PRAGUE UNIVERSITY OF ECONOMICS AND BUSINESS

Acta Informatica Pragensia
2023, Volume 12, Issue 2, pp. 275–295
https://doi.org/10.18267/j.aip.215

Article Open Access

Use of Data Mining for Analysis of Czech Real Estate Market

Ilya Tsakunov, David Chudán

Faculty of Informatics and Statistics, Prague University of Economics and Business, Prague, Czech Republic

Corresponding author: Ilya Tsakunov (tsai00@vse.cz)

Abstract

This paper analyses data from the real estate market domain. The data were scraped from the bezrealitky.cz portal. The analysis looks at both sales and rental data. A total of 3546 records and 54 attributes were obtained. A basic overview of the data was performed using exploratory data analysis where some basic characteristics of the data were identified, such as the average price of sold and rented flats. More specific results were obtained by applying data mining methods such as regression (linear regression, lasso regression and ridge regression) for predicting the flat prices and payments for utilities, classification (support vector machines, KNN, Gaussian naïve Bayes, decision tree and random forest) for estimating the PENB class (building energy performance certificate) and building condition. Lasso regression performed the most successfully ($R^2 = 0.76$) in predicting the rent price. Among the classification tasks, the best result was achieved with random forest, which had an accuracy over 80% in some cases. Other tasks included clustering (k-means and k-modes) and anomaly detection (isolation forest). The main focus was on descriptive data mining, especially on clustering. Clusters created using the k-means algorithm (silhouette score of 0.78) with flats based on geographic coordinates were identified which show that the most expensive flats are on average in Bohemian regions, followed by Silesia and the cheapest are in central Moravia. Another cluster application identified flats in the Moravian-Silesian region with very high payments for utilities (silhouette score of 0.56). The models can help estimate the value of flats based on their attributes as well as location.

Keywords

Data mining; Web scraping; Real estate market; Exploratory analysis.

Citation: Tsakunov, I., & Chudán, D. (2023). Use of Data Mining for Analysis of Czech Real Estate Market. *Acta Informatica Pragensia*, 12(2), 275–295. https://doi.org/10.18267/j.aip.215

Academic Editor: Stanislava Mildeova, University of Finance and Administration, Czech Republic

1 Introduction

Real estate prices have been rising very significantly in recent years. In the European Union, the growth between 2010 and 2022 was 45% on average, and the Czech Republic achieved the fourth highest growth in the EU of more than 120% (Eurostat, 2022). Housing affordability is becoming increasingly problematic even for higher income groups. This is why the real estate market domain of the Czech Republic is relatively interesting from the point of view of data analysis, since it may hide some interesting relationships. Discovering similar relationships is one of the tasks of data mining. There are various methods of applying data mining techniques to real estate data, including regression analysis, decision tree models and clustering algorithms. These methods can be used to predict property values, identify potential buyers or sellers and gain a better understanding of the real estate market. One key advantage of using data mining in real estate is the ability to make data-driven decisions.

Throughout the article, the CRISP-DM methodology is used. CRISP-DM stands for cross-industry standard process for data mining and describes how data mining projects can be implemented in a 6-step process: business understanding, data understanding, data preparation, modelling, evaluation and deployment (Martinez-Plumed et al., 2021). The CRISP-DM methodology is considered a successor to the knowledge discovery in databases (KDD) process, which was first introduced in 1989 (Fayyad et al., 1996). In terms of data mining methods, both classification/regression and descriptive methods are used. Regression methods, such as linear, lasso (Tibshirani, 1996) and ridge regression (Hoerl & Kennard, 1970) are used, classification methods such as support vector machines (Cortes & Vapnik, 1995), KNN (Cover & Hart, 1967), Gaussian naïve Bayes, decision trees (Quinlan, 1986) and decision forest (Ho, 1995) are used. From descriptive methods, clustering using k-means (MacQueen, 1967) and k-modes (Huang, 1998) is used as well as anomaly detection based on isolation forest (Liu et al., 2008).

Data mining models can be evaluated using different metrics. These metrics will vary depending on the model type. For instance, for measuring the performance of regression models, it is common to use metrics such as mean squared error and R-squared, while accuracy, precision, recall, F1 score and ROC AUC are usually used when evaluating classification models. Several studies have been conducted on comparison of model evaluation metrics, such as Hossin and Sulaiman (2015), which is a study of various metrics for evaluating the performance of classification models, and newer studies (Plevris et al., 2022; Thevaraja et al., 2019) which compare the metrics in regression analysis. Clustering is evaluated using the silhouette score (Rousseeuw, 1987).

Before the actual analysis, however, it is necessary to obtain data. One option to do that is to find a prepared dataset collected before, but such data will most probably be outdated. Therefore, to analyse the most recent data from a large Czech real estate portal, the web scraping approach was chosen. The aim of this work is thus to extract data from one of the leading Czech real estate portals and subsequently analyse these data using exploratory analysis and data mining methods. Retrieved data are analysed using the Python programming language along with appropriate libraries such as Pandas and Matplotlib, while modelling is performed with the help of the Scikit-learn library.

Real estate market data are a relatively rewarding topic for data mining. It is possible to find many different articles from various countries analysing their specific real estate markets (Louati et al., 2022; Oliveira et al., 2021; Sawant et al., 2018; Verma et al., 2022). There are relatively few articles in the Czech environment, most of them come from Eduard Hromada from the Faculty of Civic Engineering of the Czech Technical University in Prague. One of the available articles is Mapping of Real Estate Prices Using Data Mining Techniques (Hromada, 2015). The article is dedicated to a description of the software application called EVAL, which systematically gathers, analyses and evaluates real estate price offers advertised via real estate servers. Information was gathered automatically from 2007 to 2014, resulting in 650,000 entries with price offers for purchase or rental of flats, houses, commercial spaces and allotments.

The software enables automatic export of specific data from downloaded advertisements and then analysis of data in the analysis module. Other articles by the author are also based on this system, such as Analysis of Relationship Between Market Value of Property and its Distance from Centre of Capital (Hromada, 2018), or more recently Development of the Real Estate Market in the Czech Republic in Connection with the COVID-19 Pandemic (Hromada, 2021). In addition to scientific publications, commercial consulting companies are also involved in mapping the real estate market. For example, Deloitte presents brief reports on market developments on a quarterly basis (Deloitte, 2022a). The article extends the author's bachelor thesis (Tsakunov, 2022) by adding relevant literature research, summarizing key points from the thesis and expanding the discussion regarding the results.

In this article, we apply several data mining methods on real data from the real estate market domain. We can identify two objectives: perform data mining to get interesting results from the data and to compare the resulting metrics, such as accuracy of classification tasks or silhouette score on clustering of individual data mining methods.

2 Data Collection

There are several ways to find information about real estate. They include web pages of particular real estate agencies. Another way is to get information from portals which aggregate individual properties. In the Czech Republic, more such portals can be found, but the most popular are sreality.cz and bezrealitky.cz. For the purpose of further analysis, the latter portal was selected[1]. This portal offers advertisements directly, without mediation through real estate agencies. Data collection was performed on 22 March 2022.

The scraping itself was done using Python and the appropriate libraries, such as Selenium and Requests. The high-level data collection process included retrieving listing links from the portal. This was done via the API found on the web page. The low-level scraping was then targeted at scraping the details of properties collected at the previous level. Commercial properties as well as properties outside the Czech Republic were excluded. As a result, a dataset with more than 3500 records was received. These records could be basically divided into two groups – properties for sale (1237 records) and properties for rent (2456 records).

3 Data Understanding

The original dataset contains 3546 records and 54 attributes. However, some columns (price, surface area and layout) appeared to be duplicated due to merging datasets from high-level and low-level scraping. Other columns (developer, internal ID, administration fee, etc.) were either irrelevant, or were missing almost in all records. Both duplicated and irrelevant columns were removed, reducing the number of attributes to 35. The list of attributes used for further analysis can be found in Table 1.

Table 1. *Overview of attributes and their data types.*

Attribute	Data type
id	int
uri	string
url	string
lat	float
lng	float

[1] The reason was that the terms and conditions of the sreality.cz portal do not allow scraping of content.

Attribute	Data type
price	int
currency	string
key_offer_type	string
key_layout	string
surface	int
utilities	string
facilities	string
floor	float
balcony	string
terrace	string
cellar	string
loggia	string
parking	string
elevator	string
garage	string
public_transport_stop	string
post_office	string
store	string
bank	string
restaurant	string
pharmacy	string
school	string
kindergarten	string
sport_field	string
playground	string
condition	string
property_type	string
building_type	string
penb	string
deposit	string

Both numeric and categorical attributes can be found in the dataset. Some of the attributes, such as deposit and utilities, could only be applied to properties for rent. Another group of attributes, including *balcony, terrace, cellar, loggia, parking, elevator* and *garage*, could only acquire the values Yes or No, so they are dichotomous variables. Besides, there are ten attributes related to the distance from property to a particular location (post office, school, grocery shop, etc.), which are, however, stored as strings, because of they have units.

