Assignment 2 - Group 95

188.429 Business Intelligence

Luka Benzia
11778695
e11778695@student.tuwien.ac.at
Person B

Christian Leis
12310515
e12310515@student.tuwien.ac.a
Person A

1 Business Understanding

1.1 Definition and description of the data source

The data comes from Kaggle and contains recorded house sales along with information about the properties, from May 2014 to May 2015, in King County

(https://www.kaggle.com/datasets/harlfoxem/housesalesprediction).

This data is needed by a real estate company that rents out high-priced properties. The goal is to develop a model that, based on property characteristics such as location, features, size, etc., identifies a potential investment, independent of the listed sales price. Properties are categorized into four categories ('low', 'medium', 'high', 'luxury'), with only the last, high-priced category being relevant. The scope is purely the selection of properties; price evaluation and thus the final decision on whether to attempt to purchase the property does not fall within the scope of the model.

1.2 Business Objectives

The overarching goal from a business perspective is to maximize the company's profit in the long term by selecting only high-priced properties in King County for the portfolio that generate a high rental yield in relation to the purchase price and whose value ideally increases above average. A key advantage of the model is that properties are evaluated independently of price, apart from training the model, which means that even (heavily) undervalued properties are identified as luxury objects. These would not fall into the 'luxury' segment with a purely price-based categorization and would therefore not be considered further. Market prices also change dynamically, which would mean additional continuous effort with a purely price-based approach. In addition, the efficient identification of suitable properties optimizes the purchase process, which saves time and costs and ensures that only relevant properties are subsequently examined in detail. This targeted selection also supports the company's reputation strategy of being an expert in the luxury segment.

One limitation is that 'luxury properties' are not defined completely independently of price, because historical prices are divided into four equally sized categories and the model is trained on this basis. In addition, luxury properties often have unique features that may not appear in the data, such as a special construction method. Accordingly, there is a risk of reducing a luxury property to a few characteristics and thus classifying an actually relevant property into one of the other three categories.

The competing goal could be to avoid excessive specialization in the luxury real estate market, which could make the company vulnerable to market fluctuations and limit its long-term growth potential. This might involve diversifying the portfolio to include a wider range of property types and price points, ensuring resilience in the face of economic downturns or shifts in market preferences.

Factors and additional possible limitations that can influence the output are the quality, quantity and possible data bias (sampling bias, e.g. no luxury properties were considered) of the training data, the very limited time period of the data (2012-14), the limitation to 20 features and the definition of luxury properties (characteristics of a property that resemble the fourth quartile of the data in terms of selling price).

Related businesses are renovation companies, asset management companies and brokers who are also looking for similar luxury properties.

1.3 Business Success Criteria

A measure of the business success of the model is a growing ratio between investment size and rental income, which is an indicator of profitable property selection. Because the pure ratio can be misleading, for example, if only one property is invested in, which is extremely profitable, but cannot cover the company's costs because the volume of investments is lacking, the number of investments should also increase.

Another measure of success is employee productivity. It is successful if the time invested per property purchase is reduced in the entire process because now only selected properties are examined more closely.

1.4 Data Mining Goals

The goal of the model is to classify real estate into four categories based on features such as location, size, etc., with the highest category defining the luxury segment. The model is trained using supervised learning, where the target feature price is divided into four categories of roughly equal size, which should be highlighted. The goal is to have high precision and recall in the luxury segment to ensure that potential luxury properties are reliably identified.

1.5 Data Mining Success Criteria

The F1 score for the luxury category is used as a criterion for the success of the model. The advantage of this measure is that precision and recall are equally weighted in one metric. The threshold for success is 80% because this ensures that luxury properties are reliably identified. This is evaluated by a separate independent body within the company, through various tests such as cross-validation.

