The University of Texas at Dallas

BA with R

Final Project

Presented By

Bhupesh Kumar Srivastava

Predicting Housing Value in a Volatile Economy

Table of Contents:

1. Executive Summary

Buying a house is not just buying a property; it is like fulfilling a dream. According to a recent analysis by Zillow, the real state stock in the US is worth 43.4 trillion dollars. However, buying a house is not easy because of the high costs and the risks involved. The average house price has been rising rapidly nationwide. Home prices in the United States have increased by 16.2% from 2020 to 2021.

The variations and uncertainties in the real estate markets, makes estimating housing prices a daunting task. To develop a predictive model, we use a dataset from Kaggle.com. We utilize intrinsic house characteristics and macroeconomic indexes to predict the price of a house in a four-year window. As the methods we have developed are dataset agnostic, with minimal modification they could be used in other contexts.

Our final prediction model shows a 38 percent increase over the baseline OLS model. This improvement can benefit the users substantially, enabling them to make better decisions while investing in the real estate market.

We also use econometric tools to analyze the characteristics of a property that affect the price and provide stakeholders with general insight regarding the average behavior of the real estate market.

However as this is a preliminary work, we allude to the need for further investigation to improve the predictive and explanatory models.

2. Introduction:

In this project, we analyze and predict housing value in a volatile market over a four-year window and our aim is to predict the price of the houses based on the characteristics of houses and the Russian economy indexes. The dataset is from kaggle.com, including the characteristics of sold houses and the macroeconomics indexes. We took a total of 16 variables, 13 from house characteristics and 3 from the macroeconomy dataset to predict the housing prices.

The idea for choosing this project was to help people or businesses who want to buy or sell houses to make a more enlightened estimation. Buying a house involves one of the biggest expenses of a person's lifetime, and statistical models can reduce the uncertainties of this transaction.

The outcome of this project will be a statistical model which will predict housing prices using the provided variables. As our predictive models, we have used Linear Regression, Lasso, Elastic Net, Ridge, and XGBoost to analyze the relation between prices and other variables. We found that XGBoost has the lowest error rate and outperforms all the other models we tested by nearly 38 percent.

3. Data Description and Understanding

The dataset is from Sberbank, it was released as part of a Kaggle competition with the aim of predicting housing valuations in the volatile Russian real estate market. The main dataset has 30471 observations and 292 variables and there is a complementary dataset for Russia's macro economy status.
We limit the number of variables as this project is a demonstration and the resources (time/computation) are limited for the intended analysis.
We have selected 12 variables from house intrinsic characteristics, 2 macro-economic variables, one time variable and the target variable of property price. The following are the variables we have selected out of the 292 housing characteristics variables and 100 microeconomy variables.

Variable Name	Type	Description
Timestamp	String	Date of sale
full_sq	Numerical	Total square area including non-residential areas
life_sq	Numerical	Total square area excluding non-residential areas
floor	Numerical	Floor of the property
max_floor	Numerical	Max floor of the building
build_year	Numerical	Year the property built
num_room	Numerical	Number of residential rooms in the property
kitchen_sq	Numerical	Kitchen square footage
state	Numerical	State of property
material	Numerical	Type of material used in walls
product_type	Numerical	Either owner-occupied or investor property
full_all	Numerical	Subarea Population
usdrub	Numerical	Exchange rate of ruble to US dollar. Macro Data
unemployment	Numerical	Unemployment rate of the country. Macro Data
price_doc	Numerical	Price of property

We decided to use all intrinsic variables for the properties in the data set since they are particularly important to the analysis. We excluded variables related to the neighborhood due to limited resources. We decided to exclude most variables that detail the macro-economic situation of the area because they are highly correlated, and a lot of information could be derived from just a handful of them. We chose unemployment rate and the currency exchange rate.

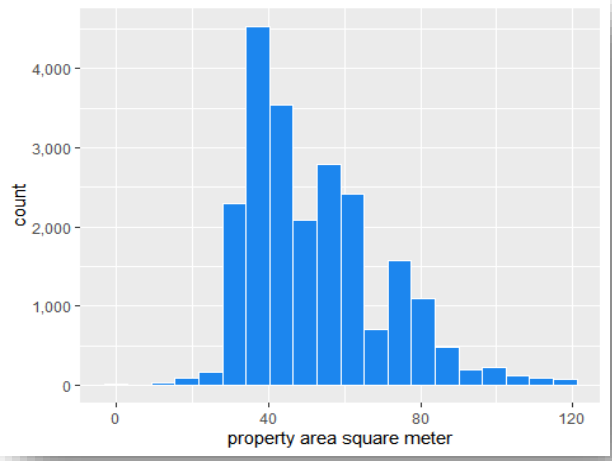
4. Data Processing

To better understand the data structure, we have incorporated the data processing with our EDA, and we look for patterns and anomalies as we inspect the data. Since our final predictive model is a tree regression, we do not need to normalise the data for that model and we skip this part.
We also split data into two parts and will use 75% of it for training and 25% for testing our models.

5. Data Visualization

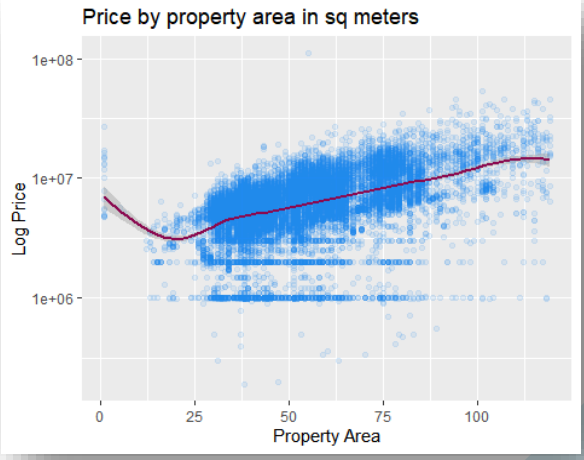
To focus on the main distributions of data, some outliers might have been removed from the graphs and they are not shown separately. As we move forward through data, cleaning might take place as needed.

Property Area



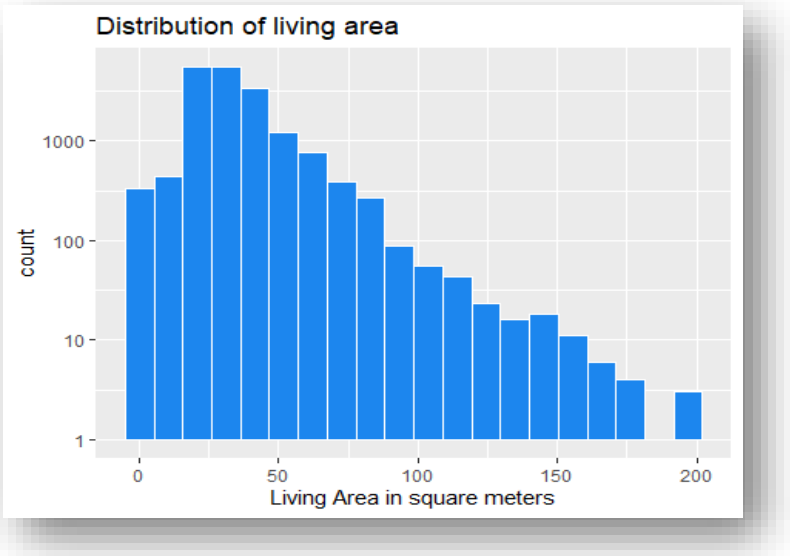
Property area is the total area in square meters, including loggias, balconies, and other non-residential areas. The histogram is similar to a Poisson distribution.

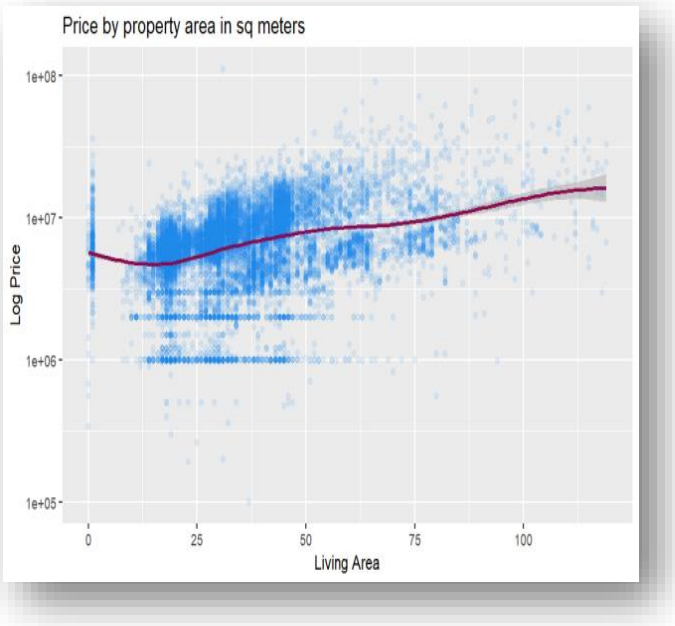
Here we have the relationship between property area and price. As one would expect, there is a general positive relation between the variables. To comment on the lower levels of data we need further information as it could be an error in the data input or something else.