It is also good to keep in mind how residential properties are classified in the Czech Republic. From a layout point of view, the property can be marked 1+kk or 1+1, 2+kk or 2+1 etc. The difference is that properties with the "kk" suffix have a kitchen corner combined with the living room, while those marked with "+1" have a kitchen as a separate space.

4 Data Preprocessing

It was found that in addition to properties for sale and rent, there were also properties looking for roommates in the dataset. Such records were removed. Properties in Slovakia and properties with a price in euros were removed as well. Some ads contained meaningless information about the surface area of the property, so they had a value of 0 for this attribute; therefore, these records were also removed.

Some categorical attributes (*balcony*, *terrace*, *loggia*, *elevator*, *garage*, *parking* and *cellar*) can take only two values (Yes or No), so it was convenient to convert them to numerical values of 1 and 0 in order to work with them in modelling.

Attributes describing the distance to individual locations are represented by text, so it was necessary to convert them into numerical values with respect to measurement units. This means that if "km" occurs in the attribute, the resulting value must be multiplied by 1000. Attributes containing currencies can also be converted to numeric values by removing the text string “Kč”.

In the obtained dataset, the number of flats with a layout of 5+kk, 5+1, 6+kk, 6+1, 7+kk and "other" totalled fewer than 35, which is why they were merged into one category “Other”. As shown in Table 2, some of the records had missing values, which had to be completed before the analysis. This was done using different methods depending on the attribute.

Table 2. *Overview of missing attributes.*

Attribute	Total records	Sale	Rent
key_layout	1	1	0
utilities	1273	994	279
facilities	439	123	316
floor	134	34	100
post_office	1	0	1
bank	67	28	39
pharmacy	14	3	11
school	12	4	8
kindergarten	42	16	26
sport_field	35	14	21
playground	5	2	3
condition	434	17	417
property_type	115	0	115
building_type	108	0	108
penb	575	128	447
deposit	1208	994	214

For instance, missing values of the attribute *floor* were replaced with random values from the range [1, 5]. Such a range was chosen due to the fact that almost 85% of the properties in the dataset are situated on floors 1–5.

On the other hand, the attributes *utilities* and *deposit* also contained missing values. Most of them occurred in properties for sale and are the illustration of an MAR (missing at random) situation, meaning that they are related to the value of another attribute – type of property (rent or sale) in this case. Missing values of the attribute *deposit* were filled with 1.6 * price, where 1.6 is a median value of deposit by price across the whole dataset.

Besides missing values, the dataset also contained outliers. The presence of outliers could be found, e.g., in the columns *price* and *utilities*, as shown in Figure 1.

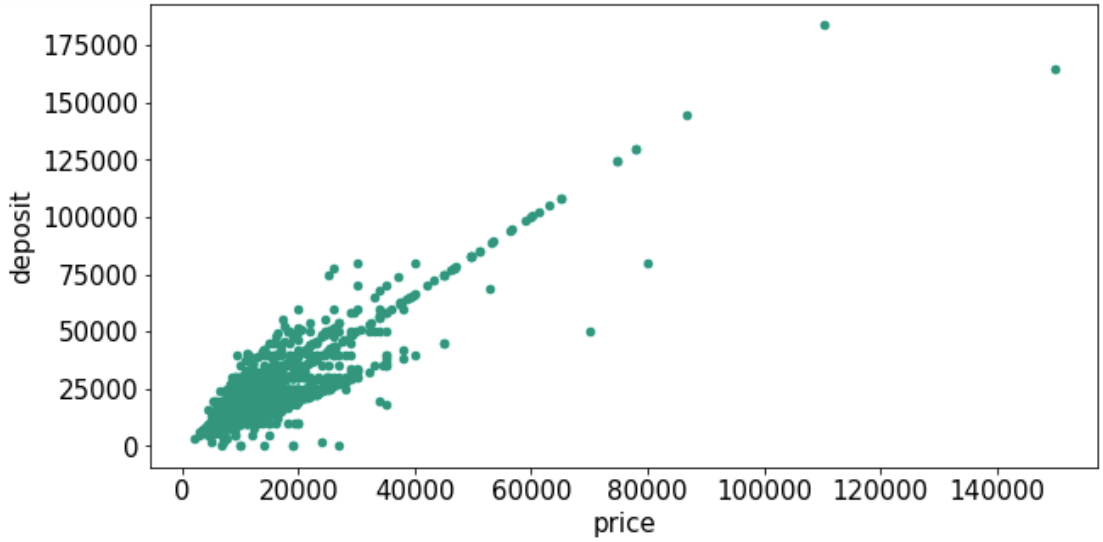

Figure 1. *Outliers in deposit/price.*

The outliers were identified using the z-score metric, which basically defines how many standard deviations away the actual value is from the mean value. The threshold value was set to 3. As for rent data, 23 outliers were identified within the attribute *deposit_by_price*. Records containing outliers were excluded from the dataset. Table 3 shows some examples of identified outliers.

Table 3. *Examples of outliers in rent data.*

price, CZK	deposit, CZK	deposit_by_price
7,500	25,000	3.333333
8,400	30,000	3.571429
9,400	40,000	4.255319
9,900	25	0.002525
26,900	2	0.000074

5 Exploratory Data Analysis (EDA)

To get familiar with the data, an exploratory data analysis (EDA) was performed. The first observation made in the EDA is extremely high utility prices in properties situated in certain places (e.g., Havířov and Poruba) in the Moravian-Silesian Region. Specifically, some properties have the utility payments as high as the rent price itself. In other cases, the utilities are even more expensive. It is possible that these are old buildings or buildings with a low energy class. However, this is not the case: all properties in these cities are either in buildings in good or very good condition. The energy class of buildings is in the range C–F.

In order to have a quick look at the properties listed on the portal, an average-listing profile was created for both the rent and sale categories (Table 4).

Table 4. *Average listing in Czech Republic excluding Prague and Prague itself.*

	Price (CZK/month)	Surface (m^2)	Utilities (CZK/month)	Price per m^2 (CZK)
Rent				
Czech Republic	13,800	52	4,100	265
Prague	17,100	53	3,500	322
Sale				
Czech Republic	6,300,000	63	-	100,200
Prague	8,260,000	63	-	131,200

The average rent price in the Czech Republic is 13,800 CZK. For that price it is possible to rent a 52 m^2 large flat with additional 4,100 CZK needed to be paid for the utilities. Normally it is also necessary to pay the deposit, which on average is 1.6 times the rent price.

