Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-03244-6 OPEN

Advanced modeling of housing locations in the city of Tehran using machine learning and data mining techniques

Ali Asghar Pilehvar[1]✉ & Arian Ghasemi[2]

This research delves into the intricate dynamics of housing location in the bustling metropolis of Tehran. It aims to gain a deeper understanding of the factors influencing housing prices across the city. Employing a descriptive-analytical method, the study utilizes the Python programming language and its libraries, along with various regression models, to analyze a comprehensive dataset of 8000 villas and apartments spread across 22 districts and 317 areas. Data obtained from official sources are used to examine the correlation between housing prices and nine key determinants. The findings reveal strong positive correlations between the total value of the houses and several factors: surface area (80%), neighborhood location (75%), presence of an elevator (44%), presence of a parking lot (43%), and year of construction (26%), these demonstrate the importance of area and neighborhood. Conversely, the distinct number shows an inverse correlation (−41%) which means the higher the distinct number is, the lower the total value will be. In its final stage, the study employs cross-validation to evaluate the performance of various learning models, achieving a maximum accuracy of 85%. The research concludes by presenting a new formulation and modeling approach for determining the total value of housing, showcasing its originality and contributions to the field.

[1] University of Bojnord, Bojnord, Iran. [2] Kharazmi University, Tehran, Iran. ✉email: pilevar@ub.ac.ir

Introduction

A widening economic and social class gap has posed a significant challenge for housing provision from a materialistic (building a shelter) and spiritual (creating peace, security, and a sense of belonging) perspective (Pilehvar, 2020). In this process, the role of the economic challenge is more significant than that of social and environmental variables (Li, 2021). Since housing is closely tied to the identity of people (Naghizadeh, 2017), serving as a critical factor in securing citizenship rights (Zarghamifard, 2019), how urban living spaces are located and identified and at the same time accounting for quantitative and qualitative dimensions of the housing are among crucial issues in ecological research concerned with the urban system (Pilehvar, 2022), which provides more significant insights into complex human-environment interactions in the housing system (Pagani, 2021). In light of this trend, the cognition, and analysis of housing dimension with quantitative and qualitative variables on the one hand and the analysis of factors affecting the location and placement, on the other hand, is of paramount importance. Also, periodic fluctuations in housing prices have escalated the investment risk, so urban planners and urban housing policymakers fail to make price projections. Therefore, modeling techniques to estimate future costs and the location and purchase of housing would be an effective and necessary strategy (Rahnama, 2014).

Location and spatial housing policy for each individual determine their individual-social identity and are the pivotal factors of spatial separation and socio-economic classification in cities (Miralaei, 2019). Thus, it can be posited that housing welfare policies have resulted in social sorts and growing inequality (Hoekstra, 2021). The housing crisis has emerged as one of the most pressing challenges in urban areas around the world. Rapid urbanization, population growth, and increasing economic disparities have contributed to a shortage of affordable housing, skyrocketing rents, and homelessness.

Palani in her research addressed the housing crisis and how social groups are located in Vienna, Austria, Vancouver, Canada, and Portland, Oregon (Palani, 2023). Elias, due to Zimbabwe's housing deficit, has included the housing problem as one of 14 national challenges to be addressed (Elias, 2023). Samarin has examined the relationship between immigrants from developing countries and rising housing rents in US metropolitan areas and believes Rent burden—often being overshadowed by terms such as housing cost burden or housing stress—has multiple potential determinants (Samarin, 2023). Alhajri in her research rapid urbanization and population growth, inadequate affordable housing, and low rates of home ownership indicates that the housing sector in the Kingdom of Saudi Arabia (KSA) cities faces significant challenges, most especially in providing adequate affordable housing for middle and low-income households (Alhajri, 2022). Moreover, Rezapour et al. analyzed spatial challenges in the Tabriz metropolis (Rezapour, 2021), and; very recently, Dolatabadi et al. explored the instability of urban spaces in Iran, especially the Tehran metropolis (Dolatabadi, 2023).

Fan et al. demonstrate that urban features extracted from street-view images through a computer vision model can effectively estimate the hidden neighborhood socio-economic status, such as travel behaviors, poverty status, health outcomes, and crime. Specifically, models using street-view features alone can estimate up to 83% of the variance in vehicle miles traveled, 64% in violent crime occurrences, and 68% in the population lacking physical activities, thereby outperforming models based on points of interest, population, and other demographic data alone (Fan, 2023). Kang et al. emphasize the potential of multi-source big geo-data in modeling house price appreciation, noting that deep features extracted from diverse datasets, including house photos and street-view images, can significantly enhance predictive models. They assert that integrating these data sources, along with machine learning models, allows for a more detailed and accurate prediction of house price trends, which is crucial for both market analysis and urban planning (Kang, 2020). Also, Kang et al. propose a novel place-oriented hedonic pricing model (P-HPM) that enhances traditional real estate valuation by integrating human dynamics and perceptions derived from extensive urban data. By analyzing house price data alongside locational amenities, human mobility patterns, and social interactions in Boston and Los Angeles, their model demonstrates significant improvements over traditional methods. The inclusion of human perceptions and mobility patterns, extracted using deep learning from large-scale street-view images, underscores the variable impacts of place-related factors on house values, offering vital insights for urban planning and policy-making. These three papers utilized Street-View Imagery and OpenStreetMap (OSM) data to enhance their research. However, these tools are not yet suitable for studies in Iran because the available data are neither extensive nor accurate enough at present. In the future, as the quantity and quality of data improve, further studies could effectively employ these tools (Kang, 2021). Tehran's metropolis, with 22 urban districts and a population of 8,693,706 people, is the most populous city in the urban network of Iran, playing a vital role in the condensation of population and activities across the country (Fathi, 2020). Also, the actual liquidity volume in Iran has a significant bearing on the housing bubble, which has triggered social and economic schisms in the 22 districts of Tehran (Nasr Isfahan, 2017). As a result of this sharp spike in prices, people dwelling in southern Tehran struggle with the problem of spatial inequality (Abdi Daneshpour, 2018). Given the unique position of Tehran (as the capital) in the urban hierarchy, it can adequately portray the housing selection method and determinants of its price in Iranian metropolises. Previous research has examined the challenges and instability of Iran's metropolitan spaces, and few studies have focused on the measurement and modeling of housing locations through data mining and valuation. Therefore, this research has focused on the location of housing and its relationship with social groups and the economic power of citizens based on some determined indicators. Using data mining and valuation of data related to 8000 samples, this article aims to measure and model the location and placement of housing in Tehran. To achieve the goal of this article, nine relevant variables affecting housing location in 22 districts of Tehran were selected, and the spatial database was created with an accuracy of 85% to measure the variables. Data incompatibility was the primary challenge in this study. Hence, attempts were made to improve the data to obtain more effective outcomes through validation and removal of outliers.

Literature review

Several theories have been proposed to ascertain the quantitative and qualitative dimensions of housing, the locating process, and location determinants. Each of these theories covers part of the facets of this research.

Sustainable urban development perspective. This perspective stresses the engineering approach to design and build construction projects (Shen, 2023) and the compression of the city in terms of physical and social cohesion and the competitiveness of all urban society groups in providing critical housing, a key achievement (Zhang, 2020). Also, it underlines optimal and maximal land use to meet the needs of current and future citizens (Zagorskas, 2007). The principles of this view are aligned with

Table 1 Results of model accuracy measurement before and after removal of outliers and irrelevant data.

Model	Lasso regression	Elastic net	Kernel ridge	Gradient boosting regressor	XGB regressor
After	0.82	0.77	0.84	0.84	0.85
Before	−0.26	−0.27	−188	0.00	−4.1

urban land use planning to achieve social, economic, and physical sustainability in the city (Yang, 2018) and emphasize the biological organization and location of various activities in the city to tailor and adapt the land to the demands of citizens (Ziyari, 2009). The theoretical achievement of this theory is that it is possible to sustainably live in harmony with the urban system so that most minor challenges and gaps are confronted in urban areas today and in the future, and metropolitan urban regions and neighborhoods are livable.

