House Sales in King County, WA
A nonparametric approach

Enrico Sartor, Jahiro Mugheddu, Luca Panzeri, Nick Usubelli.

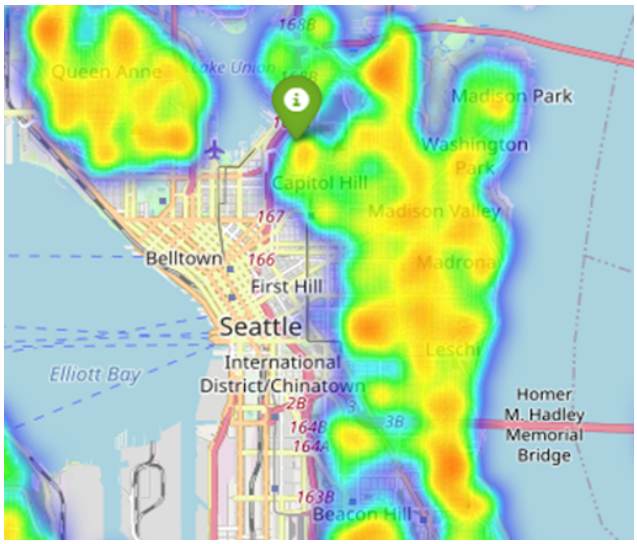


Figure 1: Heatmap of the houses around Seattle city centre.

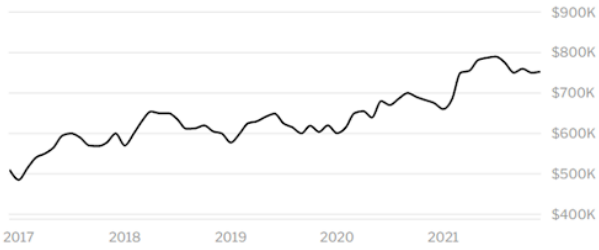


Figure 2: Median sale price in King County.

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 containing the characteristics of over 21000 houses, which were sold in the years 2014-2015 in King's County, WA, United States.[1] The region includes the city of Seattle and has known a stable growth both in the number of properties sold[2] and on the median price of sale in the past 5 years, as witnessed by Figure 2.

Dataset

The dataset we worked with contains 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;

[1] https://www.kaggle.com/harlfoxem/housesalesprediction
[2] https://www.redfin.com/county/118/WA/King-County/housing-market

- 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 iconic Space Needle, located in the city center), `ord_date` (an ordinal 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_rich` (a binary variable equal to 1 if a house is in a wealthy[3] neighborhood, 0 otherwise), and we modified some of the original ones, in particular bringing all the sizes in square meters.

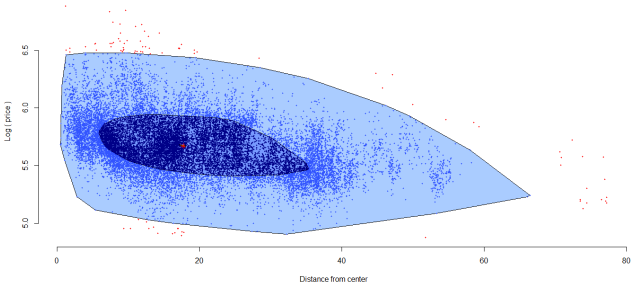
Preprocessing and Outlier Detection

We explored each variable one at a time, creating a plot, a histogram and a boxplot, to visualize graphically its behavior. We decided to make a $\log_{10}$-transformation for those variables that presented a very unbalanced distribution, such as the price and the different sizes related to house.