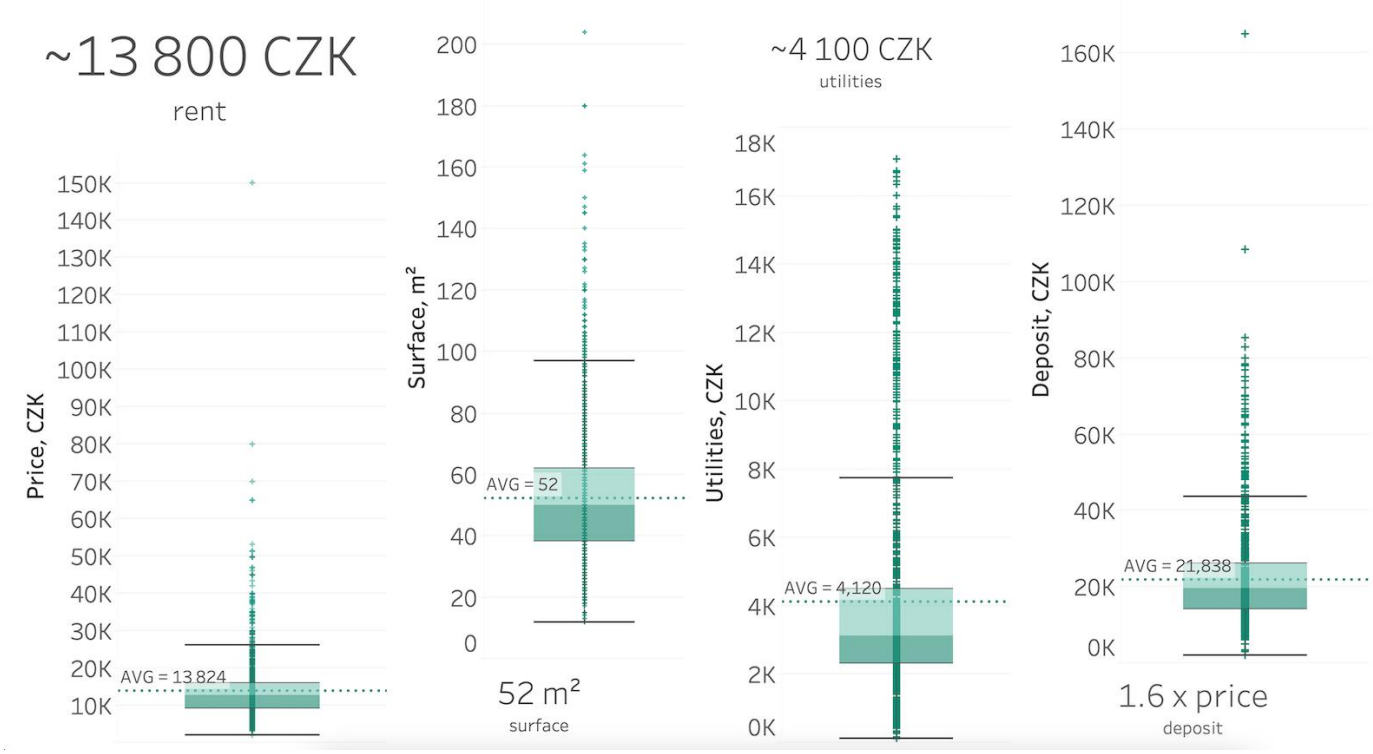

Figure 2. *Average property for rent in the Czech Republic.*

In the scope of properties for sale, the buyer needs to pay about 6,300,000 CZK for a 63 m^2 flat. The largest flat presented in the dataset had a surface area of 232 m^2, while the most expensive flat cost over 43,000,000 CZK.

Comparing properties with different layouts, another observation could be found: properties with larger surface areas cost less. For instance, in Prague properties for rent with the layout 3+kk (average surface area 78 m^2) could be rented for more than properties with the layout 3+1 (average surface area 80 m^2). The same applies to properties for sale, where properties with the layout 1+1 and 2+1 are less expensive than properties with layout 1+kk and 2+kk accordingly (Figure 4).

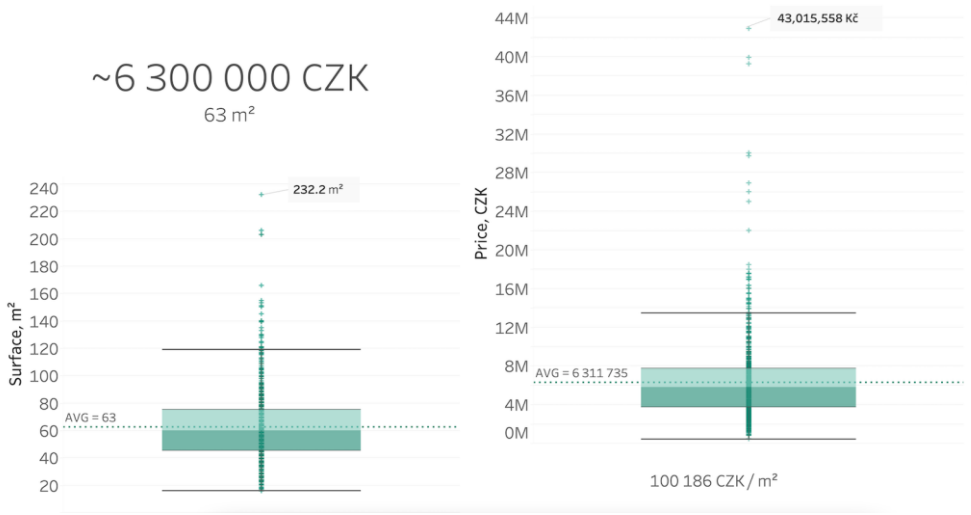

Figure 3. *Average property for sale in the Czech Republic.*

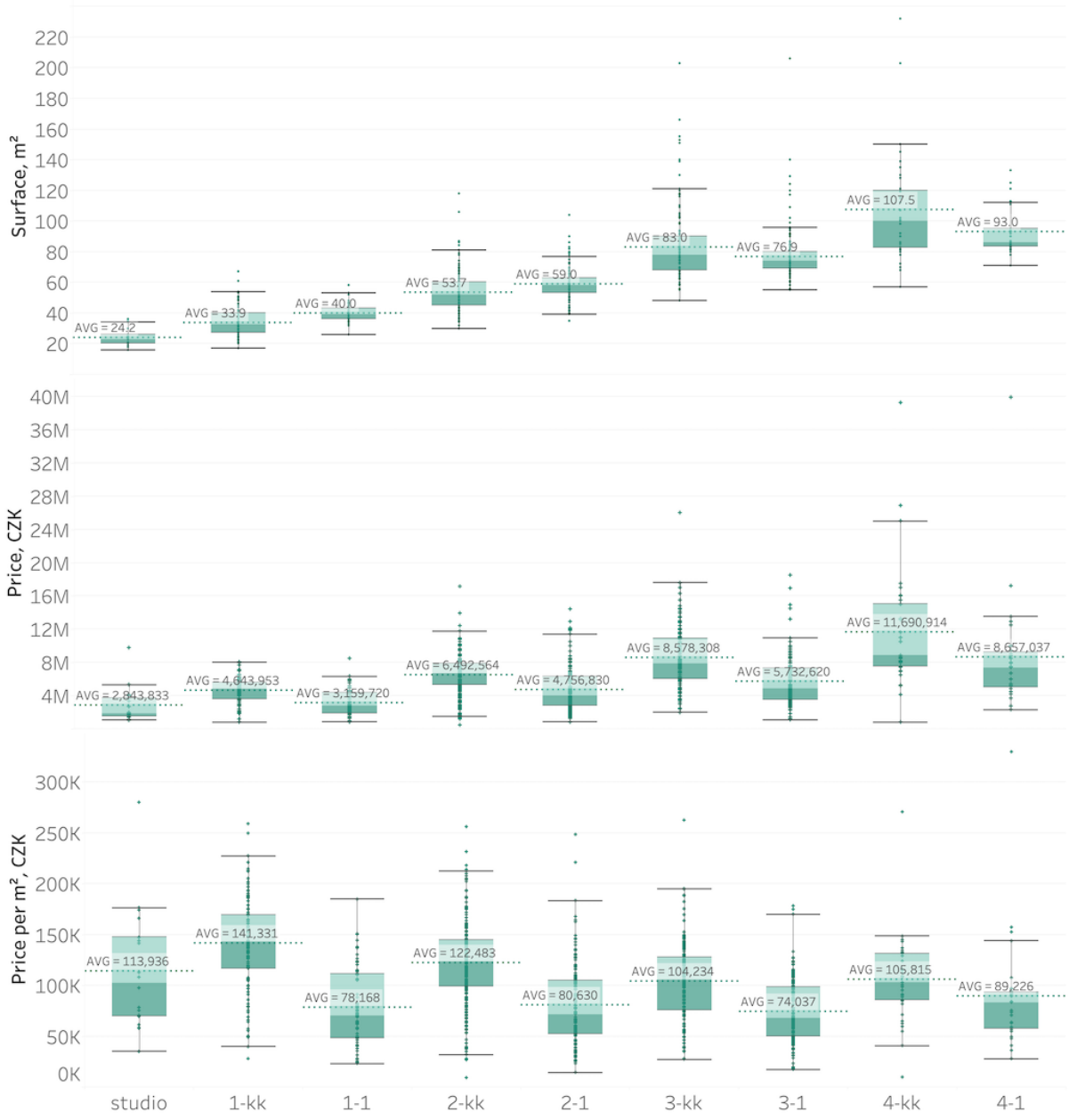

Figure 4. *Price and surface area depending on layout (properties for sale).*

The most expensive properties in the Czech Republic, as per absolute metrics (price per m^2) are the smallest ones. In the whole country, these are 1+kk properties with an average surface area of 34 m^2 and price around 143,000 CZK/m^2, while in Prague these are so called "garsoniéra" or studios, which are in fact almost the same as 1+kk, except that in 1+kk properties the kitchen corner is at least somehow separated from the rest of the room. Such a studio will cost buyers on average 156.000 CZK/m^2 and the surface area will be about 25 m^2 on average. The average difference in prices between Prague and outside Prague is in Table 5.