Valuation theory. According to this theory, two methods can be employed to evaluate or select any product or service. One is the self-declaration of individuals, and the other is the disclosure of products and services (Heldt, 2016). In the former, the applicants of a product or service voice their hypothetical opinions. In the latter, the market demand factor is a crucial principle. In the self-declaration method, the product valuation (i.e., housing in this study) is conditional and involves discrete selection. The importance of self-declaration is that it measures choices (product or service) by integrating non-use value and selective value (Miralaei, 2019). This theory's application in this research accounts for the self-declared and particular method as the hypothetical behavior of citizens in house selection. The main advantage of this method is the ease of identifying, coding, ranking, and measuring the level of interest in buying housing by citizens in any area of Tehran and the possibility of providing practical suggestions to improve the economic and social gap.

Stochastic utility theory. This theory is based on the nested model. In this model, decision-making for household location and the possible selection process are considered among discrete alternatives, emphasizing the quantitative and qualitative characteristics of housing and its social environment (Miralaei, 2019). The model's basis indicates that the housing is located hierarchically. First, the citizens determine a social setting, such as an urban area, and then suitable accommodation with desired benefits and socio-economic variables such as price, area, neighborhood, etc. are selected (Kim, 2010). The theoretical implication of this theory in the present study is that housing location is a function of region, neighborhood, and socio-economic variables in the city.

Grounded theory. This theory is a strategy and method of knowledge production based on data (Danaei Fard, 2007). The principles of this theory are founded on the inductive approach (bottom-up) and underscore discovering the relationships of variables and exploring the impact of each variable on the subject (Esfandiari, 2019). Grounded theory, through data mining, provides reliable quantitative and measurable findings for the analysis. In this theory, three techniques of open (by creating variables and presenting their properties), axial (discovering relationships between variables), and selective (integration and improvement of variables) coding are crucial to finding the interrelations of variables and presenting the analysis. In this research, three coding techniques have been used for the examination of grounded theory is particularly useful for analyzing process-oriented phenomena such as housing because it helps outline a comprehensive map of people's experiences with housing locations. The main theoretical implication of this theory in this article is extracting and revealing covert knowledge from the mass of data received from the housing system and explicating tacit knowledge of housing purchases in 22 districts of Tehran based on databases.

Choice theory. Having a psychological origin, this theory was proposed by William Glasser in 1998. The principle that inspires the behavior of human beings, from birth to death, is driven by four factors actions, thoughts, emotions, and physiology (Glaser, 2003). According to this theory, man possesses the power of choice move and thought, and feeling and physiology are a function of human choice (Wubbolding, 2004). To Glasser, when it comes to options, people behave in a way that effectively controls the situation and identifies and eliminates their needs (Aghagedi, 2013). Recent research suggests that visual characteristics and environmental variables have a pivotal role in the choices of individuals (Esfandiari, 2019). The theoretical application of this theory is in explaining how citizens choose suitable housing. They determine the accommodation based on action and thought using accurate spatial data to respond to their family's feelings and physiological needs and improve and efficiency of housing (Ahady, 2022). Despite the extensive body of research on urban housing markets, a notable gap remains in the application of advanced data mining techniques to the Tehran housing market. Previous studies have primarily leveraged traditional statistical methods and often focused on cities with different socio-economic profiles than Tehran. Our study addresses this gap by employing a comprehensive, data-driven approach using a dataset of 8000 properties in Tehran, thus providing deeper insights into the unique dynamics of housing prices in this context. This approach not only complements existing methodologies but also adds a new dimension to the understanding of urban housing markets in Middle Eastern cities.

Materials and methods

To conduct research and determine the research strategy, theoretical-applied implications of grounded theory, choice theory, evaluation, random utility, and content analysis methods were adopted. Each of these perspectives and approaches directly impacted the description and analysis of this research. Data derived from five data science-related libraries in Python programming (Online, 2022a, 2022b, 2022c, 2022d, 2022e, 2022f, 2022g, 2022h, 2022i, 2022j, 2022k, 2022l, 2022m, 2022n, 2022o) format were utilized as a reference for data discovery and optimization. Although this method seems simple, using this method for this research has helped to measure, model, and present the findings better.

The following five models were also employed for the measurement and modeling:

Lasso regression: It is a type of linear regression that draws on shrinkage. Shrinkage describes where data values are shrunk towards a central point data (e.g., mean). This model best suits data that follow multiple alignments (Online, 2022a, 2022b, 2022c, 2022d, 2022e, 2022f, 2022g, 2022h, 2022i, 2022j, 2022k, 2022l, 2022m, 2022n, 2022o).

Kernel regression: The basis of this statistical model is a non-parametric method for estimating the conditional expectation of a

Table 2 Matrix of variables and data of residential units in urban neighborhoods and areas.

Index	Name of neighborhood	Deposit[a]	Rent[a]	Area[MM]	Year of built	Elevator[b]	Parking[b]	Warehouse[b]	Distinct	Total value[a]
1	Poonak	350,000,000	8,500,000	120	2017	1	1	1	6	633,333,333
2	Heravy	200,000,000	10,000,000	110	2017	1	1	1	4	533,333,333
…	…	…	…	…	…	…	…	…	…	…
8106	Narmak	100,000,000	6,000,000	90	2020	1	1	1	4	300,000,000

[a]The value was calculated in Tomans (10 Rials), the official currency of Iran.
[b]In these columns, "1" means existence, and "0" means the absence of the selected attribute.

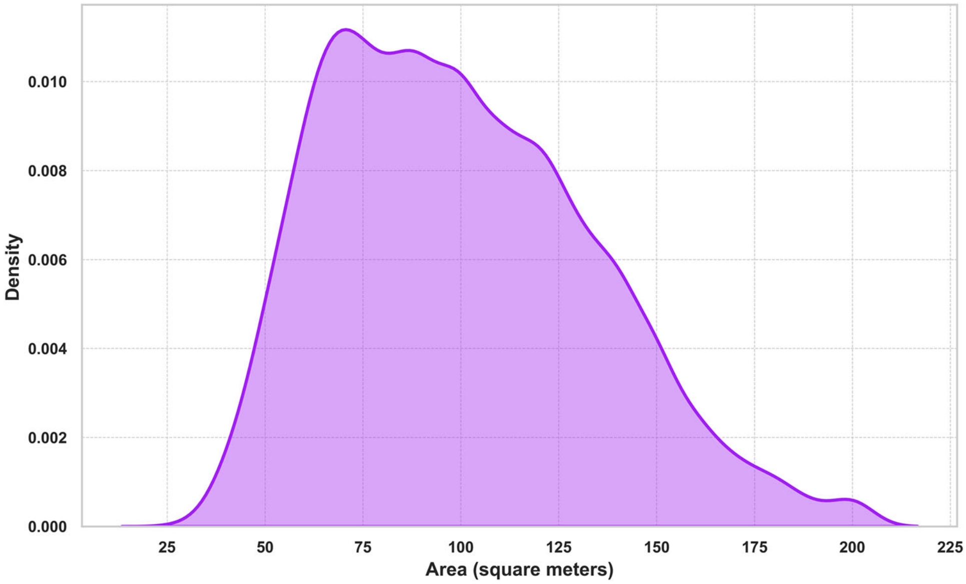


Fig. 1 Distribution of area column data after the deletion of outliers. This figure shows the density plot of the area column data after the removal of outliers. The *x*-axis represents the area in square meters, while the *y*-axis represents the density. The plot indicates the frequency distribution of area sizes within the specified range, highlighting the peak and spread of the data.

random variable, and its mission is to identify a nonlinear relationship between the two variables x and y (Online, 2022a, 2022b, 2022c, 2022d, 2022e, 2022f, 2022g, 2022h, 2022i, 2022j, 2022k, 2022l, 2022m, 2022n, 2022o).

Elastic net: It is a regulated model that linearly integrates the L1 and L2penalties of the lasso and ridge methods (Online, 2022a, 2022b, 2022c, 2022d, 2022e, 2022f, 2022g, 2022h, 2022i, 2022j, 2022k, 2022l, 2022m, 2022n, 2022o).

Gradient boosting regressor: A machine learning method that draws on the results of weaker models (e.g., decision trees) to improve learning outcomes (Online, 2022a, 2022b, 2022c, 2022d, 2022e, 2022f, 2022g, 2022h, 2022i, 2022j, 2022k, 2022l, 2022m, 2022n, 2022o).

