NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science
Bachelor's Programme "Data Science and Business Analytics"

Research Project Report on the Topic:
Analysis of the impact of real estate location on the value of the property

Submitted by the Student:

group #БПАД233, 2nd year of study Miniaitsev Ivan Vasilyevich

Approved by the Project Supervisor:

Bashminova Daria Aleksandrovna
Senior Teacher
Faculty of Computer Science, HSE University

Moscow 2025

Contents

Abstract

This study investigates the impact of geographic location of real estate properties on their market value using the Russian market as an example, with an emphasis on Moscow. I use a dataset containing more than 2.8 million apartment sale listings from various regions of Russia. A comprehensive exploratory data analysis (EDA) is performed, including handling of missing values, outlier detection, and visualization of distributions of key variables. Features considered include: total area, number of rooms, floor level, building type, average street price, coordinate cluster, and distance to the city center. Two regression models are built: linear regression and XGBoost. Quality metrics (RMSE and MAE) are evaluated on a sample for Russia as a whole and separately for Moscow. It is found that in Moscow the key factors are building type and distance to center, while nationwide area and region have the strongest influence on price. Hypotheses regarding the impact of average street price, building type and new/secondary markets are also tested. The work concludes with remarks on model applicability and recommendations for further research.

1 Introduction

The real estate market is one of the key sectors of the economy, influencing both the population's standard of living and strategic investment decisions. Understanding the factors that determine apartment prices helps not only analysts and investors, but also government bodies in forming price policies and regulating the market. Traditional real estate valuation methods often rely on expert opinions and outdated statistical data, which leads to subjectivity and inaccuracies. Modern approaches based on machine learning allow for uncovering hidden patterns in large volumes of data and improving forecasting accuracy.

The aim of this study is to analyze the impact of geographic location of real estate properties on their prices in Russia, and separately in Moscow. The objectives of this work are:

1. Collect and preprocess a large volume of apartment sale data.
2. Perform exploratory data analysis (EDA) to identify key patterns.
3. Formulate and test hypotheses about the main pricing factors.
4. Build and compare linear regression and XGBoost models for price prediction.
5. Identify differences in factors affecting prices between Moscow and other regions.

The results of this research will help to better understand which features most influence real estate prices in different parts of the country and will serve as a basis for developing automated property valuation systems.

2 Literature Overview

In recent years, predicting real estate prices has become a popular task in Data Science and machine learning. Both classic econometric approaches (hedonic models, fixed-effects regressions) and modern methods based on decision trees, ensembles, and neural networks are used.

2.1 Classical Approaches

One of the basic methods is *linear regression* [1], where a property's price is modeled as a linear combination of features: area, number of rooms, year of construction, distance to center, etc. Despite its simplicity, this method shows limited flexibility when there are strong nonlinearities and feature interactions.

Hedonic pricing models [2] assume that a price is determined by characteristics of the property itself and its environment. An important aspect is accounting for spatial dependency: proximity to city center, infrastructure level, environmental and social indicators of the city.

2.2 Machine Learning–Based Methods

With the growth of computing power, models such as *Random Forest*, *Gradient Boosting* (XGBoost, LightGBM, CatBoost) have emerged, which effectively handle large numbers of features and nonlinear relationships [4]. In [4], it is shown that XGBoost demonstrates high accuracy in price predictions.

In [3], a comparative analysis of EDA and Ridge, Lasso, and Elastic Net models was performed; it was found that regularized regressions reduce the impact of multicollinearity and overfitting.

The KNN (k-Nearest Neighbors) method is also used for price estimation: the forecast is based on the average prices of k geographically closest properties.

2.3 Geographic Influence

Studies [4], [2] emphasize that location is one of the key factors. In [5], the influence of infrastructure (proximity to subway stations, shopping centers) on price was studied, showing that adding spatial indicators improves model quality.

Other research [5], [3] focuses on major cities (New York, London) and finds that distance to center, neighborhood prestige, and availability of leisure facilities significantly influence price.