Table 5. *Prices in Prague in comparison with the Czech Republic.*

Parameter	Difference
Rent price	+ 24%
Utilities	- 17%
Buy price	+ 31%

As was mentioned above, properties for rent have the parameter *utilities*, which shows monthly expenses on gas, electricity, etc. Evidently, the size of such costs depends (but not exclusively) on the building's condition and energy efficiency. In the European Union, a rating scheme to summarize the energy efficiency of buildings exists. Such schemes are called energy performance certificates (EPCs)[2], specifically in the Czech Republic it is called PENB[3]. PENB can have a value from A to G, where A is the best possible value and G is the worst. It turns out from the dataset that the payments for utilities in properties in buildings with energy class F (almost the worst) are the lowest. Moreover, properties in buildings with class G have lower monthly expenses than properties in buildings with class D. This observation should probably be investigated more deeply, as it could depend for instance on the property layout. Another fact found is that 40% of the properties in the dataset are situated in buildings with the worst energy class (G).

6 Modelling

After the EDA, the deeper analysis went into place, which in this paper includes defining regression and classification tasks.

6.1 Regression

The Regression tasks aimed at predicting the price of both properties for rent and sale and payments for utilities of properties for rent. Talking about particular models, the following were selected:

- Linear regression
- Lasso regression
- Ridge regression

Before applying the model, some data preparation steps were needed. The attribute *floor* was transformed into categorical, with the addition of a new category "above_6", which combined all properties on floor 7 and above. The attributes *facilities*, *condition*, *property_type*, *building_type*, *penb*, *key_layout* and *floor* were converted to so-called "dummy" variables, which are variables that take the values 0 or 1 and indicate the

[2] See, https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/certificates-and-inspections_en

[3] See, https://aplikace.mvcr.cz/sbirka-zakonu/ViewFile.aspx?type=c&id=38880

presence or absence of something. After transforming and scaling of attributes, the dataset was split into a training and testing set at a ratio of 80 : 20.

Starting with prediction of the price of properties for rent, the first method (linear regression) was applied. The value of R^2 after 100 iterations was 75%.

The next method, which is lasso regression, requires definition of more input parameters. The cross-validation value was set at 5 and the maximum number of iterations at 10,000. The result of lasso regression with such parameters was 76%, which is slightly better than that of the linear regression. Unlike linear regression, this method also allows us to define the most important features, i.e., features on which the result depends the most. As for the analysed dataset, the following attributes were marked as important (Figure 5):

- deposit,
- surface,
- elevator,
- utilities,
- fully furnished, and
- terrace.

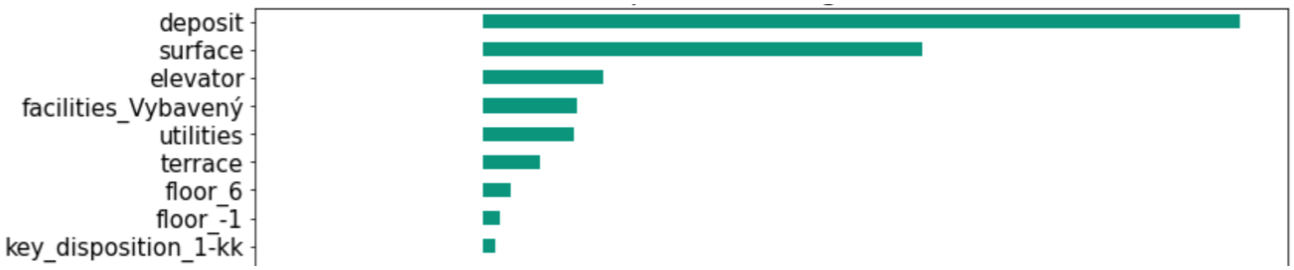

Figure 5. *Feature importance using Lasso model.*

The ridge regression did not succeed in overcoming the result of the lasso regression since it had the same value of 76%. On the other hand, additional attributes were marked as important:

- partly furnished,
- unfurnished, and
- 6th floor.

To sum up, three regression methods achieved almost the same mean result (75-76%) after 100 iterations. At the same time, these methods also allowed us to define top 5 attributes that have the greatest impact on price:

- deposit,
- surface area,
- elevator,
- utilities, and
- fully furnished.

Applying the same methods to predict the utilities gave the following results (Table 6):

Table 6. *Regression results for properties for rent.*

Method	R^2 (100 iterations)
Linear regression	0.59
Lasso regression	0.57
Ridge regression	0.57

The important attributes were price, PENB (E), PENB (D) and parking. The results of the prediction of the prices of properties for sale are shown in Table 7.

Table 7. *Regression results for properties for sale.*

Method	R^2 (100 iterations)
Linear regression	0.55
Lasso regression	0.56
Ridge regression	0.55

Here, the important attributes were surface area, elevator, PENB (A), fully furnished and garage. Evidently, the overall results of the regression tasks are not sufficient for use on other data. The only models that can be given a chance are lasso and ridge regression for predicting prices of properties for rent. Although the results are not strikingly good, this group of tasks allowed us to determine which attributes affect the price and utility payments the most.

6.2 Classification

For the next part of the modelling, the attributes PENB, building type and building condition were chosen as targets. As mentioned above, these attributes are among the missing attributes, so classification models can be potentially used to fill these missing attributes in other datasets.

To solve this group of tasks, the following methods were selected as the most usable for solving multi-class classification problems:

- support vector,
- k-nearest neighbours,
- Gaussian naïve Bayes,
- decision tree, and
- random forest.

Model training and testing was performed on the whole dataset as well as on datasets filtered by property type. For the comparison of the models, the accuracy score metric was chosen, which measures the proportion of correct predictions in a classification model. It is calculated by dividing the number of correct predictions by the total number of predictions.

First, the mentioned methods were applied for classifying the PENB class of building. However, the results obtained were not as good as desired, with the most successful method, the random forest classifier, only reaching an accuracy rate of 0.54 while predicting the PENB class in properties for sale (Table 8).

The prediction of building type (Table 9) could be regarded as more successful since the random forest classifier resulted in an accuracy rate above 0.8. This high accuracy rate suggests that the model is able to correctly identify the building type in most cases.

Table 8. *Accuracy of predicting PENB class.*

Model	Accuracy score, properties for rent	Accuracy score, properties for sale	Accuracy score, all properties
SVC	0.42	0.49	0.38
KNN	0.35	0.35	0.34
GNB	0.29	0.14	0.14
Decision tree	0.37	0.43	0.39
Random forest	0.5	0.54	0.53

Table 9. *Accuracy of predicting building type.*

Model	Accuracy score, properties for rent	Accuracy score, properties for sale	Accuracy score, all properties
SVC	0.67	0.58	0.64
KNN	0.7	0.53	0.64
GNB	0.27	0.21	0.21
Decision tree	0.67	0.64	0.67
Random forest	0.81	0.82	0.8

The results obtained from the last target attribute, which was the building condition (Table 10), were not as good as the results for the previous attribute, as the best model (again the random forest classifier) reached an accuracy rate of 0.64. This means that the given model was only able to correctly predict the building condition in 64% of the cases.

Table 10. *Accuracy of predicting building condition.*

Model	Accuracy score, properties for rent	Accuracy score, properties for sale	Accuracy score, all properties
SVC	0.57	0.5	0.54
KNN	0.48	0.44	0.43
GNB	0.52	0.26	0.38
Decision tree	0.51	0.57	0.52
Random forest	0.64	0.61	0.63

6.3 Clustering

The previous methods were part of so-called predictive analysis, and now we are moving to descriptive modelling, where clustering was done. For deeper analysis, the clustering was made based on different dimensions. K-means and k-modes algorithms were used in this part of the modelling. In addition to that, the elbow method was used to determine the optimal number of clusters.

First, k-means was applied to the latitude and longitude of each property. The elbow method defined the optimal number of clusters as 3. In that case, the silhouette score was 0.772. The resulting data divided into clusters were then displayed on a map.

