MBA Semester – IV

Research Project

Name	Maria Wilfred Melvin
USN	222VMBR02629
Elective	Data Science and Analytics
Date of Submission	15.09.2024

A study on "House Price Prediction in Urban India using Machine Learning Algorithm"

Research Project submitted to Jain Online (Deemed-to-be University)

In partial fulfillment of the requirements for the award of:

Master of Business Administration

Submitted by:

Maria Wilfred Melvin

USN:

222VMBR02629

Under the guidance of:

Mr. Sharath Srivatsa

(Faculty-JAIN Online)

Jain Online (Deemed-to-be University)

Bangalore

2023-24

DECLARATION

I, Maria Wilfred Melvin, hereby declare that the Research Project Report titled "A Study on House Price Prediction in Urban India using Machine Learning Algorithm" has been prepared by me under the guidance of Mr. Sharath Srivatsa. I declare that this Project work is towards the partial fulfillment of the University Regulations for the award of the degree of Master of Business Administration by Jain University, Bengaluru. I have undergone a project for a period of Eight Weeks. I further declare that this Project is based on the original study undertaken by me and has not been submitted for the award of any degree/diploma from any other University / Institution.

Place: Bengaluru

Date: 15.09.2024

_____

Maria Wilfred Melvin
USN: 222VMBR02629

CERTIFICATE

This is to certify that the Research Project report submitted by Mr. Maria Wilfred Melvin bearing 222VMBR02629 on the title “A Study on House Price Prediction in Urban India using Machine Learning Algorithm” is a record of project work done by him during the academic year 2023-24 under my guidance and supervision in partial fulfillment of Master of Business Administration.

Place: Bengaluru

Date: 15.09.2024

Mr. Sharath Srivatsa

ACKNOWLEDGEMENT

I would like to express my profound gratitude and special thanks to our mentor Mr. Sharath Srivatsa for the time, efforts and guidance he provided throughout the capstone project. His support and guidance in completing my project on the topic "A Study on House Price Prediction in Urban India using Machine Learning Algorithm" was very helpful. I would also like to thank all the faculty members for providing me with the required knowledge and the program manager Mr. Agneev Lahiri for providing assistance in completing the project within the stipulated timeframe.

_____

Maria Wilfred Melvin
USN: 222VMBR02629

EXECUTIVE SUMMARY

The study aims to predict rental house prices in India by analyzing various property features and employing machine learning models. The primary objective is to forecast future rental prices based on the status of the house, leveraging historical market trends and property attributes. This research explores how diverse features, such as property size, location, and furnishing status, influence rental prices and seeks to enhance the accuracy of price predictions.

Purpose and Objectives: The study's purpose is to forecast rental prices by examining historical data and property features. It seeks to predict prices with high accuracy and to understand how different property attributes impact rental values. The specific objectives are to analyze the status of houses to predict Indian rental house prices and to train a machine learning model to achieve the greatest predictive accuracy.

Scope: The research focuses on predicting rental prices based on property status and features. By analyzing a dataset containing over 4,700 entries with attributes such as BHK configuration, size, and location, the study aims to identify patterns and interactions among variables to improve understanding of rental price determinants. This knowledge can aid sellers in setting prices and assist buyers in making informed decisions.

Research Method: The study employs quantitative research methods, relying on statistical analysis to interpret data. The dataset, sourced from Kaggle, includes diverse features relevant to rental prices. Data cleaning and preprocessing involve handling missing values, correcting inconsistencies, and preparing the data for analysis. Exploratory Data Analysis (EDA) and feature selection help identify key attributes influencing rental prices.

Machine Learning Models: Three machine learning algorithms are utilized KNN, decision tree and RF.

Data Analysis Tools: The study involves classification analysis and various machine learning algorithms to predict rental prices. Models are assessed based on metrics such as mean squared error (MSE) and R-squared (R^2) to ensure accuracy.

Conclusion: The findings reveal significant variability in rental prices influenced by factors such as city location, property size, and furnishing status. Machine learning models, particularly Random Forest, are effective in predicting rental prices, with varying performance across different rent ranges. This research contributes to a better understanding of rental price determinants and provides valuable insights for both property sellers and buyers.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER-I

INTRODUCTION AND BACKGROUND

1. INTRODUCTION AND BACKGROUND

1.1. PURPOSE OF THE STUDY

The purpose of the study is to predict the rental house prices as per the status of the house. It is possible to forecast future prices by examining historical market trends, upcoming developments, and property prices. A common practice in real estate is to list the standard and general characteristics apart from the asking price and general description. These features can be easily compared because each of the attributes is listed separately in an organized manner. Since each home has its own distinct qualities, such as its view or type of sink, House sellers can list all the salient features of the property in the description. Although buyers can consider real estate features, the great diversity makes it impossible to offer an automated comparison of all variables. Conversely, house sellers must assess the worth of the property based on its features. They must compare the features with the current market price. It is challenging to forecast a fair market price and rent because of the variety of features. Thus, the information induces the curiosity of buyers to buy the house with its features.

1.2. INTRODUCTION TO THE STUDY

Housing is a key indicator of a national economy's health. Economic growth often prompts migration from rural to urban areas, increasing the urban population. As more people move to cities, the demand for housing rises, driving up prices. Housing is a pivotal component of urban development and economic stability in India. As urbanization accelerates, driven by economic growth and rural-to-urban migration, the demand for housing in Indian cities has surged. This increase in demand often leads to rising property prices, making accurate prediction of house values crucial for potential buyers, real estate professionals, and policymakers.

According to Alfiyatin et al. (2017), these include physical conditions, design, and location. Physical attributes such as property size, the number of rooms, kitchen and garage dimensions, landscaped areas, and the property's age are crucial. Kang et al. (2021) note that the internal characteristics of a house—such as its size, construction year, and number of bedrooms and bathrooms—also impact its price. Real estate developers leverage these factors in their marketing strategies to attract buyers. Location is another significant factor; properties near hospitals, markets, educational institutions, airports, and major highways generally have higher prices. Understanding these variables helps landlords, analysts, policymakers, and urban planners make informed decisions regarding property investments and development. Various factors influence house prices, including the physical characteristics of the property, its design, and its location. Key physical attributes such as property size, the number

of rooms, and the presence of amenities like gardens and parking spaces are critical in determining property value. Additionally, the property's age and the condition of its infrastructure play significant roles. The location of a property—proximity to essential services like hospitals, educational institutions, and transportation hubs—also significantly impacts its market value.

Recent advancements in machine learning, driven by increased computational power, larger datasets, and improved algorithms, have transformed property price forecasting. Machine learning now enables more accurate predictions, providing real estate professionals and property owners with reliable estimates of property values.

Historically, property value estimation depended on human expertise and traditional statistical methods. However, these methods often struggled with the complexity and non-linear nature of housing market data. Machine learning has revolutionized forecasting by extracting valuable insights from large datasets, offering a more nuanced understanding of property values.

Traditionally, property valuation relied heavily on expert opinions and statistical methods. However, these approaches often fall short in capturing the complexities and non-linear relationships inherent in housing market data. The advent of machine learning has transformed this landscape. With the ability to analyze vast amounts of data and uncover intricate patterns, machine learning algorithms offer a more sophisticated approach to predicting house prices.

In recent years, advancements in machine learning—coupled with increased computational power and the availability of extensive datasets—have revolutionized property price forecasting. These developments enable more precise and reliable predictions, aiding stakeholders in making informed decisions. This study explores the application of machine learning algorithms for predicting house prices in urban India, highlighting the impact of these advanced techniques on the accuracy and efficiency of property value estimation.

1.3. INDUSTRY PROFILE

Overview

Real estate is a substantial international industry composed of the following four essential sectors: residential, retail, hospitality, and commercial. The expansion of the corporate environment, which stimulates the need for office space and urban and semi-urban housing, is intricately connected to its growth. The construction industry occupies a critical position, with

direct, indirect, and induced effects on the economy ranking it third among the fourteen main economic sectors.

After agriculture, real estate is the second-largest employer in India. It was anticipated that increased NRI investment would be attracted to this sector over the short and long term. It was expected that Bengaluru would be the leading destination for NRI in real estate, followed by Ahmedabad, Pune, Chennai, Goa, Delhi, and Dehradun.

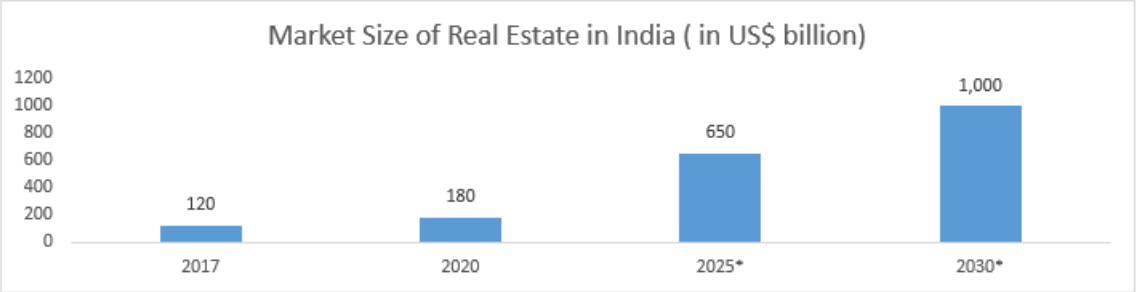
Notable has been the transformation of India's real estate industry from the traditional Zamindari system to contemporary office spaces optimized for increased work productivity. India, being the second-most populous country and the largest democracy globally, has experienced substantial expansion in its real estate industry since attaining independence, with particular emphasis on the 1990s. By 2024, the market is anticipated to be worth USD 7.9 billion. Housing demand remained robust in 2022, with Tier 2 and Tier 3 cities, as well as the top seven cities in India, experiencing substantial growth. The trajectory was bolstered by the escalating propensity for homeownership and the reduction in home loan interest rates. In the top seven cities, housing sales reached all-time highs in 2022, surpassing the previous apex set in 2014. An estimated 365,000 units were sold, compared to 343,000 units in 2014. It is worth noting that branded developers introduced over 60% of newly constructed housing units and accounted for over 55% of total sales in 2022.

Investments in the real estate industry increased by 19% in 2021 and 68% annually in 2022, respectively, demonstrating the sector's robustness and potential for expansion both domestically and internationally. With substantial growth in diverse sectors including warehouses, logistics, industrial parks, data centers, student housing, co-living, and senior assisted living, it is projected to account for 18% of India's GDP by 2030, making it one of the sector's largest contributors. It is anticipated that the Indian real estate industry will attain a valuation of USD 1 trillion by 2030, with affordable housing serving as a pivotal component in this growth. The considerable growth observed in the affordable residential segment can be attributed to the government's favourable policies that support affordable housing and the growing awareness of the advantages associated with homeownership.

Real Estate in India

The real estate industry presently accounts for an estimated 6–7% of India's gross domestic product; by 2025, that figure is expected to rise to 13%. Anticipated to attain a valuation of $1

trillion by 2030, the sector exhibits remarkable fortitude in the face of geopolitical unpredictability and the COVID-19 pandemic. Real estate, being the second-largest contributor to the Indian economy, assumes a pivotal position in stimulating economic expansion. In addition to residential, commercial, office, retail, social infrastructure (including hospitals and schools), hospitality, industrial, and logistics sectors, it also includes emergent industries such as co-working, co-living, and data centers. In addition to steel, cement, tiles, furniture, and furnishings, the sector generates activity in the electrical equipment and construction apparatus sectors.



The residential, retail, and commercial sectors of the real estate industry all experienced significant growth in 2022, notwithstanding the unanticipated decline in 2020. It is anticipated that this ascent will persist into 2023. The sector garnered USD 7.8 billion in equity investments in 2022, of which foreign investors contributed 60% and Indian institutional investors and developers contributed 40%.

It is anticipated that the real estate market will increase from USD 1.72 billion in 2019 to USD 9.3 billion by 2040. It is anticipated that the market will increase from USD 200 billion in 2021 to USD 1 trillion by 2030, and that it will contribute 13% to India's GDP by 2025. Additionally, the retail, hospitality, and commercial sectors are undergoing substantial expansion, supplying vital infrastructure to cater to the changing demands of India.

The projected growth of India's real estate sector indicates that it will account for 15.5% of the country's GDP by 2047, up from its current share of 7.3%. In FY23, residential property transactions in India reached a record-breaking USD 42 billion, representing an astounding 48% year-over-year growth. Additionally, sales volume increased by 36%, reaching 379,095 units. Around 558,000 residential properties are anticipated to be completed by developers in India's main urban centers by 2023, marking a significant milestone.

Residential Real Estate in India

The Indian residential real estate market attained noteworthy milestones in 2022, notwithstanding the multitude of challenges it encountered. Housing sales in the top seven cities established new records, new housing supply was limited but robust, and inventory overhang fell to an all-time low. Even as conditions in peripheral areas returned to near-normal levels following COVID-19, consumer interest increased.

Averaging 4-7% annual increases, housing costs rose in 2022, while higher mortgage interest rates had no discernible impact on residential sales. These patterns underscore the tenacity and sustained interest in the residential real estate industry of India.

Emerging Trends

New Launch Supply Trend

As a result of the strong yearning for homeownership, the residential real estate sector in India is witnessing an influx of newly listed properties. New unit launches increased by 18% quarterly in the top seven cities between the fourth and first quarters of 2022 and 2023, from 92,900 units in the fourth quarter of 2022 to 1.09 lakh units. In addition, the supply of newly constructed homes increased by 23% annually, indicative of a significant surge in residential construction.

The consistent demand for homeownership has fueled this spike in new residential developments. In response, leading and publicly listed developers are stepping up their efforts to meet this growing demand. Apart from Hyderabad, the other six major Indian cities experienced a rise in new launch activity in the first quarter of 2023, both quarterly and annually. MMR and Pune were the main contributors to new launches among the top seven cities, accounting for more than half of the total new supply.

City	Q1-2023	Q4-2022	Q-o-Q
National Capital Region (NCR)	12,450	5,600	122%
Mumbai Metropolitan Region (MMR)	37,300	35,300	6%
Bengaluru	13,600	9,600	42%
Pune	19,400	18,600	4%
Hyderabad	14,600	15,100	-3%
Chennai	6,400	3,100	106%
Kolkata	5,850	5,700	3%

Sales Trend

In the first quarter of 2023, approximately 1.13 lakh housing units were sold in the seven largest cities, surpassing the previous apex that was established in the first quarter of 2022. Housing sales have reached an all-time high over the past decade. This signifies a surge in sales of 14% when compared to the 99,500 units that were sold during the corresponding period of the previous year. In addition, residential sales increased by 23% in the fourth quarter of 2022 compared to the previous quarter, totaling 92,200 units sold.

