Business Analytics

Final Project

House Price Analysis for Local Armenian Market

Team Members:

Anna Martirosyan

Tamara Sedrakyan

Meri Torosyan

Professor: Hrant Davtyan

Abstract

Using the Armenian housing price data of April, 2020, this paper explores the direct relationships between house price and features such as sqm size, number of rooms, agency, district, building type, etc. The analysis finds out that from the variables present, especially the size and district have a large influence in the house price. Also, it becomes clear that some newly constructed buildings located mainly in the center have different pricing from those in the suburbs.

May 16th, 2020

American University of Armenia

Table of Contents

Introduction

Real estate is not something people buy every day, but when doing so they try to understand and find the best deal for both their needs/preferences and their financial abilities. Many times we hear more about macroeconomic factors affecting house prices, such as inflations or some bigger scale changes in the country. However, it is also important to keep in mind that houses themselves, with their attributes can have varying prices and it is crucial to understand which are those specific factors. And this is not only for the buyers, for the sellers as well, knowing how to price the house and what to expect from the market can boost their business further.

This paper will try to understand the relationships between house feature variables and its prices for the local Armenian market, based on the data available in different online platforms. Different statistical models will be constructed to find the best model explaining the variation of the price. Some basic data visualisations will also contribute to the final analysis.

Related Work

In an economy predicting the variations in house prices and factors influencing these changes have massive importance. Different parties such as tax authorities, banks and mortgage providers, real estate investors and portfolio managers rely on this information to come up with strategically made decisions. Accordingly, in academic research about real estate, the application of machine learning techniques has increased, which helps to investigate the relationship between the dependent and independent variables statistically.

Various articles and academic papers have learnt and analyzed house price variations and tried to develop a prediction model. All of them are concentrated on certain factors, however, they all have a big contribution to the academic development of this topic. As stated, the house is a heterogeneous good embodying a package of inherent characteristics relevant to the location, property attributes, and environmental amenities (Hong, Choi, Kim 2019). To find out how our model and analysis fits into the understanding of other similar research papers, we have examined the methods they have used. Numerous studies have found a house's number of bedrooms and bathrooms and its floor area to be positively related to its price. One of these articles discusses spatial and temporal dependence on the house price prediction. It gives a lot of insights about the influence of neighbourhood/location of the house on its price. Moreover, as it examines the changes based on the housing transactions, the research states that past housing transactions may proxy for the market trend of general house price development, and can also capture changes in an institutional setting, such as changes in tax laws, local amenities (Liu, 2012).

To compare the spatial and temporal dependence on the house price Xiaolong Liu examines different models. It helps to distinguish that in the long run, the housing attributes prices will be driven by the supply of these reproducible attributes with non-constant production costs (Liu, 2012). However, both in the long run and in the short run irreproducible housing attributes will be determined by the demand. Such attributes can be lake view and highway access because there is no supply to accommodate the increasing demand for such housing attributes over time. The lake view is irreproducible, so as people become better well-off over time, the demand will increase and because of fixed supply prices will rise. A research conducted in 2019 also supports this statement, claiming that location on a site with a desirable view, such as a lake or golf course, has a positive effect on the price (Hong, Choi, Kim 2019).

Another model examined by the same author shows that having a elevator in the building will positively affect the transaction by approximately 1.3%, which is reasonable since people are willing to pay for convenience in their living surroundings.

Some interesting findings from the set of variables used in these datasets include the significance of dwelling age and the fact that apartments, which are not on the ground floor, are the cheapest.Surprisingly the number of rooms and its squared term does not have a significant impact on the housing transaction prices. This model also included having a sunny living room and good maintenance for the interior and exterior, parking possibility as significant variables influencing house prices (Liu, 2012).

All of the papers and theoretical analysis help to explore that from a theoretical perspective, there are significant market complexities, which make the value determination process unable to be fully accounted for, which will also help to explain many variations in the model analyzed in this paper. There will be several factors discussed in this paper, which were analyzed in the above discussed academic papers, such as number of floors, the district of the house and sqm. Accordingly, it is interesting to analyze how the variables in this dataset perform in comparison to the other datasets discussed above. Moreover, some other variables such as “Brokers” and “Agency” are also included, which can change the performance of the model.