Living area

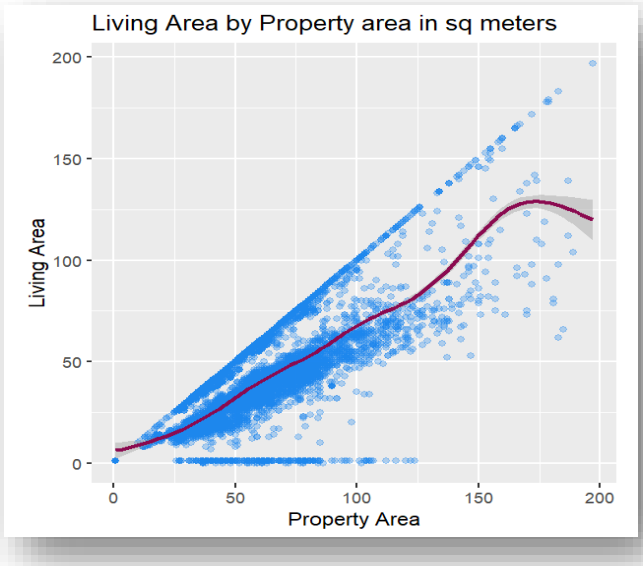
We removed values in which the property area is smaller than living area, as we are assuming the property value cannot be smaller than the living area. Now we look at the distribution of the living area.





The adjacent graph is a scatter plot of the price by living area and log price. There is a general positive correlation between the two values.

In the data set, we see the interesting pattern of round prices which is happening either at the time of the transaction or when the data is added to the dataset. The lines within the graph on the y axis show many rounded prices for different living area sizes.

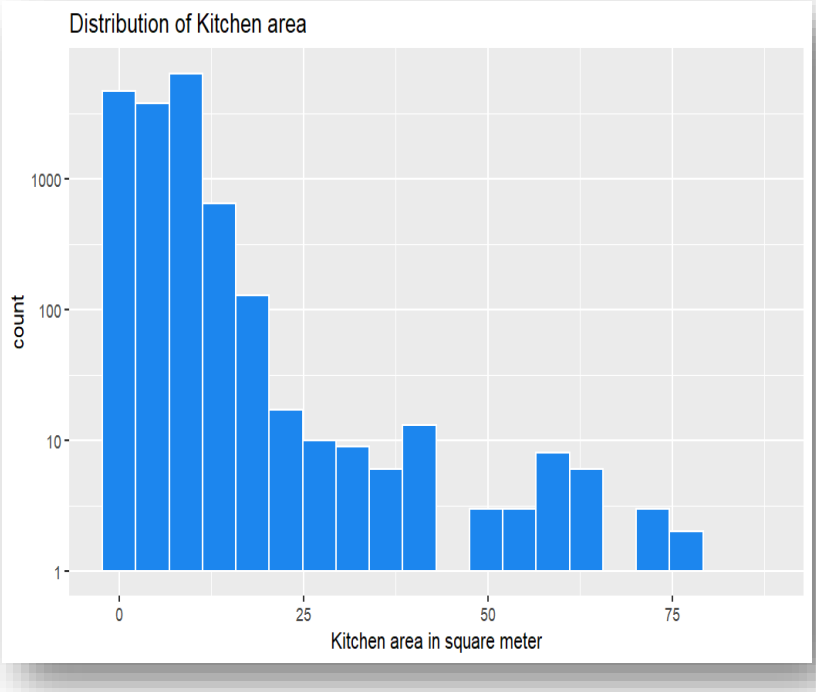


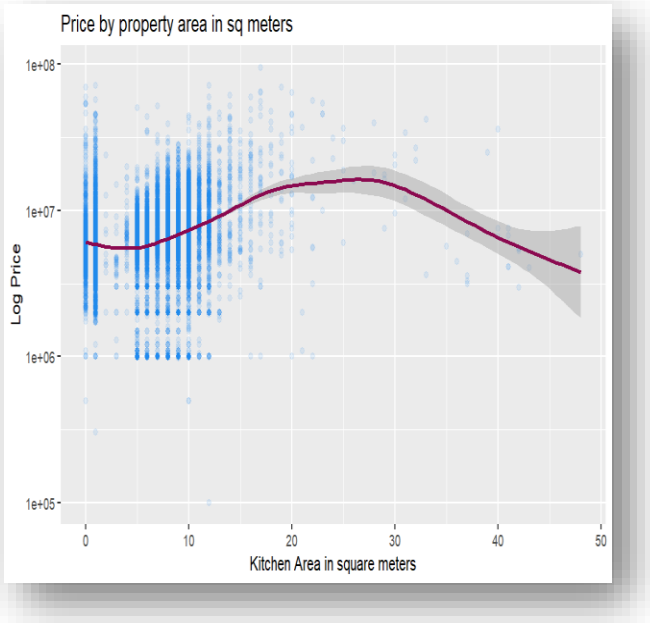
Here, we graph living area against the full property area. we expect to see all values of living are below that of property area. We see that there is a rather consistent ratio between living area and property area.

We remove outliers from the graph to have a better view of the relation.

Kitchen area

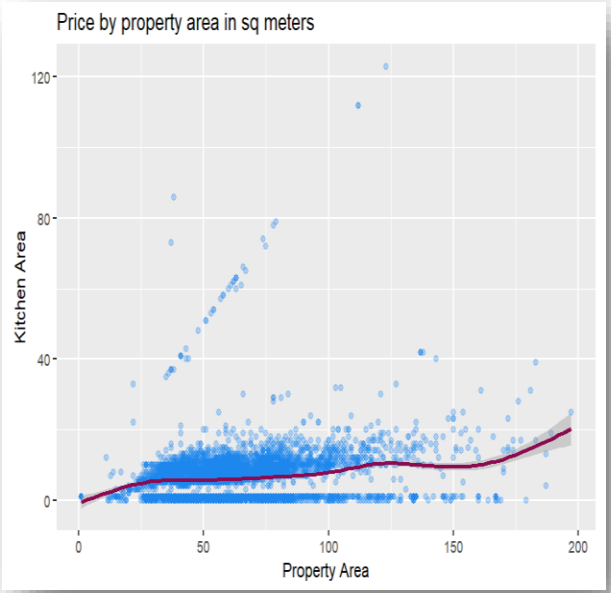
Here we have the histogram of the kitchen area.





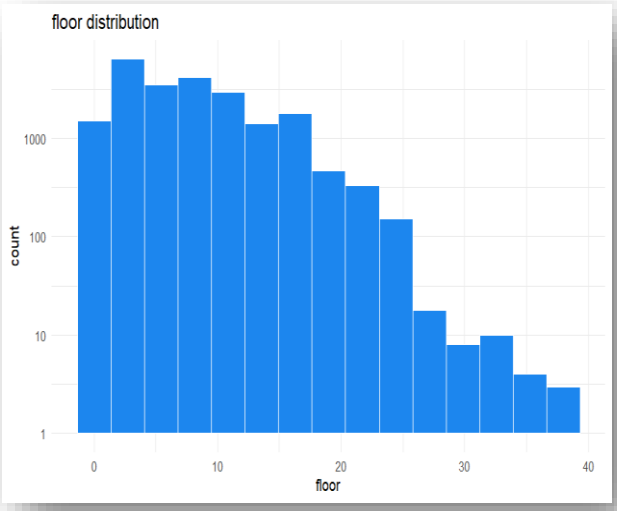
This is a scatter plot of the price by kitchen area. Although not completely linear, for a big part of the data we see a positive relationship.