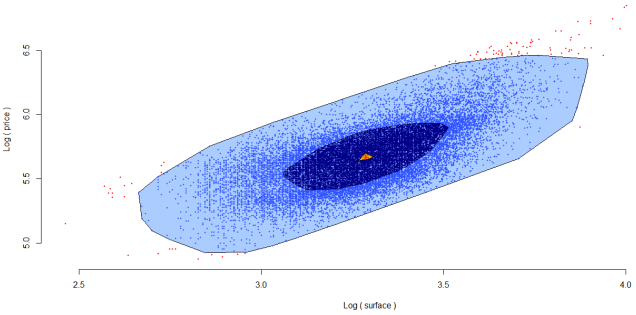
After discarding all the variables that appeared useless and those that were most highly correlated to others, we selected six final variables for the outlier detection: $\log_{10}$ `price`, our response variable, w.r.t. `bedrooms`, `bathrooms`, $\log_{10}$ `sqm_living`, $\log_{10}$ `sqm_lot`, $\log_{10}$ `sqm_living15` and `geodist_index`.

We followed an approach based on depth measures, in particular, plotting the bagplots relating the response and the other variables

[3]https://www.zipdatamaps.com/economics/income/agi/metro/wealthiest-zipcodes-in-metro-seattle-tacoma



(a) Distance from the center vs Price



(b) Living surface vs Price

Figure 3: A couple of helper boxplots for the outlier detection

mentioned above (Figure 3). This method resulted to be 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.

Testing

This part of the report collects all the tests we performed to validate our hypotheses. It was fundamental to adopt a non-parametric approach for testing, as most of the times we couldn't rely on the parametric ones, due to the non-Gaussianity of the data. All the tests use a significance level alpha equal to 0.05.

A large part of our tests is related to price variability. Here are the tests we performed:

- evaluation of the significance of the following factors on the `price`, via permutational ANOVA: `waterfront`, `view`, `condition`, `grade` and `has_bas`. They all had a significative effect on it;

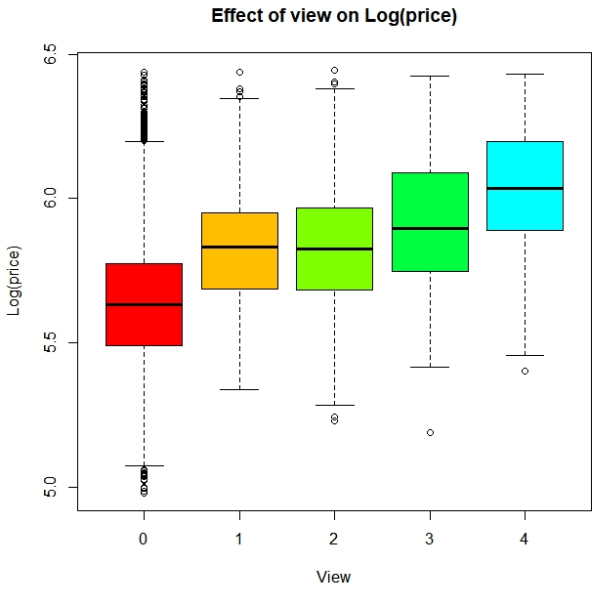


Figure 4: Log(price) grouped by "view"

- influence of the variable `geodist_index` on the `price`, via permutational ANOVA. Firstly, we binned it into 3 factors chosen in this way: "short" if `geodist_index` $\in$ [0,15), "medium" if `geodist_index` $\in$ [15,30) and "long" if `geodist_index` $>$ 30. The test returned a p-value = 0, so we can argue that the distance from the Space Needle has a significant effect on the price;

- effect of living surface on the price. We divided the logarithm of living surface (`sqm_living`) in 3 groups (1 if `sqm_living` $<$ 100, 2 if `sqm_living` $\in$ [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 living surface has a great influence on the price;