A robust supply of new residential units, enticing launch incentives from developers, and an increased fear of homeownership as individuals seek safer and more secure investments have all contributed to the robust sales momentum in the top seven cities. A prospective global recession, persistent inflation, and repo rate increases, on the other hand, may have a transient effect on real estate transactions.

Notwithstanding these challenges, significant housing sales activity transpired in the top seven cities of India during the initial quarter of 2023. The NCR, MMR, Bengaluru, Pune, and Hyderabad accounted for 90% of the overall sales.

City	Q1-2023	Q4-2022	Q-o-Q	Y-o-Y
National Capital Region (NCR)	17,100	14,600	17%	-9%
Mumbai Metropolitan Region (MMR)	34,700	28,400	22%	19%
Bengaluru	15,700	11,800	33%	17%
Pune	19,900	16,500	21%	42%
Hyderabad	14,300	11,500	24%	9%
Chennai	5,900	3,800	55%	18%
Kolkata	6,200	5,500	13%	3%

Available Inventory

As of the conclusion of the initial quarter of 2023, the inventory level in the leading seven cities amounted to 6.26 lakh units, reflecting a marginal decline from 6.3 lakh units during the fourth quarter of 2022 and an almost identical figure to 6.27 lakh units recorded in the first quarter of 2022. This represents a 1% reduction in available inventory compared to the previous quarter. Notwithstanding the quarterly decline, the annual inventory levels remained comparatively consistent in the top seven cities.

Among the seven largest cities, MMR exhibited the highest proportion of available housing inventory, comprising 32% of the overall figure. The NCR, Pune, and Hyderabad followed MMR with respective proportions of 19%, 17%, and 13%.

City	Q1-2023	Q4-2022		Q-o-Q	Y-o-Y
National Capital Region (NCR)	119,000	123,700	152,500	-4%	-22%
Mumbai Metropolitan Region (MMR)	200,500	198,000	177,600	1%	13%
Bengaluru	54,500	56,600	56,700	-4%	-4%
Pune	103,800	104,300	97,500	0%	6%
Hyderabad	83,700	83,300	71,200	0%	18%
Chennai	28,700	28,200	32,400	2%	-11%
Kolkata	36,500	36,900	39,900	-1%	-9%

Government Initiatives

RERA (Real Estate Regulatory Authority)

RERA was established through the Real Estate (Regulation and Development) Act of 2016, aimed at safeguarding home buyers, and fostering real estate investments. The Act was approved by the Rajya Sabha on March 10, 2016, and came into effect on May 1, 2016, though initially, only 52 out of 92 sections were implemented. By May 1, 2017, all other provisions were in force. With RERA's establishment, the traditionally unregulated real estate sector became subject to regulatory oversight, holding developers, builders, promoters, and agents accountable for their actions, leading to increased transparency and consumer trust. RERA has resolved over 100,000 disputes in the last five years, offering significant relief to home buyers. There has been a 21% CAGR in agent registrations from 2019 to 2023. The implementation of RERA has stimulated home sales, strengthened financial positions, reduced reliance on institutional finance, and increased the focus on mid- to premium-segment properties.

Smart Cities Mission

The Smart communities Mission, initiated by the Prime Minister on June 25, 2015, seeks to foster communities that offer necessary infrastructure and a superior standard of living for

inhabitants, while guaranteeing an environmentally friendly and sustainable atmosphere through the implementation of "smart" solutions. This project promotes economic development and improves the overall well-being of individuals by prioritizing the social, economic, physical, and institutional aspects. The mission prioritizes the achievement of inclusive and sustainable development by establishing reproducible models that can be used as standards for other communities seeking to achieve similar goals. Through a two-stage competition, 100 localities were chosen to undergo development as smart cities. These cities would get a total of USD 22 billion in funding for 8,020 approved projects.

REITs

The SEBI established Real Estate Investment Trusts (REITs) in 2014 as a mechanism for investors to contribute capital for the purchase, operation, and ownership of income-generating real estate assets. REITs are highly regulated, publicly traded entities that provide investors with consistent returns. Real estate investments that generate income enable developers to monetize such assets and reinvest the proceeds in additional development. It is anticipated that a minimum of four new REITs, namely Embassy Office Parks REIT, Mindspace Business Parks REIT, Brookfield India Real Estate Trust, and Nexus Select Trust, will be introduced on Indian stock exchanges between the latter part of 2021 and the beginning of 2025.

PMAY

The Credit-Linked Subsidy Scheme (CLSS) of the PMAY seeks to provide "housing for all" by subsidizing interest on loans taken out to purchase or construct new homes. By categorizing "affordable housing" as infrastructure, developers can procure funds via commercial borrowings from outside sources. PMAY, which debuted on June 1, 2015, advocates for affordable and sustainable housing for the urban disadvantaged. Two schemes are included in accordance with the target areas:

- PMAY-G: This scheme supports families in the economically weaker sections and lower-income groups by providing affordable financing for their homes. It has approved nearly 2.92 crore houses, with 2.32 crore completed.
- PMAY-U: This scheme encompasses over 4,300 cities and towns across India. It has approved 118.9 lakh houses, with 75.51 lakh completed.

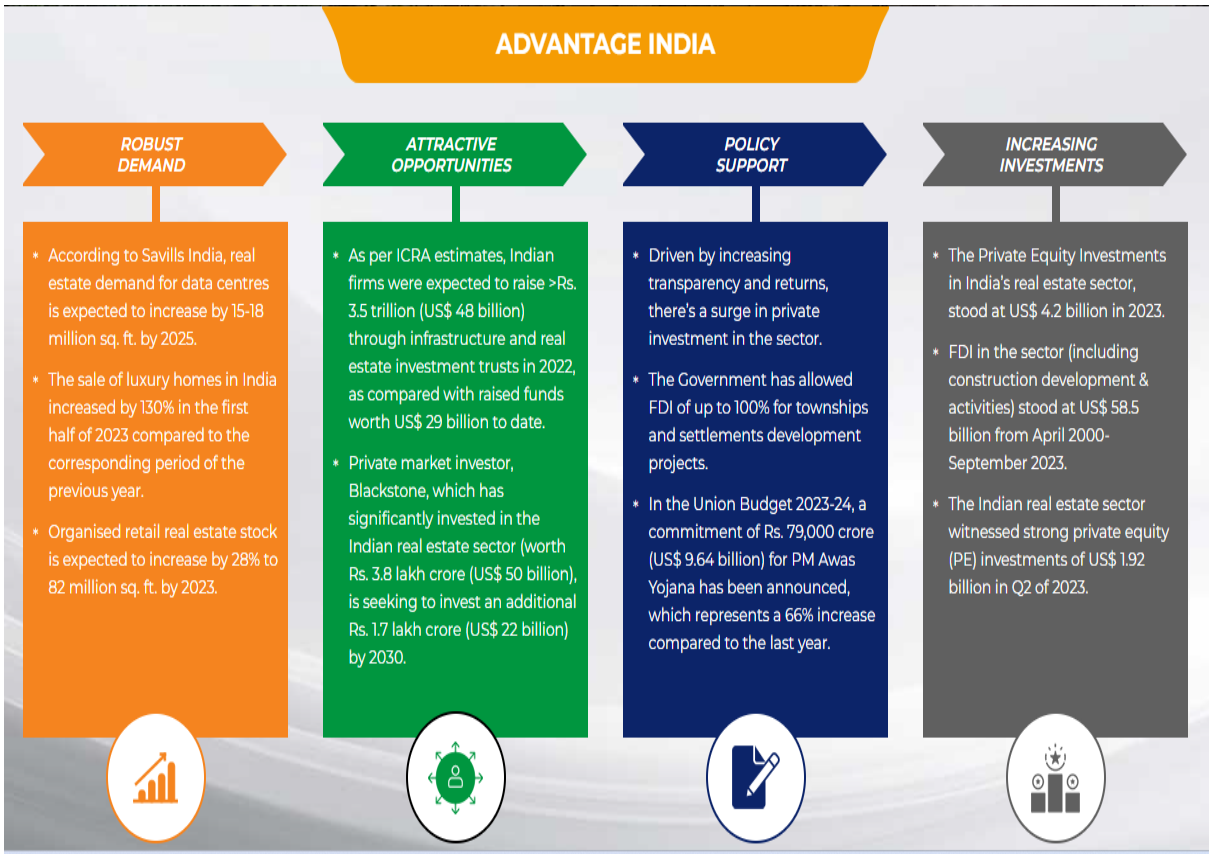
Monetization of Non-Core Real Estate

The government established the National Land Monetization Corporation in 2022 to monetize non-core real estate assets owned by public sector enterprises (PSEs). This initiative is designed to unlock the value of underutilized assets and generate additional revenue for the government.

Road Ahead

India is distinguished as a prominent global economy experiencing rapid expansion, wherein the real estate industry assumes a critical function in propelling the overall development of the economy. Significantly, it occupies the position of the second-largest employer in the nation and ranks third in terms of receipt of FDI. To optimize their operations in anticipation of a sustained high demand for residential properties in the first quarter of 2023, a growing number of real estate professionals are adopting automated solutions. The adoption of blockchain, virtual excursions, 3D property models, and AI demonstrates the industry's growing technological acceptance. AR and VR-enabled virtual excursions provide clients with interactive experiences, whereas AI assists real estate agents and brokers in precisely forecasting property values and rents within particular markets.

Furthermore, the implementation of digitalization initiatives, including the National Generic Document Registration System and the Digital India Land Records Modernization Programme, has been instrumental in mitigating land disputes and promoting the openness of land-related transactions. In general, technological progress has brought about a significant transformation in the real estate industry, empowering enterprises to maintain a competitive edge and deliver improved client experiences.



1.4. STATEMENT OF THE PROBLEM

Housing is one of the basic components that help to measure the economy of a nation. This determines the success of the country. When there is an increase in the economy, people migrate from urban to rural areas. This resulting in an increase in the population of urban society. An increase in urban society will increase the demand for accommodation. If there is an increase in demand, the price of a house will increase to a great extent. Infrastructural development in the area increases the price of houses. For example, if the area is filled with motorable roads and stable electricity, residential areas get high demand, which leads to an increase in the demand for rental houses in the specific area. Generally, it is quite a complex task to predict house rental prices for investors and valuers. They must depend on market data to assess the rental prices of the building over a period. The outcome has the power to influence the decisions of stakeholders, starting from buyers and investors to sellers. The conventional way of predicting prices is based on historical trends, expert opinions, and comparisons. This fails to capture the dynamic relationship in the real estate market. Thus, the key issue of the study is to make an accurate prediction of house rental that assists investors, prospective owners, developers, appraisers, and others in making valuable decisions.

CHAPTER-II

REVIEW OF LITERATURE

2. REVIEW OF LITERATURE

Imran et al. (2021) introduce a methodology for developing intelligent regression models to predict house prices using ML algorithms. Their approach is divided into four stages: data collection, preprocessing and transforming the data into the optimal format, developing ML models, and finally training, testing, and validating these models with house price data from Islamabad's real estate market. The validation and testing phase relies on data from online property sources, providing a robust estimate of the housing market in the city. The results of their regression analysis are promising and suggest potential for future work in predicting housing prices in Pakistan.

Thamarai & Malarvizhi (2020) address the challenge of rising house prices by developing models to assist customers in finding homes that meet their needs. Their study uses a range of house attributes, including the number of bedrooms, the age of the house, transportation accessibility, proximity to schools, and nearby shopping malls. The research focuses on predicting house prices in a small town in the West Godavari district of Andhra Pradesh, using various ML techniques implemented through Scikit-Learn. The study applies decision tree classification, decision tree regression, and multiple linear regression to model house availability and predict prices effectively.

Yağmur et al. (2022) explore two primary approaches to determining house prices in the literature. One approach forecasts house price based on macroeconomic variables in the country of production, while the other uses micro-variables related to the specific characteristics of the house. Their study aimed to predict house prices using ML techniques with a focus on micro-variables that describe the house's features. Conducted in Antalya, Turkey—a region with high foreign demand for housing—the study used data from house advertisements across various income groups. The results indicated that the ANN model provided more accurate predictions compared to SVR and MLR. This model offers a promising tool for institutions involved in housing provision, sales, and valuation, particularly in developing countries, to better predict and manage fluctuating house prices.

Adetunji et al. (2022) argues that predicting price variance rather than exact values is often more practical in real-world scenarios. They propose treating house price prediction as a classification problem rather than relying solely on the House Price Index (HPI), which measures average price changes in repeat transactions or refinancing. Since HPI does not

account for the specifics of individual homes, it is not always effective for precise price predictions. This study investigates the use of the Random Forest ML technique for house price forecasting. Using the Boston housing dataset from the UCI ML Repository, which includes 506 entries and 14 features, they found that their model produced predictions with a ±5 error margin compared to actual prices, demonstrating its practical utility.

Zulkifley et al. (2020) highlight that fluctuations in house prices are a significant concern for homeowners and the real estate market. Their literature review aims to analyze relevant attributes and identify the most effective models for forecasting house prices. The study found that artificial neural networks, support vector regression, and XGBoost were the most efficient models for this purpose. Additionally, locational and structural attributes were identified as key factors in predicting house prices. This research is valuable for housing developers and researchers as it pinpoints crucial factors influencing house prices and suggests the most effective ML models for future studies in this field.

Satish et al. (2019) investigates how ML algorithms can predict future housing prices. Their findings underscore the need for robust prediction methods capable of matching or exceeding the accuracy of existing house price models. The study also notes that housing value indices play a role in enhancing housing price prediction and informing real estate policy. The researchers developed and tested various ML models, including XGBoost, Lasso regression, and neural networks, evaluating their performance based on accuracy. They recommend Lasso regression as the most accurate model for predicting housing costs, demonstrating its superior performance compared to other algorithms in their tests.

Dipanshu et al. (2023) aims to provide an estimate of market value for land properties by considering topographical factors. Their method involves evaluating previous market models, price ranges, and upcoming developments to predict future costs. Using a decision tree regressor, their model estimates house prices in Mumbai, facilitating property additions to the market without direct interactions with sellers. The results from their assessment indicated that the decision tree regressor achieved an accuracy rate of 89%.