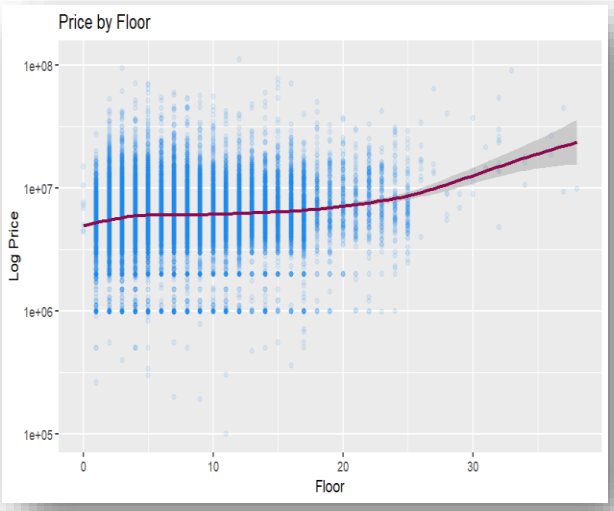
We graph the area of the kitchen against the property area. As one could easily justify, the kitchen area increases with a small slope and does not increase much as the area of the property. We remove kitchen values bigger than the property area.



Floor

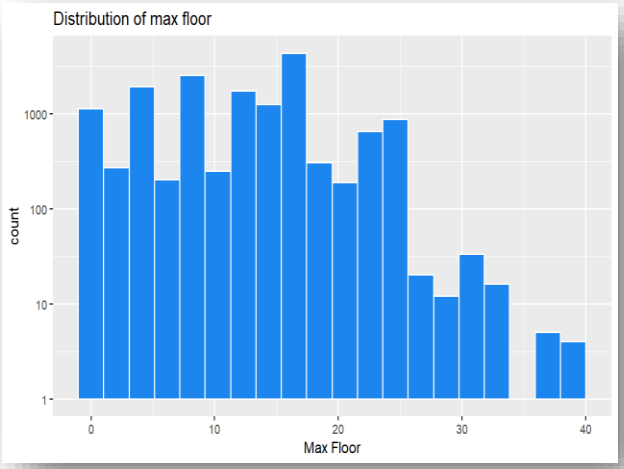
Here we have the distribution of variable floor. There is a small positive slope for relationship between price and property floor.



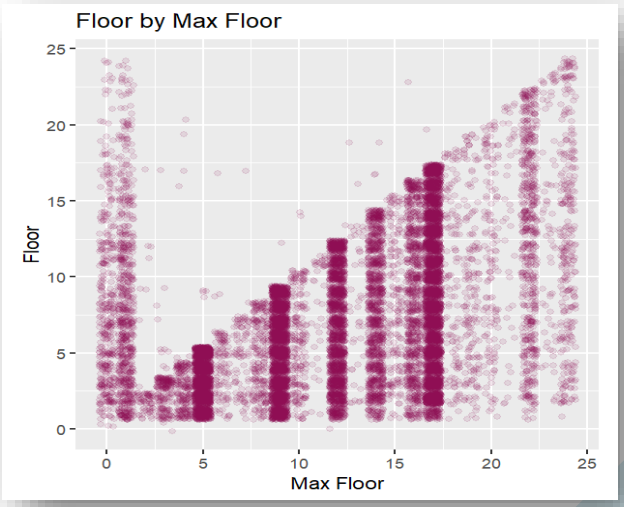


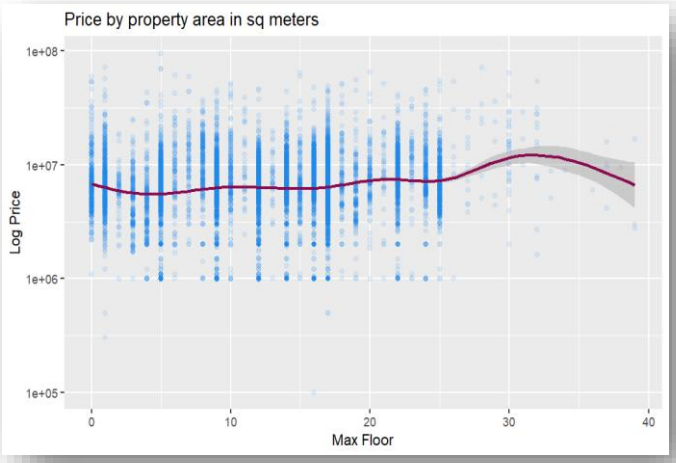
Max Floor

Here we have the max floor histogram.



We check the property floor against the maximum number of floors. We cap the graph axis on 25 floors and 25 max floors for the graph.

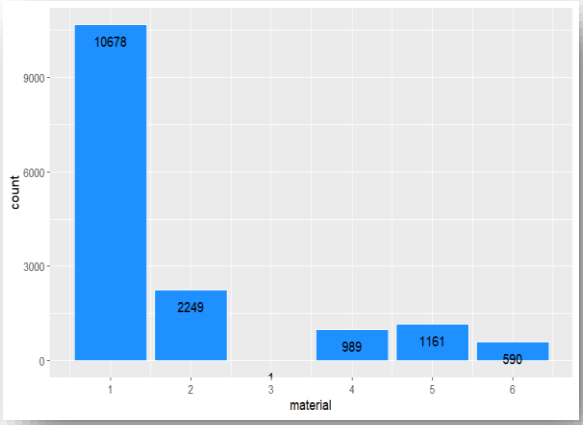


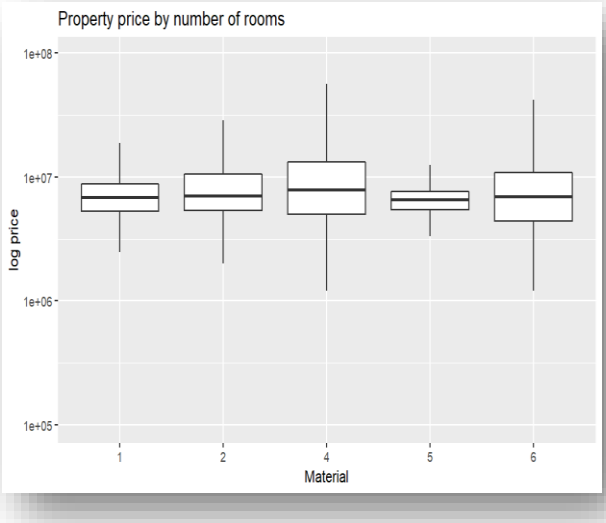


We remove max floors that are smaller than floor because these values cannot be logically correct. There is little relation between max floor and log price.

Material

Here we table the material of each house. We don't have further data to know what the materials are; hence, we keep data as it is. There is only one observation with material 3.

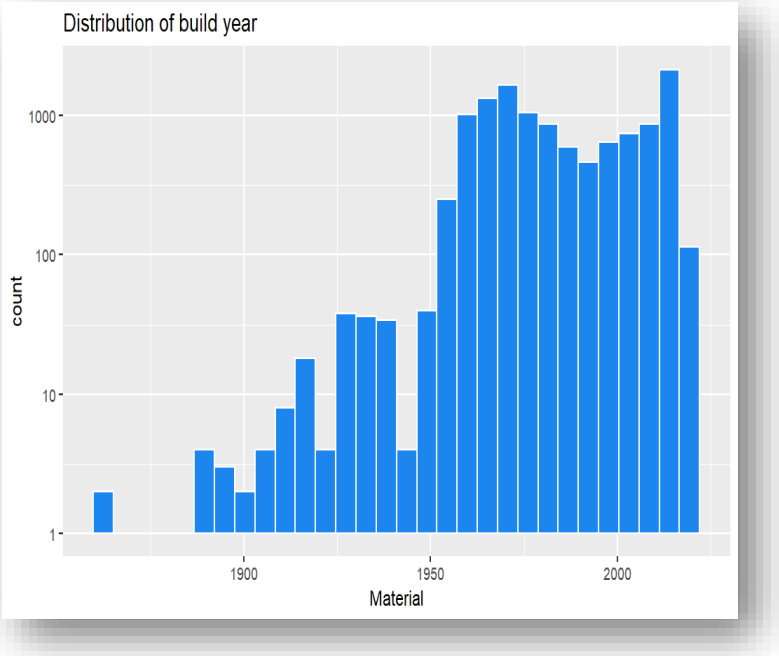


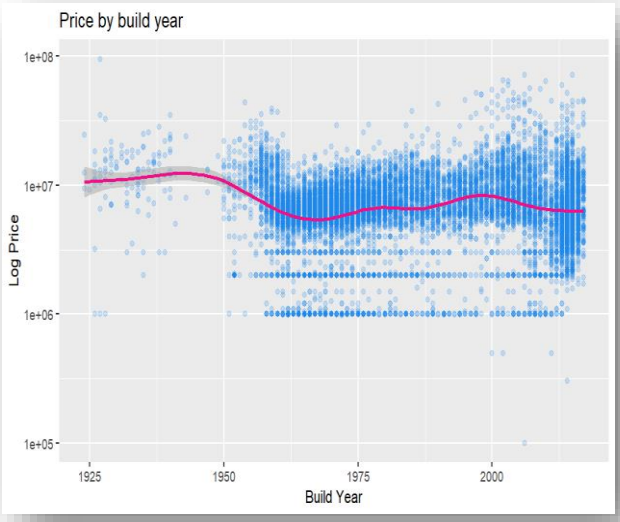


We use box plots to check the property valuation against material. The means are very close, while the variance is varying.