To conclude, the above discussed models were findings from Dutch Randstad region and South Korea. This research aims to fill the gap of such academic analysis in the real estate industry for Armenian market.

Dataset and Features

For the analysis, the local Armenian market housing information for April, 2020, will be used, scraped from three websites: List.am, MyReality.am, real-estate.am. The data has 30,520 observations, together with 16 variables in total, which are listed and explained below:

Source - originally an object type of variable, showing from which website the housing information comes from. Thus, it can have 3 values: list_am, my_reality, real_estate.
Rooms - Float variable, showing how many rooms the flat has.
City - Object variable indicating the city location of the house. All the observations are from Yerevan.
ID - Integer variable, showing the ID of the house on the website where it is posted.
nFloor - Float variable, indicating the total number of floors the building has where the house of observation is located.
Broker - Object, listing the names of brokers and agencies in charge of the house sales. 18 options in total.
District - Object variable, shows the specific district of the flat in Yerevan.
Sqm - Float variable, indicates the area of the house in square meters.
Currency - Object variable, showing which currency is used for the price. Three options available: dollar, dram, euro.
Agency - Float variable, having 2 unique values: 1 and 0.
Price - Float variable, showing the price of the house in the specific currency.
HTML_ID - Integer, HTML ID coming from the website.
Type - Object, showing the type of the building, the material it is built with.
Floor - Object, showing the number of the floor the house is at.
Price_USD - Float variable, showing the price of the house in US dollars only.
Price_per_sqm - Float, shows the price per square for each row.

Cleaning the Dataset

Before starting the analysis, it was necessary to clean the data. The data originally had 19,795 null values, no duplicates, and no single-valued columns.

Some of the columns were dropped from the data, which were “ID”, “HTML_ID”, bringing no value to the analysis. Also, the “City”, which turned out to have only one unique value, but written in different ways, was also dropped. Next, the “Price” and “Currency” were dropped together, in order to have one indicator of price only - “Price_USD”.

In the variable “Type”, one of the vales “Փայտե” was present only 3 times, those observations were dropped because the sample size was not big enough to draw conclusions.
In the variable “District” similar areas with different wording were merged together under the same names.
In the variable “Floor”, the value “Նկուղ” was replaced with 0, to represent the underground floor.

In the variable “Broker”, NaN values were replaced with “ալլ”.
The variable “Agency” was converted to a categorical one.
The variable “Floor” was converted to a numeric type.

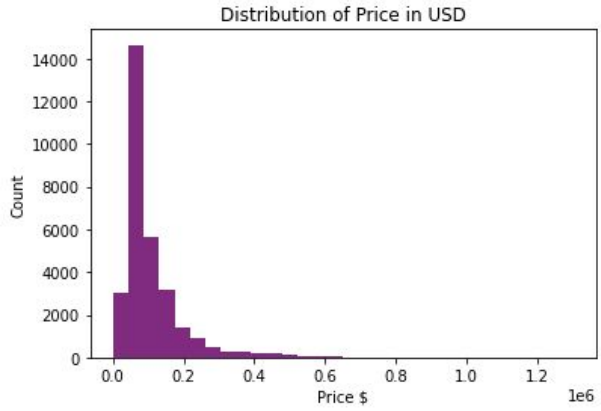
All the rest of the N/As were dropped from the data as well.

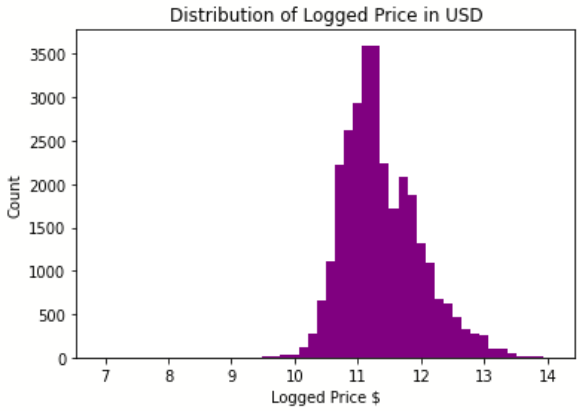
Methodological Approach

The core purpose of this analysis is to determine the real estate prices and the main factors affecting it. The model built will try to analyze the relationship between the dependent variable “Price_USD” and the rest of the independent variables mentioned above. OLS, Decision Tree Regression, and Random Forest estimators will be used.