Awonaike (2022) introduces a cumulative layering approach within the MfHPE framework. This data-driven, ML-based framework not only identifies the most effective algorithms and features for improving model accuracy but also employs a cumulative multi-feature layering technique to enhance and evaluate ML models. This approach aims to provide actionable insights for stakeholders in the housing sector, leading to more realistic house price estimates. The

MfHPE framework utilizes the DSRM and integrates transaction data from HM Land Registry, specifically 1.1 million records from London between January 2011 and December 2020, with validation performed on 84,051 transactions from 2021. The framework allows for the incorporation of new datasets and algorithms, while also considering various neighborhood and macroeconomic factors such as the locations of bus stops, rail stations, supermarkets, as well as inflation rates, GDP, employment rates, CPIH, and unemployment rates. The study demonstrates that incorporating new features across multiple layers improved performance in 50% of models, shifting the top-performing models with the introduction of additional features. Moreover, the selection of evaluation metrics should be guided by critical business objectives, feature diversity, and the specific ML algorithms used.

El Mouna et al. (2023) emphasize the importance of accurate housing price forecasts for several reasons. These forecasts help individuals make informed decisions about buying or selling real estate and setting appropriate prices. They also aid real estate agents and investors in making better investment decisions and negotiating contracts more effectively. Additionally, fluctuations in housing prices can reflect broader economic conditions, with price decreases potentially signaling an economic downturn and price increases indicating economic growth. The study proposes to address this issue by predicting house prices using ML techniques, specifically LN, RF, and XGBoost. The models were tested on the Melbourne real estate dataset, which comprises 34,857 property sales and 21 features.

Abebe (2021) seeks to explore trends in housing price forecasting through a review of published papers and identify the leading supervised learning algorithms currently in use. The study aims to develop predictive ML models for classifying and forecasting housing prices using algorithms such as support vector machines, linear regression, random forest, one-way ANOVA, and decision trees. Given the presence of many unknown values in the housing price dataset, the study uses metrics like RMSE and R-squared to evaluate model performance. The final goal is to select the most effective algorithm for housing price forecasting, offering valuable insights for both public stakeholders and real estate professionals. The main benefit of this research is its ability to support effective budget tracking, control, and management for future planning.

Yalgudkar and Dharwadkar (n.d.) highlight that real estate, along with gold and share markets, is a popular investment choice known for its substantial returns. Tracking housing price trends is crucial for both buyers and sellers, as it also reflects the broader economic conditions. Various factors influence housing prices, including the number of bedrooms, location, and floor number. Additionally, proximity to major roads, educational institutions, shopping centers, and employment opportunities can drive up house prices. In their study, Pune was chosen as the case study location to develop a model for predicting real-time house prices across different localities. They utilized data from real estate websites such as 99acres.com, magicbricks.com, and nobroker.com, focusing on features like 'area,' 'bedrooms,'

and 'bathrooms.' The study aims to create a predictive model using regression techniques, including MLR, Lasso, and XGBoost, and compares their accuracy to identify the most effective model.

Henriksson and Werlinder (2021) compared the performance of XGBoost and RF regressors for predicting housing prices using two distinct datasets. Their comparison considered training time, inference time, and three evaluation metrics: R-squared (R^2), RMSE, and MAPE. The study involved thorough data cleaning, hyperparameter tuning, and 5-fold cross-validation to ensure accurate performance estimates. The findings indicate that XGBoost outperforms Random Forest on both small and large datasets. While Random Forest can deliver comparable results, it requires significantly more training time—between 2 and 50 times longer—and has longer inference times, approximately 40 times longer, making XGBoost particularly advantageous for larger datasets.

Durganjali and Pujitha (2019) emphasize the importance of predicting the long-term resale value of a house, especially for individuals planning to live there for an extended period before selling it. This is also relevant for those aiming to minimize risks during the construction of their home. The authors use various classification techniques, including LR, DT, NB, and RF, to estimate a house's resale value. Additionally, they apply the AdaBoost method to enhance the performance of weaker models. Factors such as physical attributes, location, and several economic variables impact the resale price. The study measures accuracy across different datasets to identify the most effective approach for sellers to predict resale prices.

Sawant et al. (2018) project that India's housing market will grow by 30-35% over the next decade, second only to agriculture in terms of job creation. Pune is highlighted as a promising location for real estate investment. However, the inconsistency in housing valuation presents a challenge for buyers. The estimated price needs to strike a balance that benefits both the seller and the buyer, ensuring the price is fair. To achieve this, algorithms like DT and bagging techniques such as RF are used to select various features from the dataset for more accurate price estimation.

Wang et al. (2021) argue that property value predictions lacking consideration of all relevant factors result in inaccurate forecasts. To address this, they propose a comprehensive joint self-attention model for house price prediction. Their approach incorporates satellite imagery to evaluate the surrounding environment of residential areas and uses data on public amenities like parks, schools, and BRT stops to provide a detailed depiction of neighborhood facilities. The model leverages attention mechanisms typically used in image, speech, and translation tasks to identify key features that homebuyers consider. When provided with transaction data, the model automatically assigns weights. Unlike conventional self-attention models, this approach accounts for the interdependencies between different parameters, leading to a more accurate prediction.

Lim et al. (2016) compared the predictive performance of the ANN model, specifically the multilayer perceptron, with that of the ARIMA model in forecasting the Singapore housing market. They applied the superior model (CPI) to forecast future condominium price indexes. The ANN model demonstrated lower MSE, indicating its superior accuracy over other prediction models.

Piao et al. (2019) highlight the complexity of factors influencing residential real estate prices and the challenge of identifying useful features, which often leads to lower accuracy in traditional home price prediction models. To address this, they propose an innovative CNN-based prediction model combined with a feature selection process. Their approach, when tested with real-world property transaction data, outperforms traditional methodologies, delivering more accurate results.

Madhuri et al. (2019) focused on predicting house prices for first-time buyers, considering their financial capacity and goals. The study aimed to derive future prices by analyzing past sales, rental trends, and upcoming developments. Several regression techniques were used, including Multiple Linear, Ridge, LASSO, Elastic Net, XGBoost, and Ada Boost Regression. The model also accounted for key factors such as physical conditions, design, and location in the pricing estimation process.

Shinde et al. (2018) applied various ML algorithms to develop a predictive model for house prices. They utilized methods such as logistic regression, support vector regression, Lasso regression, and decision trees, using data from 3,000 properties. The R-squared values for these models were 0.98 for logistic regression, 0.96 for SVM, 0.81 for Lasso regression, and 0.99 for decision trees.

Dagar et al. (2020) explored different ML techniques for house price prediction based on specific features. Their dataset, consisting of 13,000 records from Bangalore, India, included nine key features. They concluded that the multivariable LR model provided the most accurate and reliable results, outperforming other methods in terms of accuracy and error rates.

Jha et al. (2020) employed multiple ML algorithms to tackle challenges in the real estate market, including LR, RF, voting classifiers, and XGBoost. They integrated these techniques with item coding to develop a model capable of accurately predicting whether the negotiated sales price would be higher or lower than the listed price. To assess model performance, they measured accuracy, precision, recall, F1 score, and error rate. Among the four algorithms tested, XGBoost demonstrated the best performance and highest robustness compared to the others.

Hjort et al. (2022) explored the use of enhanced gradient boosting trees (GBTs) for predicting house prices, focusing on the impact of different loss functions on prediction accuracy. GBTs are commonly applied in regression tasks, where the choice of loss function plays a critical role in performance. The study evaluated four loss functions: MSE, MAE, Huber loss, and quantile

loss. Using a dataset of known features, they split the data into training and testing sets, training GBT models with each loss function and evaluating their predictive accuracy on the test set.

Ho et al. (2021) applied three ML algorithms—SVM, RF, and XGBoost—to predict property prices. They tested these models using a dataset of 40,000 property transactions in Hong Kong, spanning 18 years, and compared their outcomes. The model performances were evaluated using three metrics: MSE, RMSE, and MAPE.

Zou (2023) built a predictive model utilizing methods such as logistic regression, support vector regression, lasso regression, and decision trees. This study used data from 3,000 properties in Jinan and employed R-squared to assess the effectiveness of these algorithms. The research offers valuable insights into applying ML techniques for accurately predicting house prices in Jinan, China.

CHAPTER-III

RESEARCH METHODOLOGY

3. RESEARCH METHODOLOGY

3.1. OBJECTIVES OF THE STUDY

The objective of the study is

- To predict the Indian rental house price by analysing their status of the house
- To analyse and train a machine learning model that that predict house pricing with the greatest accuracy

3.2. SCOPE OF THE STUDY

The study focuses on predicting Indian rental house prices based on key factors: the house's status. By analyzing historical data and property features such as size, location, and price per square foot, the study aims to identify patterns that can accurately forecast future rental prices. This research also examines the complexities of real estate pricing, considering the diverse and unique characteristics of each property. The study seeks to improve the understanding of how these variables interact, aiding sellers in pricing strategies and helping buyers make more informed decisions.

3.3. RESEARCH METHODOLOGY

This section is concerned with the datasets used for the ML algorithm. The section aims to outline the datasets and the detailed description of variables that exist in the datasets. The full details of the datasets and the application of various algorithms can be found in the following section.

3.3.1. RESEARCH METHOD

The present study adopts quantitative research methods. This method quantifies the results based on statistical analysis. The information is interpreted with numbers. It is highly structured and formalized. The determination of results is based on studying a few variables in many entities.

3.3.2. DATA SET

The Indian Rental House dataset is available on Kaggle and contains information about rental house price details in India. This dataset contains more than 4700 data points with 16 variables,

including BHK, Rent, Size, Floor, Area type, locality, city, Furnishing, Tenant preferred, Bathroom and Point of contact. All the variables served as features of the dataset. This helps to predict the price based on status.

3.3.3. DATA CLEANING

The data obtained from the repository was initially in the form of a text file. I connected the text file to Excel, extracted the data, and saved it as a comma-separated values (CSV) file. Data cleaning is an iterative process, and the first step involves detecting and correcting inaccurate or bad records. The dataset from the repository contained various inconsistencies and missing values. Before loading it into ML models, it was essential to clean and correct the data to achieve high prediction accuracy. Since I used different tools for predictions, the cleaning process varied depending on the tool. However, the main objective was always to improve model accuracy.

In the Indian house price dataset, some entries were incomplete, particularly the lack of state names, with only latitude and longitude values provided. By using the Python, I was able to determine the corresponding states. It turned out that all the locations were within the India. Afterward, the null values were removed to minimize inconsistencies in the data, ensuring better results in model predictions.

3.3.4. DATA PRE-PROCESSING

Data Preprocessing begins with checking for missing data. Variables with more than 50% missing values are removed from the dataset, while others are imputed with values closer to the mean. Following this, the dataset is checked for outliers, which are removed to enhance model performance. After that, the numerical values are normalized, and categorical values are encoded one at a time to prepare the dataset for analysis.

3.3.5. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) includes exploring the dataset to identify important features using a heat map. A correlation matrix is performed for all the features to uncover the most and least correlated variables. These results help recognize the interaction between the features and the target variables, allowing for a better understanding of the data.

3.3.6. FEATURE SELECTION

Feature Selection is the subsequent step. The most relevant features are selected for the study based on their correlation with the target variable. Scaling is applied to ensure the proper selection of the most appropriate features. A standard scalar function is employed to naturally distribute the data around 0 with a standard deviation of 1.

3.3.7. MODEL DEVELOPMENT

Finally, Model Development takes place, where the data is split into training (80%) and testing (20%) sets. The training set includes the target variable, which helps the model learn the relationship between features. The model is optimized for hyperparameters and assessed using metrics like mean squared error (MSE) and R-squared (R^2). Multiple ML algorithms are used to train the model, and the model with the highest accuracy is selected for predicting rental prices.

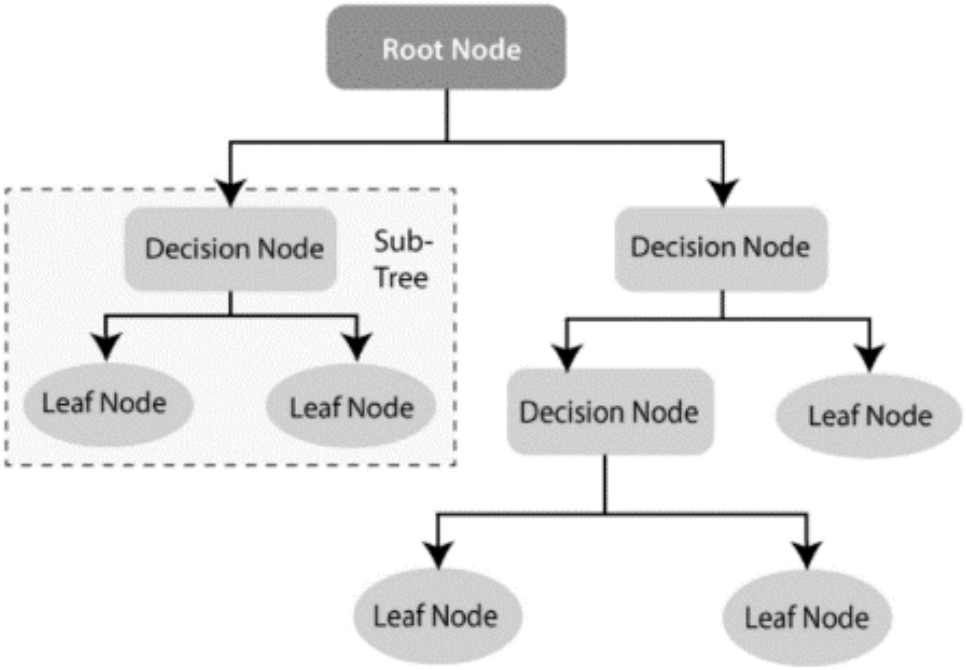
3.4. DATA ANALYSIS TOOLS

Machine learning algorithm

The present study adopts classification analysis to perform machine learning algorithms. It is also a supervised learning method. Classification analysis can be described as a predictive modelling problem. It is a mathematical mapping of a function (f) as a target from input (x) to output variables (Y). The prediction of class is available for structured or unstructured data. In this analysis, the class label is predicted for the subsequent example. The best example for classification analysis is detecting email spam that can be classified as spam and not spam in the classification problem. The present study adopts the most used methods, including KNN, Decision Tree, and RF.