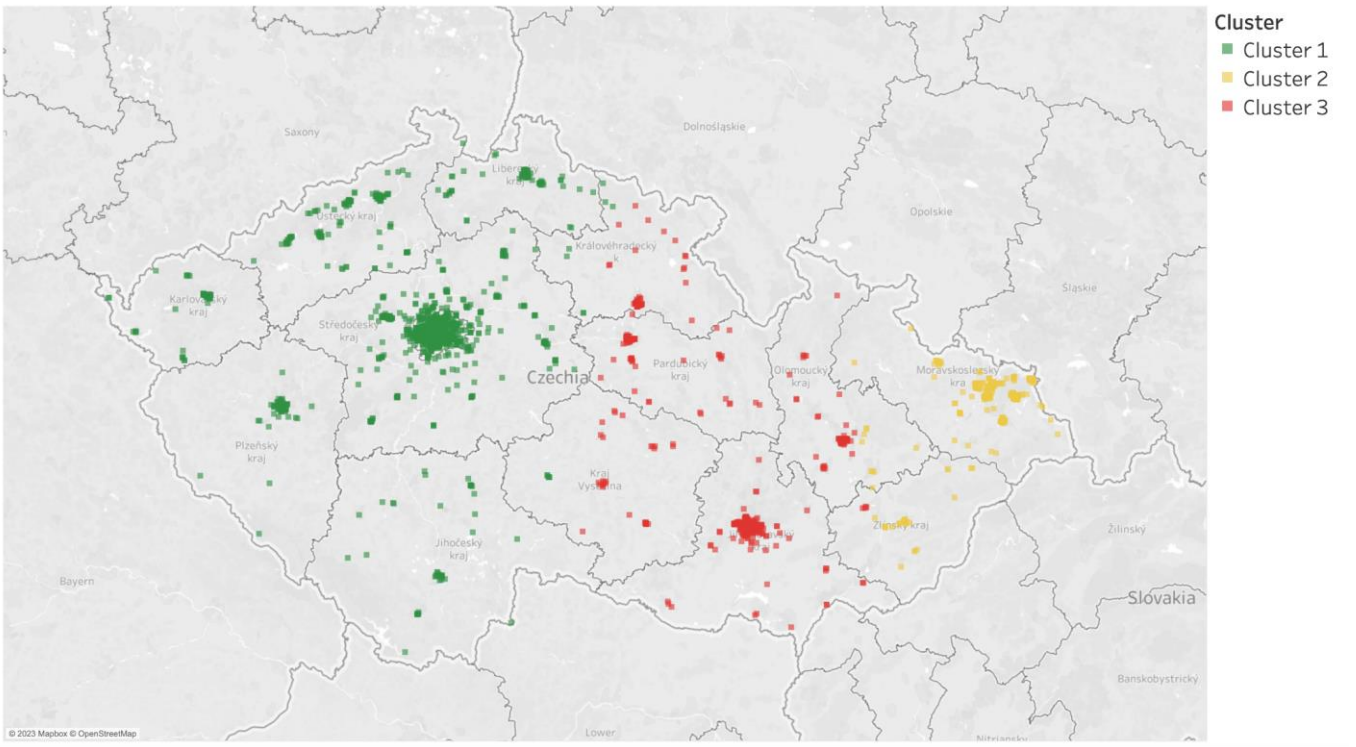

Figure 6. *Location clusters.*

The above map shows that the first cluster contains primarily properties for sale from Karlovy Vary, Ústí and Labem, Plzeň, Central Bohemian, Liberec and South Bohemian Regions. At the same time, this cluster is the biggest one. The second cluster is represented by two regions – Moravian-Silesian and Zlín. The third cluster then consists of Hradec Králové, Pardubice, South Moravian, Olomouc and Vysočina Regions. By adding more context (such as price) to the clustered data, the following boxplot could be created (Figure 7).

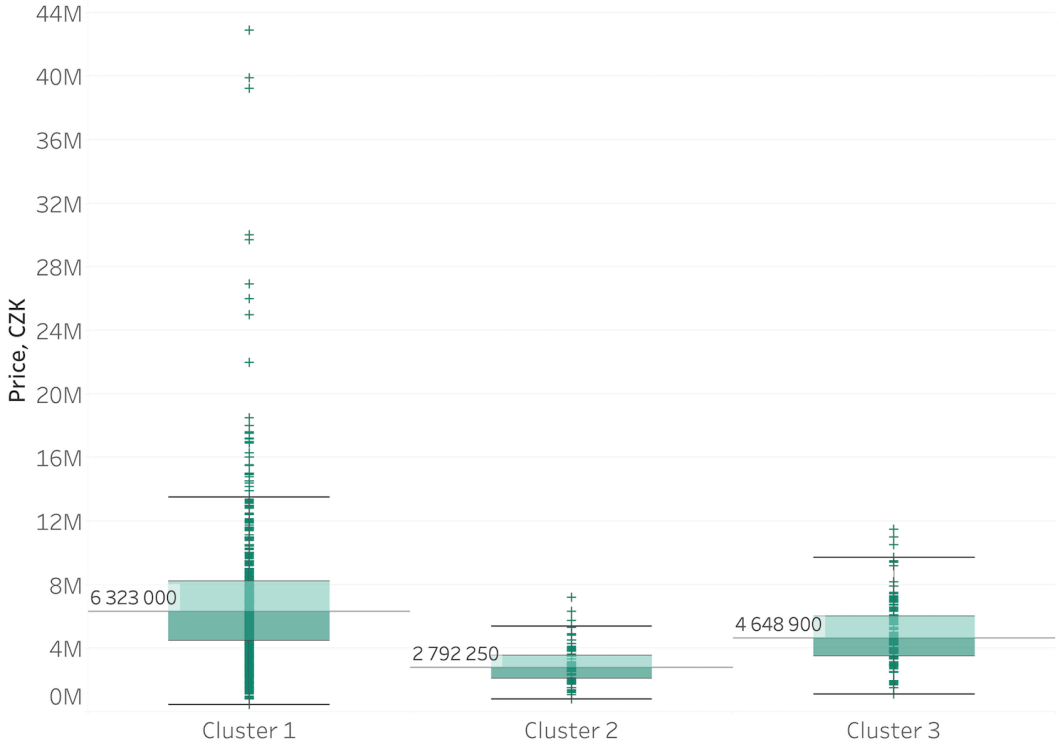

Figure 7. *Prices of properties for sale across clusters.*

Evidently, the most expensive properties are located in the first cluster, of which Prague is a part. The median property price in that cluster is CZK 6,300,000. The second most expensive cluster is Cluster 3 one with a median price of CZK 4,650,000 and the least expensive cluster has a median price of CZK 2,800,000.

The situation with properties for rent is essentially the same (Figure 8). There is a small discovery though. The difference between the prices of properties for rent in Clusters 1 and 3 is 36%, while for properties for sale the difference between the same clusters is 16%. In other words, renting a flat in the regions in Cluster 2 costs a little less than in Cluster 1. However, buying flats there is significantly cheaper.

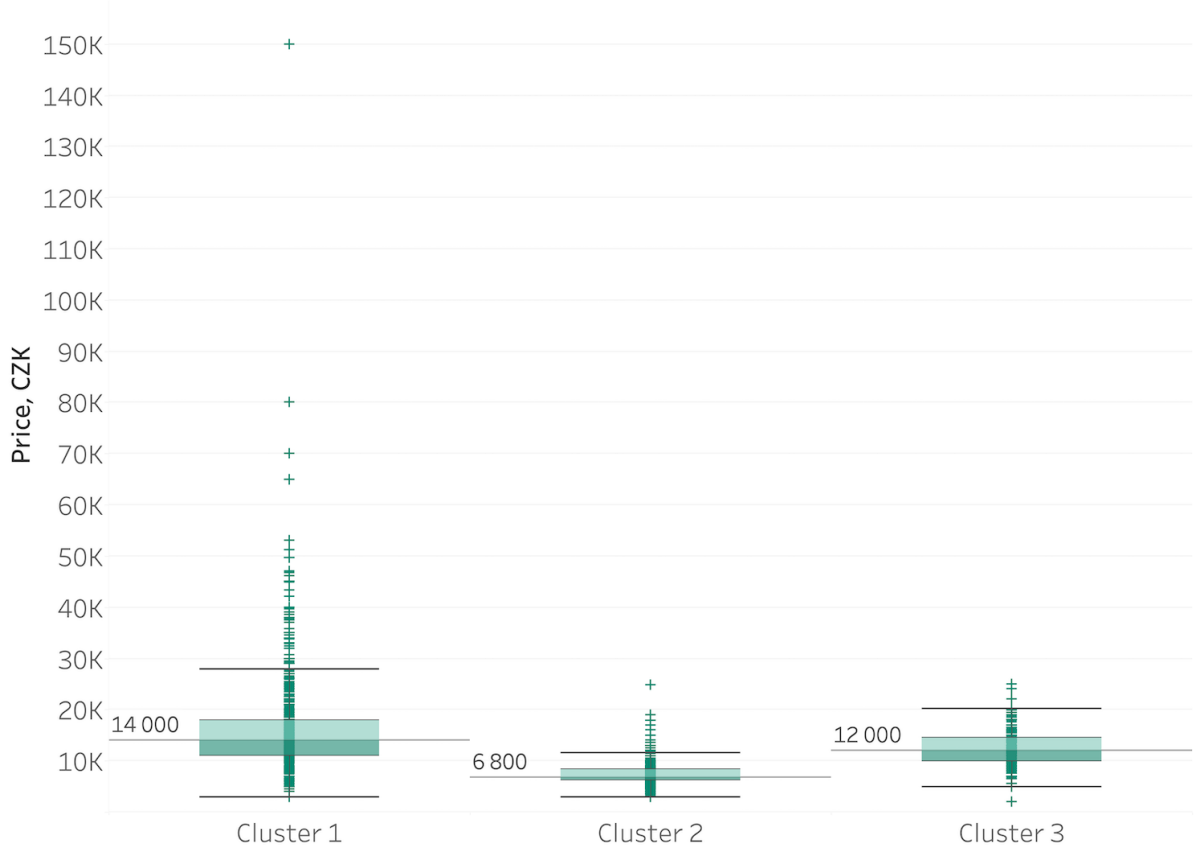

Figure 8. *Prices of properties for rent across the.*

Further cluster analysis was performed only on properties for sale. This time, the price and property size were chosen as the target attributes. Again, using the elbow method, the optimal number of clusters was determined to be two. The silhouette score of the model was 0.667. Table 11 shows an overview of the clusters including the median values of the selected attributes.

