AN INTERPRETABLE ANALYSIS OF RESIDENTIAL APARTMENT PRICES IN TASHKENT USING LINEAR REGRESSION

Abdurasul Bobonazarov, Khamidulla Khabibullaev

Turin Polytechnic University in Tashkent
e-mail: a.bobonazarov@polito.uz

Abstract: Accurate understanding of residential housing price formation is essential for urban planning, real estate market transparency, and informed decision-making by buyers and sellers. In recent years, online real estate platforms have emerged as valuable sources of large-scale market data; however, empirical studies based on local housing data in Uzbekistan remain limited. This paper presents an interpretable analysis of residential apartment prices in Tashkent using multiple linear regression applied to real listing data collected from an online real estate platform.

The dataset consists of 7,421 apartment listings and includes structural characteristics (apartment size, number of rooms, floor level, and total building floors) as well as geographic attributes represented by latitude and longitude. A systematic preprocessing pipeline involving outlier removal and train–test splitting is employed. Several linear models, including ordinary least squares, Ridge, Lasso, and Elastic Net regression, are evaluated and compared using standard performance metrics.

Experimental results show that all linear variants achieve comparable performance, with an R^2 value of approximately 0.67 on the test set, indicating that a substantial proportion of price variability can be explained using linear relationships. The analysis highlights the dominant role of spatial location and apartment size in price formation, while regression coefficients provide clear interpretability of individual feature effects. Residual analysis confirms the absence of systematic prediction bias, with increased variability observed for higher-priced apartments.

Overall, the findings demonstrate that simple and interpretable linear regression models provide a robust baseline for residential price analysis in the Tashkent housing market. The study offers practical insights for market participants and lays the groundwork for future extensions incorporating nonlinear and spatially explicit modeling approaches.

Keywords: *Residential property prices; Linear regression; Housing market analysis; Spatial factors; Real estate listings; Data-driven analysis; Tashkent housing market.*

INTRODUCTION

Residential housing markets play a critical role in urban economic development, influencing household welfare, investment decisions, and city planning strategies. Understanding the factors that determine apartment prices is therefore essential for policymakers, real estate professionals, and individual buyers. In recent years, the rapid growth of online real estate platforms has enabled access to large volumes of market data, creating new opportunities for data-driven analysis of housing price formation.

A common and well-established approach to housing price analysis is the use of hedonic pricing models, where property prices are expressed as a function of structural, locational, and environmental attributes. Among these approaches, multiple linear regression remains widely used due to its transparency, ease of interpretation, and solid theoretical foundation. Several studies have demonstrated that linear regression models can effectively capture key price determinants such as apartment size, number of rooms, building characteristics, and location variables [1-3]. Despite the growing popularity of complex machine learning and deep learning models, linear regression continues to serve as a strong baseline for housing price analysis, particularly when interpretability is a primary objective [4,5].

Recent literature has increasingly explored advanced regression techniques and machine learning models for housing price prediction, including regularized linear models, ensemble methods,

and neural networks [6-10]. While these approaches often achieve improved predictive accuracy, they may sacrifice interpretability, making it difficult to understand the economic significance of individual features. For emerging housing markets and policy-oriented studies, interpretable models remain especially valuable, as they allow stakeholders to clearly assess how specific attributes contribute to price variation [7,11].

Despite the extensive global literature on housing price modeling, empirical studies based on real estate data from Uzbekistan remain scarce. The Tashkent housing market exhibits unique structural and spatial characteristics shaped by rapid urban growth, heterogeneous neighborhood development, and evolving market dynamics. The lack of systematic, data-driven analyses using local housing data represents a notable research gap, particularly with respect to interpretable models that can provide actionable insights into price formation mechanisms.

In this study, we address this gap by conducting an interpretable analysis of residential apartment prices in Tashkent using multiple linear regression applied to real listing data collected from an online real estate platform. The proposed approach incorporates both structural apartment attributes and geographic coordinates to account for spatial effects. Several linear model variants, including ordinary least squares and regularized regressions, are evaluated and compared. The primary objective is not to achieve maximum predictive accuracy, but rather to identify and interpret the dominant factors influencing apartment prices in the Tashkent housing market, thereby providing a transparent and reliable baseline for future research and more advanced modeling approaches.

Dataset

The dataset used in this study was collected from publicly available residential apartment listings published on an online real estate platform. The analysis focuses exclusively on apartments located in Tashkent, the capital city of Uzbekistan. After initial data collection, the dataset consists of 7,421 listings, each representing a unique apartment offering. The listings include information provided by sellers or agents at the time of publication and reflect asking prices rather than finalized transaction prices. All prices are reported in United States dollars (USD), which is a commonly used reference currency in the local real estate market.