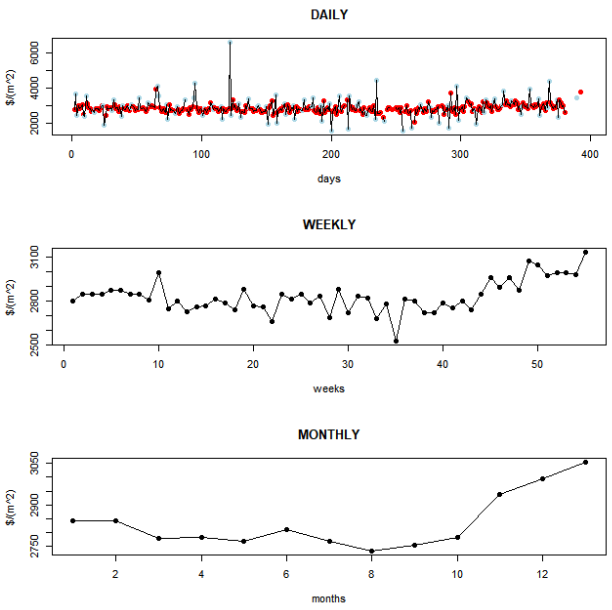
- equality in distribution between old but renovated houses and recently built houses via permutational test. We used the L^2 norm of the multivariate mean as test statistic. As expected, the hypothesis of the two samples coming from the same distribution was rejected at the given significance level;

- location of expensive houses with respect to popularly known wealthy neighborhoods. We divided the dataset in two groups based on the variable `is_rich`, then we performed a permutational test between the prices in the groups and we got a null p-value. We concluded that the price of the houses in wealthy neighborhoods is considerably higher than that of houses in poor neighborhoods;

- correlation between `view` and `condition`. 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 are correlated.

Analysis on the price per square meter

Whenever buying a house, the first thing that comes to mind for a rough estimation of the cost of the investment is the price per square meter. For this reason, we studied how the price changed over space and time in King's County. At first glance, the prices per square meter presented a high daily variability, hence we decided to consider them on a weekly and monthly basis. The available time span was from May 2014 to May 2015, where the last month was excluded due to the poor data availability. In some zipcodes, the number of sales was quite low, so we grouped multiple neighboring zipcodes with a K-means algorithm into 4 clusters, to minimize the within sum of squares (Figure 5). We noticed that the median price per square meter changes in the same fashion for each cluster, that is: lower prices in fall and higher prices in the springtime. Moreover, it seems that the cheapest area in the whole county is the South-Eastern, while the North-Western is the most expensive.



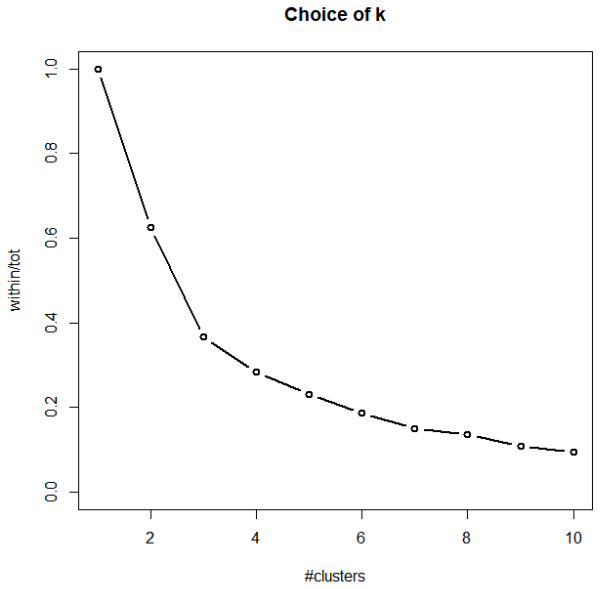


Figure 5: Price per square meter variation (top) and knee-elbow analysis for the clustering (bottom).