2.4 Conclusion of Review

The existing literature confirms the importance of using modern machine learning models and integrating geographic features (coordinates, distance to center, clustering) to improve prediction accuracy. This work builds on those approaches but focuses specifically on the Russian market and includes a separate analysis for Moscow.

3 Data Collection

Data for this study were collected from open sources of real estate sale listings in Russia, available on Kaggle. The main dataset includes more than 2.8 million records from regional portals and aggregators. Each listing contains the following key features:

- `price` (in rubles) — property price;
- `area` (in square meters) — total area;
- `rooms` — number of rooms;
- `level` — floor on which the apartment is located;
- `levels` — total number of floors in the building;
- `kitchen_area` — kitchen area;
- `building_type` — building type (1=panel, 2=brick, 3=monolithic, etc.);
- `object_type` — new building (2) or secondary market (0);
- `id_region` — region code;
- `street_id` — unique street identifier;
- `geo_lat`, `geo_lon` — coordinates of the property.

3.1 Data Preprocessing

1. **Missing Value Handling.** The features `price`, `area`, `rooms`, and `level` had the most missing values—records without these values were removed. Missing values in `building_type` were filled with 0 and subsequently removed as invalid.
2. **Outlier Adjustment.** Prices above the 99th percentile and areas below 10 m^2 were considered outliers and excluded as they were clearly unrepresentative. For visualizations, apartments with area over 200 m^2 (likely luxury apartments) were excluded to avoid skewing and to better reveal the general patterns.
3. **Street Aggregation.** I computed `street_avg_price`—the average price per `street_id`:
$$\texttt{street\_avg\_price} = \frac{1}{N_{\text{street}}} \sum_{i \in \text{street}} \texttt{price}_i.$$
4. **Price per Square Meter.** I computed `price_per_m2`—the price per square meter for each apartment:
$$\texttt{price\_per\_m2} = \frac{\texttt{price}}{\texttt{area}}.$$
5. **Splitting into Subsamples.** The dataset was split into two subsets:
   - **df_moscow** — records with `id_region` = 77 (Moscow).
   - **df_not_moscow** — all other regions.
6. **K-Means Clustering of Moscow Coordinates.** For records in Moscow (`id_region` = 77), I applied K-Means clustering on coordinates (`geo_lat`, `geo_lon`) to identify $n = 10$ clusters. The feature `geo_cluster` denotes the cluster label.

7. **Distance to Moscow Center.** For Moscow records, I computed:

$$\texttt{distance\_to\_center} = \sqrt{(\texttt{geo\_lat} - 55.7539)^2 + (\texttt{geo\_lon} - 37.6208)^2}.$$

4 Hypothesis Formulation

Based on the EDA and knowledge of the Russian market, I formulated the following main hypotheses:

1. **Hypothesis 1:** Average street price (`street_avg_price`) significantly affects the price of an individual property.
2. **Hypothesis 2:** Building type (`building_type`) has a significant influence on price: monolithic buildings are more expensive than panel buildings.
3. **Hypothesis 3:** New buildings (`object_type` = 2) have higher price per square meter than secondary market properties (`object_type` = 0).

5 Models

Two main regression models are built for predicting apartment prices:

5.1 Linear Regression

Linear regression assumes that the dependent variable y (price) can be approximated by a linear combination of features X:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

Features used for Moscow:

$$X = \{\texttt{area}, \texttt{rooms}, \texttt{level}, \texttt{building\_type}, \texttt{object\_type},$$

$$\texttt{street\_avg\_price}, \texttt{geo\_cluster}, \texttt{distance\_to\_center}\}.$$

Features for other regions:

$$X = \{\texttt{area}, \texttt{rooms}, \texttt{level}, \texttt{building\_type}, \texttt{object\_type}, \texttt{street\_avg\_price}\}.$$

Categorical variables (`building_type`, `object_type`, `geo_cluster`) are converted to dummy variables using `pd.get_dummies()` to avoid multicollinearity. The model is trained on the training set using ordinary least squares (`LinearRegression` from `sklearn`).