Table 11. *Clusters by price and size (properties for sale).*

Cluster	Number of records	Price, CZK	Size, m²
Cluster 1	818	5,299,450	57
Cluster 2	113	12,420,000	91

The model essentially divided flats into two quite logical types:

- larger and more expensive, and
- smaller and cheaper.

Figure 9 shows the location of properties in Cluster 2.

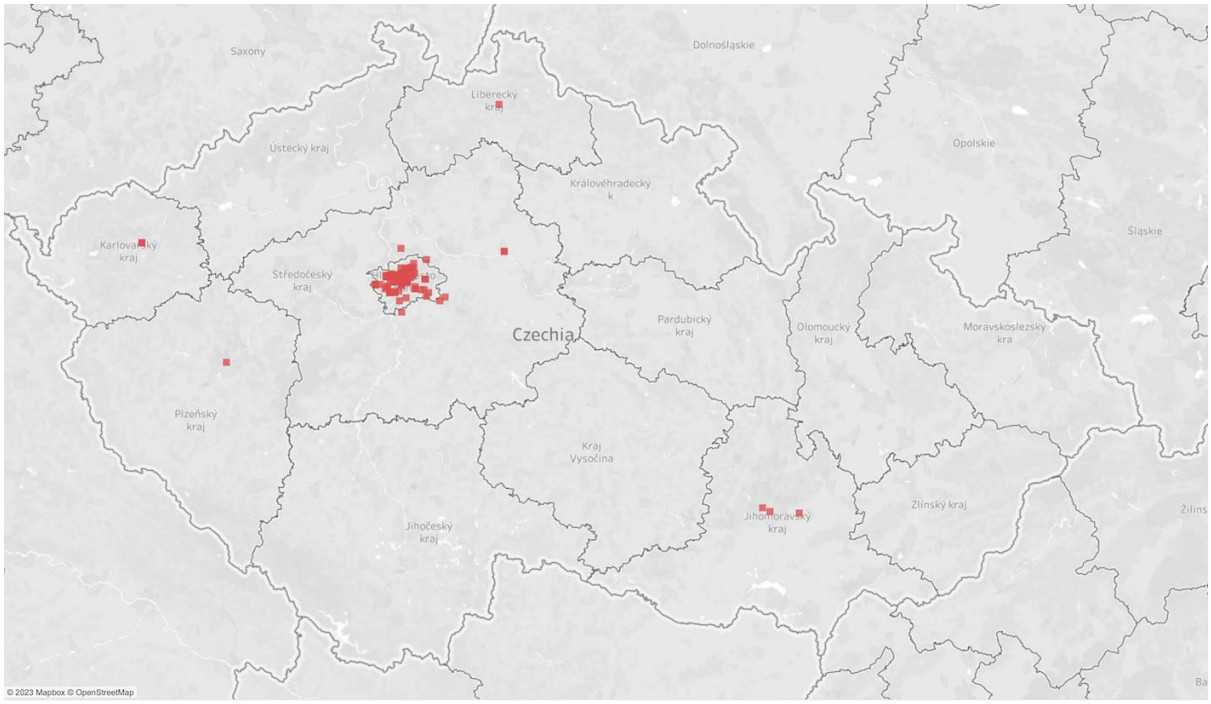

Figure 9. *Expensive properties for sale.*

Most of the expensive properties were concentrated in Prague. It is clearly visible on the scatter plot (Figure 10) that a straight line can be drawn at the level of CZK 10,000,000, which would be the border between the created clusters.

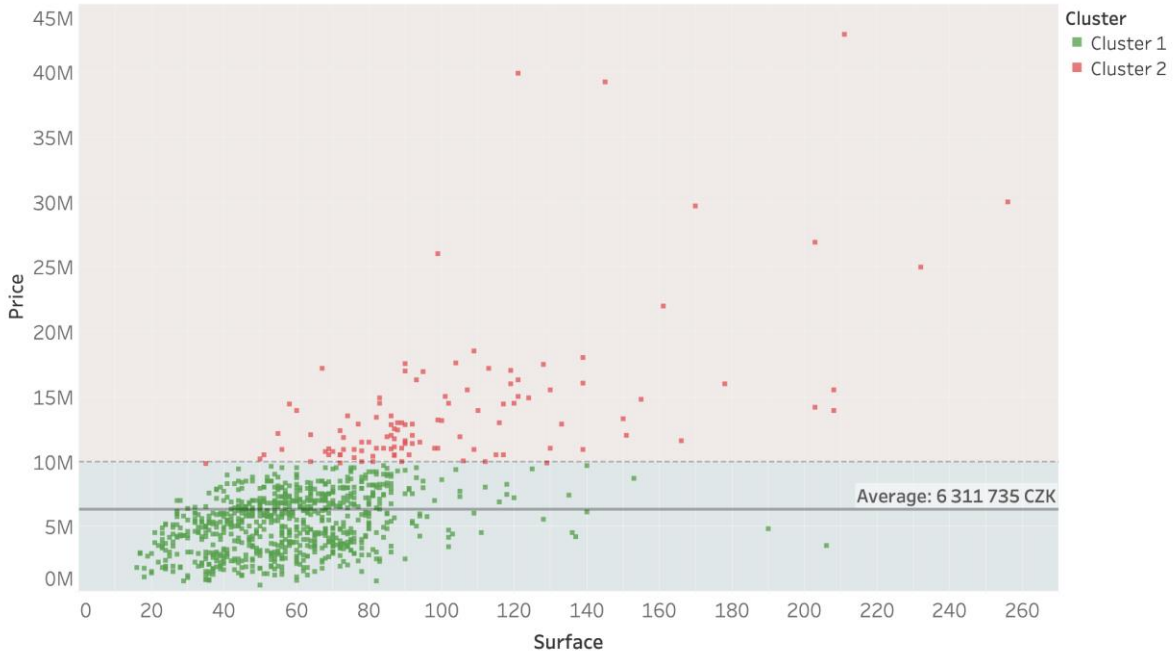

Figure 10. *Price border between clusters.*

For clustering properties for rent, four attributes were selected, namely price, surface area, utilities and deposit. Because of that, PCA (principal component analysis) was needed to adjust the values to ensure that they are on the same scale. The (Table 12) model resulted in three clusters, with a silhouette score of 0.564.

Table 12. *Summary of clusters.*

Cluster	Number of properties	Price (CZK)	Surface (m²)	Utilities (CZK/month)	Deposit (CZK)
Cluster 1	279	24,000	83	5,000	40,000
Cluster 2	1694	12,500	47	3,000	19,000
Cluster 3	247	6,550	51	10,917	13,100

Cluster 1 contains the most expensive and largest properties with a high deposit value. The largest Cluster 2 has the smallest surface area and payments for utilities among all the clusters. The last cluster was interesting because it concentrated the properties with the lowest prices, but at the same time with the highest utility payments. This cluster can be characterized as properties where the utilities are more expensive than the price of the flat. Showing the values from the last cluster on the map (Figure 11) immediately catches the eye in that most of the properties in this cluster are located in the Moravian-Silesian Region. The same observation about utilities being more expensive than the rental price was found during the exploratory analysis.