Build Year

Here we have the distribution of year that the buildings were built.

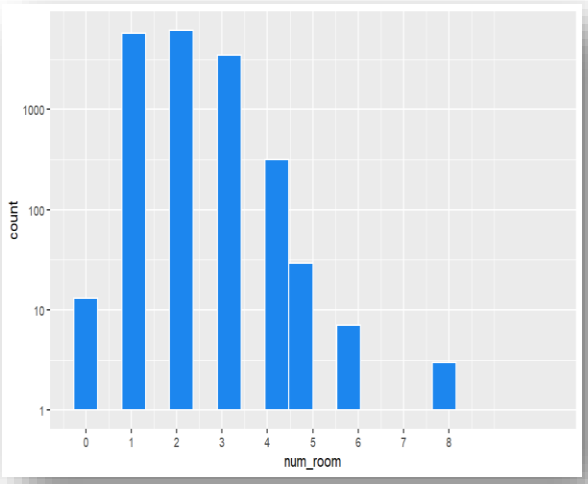




The plot of price against the built year is as follows. The relationship is not completely linear. And we can see is the increase in the spread of price for newer properties.

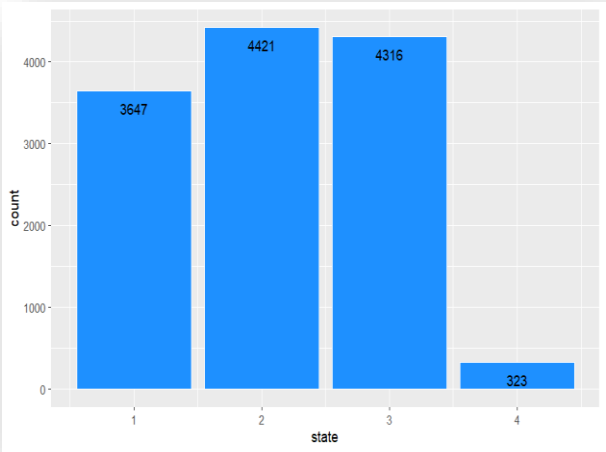
Number of rooms

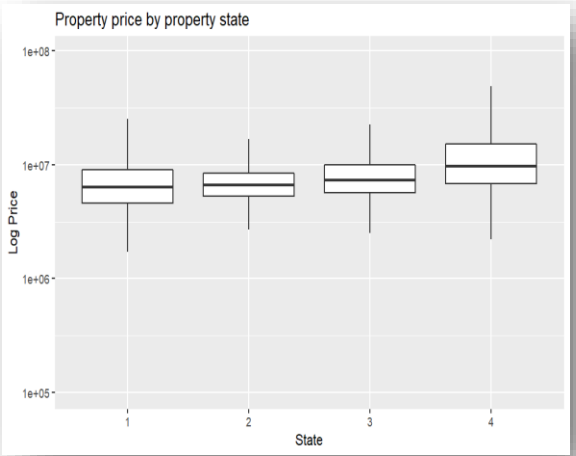
We use a histogram plot to investigate the distribution of number of rooms. We check the property price by the number of rooms, and as expected there is a positive correlation.



State

Here we check the property condition. We see a slight increase in the price by state.

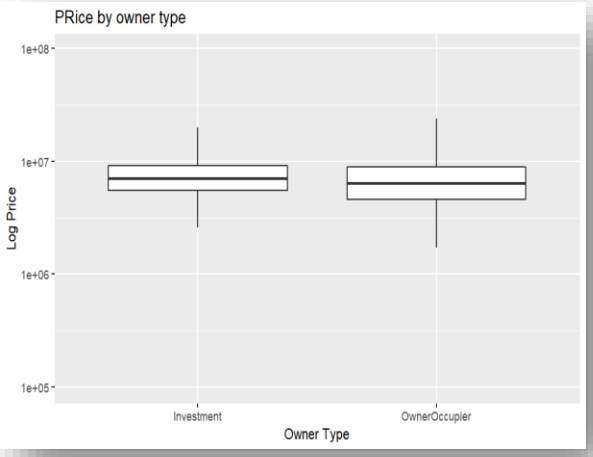


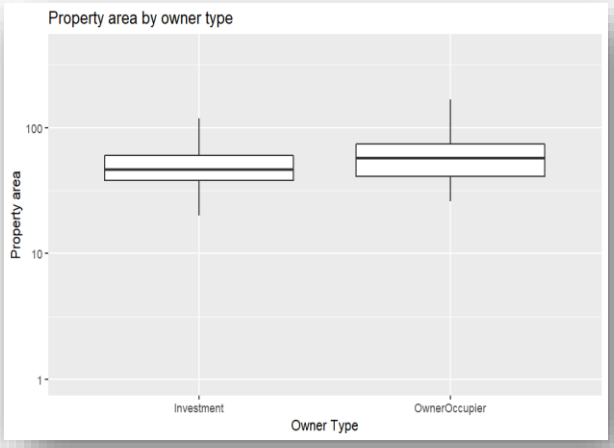


Product type

Here we have property value by owner and investors. From this graph we can see that the investors are buying more expensive properties.

Here we have property areas by owner types. Occupiers are buying bigger houses which can be justified by the fact that they are getting both the utility of living on the property and having it as an investment.

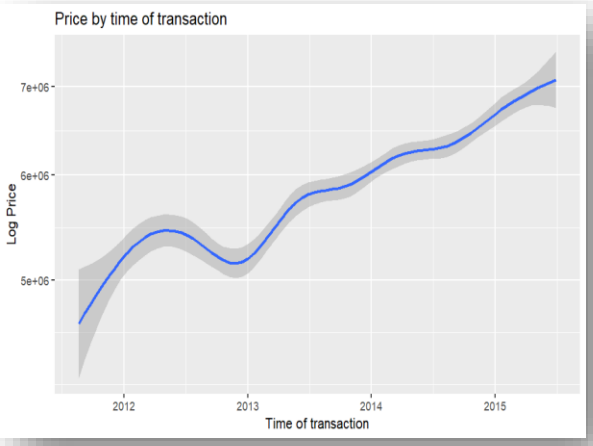




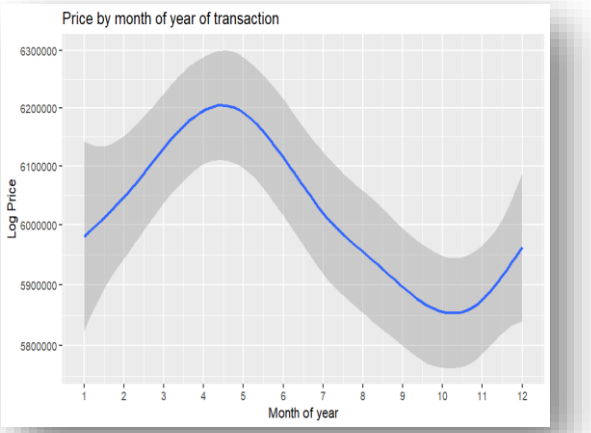
Macro Economics Data

Of the columns of Macro data, we have picked those we found most relevant.

Time of transaction



Here we check the price trend in our dataset and as we see the transaction value is continuously increasing.