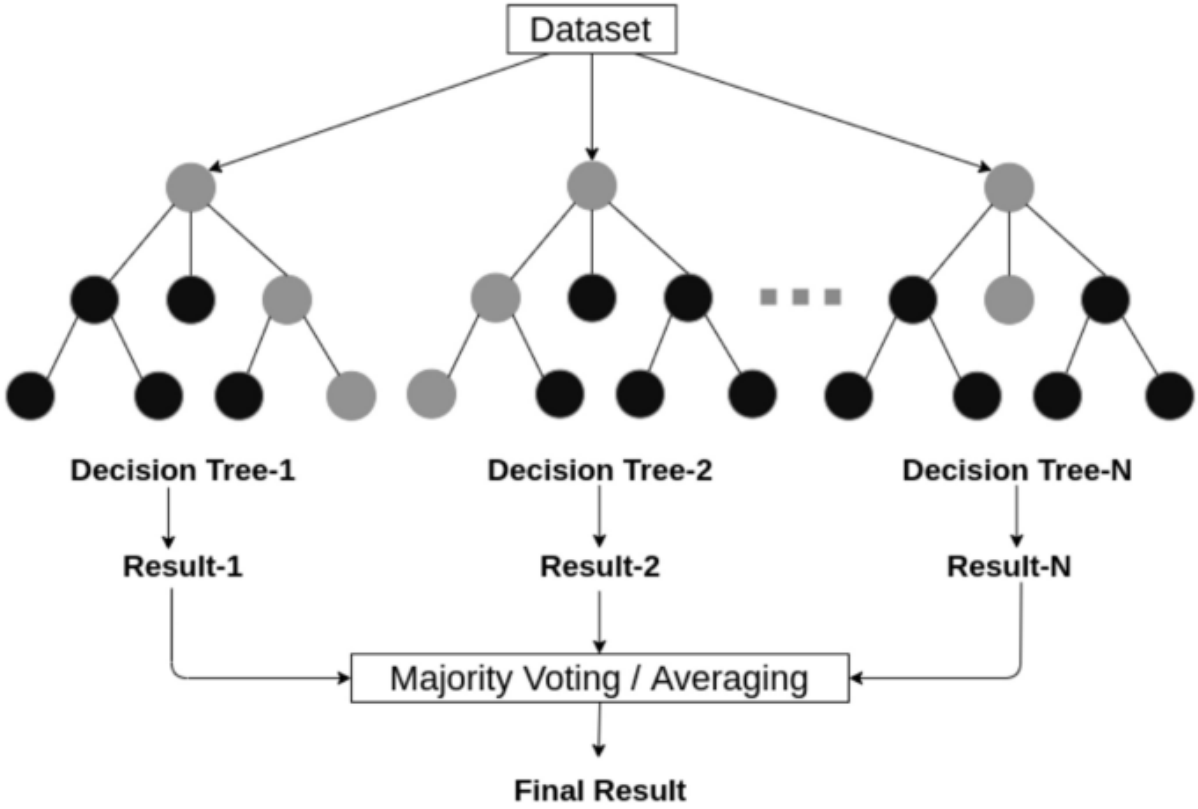
- KNN, abbreviated as K-Nearest Neighbors, is a non-generalizing learning method. It is also an instance-based learning method, also known as a lazy learning algorithm. Instead of concentrating on building a general internal model, it stores all instances in n-dimensional space that correspond to the training data. The classification of this algorithm is to determine new data points on the basis of similarity measures. The advantage of this technique is that it is suited for noisy training data. The accuracy of the algorithm relies on the quality of the quality of the data. On the other hand, the most significant problem is determining the optimal number of neighbors.

- A decision tree is a supervised learning method. This is also suited for classification tasks. Some of the well-known algorithms include CART, C4.5, and ID3.



- The following figure illustrates how to sort the tree down to a few leaf nodes from the root. This is known as an instance, which is also categorized by DT. Instances are classified by examining the attribute defined by the node. Starting at the root node of the tree, work down the branch that corresponds to the attribute values. The most popular criteria include gini for gini impurity and entropy for information gain.

- Random forest is an ensemble classification technique. It is widely used in the field of machine learning. This is suitable for both classification and categorical values. Parallel ensembling is a technique used in this method. This fits multiple decision tree classifiers concurrently. It is illustrated in the subsequent figure. Utilizing sub-samples from various data sets, majority voting or averages are used to determine the outcome. As a result, it can diminish the issue of overfitting and improve prediction accuracy.



3.5. LIMITATION OF THE STUDY

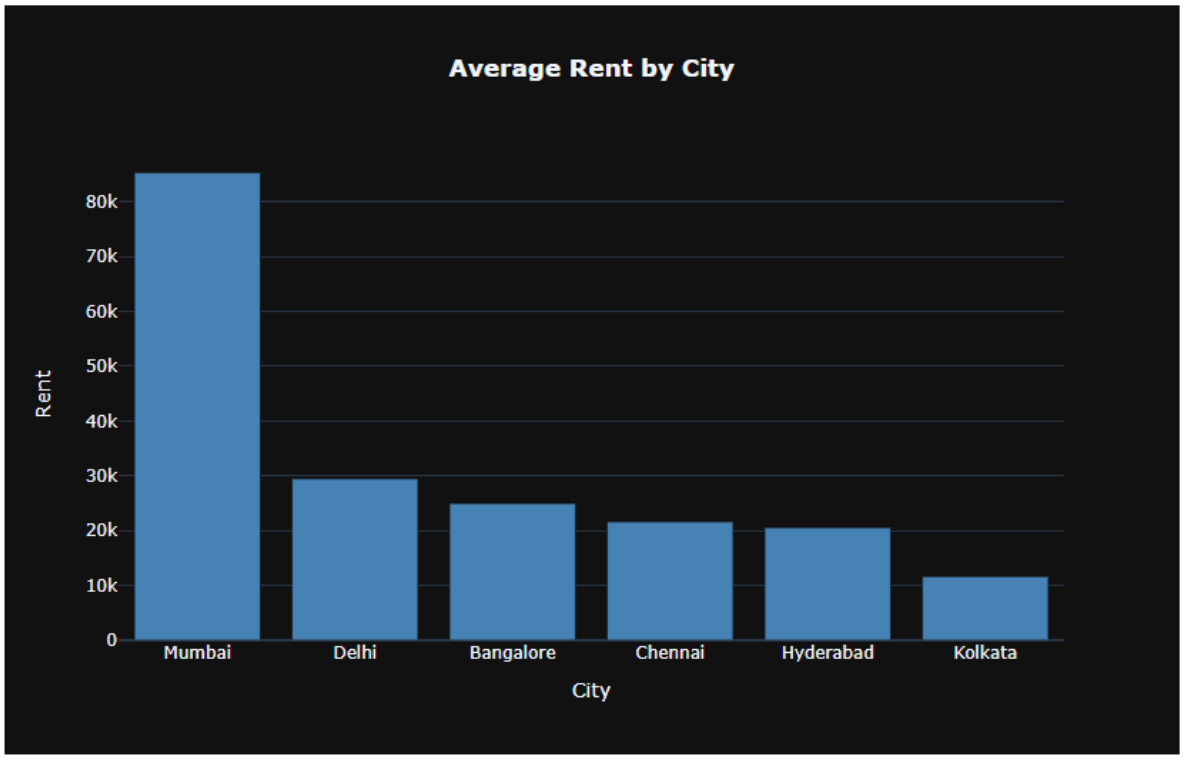
The data is sourced from the open data source, i.e., Kaggle. The determination of a model is based only on house characteristics. The optimisation of the model is based on feature selection, engineering, and tuning methods. The scope of the research is limited to sixteen features. The comparison of model performance is only on the basis of mean absolute error and absolute percentage error. These features may have an explanatory effect on house rental prices. This did not cover all the areas of house rental prices in which ML intervention is possible. This considers the tasks in the rental house price that are suitable for ML techniques. It did not intend to present an exhaustive list of ML techniques. This study fails to explain how each function performs at ML. This study fails to consider the neighbourhood features, including the seaside, schools, recreation, road network, restaurants, cafes, fire stations, crime rate, and police stations. When considering the features for ML, the accuracy is not determined on the basis of macroeconomic factors including the stock market, balance of trade, producer price index, interst rate, housing starts, and other factors.

CHAPTER-IV

DATA ANALYSIS AND INTERPRETATION

4. DATA ANALYSIS AND INTERPRETATION

Figure 4. 1: Average rent by city



The above figure clearly shows the average house rent price by cities. It is found that Mumbai has the highest average house rent price with values exceeding ₹80,000. Delhi follows with a significantly lower average rent, slightly above ₹30,000. Bangalore, Chennai, and Hyderabad have comparable rent averages, all falling between ₹20,000 and ₹30,000. Kolkata has the lowest average rent among the cities, with values just below ₹20,000. Hence it is concluded that Mumbai house rent prices are much higher than those in the other cities, indicating its status as the most expensive city.

Figure 4. 2: Percentage distribution of data Indian cities

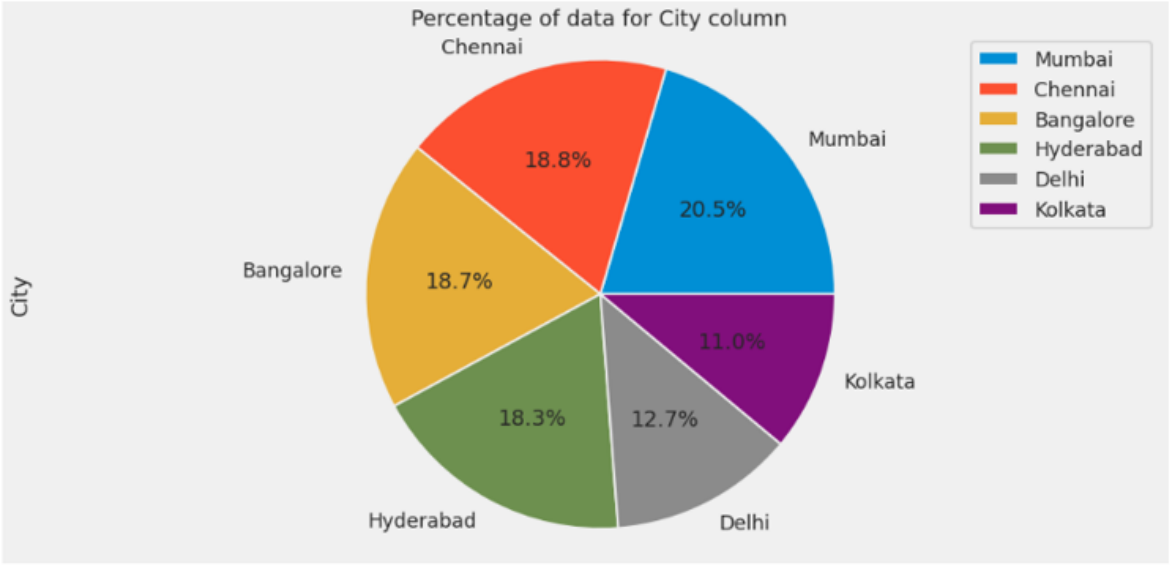


Figure 4.2 showing the percentage distribution of data Indian cities. Mumbai holds the largest percentage, accounting for 20.5% of the data. Chennai comes in second with 18.8%, followed closely by Bangalore at 18.7%. Hyderabad is slightly lower at 18.3%. Delhi accounts for 12.7% of the data. Kolkata holds the smallest portion, representing 11.0% of the data. Thus, Mumbai contributing the highest percentage and Kolkata the lowest.

Figure 4. 3: Impact of city on rent price

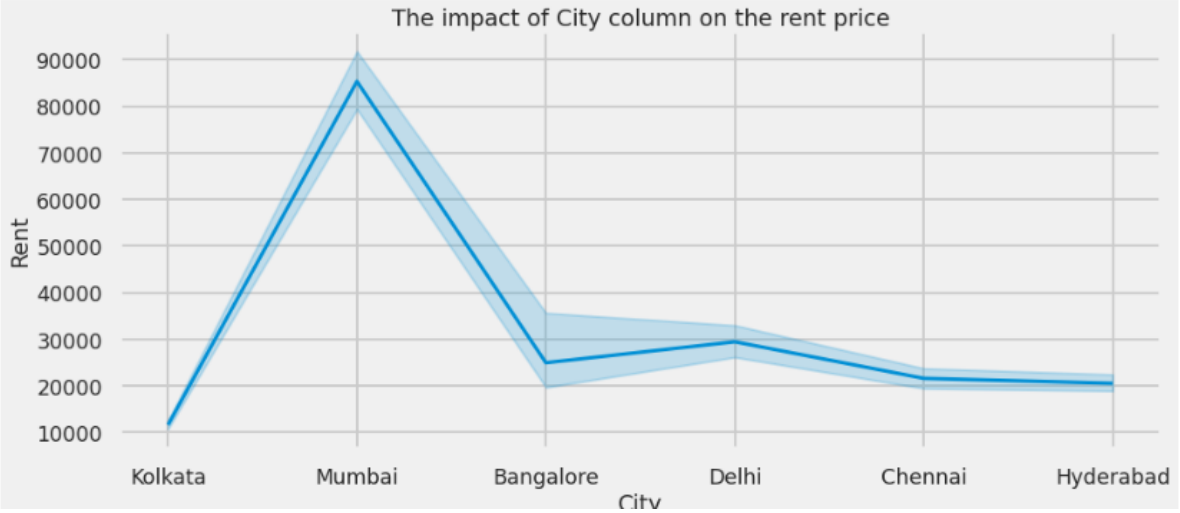
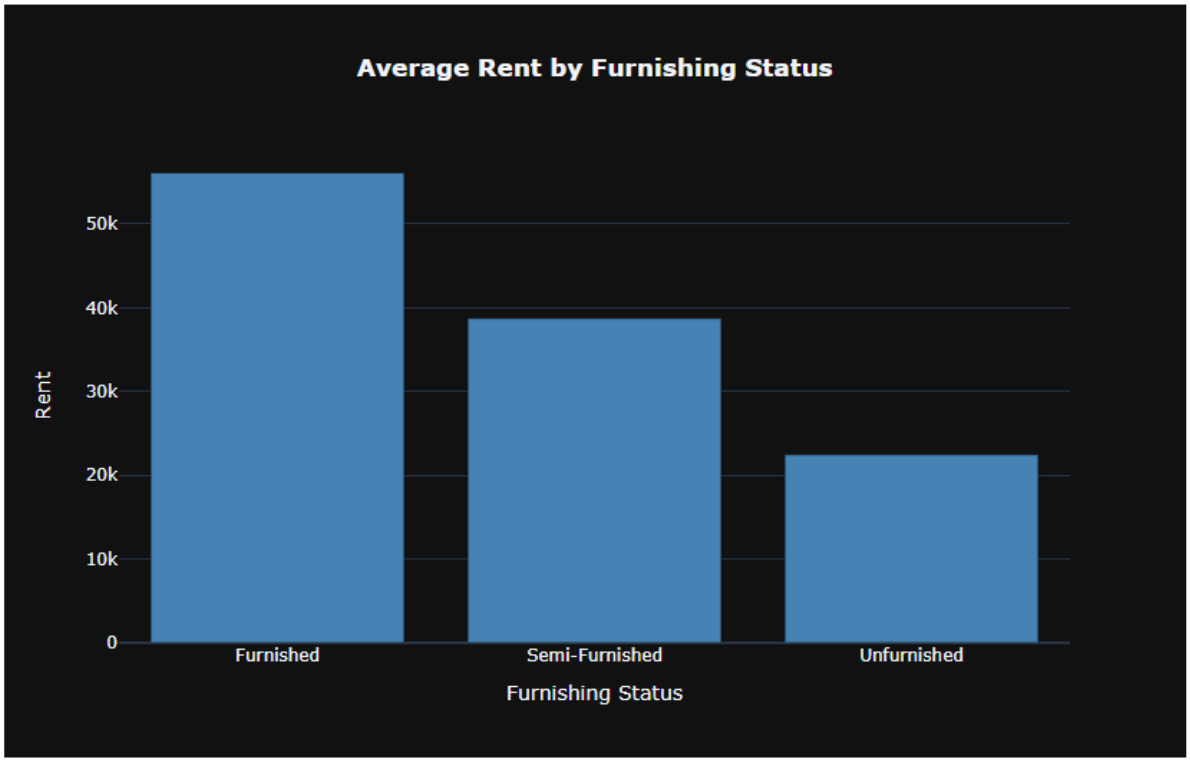


Figure 4.3 shows that Mumbai are highly impact the house rent compared to other cities. While Kolkata is lowly affecting the house rent.