XGB Regressor: A more powerful version of Gradient boosting regressor (Online, 2022a, 2022b, 2022c, 2022d, 2022e, 2022f, 2022g, 2022h, 2022i, 2022j, 2022k, 2022l, 2022m, 2022n, 2022o).

This is an exploratory study that adopts a descriptive-analytical perspective. The research sampling is also theoretical. That is a purposive sampling method in which the researcher tries to perform data mining and explore the phenomenon by drawing on the knowledge and opinions of the subjects (Kopai, 2015). Purposive sampling was also used to collect data, mainly extracted from the official sources and statistics (Online, 2022a, 2022b, 2022c, 2022d, 2022e, 2022f, 2022g, 2022h, 2022i, 2022j, 2022k, 2022l, 2022m, 2022n, 2022o). Also, the research data was derived from a systematic review of documents and techniques over 2 years. Data analysis was conducted based on the grounded theory and coding to discover priority variables in housing locations. Also, to convert nominal data to numerical data (the column related to the neighborhood), the One Hot Encoding method and Python programming language as content and data mining were used. Converting nominal data to numerical one is a requirement for learning models. The rationale for using data mining is to expand the size of existing and future data. Although data mining, like other techniques, could only be conducted with human intervention, it enables analysis, who may need to be more expert in statistics or programming, to manage the knowledge extraction process effectively (Wickramasinghe, 2005). The study population consisted of 18,000 samples of villas and apartments selected. After extracting and deleting duplicate data, data distribution on the map of Tehran was determined and data analysis was carried out in 3 steps. First, after validation, 8,000 data from 22 districts and 317 neighborhoods of Tehran were selected and evaluated in terms of 9 variables of the warehouse, elevator, parking lot, surface area, neighborhood, rent, mortgage, year, and total secure deposit affecting the housing prices. Then, the extent of positive or negative correlation of the selected indicators was measured using the Dython Library in the Python programming language. Finally, the learning models were estimated in the existing data using the cross-validation method.

Finally, five regression-based models were implemented on the research data to achieve 85% accuracy to enhance research validity. Therefore, based on Table 1, the accuracy of these models was measured using cross-validation (Online, 2022a) in two stages, before deleting the outliers and the warehouse

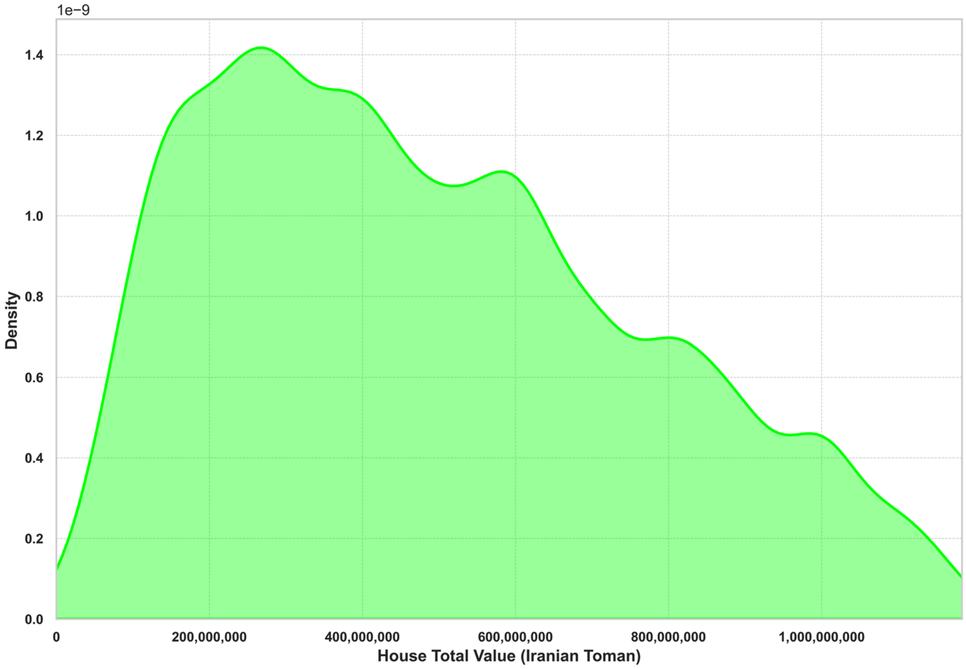


Fig. 2 Distribution of total value column data after the deletion of outliers. This figure illustrates the density plot of the total value column data after removing outliers. The *x*-axis represents the house total value in Iranian Toman, while the *y*-axis represents the density. The plot highlights the frequency distribution of house values, showing the range, peak, and overall distribution pattern of the data.

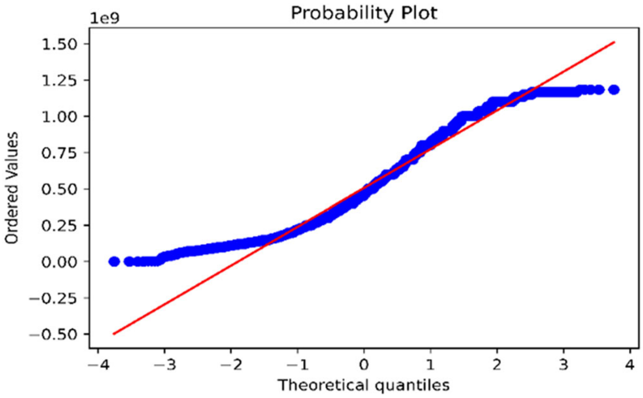


Fig. 3 The probability function of the total value column after the deletion of outliers. This figure presents a Q-Q (quantile–quantile) plot comparing the probability distribution of the total value column data, post-outlier removal, against a normal distribution. The *x*-axis represents the theoretical quantiles, while the *y*-axis represents the sample quantiles. The blue points indicate the observed values, and the red line represents the reference line for a normal distribution. The plot demonstrates how well the data conforms to a normal distribution, with deviations indicating departures from normality.

column, and after deleting the data and the warehouse column (Table 1).

Negative values in Table 1 suggest very low accuracy of models (Online, 2022e), and the closer the precision of a model is to 1 (assuming a maximum accuracy of 100%), the results would be better, and vice versa. A significant improvement in the accuracy of the models is because the skewness and kurtosis of the value distribution forms of each data column were optimized by deleting the outliers, which was essential for the modeling. The skewness and kurtosis optimization does not improve the accuracy of each model (Online, 2022). Still, the models adopted in this paper benefited from this optimization in the best possible manner. Since data with a surface area of more than 200 m^2 had an asymmetrical distribution, settlements with a maximum area of 200 m^2 were evaluated and measured. In this research, each data includes the house price, presented by each seller according to the determinants of residential housing prices. After selection, the research data was organized into a database, and several columns formed a matrix for valuation and encoding. Each column contains nine variables: Warehouse, elevator, parking lot, area, neighborhood, rent, mortgage, year, and total deposit, and the amount of data is shown in each row. Each of these variables plays a significant role in housing pricing and location. The neighborhood name column was converted into columns with numerical variables in the research process using the One Hot Encoding (Online-retrieved, 2022) method. For clarity, Table 2 displays the matrix of variables and the data values for selling or buying housing in some Tehran neighborhoods and urban areas (Table 2).

According to the data analysis, some values of the total value column were zero because they had been put on sale for a “negotiated” price. Therefore, the equivalent rent and deposit were zero. Thus, containing this value in this column was deleted because prices outside the natural range interrupt the learning process of models and yield false predictions.

For example, Figs. 1 and 2 show outliers for the columns relate to Area and Total values after the preprocessing data step (Figs. 1 and 2).

The bulk of data has a relative value of zero compared to other data, indicating that the data is too large with low frequency. In the research data section, by limiting the range of values, attempts have been made to bring the distribution of importance of these columns closer to the normal distribution. Also, the probability function pertained to the area columns, and the Total value before removing outliers caused by data that are too large or have low frequency was plotted this way. Figures 3 and 4 reveal the results after omitting outliers (Figs. 3 and 4).

The statistical studies suggested that the column dedicated to the year of construction also contained abnormal data; hence, houses built earlier than 1995 were removed as outliers. In addition, the skewness and kurtosis of the distribution curve related to the area, year of construction, and Total value, before and after the omission of outliers, are presented in Table 3 (Table 3).