Each apartment listing contains a set of structural and geographic attributes used for modeling. The target variable is the total apartment price, while explanatory variables capture physical characteristics of the apartment and its location. A summary of the variables is provided below:

- *Price:* Total listed price of the apartment (USD), used as the target variable.
- *Size:* Apartment area in square meters (m^2).
- *Rooms:* Number of rooms in the apartment.
- *Level:* Floor number on which the apartment is located.
- *Max_levels:* Total number of floors in the building.
- *Latitude (lat):* Geographic latitude of the apartment location.
- *Longitude (lng):* Geographic longitude of the apartment location.

Geographic coordinates are used as continuous spatial proxies to capture location-related price variation across different parts of the city.

Prior to model training, several preprocessing steps were applied to ensure data quality and robustness of the analysis. Listings with missing or invalid values in essential fields (price, size, number of rooms, floor level, building height, and geographic coordinates) were removed.

To reduce the influence of extreme values, outlier filtering was performed using a quantile-based approach. Specifically, observations with prices or apartment sizes below the 1st percentile or above the 99th percentile, computed on the training subset, were excluded from the analysis. This procedure preserves the core structure of the housing market while mitigating the impact of atypical luxury listings and data anomalies. The effect of this filtering is illustrated in *Figure 1*.

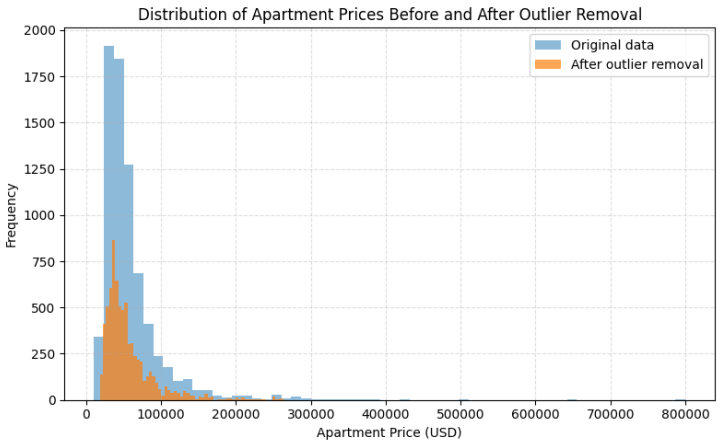

Figure 1. Distribution of apartment prices before and after outlier removal.

The dataset was then randomly shuffled and divided into training (75%) and test (25%) subsets to enable an unbiased evaluation of model performance.

Preliminary exploration of the cleaned dataset reveals substantial heterogeneity in both apartment prices and spatial distribution. *Figure 2* illustrates the geographic dispersion of listings across Tashkent, highlighting pronounced spatial price gradients and clustering effects. These observations motivate the inclusion of geographic coordinates as explanatory variables in the regression models.

Figure 2. Spatial distribution of apartment listings in Tashkent after outlier filtering.

Overall, the resulting dataset provides a representative and sufficiently large sample of the Tashkent apartment market, enabling interpretable regression-based analysis of residential price formation.

METHODOLOGY AND RESULTS

This study employs multiple linear regression to analyze residential apartment prices in Tashkent. Linear regression is selected due to its interpretability, transparency, and suitability for assessing the marginal contribution of individual housing attributes to price formation. The general regression model is defined as:

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon \quad (1)$$

where y denotes the apartment price, x_i represents the explanatory variables, β_i are the regression coefficients, and ε is the error term.

The explanatory variables include apartment size, number of rooms, floor level, total number of floors in the building, and geographic coordinates (latitude and longitude), which are used as continuous proxies for spatial location. Prior to model training, the dataset was randomly divided into training (75%) and test (25%) subsets. To reduce the influence of extreme observations, a quantile-based outlier filtering strategy was applied, excluding values below the 1st percentile and above the 99th percentile for apartment price and size. Importantly, percentile thresholds were computed using the training set and subsequently applied to the test set to avoid data leakage.

In addition to ordinary least squares (OLS) regression, several regularized linear models, Ridge, Lasso, and Elastic Net, were evaluated to assess whether coefficient regularization improves predictive performance or model stability. All models were trained using identical feature sets and evaluated under the same conditions. Model performance was assessed using the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE).

Table 1 summarizes the predictive performance of the evaluated linear regression models on the test dataset. The OLS model achieves an R^2 value of approximately **0.67**, indicating that a substantial portion of apartment price variability is explained by the selected explanatory variables. The RMSE and MAE values are consistent with expectations for real estate listing data, which typically exhibit inherent noise and heterogeneity.