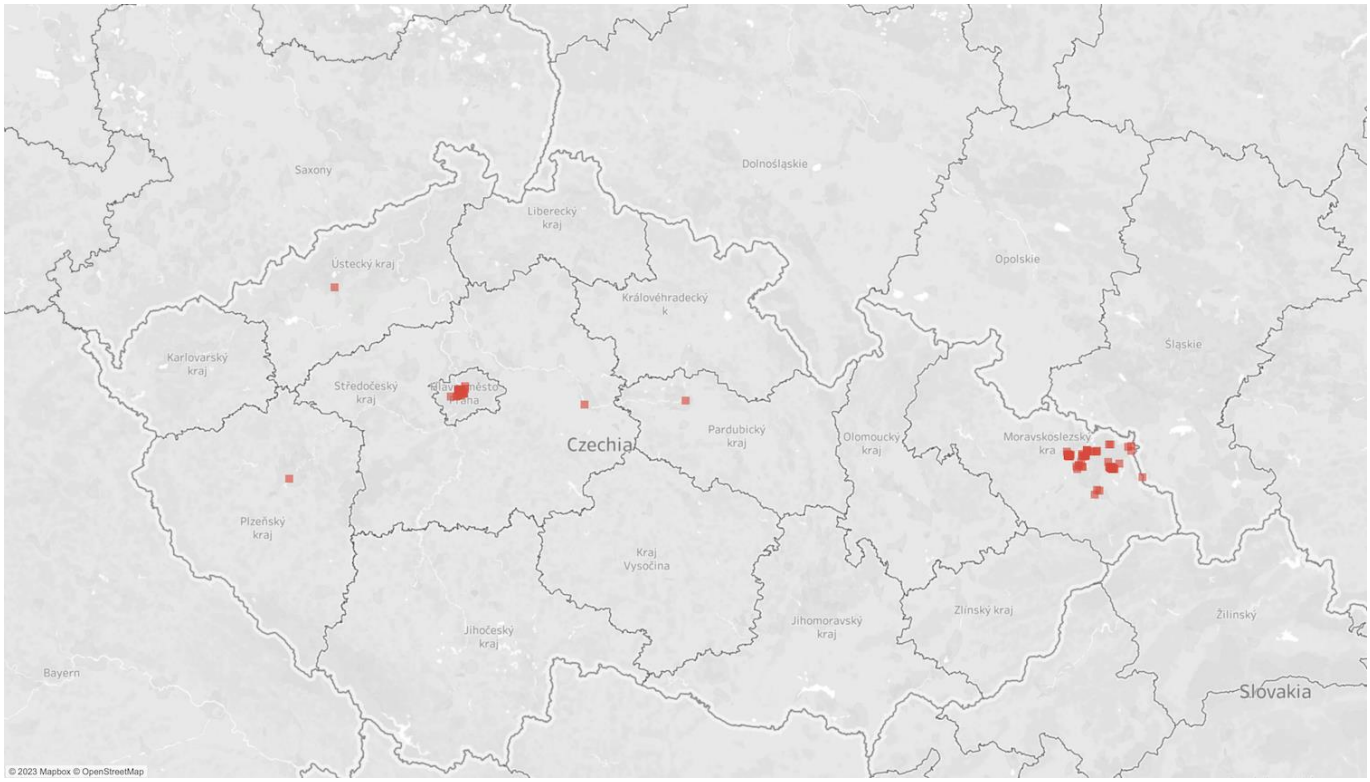

Figure 11. *Extremely high utility payments in Moravian-Silesian Region.*

The last cluster analysis differs from the previous one by using categorical attributes instead of numerical. Thus, the k-means algorithm was not suitable and another algorithm – k-modes – was used. The number of iterations was set at 5. Before creating the model, selected numerical attributes (price, surface area, fees and deposit) were discretized as follows:

- *price*: bins of size 5000, starting from 0 up to 50,000 + bin with price > 50,000
- *surface*: bins of size 10, starting from 0 up to 110 + bin with surface > 110
- *utilities*: bins of size 1000, starting from 0 up to 10,000 + bin with utilities > 10,000
- *deposit*: bins of size 5000, starting from 0 up to 40,000 + bin with deposit > 40,000

The result is shown in Table 13.

Table 13. *Summary of clusters.*

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Number of records	1085	361	246	238	290
Layout	2–kk	1–kk	2–kk	2–1	2–1
Furnished	Partly	Fully	Partly	Partly	None
Floor	2	4	1	3	1
Balcony	No	No	Yes	Yes	No
Terrace	No	No	No	No	No
Cellar	No	Yes	No	Yes	Yes
Loggia	No	No	No	No	No
Parking	No	No	Yes	No	Yes
Elevator	No	Yes	Yes	Yes	No
Garage	No	No	No	No	No
Building condition	Very good	Very good	New construction	Very good	Good
Property type	Private	Private	Private	Private	Private
Building type	Brick	Panel	Brick	Brick	Brick
PENB	G	G	B	G	E
Price	10K–15K	5K–10K	15K–20K	15K–20K	5K–10K
Surface	40–50	30–40	50–60	60–70	50–60
Utilities	2K–3K	1K–2K	3K–4K	3K–4K	>10K
Deposit	15K–20K	10K–15K	20K–25K	25K–30K	10K–15K

6.4 Anomaly detection

The last descriptive task was anomaly detection. The isolation forest algorithm was used for the task. The attributes chosen were price, surface, utilities and all attributes about the surroundings. Optimal parameters for the model were selected using GridSearchCV (Table 14).

Table 14. *Model parameters.*

Parameter	Value
Number of estimators	50
Maximum number of samples	10
Contamination	Auto
Maximum number of features	5

The result of the model from the previous step was 760 rows. There are properties that are located close (< 100 m) to some places, or, on the contrary, far from them. In addition, as a result, there are large flats (120 m^2) at a low price (CZK 16,000).

By changing the "contamination" parameter (proportion of outliers in the dataset) to 0.005, i.e., so that the resulting outliers represent 0.5% of the dataset, and increasing the "max_features" parameter by 5, the result will only contain 11 properties, which are shown below (Figure 12).

	price	surface	utilities	public_transport_stop	post_office	store	bank	restaurant	pharmacy	school	kindergarten	sport_field	playground	ano
175	9000	60	2500.000000	5000	2200	7200	8400	6900	8400	4400	7900	7300	4700	
184	12000	86	4000.000000	1488	1500	2700	9500	2900	1700	4000	2800	10200	1900	
341	6000	30	2800.000000	653	2300	2300	8900	3200	3300	3500	3500	3900	2400	
443	5500	20	9166.666667	578	1900	1900	6600	1102	2000	1700	2600	3200	3100	
968	18000	75	3500.000000	1325	1600	1322	6000	1600	4700	4600	2000	5000	1283	
1288	12000	42	2000.000000	1161	4800	2600	5800	8100	6000	5700	6600	6200	7000	
1356	7000	33	3500.000000	4400	3700	4500	10500	4800	5300	4800	8900	4700	4200	
1406	21000	79	3500.000000	90	3200	3100	9100	3100	6200	3000	3100	3700	1800	
1527	8000	30	3000.000000	3700	3900	7700	7700	7100	7400	8000	8600	3200	228	
1610	11000	80	4000.000000	2200	3500	710	3600	2800	4300	2500	2600	2900	4000	
2063	9500	36	2500.000000	874	1308	36100	6200	37200	6200	5400	5000	6400	1009	

Figure 12. *Example of records from clusters.*

7 Discussion

The discussion can be divided into two parts: discussion of use of individual data mining methods and discussion of the obtained results.

7.1 Discussion of data mining methods

From the perspective of use of individual data mining method, we use many different data mining methods. In the regression tasks we used linear, lasso and ridge regression. The results of the methods were very similar: the R^2 error differs by 0.02 (3.51%, Table 6, Properties for rent) and by 0.01 (1.82%, Table 7, Properties for sale). These small differences are consistent with relevant studies, for example Thevaraja et al. (2019) indicates the residual sum of squares deviation between the best and worse methods as 2.14%.

In the classification tasks, the best result in terms of accuracy was achieved by the decision forest method. Generally, ensemble methods (to which random forest belongs) yield better results than individual methods, which is confirmed by many relevant studies (e.g., Chaurasia et al., 2022; Dželihodžić & Đonko, 2016).

7.2 Discussion of obtained results

From the perspective of the obtained results, if we compare available statistics (Deloitte, 2022b; Realitymix.cz, 2022) with the results of our analysis, we reached very similar results. According to (Realitymix.cz, 2022), the average price per square metre in Prague in Q1/2022 was 118,500 CZK and according to Deloitte (2022b) the average price of an older flat in the Czech Republic exceeded 90,000 CZK per square metre. From the results of our analysis, the average price per square metre in Prague was 131,200 CZK. This difference is due to the difference between the offer price and the realized property sale price. This difference depends on the total time that the advert is available before the deal goes through, and ranges from around 4% to 15% (in some cases, however, the realized price may be higher than the original asking price) (Šitera, 2020).

In the exploratory analysis, we found that there are two cities in the Moravian-Silesian Region (Havířov and Poruba) where payments for utilities are particularly high. It could be assumed that these would be very old buildings with poor insulation. However, this is not the case: all the properties in these cities are in buildings in either good or very good condition. The energy class of the buildings is in the range C–F. We have not been able to clarify the reason for this anomaly: it would require consultation with a domain expert from the real estate market.

In addition, we also found that the building energy class (PENB) is not related to the price of utilities, i.e., buildings in energy class F or G (the worst) have lower or the same utility payments as buildings in classes A–C. In addition to the above points of interest, we also found that 40% of the properties in the dataset are located in buildings with the worst energy class (G).