The skewness and kurtosis of the distribution curve of each column exert a direct effect on the learning of prediction models, diminishing or improving the accuracy of the models. The closer the skewness and kurtosis are to their optimal value, the more accurate the models' prediction will be. The skewness and kurtosis of the other columns were not investigated due to the lack of continuous data. On the other hand, skewness in the range of −0.5 to 0.5 means that the data are relatively symmetric, and the kurtosis between −2 and 2 is acceptable (George, 2010). Therefore, skew values after removing remote data help the model learn more.

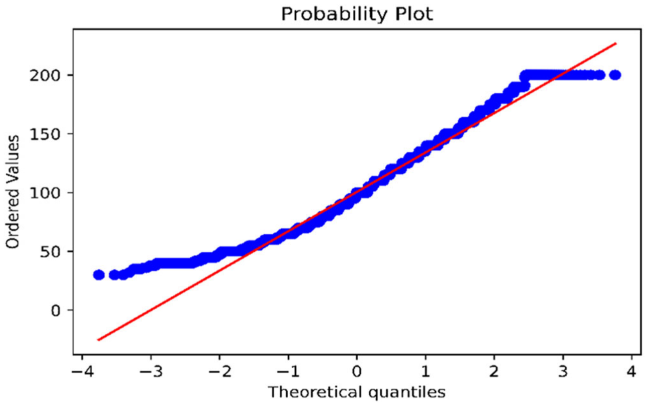


Fig. 4 The probability function of the area column after deleting outliers. This figure shows a Q-Q (quantile–quantile) plot comparing the probability distribution of the area column data, after the removal of outliers, to a normal distribution. The *x*-axis represents the theoretical quantiles, while the *y*-axis represents the sample quantiles. The blue points indicate the observed values, and the red line represents the reference line for a normal distribution. The plot assesses how closely the area data follows a normal distribution, with deviations highlighting discrepancies from normality.

Results

Correlation status and variables sustainability. Finding connections between columns helps strengthen the model and remove columns that are not irrelevant to the target column. The target column is the one that is subject to the prediction. The target column and the final price column are critical. Using this column, the last housing price in 22 urban areas can be modeled. Figures 5 and 6 show the degree of correlation and non-

Table 3 The skewness and kurtosis before and after the removal of outliers.

Factors	Skewness (before)	Kurtosis (before)	Skewness (after)	Kurtosis (after)
Area	86.79	8180.20	0.47	−0.33
Year	−1.11	1.68	−0.42	−0.73
Total value	55.55	3541.40	0.47	−0.70

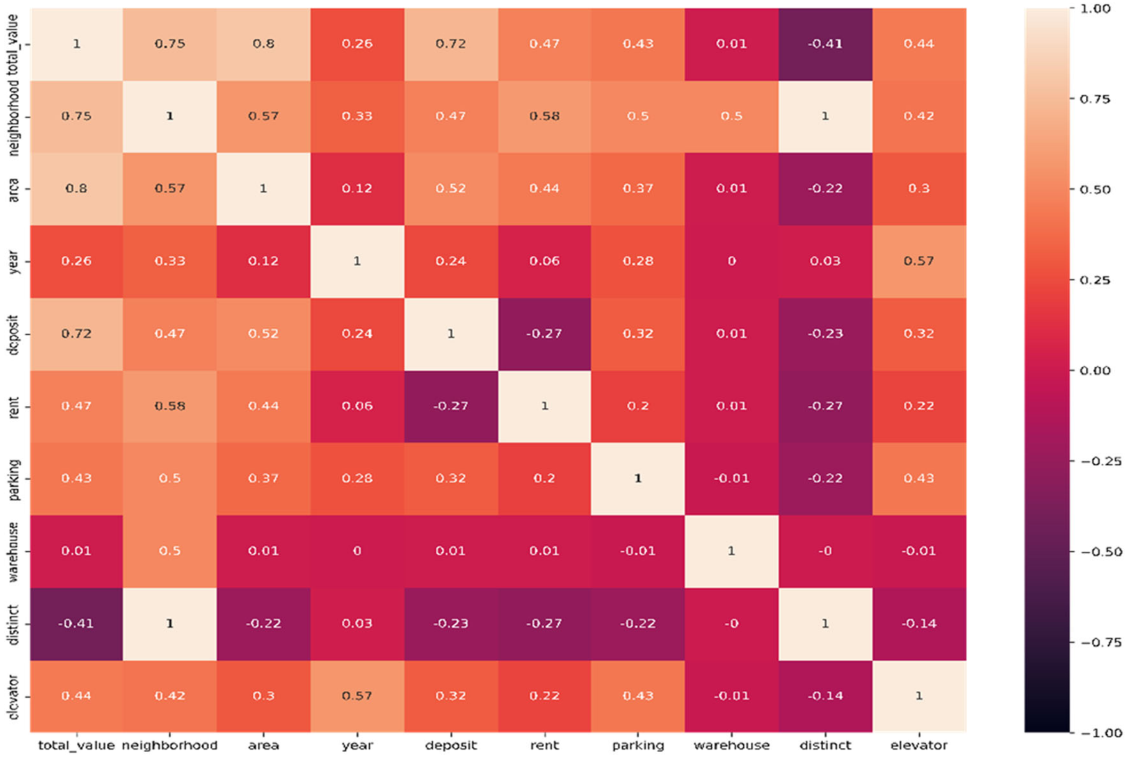


Fig. 5 Thermal matrix of the interrelation of variables. This figure depicts a heat map showing the correlation matrix of various variables. Each cell in the matrix represents the correlation coefficient between two variables, with the color intensity indicating the strength and direction of the relationship. The color bar on the right provides a scale for interpreting the correlation values, ranging from negative to positive correlations. High positive correlations are shown in lighter colors, while negative correlations are displayed in darker colors. The matrix helps visualize the degree of interrelation between the variables in the dataset.

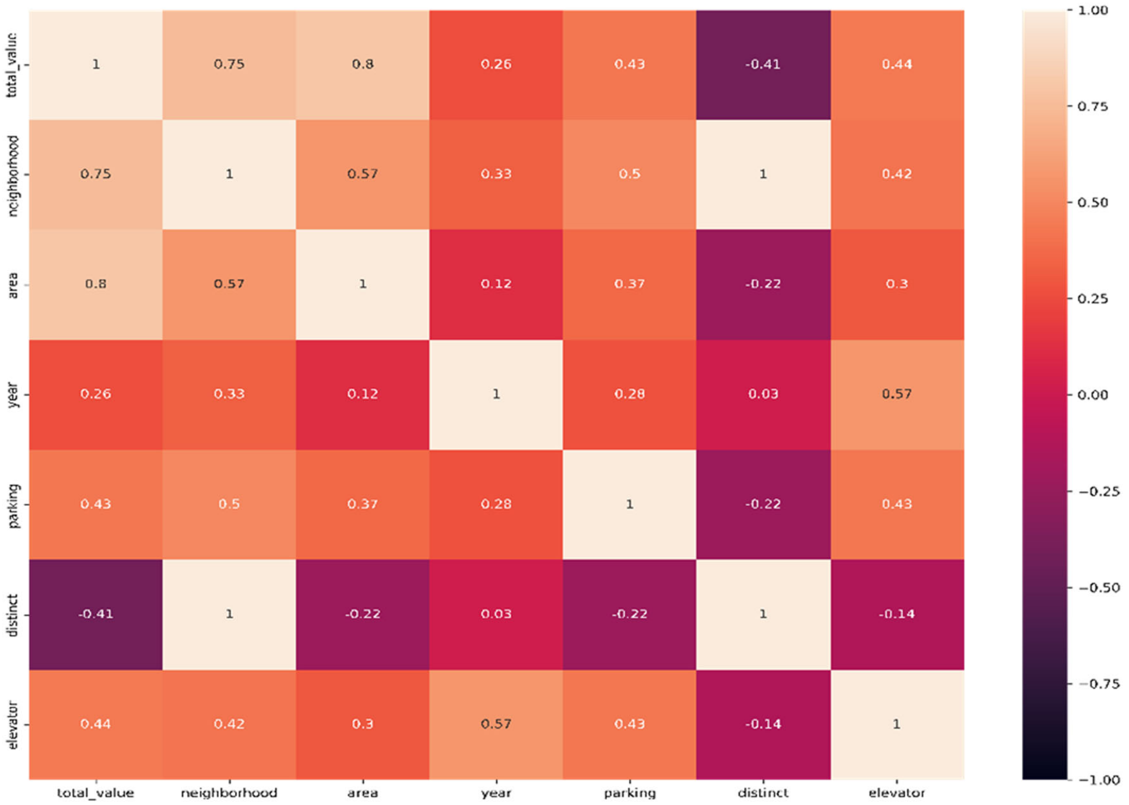


Fig. 6 Thermal matrix of the interrelation of variables after the removal of unnecessary columns. This heatmap displays the correlation matrix for variables after removing unnecessary columns. Each cell shows the correlation coefficient between two variables, with color intensity indicating the strength and direction of the relationship. The color bar on the right provides a scale for interpreting the values.

