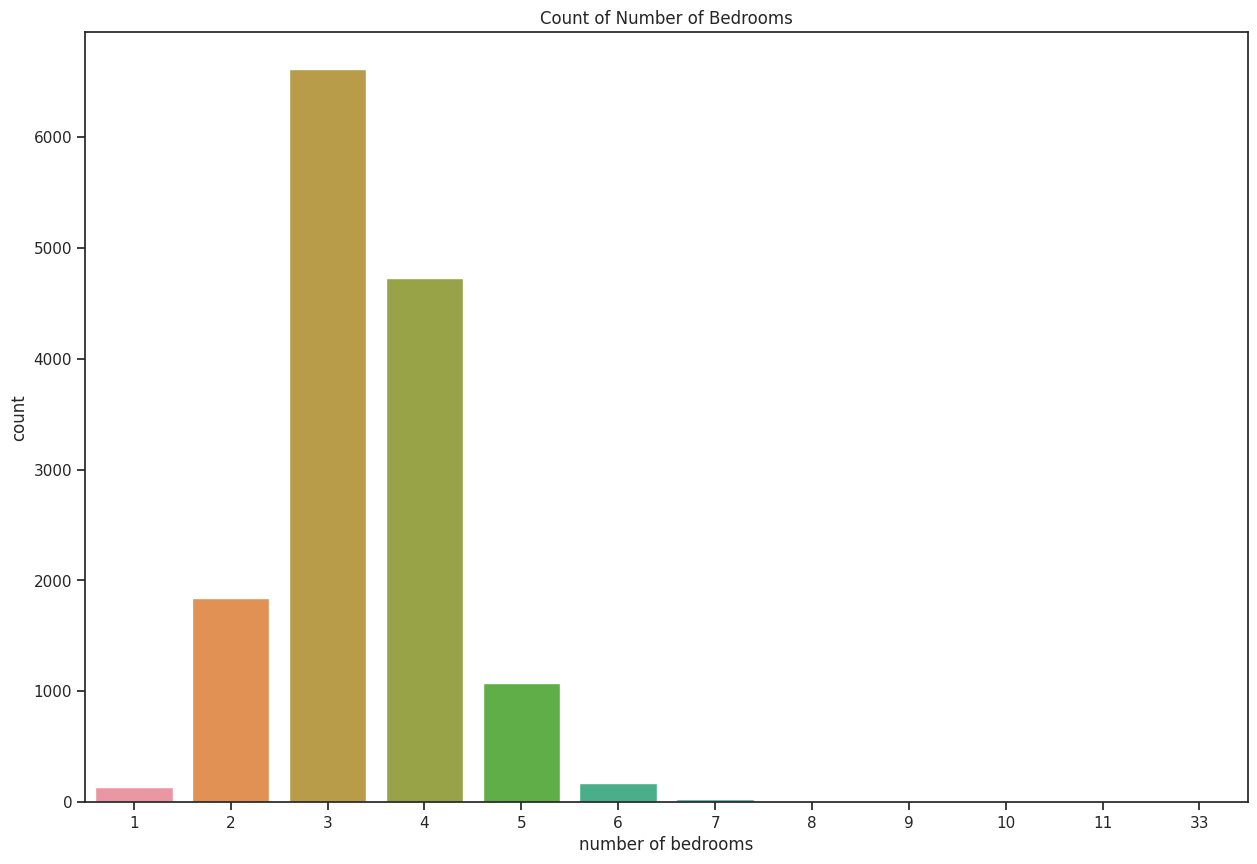
* Input variables such as the number of bathrooms, living area, grade of the house, Area of the house, and the living area renov are highly correlated to the price of the house.
* Low correlation between the price of house and input variables such as the date, built year, postal code, Longitude, number of schools in the nearby, and distance from the airport was observed.