Figure 4. 4: Average rent by furnishing status



The above figure clearly shows the average rent by furnishing status. It is found that furnished has the highest average house rent price with values exceeding ₹50,000. Semi - furnished follows with a significantly lower average rent, slightly above ₹30,000. Unfurnished has the lowest average rent among the cities, with values just above ₹20,000. Hence it is concluded that furnished house has the highest average rented house.

Figure 4. 5: Percentage of data by furnishing status

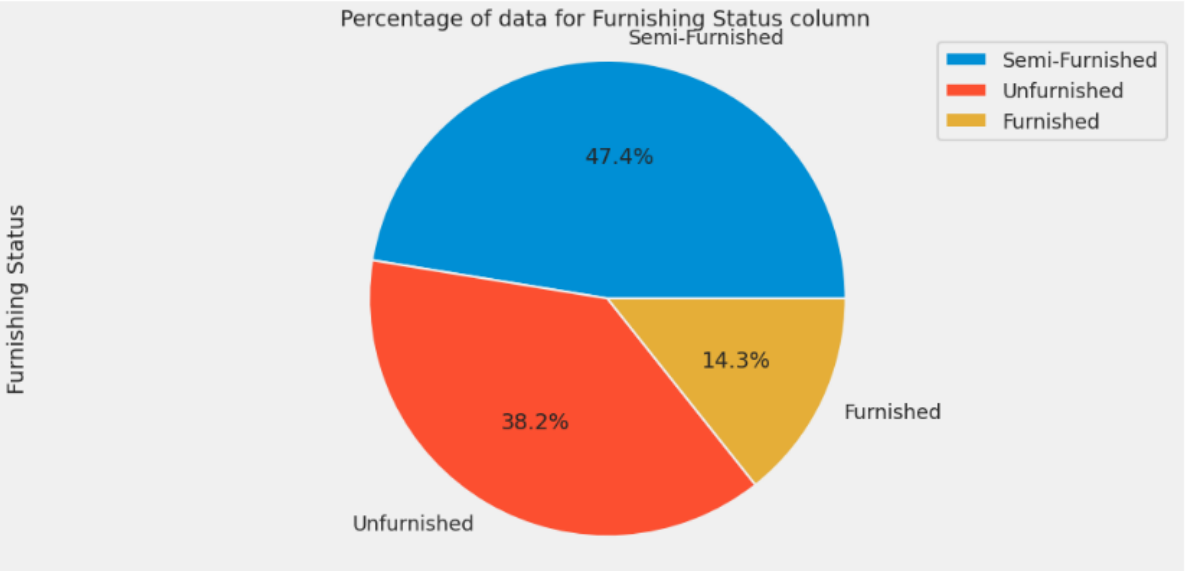
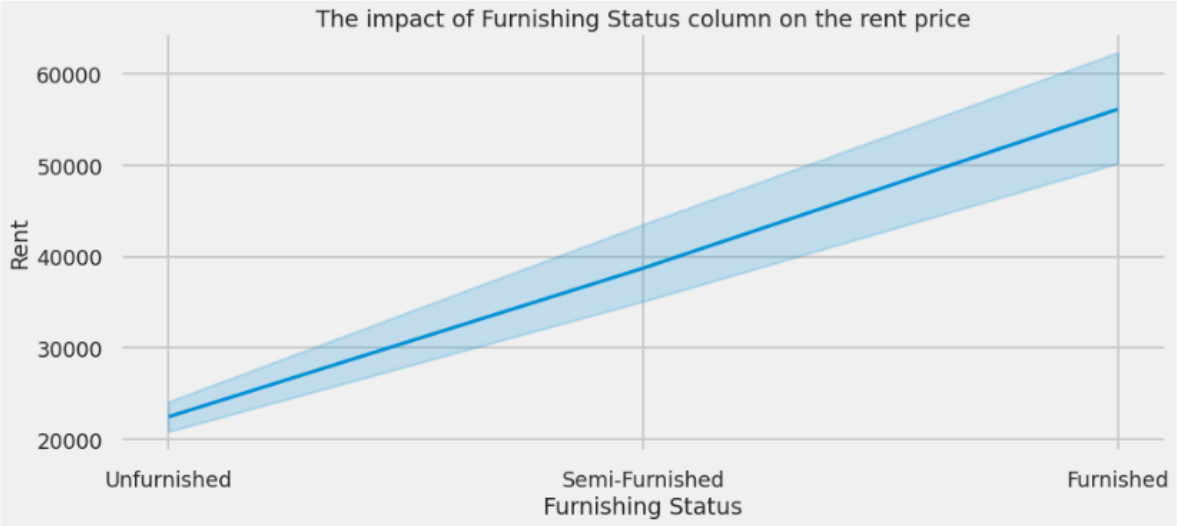


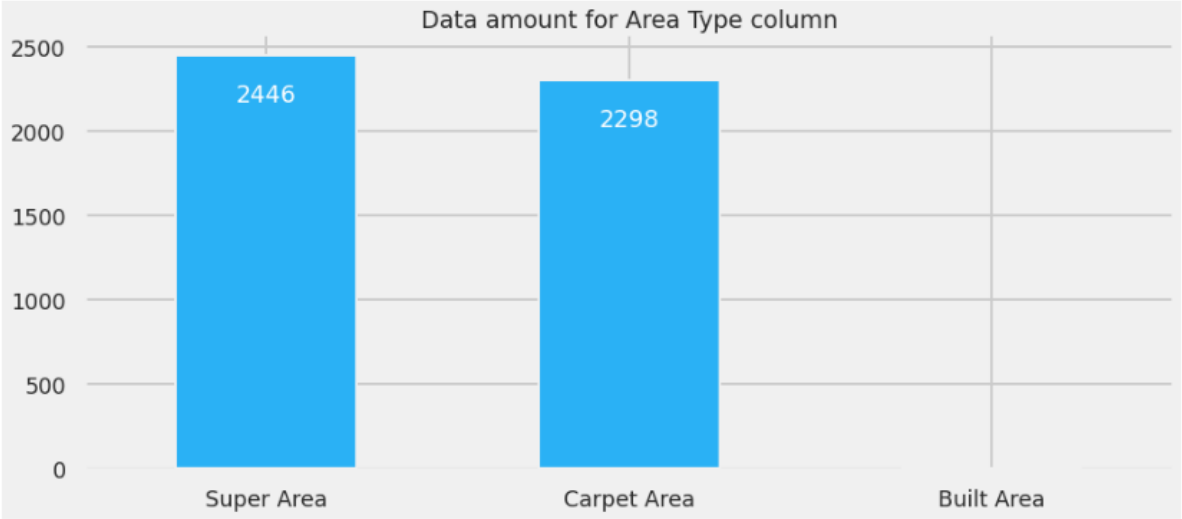
Figure depicts the percentage distribution of data by furnishing status. Semi-furnished house hold the highest percentage accounting for 47.4%. While Unfurnished house hold the second largest percentage with 38.2%. Furnished house holds the smallest portion, representing 14.3% of the data. Thus, Semi-furnished house are highly recommended house.

Figure 4. 6: Impact of furnishing status on rent price



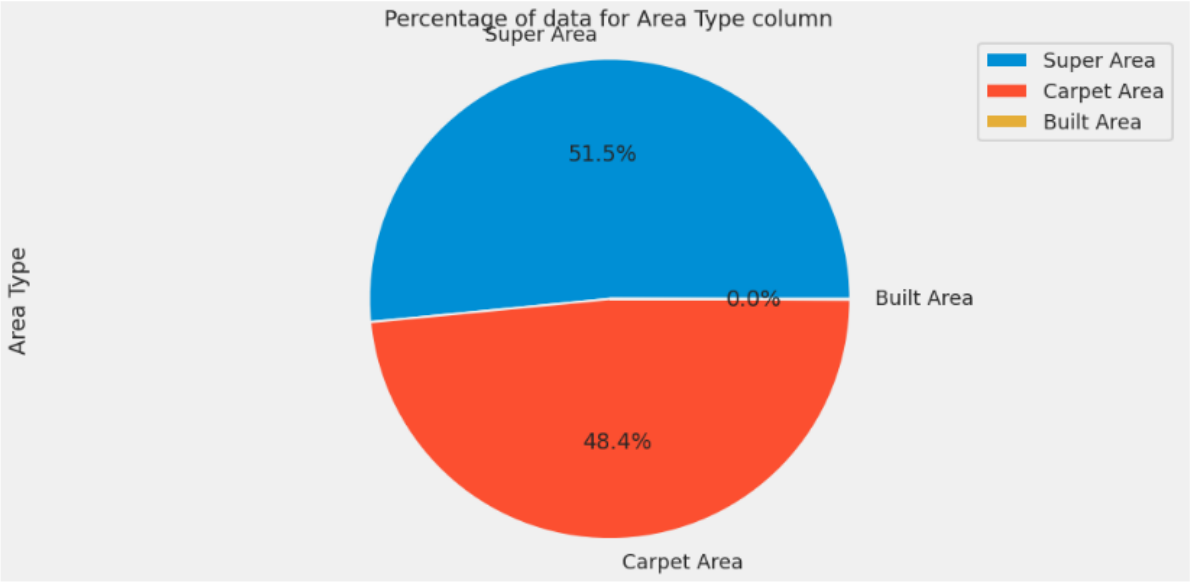
The figure shows that the house is semi-furnished and furnished, then the price goes up.

Figure 4. 7: Area type



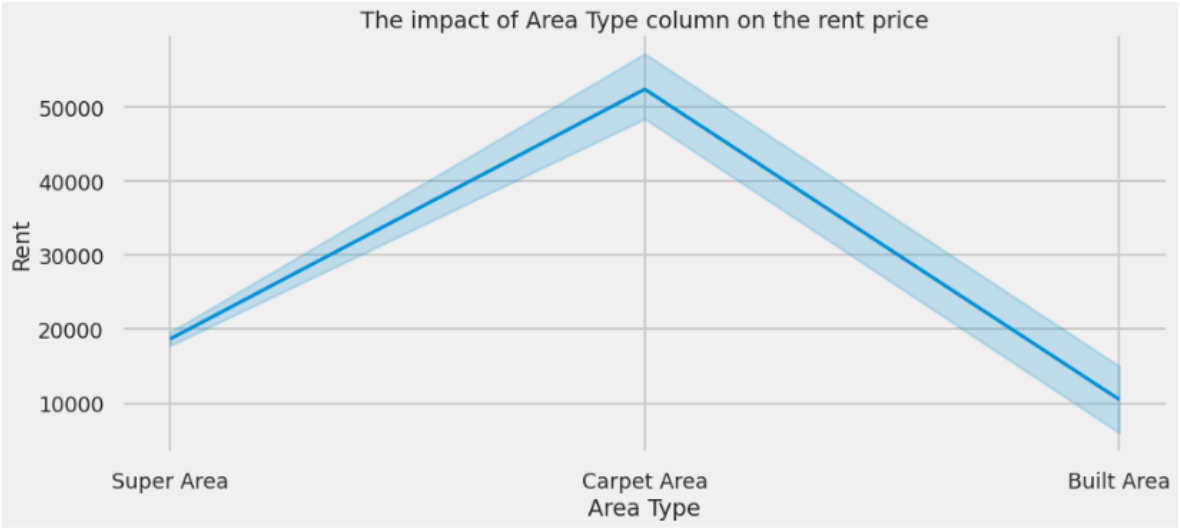
The above figure clearly shows the data amount for area type. It is found that super area has the highest average house with 2446 sq.ft. Carpet area follows with a significantly lower average rent, slightly with 2298 sq.ft. Hence it is concluded that super area has the highest average rented house.

Figure 4. 8: Percentage of data for Area type



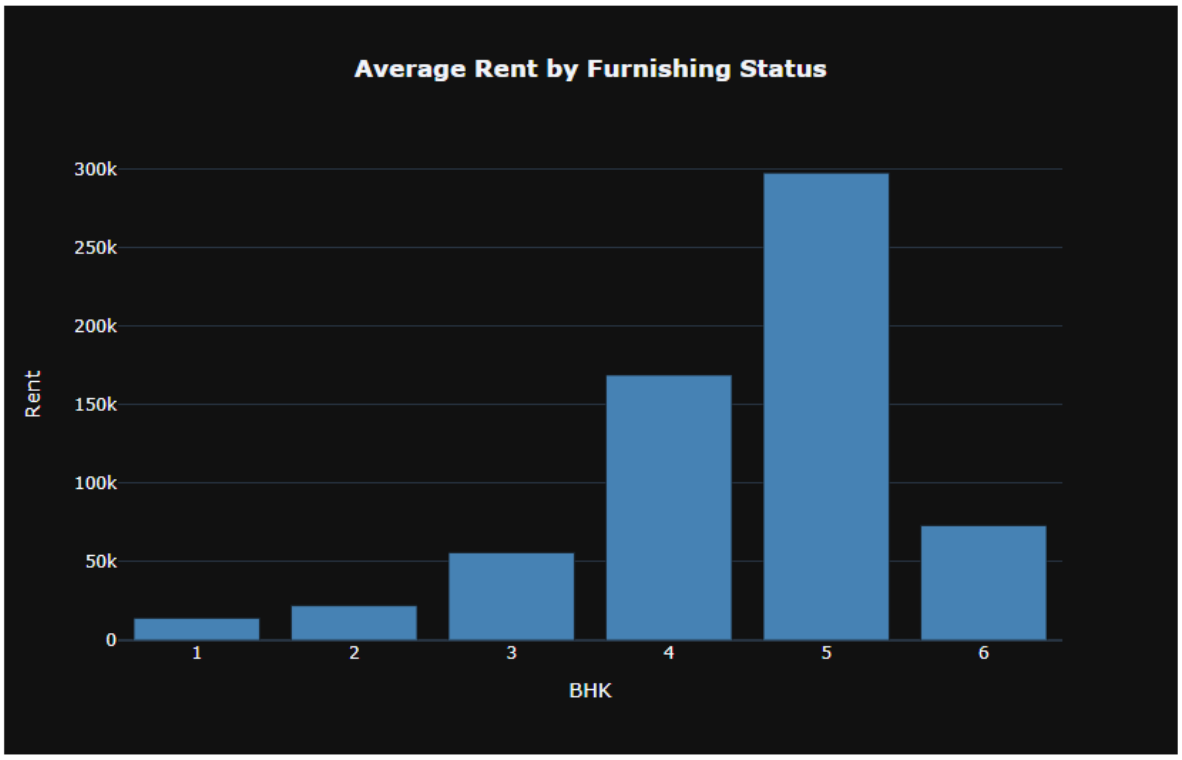
The figure shows that half of the house area type was super area with 51.5%, and the remaining were indicating the carpet area with 48.4%. It is then concluded that Size of the Houses calculated on Super Area or Carpet Area.