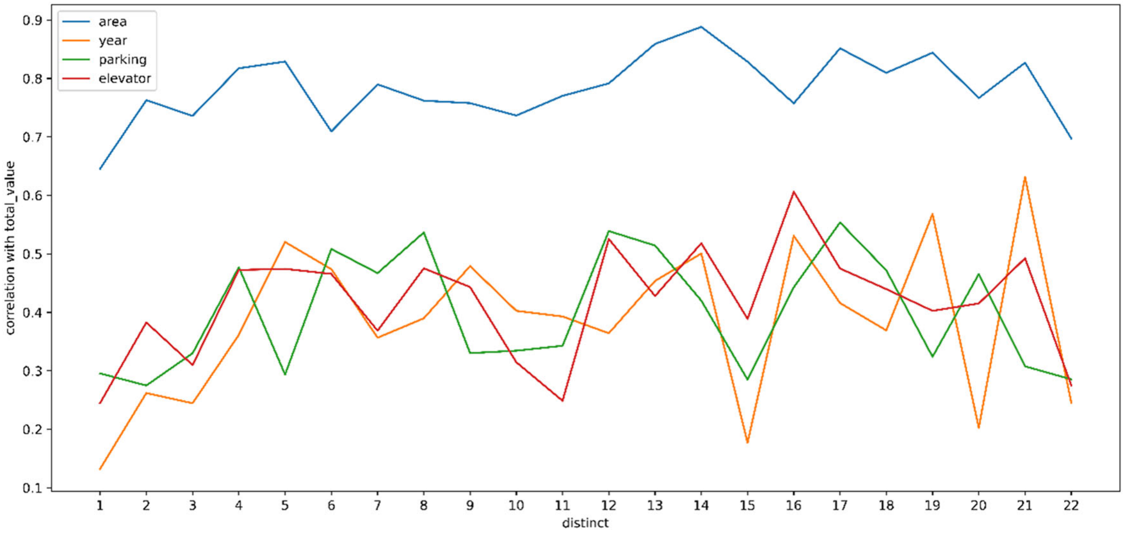
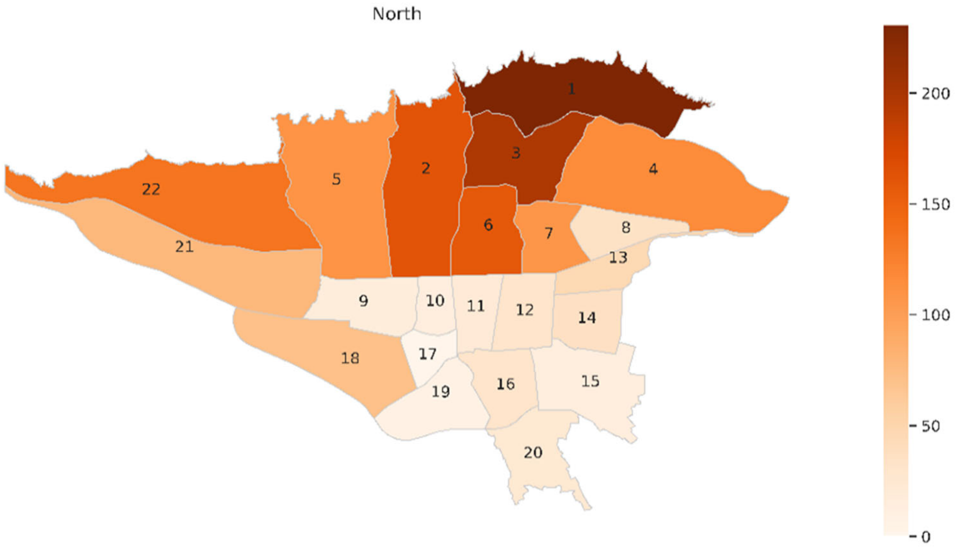


Fig. 7 Correlation of area, year of construction, parking lot, and elevator columns with total value. This figure shows a line plot illustrating the correlation of area, year of construction, parking lot, and elevator columns with the total value.Each line represents a different variable, showing how each correlates with the total value across different data points. The legend identifies the lines corresponding to each variable.

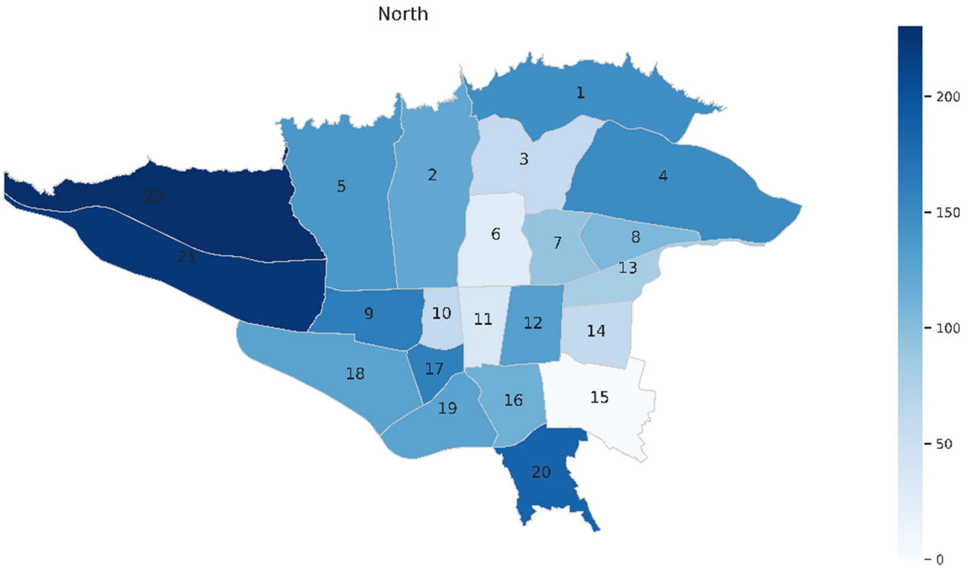
correlation of selected variables for the total data in the research scope and the interrelationships. The thermal matrices of the outputs (correlations) derived from this library are shown in Fig. 5 (Fig. 5).

A value of 1 indicates the highest positive correlation, 0 means lack of correlation, and −1 suggests the highest negative correlation. A positive correlation between two columns means that any increase in the values of one column would lead to a rise in the other column. In contrast, a negative correlation between two columns implies that a rise in one column would be associated with a reduction in the values of the other column.

According to Fig. 5, the absence or presence of a warehouse column is weakly related to the total value, but the neighborhood and area are strongly related to the final price. The rental and deposit columns are also inversely correlated, which is reasonable. In this process, the rental fee for three months (rent) and deposit



Map 1. The mean of houses price.



Map 2. The mean year of construction.

(deposit) are removed because the Total value column includes both. Figure 6 shows the final thermal matrix (Fig. 6).

Finding connections between columns is crucial to strengthening the model and removing columns unrelated to the target column. The target column is the one for which the prediction is made. The target and total value columns are significant, as the final housing price in 22 urban areas can be modeled and predicted based on these columns. Figures 5 and 6 illustrate the degree of correlation and non-correlation of selected variables for all data in the scope of research.