Data Visualization



From the above plot, it is observed that,

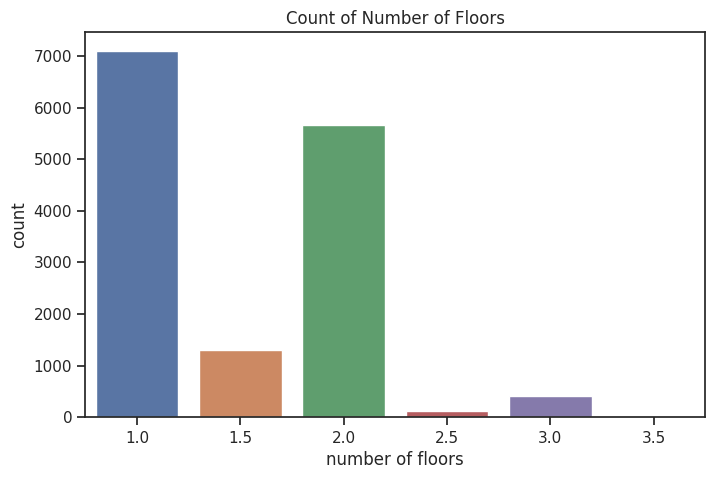
* Many of the houses have 2.5 bathrooms.
* The second in the graph is the houses with only 1 bathroom.
* Houses with 0.75, 4.0, 4.25, and 4.5 bathrooms are less in count.



From the above countplot, the following inferences can be made:

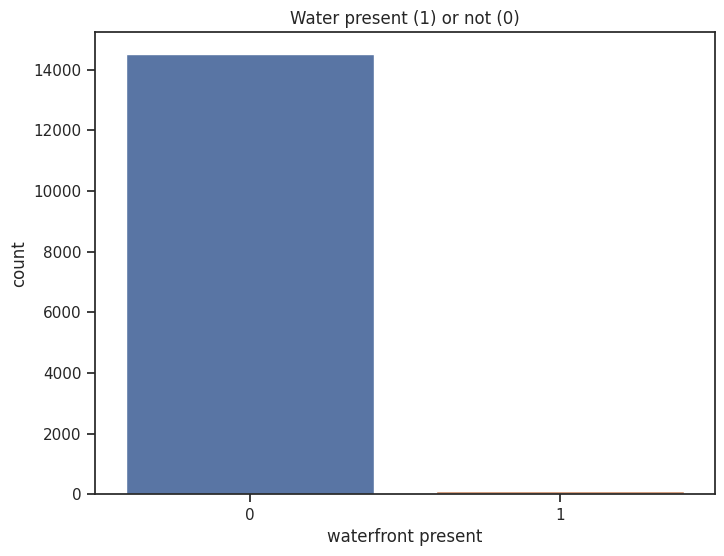
* Most of the houses have three bedrooms. It is then followed by four bedroom house.
* Houses with only one bedroom were the least in the data.

This implies that most of the houses were bought by individuals with families. For example, a three bedroom house would mean that, bedroom one is a master bedroom, the second is the kids' bedroom, and the third bedroom might be the guest bedroom.

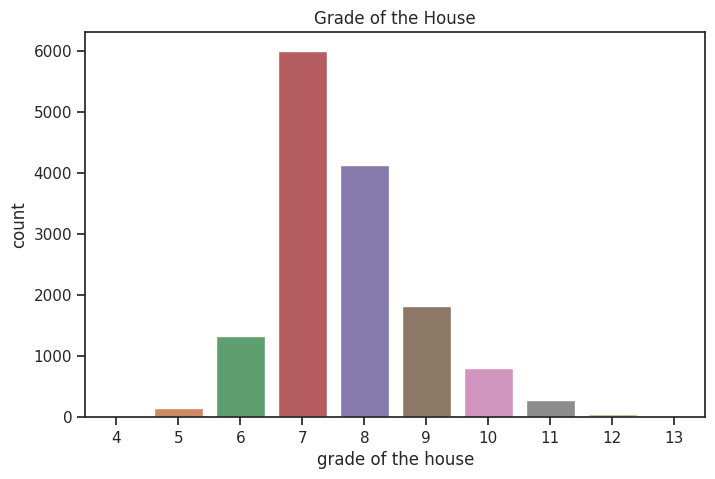


The following inferences can be made from the above graph:

* Most of the house have only 1 floor.
* The second largest count is house with only two floors.
* Houses with 2.5 and 3.0 floors are of the least number of counts. This is because these houses may be very expensive and too big for a household family.