2 Business Understanding

2.1 Attributes types

Feature	Scala	Semantics
id	Nominal	Unique identifier for each property
date	Ordinal	Timestamp of the transaction (YYYYMM)
price	Ratio	Sale price (USD)
bathrooms	Ratio	Number of bathrooms
sqft_living	Ratio	Total living area in square feet
sqft_lot	Ratio	Size of the lot in square feet
floors	Ordinal	Number of floors
waterfront	Nominal	Whether the property has a waterfront
view	Ordinal	Rating of the view (0 - 4)
conditions	Ordinal	Rating of the condition (1 - 5)
grade	Ordinal	Rating based on construction and design (1 - 13)
sqft_ above	Ratio	Living area above ground level in square feet
sqft_ basement	Ratio	Basement area in square feet
yr_built	Interval	Building year
yr_renovated	Interval	Renovated year
zipcode	Nominal	Zipcode
lat	Interval	Latitude
long	Interval	Longitude
sqft_living15	Ratio	Average living area in square feet of the 15 nearby properties
sqft_lot15	Ratio	Average lot size of the 15 nearby properties

Table 1: **Attributes types**

The criteria for the ranking of view, conditions and grade are not specified and should therefore be viewed with caution.

2.2 Statistical properties and plausibility

The data set contains a total of 21,613 transactions.

Feature	Mean	Std	Median	Range
price	540.088	36.7127	450.000	72.000 - 770.0000
bedrooms	3	1	3	0 - 33
bathrooms	2	1	2	0 - 8
sqft_living	2.080	918	1.910	290 - 13.540
sqft_lot	15.107	41.421	7.618	520 - 1.651.359
floors	1	1	2	1 - 4
waterfront	0	0	0	0 - 1
view	0	1	0	0 - 4
condition	3	1	3	1 - 5
grade	8	1	7	1 - 13
sqft_above	1.788	828	1.560	290 - 9.410
sqft_basement	292	443	0	0 - 4.820
yr_built	1971	29	1975	1900 - 2015
yr_renovated	84	402	0	0 - 2015
sqft_living15	1.987	685	1.840	399 - 6.210
sqft_lot15	12.768	27.304	7.620	651 - 871.200

Table 2: **Statistical properties**

Firstly, it is noticeable that yr_renovated has a value of 0 if the property has most likely never been renovated. This should be kept in mind for the later preparation of the data. It is also noticeable that there is a large variance in the type of property, for example the range for lot 520 - 1.651.359 and bedrooms 0 - 33. It is possible that not only private purchases are shown here but also business purchases, e.g. those of a hotel.

2.3 Distribution

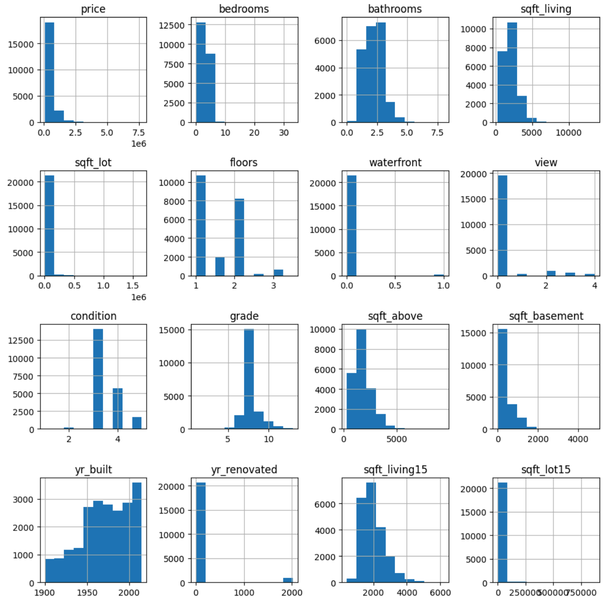

Figure 1: **Distribution of the data**

It is noticeable that there are many right-skewed distribution (Bedrooms, bathrooms, sqfr_living, sqft_lot, waterfront, view, sqft_basement, sqft_above, sqft_living15, sqft_lot15), which also makes sense if you look at the distribution of price. There is a similar distribution there, which indicates that price strongly influences the above-mentioned features. Thus, there are many cheap properties and few in the middle or upper price segment. It is also noticeable that the older the property, the more of them there are.