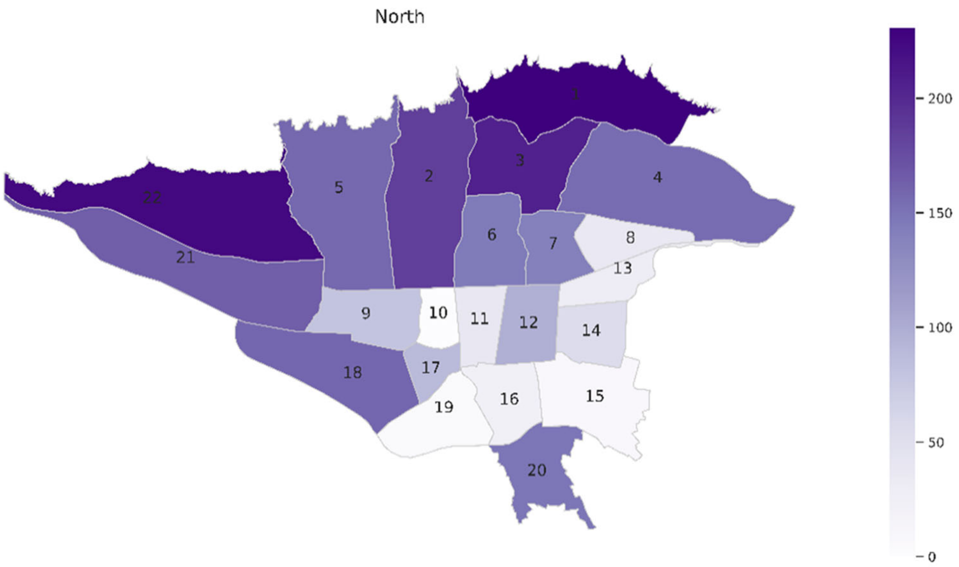
In addition to quantitative factors, the identity variable plays a crucial role in housing prices. This encompasses how residents' socio-economic status, cultural background, and lifestyle preferences influence their choice of neighborhood in Tehran. This aspect could further explain the variations in housing prices across different districts, reflecting the city's socio-cultural diversity. Integrating this qualitative dimension adds depth to our understanding of the housing market, bridging the gap between numerical data and the lived experiences of Tehran's residents.

The relationship of variables with the total value. In Fig. 7, the variables of area, year of construction, elevator, and parking lot are evaluated about the Total value. These results provide significant insights into the situation of each neighborhood and the urban area of Tehran. It is because, for example, parking does not have the same effect on the price of houses in all areas. Therefore, a separate study can draw a correlation more accurately. Figure 7 shows the correlation of different variables with the Total value column in the 22 areas of Tehran relative to each district (Fig. 7).

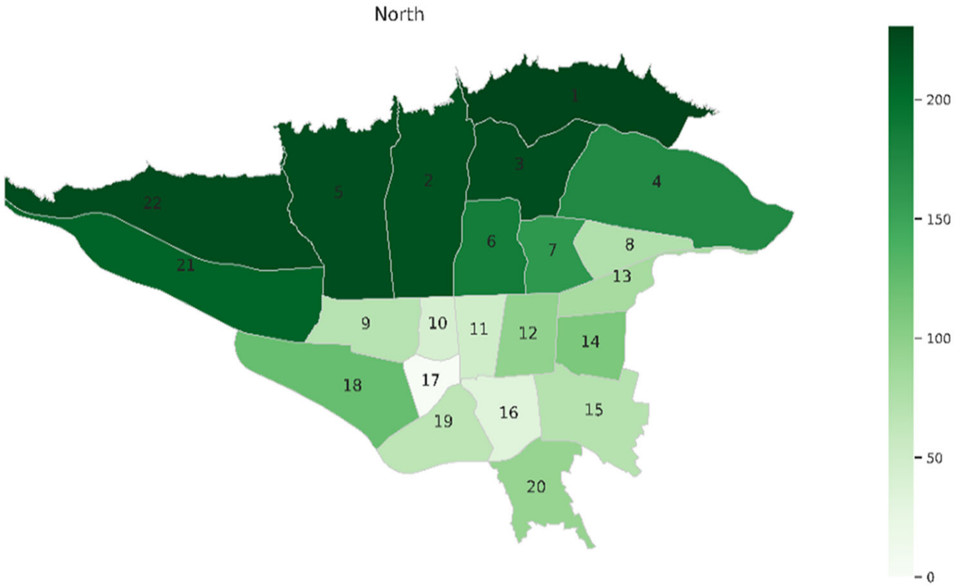
Figure 7 displays some research findings:

First, in District 14 of Tehran, there is a strong relationship between area and Total value, and the weakest relationship between area and house price is observed in District 1. Buying a house in District 14 is directly related to the area.

Second, the year of construction in District 21 has the most significant effect on the price, but in District 1, this relationship is reversed. The geographical location, air quality, and spatial value of District 1 can explain this.



Map 3. The density of houses having elevators.



Map 4. The density of houses having parking.

Third, the presence/absence of a parking lot has a massive bearing on the housing price in District 17, which the insufficient parking area can explain in this district.

Fourth, an elevator can wield a considerable influence on increasing housing prices in District 16. Still, it has a slight effect on house prices in District 1, which can be attributed to the high spatial value of District 1.

Fifth, the values of the horizontal chart in Fig. 7 suggest the study area and vertical chart values display the degree of correlation with the Total value variable.

Geographical distribution and location. Variations in locating and site selection patterns of Tehran citizens based on the mean of selected variables (price, year of construction, density of houses with elevators, and density of homes with parking) on the map of urban districts of Tehran portray a meaningful trend. The geographical distribution of 4 variables is displayed on the map because the warehouse variable was slightly correlated with the target variable and therefore removed. Also, because the two variables of rent and deposit are included in the Total value, they are not illustrated on the map. Darker areas on the map mean that a feature is more pronounced in that area. In maps 1, 2, 3, and 4, dark areas suggest that house prices are higher, houses are more recently constructed, and most homes have elevators and parking lots. According to Map 1, the highest and lowest housing prices belong to Districts 1 and 17, respectively. Therefore, District 1, due to its favorable climate and spatial value, has the highest price in the transaction of housing and commercial properties in Tehran (Map 1).

As depicted in Map 2, Tehran's physical expansion and development are defined based on a detailed plan in its western area, and there is enormous potential for urban growth in this part. Map 2 shows which distinct in Tehran contains more newly built houses. Thus, in Districts 21 and 22, the suitable topographic situation of urban planning has paved the way for

Table 4 Formula assessment by calculating the total value of housing in some neighborhoods and 22 districts of Tehran.

Total value	Neighborhood	Area MM	Year	Parking	Elevator	District
900,000,000	Gheitariye	150	1399	1	1	1
633,333,333	Poonak	120	1395	1	1	2
250,000,000	Mirdamad	40	1392	0	1	3
366,666,666	Narmak	110	1375	1	0	4
333,333,333	Kooye Ferdos	70	1387	1	0	5
683,333,333	Amir Abad	127	1398	1	1	6
666,666,666	Sohrevardi	101	1380	0	1	7
350,000,000	Kerman	70	1392	1	1	8
120,000,000	Ostad Moein	45	1384	0	1	9
180,000,000	Salsabil	50	1399	1	1	10
155,000,000	Navab	65	1384	0	1	11
270,000,000	Darvazeh Shemiran	70	1390	1	1	12
400,000,000	Piroozy	110	1399	1	1	13
170,000,000	Shokoofe	70	1385	0	1	14
110,000,000	Aboozar	40	1382	0	0	15
160,000,000	Ali Abad	85	1389	1	0	16
300,000,000	Emamzadeh Hassan	125	1395	1	1	17
683,333,333	Shams Abad	123	1396	1	1	18
100,000,000	Nemat Abad	70	1390	0	1	19
280,000,000	Dolat Abad	100	1395	1	1	20
360,000,000	Tehransar	100	1397	1	1	21
300,000,000	Shahrak Rahahan	87	1393	1	1	22

Under such conditions, it is possible to determine the location in the desired area.

the construction of residential complexes, and the trend is gaining momentum (Map 2).

Map 3 shows that due to the vertical expansion of urban planning and the existence of residential and commercial towers, the highest density of houses with elevators is in Districts 1 and 22. Map 3 shows the density of homes having elevators in Tehran. In District 1, there is a strong incentive for investment and building due to its spatial desirability. However, in District 22, new urban planning based on engineering principles has enabled high-rise tower construction (Map 3).