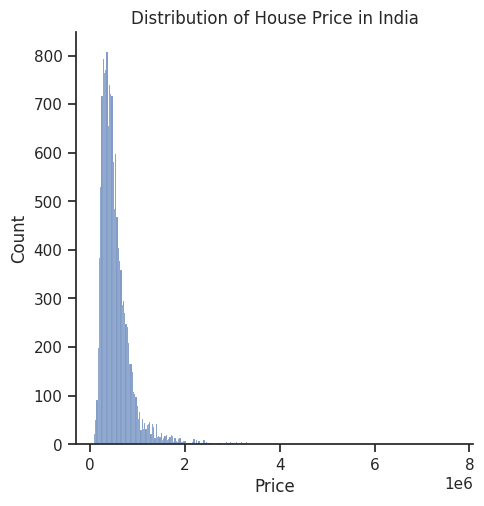


From the above graph, it is found that more than 14000 house lack water in front.



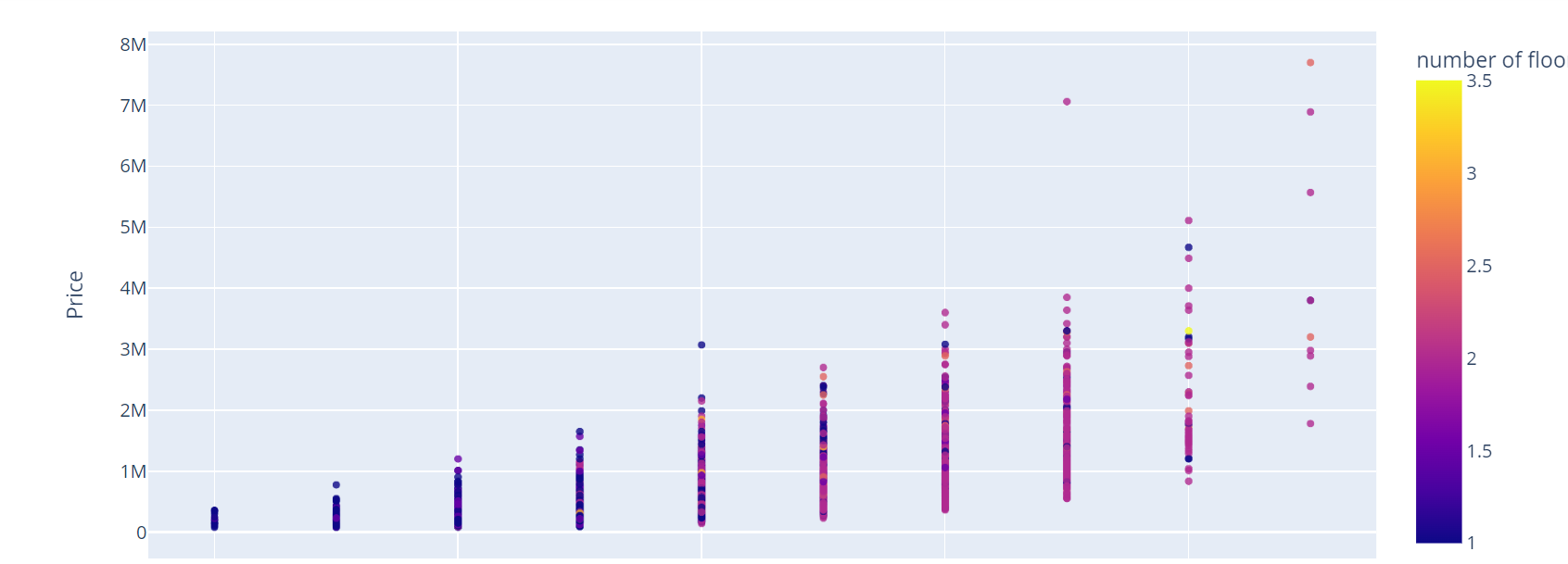
The following observations can be made from the graph:

* Most of the house are of grade 7.
* The second in the rank are houses of grade 8.
* Houses of grade 4, 12, and 13 are the least in the ranking.