Figure 4. 9: Impact of area type on rent price



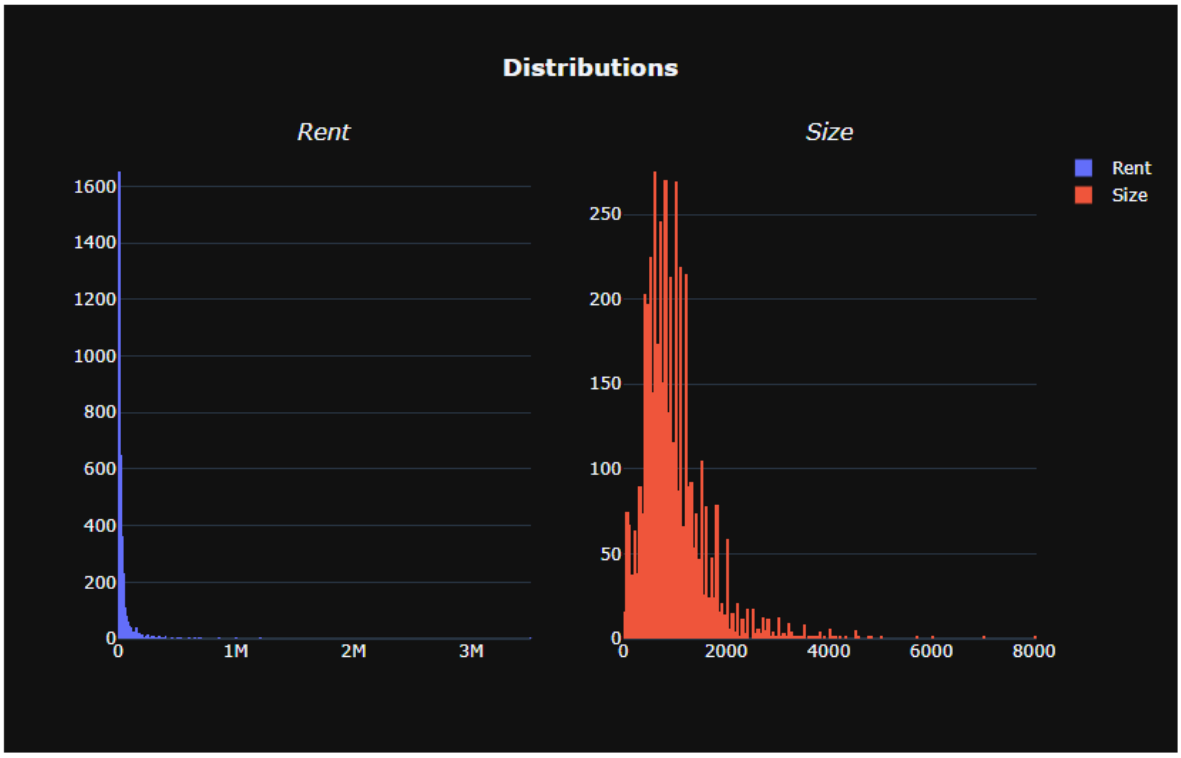
The above figure clearly shows that the high house rent price is more differed by carpet area than the super and Built area.

Figure 4. 10: Average rent by furnishing status



The above figure clearly shows the average rent by furnishing status. It is found that Five BHK has the highest average house rent price with values nearly 300K. Followed by 4 BHK has the average house rent reached approximately 200K, 6 BHK has the house rent reached nearly 100K, 3BHK has rent approximately 75K, and the least rent for one two BHK.

Figure 4. 11: Distribution between rent and Size



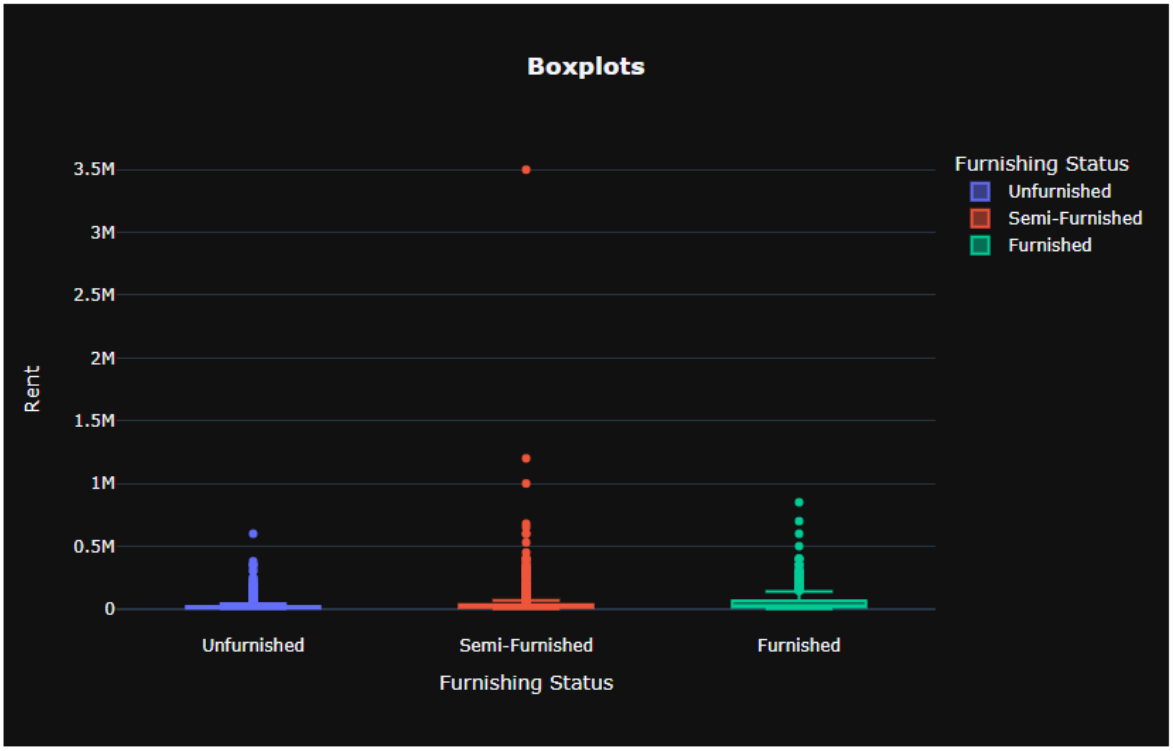
The figure shows two distribution graphs side by side, one for Rent and the other for Size of properties.

The Rent distribution appears heavily skewed to the right. Most rent values are concentrated between lower values (close to zero) and taper off quickly as the rent increases. A few extreme outliers with very high rent values (above 1 million) are visible, but these occur with very low frequency.

The Size distribution also shows a right-skewed pattern, though less extreme compared to the rent distribution. The majority of properties fall between 0 and around 2000 square feet, with a sharp decline in frequency for properties larger than 2000 sq. ft. A few properties have larger sizes extending up to 8000 sq. ft., but they occur rarely.

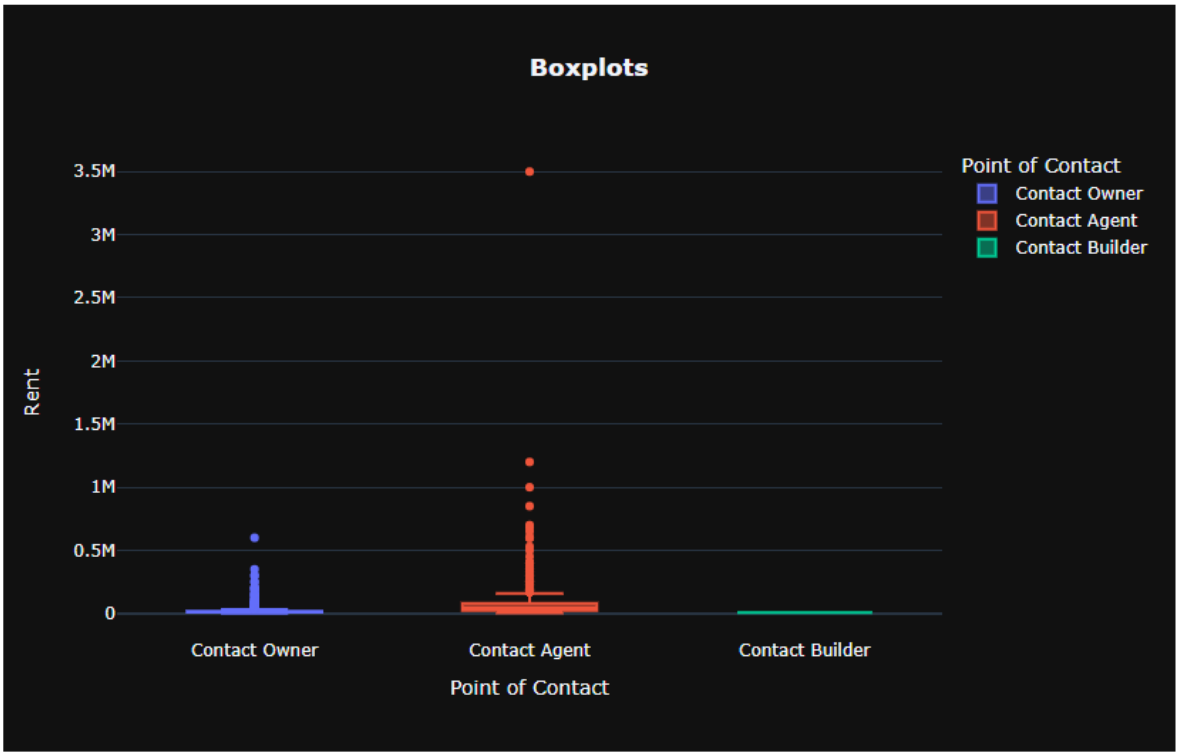
In summary, both rent and size distributions are right-skewed, with most properties having lower rent and size, while a small number of high-rent or large-sized properties exist as outliers.

Figure 4. 12: Boxplots of rent and furnishing status



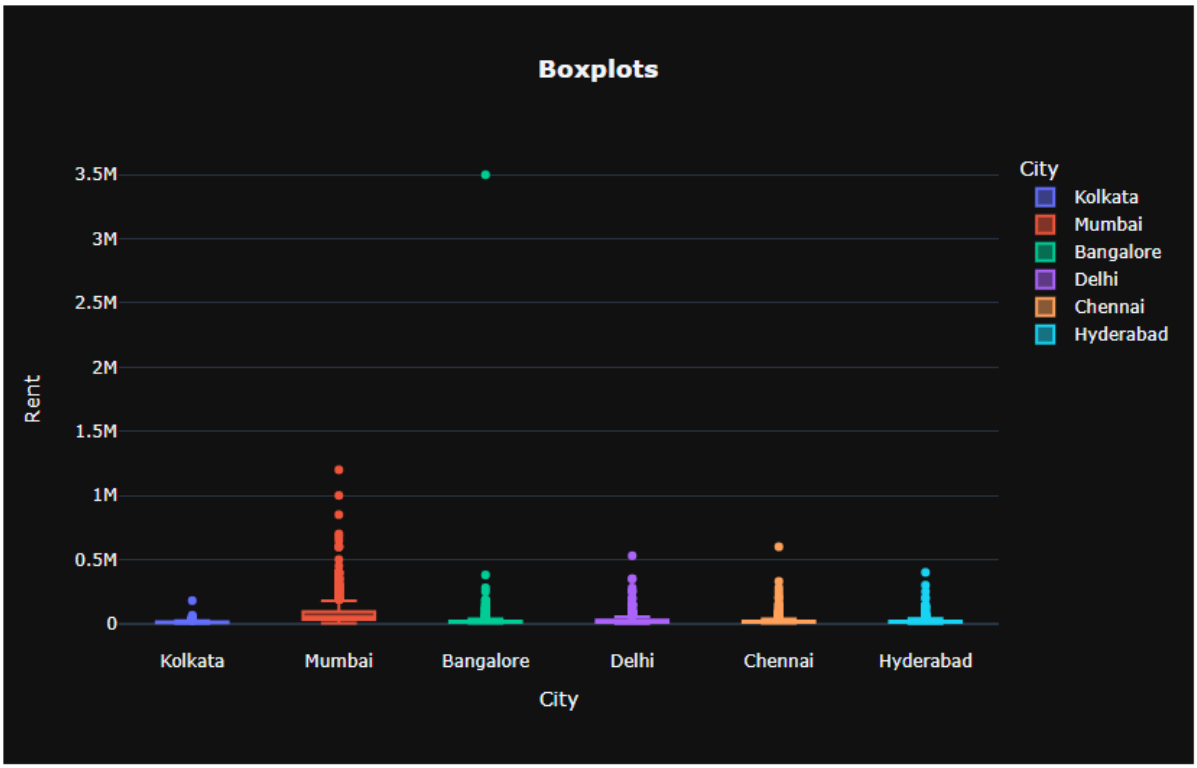
The boxplot provides a comparative view of the rent distribution across different furnishing statuses: Unfurnished, Semi-Furnished, and Furnished. The majority of rents for unfurnished properties are clustered around a low range, with a few outliers extending beyond 500K. Similar to the unfurnished category, rents are generally low, but there are more frequent and higher outliers, some reaching 1M. Furnished properties also have a low rent concentration in the core, but show a few extreme outliers, some going beyond 3M. The boxplots show several outliers, especially for Semi-Furnished and Furnished properties, indicating that some properties in these categories have significantly higher rent. Most properties, regardless of furnishing status, seem to have rents concentrated in the lower ranges. Semi-furnished and furnished properties exhibit higher outlier rents, indicating that certain high-end listings are present in these categories. Unfurnished houses for rent are having less rent as compared to others.