2.4 Correlations

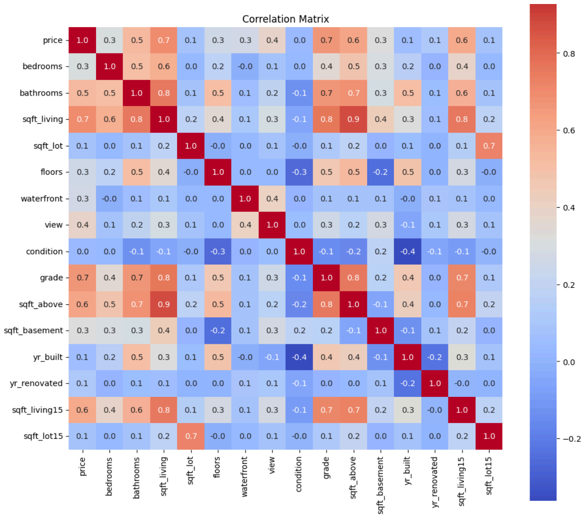

Figure 2: **Correlations Matrix**

If you look at price, the highest correlation is 0.7 with sqlt_lot and grade. Also strong with a value of 0.6 is sqft_above and sqft_living15, indicating that price depends mainly on the size of the property. Surprisingly, price correlates to 0.0 with conditions, 0.1 with yr_built, and 0.1 with sqft_lot. Also interesting, the grade and sqft_lot is 0.7, which emphasizes how important lot size is for the grade. Surprisingly, the number of bedrooms is not so important for price.

2.5 Dependencies

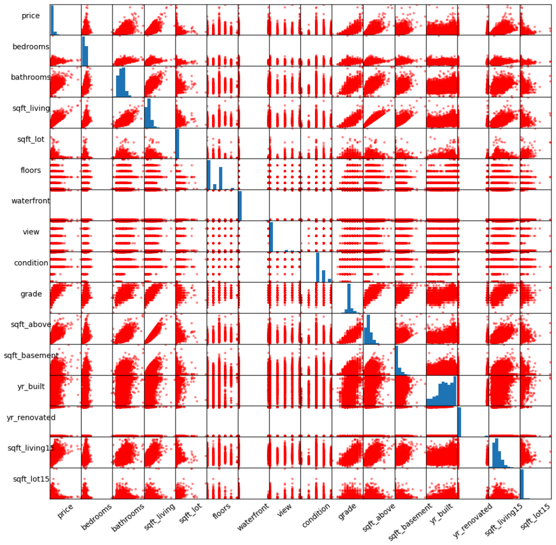

Figure 3: **Dependencies of all features**

The higher x the higher the price applies to the x: bathrooms, sqft_living, grade, sqft_above, sqft_basement, sqft_living15 and lot15, which was to be expected. What is surprising, however, is that Built year is relatively evenly distributed in terms of price, but with outliers on both sides; more outliers for newer properties Also interesting with regard to price is that, as already noted above, Bedrooms are not so relevant for price. The highest price is in the middle of the distribution, at around 3-4, after which the price tends to decrease. For floors, view, and condition rating, the price is relatively equally graded across the ratings, with condition being most positively related to a higher price.

We will now take a closer look at the connection between bedrooms and price.

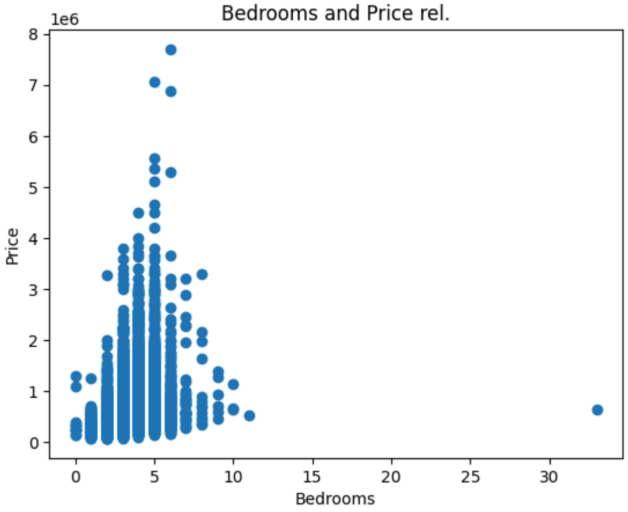

Figure 4: **Relation bedrooms and price**

It can be seen that we have one or a small group of outliers, with a low price for disproportionately large beds (33). These outliers should be treated accordingly in the data preparation to avoid distortions.