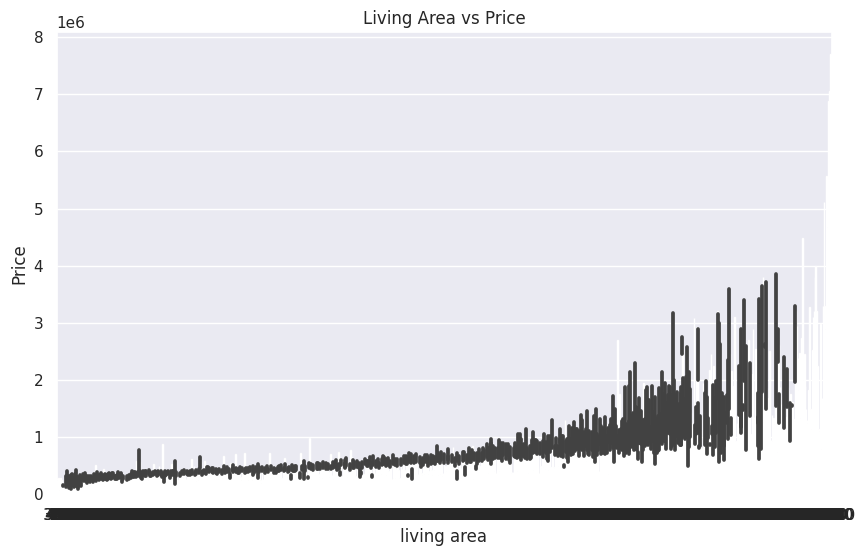
From the above plot, one can notice that the price distribution is right skewed implying that most of the houses are worth less i.e., less than one million Indian Rupee.

Grade of the house vs Price



From the above scatter plot, the following inferences can be made:

* The house prices increase with increase in the grade of the houses.
* The most expensive house costs 7.7 million Rupee. The house has no water front, is of grade 13, and has 2.5 floors.
* The cheapest house costs 78 thousands Rupee. It has the following features; no water front present, grade 5 house, has one floor number.