	R^2	RMSE	MAE
Ordinary Least Squares	0.67	24731	14356
Lasso Regression	0.67	24731	14356
Ridge Regression	0.67	24732	14363
Elastic Net Regression	0.67	24732	14360

Table 1. Predictive performance of linear regression models on the test dataset.

The regularized models, Ridge, Lasso, and Elastic Net, exhibit nearly identical performance to the OLS model, with differences observable only at the third or fourth decimal place. This indicates that multicollinearity effects are limited and that regularization does not provide a meaningful advantage for this dataset. Consequently, the standard linear regression model is sufficient for capturing the dominant price formation mechanisms while maintaining maximum interpretability.

The relationship between actual and predicted apartment prices is illustrated in *Figure 3*. Most observations lie close to the 45° reference line, confirming good agreement between predicted and observed values. Increased dispersion is observed for higher-priced apartments, reflecting greater heterogeneity and the presence of unobserved qualitative attributes such as interior quality or neighborhood amenities.

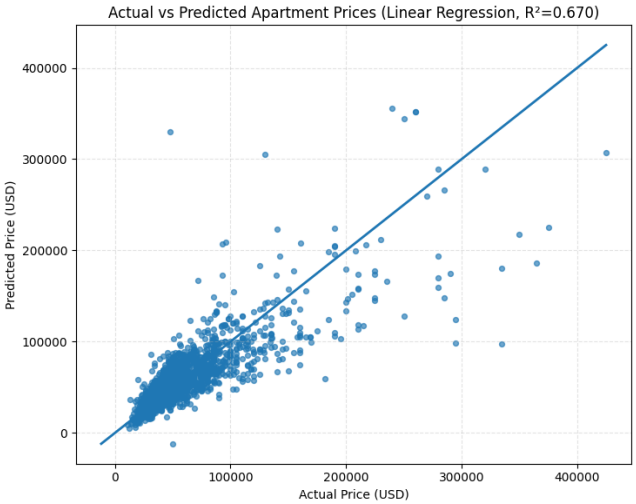

Figure 3. Scatter plot of actual versus predicted apartment prices for the linear regression model.

To support interpretability, the estimated coefficients of the linear regression model for numeric explanatory variables are presented in *Figure 4*. Apartment size exhibits a positive coefficient, confirming its role as one of the primary determinants of housing prices. Geographic coordinates (latitude and longitude) display the largest coefficient magnitudes, highlighting the dominant influence of spatial location on price variation across the city.

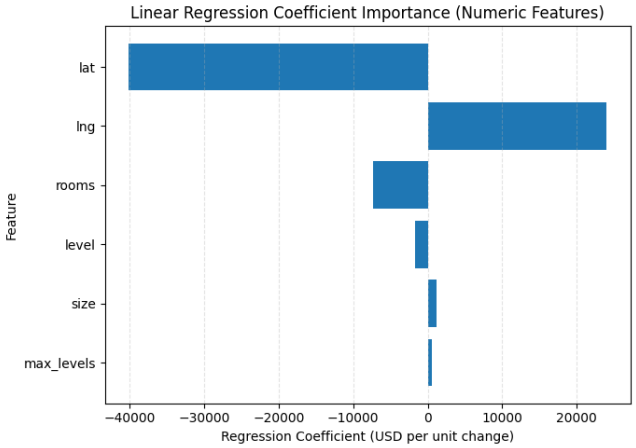

Figure 4. Estimated coefficients of the linear regression model for numeric explanatory variables.

The number of rooms shows a negative coefficient when apartment size is held constant, indicating that, for a fixed total area, a higher number of rooms may correspond to smaller individual room sizes and reduced perceived value. Floor level and total building height exhibit smaller but consistent effects, suggesting secondary structural influences on apartment prices.

It is important to note that the coefficients associated with latitude and longitude do not represent causal geographic effects, rather, they serve as spatial proxies capturing neighborhood-level price gradients and location-specific characteristics.

The distribution of prediction residuals for the linear regression model is shown in *Figure 5*. The residuals are centered around zero and exhibit an approximately symmetric distribution, indicating the absence of systematic prediction bias. Slightly heavier tails are observed, particularly for higher-priced apartments, which is consistent with the increased heterogeneity and unobserved attributes associated with the luxury segment of the housing market.

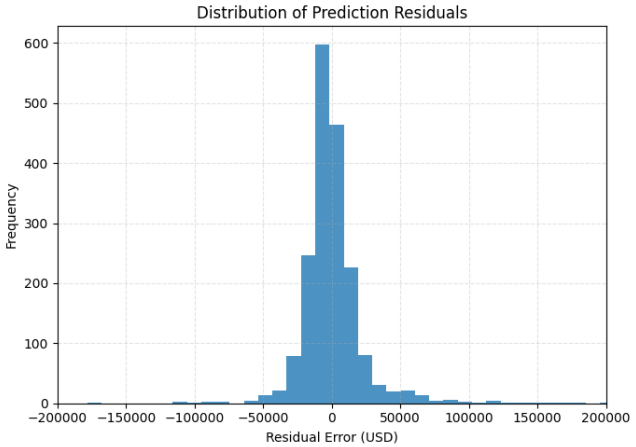

Figure 5. Distribution of prediction residuals for the linear regression model.