According to Map 4, Districts 1, 2, 3, 5, and 22 in northern Tehran have the highest share of the parking lot on account of renovation and planned expansion. Map 3 shows the density of houses having parking in Tehran. Hence, it is a point considered by people interested in housing investment in these districts. The old textures, marginalization, and low-density organic expansion in the city's south have influenced the parking variable (Map 4).

Formulation and modeling. One of the goals of this paper is modeling to predict and determine housing prices in the metropolis of Tehran. Therefore, the formulation and modeling of research findings represent the innovative section of this study. The analysis of various models indicates that the Lasso model provides the best features for predicting housing prices and site selection of the buyers in 22 districts of Tehran. Therefore, Formula 1 and Equation (A) were obtained to predict the final cost of housing by extracting information from the Lasso model.

$$\begin{aligned} \mathrm{A} = \mathrm{Total\_value} = {} & -9681055555 + 4593786(\mathrm{Area}) \\ & +6947200(\mathrm{Year}) + 30157955(\mathrm{Parking}) \\ & +47956128(\mathrm{elevator}) + \mathrm{Coefficient\ neighborhood} \end{aligned}$$

In addition, in the above Formula and Equation, the house area and year of construction must be substituted with area and year, respectively. Also, if there is a parking lot in the house, the parking value will be set to 1 and 0 otherwise. The variable of the elevator is also set to 1 when the home is equipped with an elevator and 0 otherwise. The coefficient neighborhood variable also indicates the coefficient value pertained to the area where the house is located, which is extracted from the table in the attachment. This formula is well suited for homes with a maximum area of 200 square meters built after 1994. Its adoption was justified by removing low-frequency data at the outset of this process to maximize the model's accuracy. Table 4 allows housing buyers to select one house from each district in the initial data, add values to the formula, and compare the result with the actual value (Table 4).

For ease of understanding and application, the first row of the attached table has been tested:

$$\begin{aligned} \mathrm{Total\ value} = {} & -9681055555 + 4593786(150) + 6947200(1399) + 30157955(1) \\ & +47956128(1) + 164924468 = 970183696 \end{aligned}$$

There is approximately a 70-million toman difference from the actual price. Thus, the accuracy of this formula for this particular calculation is as follows:

$$1 - (970183696 - 900000000)/900000000) = \%92$$

Discussion

Further enriching our analysis, a comparative study on the Chinese housing market employing multivariate linear regression across 31 provinces and cities in China reveals significant impacts of land purchase prices and residents' posable income on housing prices (Yuxi Jiang, 2022). These insights provide a valuable perspective for our study, highlighting the variability and universality of economic factors influencing urban housing markets. For example, a comparison between Tehran and key Chinese cities underscores the diverse yet similar economic dynamics shaping housing prices in different urban environments. The review of previous research on housing challenges in different countries and some Iranian metropolises shows the need to pay attention to housing challenges at the national level. As Elias (2023) believes, housing challenges in Zimbabwe are a national problem. Also, Samarin (2023) has paid attention to the effect of immigration from developing countries on housing problems in American cities. Furthermore, Alhajri (2022) has discussed the rapid urbanization and population growth in the emergence of housing problems in the big cities of Saudi Arabia. In these studies, housing is considered an important challenge and problem that has socio-economic and cultural dimensions. But in terms of geography and location of housing, no research has been done on the purpose of place value in residence. The existing research about Iran's metropolises is generally focused on public space, social and economic analysis, and urban instability.

In Iran, the latest research of Khademi (2021), Rezapour (2021) considers the causes of housing problems in the economic, social, and rural-urban migration flows, which have intensified housing lessness and marginalization around Iran's metropolises over several decades. In none of these studies has attention been paid to data mining and valuation of housing and how to locate according to location value. This article and its findings have specifically shown the relationship between location and housing value in Tehran. Also, with the help of modeling, it has investigated the spatial differences in terms of the residential system in Tehran. It is important to understand the behavior of Tehran residents to choose their housing according to the indicators examined in this article. Although the findings of this research have raised a socio-economic challenge regarding housing and location, so far, no research has been done on housing modeling based on data mining and valuation in the Tehran metropolis. Therefore, the findings of this article show that there is a significant relationship between housing location, geographic location, facilities and infrastructure, and family purchasing power, which can be understood in the figures, maps, and tables of this process. Also, this article has been able to model housing with the help of data, which is an innovation in the housing problems of the Tehran metropolis. The findings of this research can reveal the social, economic, and cultural aspects related to the positioning and environmental behaviors of citizens, and urban managers can implement urban development programs based on these behaviors.

Conclusion

The analysis of variables affecting the method of locating and site selection of Tehran citizens provided promising results, which are presented in two categories of theoretical and objective findings (spatial inequality in location):

- The theoretical findings of this paper highlight the importance of access to housing as one of the social rights of citizens. Housing policy and planning need to support all social and economic groups. Therefore, a model and formula for price prediction were presented to improve the housing situation and locating methods in Iranian cities, especially the metropolis of Tehran.

- The objective results of identifying the effects of housing policies suggest spatial inequality in Tehran. Regarding the analysis of variables influencing housing and inequality in location, the following results were obtained:

(1) This is the first large-scale research on the 22 urban districts of Tehran to investigate the socio-economic and environmental dimensions of housing from the applicants' perspective. Its innovation lies in proposing a model and formula for price prediction.

(2) Preparing the environmental database allowed comparing selected variables in each district of the Tehran metropolis and determining the weight and impact of each variable on the tendency of housing buyers and the incentives of the trends.

(3) Positive correlation between new urban planning and the observance of urban engineering principles has influenced the housing prices and site selection in the west of Tehran.

(4) spatial value and spatial desirability in northern Tehran have a direct and positive effect on building density, high price, and residence tendency compared to the southern half of Tehran.

(5) Housing buyers prefer urban areas with more excellent facilities and better air quality. Therefore, according to the thermal matrix, figures, and maps, policymakers and investors of urban housing in Tehran attach greater priority to environmental factors and the housing situation.The findings of this study revealed that housing supply is significantly related to the density of elevators and parking lots in the northern areas of Tehran, such as districts 1, 2, 3, 5, and 22. The tendency to dwell in villas rather than apartments as a residential pattern is more dominant in the northern part of Tehran. Also, the economic and social concentration is more pronounced in the north of Tehran than in southern areas. The expansion of marginalization and socio-economic problems in the southern regions has given rise to a class gap between the south and the north, the rich and the poor. This schism in housing has maintained its upward trend until now. Based on our findings, we recommend urban planners and policymakers in Tehran, and similar cities, consider the following strategies:

(1) Targeted urban development: Focus on enhancing infrastructure and amenities in neighborhoods showing a positive correlation with higher housing prices, to promote balanced urban growth.

(2) Data-driven policy-making: Utilize data mining techniques similar to those applied in this study for informed decision-making in housing and urban development policies.

(3) Affordable housing initiatives: Implement policies to support affordable housing in areas where rent and deposit have an inverse correlation with housing prices, addressing potential issues of housing affordability and social inequality.

4) Investor guidance: Provide data-driven insights to real estate investors for identifying potential areas of high growth and return on investment.

These strategic recommendations aim to leverage the insights from our study to foster a more equitable, data-informed approach to urban planning and housing market regulation.

While our study offers valuable insights into Tehran's housing market, it's essential to recognize its limitations. The methodology and data used, while robust, are specific to Tehran and may need adjustments for applicability in different urban contexts, as highlighted in our comparative analysis with other cities. Furthermore, the study focuses on quantitative data, leaving room for future research to explore qualitative factors affecting housing prices. These limitations offer avenues for further study, emphasizing the need for context-specific adaptations in housing market analysis.

Data availability

The dataset analyzed during the current study is available as a supplementary file attached to this manuscript. The supplementary file, in.csv format, includes comprehensive data on housing locations, prices, surface areas, neighborhood characteristics, and other relevant factors used in the modeling and analysis.