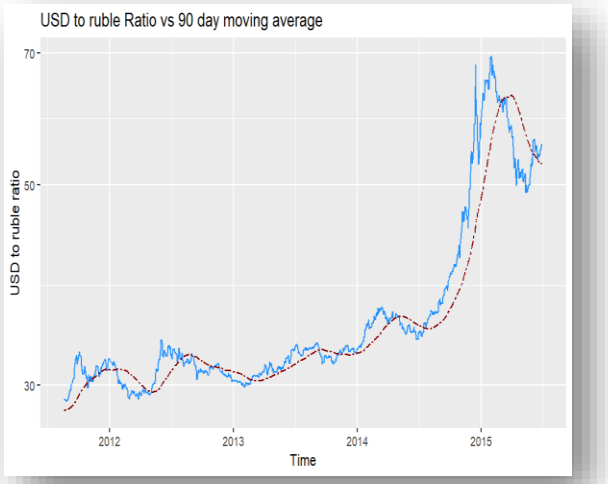


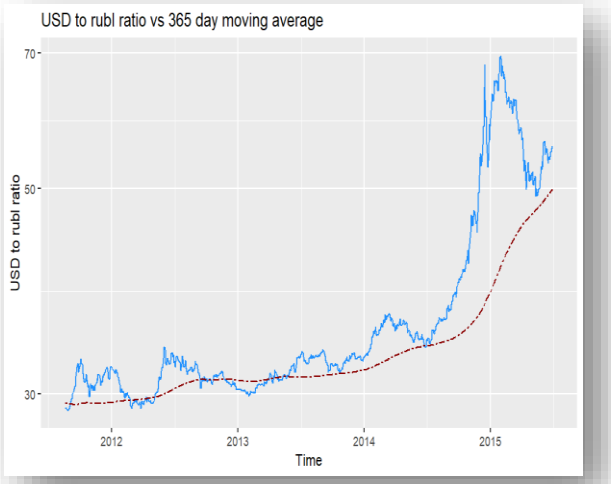
Now we check the scatter plot of price by month of the transaction, to check seasonality. The transactions in spring are of a higher value compared to winter. We create dummy variables for the months of the year to account for this seasonality.

US Dollar to Ruble

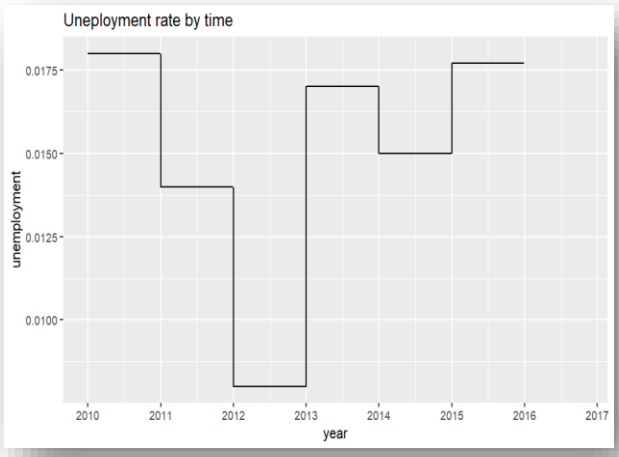
This graph is a proxy measurement of the Russia's economy. The upward trend shows how the currency has lost its value and hence, ballooning real state value as an asset that is more resilient to inflation, which could increase the demand for assets however it simultaneously decreases the purchase power, so we avoid driving any conclusion with the data at hand.

The following graphs are the moving average of the currency exchange rate against the daily exchange rate. As the shorter time periods were very close to the main graph, we skipped plotting them and instead limit the graph to the smooth versions. These moving averages are calculated as we suspect that stakeholders do consider historical volatilities.





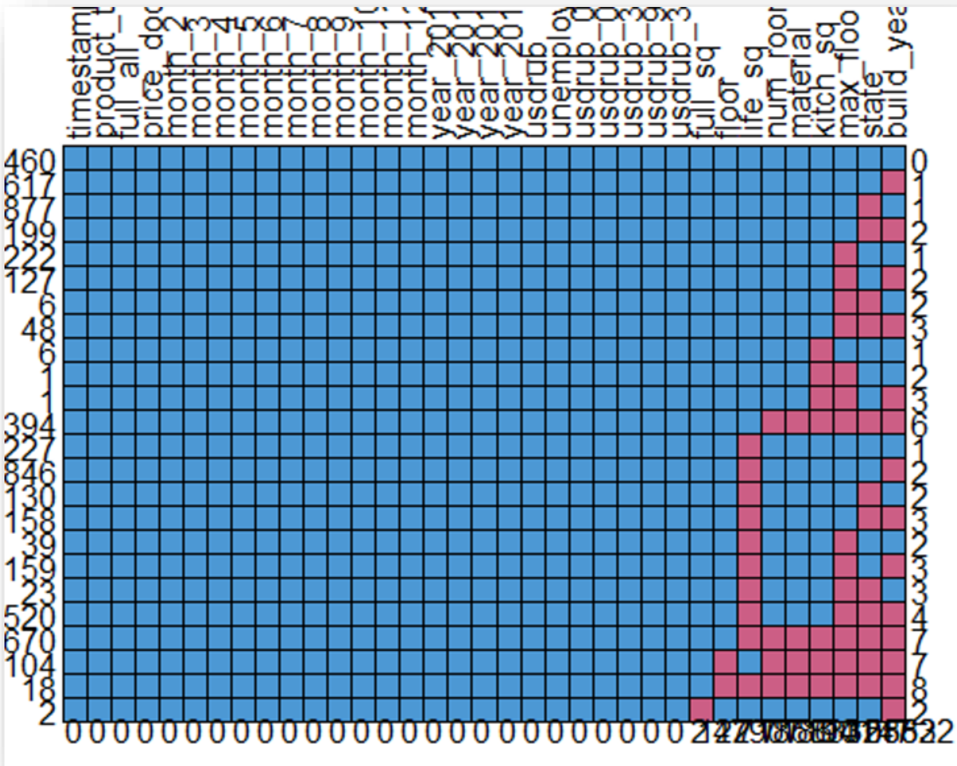
Unemployment



Unemployment is another macroeconomic factor which can affect the housing market. When unemployment is high it is a sign of slow economics and less purchase power and probably lower property prices.

6. Missing Value Imputation

Here we check the pattern of missing data, as we can see we have a case of multivariate missing values. In the graph, on the left we have the frequency of each pattern and on the right side the number of missing values.



Now we start imputing the missing variables using "Multivariate Imputation by Chained Equations" MICE is an imputation method that uses multivariate regressions to predict the missing values. We set a prediction matrix through which we select which variables should be used to impute each column. Since this is a stochastic procedure, there is not a best value and by running the random procedure, we produce a set of candidates that could be potential candidates and on average tend to be close to the real value. In practice in most cases 5 iterations do suffice and we are going the same number.

The imputation method used is Predictive mean matching (PMM). It is an advanced imputation method that relatively preserves the distribution of data. It tries to reduce the bias introduced in a dataset through imputation, by drawing real values sampled from the data. We also make several draws as our attempt is a stochastic procedure and we are not trying to find the best fit but a series of possible values.

With missing data imputed we check the correlation of columns. As the table is very large, we leave it for the appendix, and we just mention that there is a high correlation between time-related variables and measures of area, which could lead to collinearity, however since the main objective of the project is prediction, it does not hinder our efforts. We skip further technical descriptions.

7. Explanatory Model

The following table is the results of the robust regression table to further investigate the patterns in data. However, we must assert that all the observations are just correlations, and no causation should be assumed as there are many missing variables. There is the possibility of reverse causality as the price has a two-way relationship with "independent variables". In addition, the imputation method that we used has created a five-fold dataset which is fine for the purpose of prediction but does lead to lower p values for an explanatory model. Hence, we should consider the following table a correlation table with a dependent variable.