Since there are many observations and comparably fewer number of columns, the first model was constructed using all the variables available after the cleaning process. Afterward, backward elimination was used to eliminate insignificant ones. Apart from this, some other feature engineering was conducted, such as adding interaction points: Sqm/rooms, Floor/nFloor, Floor_lg (Logarithm of Floor), Sqm_lg (Logarithm of Sqm), Rooms_Floor (Rooms times logarithm of the Floor), cent_penthouse (the last floor houses in the center, referred to as penthouses).

As a dependent variable “Price_USD” was used, however, because of its right-skewed distribution its log was calculated, making the distribution a bit more normal.





In the end, the best possible linear model was left with the following variables only: “nFloor”, “Sqm_lg”, “Floor”, “Price_USD”, “Sqm/rooms”, “Source”, “District”, “Type”, “Agency”, “cent_penthouse”. The R-Squared was 0.796, meaning that the independent variables chosen were able to explain 79.6% of the price variation.

R-square Adjusted has the exact same value as the R-squared itself, which means that the model has a good performance and is not changing its results based on the number of independent variables added. Also, with the variables mentioned and with a p-value threshold of 0.05 we can see that all the variables are statistically significant, except for the "Floor".

Besides showing good results for the sample, the model needs to be checked for the outside population as well, this is why train and test split was used next. Fitting the linear OLS model again, the R-squared for train and test were 0.754 and 0.793 accordingly, which means that the model has almost a similar performance for both groups.

To get better results, a polynomial second degree model was used next, helping to raise the R-squared of the test and train to 0.817 and 0.813 accordingly. Polynomial second degree regression was done because some quadratic relationships between variables and Price_USD were detected in the pairplots.

For comparison purposes, Mean Squared Error (MSE) for the model was calculated as 0.0756 and 0.0739 for train and test. Decision tree regression model showed an MSE of train 0.0784 and of test 0.0786. While Random Forest, which unlike the decision tree tries to reduce variance by training on multiple data samples, showed results of 0.0802 and 0.0777 for the train and test.

The cross validation score mean was the following: Linear Regression 0.0792, Decision Tree 0.0865 and Random Forest 0.0882.

It could be concluded that polynomial regression gave us the best model possible for the data provided, because of its MSE being lower than the rest and for having almost the same value for both train and test.

Internal Validation: LINE

After getting the best possible model, we do internal validation to see whether the underlying assumptions are held. First we checked the LINE assumptions visually.

Linearity and homoscedasticity assumption is checked visually by drawing a scatter plot between the fitted values and error terms. We can see that visually there are not any particular distribution patterns for the residuals vs predicted values relationship. However there is a slight difference in the variance over different predicted values. Variance of residuals is smaller for the lower priced flats. In spite of these small points, visually the linearity and equal variance assumption is satisfied.

Normality assumption is analyzed perceptibly by plotting the histogram of residuals and by QQ-plot. With both plots we can detect violation of normality assumption. With the residuals of the second degree polynomial model, the results are better, however still there is some

violation. To meet the normality assumption we tried to normalize the Price USD variable via box-cox transformation. However a little negligible improvement in normalizing assumption, resulted in significant sacrifice of accuracy.

We also conducted tests to check the LINE assumptions.

1. Linearity - Rainbow test was conducted, giving a large p-value of 2.25 (threshold = 0.05), which means that the regression is correctly modelled as linear.
2. Independence - Variance Inflation Factor is the way to check multicollinearity. Some of the variables are correlated, with a VIF value larger than 10.
3. Normality - Shapiro-Wilk test was conducted, where the p-value was 0, significantly lower than the threshold. This means that the normality assumption is not held, as already discussed above as well.
4. Equal Variance - Breusch-Pagan test was conducted to make sure there is no heteroscedasticity. P-value was 3.65, higher than the threshold, which means we fail to reject the null hypothesis of homoscedasticity.

Taking into consideration the factors mentioned above, not all assumptions are held, however we have got a model with quite good accuracy. We do not have the Best Linear Unbiased Estimator(BLUE), but the model is still a good estimator of reality.

Experiments and Results

As explained above, except for Floor variable, all the rest were statistically significant, meaning that they are able to explain the change of the house price to some extent. Some of the most influential and interesting variables are discussed below.