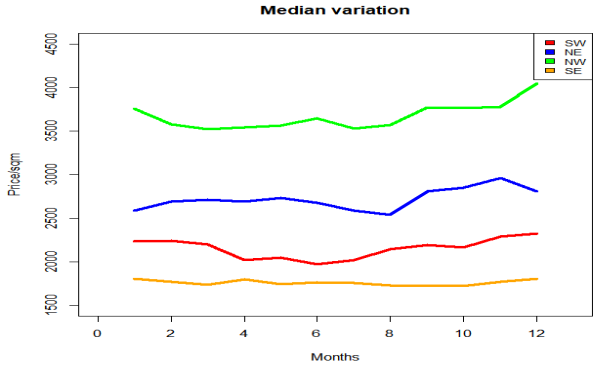


Figure 6: Median price per square meter for each of the clusters

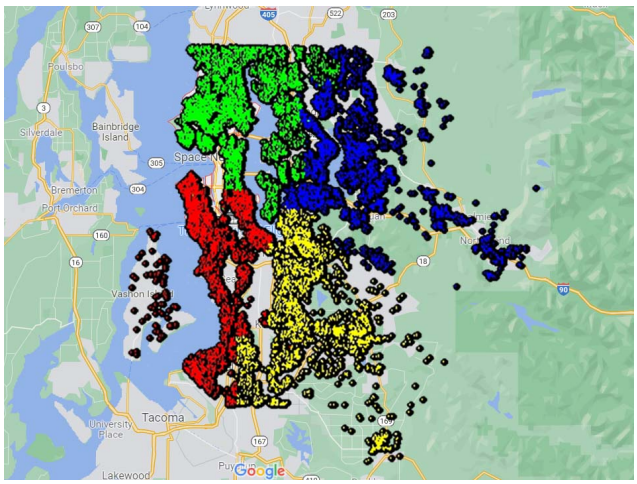


Figure 7: Macroregion: NW, NE, SW, SE

Price Modelling

Despite the findings of the previous point, we decided to take the simplifying assumption of time-invariant distribution for the price. Otherwise, we would have had to model the prices with time-series methods.

The price model went through several design phases. At first, each of the variables is modelled individually against the price with nonparametric 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 distance from the Space Needle (geodist-index) is modelled with a degree 2 spline; the latitude is broken down into different linear models at different values, which are uneven. The following variables are

Method	MAPE(%)	MAE($)
Linear Regression	17.84	89351
Nonparametric Robust Regression	16.05	80897
Nonparametric Robust Regression (standard + expensive)	15.56	75436
XGBoost	11.65	58455

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

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 size of the upper floor, living space 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, 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. The models were all trained on the same 80% split of the data and trained on the remaining 20%. As expected, our model outperforms a basic linear regression but is not as good as XGBoost.

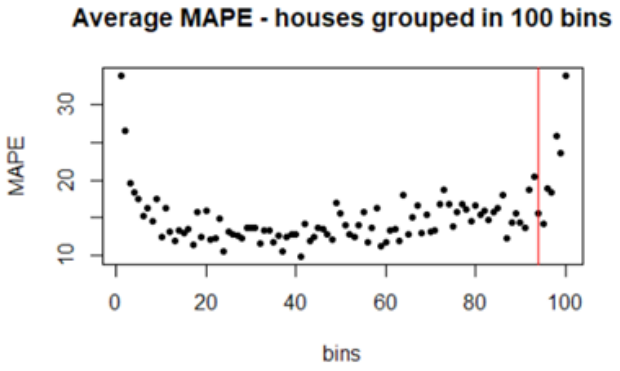


Figure 8: Bins from 100k to 2M dollars. The red line marks the boundary for expensive houses (greater than 1M dollars).

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 8. The model is described in Appendix. Despite the custom modelling, the variability in the expensive houses is difficult to model (the R^2 is equal to 0.44). Hence, for people who want to buy expensive houses, our study cannot be of much help.

Conclusion

Our study highlights which of the features influence most the house prices and what underlying nonlinear relationships link them to the prices. It proves that the qualitative variables in the dataset, such as the grade, condition and view score, are significant in determining the price of the house. So, we can conclude that the first impression plays its role on the sale price. Our tests show that having a basement matters, as well as facing the water or being close to the city main attraction. It is also important to renovate the house as this will increase significantly its price. It is also clear that houses located in wealthy neighborhoods are necessarily expensive.

This work confirms the relevance of nonparametric and robust techniques for modelling house prices. In particular, it brings to light the inadequacy of a simple regression model for such a complex problem and suggests the use of more advanced models to get more reliable predictions. However, expensive properties proved to be difficult to model, so this could be a limit of our study.