5.2 XGBoost Regressor

XGBoost (eXtreme Gradient Boosting) is an ensemble model based on gradient boosting over decision trees. Training involves adding trees iteratively to minimize the loss function. Key hyperparameters:

- `n_estimators` = 100 (number of trees),
- `max_depth` = 6 (maximum tree depth),
- `learning_rate` = 0.1 (learning rate).

6 Metrics

To evaluate model quality, I used the following standard regression metrics:

- **RMSE** (Root Mean Squared Error):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}.$$

  RMSE is sensitive to large errors and is measured in the same units as the target variable (rubles).

- **MAE** (Mean Absolute Error):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|.$$

  MAE reflects the average absolute error regardless of direction and is more robust to outliers.

For each dataset (Russia as a whole and Moscow separately), both metrics are computed on the test set.

7 Hypothesis Testing

7.1 Hypothesis 1: Influence of Average Street Price (`street_avg_price`) on Property Price

Hypothesis Statement: Average street price (`street_avg_price`) significantly affects the price of a specific apartment.

Method: Performed a log-linear regression:

$$\log(\text{price}) = \beta_0 + \beta_1 \log(\text{street\_avg\_price}) + \varepsilon.$$

Regression Results:

Table 1: Log-Linear Regression Results: $\log(\text{price}) \sim \log(\text{street\_avg\_price})$

Parameter	Estimate	Std Err	t-Statistic	p-value
β_0 (Intercept)	2.31	0.01	231.00	< 0.001
β_1 (Coefficient for log(street_avg_price))	0.52	0.00	1448.00	< 0.001

$R^2 = 0.48$, $N = 2{,}344{,}254$.

Interpretation: Coefficient $\beta_1 \approx 0.52$ with $p < 0.001$ implies that a 1% increase in average street price corresponds to a 0.52% increase in the individual apartment price on average. The R^2 of 0.48 indicates that about 48% of the variation in log(price) is explained by log(street_avg_price). Thus, the hypothesis is confirmed: properties on more expensive streets indeed have higher prices.

Figure: Figure 1 shows points from a random sample of 20,000 properties and the regression line in log-space.

Figure 1: Relationship between $\log(\text{price} + 1)$ and $\log(\text{street\_avg\_price} + 1)$ (Russia). The red line represents the log-linear regression.

7.2 Hypothesis 2: Difference in Average Price per m^2 by Building Type

Hypothesis Statement: Different `building_type` categories exhibit statistically significant differences in average price per square meter.

Method: Grouped data by `building_type` and calculated the median or mean `price_per_m2` for each group. Performed a one-way ANOVA test to compare all groups simultaneously.

Groups:

- `building_type = 1` — Others
- `building_type = 2` — Panel
- `building_type = 3` — Monolithic
- `building_type = 4` — Brick
- `building_type = 5` — Block
- `building_type = 6` — Wooden

ANOVA Results:

Table 2: ANOVA Results for `building_type` Groups

Statistic	Value
F-statistic	60,978.89
p-value	< 0.0001

Interpretation: The high F-statistic of 60,978.89 and $p < 0.0001$ indicate that there are statistically significant differences in average price per m^2 among the `building_type` categories. Figure 2 displays a bar chart showing the average price per m^2 by building type.