When the building where the house is located has one more floor (nFloor), this will lead to 100*0.0023 = 0.23 percent change in the price.

On average 59.41 percent increase in mean price is expected if we observe “Աղյուսե” type of flat compared to the “Մոնոլիտ” type.

When the district of the house is changed from “0” (not indicated) to “Դավթաշեն”, its mean price will decrease with on average 10.36 percent.

When the district of the house is changed from “0” (not indicated) to “Կենտրոն”, its mean price will increase with on average 40.93 percent.

When there is a one percent change in square meter of the house, it leads to on average 0.8112% increase in the mean price.

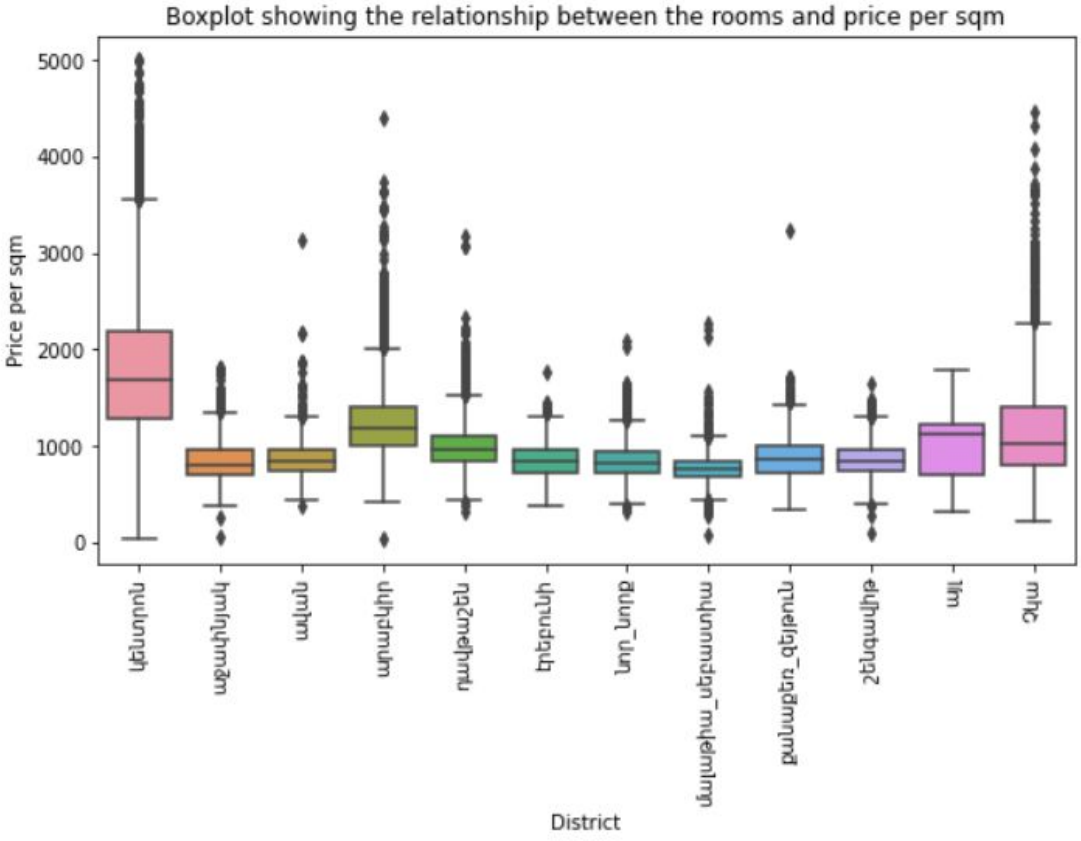
Since the decision tree model does not have coefficients to explain the variables’ specific influence on the price, their importances were calculated instead. It turned out that logged square meter has the highest importance (0.5223), followed by “Կենտրոն” district (0.2058), type

“Մոնոլիտ” (0.0519) and district “Արաբկիր” (0.0461). Type “Կասետային” has the least importance with a value of 0.000069.