Figure 4. 13: Boxplots of rent and point of contact



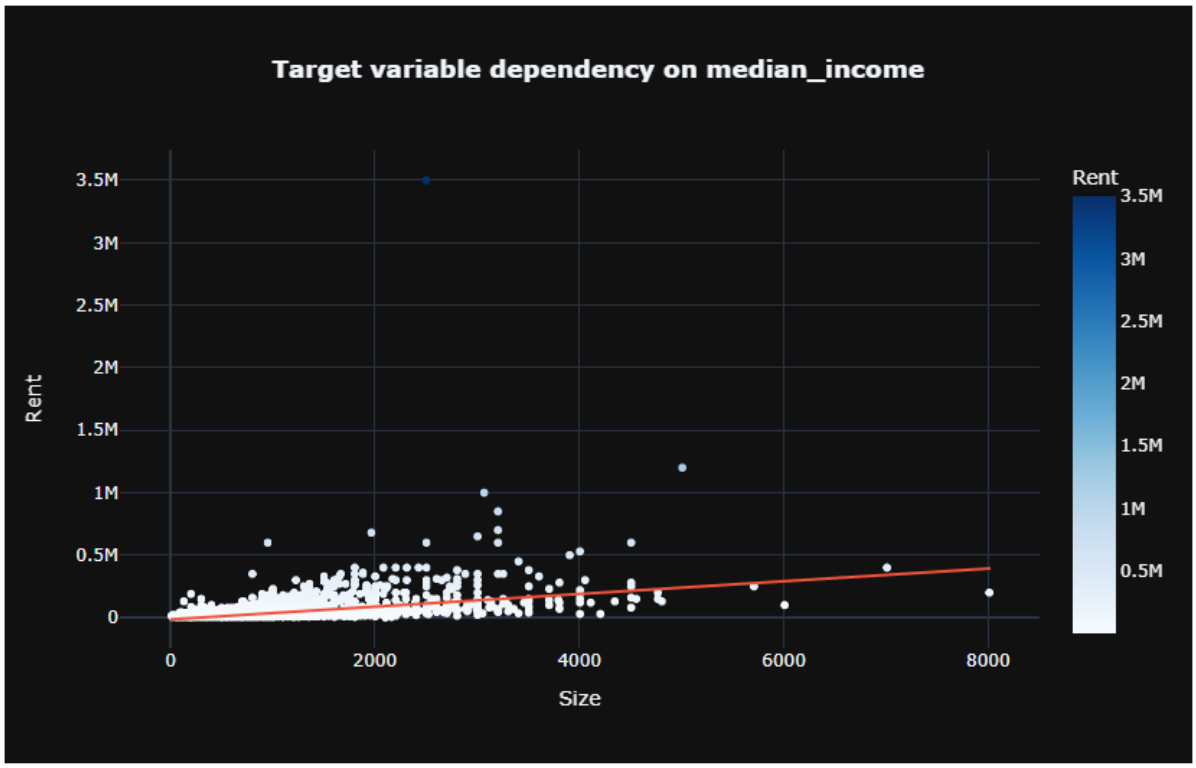
This boxplot visualizes the rent distribution based on the point of contact. The higher rent for properties obtained through a contact agent may reflect additional costs to compensate the agent's services. In contrast, renting directly from builders tends to be less expensive.

Figure 4. 14: Boxplots of rent and city



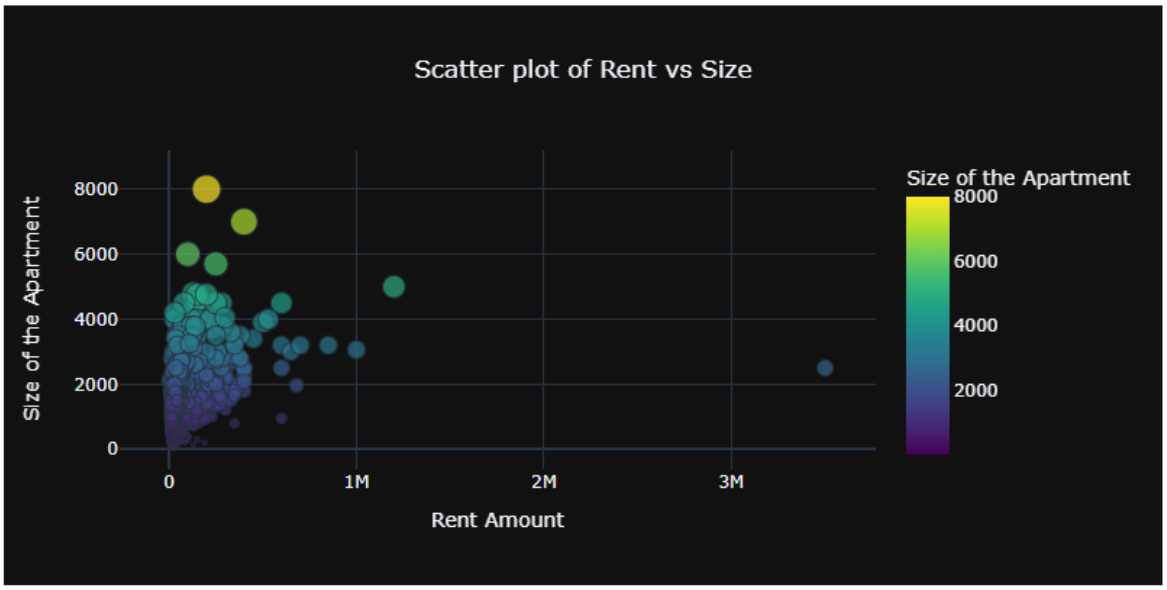
Mumbai has a high demand for housing, leading to higher rents, likely due to the high influx of job seekers and corporate relocations. Other cities, except Kolkata, have relatively equal rent levels. Kolkata has lower rents, which can be attributed to its comparatively less developed job sectors and lifestyle, resulting in lower demand for rental properties.

Figure 4. 15: Target variable based on income



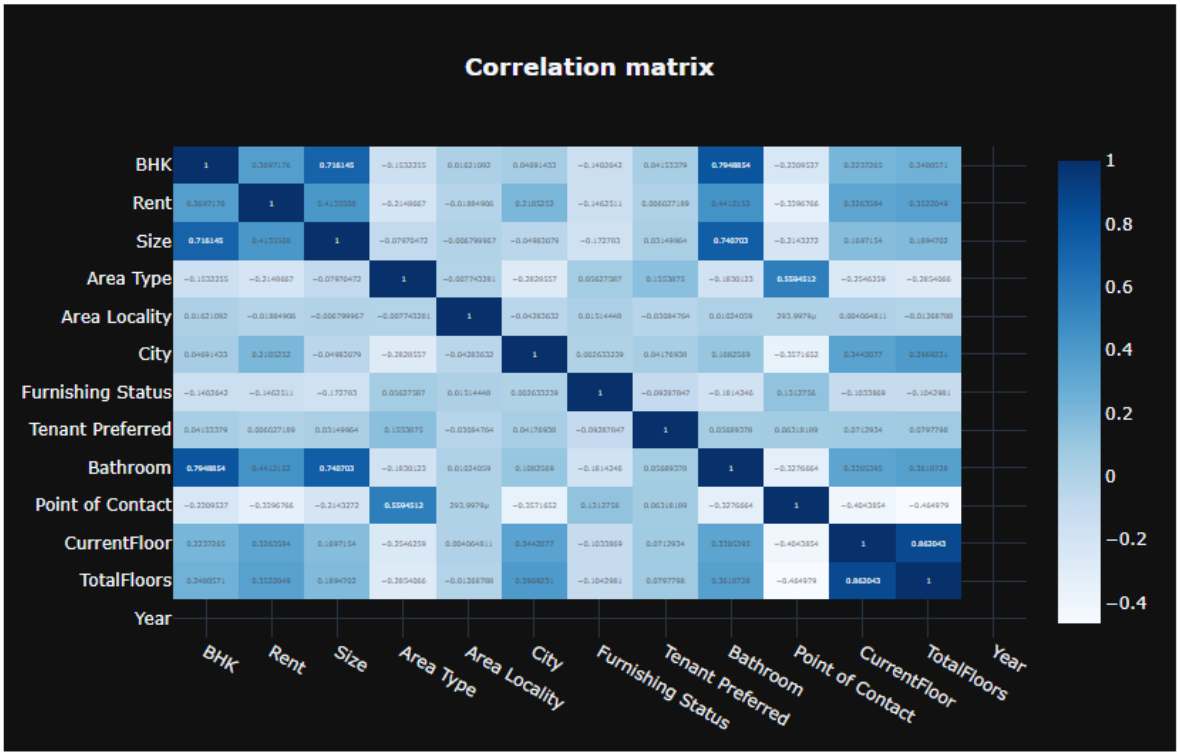
The scatter plot illustrates the relationship between property size and rent, showing a modest positive correlation between the two variables. As property size increases, rent also tends to rise, but the trendline is relatively flat, indicating that size alone does not have a strong impact on rent. Most properties, especially those under 2000 square units, are clustered around lower rent values, with most rents falling below 500K. As the size increases beyond 2000, there is more variation in rent, although few properties command extremely high rents, with some outliers reaching up to 3.5 million. These outliers suggest that certain large or high-end properties are exceptions, with much higher rents. Overall, while larger properties generally cost more, factors beyond size likely play a significant role in determining rent prices. This chart indicates that while there is a positive correlation between property size and rent, the relationship is not particularly strong. Most properties, even those larger in size, tend to cluster in lower rent ranges, suggesting that factors other than size (e.g., location, furnishing status, etc.) may significantly influence rent prices. The high outliers suggest the presence of premium properties where size may contribute to much higher rent values.

Figure 4. 16: Scatter plot of Rent vs Size



The scatter plot titled "Rent vs Size" illustrates the relationship between the rent amount and the size of apartments. The x-axis represents the rent amount, ranging from 0 to over 3 million, while the y-axis represents the size of the apartment in square feet, going up to 8000 square feet. Each data point is depicted as a circle, and the size of the circles represents the size of the apartments. Larger circles correspond to larger apartment sizes. A color gradient, ranging from purple to yellow, is used to show the apartment size, with yellow indicating larger apartments (closer to 8000 sq. ft.) and purple indicating smaller ones. Most of the data points are concentrated in the lower range of both rent and size, suggesting that most apartments in the dataset are relatively small and have lower rents. A few points are scattered toward higher rent and size, indicating some larger, more expensive apartments, but these are fewer in number.

Figure 4. 17: Correlation between the house rent variable and rent



The image presents a correlation matrix, which displays the relationships between various variables involved in a real estate dataset. Each cell in the matrix contains a value between -1 and 1, indicating the strength and direction of the correlation between two variables. The color gradient, ranging from light blue to dark blue, represents the correlation values, where darker shades correspond to stronger positive correlations, and lighter shades indicate weaker or negative correlations.

Rent and Size also have a moderate positive correlation (0.44), suggesting that larger apartments tend to have higher rent prices. Rent and BHK have a positive correlation (0.36), showing that more bedrooms generally lead to higher rents. Size and Bathroom have a strong positive correlation (0.74), indicating that larger apartments tend to have more bathrooms. Point of Contact and Current Floor have a moderate positive correlation (0.34), but its real-world interpretation is unclear without further context. Area Type, City, and Area Locality exhibit weak to negative correlations with most other variables, indicating that these variables are not strongly related to rent or apartment size. Year and Total Floors show a moderate positive correlation (0.48), meaning newer buildings tend to have more floors.

Overall, variables such as BHK, Size, and Bathroom are highly correlated, while Rent shows a moderate relationship with both Size and BHK, indicating these are key factors that influence rent prices. Other features such as Furnishing Status and Tenant Preferred appear to have weaker or insignificant correlations with rent or apartment size. Area type is the least dependent upon target variable (Rent)

Figure 4. 18: Model result

```
Tree RMSE: 43531.73929056804
Tree R-squared: 0.4660780572655129

KNN RMSE: 48831.390488732075
KNN R-squared: 0.3281630081901137

Random Forest RMSE: 32265.721872206814
Random Forest R-squared: 0.7066753897581715
```

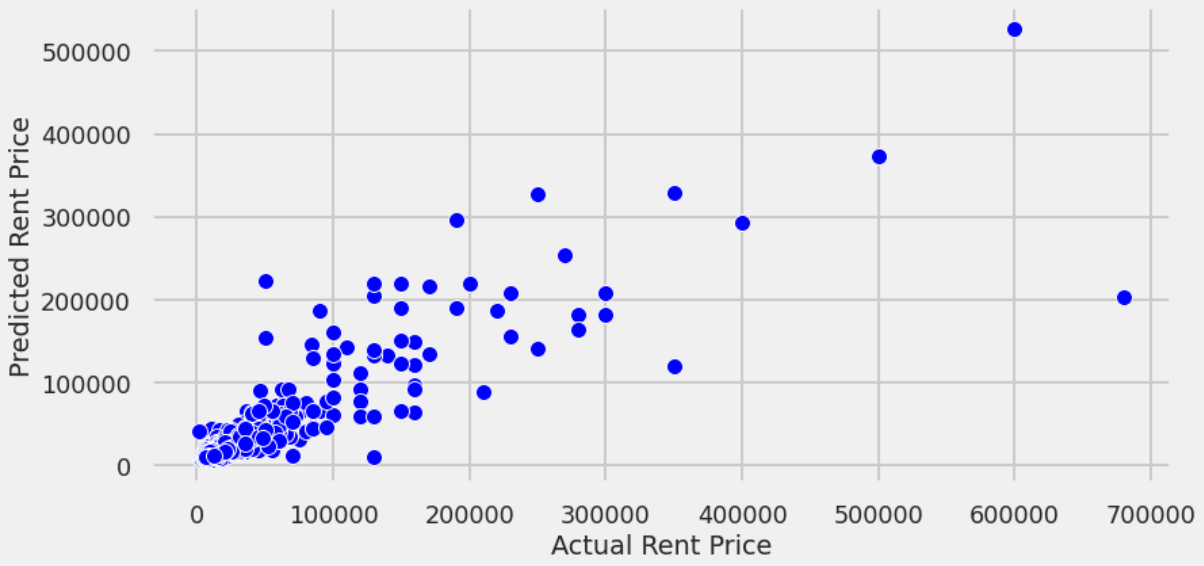
At first glance, the RMSE value appears quite large, but considering the rent range—where the minimum value is 1,200 and the maximum reaches 1,200,000 (a difference of three orders of magnitude)—these RMSE values are understandable. The comparison of the KNN and Decision Tree models shows similar performance, with Decision Tree performing slightly better due to its higher r^2 score and lower RMSE. However, the Random Forest Regressor stands out as the best model among the three. It demonstrates a significantly lower RMSE (by over 10,000) and a considerably higher r^2 score (almost 30% higher than the Decision Tree). Therefore, we can confidently conclude that Random Forests are the superior model for predicting house rents and will proceed with this model for further analysis.