Received: 22 March 2023; Accepted: 29 May 2024;
Published online: 21 June 2024

References

Abdi Daneshpour Z (2018) Analysing spatial inequality in Tehran's housing system, via changing prices during 1992–2016. Geogr. Dev. Iran. J. 16:267–292

Aghagedi P (2013) Studying the effectiveness of training Glasser's Choice and control theory on identity pattern evolution of adolescent. J. Psychol. Stud. 8:33–56

Ahady EA (2022) Urban residential buildings' energy consumption pattern and efficiency. Iran. J. Sci. Technol. Trans. Civ. Eng. 46:3963–3978

Alhajri (2022) Housing challenges and programs to enhance access to affordable housing in the Kingdom of Saudi Arabia. Ain Shams Eng. J. 6:101798

Danaei Fard H (2007) Qualitative research strategies: a reflection on data theory. Strateg Manag. Thought 1:69–97

Dolatabadi (2023) Examining the problems and challenges of regeneration of inefficient urban areas with the approach of sustainable development of Tehran. J Geogr Region Dev, Articles in Press. 1–24. https://doi.org/10.22067/jgrd.2023.81200.1249

Elias (2023) Housing crisis, affordable housing. Linkedin. pp. 1–4

Esfandiari MI (2019) Customer behavior analysis of the bank industry: grounded theory approach. Econ. Model 13:93–114

Fan ZZ (2023) Urban visual intelligence: uncovering hidden city profiles with street view images. Proc. Natl Acad. Sci. USA 120(27):e2220417120

Fathi E (2020) Trend of population changes in Tehran: from the past to the future. Stat. Month 6:32–35

George MA (2010) SPSS for Windows step by step: a simple guide and references. Pearson, Boston

Glaser W (2003) Selection theory: the new psychology of individual freedom (translated by Mehrdad Firooz Bakht). Resa, Tehran

Heldt BG (2016) Determination of attributes reflecting household preferences in location choice modes. Trans. Resh Procedia 19:119–134

Hoekstra JD (2021) Attitudes towards housing equity release strategies among older home owners: a European comparison. J. Hous. Built Environ. 36:1347–1366

Kang YZ (2021) Understanding house price appreciation using multi-source big geo-data and machine learning. Land Use Policy 111:104919

Kang YZ (2021) Understanding human settlement value assessment from a place perspective: considering human dynamics and perceptions in house price modeling. J. Cities 118:103333

Khademi (2021) An analysis of the future challenges of social, institutional, and economic sustainability of Iranian metropolises. Region Plann 1–17

Kim MJ (2010) Residential location decisions: heterogeneity and the trade-off between location and housing quality. The Ohio State University, Ohio

Kopai M (2015) Paradigm model of Jihadi management using data foundation theory. Gov. Manag Perspect. 5:109–128

Li J (2021) Assessing economic, social and environmental impacts on housing prices in Hong Kong: a time-series study of 2006, 2011 and 2016. J. Hous. Built Environ. 37:1–25

Miralaei SM (2019) Housing choice based on the trade-off between residential location characteristics and housing quality using choice experiment method from homeowner's household viewpoint in Isfahan. Urban Econ. 4:55–70

Naghizadeh M (2017) Islamic aboding: from theory to practice in the past and present. J Architect. Thought 1:47–67

Nasr Isfahan RS (2017) Analysis of economic effective factors on the housing price bubble (Case study: Tehran). J. Econ. Res 52:163–186

Online (2022a, December 14) Cross-validation. Retrieved from en.wikipedia.org: https:/en.wikipedia.org/wiki/Cross-Validation_(statistics)

Online (2022b, December 14) Dython. Retrieved from shakedzy.xyz: https://shakedzy.xyz/dython/

Online (2022c, December 14) Elastic. Retrieved from en.wikipedia.org: https://enwikipedia.org/wiki/Elastic_net_regularization

Online (2022d, December 14) Gradient. Retrieved from en.wikipedia.org: https://en.wikipedia.org/wiki/Gradient_boosting

Online (2022e, December 14) kaggle. Retrieved from kaggle.com: https://www.kaggle.com/amiralimadadi/tehran-housing

Online (2022f, December 14) Kaggle. Retrieved from kaggle.com: https://www.kaggle.com/questions-and-answers/152500

Online (2022g, December 14) kernel. Retrieved from en.wikipededia.org: https://en.wikipedia.org/wiki/Kernel_regression

Online (2022h, December 14) Matplotlib. Retrieved from matplotlib.org: https://matplotlib.org

Online (2022i, December 14) Numpy. Retrieved from numpy: https://numpy.org

Online (2022j, December 14) Pandas. Retrieved from pandas.pydata: https://pandas.pydata.org/docs

Online (2022k, December 14) Python. Retrieved from python.org: https://www.python.org

Online (2022l, December 14) researchgate. Retrieved from researchgate.net: https://www.researchgate.net/publication/334309178_The_relationship_between_data_skewness_and_accurancy_of_Aarticial_Neural_Network_predictive_modle

Online (2022m, December 14) Scipy. Retrieved from docs.scipy: https://docs.scipy.org.doc

Online (2022n, December 14) shirinsplayground. Retrieved from shirinsplayground.netlify.app: https://shirinsplayground.netlify.app/2018/11/ml_basics_gbm/

Online (2022o, December 14) statisticshowto. Retrieved from statisticshowto.com: https://statisticshowto.com/lasso-regression

Online-retrieved (2022, December 14) wikipedia. Retrieved from en.wikipedia.org: https://en.wikipedia.org/wiki/One-hot

Pagani AB (2021) Tenants' residential mobility in Switzerland: the role of housing functions. J. Hous. Built Environ. 36:1417–1456

Palani (2023) Housing crisis in cities: causes, consequences, and solutions. Linkedin, Plann Eng 1–3

Pilehvar (2020) Urban unsustainability engineering in metropolises of Iran. Iran. J. Sci. Technol. Trans. Civ. Eng. 44:775–785

Pilehvar A (2022) Investigating the relationship between informal economy and competitiveness in Iran's metropolises. J. Knowl. Econ. 14:1–24

Rahnama MA (2014) Estimation the housing price in holy city of Mashhad using the Kaplan Meier model (survival curve). J. Urban Econ. Manag 2:31–46

Rezapour (2021) Study of spatial planning challenges of Iranian metropolises with the focus on the new economic system (case study of Tabriz metropolis). J. Geogr. Plann 25:113–127

Samarin (2023) A typology of U.S. metropolises by rent burden and its major drivers. GeoJournal 88:4887–4906

Shen EA (2023) Evaluating the engineering-procurement-construction approach and whole process engineering consulting mode in construction projects. Iran. J. Sci. Technol. Trans. Civ. Eng. 47:2533–2547. https://doi.org/10.1007/s40996-023-01040-x

Wickramasinghe NG (2005) Creating knowledge-based healthcare organizations. Idea Group Publishing, Hershey

Wubbolding RE (2004) Reality therapy a global perspective. Int J. Adv. Couns. 26:219–228

Yang JG (2018) Sustainability article how to measure urban land use Intensity? A perspective of multi-objective decision in Wuhan urban agglomeration. China Sustainability J. 10:1–15

Yuxi Jiang LQ (2022) Empirical study on the influencing factors of housing price—based on cross-section data of 31 provinces and cities in China. Procedia Comput Sci. 199:1498–1504

Zagorskas JB (2007) Urbanistic assessment of city compactness on the basis of GIS applying the COPRAS method. Ekologija 53:55–63

Zarghamifard MM (2019) Determining the adequate housing indicators from islamic school viewpoint. J. Stud. Islam-Iran. 9:33–45

Zhang B (2020) Social policies, financial markets and the multi-scalar governance of affordable housing in Toronto. Urban Stud. 57:2628–2645

Ziyari Y (2009) Study and analysis of urban land use and weighting of location criteria for CNG stations using AHP model case study: Tehran 4 gas district. Q J. N. Attitude Geogr. Hum. 2:39–52

Author contributions

Ali Asghar Pilevar: provided information about recent Geo-Planning related articles and theories, advised on the appropriate article format, corrected reference formats, supplied some reference articles, and found the best compatible journal for publication. Arian Ghasemi: programmed and validated machine learning models for this research, gathered data, generated output figures and tables, applied revisions to the article, handled multiple submissions, and supplied some reference articles.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

Informed consent was not required as the study did not involve human participants.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-024-03244-6.

Correspondence and requests for materials should be addressed to Ali Asghar Pilehvar.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.