Overall, the predictive performance metrics, coefficient analysis, and residual diagnostics demonstrate that linear regression provides a stable, interpretable, and reliable baseline for analyzing residential apartment prices in the Tashkent housing market.

CONCLUSION

This paper presented an interpretable analysis of residential apartment prices in Tashkent using multiple linear regression applied to real listing data from an online real estate platform. By incorporating both structural apartment characteristics and geographic coordinates, the study examined key factors influencing housing price formation in an emerging urban market context.

The experimental results demonstrate that linear regression models explain a substantial proportion of price variability, achieving an R^2 value of approximately 0.67 on the test dataset. Comparisons with regularized linear variants show that Ridge, Lasso, and Elastic Net regressions provide no significant performance improvement over ordinary least squares, indicating that the underlying relationships between explanatory variables and apartment prices are largely linear and stable. The analysis highlights apartment size and spatial location as the dominant determinants of price variation, while other structural attributes exhibit secondary but consistent effects.

Residual diagnostics confirm the absence of systematic prediction bias, with increased variability observed for higher-priced apartments due to greater heterogeneity and unobserved qualitative factors. Importantly, the use of linear regression enables transparent interpretation of individual feature effects, offering practical insights for buyers, sellers, and real estate analysts.

Overall, the findings suggest that simple and interpretable linear regression models provide a reliable baseline for residential price analysis in the Tashkent housing market. Future work may extend this study by incorporating nonlinear machine learning models, explicit spatial regression techniques, or temporal dynamics to further improve predictive performance and capture more complex market behavior.

REFERENCES

1. Q. Zhang. “Housing Price Prediction Based on Multiple Linear Regression.” Scientific Programming, vol. 2021, no. 1, pp. 1-9, 2021, doi:10.1155/2021/7678931.
2. M.A.K. Shukla. “House Price Prediction Using Regression”. *International Journal For Science Technology And Engineering*, vol. 10, no. 2, 2022, pp. 344-346, doi: 10.22214/ijraset.2022.40272.
3. V. Amaresh, R.R. Singh, R. Kamal; A. Kulkarni. “Linear Regression Models Based Housing Price Forecasting”. *2022 International Conference on Industry 4.0 Technology (I4Tech)*, pp. 1-5, 2022, doi: 10.1109/i4tech55392.2022.9952397.
4. T. Mao. “Real Estate Price Prediction Based on Linear Regression and Machine Learning Scenarios”. *BCP Business & Management*, vol. 38, pp. 400-409, 2022, doi: 10.54691/bcpbm.v38i.3720.
5. X. Xu. “The real estate price prediction of US prediction based on multi-factorial linear regression models”. *BCP Business & Management*, vol. 36, pp. 1-6, 2023, doi: 10.54691/bcpbm.v36i.3378.
6. H. Prakash, K. Kanaujia, S. Juneja. “Using Machine Learning to Predict Housing Prices”. *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, pp. 1353-1357, 2023, doi: 10.1109/AISC56616.2023.10085264.
7. Verma, P. Gera, S. Singhal, A.K. Mohapatra. “Advanced Regression Models for Accurate House Price Prediction: An Analysis of Performance and Interpretability”, *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1-7, 2023, doi: 10.1109/icccnt56998.2023.10307946.
8. Z. Liu, Y. Wang. “City’s House Pricing Prediction Based on Machine Learning Algorithms”. *International Symposium on Robotics, Artificial Intelligence, and Information Engineering (RAIIE 2022)*, vol. 12454, pp. 1-5, 2022, doi: 10.1117/12.2658679.
9. Z. Liu. “Real Estate Price Prediction Based on Supervised Machine Learning Scenarios”. *Highlights in Science, Engineering and Technology*, vol. 39, pp. 731-737, 2023, doi: 10.54097/hset.v39i.6637.
10. N. Chuhan. “House price prediction based on different models of machine learning”. *Applied and Computational Engineering*, vol. 49, pp. 47-57, 2024, doi: 10.54254/2755-2721/49/20241058.
11. S.J. Wawge. “Evaluating Machine Learning and Deep Learning Models for Housing Price Prediction: A Review.” *International Journal of Advanced Research in Science, Communication and Technology*, vol. 5, no. 11, pp. 367-377, 2025, doi:10.48175/ijarsct-25857.