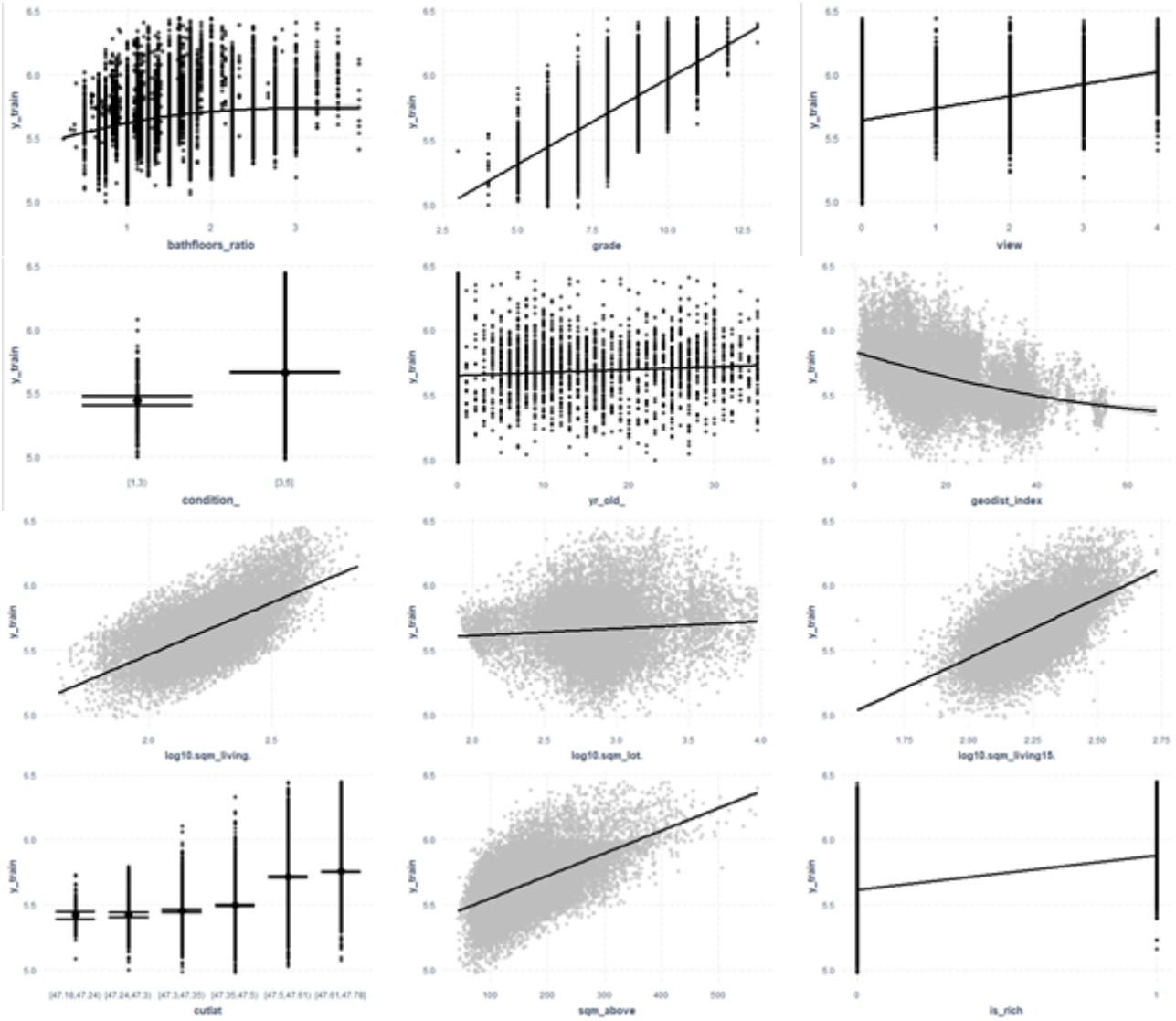


Figure 9: Here are the plots of how each variable was modelled individually.

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, size of the upper floor are kept as linear predictors. The model is a Robust Regression with MM-type estimators.