Now let's look again at the relationship between Year renovated and price, but excluding 0 values.

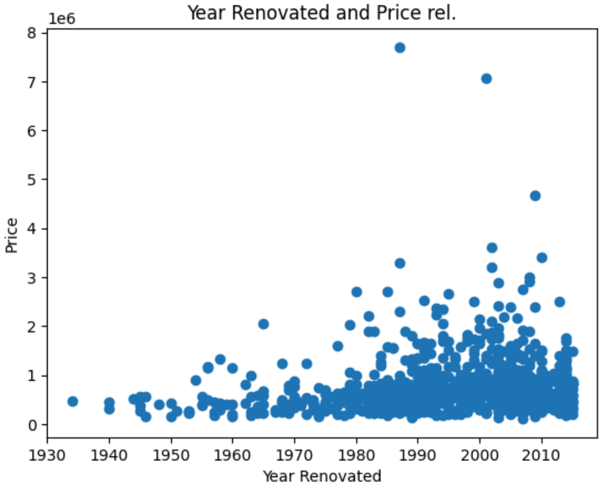

Figure 5: **Relation yr_renovated and price**

There is a clear relationship, namely that the later the property was renovated, the higher the price.

2.6 Spatial analysis

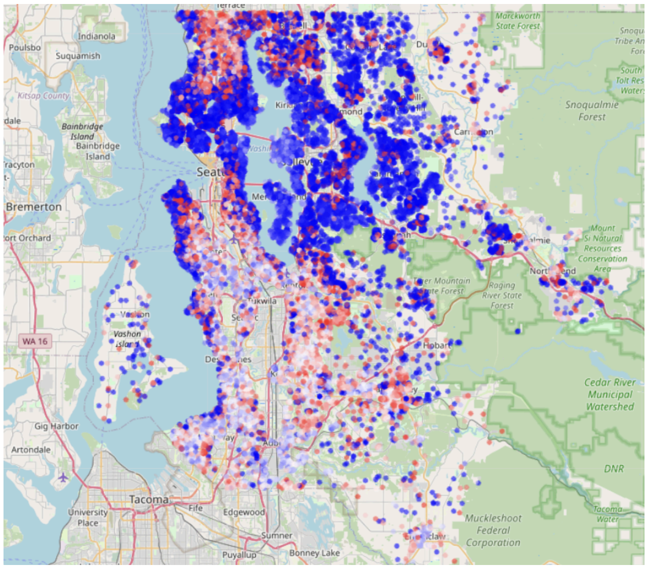

Figure 6: **Spatial graph**

The price is shown in the color spectrum from blue (cheap) to blue (expensive), with the middle spectrum corresponding to the median. It can be seen that more expensive properties tended to be sold in the north and south, while cheaper properties tended to be sold on Bainbridge Island.

2.7 Missing Values, data provenance and ethical issues

There are no missing values in the data and the data provenance is missing. There are also no ethical restrictions.

2.8 Bias in the data

Population bias and self-selection bias: Due to the unclear way in which the data was collected, the data set may not represent the actual market.

Measurement Bias: It remains unclear how the assessments of certain characteristics such as condition, size or viewpoint were arrived at, as these subjective assessments may not have been standardized or objectively recorded.

Omitted variable bias: Important variables such as the interior fittings of the apartments, for example the quality of the kitchen, the floor or the apartment renovation, may not be represented in the data.

Data bias and non-normality: There could also be a data bias, as non-normality is often observed in the data.

Questions for an expert: How was the data collected and does it represent the entire market or just a specific part? Are the subjective assessments such as condition and outlook based on standardized criteria or do they vary depending on the assessor? Are there missing features that could represent important features such as interior design or renovation quality?

2.9 Hypotheses

We hypothesize that Price is significantly influenced by bathrooms, sqft_living, sqft_above, sqft_basement, sqft_living15, and lot15; however, bedrooms play a subordinate role, because luxury does not mean an accumulation of bedrooms, but rather the size of the rooms. These properties tend to be located in the south or north-west of the city.

