Predicting house prices in King County

**1 Introduction**

The price of an apartment is affected by several factors and the purchase decision is rarely easy. In this project, we try to facilitate this process and aim to predict the house prices for houses with different features in King County.

We aim to make predictions about the price based on some simple parameters. Our model should learn from the data and be able to predict the price of the house with a new combination of features. The predictions could be done using different models such as linear regression and polynomial regression.

The structure of this report is as follows: the machine learning problem is explained in more detail in the problem formulation section. In the methods section, the feature selection and used machine learning models are described. More detailed information on data processing is also provided in the methods section. In the result section, the results are compared and there is a summary of the results and the report in the conclusion section.

**2 Problem formulation**

The data points are houses in King County (USA). There are 16 different features. There is a lot of basic information such as the number of bedrooms, floors, and bathrooms. 0.5 bathrooms means that there is a toilet in the room without a shower. Building year, possible last renovation, and zip code are also included. There is also information about square footage of land and interior living space. Interior living space is divided into above square footage and basement square footage. Above square footage means the square footage above ground level and basement square footage the opposite. Waterfront is described with a dummy variable for whether there is a water view or not. Other features of the apartment are described on different scales. The view is described with an index from 1 to 4 of how good the view is. Condition is an index from 1 to 5 and grade of the apartment from 1 to 13. Grade 1-3 has poor construction and design, 7 is average and from 11 to 13 is high-level quality. There is also included the average square footage of interior and land of the 15 nearest neighbours.

**2.1 Summary of the problem**

Kuva, joka sisältää kohteen pöytä

Kuvaus luotu automaattisesti**Label**: the price of the house. **Features**: number of bedrooms, number of bathrooms, square footage of the apartments interior living space, square footage of the land space, number of floors, whether there is a water view or not, view(0-4), condition(1-5), grade(1-13), the square footage of interior living space that is above ground level, the square footage of interior living space that is below ground level, building year, year of the house’s last renovation, zip code, the square footage of interior living space that for the nearest 15 neighbours and the square footage of land of the nearest 15 neighbours. The data types are shown in figure 1.

Figure 1. Datatypes of the features.

**3 Methods**

**3.1 Dataset**

For the project, I found the data from the Kaggle [1]. The dataset includes data for 21 614 apartments in different locations in King County, so we have enough data to solve this problem with machine learning methods.

**3.2 Feature selection**

We use price as a label and 21 other columns could be used as a feature. To get an idea of the correlation of different properties and to find out which features have high correlations with the price, we are using a correlation matrix that shows correlations between different variables. The correlations are shown in figure 2.

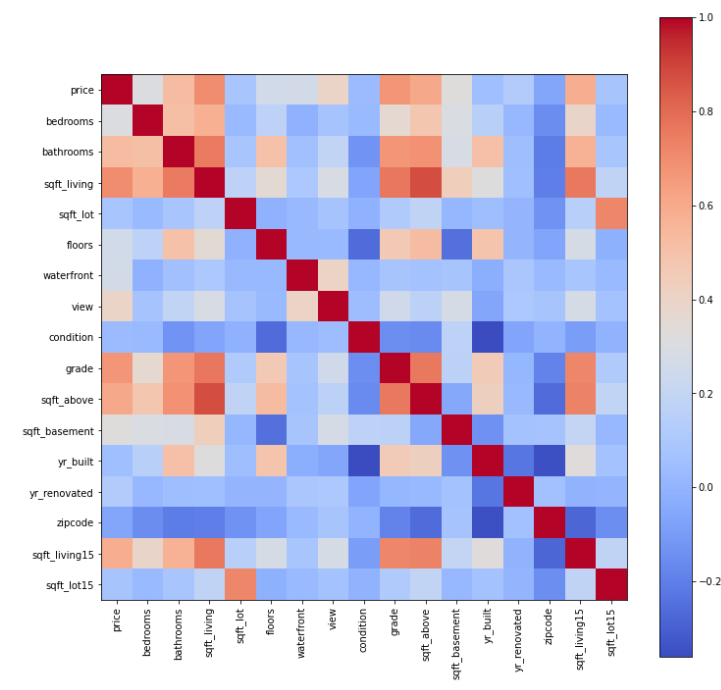


Figure 2. Correlation heatmap.