Descriptive Results

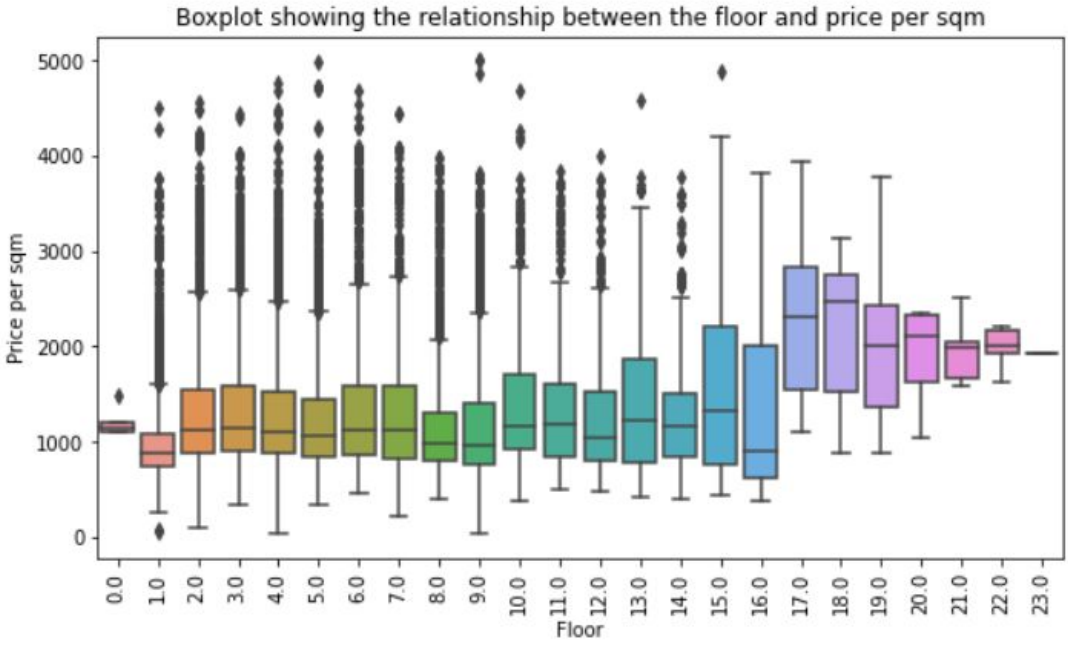
To analyze the price movement itself, chart analysis can be very helpful. The visualizations help to determine patterns and draw more accurate conclusions. Accordingly, in this section we will demonstrate some interesting findings from the charts.

1. The chart below is a boxplot describing the relationship between price per sqm in different districts of the dataset. As shown, the highest mean price is for “Կենտրոն” district. The second high-priced apartments based on sqm are found in “Արաբկիր”.

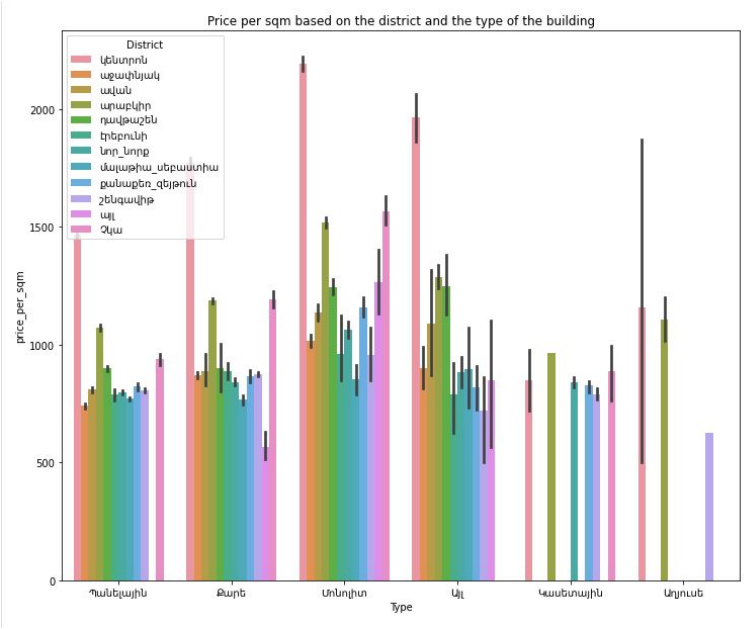


2. The second boxplot we would like to interpret shows the relationship between the floor and price per sqm. There is a change in the pattern after the 17th floor. We see that the mean price drastically increases after that. It is assumed that based on the previous chart analysis these

apartments can be newly constructed and from "Կենտրոն" district.