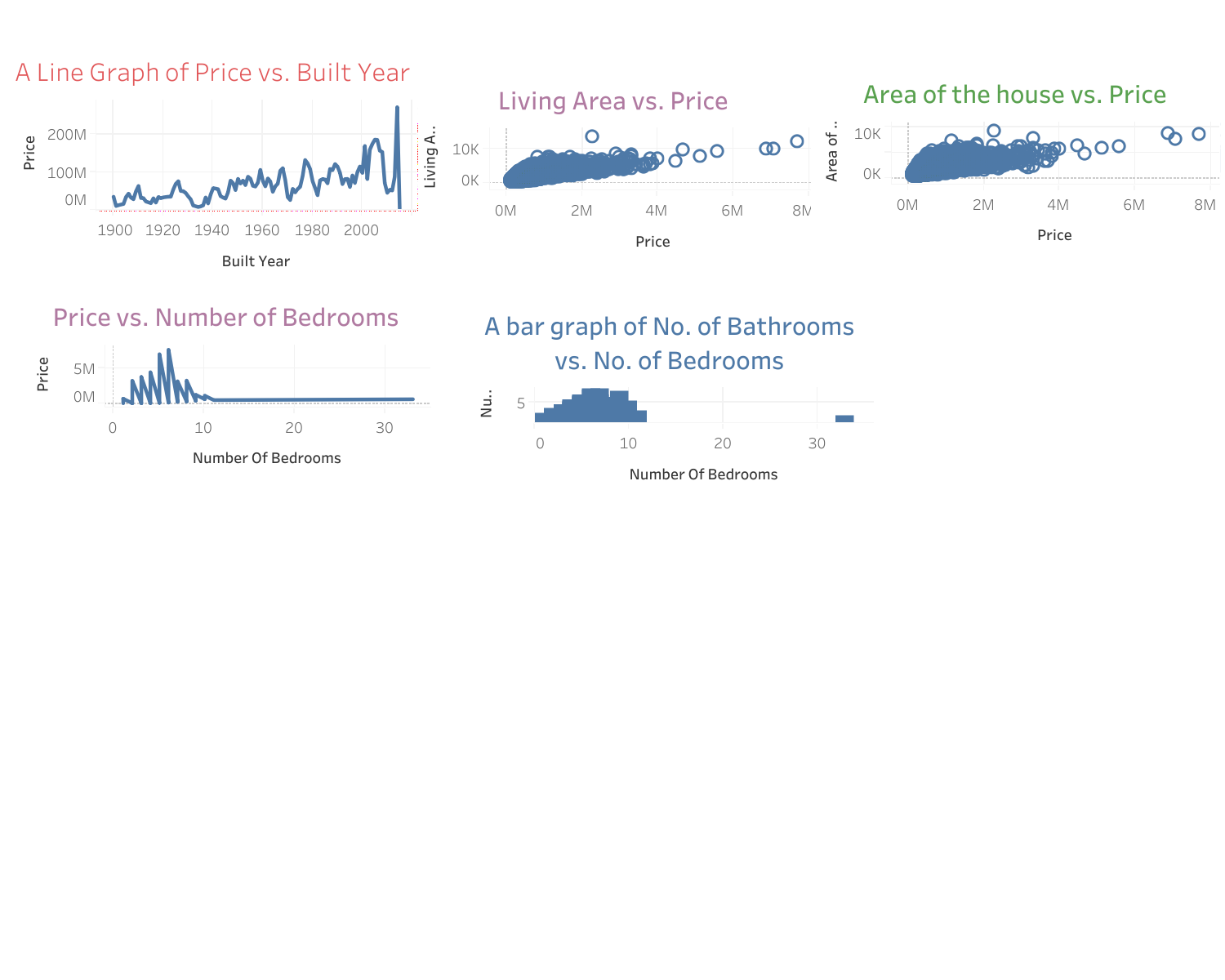


From the above graph, it is found that the house prices in India increase as a result of increase in the size of the living area.



From the graph, the prices of the houses generally decreases from June, 2016 to February, 2017. This may be due to the low demand of houses by the Indians.

**Visualizations with Tableau**

****

**Models**

Ridge Regression Technique

print('The RMSE loss for the training set is $ {}.'.format(train\_rmse))

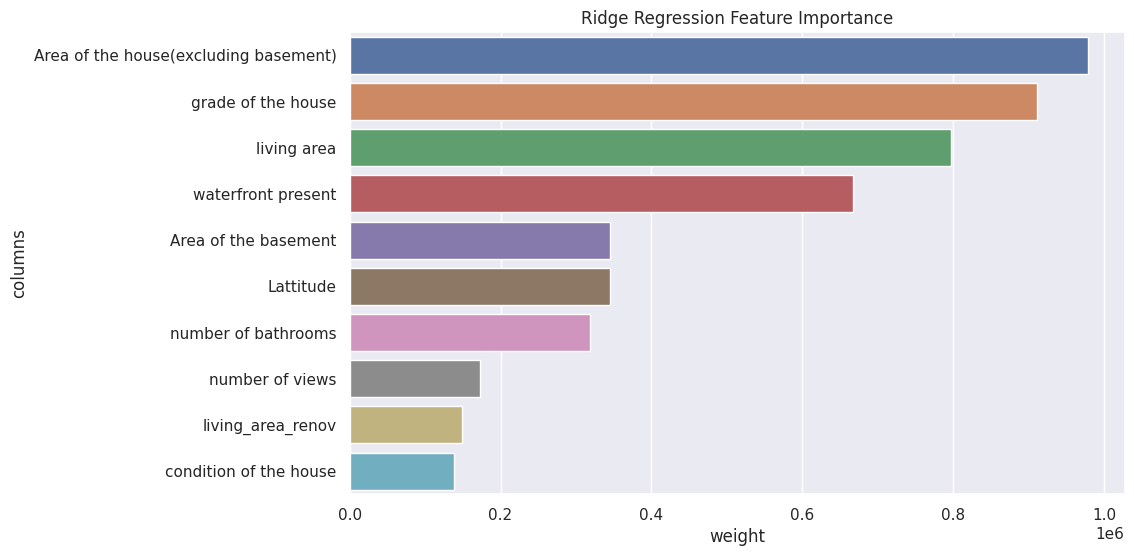
The RMSE loss for the training set is $ 192992.43862855725.

print('The RMSE loss for the validation set is $ {}.'.format(val\_rmse))

The RMSE loss for the validation set is $ 222351.40444933224.

Ridge Regression Feature Importance

The bar plot below showcases the weights assigned to different features of our dataset. It was found out that features or columns such as the grade of the house, living area, and the area of the house were the most important features in during the model development.