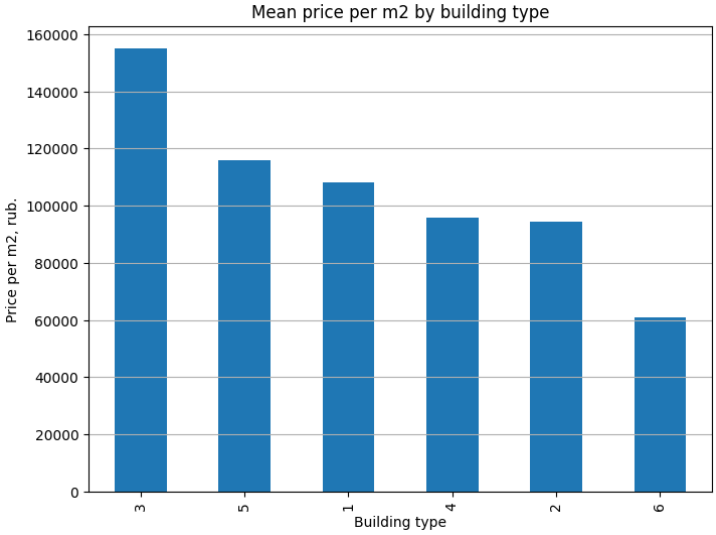


Figure 2: Average price per m^2 by building type (Russia).

Conclusion: Monolithic buildings (`building_type = 3`) have the highest average price per m^2, while panel buildings (`building_type = 2`) have the lowest. Hence, the hypothesis that building type influences price is validated.

7.3 Hypothesis 3: New Buildings (`object_type` = 2) vs Secondary Market (`object_type` = 0)

Hypothesis Statement: Apartments in new buildings (`object_type` = 2) have higher average price per m^2 than those on the secondary market (`object_type` = 0).

Method: Selected two groups:

$$\text{New Buildings: } \{\texttt{object\_type} = 2\},$$
$$\text{Secondary Market: } \{\texttt{object\_type} = 0\}.$$

Computed sample means $\overline{x}_{\text{new}}$ and $\overline{x}_{\text{sec}}$ with their standard errors. Then performed a Mann–Whitney U test to verify the difference between the two distributions.

Comparison of Means (Russia):

Table 3: Comparison of Average Price per m^2 for New Buildings and Secondary Market (Russia)

Parameter	New Buildings (`object_type` = 2)	Secondary Market (`object_type` = 0)
Mean ± SE, RUB/m^2	$94{,}796 \pm 59$	$76{,}832 \pm 23$
Mann–Whitney $U = 5.0657 \times 10^{11}$, $p < 0.0001$.		

Interpretation: The average price per m^2 for new buildings ($\approx$ 94,796 RUB/m^2) significantly exceeds that of the secondary market ($\approx$ 76,832 RUB/m^2). The Mann–Whitney U statistic of 5.0657×10^{11} with $p < 0.0001$ confirms that this difference is statistically significant.

Figure: Figure 3 shows boxplots comparing distributions of price per m^2 in the "New Buildings" and "Secondary Market" groups.

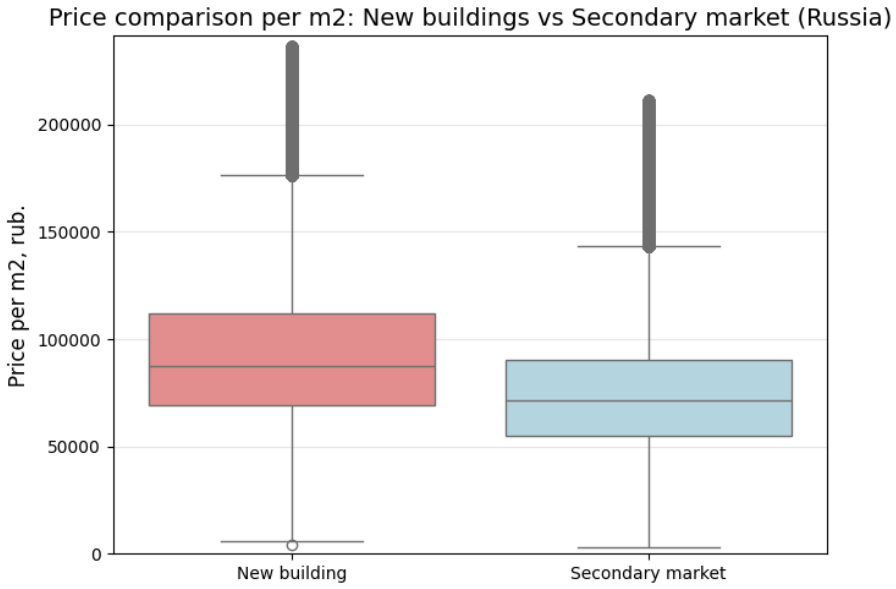


Figure 3: Comparison of Price per m^2: Secondary Market vs New Buildings (Russia).