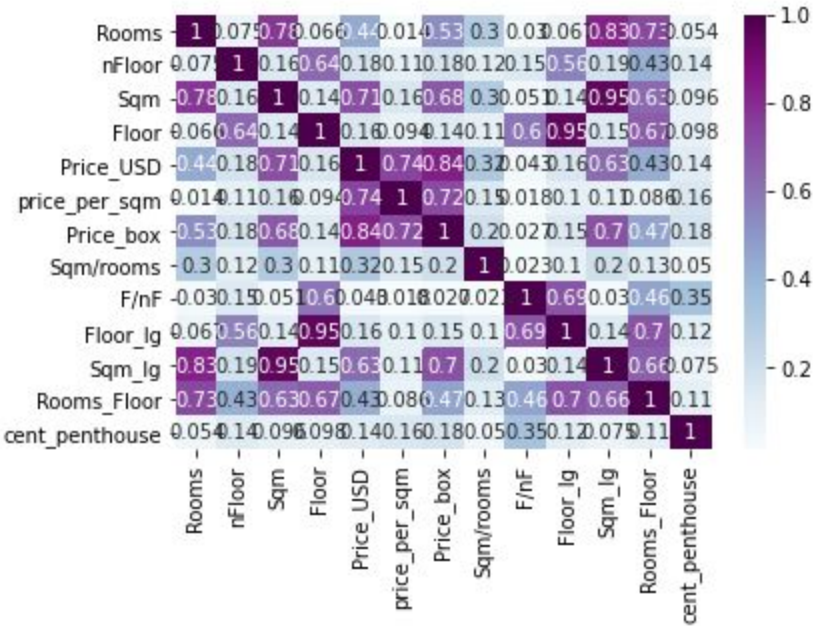


3. As we know, the type of the building is a crucial factor in house price prediction.



It was assumed that the type of the building would change the price drastically, however, we see that again in most of the cases the apartments from "Կենտրոն" have the highest price. The "Մոնոլիտ" type of houses have on average higher price per sqm regardless of the district. It may be explained with the fact that such buildings are considered to be more seismic resistant.

4.The last visualization chosen for interpretation in this paper shows the correlation between all of the factors of the dataset.



As expressed, there is a high correlation (0.83) between Sqm_lg and Rooms, which can be interpreted by the fact that the larger the area of the house the more rooms. A correlation of 0.71 is also found for Price_USD and Sqm, which can be explained by the fact that the area of the house plays a huge role in house price prediction. The heatmap encompasses newly constructed interaction terms as well. Afterwards we tried to avoid multicollinearity by not choosing two variables in the model that have higher than 70% correlation.

Review and Future Plans

As we assumed, the factors examined do influence the housing prices in the Armenian market. Some of them have larger influence, such as the size (square meter) or the location - houses in the center are far more expensive than the ones in suburbs. Also, the type of the building turns out to have a distinguishing impact on the price as well. Some descriptive findings were also interesting, connected to the higher house prices in the center, especially for the top floors.

Findings in this paper can help both buyers and sellers in the market to know what to expect and to understand where those price changes are mostly coming from. It can also be a

help for governmental decision makers when making construction plans in different parts of the city.

As of limitation of the study, which can be concluded from the literature review as well, the presence of some other variables might help to have a better model in terms of accuracy. Because when we think about it, there are many other factors such as the availability of a garage, elevator, windows, the renovation condition or the distance from transportation means could be a big influencer in the price of the houses in the market. For the future research, in case of data availability, things like those should be analysed as well. Moreover, we can expand the location and construct/apply the same model on the data with districts of different cities (e.g. Gyumri, Vagharshapat, Yeghegnadzor).

References

Jengei HONG, Heeyoul CHOI, & Woo-sung KIM. (2020). *A House Price Valuation Based on the Random Forest Approach: The Mass Appraisal of Residential Property in South Korea.* International Journal of Strategic Property Management, 24(3), 140–152. https://doi.org/10.3846/ijspm.2020.11544

Liu, X. (2013). *Spatial and Temporal Dependence in House Price Prediction.* Journal of Real Estate Finance & Economics, 47(2), 341–369. https://doi.org/10.1007/s11146-011-9359-3