**House Sales in King County, USA**

**A nonparametric approach**

**Introduction**

House prices in the US have experienced constant growth over the past years, therefore, it is becoming ever more important to take with care the decision to buy a house. Few people know what characteristics of the house are most highly valued in the market, hence the aim of this study is to explain how such features interplay in the final sale price. The solution consists of an analysis of the statistical significance of a range of variables and follows with the construction of a nonparametric model to predict house prices. Such a model is free from assumptions on the distribution of the features and it is statistically robust to outlying observations.

The work is carried out on a dataset[[1]](#footnote-1) containing the characteristics of 20000 houses, which were sold in the years 2014-2015 in King’s County, WA, United States. The region includes the city of Seattle and has known a stable growth both in the number of properties sold[[2]](#footnote-2) and on the median price of sale in the past 5 years, as witnessed by Figure 2.

Map

Description automatically generated

Figure 1. Heatmap of the houses around Seattle city centre.

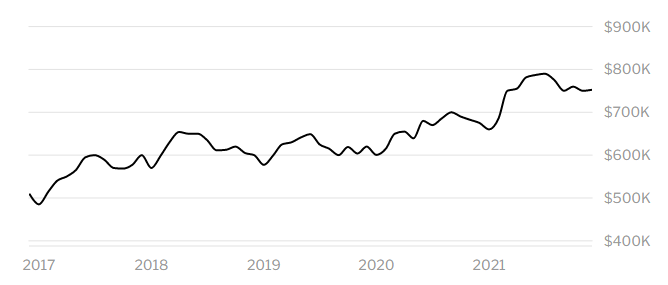


Figure 2. Median sale price in King's County

**Dataset**

The dataset we worked with contains, in principle, features of houses sold from May 2014 to May 2015 in the King County area, Washington State. There are, at the beginning, 21 variables:

* ID and location (zipcode, latitude and longitude)
* date and price of the sale
* number of floors, bedrooms and bathrooms
* size of the interior living space and size of the land lots in square feet
* average size of the interior living space and average size of the land lots for the closest 15 houses, in square feet
* size of the interior living space above and below the ground level, in square feet
* year of construction and year of renovation
* categorical variables about waterfront, view rating, current condition and construction grade

We, firstly, created some new variables of interest, such as bathfloors\_ratio (bathrooms/floors), bedfloors\_ratio (bedrooms/floors), geodist\_index (the distance of each house from the most southern point of Lake Union (DIRE PERCHE ABBIAMO PRESO PROPRIO QUESTO PUNTO??), ordinal date (a clear version of the date of the sale, to take into account the market evolution), has\_ren and has\_bas (if the house was renovated and has basement, respectively), is\_reach (a binary variable equal to 1 if a house is in a rich neighborhood, 0 otherwise. In particular, we took from a site the richest Zipcodes in King’s County and compared them to the houses in the dataset), FORSE NE MANCA QUALCUNA MA NON SAPEVO SE METTERLE TUTTE, and we modified some of the original ones, in particular bringing all the sizes in square meters.

**Preprocessing and Outlier Detection**

We explored each variable one at a time, creating a plot, a histogram and a boxplot, to visualize graphically its behavior. We decided to pass through the log10 transformation for those variables that presented a very unbalanced distribution, such as the price and the different sizes related to house.

After discarding all the variables that appeared completely useless and those that exhibited the same behavior as others, we selected six final variables for the outlier detection: log10(price), our response variable, w.r.t. bedrooms, bathrooms, log10(sqm\_living), log10(sqm\_lot), log10(sqm\_living15) and geodist-index.

We followed an approach based on the automatic baglot analysis, in particular, looking at the bagplots relating the response and the other variables mentioned above. This method is result in being very effective, since all the problematic points we found in the preliminary variable exploration have been spotted and the number of discarded points was kept at an acceptable level.

INSERIRE LE IMMAGINI DEI DUE BAGPLOT PIU SIGNIFICATIVI

Obviously, we then created a new dataset, which has been used from now on, removing all the outliers detecting with the analysis just described.

**Testing**

In this part of the report are collected and presented all the tests we performed to validate our beliefs. Adopting a non-parametric approach for testing was fundamental, as most of the times we weren’t in the position to rely on parametric ones.

A large part of our tests is related to price variability. In this section, indeed, we performed a lot of permutational ANOVA (we had non normality between groups) to evaluate the significance of some factors on the price, finding that the categorical variables waterfront, view, condition, grade and has\_bas, had significative effect on it.

Regarding the test for view significance, we did also sub-tests on the single classes belonging to this factor:

METTERE L’IMMAGINE DEL BOXPLOT DI LOGY W.R.T. VIEW

* class 1 vs class 2: effect was not significant at any reasonable level (p-value of 0.60)
* class 2 vs class 3: effect was significant with confidence greater than 99% (p-value 0)
* p-value 0 also for all the other tests with one class against another

Thus, apart from the price of the house with a medium-low (class 1) and medium (class 2) rating of the visit that is equal, for all the other classes the price is different.

We performed a permutational ANOVA test to verify an eventual influence of the variable geodist\_index on the price. Firstly, we transformed geodist\_index in a new variable distance, consisting on 3 factors chosen in this way: “short” if geodist\_index ∈ [0,15), “medium” if geodist\_index ∈ [15,30) and “long” if geodist\_index>30. The test returned a p-value=0, so we can argue that the distance has a significant effect on the price.