Conclusion: The hypothesis that new buildings are more expensive per square meter than secondary market properties is confirmed with high statistical significance ($p < 0.0001$). The difference in sample means is more than 17,964 RUB/m^2, which is substantial for the real estate market.

8 Model Results and Metrics

Below are the results of training and testing models on two datasets: `df_not_moscow` (all of Russia except Moscow) and `df_moscow` (only Moscow).

8.1 Data Preparation

Prices above the 95th percentile and below the 5th percentile in other regions, and above the 90th percentile and below the 5th percentile in Moscow, were considered outliers and excluded, as they significantly distorted calculations and weakened predictive performance.

For Moscow, the features used were:

$$X = \{\texttt{area}, \texttt{rooms}, \texttt{level}, \texttt{building\_type}, \texttt{object\_type}, \texttt{street\_avg\_price}, \texttt{geo\_cluster}, \texttt{distance\_}$$

For other regions, the features used were:

$$X = \{\texttt{area}, \texttt{rooms}, \texttt{level}, \texttt{building\_type}, \texttt{object\_type}, \texttt{street\_avg\_price}\}.$$

Categorical variables were converted to dummy variables. A random 80%/20% train/test split was then performed.

8.2 Results for Russia (excluding Moscow)

Model	RMSE, RUB	MAE, RUB
Linear Regression	1,381,043	1,005,419
XGBoost Regressor	750,114	530,922

Table 4: Model performance on the *Russia excluding Moscow* dataset

XGBoost achieved an RMSE improvement of about 800,000 RUB compared to linear regression. The error distributions are shown in Fig. 4.

```
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)
y_pred_lr = lr_model.predict(X_test)

rmse_lr = np.sqrt(mean_squared_error(y_test, y_pred_lr))
mae_lr = mean_absolute_error(y_test, y_pred_lr)

print("Basic model (Linear Regression)")
print("RMSE:", str(int(rmse_lr)))
print("MAE:", str(int(mae_lr)))
```

```
Basic model (Linear Regression)
RMSE: 1381043
MAE: 1005419
```

(a) Linear Regression

```
[45] xgb_model = XGBRegressor(n_estimators=100, max_depth=6, learning_rate=0.1)
     xgb_model.fit(X_train, y_train)
     y_pred_xgb = xgb_model.predict(X_test)

     rmse_xgb = np.sqrt(mean_squared_error(y_test, y_pred_xgb))
     mae_xgb = mean_absolute_error(y_test, y_pred_xgb)

     print("\nAdvanced model (XGBoost)")
     print("RMSE:", str(int(rmse_xgb)))
     print("MAE:", str(int(mae_xgb)))
```

```

Advanced model (XGBoost)
RMSE: 750114
MAE: 530922
```

(b) XGBoost

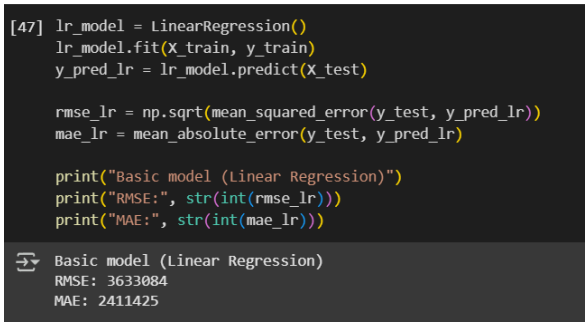
Figure 4: Residuals of the models on the *Russia excluding Moscow* dataset