Features	Value	Std. Error	t value
(Intercept)	-2.44e+07	1.55e+06	-1.57e+01
full_sq	1.94e+06	2.35e+04	8.27e+01
life_sq	-1.97e+03	3.93e+02	-5.00e+00
floor	4.03e+04	1.24e+03	3.23e+01
max_floor	6.94e+04	1.31e+03	5.29e+01
build_year	1.06e+04	7.81e+02	1.36e+01
num_room	-2.14e+05	9.89e+03	-2.16e+01
kitch_sq	1.23e+04	1.34e+03	9.12e+00
state	3.29e+05	9.79e+03	3.36e+01
material	1.02e+04	3.75e+03	2.72e+00
product_typeOwnerOccupier	-1.24e+06	2.49e+04	-4.99e+01
full_all	-4.80e-01	4.59e-02	-1.04e+01
month_2	1.82e+04	3.07e+04	5.93e-01
month_3	1.04e+05	3.07e+04	3.40e+00
month_4	5.96e+04	3.24e+04	1.83e+00
month_5	1.99e+05	3.40e+04	5.85e+00
month_6	1.63e+05	3.39e+04	4.79e+00
month_7	2.10e+05	3.55e+04	5.90e+00
month_8	1.17e+05	3.62e+04	3.25e+00
month_9	1.99e+05	3.55e+04	5.60e+00
month_10	3.97e+04	3.60e+04	1.10e+00
month_11	8.43e+04	3.72e+04	2.26e+00
month_12	3.72e+04	4.06e+04	9.18e-01
year_2012	3.14e+05	4.00e+04	7.86e+00
year_2013	5.85e+05	4.13e+04	1.41e+01
year_2014	8.10e+05	5.33e+04	1.51e+01

year_2015	5.91e+05	1.28e+05	4.59e+00
usdrub	-1.13e+04	1.08e+04	-1.04e+00
usdrub_03da	5.33e+04	1.89e+04	2.81e+00
usdrub_07da	-1.82e+04	1.47e+04	-1.24e+00
usdrub_30da	-1.51e+04	9.47e+03	-1.60e+00
usdrub_90da	1.67e+04	6.53e+03	2.56e+00
usdrub_365da	1.92e+04	7.47e+03	2.57e+00
poly(full_all, 2)2	-9.68e+07	4.058e+06	-2.38e+01
(full_sq * build_year)	-9.12e+02	1.17e+01	-7.77e+01
(full_sq * full_all)	1.05e-02	8.00e-04	1.30e+01

Still there are interesting patterns present in the table, which could be studied in another project.

The Max floor has a larger coefficient than the floor, which could be interpreted as a sign that the max floor is carrying information about the quality of the property and its neighborhood. It is reasonable to hypothesize that taller buildings are better built, in better condition, and in better neighborhoods.

Owner occupiers are paying less for the properties that they are going to live in. In other words, we could say that investor buyers tend to buy more expensive properties. Hence it could be deemed as a sign that which houses should be advertised to whom.

We see that controlling for neighborhood population we get a statistically insignificant coefficient, however when we control for the second degree by squaring the values, we get a positive first-degree population coefficient and a negative second-degree. It shows that there is a nonlinear relationship between the variables. Considering the magnitude of the coefficients, the neighborhood population has a negative effect on property prices.

Investigating the dummy variables for the years 2012 to 2015, we see that these dummy variables also have statistically significant relations with the price. Compared to the prior year, all following years see an average increase in the valuation of properties except for 2015.

For seasonality, and trends within a year, we see that compared to the base month of January, prices do increase in months 5 and 9. However we do not observe what was shown in the graphs in our coefficients numerically and other months have statistically insignificant coefficients.

The total of the currency exchange rate for the day of transactions, moving average of last three days and last seven-day days have a positive effect on the price of the transactions. As this is more of a time series concept, we are not going to read too much into it, and we would just mention that it is justifiable that the property price would increase as Ruble gets weaker.

Moving average of exchange rate for the last 30, 90 and 365 days (about 12 months) are statistically insignificant, and it is somehow in contradiction with the gains made by XGBoost regressors, in which these are more prominent variables.

If we add the multiplication of area population and building area, we get a negative coefficient which could hint less desire for bigger houses in populated areas. The coefficient itself is not very large but the values that it represents are of magnitude of hundred thousand and hundreds, hence it could significantly affect the dependent variable.

We also can add the multiplication of full area and build year which has a negative coefficient, hinting that buyer are paying more for smaller new houses. Although the coefficient is comparatively small, the value of the year is in the 1000s and the value of the property area is in the 100s, hence the effect is interesting.

Although we are aware that to further solidify our findings, we should also consider multicollinearity and VIF matrix, as it is a prediction project, we are going to skip those parts. However, we find it illuminating to once again point out that correlation matrix could show different results from regression table, as regression helps us to isolate the effect of explanatory variables independently.

8. Prediction Modeling

The aim of this project is to make the best possible predictive model. Hence, we are going to use a variety of approaches. We use linear regression, penalized regressions, and Extreme Gradient Boosted Tree regression (XGBoost). These methods are some of the most used methods in data analysis. Linear regression has the advantage of being simple and interpretable and its wide usage makes it reasonable to use it as the base model. Penalized regressions are more recent and put a degree one or degree two penalty on coefficients removing those that are less relevant, and in many cases, they prevent model overfit. Extreme Gradient Boosted Tree regressions have the advantage of fitting data very well and providing accurate predictions on the learned intervals.

Linear Regression

We begin with linear regression. The dependent variable is the documented price, and we run the regression on all other variables. The sum of square prediction errors is as follows.

1.391945e+13

To make a better sense of the error rate we sum up the absolute errors and divide them by the total value of the property. The ratio is 0.307.

Shrinkage Model

The standard linear model does not perform well in situations where you have a lot of variables in your data set. A better alternative is penalized regression, allowing us to create a linear regression model that is penalized, for having too many variables in the model, by adding a penalty in the equation. The idea behind penalizing coefficients is the prevention of over fitting and better performance for unseen data. This allows the less contributive variables to have a coefficient closer to zero or equal to zero.

Ridge regression

Ridge regression penalizes the coefficients by using a term called L2-norm, which is the sum of the squared coefficients.

The amount of the penalty can be created using a constant called lambda. When alpha = 0, the penalty term has no effect, and ridge regression will

produce the original linear regression. However, as lambda increases to infinite, the impact of the shrinkage penalty grows, and the ridge regression coefficients will get closer to zero. Ridge regression will not set any of them exactly to zero.

Lasso Regression

Lasso stands for Least Absolute Shrinkage and Selection Operator. Lasso penalizes the regression model with a term called L1-norm, which is the sum of the absolute coefficients. In the case of lasso regression, it can cause variables with low significance to be exactly equal to zero.

Generally, lasso might perform better in a situation where some of the variables have large coefficients, and other variables have very small coefficients while ridge regression will perform better when the variable's coefficients are roughly equal size.

Elastic Net

Elastic Net is a combination of Ridge and Lasso. Elastic Net creates a regression model that is penalized with both the L1-norm and L2-norm. This method shrinks coefficients (like in ridge regression) and sets some coefficients to zero (as in LASSO).

With an alpha of 1, we start from Lasso and go to alpha 0 for Ridge and in between cover different measures for Elastic Net.
The glmnet package has the convenience of automatically normalizing and then scaling back the data.

Alpha	Error
[0]	1.461248e+13
[0.1]	1.454548e+13
[0.2]	1.448947e+13
[0.3]	1.435673e+13
[0.4]	1.445017e+13
[0.5]	1.457345e+13
[0.6]	1.447969e+13
[0.7]	1.440332e+13
[0.8]	1.45336e+13

[0.9]	1.483017e+13
[1]	1.452096e+13

Although lasso and ridge are supposed to help with the problem of over fitting it seems that our dataset is not currently facing this issue and these methods prediction is worse than OLS. Hence, we are not going to further inspect them. Including more variables in another analysis could change this outcome.

XGBOOST Regression

XGBoost stands for Extreme Gradient Boosting and is an implementation of gradient boosted decision trees with several advantages. It is fast and has been shown to outperform other models in many cases.

XGBoost creates many trees and reduced the residuals in each new tree. This model will also stop over fitting by pruning the tree with a gamma value. This is done by creating an initial tree using similarity to calculate gains and then pruning the tree by deciding to take off branches if they are less than the gamma. After a tree is created, we multiply results by a learning rate and make another tree based on those new results. Each new tree will have fewer total residuals and it stops when the predetermined number of iterations is done.

Tuning such models needs advanced expertise, however we can get reasonable results using random draws. We can set values that do cover a wide range and pick the one that shows the best performance on the validation set.

The other advantage is that the model does not need any normalization and scaling and addresses these issues automatically.

The grid of parameters that we have used is as follows. We used a "gbtree" booster, and we set our objective to be “reg:squarederror", the maximum depth for the trees have been drawn using random samples from integers 3 to 20. For eta we had the random number from the range (0.01, 0.3), subsample (0.7, 1), colsample_bytree (0.6, 1), min_child_weight (0,20).
The total Mean Squared Error value is as follows:
8.007801e+12
Which shows a 38 percent increase over linear regression model.

The next table gives a set of importance metrics for each variable in the XGBoost model.