In a similar way, we divided the logarithm of sqm\_living in 3 groups (1 if sqm\_living<100, 2 if sqm\_living∈ [100,250) and 3 if sqm\_living>=100). Doing a permutational test again, we obtained a p-value=0, so we can assert that the sqm\_living has a great influence on the price.

These tests confirm that the price has a great variability.

We performed a permutational test to assess equality in distribution between the cluster of old built but renovated houses and the cluster of recently built but not renovated houses. norm of the multivariate mean was considered as test statistic. As expected, the hypothesis of the two samples coming from the same distribution was rejected at any significance level.

We wanted to compare sqm\_living and sqm\_living15. We computed the differences between the two variables, checking their distribution to be symmetric by visualization (we still working with the log10 transformation of these variables), and then perform a center of symmetry test. The squared Euclidean distance between the difference in means and the hypothesized value (zero) was considered as test statistic. The p-value of the test is 0.324 and thus I can argue, at any reasonable level, that the size of the interior living space is equal to the average size of the interior living space for the closest 15 houses, therefore it is easier to find a large (small respectively) house in a group of large (small respectively) houses than a large (small respectively) house in a group of small (large respectively) houses.

INFATTI, ABBIAMO TROVATO QUARTIERI RICCHI (DA ALTRI TEST CHE SONO QUELLI CON CASE PIU GRANDI E CHE COSTANO DI PIU) VICINI, NON CASE RICCHE SPARSE QUA E LA

We also did a test to verify if the richer neighborhoods correspond to those where the houses cost more. We divided the dataset in 2 groups based on the variable is\_rich, then we performed a permutational test between the prices in the 2 groups and we got a null p-value. We concluded that the price of houses in wealthy neighborhoods is much higher than that of houses in poor neighborhoods.

Another interesting test (looking at the previous results) is the variables view and condition are correlated in some way. Since both categorical variables consist of 5 factors, we performed a chi-square test of independence to determine if the two are related. We got a p-value= 1.79e-08, thus we can say that house condition and view have no correlation with a confidence level higher than 99%.

**Price Modelling**

The price model went through several design phases. At first, each of the variables is modelled individually against the price with non-parametric methods, such as B-splines, natural splines, step regression, piecewise linear regression. The univariate models’ regressors that look more promising are then joint into a multivariate model, initially a Generalized Additive Model that we replaced with a Robust Linear Regression based on MM-type Estimators, which had better performances.

Here’s a recap of the ad-hoc modelling we performed on the variables that were introduced in the nonparametric model: the ratio between the number of bathrooms and floors is modelled with B-spline with 2 degrees of freedom; the overall condition of the house, which ranges from 1 to 5 has two different steps with a discontinuity at 3; houses older than 80 years have a different linear model than those that are younger; we include an interaction term between the age of the house and the fact that it was renovated; the index of distance from the lake shore (“geodist”) is modelled with a degree 2 spline; the latitude is broken down into different linear models at different valies, which are uneven. The following variables are kept as they are, so they give a linear contribution: the view score, the grade of the property, whether a house is in a rich neighborhood, the zipcode, the size of the upper floor, living space, lot and average living space of the neighbors.

The robust model obtained is compared to both a linear regression model that uses the same variables and a state-of-the-art regression method such as XGBoost [1]. The results of each of the models is reported below, in Table 1.

The metrics we used to compare the model are MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error) on the price in dollars. As expected, our model outperforms a basic linear regression but is not as good as XGBoost.

The diagnostics on the results of the regression reveal that the model doesn’t perform well in neither very expensive houses nor very cheap houses. As a consequence, we develop a further nonparametric model in the same fashion as the one presented above, but aimed at modelling only the 1218 most expensive houses. We defined the houses to be expensive when their sale price is over one million dollars, as inferred from the data on Figure 3. The model is described in Appendix.

The results in Table 1 show a clear improvement on the predictions thanks to separate modelling of expensive houses.

|  |  |  |
| --- | --- | --- |
| Method | MAPE (%) | MAE ($) |
| Nonparametric Regression | 16.12 | 79728 |
| Linear Regression | 17.42 | 87147 |
| XGBoost | 11.65 | 58455 |
| NonParametric Regression (standard+expensive) | 15.18 | 73391 |

Table 1.

Chart, scatter chart

Description automatically generated

Figure 3. Bins from 100k to 2M dollars. The red line marks the boundary for expensive houses (greater than 1M dollars).

**Conclusion – Future steps.**

**References**

[1] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: ACM. <https://doi.org/10.1145/2939672.2939785>

**Appendix**

Model for expensive houses

The regressors were modelled as follows: the ratio between number of bathrooms and floors and the average lot size of the neighboors is a natural spline with 2 degrees of freedom; geodist\_index, the living surface and the lot surface with a degree 2 spline; view score, grade, latitude, zipcode and size of the upper floor are kept as linear predictors. The model is a Robust Regression with MM-type estimators.

1. https://www.kaggle.com/harlfoxem/housesalesprediction [↑](#footnote-ref-1)
2. https://www.redfin.com/county/118/WA/King-County/housing-market [↑](#footnote-ref-2)