Figure 4. 19: Prediction of House rent price on cities

	Rent Price	Predicted Rent Price
count	712.0	712.0
mean	34084.7	32201.7
std	59791.6	48179.6
min	1200.0	5811.5
25%	10000.0	11398.4
50%	15750.0	16065.0
75%	32000.0	31146.4
max	680000.0	526398.2

MAE for Predicted 10962.8

r2_score for Predicted 0.76



This scatter plot compares the actual rent price with the predicted rent price. Each blue point represents a data pair: one actual rent price and the corresponding predicted value. Most of the points are clustered near the lower end of both axes, meaning that the model performs well for predicting lower rent values (below 100,000). The spread increases as the actual rent price rises, indicating that the model's predictions become more varied or less accurate for higher rents. A few points are far from the diagonal line that would represent perfect predictions, particularly at higher rent prices, which suggests that the predictions deviate significantly from the actual values at the upper end of the rent spectrum.

CHAPTER V

FINDINGS, RECOMMENDATIONS AND CONCLUSIONS

5. FINDINGS, RECOMMENDATIONS AND CONCLUSIONS

5.1. FINDINGS

- It is found that Mumbai has the highest average house rent price with values exceeding ₹80,000. Delhi follows with a significantly lower average rent, slightly above ₹30,000. Bangalore, Chennai, and Hyderabad have comparable rent averages, all falling between ₹20,000 and ₹30,000. Kolkata has the lowest average rent among the cities, with values just below ₹20,000. Hence it is concluded that Mumbai house rent prices are much higher than those in the other cities, indicating its status as the most expensive city.

- Mumbai holds the largest percentage, accounting for 20.5% of the data. Chennai comes in second with 18.8%, followed closely by Bangalore at 18.7%. Hyderabad is slightly lower at 18.3%. Delhi accounts for 12.7% of the data. Kolkata holds the smallest portion, representing 11.0% of the data. Thus, Mumbai contributing the highest percentage and Kolkata the lowest.

- Mumbai are highly impact the house rent compared to other cities. While Kolkata is lowly affecting the house rent.

- It is found that furnished has the highest average house rent price with values exceeding ₹50,000. Semi - furnished follows with a significantly lower average rent, slightly above ₹30,000. Unfurnished has the lowest average rent among the cities, with values just above ₹20,000. Hence it is concluded that furnished house has the highest average rented house.

- Semi-furnished house hold the highest percentage accounting for 47.4%. While Unfurnished house hold the second largest percentage with 38.2%. Furnished house holds the smallest portion, representing 14.3% of the data. Thus, Semi-furnished house are highly recommended house.

- It is found that super area has the highest average house with 2446 sq.ft. Carpet area follows with a significantly lower average rent, slightly with 2298 sq.ft. Hence it is concluded that super area has the highest average rented house.

- half of the house area type was super area with 51.5%, and the remaining were indicating the carpet area with 48.4%. It is then concluded that Size of the Houses calculated on Super Area or Carpet Area.
- high house rent price is more differed by carpet area than the super and Built area.
- It is found that Five BHK has the highest average house rent price with values nearly 300K. Followed by 4 BHK has the average house rent reached approximately 200K, 6 BHK has the house rent reached nearly 100K, 3BHK has rent approximately 75K, and the least rent for one two BHK.
- both rent and size distributions are right-skewed, with most properties having lower rent and size, while a small number of high-rent or large-sized properties exist as outliers.
- Unfurnished houses for rent are having less rent as compared to others.
- The higher rent for properties obtained through a contact agent may reflect additional costs to compensate the agent's services. In contrast, renting directly from builders tends to be less expensive.
- Mumbai has a high demand for housing, leading to higher rents, likely due to the high influx of job seekers and corporate relocations. Other cities, except Kolkata, have relatively equal rent levels. Kolkata has lower rents, which can be attributed to its comparatively less developed job sectors and lifestyle, resulting in lower demand for rental properties.
- Overall, while larger properties generally cost more, factors beyond size likely play a significant role in determining rent prices. This chart indicates that while there is a positive correlation between property size and rent, the relationship is not particularly strong. Most properties, even those larger in size, tend to cluster in lower rent ranges, suggesting that factors other than size (e.g., location, furnishing status, etc.) may significantly influence rent prices. The high outliers suggest the presence of premium properties where size may contribute to much higher rent values.
- Most of the data points are concentrated in the lower range of both rent and size, suggesting that most apartments in the dataset are relatively small and have lower rents.

A few points are scattered toward higher rent and size, indicating some larger, more expensive apartments, but these are fewer in number.

- Overall, variables such as BHK, Size, and Bathroom are highly correlated, while Rent shows a moderate relationship with both Size and BHK, indicating these are key factors that influence rent prices. Other features such as Furnishing Status and Tenant Preferred appear to have weaker or insignificant correlations with rent or apartment size. Area type is the least dependent upon target variable (Rent)
- Random Forests are the superior model for predicting house rents and will proceed with this model.
- Most of the points are clustered near the lower end of both axes, meaning that the model performs well for predicting lower rent values (below 100,000). The spread increases as the actual rent price rises, indicating that the model's predictions become more varied or less accurate for higher rents. A few points are far from the diagonal line that would represent perfect predictions, particularly at higher rent prices, which suggests that the predictions deviate significantly from the actual values at the upper end of the rent spectrum.

5.2. RECOMMENDATIONS BASED ON FINDINGS

- Prioritize Mumbai's Rental Market: Given Mumbai's highest average rent, exceeding ₹80,000, investors and developers should focus on this city for premium property offerings and target high-demand housing markets.

- Promote Furnished and Semi-Furnished Properties: Furnished homes command the highest rents, while semi-furnished units represent the largest market share. Property owners should prioritize these categories to maximize rental income and meet tenant demand.

- Invest in Larger Super Area Properties: Super area properties attract higher rents and should be emphasized in marketing luxury properties. Investors should focus on developing and promoting homes with larger super areas to capture higher-end renters.

- Tailor Offerings Based on BHK Configurations: 5 BHK units have the highest rents, followed by 4 and 6 BHK. However, more accessible options like 2 BHK and 3 BHK should be prioritized to cater to most renters, especially in cities like Bangalore and Chennai.

- Target Affordable Rental Market in Kolkata: With the lowest average rents, Kolkata presents an opportunity for affordable housing developments. Developers should focus on budget-friendly projects to attract renters in this market.

- Focus on Location and Amenities: While size correlates with rent, location, furnishing, and amenities play crucial roles in rent determination. Property owners should optimize these aspects to command higher rents, especially for smaller or mid-sized units.

- Improve Prediction Models for High-Rent Properties: Refine machine learning models to improve rent prediction accuracy, particularly for premium properties where current models underperform.

5.3. SCOPE FOR FURTHER RESEARCH

This study has provided valuable insights into house rent prediction using machine learning models, particularly in urban India. However, several areas remain for further exploration and improvement. Future research could expand the dataset to include more cities, particularly smaller urban centers and towns, to gain a broader understanding of rental price trends across India. In terms of methodology, exploring more advanced machine learning models like XGBoost, Ridge Regression, or even deep learning techniques may provide improved predictive accuracy. Hyperparameter tuning and the use of ensemble methods can further optimize model performance. Another potential area of research involves the application of time-series analysis to predict future rent trends based on historical data, which could be beneficial for real estate planning and investment. Finally, integrating a broader range of features, such as environmental factors, public transport accessibility, and future development plans, could provide a more holistic model. This approach may enhance the accuracy and reliability of house rent predictions, offering better tools for urban planners, policymakers, and real estate stakeholders.

5.4. CONCLUSIONS

This study aimed to predict house rent prices in urban India using machine learning algorithms. Through the analysis of various features such as property size, location, number of bedrooms, furnishing status, and BHK configurations, it was found that Random Forest outperformed other models in terms of prediction accuracy. Mumbai consistently showed the highest rent prices, while Kolkata had the lowest, highlighting regional disparities in the rental market. Additionally, the study found that factors such as the number of BHKs, property size, and furnishing status significantly influence rent prices, while others like area type had a lesser impact. Although the model performed well for lower rent values, it showed some variance in predicting higher rent prices, indicating that factors beyond those used in this study could also play a significant role.

The Random Forest model proved to be the most accurate for predicting rents, particularly for lower rent values. However, its predictions become less reliable for higher rents, indicating more variation at the upper end of the spectrum. The data also shows a right-skewed distribution, with most properties being smaller and less expensive, and few high-rent properties significantly influencing the overall rent data. Additional factors such as agent fees contribute to higher rents when renting through agents, whereas direct rentals from builders tend to be more affordable. Overall, while BHK configuration and property size are significant factors, the study highlights the importance of considering other elements like furnishing status and additional costs in predicting rental prices. The findings suggest that machine learning models can be powerful tools in understanding and forecasting rental markets, but further refinement and exploration of additional variables could improve model performance. This study provides a foundation for future research to enhance house rent prediction models and better understand the dynamics of the rental market in urban India.

BIBLIOGRAPHY

- Imran, I., Zaman, U., Waqar, M., & Zaman, A. (2021). Using machine learning algorithms for housing price prediction: the case of Islamabad housing data. *Soft Computing and Machine Intelligence*, *1*(1), 11-23.
- Thamarai, M., & Malarvizhi, S. P. (2020). House Price Prediction Modeling Using Machine Learning. *International Journal of Information Engineering & Electronic Business*, *12*(2).
- Yağmur, A., Kayakuş, M., & Terzioğlu, M. (2022). House price prediction modeling using machine learning techniques: a comparative study. *Aestimum*, *81*.
- Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House price prediction using random forest machine learning technique. *Procedia Computer Science*, *199*, 806-813.
- Zulkifley, N. H., Rahman, S. A., Ubaidullah, N. H., & Ibrahim, I. (2020). House price prediction using a machine learning model: a survey of literature. *International Journal of Modern Education and Computer Science*, *12*(6), 46-54.
- Satish, G. N., Raghavendran, C. V., Rao, M. S., & Srinivasulu, C. (2019). House price prediction using machine learning. *Journal of Innovative Technology and Exploring Engineering*, *8*(9), 717-722.
- Awonaike, A. (2022). *Estimating UK house prices using machine learning* (Doctoral dissertation, University of East London).
- El Mouna, L., Silkan, H., Haynf, Y., Nann, M. F., & Tekouabou, S. C. (2023). A comparative study of urban house price prediction using machine learning algorithms. In *E3S Web of Conferences* (Vol. 418, p. 03001). EDP Sciences.
- Abebe, K., & Patil, P. V. (2021). HOUSING PRICE FORECASTING USING MACHINE LEARNING ALGORITHMS (IN CASE OF REAL STATES IN BANGALORE CITY).
- Yalgudkar, S. S., & Dharwadkar, N. V. A Literature Survey on Housing Price Prediction.

- Henriksson, E., & Werlinder, K. (2021). Housing Price Prediction over Countrywide Data: A comparison of XGBoost and Random Forest regressor models.
- Durganjali, P., & Pujitha, M. V. (2019, March). House resale price prediction using classification algorithms. In *2019 International Conference on Smart Structures and Systems (ICSSS)* (pp. 1-4). IEEE.
- Sawant, R., Jangid, Y., Tiwari, T., Jain, S., & Gupta, A. (2018, August). Comprehensive analysis of housing price prediction in pune using multi-featured random forest approach. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)* (pp. 1-5). IEEE.
- Wang, P. Y., Chen, C. T., Su, J. W., Wang, T. Y., & Huang, S. H. (2021). Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism. *IEEE access*, *9*, 55244-55259.
- Lim, W. T., Wang, L., Wang, Y., & Chang, Q. (2016, August). Housing price prediction using neural networks. In *2016 12th International conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD)* (pp. 518-522). IEEE.
- Piao, Y., Chen, A., & Shang, Z. (2019, August). Housing price prediction based on CNN. In *2019 9th international conference on information science and technology (ICIST)* (pp. 491-495). IEEE.
- Madhuri, C. R., Anuradha, G., & Pujitha, M. V. (2019, March). House price prediction using regression techniques: A comparative study. In *2019 International conference on smart structures and systems (ICSSS)* (pp. 1-5). IEEE.
- Shinde, N., & Gawande, K. (2018). Valuation of house prices using predictive techniques. *Journal of Advances in Electronics Computer Science*, *5*(6), 34-40.
- Dagar, A., & Kapoor, S. (2020). A Comparative Study on House Price Prediction. *International Journal for Modern Trends in Science and Technology*, *6*(12), 103-107.
- Jha, S. B., Pandey, V., Jha, R. K., & Babiceanu, R. F. (2020). Machine learning approaches to real estate market prediction problem: a case study. *arXiv preprint arXiv:2008.09922*.

- Hjort, A., Pensar, J., Scheel, I., & Sommervoll, D. E. (2022). House price prediction with gradient boosted trees under different loss functions. *Journal of Property Research*, *39*(4), 338-364.
- Ho, W. K., Tang, B. S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, *38*(1), 48-70.
- Zou, C. (2023). The House Price Prediction Using Machine Learning Algorithm: The Case of Jinan, China. *Highlights in Science, Engineering and Technology*, *39*, 327-333.
- Alfiyatin, A. N., Febrita, R. E., Taufiq, H., & Mahmudy, W. F. (2017). Modeling house price prediction using regression analysis and particle swarm optimization case study: Malang, East Java, Indonesia. *International Journal of Advanced Computer Science and Applications*, *8*(10).
- Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2021). Understanding house price appreciation using multi-source big geo-data and machine learning. *Land use policy*, *111*, 104919.

PLAGIARISM REPORT

ORIGINALITY REPORT

11%	4%	5%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Lale El Mouna, Hassan Silkan, Youssef Haynf, Mohamedade Farouk Nann, Stéphane C. K. Tekouabou. "A Comparative Study of Urban House Price Prediction using Machine Learning Algorithms", E3S Web of Conferences, 2023 Publication	3%
2	www.ibef.org Internet Source	2%
3	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	1%
4	Submitted to University of East London Student Paper	<1%
5	Submitted to University of Hertfordshire Student Paper	<1%
6	Submitted to Liverpool John Moores University Student Paper	<1%
7	Submitted to Riga Technical University Student Paper	