Random Forests

Random forests are machine learning techniques that work by averaging the results of several decision trees. The whole idea is that, each decision tree present in forest will generate several types of errors and on averaging, many of these errors will cancel out. For this project, we'll use the RandomForestRegressor class from sklearn.ensemble.

print('Train RMSE: {}, Validation RMSE: {}'.format(rf1\_train\_rmse, rf1\_val\_rmse))

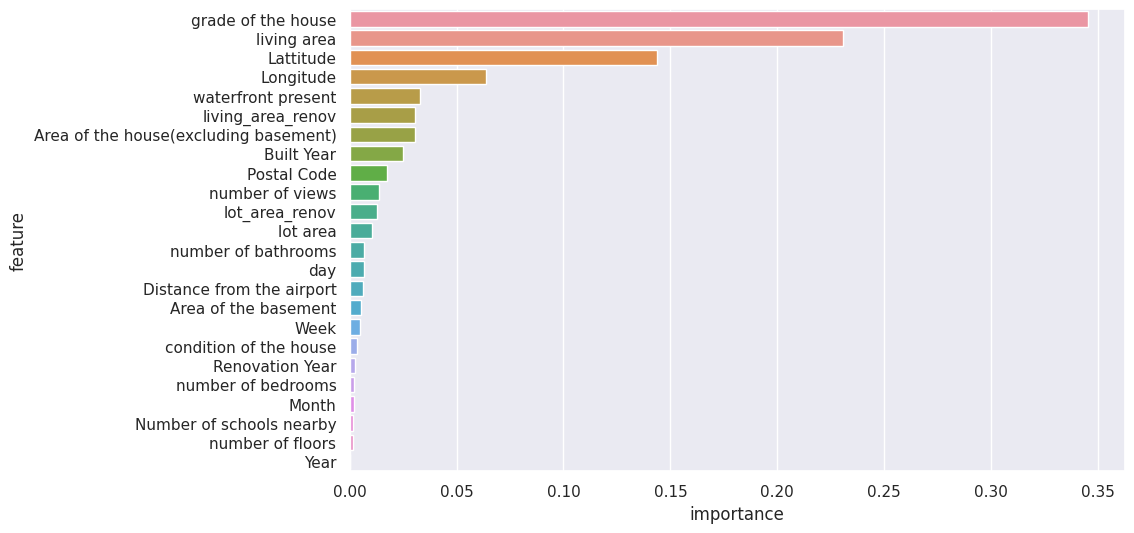
The train root mean squared error is 48062.023716204996 while the validation root mean squared error is 140709.18596951824.

rf1.score(train\_inputs, train\_targets), rf1.score(val\_inputs, val\_targets)

(0.9819853180475872, 0.872705659867272)

From the above errors, one can observe that the loss function for the train data is 48062.023716204996, and the loss function for the validation data is 140709.18596951824.

To improve the validation score, hyperparameter tuning has to be conducted. This will help reduce overfitting of the curve which in turn reduces the validation errors.



From the graph one can notice that the grade of the house, living area, latitudes, and longitudes are the most important input variables for this model.

In general the results of the analysis portrays that most of the affordable houses have fewer bedrooms, bathrooms, and are a distance far from the airport. Nonetheless, people of India do not have to move to urban areas as they can afford properly constructed houses in the rural areas or suburbs. Also, the analysis of house price prediction will assist the government to capture real estate companies that hike or rapidly inflate the prices of houses. This will increase the number of homebuyers as many will feel comfortable purchasing the house units.

**Conclusion**

In conclusion, the objectives of the analysis were met as the model techniques used (ridge and random forests) predicted the house prices in India. Random forests method generated better house price predictions than ridge regression because it has additional and sophisticate features which deeply analyze the data. Input variables such as the grade of the house, living area, area of the house, latitudes, and longitudes play a major role in determining the house prices. A validation score of 87.27% was obtain from random forests model. This implies that the model is good and can be used to make decisions and policies on housing. However, the accuracy score of house prices prediction can be further improved by applying other techniques like XGBoost gradient descent, and hyperparameter tunings as used by Osman et al., (2021) to predict the groundwater levels in Selangor Malaysia.

**References**

Kaggle website: <https://www.kaggle.com/search?q=House+price+in+India>

Malek, M., Mohibali, S., & Bachwani, D. (2021). Identification, evaluation, and allotment of critical risk factors (CRFs) in real estate projects: India as a case study. Journal of Project Management, 6(2), 83-92.

Ministry of Housing and Urban Affairs: <https://mohua.gov.in/index.php>

Osman, A. I. A., Ahmed, A. N., Chow, M. F., Huang, Y. F., & El-Shafie, A. (2021). Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. *Ain Shams Engineering Journal*, *12*(2), 1545-1556.