8.3 Results for Moscow

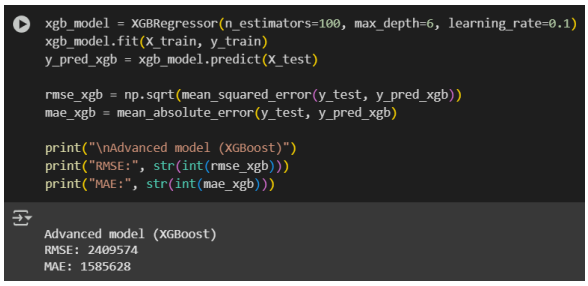
Model	RMSE, RUB	MAE, RUB
Linear Regression	3,633,084	2,411,425
XGBoost Regressor	2,409,574	1,585,628

Table 5: Model performance on the *Moscow* dataset

XGBoost achieved an RMSE improvement of about 600,000 RUB compared to linear regression. The error distributions are shown in Fig. 5.



(a) Linear Regression



(b) XGBoost

Figure 5: Residuals of the models on the *Moscow* dataset

9 Limitations

This study has the following limitations:

- **Data Quality.** The data are collected from public sources, so input errors, outdated listings, or missing attributes may exist.
- **Incomplete Features.** There is no information on infrastructure: proximity to subway, number of schools and kindergartens, environmental quality.
- **Outliers and Extreme Values.** Despite filtering, extremely expensive apartments may remain in the test set, distorting the metrics.
- **Sample Segmentation.** Moscow and other regions are analyzed separately, but internal heterogeneity within cities (districts, neighborhoods) is not accounted for in detail.

10 Future Work

For further development of this project, I recommend:

- **Incorporate Textual Data.** Include analysis of listing descriptions (NLP) to extract keywords ("renovated", "subway nearby", "new building").
- **Enhance Geographic Features.** Use actual road distances, data on proximity to subway, traffic; apply DBSCAN for identifying dense neighborhoods.
- **Additional Models.** Test CatBoost, LightGBM, neural networks; compare their performance against XGBoost.
- **Temporal Analysis.** Add a time feature (listing date) to analyze price trends and seasonal fluctuations.

11 Conclusion

This work presents a comprehensive analysis of the Russian real estate market, taking into account geographic location of properties. The main conclusions are:

1. Nationwide, the most significant price drivers are area, number of rooms, and average street price.
2. In Moscow, in addition to these factors, distance from the city center plays a crucial role.
3. Properties on more expensive streets have higher prices than comparable properties on cheaper streets.
4. The highest price per m^2 is observed in brick buildings, while wooden buildings have the lowest.
5. New buildings are significantly more expensive per square meter than secondary market properties.

6. XGBoost outperforms linear regression in terms of RMSE and MAE for both the overall dataset and for Moscow specifically, due to its better handling of nonlinearities and outliers.

The findings can be applied to develop automated property valuation systems, assist analysts in investment decision-making, and build tools for mortgage scoring.

References

[1] Chih-Ling Tsai, Xiaogang Su, Xin Yan, *Linear Regression*, WIREs Computational Statistics, 2012.

[2] Homer Erekson, Ann D. Witte, Howard J. Sumka, *An Estimate of a Structural Hedonic Price Model of the Housing Market*, Econometrica, 1979.

[3] G. Dwilestari, F. M. Basysyar, *House Price Prediction Using Exploratory Data Analysis and Machine Learning with Feature Selection*, Acadlore Trans. Mach. Learn., 2022.

[4] Shuhui Shi, Yaping Zhao, Ramgopal Ravi, *PATE: Property, Amenities, Traffic and Emotions Coming Together for Real Estate Price Prediction*, 2022.

[5] M. S. Aditya Narhari Khobragade, N. Maheswari, *Analyzing the Housing Rate in a Real Estate Informative System: A Prediction Analysis*, IAEME, 2018.