2.10 Actions in data preparation

The following steps are taken in the data preparation: 1. The numerical values of the target variable are divided into 4 categories. 2. Outliers in the data are processed. 3. The variable yr_renovated is transformed so that the model can better understand the data. 4. An explanation of what is done with highly correlated features. 5 . Finally, we explain how to handle the non-normal distribution of the data and how to scale.

3 Data Preparation

3.1 Necessary Preprocessing actions

Based on the findings in the previous part, the business understanding, we performed some necessary preprocessing actions on the data:

3.1.1 Encoding the target feature

We started by encoding the 'price' column into four same size quartiles. This created a new ordinal categorical feature, y_quartile, which serves as our target feature for the classification task. The quartiles were labeled as 'tier 1', 'tier 2', 'tier 3', and 'tier 4'. This strategy makes sense because, relatively speaking, we want to serve the upper price segment. It is also advantageous for the training model to have an equal distribution.

3.1.2 Handling outliers

Observing the presence of outliers in the 'bedrooms' feature, we decided to remove data points exceeding 30 bedrooms, because otherwise it would distort the model. The new distribution can be seen in the picture 7.

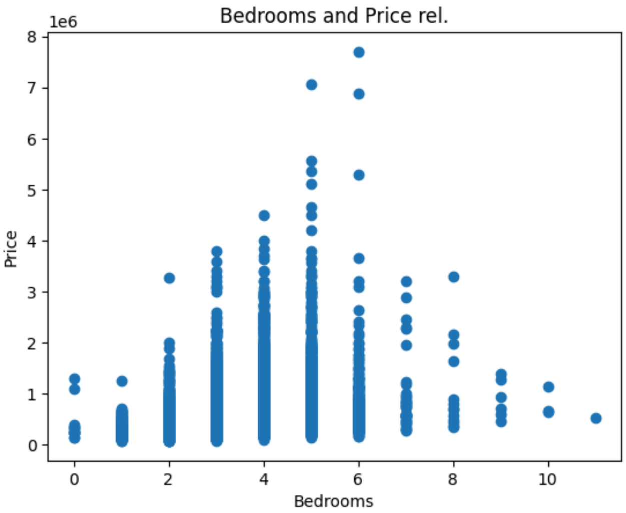

Figure 7: **Scatterplot showing the bedrooms feature without the outlier**

3.1.3 Handling 'yr_renovated'

The yr_renovated column contained numerous 0 values, signifying no renovation, which is misleading for the model, by assuming a renovation in year 0. To address this, we engineered a new binary feature named renovated. This feature indicates whether a house underwent renovation (1) or not (0). In total the data has 20.689 non-renovated properties and 914 renovated ones. As we figured out, that there is a relation regarding the price and the year (Figure 5), we want to keep the feature, but replace it with NaN, because it will tell the model that there was no renovation. This also makes sense in terms of model selection, because KNN or decision trees, for example, which can process NA values, are suitable for categorization compared to regression models.

3.2 Preprocessing steps that were omitted

3.2.1 Removing Highly Correlated Features

We observed strong correlations between some of the features, particularly between the base features (e.g., 'sqft_living') and their corresponding '_15' counterparts (e.g., 'sqft_living15'). While highly correlated features can lead to multicollinearity issues, we decided to retain all features for now and potentially explore feature selection techniques later in the model-building process, by testing several feature selection parameters.

3.2.2 Transforming Skewed Features

Several features exhibited right-skewed distributions, as evident from the histograms. We chose not to apply transformations at this stage because many multi-classification models are robust to feature skewness and do not require normally distributed input features.

3.2.3 Binning Continuous Features

We briefly considered binning some of the continuous features, such as 'house_age' or 'sqft_living', to create categorical features. However, we decided against binning at this point because we wanted to preserve the granular information present in the continuous features. We may explore binning later if we find it beneficial for improving the model's performance.

3.2.4 Scale the data

We have chosen to scale the numerical features, even if it is not necessary in the case of e.g. tree-based models, it has no disadvantages for the model itself. The advantage is that we are more independent in the model selection and can therefore expect a better result with e.g. neural networks. A standardization is used. Features that are scaled are 'sqft_living', 'sqft_lot', 'sqft_above', 'sqft_basement', 'house_age', 'sqft_living15', 'sqft_lot15', 'bathrooms', as they are numeric with a wide range of values. Not scaled are ordinal and binary features.