7.3 Limitations

A limitation of the work is that it focuses only on direct property sales without the participation of real estate agencies – the portal bezrealitky.cz offers a direct connection between the seller (landlord) and the buyer (tenant). Therefore, it cannot be said that it maps the entire real estate sales market. On the other hand, the information on the prices of flats should be more accurate, as it does not include the real estate agents' sales brokerage fee. Another limitation of the work is the fact that the data were obtained on a single day, so the factor of change over time is not taken into account. However, we expect to re-run the analysis in 2023 to eliminate this shortcoming.

7.4 Future work

The data for the analysis were scraped on 22 March 2022. Since then, the real estate market has been fundamentally influenced by two events: a sharp increase in mortgage prices and energy prices. According to the Czech Banking Association, the volume of mortgages granted in October 2022 decreased by 80% year-on-year (Česká Bankovní Asociace, 2022). The overall change in the property market was also reflected in an increased supply of flats and a significant growth in rental interest. As of today, (16 November 2022), the portal bezrealitky.cz offers 2630 advertisements for flats for rent and 3243 advertisements for flats for sales. While the number of rental ads remains about the same, the number of ads for sale has increased by 262%. Obviously, it would be interesting to perform the analysis again and identify changes and trends in this new market situation.

8 Conclusion

The paper aimed to obtain interesting findings using different data mining methods on real estate data. With the help of web scraping, data on more than 3500 properties were acquired, but after data processing, 3131 of them remained, of which 2220 were properties for rent and 931 were properties for sale. Various data mining methods, such as classification, regression, clustering and anomaly detection, were applied on this dataset.

The lasso regression for predicting the rent price was the most successful model among others with an R^2 value of 0.76. The model can be useful for estimating property prices based on property parameters, such as the number of rooms, surface area, floor, utilities, location, etc. The random forest model a with success rate of 0.8 allows us to predict the property's PENB class, which can also be useful when dealing with missing values in datasets.

Besides predictive analytics, descriptive data mining methods were applied, which also led to some insights regarding the Czech real estate market. Some of the findings are consistent with well-known facts (domain knowledge) about the real estate market, such as the fact that the most expensive prices are in Prague and other large cities, or that the highest prices per square metre are for relatively small flats. Other findings are relatively interesting, especially the very high level of utility payments in the Moravian-Silesian Region.

Overall, data mining has the potential to affect the real estate industry by providing insights and enabling data-driven decision making. By applying data mining techniques, it is possible to discover patterns and relationships that can help predict property values and better understand market trends. The use of data mining in the real estate domain is about to become really important as more data become available.

Additional Information and Declarations

Conflict of Interests: The authors declare no conflict of interest.

Author Contributions: I.T.: Conceptualization, Methodology, Software, Data curation, Writing – Original draft preparation, Visualization, Investigation. D.C.: Supervision, Validation, Resources, Writing – Reviewing and Editing.

Data Availability: The data that support the findings of this study are available from the corresponding author.

References

Chaurasia, V., Pandey, M., & Pal, S. (2022). Chronic kidney disease: a prediction and comparison of ensemble and basic classifiers performance. *Human-Intelligent Systems Integration*, 4(1–2), 1–10. https://doi.org/10.1007/s42454-022-00040-y

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. https://doi.org/10.1007/bf00994018

Cover, T. M., & Hart, P. D. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27. https://doi.org/10.1109/tit.1967.1053964

Česká Bankovní Asociace. (2022). ČBA Hypomonitor říjen 2022: Úroková sazba mírně vzrostla. https://cbaonline.cz/cba-hypomonitor-rijen-2022

ČTK - České Noviny. (2022) Průměrná cena staršího bytu v Česku přesáhla 90.000 korun za metr čtvereční. https://www.ceskenoviny.cz/zpravy/prumerna-cena-starsiho-bytu-v-cesku-presahla-90-000-korun-za-metr-ctverecni/2203149

Delloite. (2022a). Deloitte Real Index: How do real prices of flats in the Czech Republic develop? https://www2.deloitte.com/cz/en/pages/real-estate/articles/cze-real-index.html

Deloitte. (2022b). Deloitte Real Index Q1 2022, Skutečné ceny prodaných bytů v ČR. https://www2.deloitte.com/content/dam/Deloitte/cz/Documents/real-estate/Real-index-1Q-2022-CZ.pdf

Dželihodžić, A., & Đonko, D. (2016). Comparison of ensemble classification techniques and single classifiers performance for customer credit assessment. *Modeling of Artificial Intelligence*, 11(3), 140–150. https://doi.org/10.13187/mai.2016.11.140

Eurostat. (2022) Rents up by 17%, house prices by 45% since 2010. https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20220708-1

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). *Advances in knowledge discovery and data mining*. AAAI press MIT press.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE. https://doi.org/10.1109/ICDAR.1995.598994

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55. https://doi.org/10.2307/1267351

Hossin, M., & Sulaiman, M. (2015). A review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2), 1–11. https://doi.org/10.5121/ijdkp.2015.5201

Hromada, E. (2015). Mapping of Real Estate Prices Using Data Mining Techniques. *Procedia Engineering*, 123, 233–240. https://doi.org/10.1016/j.proeng.2015.10.083

Hromada, E. (2018). Analysis of relationship between market value of property and its distance from center of capital. In *17th International Scientific Conference Engineering for Rural Development* (pp. 646–651). Engineering for Rural Development. https://doi.org/10.22616/erdev2018.17.n305

Hromada, E. (2021). Development of the Real Estate Market in the Czech Republic in Connection with the Covid-19 Pandemic. In *Proceedings of the 15th Economics & Finance Conference* (pp. 169–176). IISES. https://doi.org/10.20472/EFC.2021.015.014

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, 2(3), 283–304. https://doi.org/10.1023/a:1009769707641

Liu, F., Ting, K. M., & Zhou, Z. (2008). Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 413-422). IEEE. https://doi.org/10.1109/icdm.2008.17

Louati, A., Lahyani, R., Aldaej, A., Aldumaykhi, A., & Otai, S. (2022). Price forecasting for real estate using machine learning: A case study on Riyadh city. *Concurrency and Computation: Practice and Experience*, 34(6). https://doi.org/10.1002/cpe.6748

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297). University of California Press.

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 33(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Oliveira, T. C., De Medeiros, L., & Detzel, D. H. M. (2021). Applying data mining algorithms to real estate appraisals: a comparative study. *International Journal of Housing Markets and Analysis*, 14(5), 969–986. https://doi.org/10.1108/ijhma-07-2020-0080

Plevris, V., Solorzano, G., Bakas, N., & Ben Seghier, M. (2022). Investigation of performance metrics in regression analysis and machine learning-based prediction models. In *8th European Congress on Computational Methods in Applied Sciences and Engineering*. Scipedia. https://doi.org/10.23967/eccomas.2022.155

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106. https://doi.org/10.1007/bf00116251

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Realitymix.cz (2022). Průměrná cena za 1 m² bytu. https://realitymix.cz/statistika-nemovitosti/

Sawant, R., Jangid, Y., Tiwari, T., Jain, S., & Gupta, A. (2018). Comprehensive Analysis of Housing Price Prediction in Pune Using Multi-Featured Random Forest Approach. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA),* (pp. 1–5). IEEE. https://doi.org/10.1109/ICCUBEA.2018.8697402

Šitera, R. (2020). Nabídkové vs. Realizované ceny – jaký je skutečný rozdíl. *Valuo*. https://www.valuo.cz/blog/nabidkove-vs-realizovane-ceny-jaky-je-skutecny-rozdil/

Thevaraja, M., Rahman, A., & Gabirial, M. (2019). Recent developments in data science: Comparing linear, ridge and lasso regressions techniques using wine data. In *International Conference on Digital Image and Signal Processing 2019: DISP 2019* (pp. 1-6). University of Oxford.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1), 267–288.

Tsakunov, I. (2022). *Využití data miningu pro analýzu českého realitního trhu*. Prague University of Economics and Business.

Verma, A., Nagar, C., Singhi, N., Dongariya, N., & Sethi, N. (2022). Predicting House Price in India Using Linear Regression Machine Learning Algorithms. In *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM),* (pp. 917–924). IEEE. https://doi.org/10.1109/ICIEM54221.2022.9853185

Editorial record: The article has been peer-reviewed. First submission received on 23 November 2022. Revisions received on 15 January 2023 and 23 March 2023. Accepted for publication on 19 April 2023. The editor in charge of coordinating the peer-review of this manuscript and approving it for publication was Stanislava Mildeova.

Acta Informatica Pragensia is published by Prague University of Economics and Business, Czech Republic.

ISSN: 1805-4951