	Feature	Gain	Cover	Frequency
1	full_sq	0.4940167436	1.953522e-01	0.1515534226
2	full_all	0.1704204475	2.154194e-01	0.1758019702
3	build_year	0.0564605609	7.859183e-02	0.0929527658
4	life_sq	0.0421239630	4.782975e-02	0.0745137661
5	usdrub_90da	0.0419438837	6.824914e-02	0.0596110129
6	max_floor	0.0253345766	4.024231e-02	0.0444556706
7	floor	0.0250855264	5.974993e-02	0.0636524375
8	num_room	0.0212703116	2.802260e-02	0.0171760546
9	usdrub_365da	0.0189577574	4.242886e-02	0.0409194241
10	kitch_sq	0.0188203949	3.750879e-02	0.0442030816
11	state	0.0176277257	1.718803e-02	0.0224804243
12	usdrub_30da	0.0130104462	3.859387e-02	0.0381409447
13	usdrub	0.0112584498	2.213462e-02	0.0421823693
14	product_type	0.0094632709	5.763144e-03	0.0113665067
15	usdrub_07da	0.0080983036	3.926705e-02	0.0303106845
16	usdrub_03da	0.0078163284	3.036030e-02	0.0328365749
17	material	0.0055616659	1.111162e-02	0.0174286436
18	unemployment	0.0024988483	7.320724e-03	0.0030310685
19	month_9	0.0019044494	3.665741e-03	0.0042940136
20	year_2012	0.0010297993	1.689341e-04	0.0032836575
21	month_11	0.0009994174	4.052926e-04	0.0025258904
22	month_5	0.0009864436	1.233453e-03	0.0027784794
23	month_3	0.0009735116	4.961293e-03	0.0030310685
24	year_2013	0.0007683067	8.076028e-04	0.0047991917
25	month_12	0.0007033689	2.789649e-04	0.0027784794
26	year_2014	0.0006625337	3.654557e-04	0.0035362465
27	month_8	0.0005983229	3.036765e-04	0.0020207123
28	month_6	0.0004320488	2.686328e-04	0.0022733013
29	year_2015	0.0003240998	1.054506e-05	0.0005051781
30	month_2	0.0003124575	7.463562e-04	0.0015155342
31	month_4	0.0002332222	3.323293e-05	0.0012629452
32	month_10	0.0001627496	3.055938e-04	0.0015155342
33	month_7	0.0001400641	1.311103e-03	0.0012629452

Importance provides a score that indicates how useful or valuable each feature was in the construction of the boosted decision trees within the model. The more a feature is used to make key decisions with decision trees.

The gain means the relative contribution of the variable to the model calculated by taking each variable's contribution for each tree in the model. The Gain is the most important feature in the importance table. The coverage equals the relative number of observations related to this variable. The frequency is the percentage representing the number of times a particular variable occurs in the trees of the model.

The model that we have created for our project is used for computing the value/price of the housing using only certain characteristic features and macro-economic variable data. We have included certain limited intrinsic characteristics of the house and macro-economic variables due to time and resource constraint since this is an educational project, but these can be further expanded depending on the needs and requirements of the situation.

Total area including non-residential (in m2) holds the maximum contribution to our model (49.4%) and therefore is the most important housing characteristic affecting the price. This is followed by subarea population (17%), year of construction (5.64%), total built up area in m2(4.21%), exchange rate of ruble to us dollars with its 90-day moving average (4.19%) and so on.

The variables mentioned above hold the highest volume of contribution to our model in predicting the price of the houses in the respective order. The main use of this model is to calculate the housing price as accurately as possible with relatively fewer errors using statistical computations and model.

9. Conclusion

Our project goal was to predict the housing value based on several housing characteristics and macroeconomic variables from the data available and then create a model that can predict the values of the houses from unseen data.

We selected 12 variables from the House characteristics dataset and 2 variables from the macroeconomics dataset and did exploratory data analysis on the important variables to get the data insights. By Data Visualization we got some expected results such as with an increase in property as the prices increase, etc. but also some surprising outcomes like the state of the property doesn't affect the price of the property and there's a seasonality by the month in the number of transactions. The number of transactions increases in spring and decreases in the winter.

We tried to create the best models after running various models based on simple OLS regression, generalized linear model (ridge, lasso and elastic net) and the XGBoost model and selected the model that gives us the least amount of errors. The XGBoost model outperforms the simple linear regression by almost 38 percent.

The linear model had a smaller residual error compared to penalized models. This is most likely because we did not have many columns for the models to better fit. The XGBoost model had a lower residual because our data is not very linear. The XGBoost allow more flexibility in complex relationships in the data and accounts for over fitting.

10. Applications, Future use and Recommendations

Our model for predicting the housing value/price has a wide range of applicability and can be used by various stakeholders of the real estate industry including any individual or institution or a company interested in knowing and predicting the housing values. Some of the users for our models are as follows:

- **Banks and banking corporations** – The institutes can be interested in estimating housing prices for mortgage purposes. Market valuation of the houses as the housing wealth has a much larger impact on consumption than do changes in the stock market wealth.

- **Individual Buyers** - Our model can be a good starting point for individual buyers interested in prices of houses for their residential or investment purposes.

- **Real estate agents/agencies** – Estimating prices based on our prediction model can be useful source for real estate agencies and agents involved in the housing business.

- **Construction companies/Individuals/Developers** – companies and individuals involved in construction may use this model to predict the value of the houses they are building for the calculations of estimated profit based on the characteristics they are hoping to provide in the house.

- **Housing census Bureau** - The model can be used by the housing census bureau to predict the housing patterns in the country.

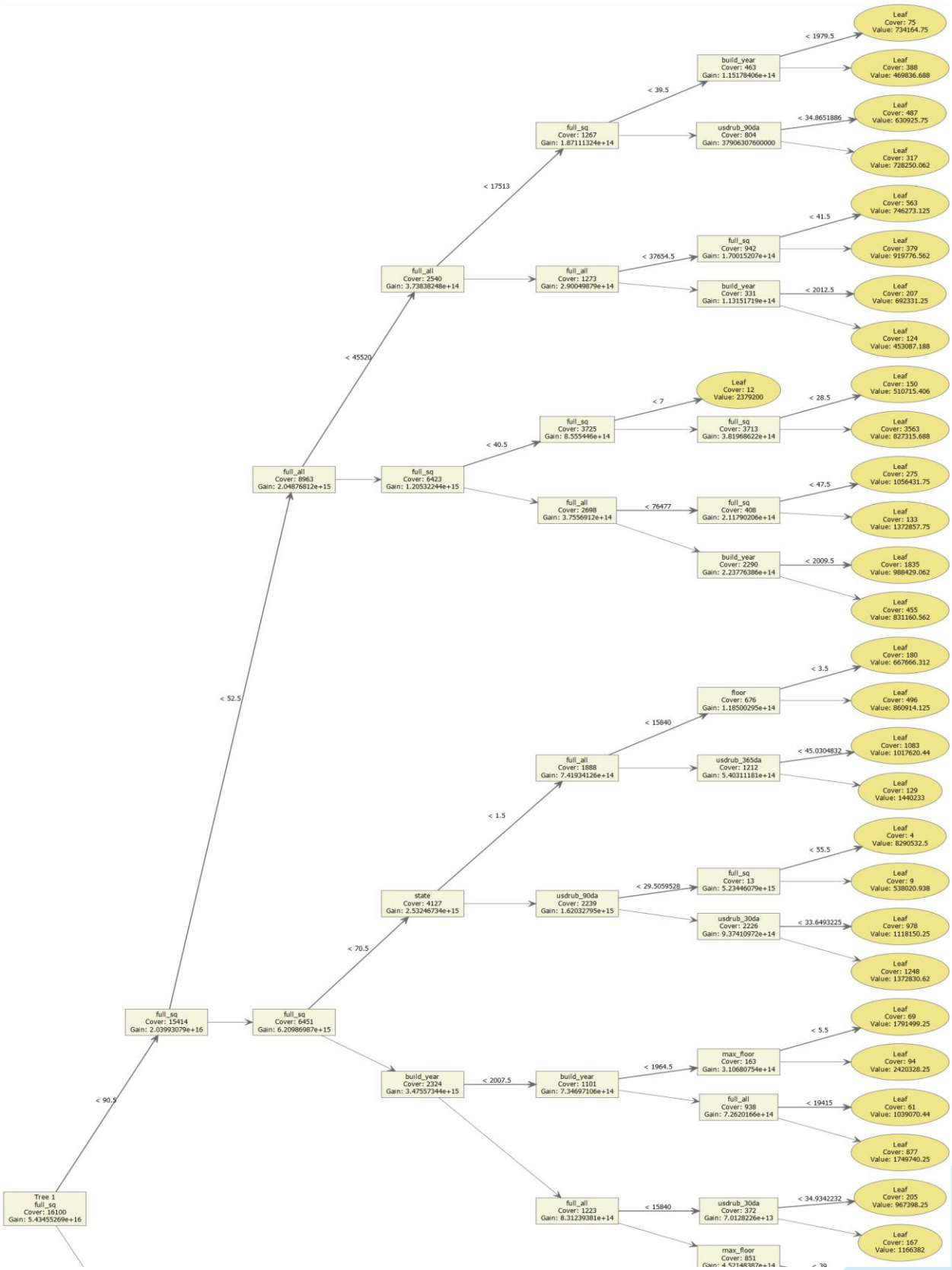
- **Insurance companies** - Insurance companies can use the model for the purpose of pricing their securities and compensating insurers for their damages.