Based on the correlation matrix, the price has a high positive correlation with number of bathrooms, square feet of living, grade, the square footage of interior living space that is above ground level and, the square footage of interior living space for the nearest 15 neighbour. Price has a low correlation with the number of bedrooms, floors, waterfront, and sqft basement. We take a closer look at correlations between price and square foot living, price and grade and price and number of bedrooms in figure 3. Based on figure 3, we make the same conclusion about the correlations as from the correlation matrix.

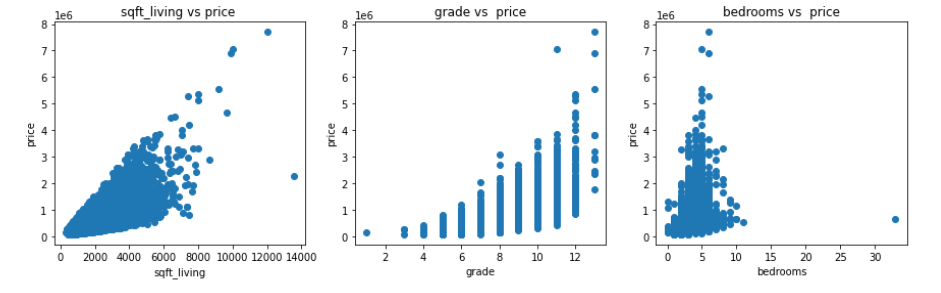


Figure 3. Correlations between different features with price.

Some features are dropped because they are not relevant to the problem. The id, date of sale, latitude, and longitude was removed from the dataset because we do not want to study their effect on price. If we wanted to include the date, we should have data from the longer term. We also get a good understanding of the apartment’s location from the postal code and thus we do not need the exact location of the apartment. There were no missing features or labels in the dataset. All the features were scaled between 0 and 1 because it makes it easier to compare different features. After comparing the correlations in the correlation matrix and deleting the extra features we are left with 16 features.

**3.3. Construction of training and validation sets**

Datapoints were split into training (80%), validation (10%), and test (10%) sets. The validation set is a separate section of the dataset that is used to evaluate the model trained on the training set to get a little sense of how the model works. In this case for example we can compare the success of the training process with different training sizes with the validation set. We can evaluate the final model performance of the chosen model with the test set. We want to keep the training set as big as possible because the more samples we have in the training set the better opportunity the algorithm has to understand the dependencies of the features.

**3.4 Linear regression model**

The first model I used was the linear regression model. We try first a linear model because if thinking with common sense, there could be a relatively strong correlation between these different features and price.

In the linear regression model, we assume that there is a linear relationship between continuous variable y and one or more independent variables X. The idea is to find the best fit linear line and coefficients so that the error is as small as possible. The model takes the form:

where y is the predicted label, x’s are the values of different labels, is a constant term and ‘s are weights. We use linear regression to search optimal weights and from a linear hypothesis space. As a loss function, we use mean squared error. We try to minimize the error by trying different weights . Mean squared error is a convenient way to determine how “good” a model is. Also, according to the literature, it is smart to use mean squared error with linear regression [2]. Mean squared error is defined as:

**3.5 Polynomial regression**

Polynomial regression is the next model to be used. The problem may be too complex for the linear model so we could assume that polynomial regression performs better. [4]

The idea is to reuse the algorithm for linear regression. Polynomial regression is a form of linear regression where due to the non-linear relationship, we add polynomial terms to the model to convert it to polynomial regression. [5] We fit the data to polynomial regression models with different degrees. The model takes the form:

Mean squared error is used also with this model because it gives a good picture of the performance of the model.