3.3 Potential for derived attributes

In our exploration of the data, we identified opportunities to engineer new features that could potentially enhance the model's performance.

3.3.1 Feature Engineering: House Age

To potentially enhance our model's performance, we engineered a new feature, house_age. This was calculated by subtracting the year the house was built (yr_built) from the year of the sale, extracted from the date column. The reason for this is that age is the better choice for models and scaling, as it directly shows a relevant, often non-linear relationship to the target variable and is therefore easier for the model to interpret. In addition, age has a smaller value range, which makes the scaling more stable, while the year of birth can favor scaling errors due to its large value range. Accordingly, the year of birth is then threatened.

3.3.2 Feature Engineering: Date

We observed that the 'date' column contains valuable temporal information. By extracting the 'year', 'month', and ‘day’ of the sale, we can create new features that capture potential seasonality or trends in the real estate market.

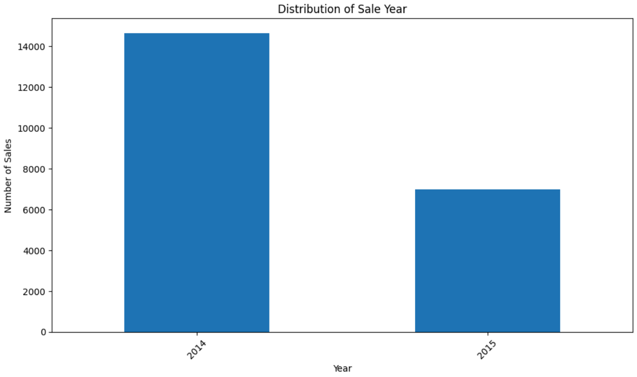

Figure 8 : **Outcome of ‘Date’ Feature Engineering Step**

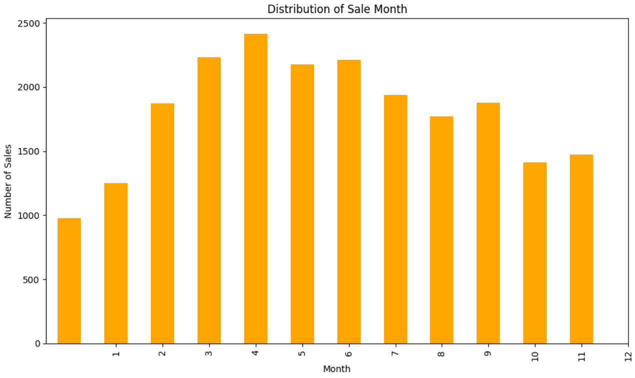

Figure 9: **Outcome of ‘Date’ Feature Engineering Step**

We believe these engineered features will provide the model with more relevant information and improve its ability to identify valuable properties.

3.4 Additional External Data Sources

While the provided dataset offers a solid foundation for our analysis, we recognize that incorporating additional external data sources could significantly enrich our understanding of the factors influencing house prices and potentially improve the predictive power of our model. Here are a few external data sources we believe could be valuable.

3.4.1 Economic Data

Integrating economic indicators like unemployment rates, interest rates, and inflation rates for the corresponding zip codes and time periods could provide valuable insights into market dynamics and their influence on housing prices.

3.4.2 School District Data

Incorporating information on school districts, such as ratings, test scores, and proximity to schools, could be highly relevant, especially for predicting the value of family-oriented homes.

3.4.3 Amenities Data

Enriching the dataset with neighborhood-level demographics (e.g., median income, population density, crime rates) could help capture the impact of neighborhood characteristics on house prices.

3.4.4 Neighborhood Demographics

Adding data on proximity to amenities like parks, libraries, and community centers could enhance our understanding of location-based factors influencing house values.

It's important to note that the usefulness of these external data sources would depend on the specific goals of our analysis and might require careful consideration of privacy and ethical implications. However, we believe that exploring these additional data sources could provide a more comprehensive and nuanced understanding of the real estate market in King County.