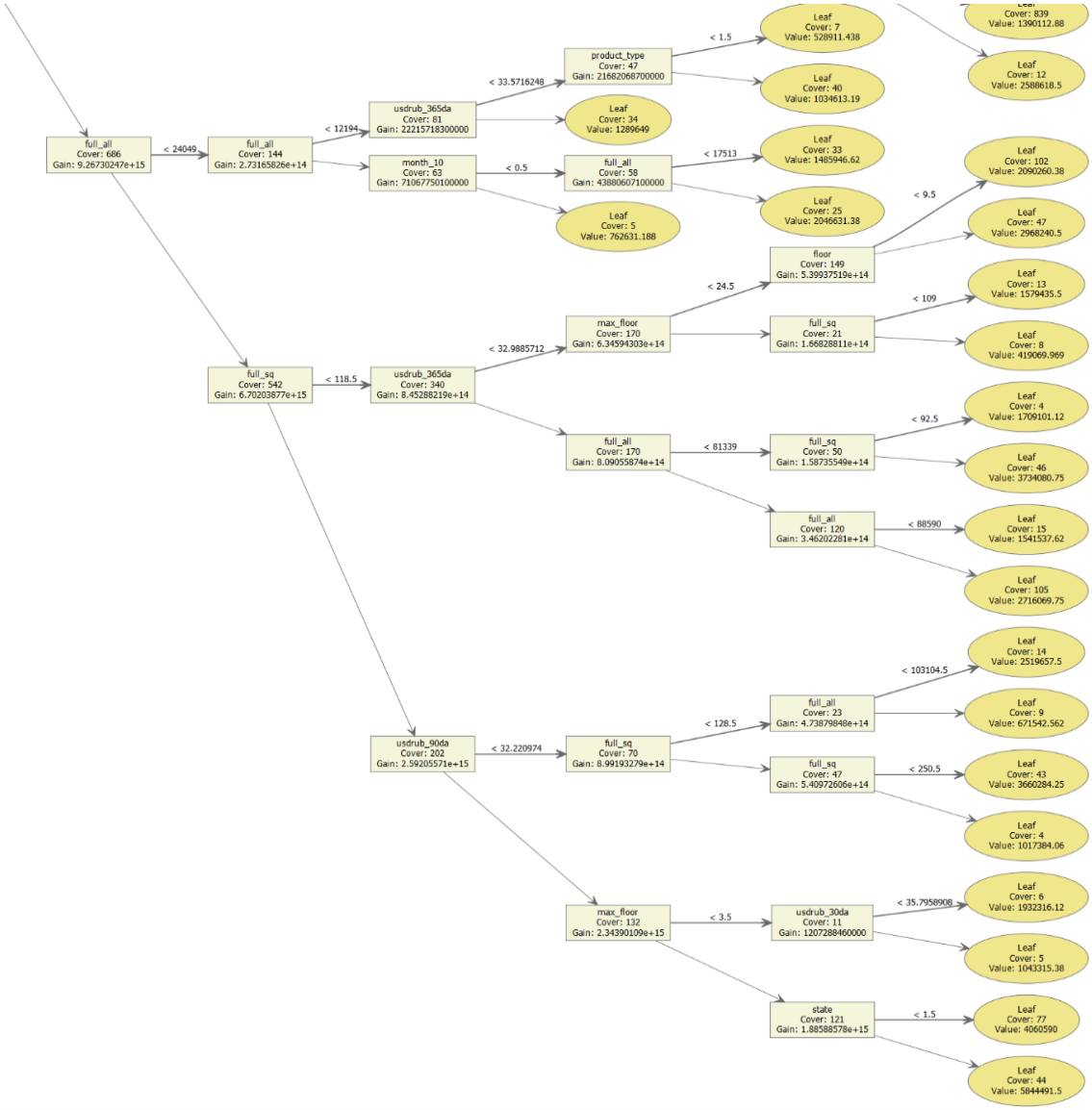
- **Hedge funds** – Hedge funds interested in investing money in the housing market can use our model to leverage their risks.

- **Landowners** – can use the model for predicting the value of the housing property if they are interested in developing their land to build a house either for residential or investment purposes.

This is a preliminary model and can be used as a basis for more advanced and detailed models.

11. Decision Tree




full_all
Cover: 686
Gain: 9.26730247e+15
< 24049
full_all
Cover: 144
Gain: 2.73165826e+14
< 12194
usdrub_365da
Cover: 81
Gain: 22215718300000
< 33.5716248
product_type
Cover: 47
Gain: 21682068700000
< 1.5
Leaf
Cover: 7
Value: 528911.438
Leaf
Cover: 40
Value: 1034613.19
Leaf
Cover: 34
Value: 1289649
month_10
Cover: 63
Gain: 71067750100000
< 0.5
full_all
Cover: 58
Gain: 43880607100000
< 17513
Leaf
Cover: 33
Value: 1485946.62
Leaf
Cover: 25
Value: 2046631.38
Leaf
Cover: 5
Value: 762631.188
full_sq
Cover: 542
Gain: 6.70203877e+15
< 118.5
usdrub_365da
Cover: 340
Gain: 8.45288219e+14
< 32.9885712
max_floor
Cover: 170
Gain: 6.34594303e+14
< 24.5
floor
Cover: 149
Gain: 5.39937519e+14
< 9.5
Leaf
Cover: 102
Value: 2090260.38
Leaf
Cover: 47
Value: 2968240.5
full_sq
Cover: 21
Gain: 1.66828811e+14
< 109
Leaf
Cover: 13
Value: 1579435.5
Leaf
Cover: 8
Value: 419069.969
full_all
Cover: 170
Gain: 8.09055874e+14
< 81339
full_sq
Cover: 50
Gain: 1.58735549e+14
< 92.5
Leaf
Cover: 4
Value: 1709101.12
Leaf
Cover: 46
Value: 3734080.75
full_all
Cover: 120
Gain: 3.46202281e+14
< 88590
Leaf
Cover: 15
Value: 1541537.62
Leaf
Cover: 105
Value: 2716069.75
usdrub_90da
Cover: 202
Gain: 2.59205571e+15
< 32.220974
full_sq
Cover: 70
Gain: 8.99193279e+14
< 128.5
full_all
Cover: 23
Gain: 4.73879848e+14
< 103104.5
Leaf
Cover: 14
Value: 2519657.5
Leaf
Cover: 9
Value: 671542.562
full_sq
Cover: 47
Gain: 5.40972606e+14
< 250.5
Leaf
Cover: 43
Value: 3660284.25
Leaf
Cover: 4
Value: 1017384.06
max_floor
Cover: 132
Gain: 2.34390109e+15
< 3.5
usdrub_30da
Cover: 11
Gain: 1207288460000
< 35.7958908
Leaf
Cover: 6
Value: 1932316.12
Leaf
Cover: 5
Value: 1043315.38
state
Cover: 121
Gain: 1.88588578e+15
< 1.5
Leaf
Cover: 77
Value: 4060590
Leaf
Cover: 44
Value: 5844491.5
Leaf
Cover: 839
Value: 1390112.88
Leaf
Cover: 12
Value: 2588618.5

12. References:

1. https://www.kaggle.com/c/sberbank-russian-housing-market

2. https://www4.stat.ncsu.edu/~post/josh/LASSO_Ridge_Elastic_Net_-_Examples.html

3. https://stefvanbuuren.name/fimd/ch-introduction.html

4. https://amices.org/Winnipeg/

5. https://www.gerkovink.com/miceVignettes/

6. https://www.prnewswire.com/news-releases/us-housing-market-has-doubled-in-value-since-the-great-recession-after-gaining-6-9-trillion-in-2021--301469460.html

Thank You