**4 Results**

Errors and inaccuracies of linear regression are shown in figure 4.

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti

Figure 4. The result from linear regression.

From the results, we can observe that we can analyse and predict housing market prices relatively well with linear regression. However, we have to admit that there is room for improvement in accuracy.

The results from polynomial regression are shown in figure 5.

Kuva, joka sisältää kohteen pöytä

Kuvaus luotu automaattisesti

Figure 5. The results from polynomial regression.

From the polynomial regression results, we notice that the model starts dramatically to overfit after when poly degree is 3 but there is a little overfitting already when poly degree is 3.

Errors and inaccuracies of polynomial regression with poly degree 2 are shown in figure 6.

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti

Figure 6. Results from polynomial regression with polynomial 2.

**5 Conclusion**

If thinking about what we want from the model, there are a couple of things we value. The model should maintain its performance when applied to new data. The model should also be parsimonious and as simple as possible. [6] We have now tried two models with different properties. Accuracy from polynomial regression was better compared to linear regression. Also, the mean squared errors are smaller in polynomial regression. Thus, polynomial regression with polynomial 2 was the model that performed the best and that is the model we choose to use.

Many things could be developed further in this project. It might be that very high house prices distort the operation of the model and we get some weird results. Because of that, the model could only use apartments of a certain price. Also, it would be interesting to study the effect of distance to schools, kindergartens, and shops on house prices.

**6 Appendices**

The code can be found from Git:

<https://github.com/Valdde/Machine-Learning.git>

**References**

[1] <https://www.kaggle.com/harlfoxem/housesalesprediction>

[2] <https://vitalflux.com/mean-square-error-r-squared-which-one-to-use/>

[3] <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html>

[4] https://www.analyticsvidhya.com/blog/2021/07/all-you-need-to-know-about-polynomial-regression/

[5] <https://www.slideshare.net/PawanShivhare1/predicting-king-county-house-prices>

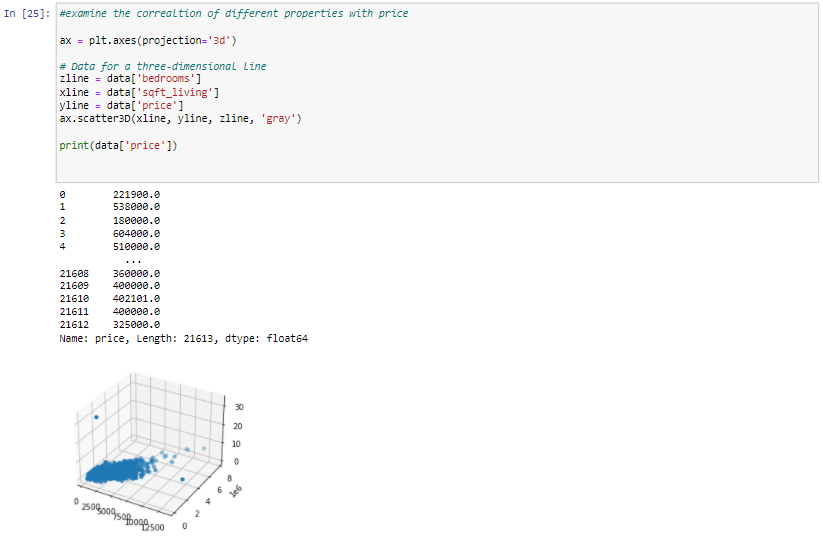
[6] <https://jofalu.github.io/>

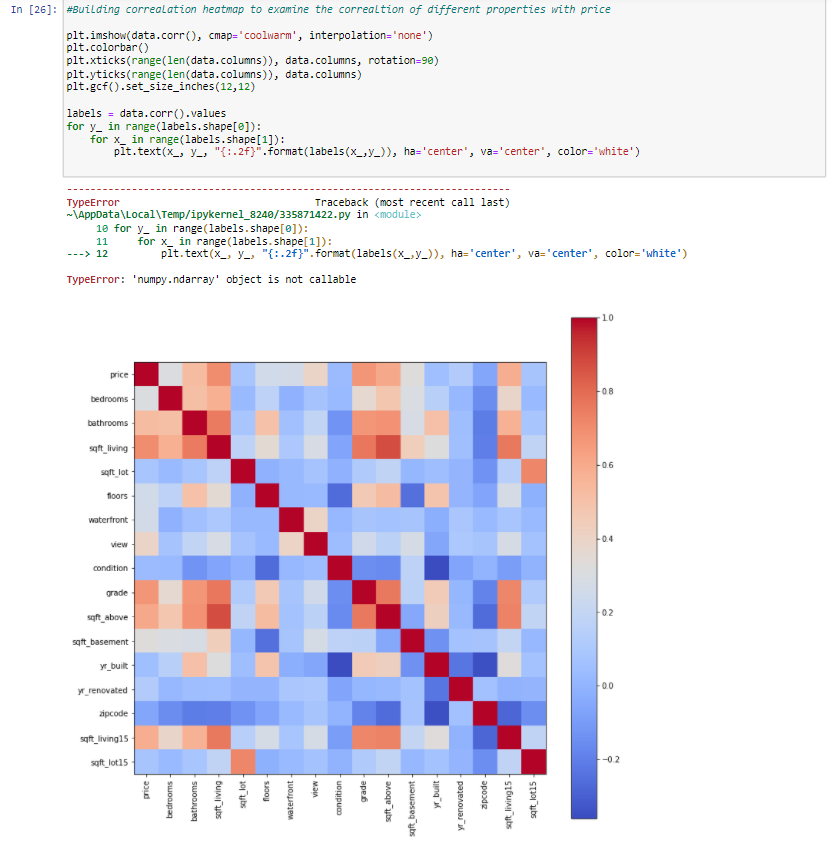
Kuva, joka sisältää kohteen teksti

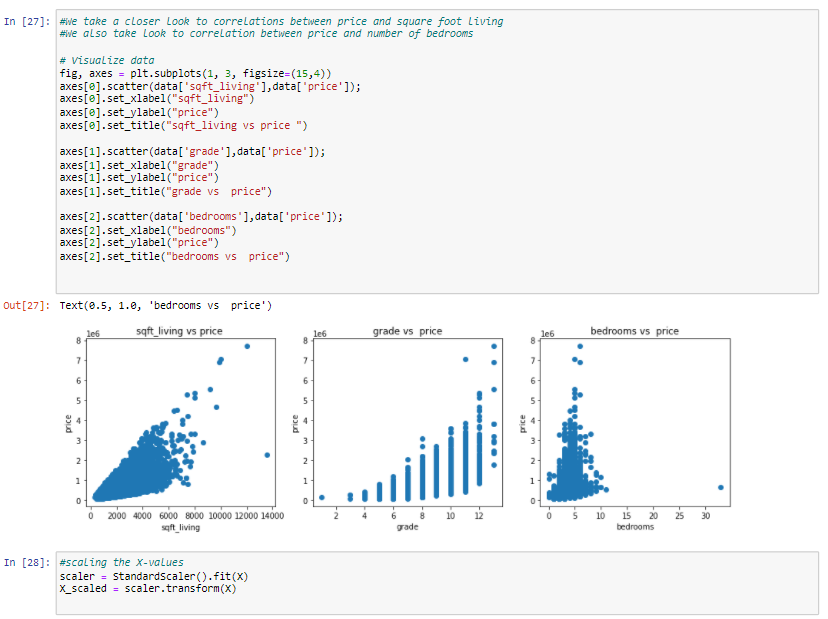
Kuvaus luotu automaattisesti

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti







Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti