KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

FACULTY OF ECONOMICS AND BUSINESS

The impact of POI-data on price prediction models

Jens Vanbever

Mattias Verbruggen

Silvio Sopić

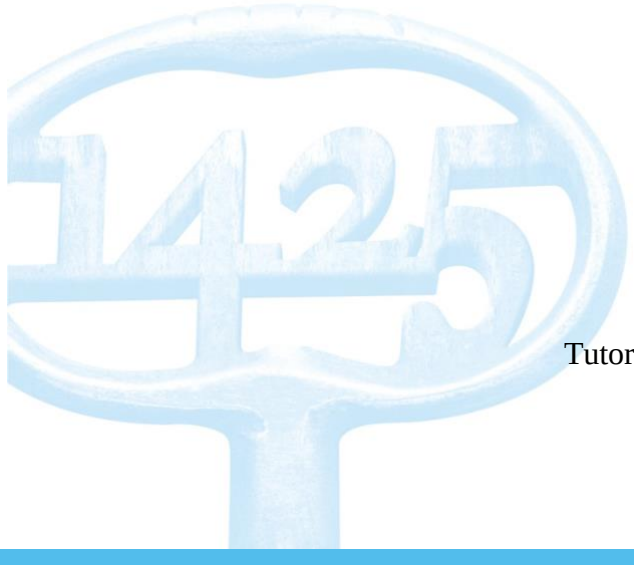
R0581532

R0673362

R0643091

Thesis submitted to obtain the degree of

MASTER OF INFORMATION MANAGEMENT



Promoter: Prof. Dr. Jochen De Weerdt

Tutor: Steven Van Goidsenhoven and Jari Peeperkorn (pro forma)

Academic year 2020-2021

KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

FACULTY OF ECONOMICS AND BUSINESS

The impact of POI-data on price prediction models

Real estate price prediction has been a hot topic for countless years and will continue to be so. While already being extensively researched the focus has only shifted towards the inclusion and study of external influences, such as location on a property in the more recent years. Continuing with this trend our thesis captured the influence of points of interest out of the Yelp database on our properties dataset from King County, USA. Having to cope with the impressive size of the Yelp database this thesis focussed mainly on the possible influential presence of 35 big categories of points of interest, within certain ranges of a property and their corresponding ratings, on the price prediction model. The analysis resulted in a list of important features in the surroundings of a property with the amount of parks within 2 kilometres, the quality of home services within 2 kilometres, the amount of elementary schools within 2 kilometres and the amount of gyms within 2 kilometres being the most important in predicting the price of real estate value. These features improved model performance in both accuracy and error values.

Jens Vanbever

Mattias Verbruggen

Silvio Sopic

R0581532

R0673362

R0643091

Thesis submitted to obtain the degree of
MASTER OF INFORMATION MANAGEMENT

Promoter: Prof. Dr. Jochen De Weerdt
Tutor: Steven Van Goidsenhoven and Jari Peeperkorn (pro forma)

Academic year 2020-2021

Acknowledgements

This master's thesis is offered as part of our master in information management. After a year of hard work and dedication, under exceptional circumstances, this acknowledgment feels like the best place to look back and be grateful for everyone that made this possible.
We would like to thank our promotor, Professor Jochen De Weerdt, for his support throughout this thesis. Thanks to his clear view on the topic and attention to detail, we were able to consider his feedback and improve our thesis tremendously.

In addition, we would like to thank the person who helped us on our way at the start of our thesis, namely Steven Van Goidsenhoven. Due to unforeseen circumstances our contact with Steven was cut short after the first few months but not after assisting us with the necessary information to get off to a flying start. We could count on him for assistance for both the technical and the literary side of the thesis. He provided us with the right kind of motivation and encouraging words which improved the progress of our paper. Finally, we would like to thank Professor Jochen De Weerdt again because he took over the job as daily supervisor from Steven for which we are very grateful.

Lastly, we would like to thank each other for the hard work and partnership during the past year. The current pandemic did not make things easier, but nevertheless we made it work. No roadblock was insurmountable, and we kept each other motivated throughout the year. We can all agree that this was a very good and complementary group and are proud of the final product we delivered.

Table of Contents

General Introduction

Houses are stationary, which means that their location plays a big part in their values. This explains the cliché that three things determine the price of a house: location, location and location. In this thesis however, we take it a step further by looking into why this location is such an important aspect of house price valuation. The way this is done is by taking into account the effect of points of interest (POI) in the immediate vicinity of these houses. We will be exploring how to use both internal and external variables when evaluating house prices. The former being features such as m² or number of bedrooms and the latter being the distance to particular points of interest.

The goal of this thesis is to develop a prediction model and to clarify how specific types of external data related to location play a role in the estimation of housing prices. The location based variables of the dataset were categorical variables which we believed did not contain all the information that would influence a purchase decision. For this purpose, we extended the dataset by downloading information on the surroundings using Yelp Fusion. We will be examining the effect of different POI by using both the distance of a house to these points within a certain radius, as well as the quality of these points, which is expressed as a rating from 1 to 5. The prediction model used in this thesis is the Random Forest Regressor, subsequently by using a basic house price prediction model we are able to see what adding these distances and ratings does to the accuracy of the model. Therefore, this thesis will add to previous research regarding prediction models for housing prices and the results can be useful for real estate agencies or notarial government branches to implement in their house price estimation applications.

Predicting the price of houses is a practice that has been around for years or even centuries. With the advent of all kinds of data, this phenomenon has become not only more precise but also more complex. Simple linear models have dominated the research field of house price evaluations for a long time but nowadays the more accurate predictions tend toward models that take non-linearity into account, such as machine learning techniques and the Classification and Regression Trees (CART). The practical usability of such models will also be explored in this paper.

In the Literature review we explore existing theory on which external features influence the house price. We found the external elements of the real-estate to be the most interesting because they explain the influence of the variables that are native to the dataset received. Even though the internal variables were of high quality, the external ones were extremely lacking. Therefore, we decided to enrich them using data from Yelp. Numerous studies have already researched the effect of different models and features on the accuracy of house price predictions. While older research (Chica-Olmo, 2007; Clapp et al., 2002) focused mainly on linear models such as kriging, linear regression, ordinary least squares and standard hedonic price equations. More recent research (Ma, J., 2020; Hong et al., 2020) shows that non-linear models tend to have better results than basic linear regression models. For this purpose, this paper will make use of the Random Forest ensemble learning method to predict the house prices of houses in King County, USA. The dataset, which contains house sale prices for King

County homes sold between May 2014 and May 2015 is publicly available at kaggle ("House Sales in King County, USA", 2021).

The results of this dissertation are encouraging and could provide a meaningful base for future research regarding real estate valuation. Several improvements in prediction accuracy have been found and the effect of distance to certain points of interest is promising. With more research into different kinds of population densities and more relevant data, the prediction model along with the features that are found could improve even further. This can create opportunities for integration into real life price prediction models

1 Literature review

The literature review is divided into two parts. Firstly, an exhaustive examination of previous research regarding prediction models for mass appraisals is conducted. This section provides information on which models have been used in the past as well as the state of the art. In addition, it is used to explain our reasoning behind the utilization of the Random Forest Regressor as our prediction model. Secondly, a literature study is provided of existing research regarding points of interest and their influence on real estate prices. Due to the enormous possibility of usable categories for POI data, a thoughtful selection has to be made. This is a challenge because selecting the right categories has a direct influence on the quality of the data and therefore, garbage in, garbage out, applies.

1.1 Prediction models for house price valuation

According to Hong et al. (2020), in the practice of mass appraisal or the automatic valuation of real estate assets, the stability and accuracy of hedonic pricing models based on linear regression remain questionable. With hedonic pricing models the price of a building or piece of land is determined by the characteristics of both the property itself (e.g., internal factors like its size, appearance and condition), as well as characteristics of its surrounding environment (e.g., external factors such as if the neighborhood has a high crime rate and/or is accessible to schools and a downtown area or the value of other homes close by).

Xiao et al. (2017) indicated that house prices are often depicted as a combination of internal and external amenities. An amenity being a feature of a property that makes it more valuable to potential buyers or tenants. Internal amenities are structural features of the house itself (e.g., size, age of building, number of bathrooms, number of bedrooms). External amenities are locational (e.g., availability of public transit, accessibility to the Central Business District) and environmental factors (e.g., availability of green spaces, scenic views) near the property. Although it is clear that certain amenities in the area of housing locations have positive effects on its price, it is not clear which type of external amenities have the biggest priority and effect and furthermore which amenities could result in lower housing prices. This paper will therefore not only contribute to the existing research regarding house price predicting models and how to model with POI data but also give an overview of the most important POI and how the distance to these points could increase the quality of prediction models.

As mentioned, older research into house price predictions focused mainly on the use of hedonic pricing models based on linear regression for the valuation of real estate prices and mass appraisals. Hong et al. (2020) showed that the accuracy and stability of these models, in this case the ordinary least squares (OLS) regression model, remain questionable. The probability of the predicted price being within 5% of its actual price was only 17.5% with the regression based models while the Random Forest model led to a tremendous 72%. According to McMillan, Reid, & Gillen (1980), the reason for the popularity of hedonic pricing models is that the marginal implicit impact of the features can be obtained by

differentiating the price function with respect to each attribute. This means that these kinds of price models make it clear which variables have a positive or negative impact on the price such as type, number of rooms and other amenities within the property. These other amenities being a desirable view such as a lake or a golf course which has been found to have a positive effect on the price in Benson, Hansen, Schwartz, and Smersh (1998), Gillard (1981), and Darling (1973).

According to Hong et al. (2020), in spite of the wide application of machine learning techniques in house price valuation, there have been few studies using Random Forest (RF) techniques for appraisal. This resulted in their research where they used a default RF model combined with latitude, longitude and distance to nearby facilities. The facilities considered are national park, high school, redevelopment area, university, general hospital, museum, and subway station. The goal of this research was to indicate the differences in accuracy of RF and basic hedonic linear models. With conventional hedonic pricing models assuming that each attribute is separable and its influence constant, leading to an extremely simplified effect of each attribute in the OLS based model. On the other hand, since the predictor in the RF model explores the hierarchical structure of features, the RF model can more sensitively track the possibility that the effect of each attribute on price varies by context. Thus, meaning that random forests are more capable in detecting the non-linearities that are specific to real estate. For example, if the real estate market is organized into a sequence of sub-markets by housing size or income or if there is a non-linearity in a household's preference regarding an attribute, the predictor that is obtained from a single regression would not be able to capture the complexities.

The problem of non-linearity in the housing market arises since we cannot directly observe the structure of preference and capture all the market characteristics causing the complexity in a market. In the real world, many characteristics of the market may intermingle but there is no flexibility in the conventional hedonic pricing model to explore such complexity (Hong et al., 2020). Therefore, this implies that the simplified nature of the OLS-based model leads to substantial losses and that some of these losses can be recovered using the RF predictor. This shows that the Random Forest method could be a useful complement to hedonic models, due to machine learning techniques being able to more adequately capture the complexity or non-linearity (external factors) of actual housing markets.

1.2 Points of interest and their influence

In terms of POI, information can be found on which facilities could have the biggest effect on house price valuations and thus which points of interest will be important to focus on in this paper. Looking at one of the biggest categories, namely transportation, a few points jump out. Having an airport near your property means bad business in most cases, the question however is: at which point or certain distance does this point stop being a burden and starts becoming a benefit? According to Limlomwongse Suksmith & Nitivattananon (2015), the only factors that demonstrate significant negative relations with property value are noise and air pollution. With the effect of noise having a bigger impact on property price than the effect of air pollution. In addition, using the NEF (noise

exposure forecast) indicator as a variable for noise and a 70 km² perimeter around the airport, the paper shows that indeed noise and air pollution are considered to be significant factors in a downwards percentage change in property value.

Train stations are another big category which could have different impacts on housing prices due to ranging distances. Debrezion, Pels & Rietveld (2006) indicated this effect on property values after correcting for a wide range of other determinants of house prices. Hence, they found that dwellings (house, flat or other place of residence) very close to a station are on average about 25% more expensive than dwellings at a distance of 15 kilometres or more. This percentage ranges between 19% for low frequency stations and 33% for high frequency stations, with frequency being calculated as trains per day. In fact, a doubling of frequency leads to an increase in house values of about 2.5%, ranging from 3.5% for houses close to the station to 1.3% for houses far away. For a negative effect of distance to railways due to noise effects, we need to look within the zone up to 250 meters around a railway line. Here prices are about 5% lower compared with locations further away than 500 meters.

Regarding POI that could potentially reduce the surrounding estate values a few points come to mind, one of them are junkyards or landfills. Research (Ready, 2010) shows that the impact of small sized landfills have little to no impact on house price deprecation, while large prominent landfills depressed nearby property values. Using different ranges of distances, Ready indicated that landfills that accept high volumes of waste (500 tons per day or more) have a greater impact on nearby property values than landfills that accept low volumes. On average, a high-volume landfill will reduce adjacent property value by 13.7%. Moreover, this impact decreases as distance from the landfill increases at a gradient of 5.9% per mile. A low-volume landfill will depress the value of an adjacent property by only 2.7%, on average, with a gradient of 1.3% per mile.

Logical reasoning regarding the effect on housing prices of facilities such as strip clubs, sexually oriented businesses (SOBs) and jails or prisons would result in property value being negatively affected by such amenities. However, little empirical evidence exists that SOBs generate such negative externalities (Brooks, Humphreys & Nowak, 2018). Despite claims based on anecdotal evidence, or rudimentary statistical analyses carried out by local planning agencies, there is no systematic evidence that supports the idea that strip clubs in Seattle generate any secondary effects in terms of negative impacts on nearby residential property values. Furthermore, empirical evidence on the possible negative effects of jails and prisons were not scientifically researched and can thus not be assumed in this paper.

Another controversial POI that needs to be acknowledged are stadiums and arenas. Neighbourhood activists often follow the NIMBY (not-in-my-back-yard) principle, arguing that the construction and implementation of a stadium will result in inconveniences such as traffic congestion, air and noise pollution, and undesirable crowds to the neighborhood. Consequently, this would cause property values to decrease. However, Tu (2005) indicated through basic hedonic models that properties close to the stadium undergo price improvements by creating benefits for local residents that offset the inconvenience caused by the stadium. The results also showed that the closer the property is to the stadium,

the greater the price improvement, and that the impact is minimal when the property is more than 2.5 miles (4 km) away from the stadium.

Education and healthcare are also two very important facilities that need to be taken into account. Rivas, Patil, Hristidis, Barr & Srinivasan (2019) analyzed several measures of property valuation near universities and hospitals based on both individual home sales and ZIP code level aggregates. Generally, the ZIP code-level analysis showed that ZIP codes with universities tend to have median home and rent prices that are above the average, especially those with medium-sized universities. Meanwhile, ZIP codes with hospitals tend to have below average median home price and median rent, with the exception of those with large hospitals. These findings regarding education and property valuation are consistent with the analysis of Wada & Zahirovic-Herbert (2013) concerning the impact of distance to schools and housing prices. It seems intuitive that smaller house-to-school distances lead to a bigger desire for this house to families with school aged children due to commuting and safety concerns. Additionally, education provided by schools could enhance economic performance and reduce crime rates in the area nearby. Finally, findings show that distance to elementary school and middle school plays a more important role than distance to high school in influencing house prices.

Other factors that should be taken into account when looking at house price valuation are distance to grocery & convenience stores, pubs & wine bars and parks. The effects of grocery and convenience stores can result in a 4 to 7 percent increase in property value for properties within 400-800 meters and 0-399.9 meters (Cerrato Caceres & Geoghegan, 2017). Next, the density of convenience stores may offer benefits to local residents in the lower-priced house neighbourhoods. On the other hand, convenience stores might reduce the local living quality (noise, potential crime), which are more highly valued by the residents in neighbourhoods with higher-prices houses (Chiang, Peng & Chang, 2015). Last, Gibbons (2004) indicated that households may like a range of pubs and wine bars in their neighbourhood but are not concerned with the distance to the nearest pub. In particular, having as many as 10 local pubs can boost the property prices by 2.8 percent. However, living very close to one specific pub could depress the price as it seems to be associated with higher levels of criminality.

Given the variation in size, usage and design of parks it is not possible to conclude a generalizable answer on the impact a park could have on house prices in its proximity. However, Crompton suggests that a positive impact of 20% is a reasonable starting point. It is consistently demonstrated by the real estate market that many households are willing to pay a larger amount of money for real estate which is located closer to a park than a house which does not have this amenity. Unless the park is, not well maintained, not easily visible from the street leading to antisocial behaviour and/or the privacy of certain properties can be violated by park users (Crompton, 2005).

2 Research questions

The goal of this thesis is to develop a prediction model and to clarify how specific types of external data related to location play a role in the estimation of housing prices. With the research questions we try to provide an answer to this goal. Furthermore, formulating clear answers to our research questions will supplement previous research regarding house price prediction models and the effect of points of interest. The following research question will be central throughout this thesis:

RQ: '*How can we build house price prediction models taking into account POI data?'*

To formulate an answer to the given research question, a number of sub questions were extracted and will be examined accordingly:

RQ1.1: 'How do we build or derive variables using POI data? Which ranges and what distances should be used to create the most accurate prediction? Which POI types have the biggest effect on real estate value?'

RQ1.2: 'Which features to use to supplement the dataset? How to select specific POI-types from Yelp based on the literature review?

RQ1.3: 'What insights can be gained regarding how POI-data derived variables impact the predictive accuracy? More specifically, what is the impact on the predictive power a single type of POI and its distance to a property have on the property's value prediction?'

RQ1.4: 'Which locational features are selected by the model predicting the price? For the purpose of post processing visualization, how can these features be visualized and what knowledge can be gained in doing so? What information regarding location can be uncovered in additional visualization efforts? How do these behave in comparison to how they should behave given the literature research?'

3 Research methodology

In conjunction with the coordinates of the real estate sold from the King County dataset we use Yelp Fusion, an API used to download location data from Yelp. Both the Yelp Fusion and the extracted POI data are discussed in the following section. Finally, we complete the research methodology section with information on how the POI data was parsed to obtain additional variables for the KC dataset.

3.1 Yelp Fusion API

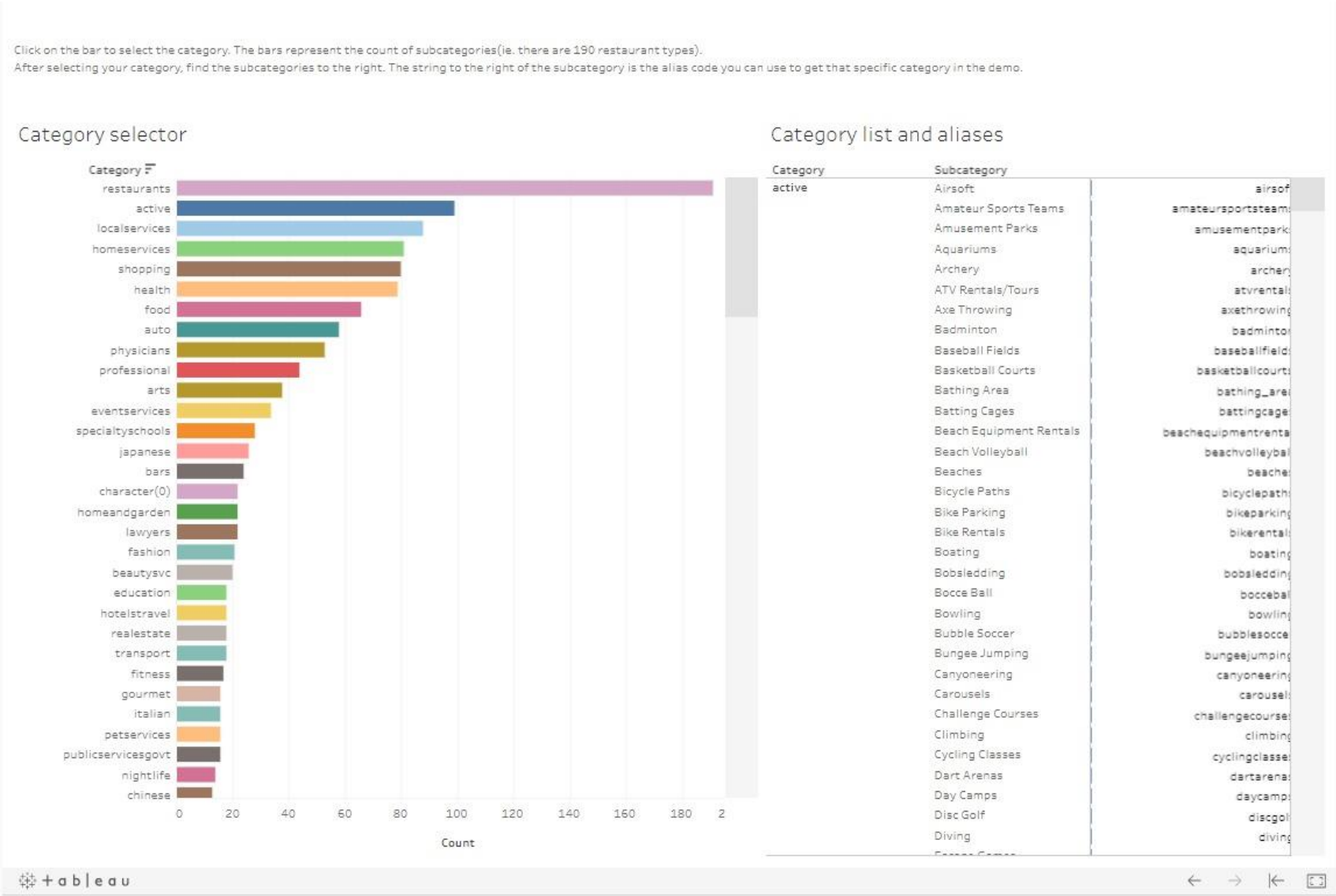
Using the coordinate information in combination with the Yelp Fusion API provided information on multiple POI types. Downloading the necessary data from Yelp Fusion is based on multiple steps. First, information such as coordinates and range are used to select the location. Then, a string of text and/or POI category can be used for the search procedure. The usage of strings for searching is out of the scope for this study. Due to this, the search is based on category and subcategory information.

The categories we will be looking at are those that influence which house and at what price a family or single person is going to purchase. Furthermore, groups of people purchasing real estate can be segmented. For example, families with five children purchase different houses than singles. Some information on these types of segments is already in the dataset through the information on the number of rooms, size, etc. A single person could select between those houses probably based on the closeness of a gym or nightlife center. Whilst a family with 3 children would want a property close to some kind of education such as primary schools or daycare. Different quantities of people in each segment buy different types of real estate, meaning that they influence the price naturally based on their own needs. Therefore, information on the locations surrounding the real estate could improve predictions of the price. The exact decision on which variables to use was grounded in the literature review

We will create more than 200 variables using this scrapping. If this was a grid search, one dimension would be the categories we use. The other dimension would be the ranges, discussed later. These categories are selected based on deep theoretical research not based only on which factors improve model performance.

For downloading purposes we used the helper dashboard in figure 1. Yelp provided a file containing all of its POI categories and subcategories in json. Next, this was flattened and then used in Tableau. For example, restaurants, the largest category has 191 subcategories. Smaller categories such as airports have only one subcategory, namely airport terminals.

Figure 1: Yelp subcategory proportions



The data was obtained using the yelp Fusion API. The process of extracting the data from Yelp starts from the authorization (tokenization). Afterwards, as the method for querying in Fusion is URL based, another standard was modified to create the initial url. The query URL can consist of a category of POI, location (lat/long), query radius, query limit, zip code, country, search term, price range, etc. Next, each of these could be used as a variable to input into the query. Ma (Ma, Cheng, Zhang, 2020) suggests using the output of the query as a dataset by using the coordinates of each POI data object from a specific category. We queried the Yelp database using a sample of 5000 King County latitude/longitude combinations and the categories suggested by the theory. This resulted in each category now becoming a dataset. All of the categories had between 75% and 80% of the data non distinct. This is probably due to the sampling being too large as each of the sampled KC locations obtained data on 50 businesses within a 20 mile radius. Therefore, the sampling was successful. A pybook demo link on how to data from Yelp was extracted can be found here. In addition, an example of the URL call used for the extraction is shown in figure 2.

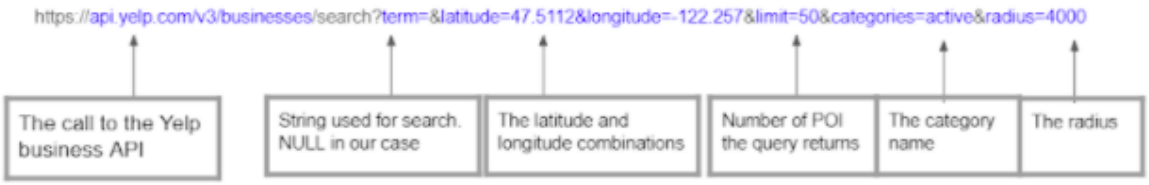


Figure 2: URL call structure

For each scrapped category we obtained the id, name, rating, the count of ratings, coordinates, address, city, state and distance. A snapshot of this table is pictured in figure 3.

```
head(Business_data_Total)
```

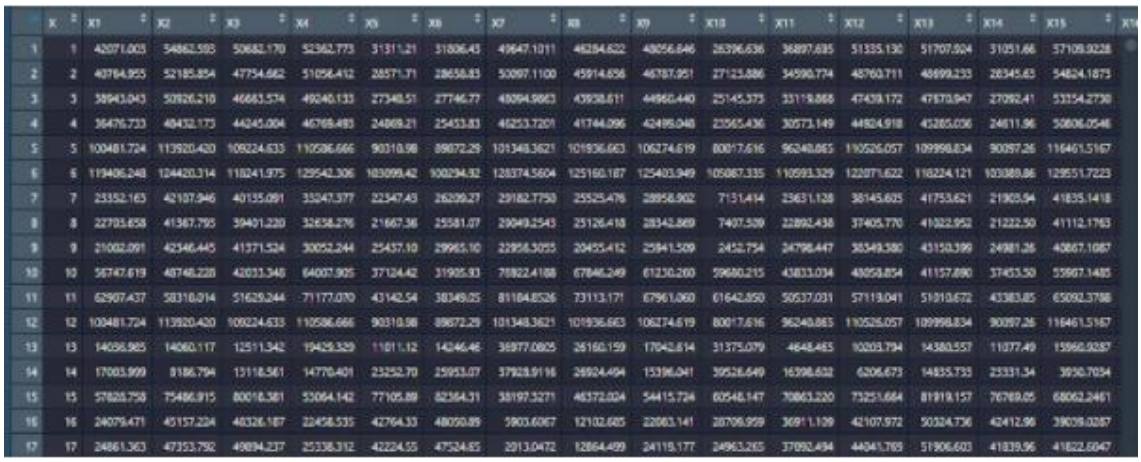
id	name	rating	review_count	latitude	longitude	address1	city	state	distance
EK96wCmNDDUik4pdrlqUlw	Holman's Body & Fender Shop	4.5	130	47.68176	-122.3178	7301 Roosevelt Way NE	Seattle	WA	1377.215
_ACODTT4fG6aiksVpi0KzlA	Grease Monkey	4.5	192	47.66833	-122.3003	2501 NE 55th St	Seattle	WA	2267.547
_5YDJBhA0yn4XUjy3XG1mQ	Quality Auto Glass	4.5	166	47.68559	-122.3446	7801 Aurora Ave N	Seattle	WA	1819.491
2oCuGfM2MurN9grimH5tCQ	Matt's Greenwood Auto Care	5.0	117	47.68642	-122.3550	7900 Greenwood Ave N	Seattle	WA	2433.267
UfOusL4tcaorQb4SerUsA	D & D Brakes	5.0	222	47.70620	-122.3443	10538 Aurora Ave N	Seattle	WA	3902.675
gB6chno2PVM0XHyXMZfcfbA	Rack N Road Car Racks & Trailer Hitch Superstores	4.5	108	47.68667	-122.3441	7918 Aurora Ave N	Seattle	WA	1898.158

Figure 3: Example of category dataset

The distance is an artifact of the scrapping process and is the distance from the houses of the King County dataset and the categories of the Yelp Fusion scrap. As our main method for parsing the location are the coordinates, the variables address, city and state are removed. Next, the id is used to make sure that each observation is unique and will not be used for other purposes. Finally, as the rating count will also be removed the remaining variables are rating, latitude and longitude. As this process is completed for each category, an additional variable, the category (name) was added.

After obtaining the table for an object type POI we connected them to the original KC dataset. Then, for each of the observations of the KC dataset we used the latitude and longitude variables to find the distance between them and each Yelp object. Therefore, an empty matrix (nrows of POI * nrows of KC) was created with the rows as the Yelp identifications and the house id's in the columns as house

identifications. This matrix was then filled with the Euclidean distances between each. With data in this form we obtained the number of objects within 50, 100, 500, 1000, and 2000 meters by making the sum of the respective number of observations within that range. An example of how this distance matrix for a specific Yelp category looks like is shown in figure 4.



	X	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15
1	1	42071.003	54861.595	50682.170	52362.773	31311.21	31806.43	49647.1011	46284.622	48056.846	26396.636	36897.635	51335.130	51707.924	31051.66	57109.9228
2	2	40764.955	52185.854	47754.662	51056.412	28571.71	28658.83	50897.1100	45914.656	46787.951	27123.886	34590.774	48760.711	48699.233	28345.63	54824.1873
3	3	38943.043	50926.210	46663.574	49240.133	27348.51	27746.77	48094.9863	43938.611	44960.440	25145.373	33119.868	47439.172	47870.947	27092.41	53354.2730
4	4	36476.733	48432.173	44245.004	46769.483	24868.21	25453.83	46253.7201	41744.096	42895.048	23565.436	30573.149	44824.918	45285.036	24611.96	50806.0546
5	5	100481.724	113920.420	109224.633	110586.666	90310.98	89872.29	101348.3621	101936.663	106274.619	80017.616	96240.865	110526.057	109998.834	90097.26	116461.5167
6	6	119406.248	124420.314	118241.975	129542.306	103099.42	100294.32	128374.5604	125160.187	125403.949	105867.335	110593.329	122871.622	118224.121	103089.86	129551.7223
7	7	23352.163	42107.946	40135.091	33247.377	22547.43	26209.27	29182.7750	25525.476	28958.902	7151.414	23631.128	38145.605	41753.621	21903.94	41835.1418
8	8	22703.658	41367.795	39401.220	32638.276	21667.36	25581.07	29049.2543	25126.418	28342.869	7407.509	22892.438	37405.770	41302.952	21222.50	41112.1763
9	9	21002.091	42346.445	41371.524	30052.244	25437.10	29965.10	28958.3055	26455.412	25841.509	2452.754	24796.447	38349.380	45158.399	24981.26	40867.1087
10	10	56747.619	46748.228	42033.348	64007.905	37124.42	31905.93	78822.4188	67846.249	61230.260	59680.215	43833.034	43058.854	41157.890	37453.50	55997.1485
11	11	62907.437	58310.014	51629.244	71177.070	43142.54	38348.05	81184.8526	73113.171	67961.060	61642.850	50537.031	57113.041	51010.672	43383.85	65092.3788
12	12	100481.724	113920.420	109224.633	110586.666	90310.98	89872.29	101348.3621	101936.663	106274.619	80017.616	96240.865	110526.057	109998.834	90097.26	116461.5167
13	13	14056.985	14060.117	12511.342	19429.329	1101.12	14246.46	36977.0805	26160.159	17042.814	31375.079	4648.465	10203.794	14380.557	11377.49	15960.9287
14	14	17003.999	8186.794	13118.561	14770.401	23252.70	25953.07	37928.9116	26924.494	15396.041	39528.649	16398.632	6206.673	14835.733	23331.34	3930.7054
15	15	57828.758	75486.915	80016.381	53064.142	77105.89	82364.31	38197.3271	46372.024	54415.724	60548.147	70863.220	73251.664	81919.157	76769.05	68062.2461
16	16	24079.471	45157.224	46326.187	22458.535	42764.33	48050.89	5803.6067	12102.685	22083.141	28706.959	3691.1.109	42107.972	50324.736	42412.96	39038.0287
17	17	24861.363	47353.792	49894.237	25338.312	42224.55	47524.65	2913.0472	12864.499	24119.177	24963.265	37092.494	44041.769	51906.603	41839.96	41822.6047

Figure 4: Final distance matrix for a Yelp category

An additional variable was created which contained the average rating of the objects within each distance. In cases where no objects of that category existed, a value of -1000 was imputed into the average rating variable. Our model will therefore be able to distinguish between the quantity of the observations, as well as the average quality within a range. Figure 5 shows the naming conventions used for each variable.

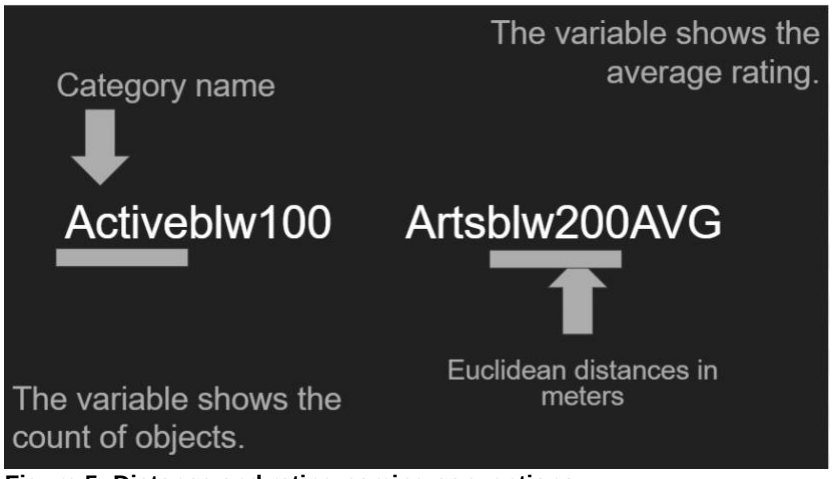


Figure 5: Distance and rating naming conventions

3.2 King County variables feature engineering

In this section we discuss the preparation of the KC variables for predictions using the Random forest model (RF). Thanks to the robustness of a RF and ease of use we opted to use this technique to construct our price prediction model. When creating a base dataset for our RF, one has to carefully consider each feature of the dataset and the type of information that it represents. When

working with a large amount of features, chances are that not every feature will contribute in a positive way to the predictive model. With this thought in mind we started by removing features like the *ID* and *date*. Since our housing sales data is a collection of sales between May 2014 and May 2015 and we can see that over the course of the year (table 1), no obvious differences in average price for every month can be seen. In conclusion, this supports our decision of removing the ID variable.

Table 1: Sales between 2014-2015

Month	Number of sales	Average sale price
05/2014	1768	€ 475 099,66
06/2014	2180	€ 480 971,61
07/2014	2211	€ 472 442,23
08/2014	1940	€ 472 711,73
09/2014	1774	€ 467 069,98
10/2014	1878	€ 463 481,31
11/2014	1411	€ 454 644,79
12/2014	1471	€ 453 161,58
01/2015	978	€ 452 815,77
02/2015	1250	€ 450 067,32
03/2015	1875	€ 468 109,54
04/2015	2231	€ 490 006,55
05/2015	646	€ 471 632,46

Similar to the *date*, *ID* holds no value to us. As the next step we took a closer look at some specific features that we thought would be very useful but weren't necessarily in the right format. If we take a look at the *yr_built* variable this represents the year that the property was built. This approximately ranges from the 1900's up until 2015 so we can easily derive a variable that is much easier to understand, namely the age of the house/appartement by subtracting the year of the sale *date* with *yr_built* to create *age.* Furthermore, we also created a new variable based on the feature *yr_renovated* which is either 0, if the property was never renovated, or the year of the renovations. This leaves us with a feature that is mostly 0 or suddenly nearly 2000. Just like the *age* variable we tried to optimize this by creating a new boolean variable *renovated*, as the name suggests the variable equals 1 if the property has been renovated or 0 if it hasn't been. All the other features were left as found in the original dataset. Finally, for every RF model the dataset was divided in a training and test set using a 70/30 split.

3.3 Additional preprocessing

Genesove & Mayer (1994) have confirmed the existence of a feature which helps predict the price of the real-estate, namely the count of times sold. This feature highlights how popular a certain house is. As such, this could become a source of noise for our data. The count of these houses in our dataset is less than 200 observations for a dataset of 21000 observations. Therefore, it seems important to either remove these observations or create a dummy variable.

3.4 Feature evaluation

After combining all these features in our predictive model, we quickly realised that maybe not all of them were of that much importance. Therefore, we decided to investigate this using the built in function of the RandomForestRegressor class from the Python SciKit-Learn library. For a classification model, the feature importances would be based on the decrease in impurity which would be calculated using the gini impurity divided by information gain that results in the entropy. However, for our regression tree these importances are based on the variance. These results showed us that the features “floors”, “yr_renovated” and the newly derived feature “renovated” had an importance of less than 0.01 and thus were removed. In the results section, more information about the base model results is provided.

4 Implementation

We ran our first predictive model with only the internal features with an importance higher than 0.01 using the sensible default parameters (Buitinck et al., 2021) from the Scikit-learn library. This made us look into the direction of hyperparameter tuning even though this would lead us into a direction of trial and error experiments instead of the easy to use, accessible baseline random forest. With hyperparameter tuning we want to optimize the parameters the model uses to train the model.

4.1 Cross validation

Due to the experimental nature of this method we had to be careful not to fall victim to overfitting (high performance on training set but poor performance on test set) if we would only search for parameters based on our training set. As a result, we immediately opted for Cross Validation (CV). Due to the size of our dataset and limited resources we used 3 fold CV. This means we trained our model 3 times, each time using a combination of 2 out of 3 folds and evaluation on the third fold. Afterwards we evaluated the performance of the complete model by averaging the performance of each fold. This resulted in a large amount of model iterations but luckily once again the Scikit-learn library had us covered. The RandomizedSearchCV method allowed us to perform the 3 fold CV for random samples from the parameter ranges.

4.2 Optimal parameters

Following the example of (Khandelwal, Chaturvedi & Gupta, 2020) we focussed on the parameters responsible for the size of the trees and we achieved an improvement in accuracy of 0.65% for the complete model with all Yelp derived features, for the following optimal parameters in table 2.

Table 2: Optimal parameters RF

Parameters	Ranges	Optimal value
n_estimators	[200 ; 2000]	1400
min_samples_split	[2, 5, 10]	2
min_samples_leaf	[1, 2 , 4]	1
max_features	[Auto, Sqrt]	Auto
max_dept	[10 ; 110]	100
bootstrap	[True, False]	True

5 Results

In this results section a descriptive dataset analysis of both datasets will be conducted. After the data analysis the price predictions for the base model, locational features, distances and subcategories will be discussed.

5.1 Exploratory dataset analysis

For the exploratory data analysis we produced two dashboards. One is the exploratory analysis of the provided King County dataset while the other contains information on all of the objects downloaded through Yelp Fusion. Multiple settings were used so we limit our examples.

5.1.1 King County dataset

For the KC dataset exploratory analysis we will be discussing outliers. The dataset provided was the King County (KC) dataset on real-estate price prediction. It has both the internal (number of rooms, size etc.) and external (view quality, waterfront etc.) hedonic variables mentioned in the literature review. Below the dashboard for the exploratory King County dataset analysis of the squared feet living size and price can be found in figure 6.

We see a positive correlation between the price and size. In the histogram we can notice how there is a tail towards higher prices and quite some outliers. These outliers could be errors in data input.

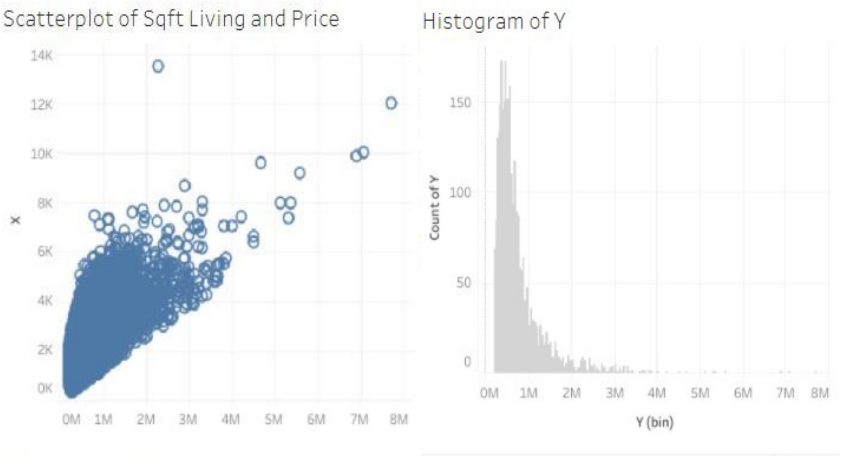


Figure 6: Squared feet living size and price exploratory analysis

There is a positive correlation between price and both grade and number of bathrooms (figure 7). The same observations that are outliers in the price histogram seem to be outliers in these scatterplots as well (information available as tooltip on dashboard).

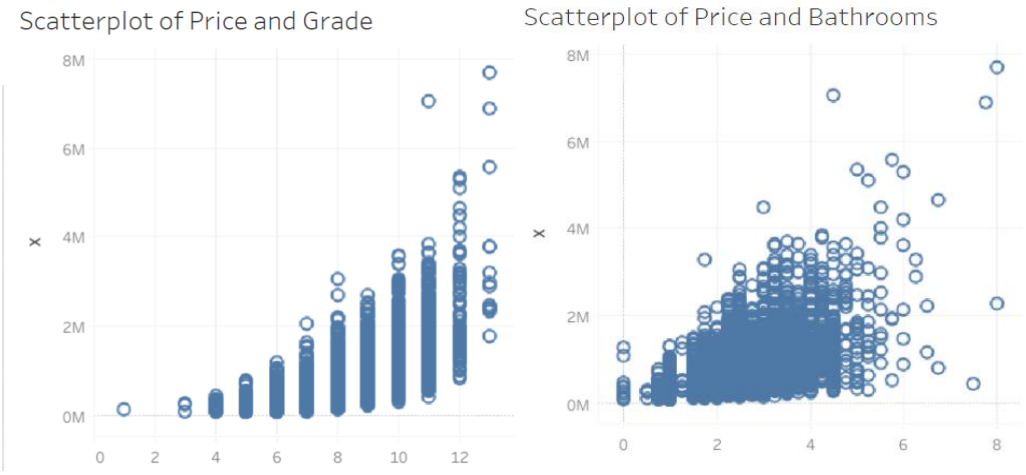


Figure 7: Number of bathrooms, price and grade exploratory analysis

The next item we viewed was the view variable. There are two charts below (figure 8), one showing the data with prices below 2 million and the other above (max is 7.7 mil). This cut off was decided based on the tail of the price histogram. This allows for direct observation of the tail of the price histogram. The bubble size is defined through the variable view. The color intensity pictures the price. The darker circles locations on the left have a higher rating for the view variable. The similar pattern can be observed to the right. That chart shows only the observations whose price was above approximately the 3rd quartile.

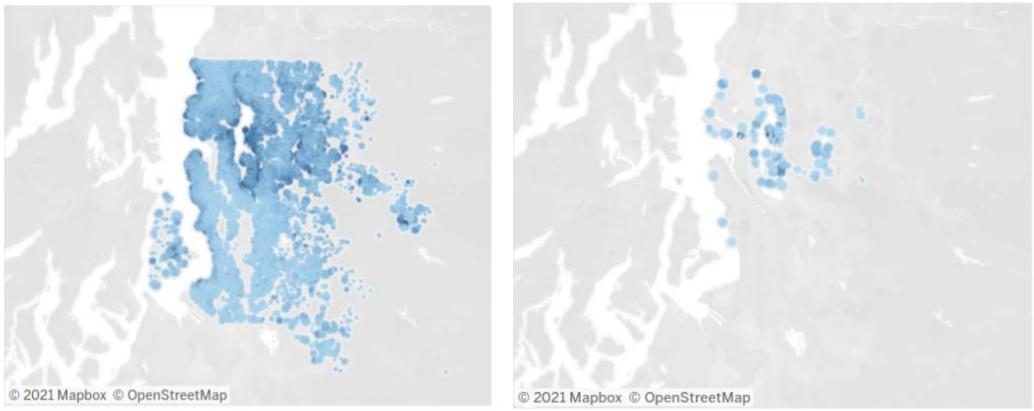


Figure 8: Map graphs with price as color intensity and view as size.

In conclusion, high levels of attributes which should positively influence the prices have the same outliers. As such we understand that these are not errors and therefore do not continue with outlier removal.

5.1.2 Yelp dataset

For this section of the results, we explore the data downloaded from the Yelp API through a dashboard (figure 9). This dashboard is an amazing method to filter out specific information about a Yelp POI category or find it on the map.

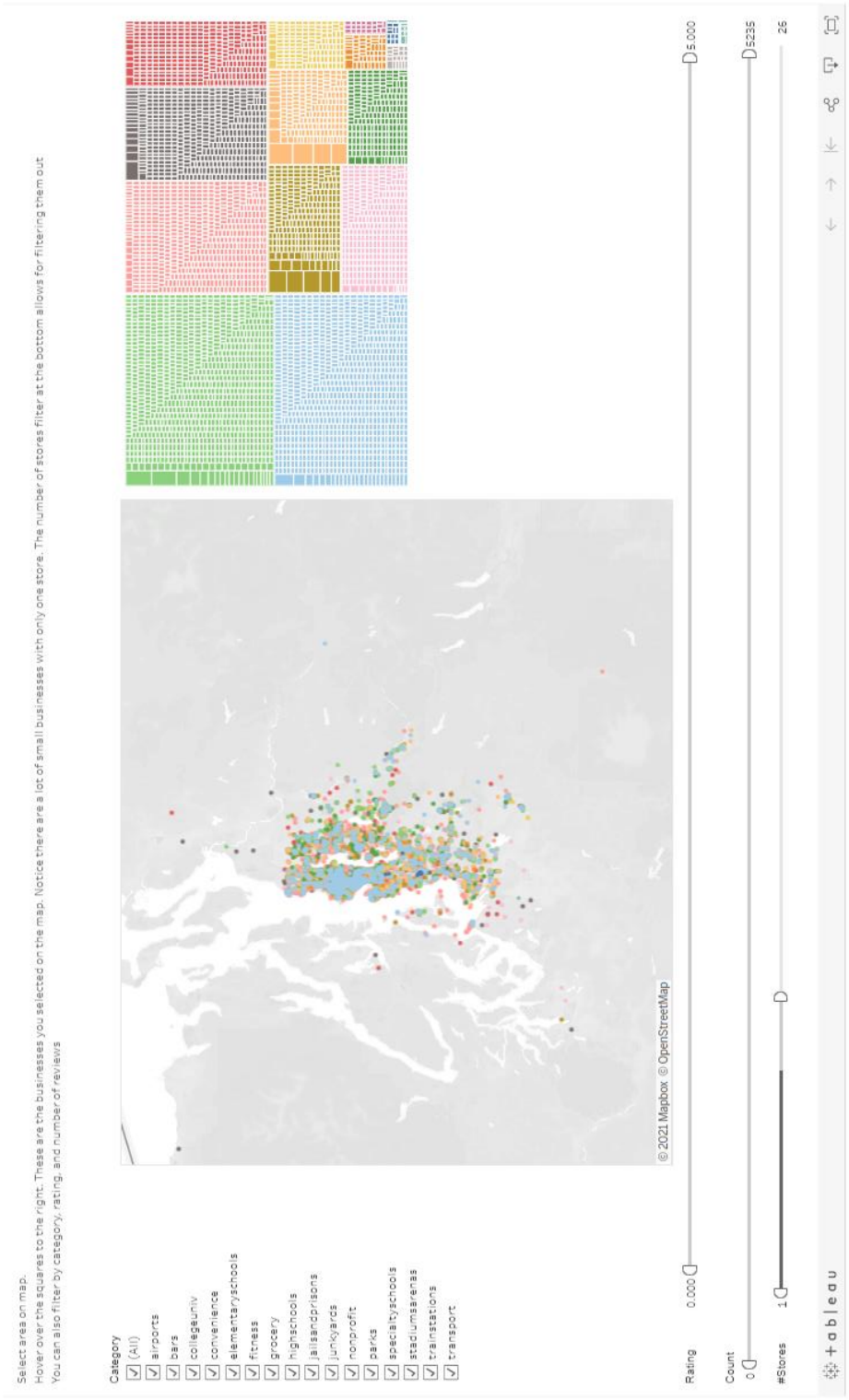


Figure 9: Yelp dataset exploration

The variables present in this dataset are available as a filter to the left. They include airports, junkyards, fitness centers, grocery stores, etc. In addition, they are mapped onto the map in the middle. To the right are the filtered observations, which are also proportions of the subcategories in each category. Selecting only airports on the category selector plots them on the map (figure 10). Next, when we compare this image to the plot of the most expensive houses (figure 11) we can clearly see that most of those are further away from them. This is a clear example how a negative external variable can make the price of surrounding real estate plummet.

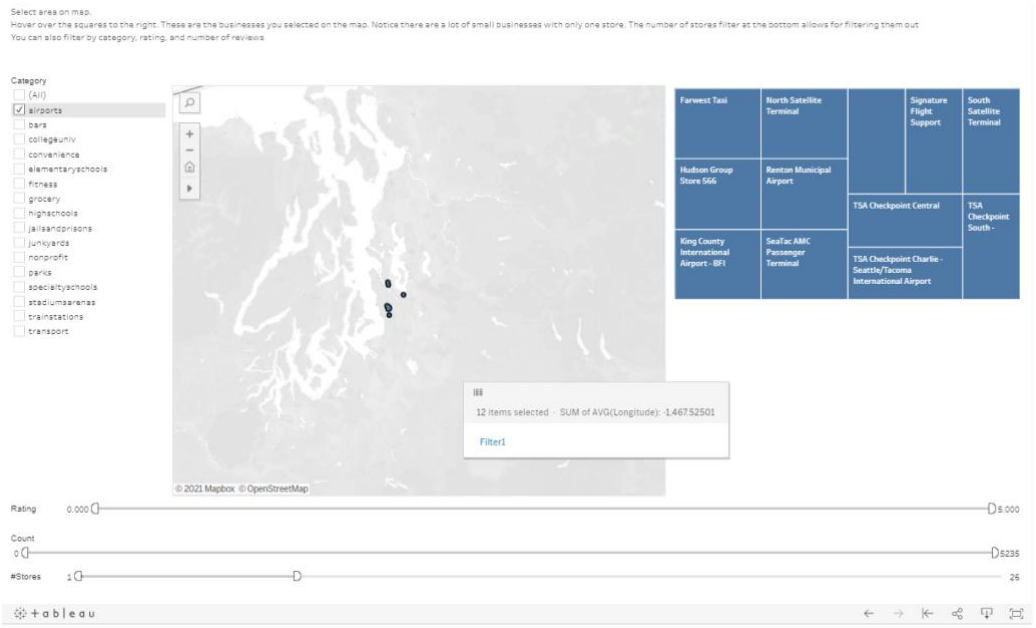


Figure 10: Airport location

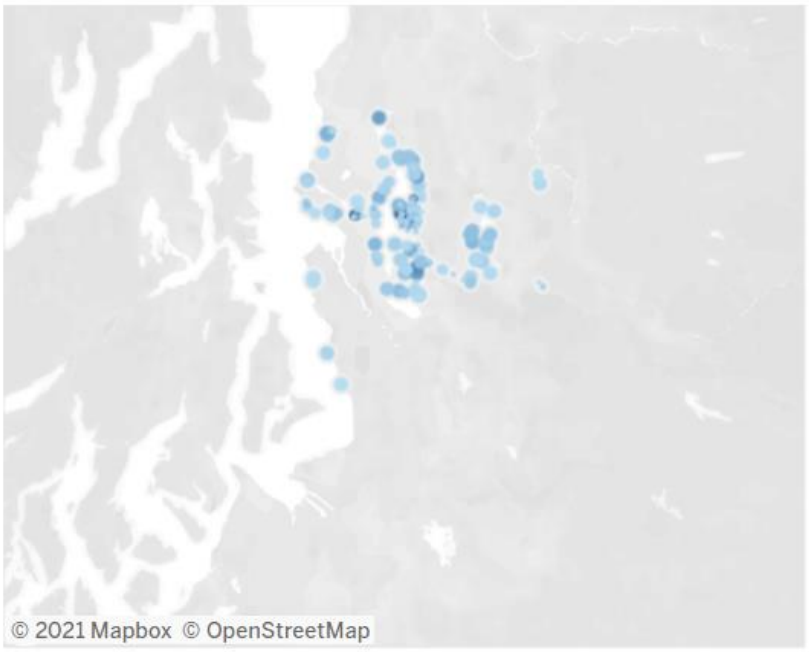


Figure 11: Location of most expensive houses

5.2 Price predictions

The results that follow will be discussed on the basis of the mean absolute percentage error (MAPE), the mean absolute error (MAE) and the R^2 score. Firstly, the MAPE is a measure of how accurate a forecast system is and will be shown as a percentage. Since MAPE is a measure of error, lower numbers are better. However, it is not responsible to set arbitrary forecasting performance targets such as MAPE < 20% is good and MAPE < 30 is less good. Furthermore, the MAPE can be translated to an accuracy number by subtracting it from 100 although this is not an industry recognized acronym. Secondly, the MAE displays the mean error between the price predicted by our prediction model and the official house price. Lastly, the R^2 score which is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

5.2.1 Base model

Before going into the analysis of the effect of distance to POI, we establish a base model which takes into account all internal features of a house. More specifically, this model consists of properties specific to the house such as the grade, age, the number of bathrooms, square footage, view, etc. Consequently, this base model will be used as a benchmark throughout the results section due to the exclusion of locational features such as longitude, latitude and distance to POI. As is shown in table 3, our base model has a MAPE of 23,94%. This means that without any locational or distance metrics, the base model can predict the house price with an accuracy of 76,06% on the basis of internal house features alone. Furthermore, the MAE indicates that the prediction of the house price with the base model differs, on average, 118906 dollars from the actual value. Finally, seeing as the R^2 has a value of 0,7276, the features of the base model declare the independent price variable fairly well, with an R^2 of 0,70 to 1 being a good fit.

Table 3: Base model metrics

Model	MAPE (%)	MAE ($)	R^2
Base model	23,94	118905,71	0,7276

The importance of every variable in predicting the price with the base model is shown in figure 12. It is clear that the grade of the house, being an index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design, has the biggest impact on the model with an importance of 0,37. Followed shortly by sqft_living, which is the square footage of the estate's interior living space with an importance of 0,27. The third most important variable is age, which is our own manipulated variable of the initial yr_built variable with an importance of 0,10. Ultimately, other important variables are sqft_living15 (0,06), sqft_lot15 (0,04), sqft_lot (0,03) and sqft_above (0,03).

```
Variable: grade                Importance: 0.37
Variable: sqft_living          Importance: 0.27
Variable: age                  Importance: 0.1
Variable: sqft_living15        Importance: 0.06
Variable: sqft_lot15           Importance: 0.04
Variable: sqft_lot             Importance: 0.03
Variable: sqft_above           Importance: 0.03
Variable: bathrooms            Importance: 0.02
Variable: waterfront           Importance: 0.02
Variable: view                 Importance: 0.02
Variable: bedrooms             Importance: 0.01
Variable: floors               Importance: 0.01
Variable: condition            Importance: 0.01
Variable: sqft_basement        Importance: 0.01
```

Figure 12: Base model variable importance

5.2.2 Locational features

We want to expand our base model by adding features based on location, namely longitude, latitude and zipcode. These variables were made available through the Kaggle database. As a result of adding these features it will be possible to enhance our prediction model by being able to take the location of houses into account. In addition, it will be easier to predict the price of houses in similar locations or clusters, thus making our predictions more accurate. Find below (figure 13) the dependent variable price, log transformed (only for visualization purposes) and then plotted as color gradient over each location below.The dark blue elements are houses of lower price while the green are of higher. The inflection is portrayed by the white gradient.

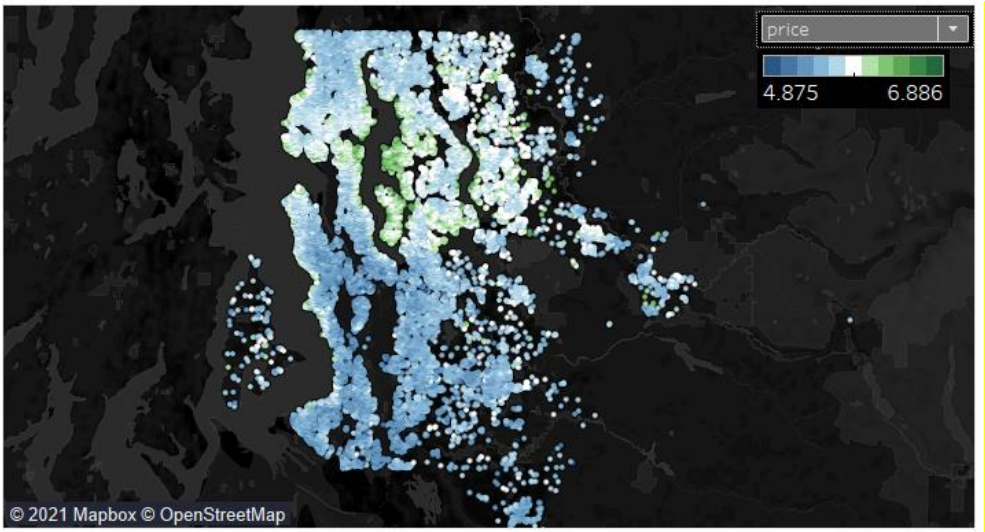


Figure 13: Dependent variable price

First of all a prediction model using only latitude, longitude and zipcode was used to create a comparison between solely locational and internal house features. Table 4 indicates that the locational model realises a MAPE of 22,54% compared to the 23,94% of the internal features prediction model. This immediately suggests the importance of using characteristics based on location when predicting real estate value, as the prediction model based on location has a slightly lower MAPE than the prediction model using only features of the house itself. However, looking at the R^2 value we can clearly see a significant difference from the internal feature predictor (0,7276) down to 0,6035 for the locational model. Further, an MAE of 125855 is higher compared to the base model,

despite the locational model having a superior MAPE value. The reason for this difference being the fact that the MAPE utilises real measures while the mean absolute error uses absolute values. In conclusion, it is apparent that both the base and locational model are fairly good prediction models, nevertheless the most accurate model will be a combination of both.

As mentioned above, a combination of both the base model and the locational model will result in a high predictive quality. This statement is validated by a MAPE of 13,52% for the combined model, which is a big improvement compared to the 23,94% of the internal base model and the 22,54% of the locational model. A mean absolute error of 73259 can be seen as a good outcome taking into account the nature of the price variable and the many 'outliers' that come with real estate prices. In comparison to the base model, this is an improvement of 45647 dollars for the MAE. Furthermore, the R^2 has seen a jump of 0,1297 to 0,8573. This suggests that almost 86% of the variance of the house prices can be explained by the variance of the internal house and locational features. Also remarkable, is the very small decrease in accuracy when we leave zipcode out of the equation. This can be explained due to zipcode being a worse version of latitude and longitude as they are arbitrary values. Think of this as comparing the price of houses in different parts of a big city. Some neighbourhoods would have a bigger price but it would be better to also have information on parts of the neighbourhoods.

Table 4: Locational features

Model	MAPE (%)	MAE ($)	R^2
Latitude, longitude and zipcode	22,54	125855,03	0,6035
Base + longitude and latitude	13,52	73812,7	0,8563
Base + longitude, latitude and zipcode	13,43	73258,77	0,8573

Figure 14 further indicates the importance of adding locational variables such as longitude and latitude. With grade (0,33) and sqft_living (0,27) still being the most important features. As explained above, latitude and longitude take third and fourth spot with a 0,14 and 0,06 importance respectively, causing age to a 0,06 drop from 0,10 to 0,04. The condition of the house and the number of bedrooms/floors now completely drop off and have no more importance in predicting the price. This could be due to bedrooms and floors being indirectly measured in the sqft_living variable and condition being integrated into the grade of the real estate. However, the 0,0 importance should always be taken with a grain of salt as this does not mean that these features have absolutely no effect on the prediction model.

Figure 14: Locational feature importance

5.2.3 Distance to POI

The aim of this dissertation is to examine the influence the distance from different points of interest to real estate locations has on our prediction model. Firstly, in total 35 categories have been extracted from Yelp and the distance from every POI to every house has been calculated. These distances were then bundled in features ranging from 100, 200, 500, 1000 and 2000 meters. Secondly, besides distance, the average rating of every POI within a certain range was calculated and added to the model as features. Therefore, for instance, not only is the distance from a certain house to every park within a 100 meters radius calculated but the average rating of those parks as well. Lastly, the effect of distance to a certain POI on the prediction model has been calculated and is shown in table 5.

The goal of this overview is to look into which POI could potentially play an important role in predicting the house prices. Since the MAPE values are ordered from low to high we can directly see that the base model with just the distance to parks (19,55%), active life categories (19,61%), real estate (19,84%), fitness (19,91%), financial services (20,03%), elementary schools (20,16%) and shopping facilities (20,20%) have a fairly low MAPE. Furthermore, all models have a lower MAPE value compared to the base model with just the internal features (23,94%), with junkyards closing the list.

Not only is the MAPE better but the mean absolute error and R^2 of every category is higher than the values for the base model, meaning that not a single distance to a POI decreases the quality of our prediction model. However, the MAPE of airports (23,77%), stadiums and arenas (23,56%), colleges and universities (23,10%), train stations (22,58%) and convenience stores (22,10%) are lower than expected based on the literature review. The reason for this can be due to many different factors such as using datasets of other cities, less densely populated areas, more qualitative data available than the Kaggle database used in this paper, etc.

Table 5: Distance to POI

Model	MAPE (%)	MAE ($)	R^2
Base + parks	19,55	101313,01	0,7831
Base + active	19,61	102587,08	0,7668
Base + real estate	19,84	102714,89	0,7728
Base + fitness	19,91	103259,53	0,7737
Base + financial services	20,03	104265,25	0,7652
Base + elementary schools	20,16	103465,31	0,7814
Base + shopping	20,20	103630,78	0,7699
Base + specialtyschools	20,34	106056,63	0,7604
Base + professional services	20,56	106745,53	0,7536
Base + home services	20,61	105871,79	0,7671
Base + food	20,69	105849,86	0,7657
Base + arts	20,84	107214,9	0,7655
Base + bars	21,11	107145,86	0,763
Base + restaurants	21,18	107350,76	0,7577
Base + hotel and travel	21,35	107968,31	0,7629
Base + education	21,41	109278,39	0,7527
Base + highschools	21,44	108190,34	0,7676
Base + grocery stores	21,47	108524,59	0,7589
Base + non-profits	21,73	109790,1	0,7519
Base + convenience stores	22,10	110838,05	0,7499
Base + train stations	22,58	112425,14	0,7531
Base + college and universities	23,10	114255,15	0,757
Base + adult entertainment	23,51	117409,39	0,7311
Base + stadiums and arenas	23,56	117589,43	0,7304
Base + airports	23,77	118224,17	0,7292
Base + junkyards	23,81	118430,65	0,7284

Although the effects of singular POI on the prediction model are interesting, the intention is still to obtain a general overview of the most important POI when taking all factors into account.

This overview is shown in table 6 where all distance and rating features are grouped and the distance radius is incremented accordingly. Notice that with every increase of the radius the quality of the model will improve as well, due to more facilities being added/found within a certain distance. Starting with the smallest distance of a 100 meters a small improvement in MAPE can be observed going from 23,94% for the base model to 23,84%. This can be seen as straightforward as little facilities can be found within a 100 meter radius, thus only giving the prediction model little to work with. The biggest effect can be seen in the last row when the base model is combined with the distance of all POI within 2000 meters and the corresponding average. As a result, a MAPE value of 15,09% for the POI distance prediction model can be noticed in comparison to the 23,84% for the base model. In addition, with a mean absolute error of 80819, the POI model performs way better than the base model with a 118905,71 MAE. The R^2 value of 0,8438 is also a good sign of prediction by the POI distance features, meaning that the variance of the house prices can be explained by the variance of the distance between real estate and the points of interest.

Comparing our prediction model with POI distances, to the prediction model including the latitude, longitude and zipcode we can declare that the POI distance model holds up very well. Taking into account that latitude and longitude are very precise measurements in a random forest regressor, the POI model is inferior to the locational model with a miniscule difference of 1,57% MAPE. Finally, a difference of -7560 dollars mean absolute error and -0,0135 for the R^2 value all lead to the conclusion that the prediction model with distance and rating of POI within 2000 meters comes very close to the latitude/longitude model.

Table 6: Distance segments

Model	MAPE (%)	MAE ($)	R^2
Base + below100 and below100AVG	23,84	118599,05	0,7286
Base + below200 and below200AVG	23,52	117216,49	0,7313
Base + below500 and below500AVG	21,64	109722,15	0,7543
Base + below1000 and below1000AVG	17,95	94448,93	0,7967
Base + below2000 and below2000AVG	15,09	80818,65	0,8438

For completion, the importance of the variables from the below2000 and below2000AVG model is given in figure 15. What is clear is the importance of the parks category with a 0,03 importance which, as expected in table 5, is the highest of all features. Followed by home services (0,02) which exists out of facilities such as electricians and real estate, elementary schools (0,02) which is a subcategory of the education subcategory and was isolated because of the importance indicated in the literature review, and fitness (0,02).

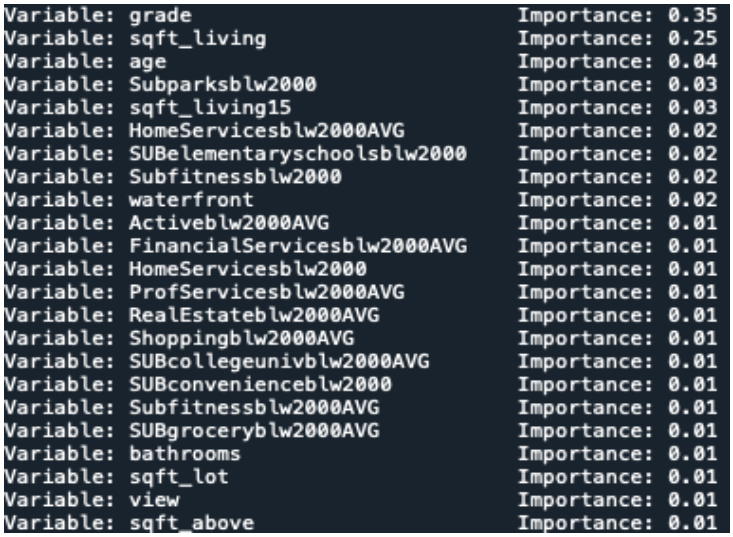


Figure 15: POI importances

5.2.4 Subcategories

In accordance to the literature study, the following section explores whether the importance of some subcategories would go lost if we only included the main categories. As a results, a few subcategories were extracted and focused on: adult entertainment, airports, bars, colleges and universities, convenience stores, elementary schools, fitness, grocery stores, high schools, junkyards, non-profits, parks, specialty schools, stadium and arenas and train stations. Table 7 shows the outcome of these results.

Running the model with the distance and average score of every subcategory named above within 2000 meters gave a MAPE of 15,46% which, in comparison to the outcome of the model in table 6 (15,09) is only a -0,36 difference. Hence, our subcategory selection practically has the same quality as the model with every category included. This indicates that a carefully chosen selection of subcategories, based on literature review can greatly enhance the prediction model.

The final row of the table indicates another popular measure for house price predictions namely, the distance from a certain property to the central business district (CBD) or city centre. Since Yelp did not have the option to extract the CBD from their database we had to artificially create this feature. Thus, this approximation was made by using the distance to the following categories/subcategories: Arts (culture center, musea), shopping (shopping centre), bars, college & universities and fitness. The final outcome is a MAPE of 17,17%, a R^2 score of 0,8147 and an MAE of 90458. In conclusion, these results indicate that not only do the selected categories create a good approach to the CBD district but that the distance to the central business district is a good feature for house price prediction as well.

Table 7: Subcategories

Model	MAPE (%)	MAE ($)	R^2
Base + every subcategory	15,46	82585.44	0,8381
Base + distance to city centre	17,17	90457,86	0,8147

Finally, figure 16 pictures the importance of every subcategory when isolated from the main categories. It is clear that the distance to a park within a 2000 meter radius remains the most important POI (0,04), even gaining in importance in comparison to the full model (0,03). Furthermore, elementary schools (0,03), college and universities (0,02) and fitness (0,02) follow closely after. What is also clear is that most of the subcategories play some sort of role (importance) when predicting house prices and most subcategories and their importance align with previous research which is further elaborated in the discussion.

Figure 16: Subcategory importance

6 Discussion

The basic principle of this study originates from earlier research which shows that the valuation of a house is dependent on the location and different points of interest in its nearby vicinity. This indicates the concept of a ‘neighbourhood’. According to Kiel & Zabel (2008) “individuals care about their very local surroundings such as the general upkeep of their street and possibly their neighbors’ characteristics (cluster variables), and a broader area such as the school district and/or the town that accounts for school quality and crime rates (tract variables)”. These tract variables which require more focus and are more difficult to map the specific impact of, since the influence of cluster variables is more easily researchable with the help of locational variables such as longitude and latitude. As our literature review suggests, research regarding different categories of POI-data and its influence are plenty. Thus, a general and clear view is necessary of the most important points of interest. As a result, this discussion will consider if the impact of singular features stays the same when combined with other attributes.

6.1 Application of results and post processing visualization

The way the results in this paper differ from the research reported on in the literature review depends on the fact that previous research mainly focused on distance to singular features and their impact on the specific house price. Meanwhile, our research imposes the focus on a complete dataset of features, their distance and the effect on the accuracy of the prediction model. However, interesting conclusions can be drawn when comparing the results with the literature regarding the importance certain features have on the price and inspecting if these importances remain valid when combined. Find below a graph showing the real estate sold with the log-transformed price as gradient (figure 17). This graph is compared to other graphs showing location features through this section.

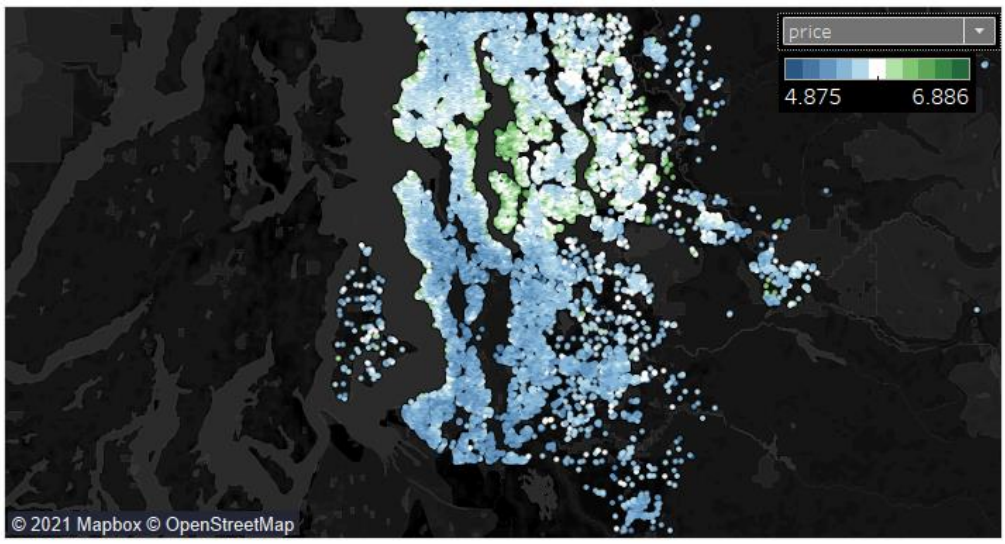


Figure 17: Price graph used for comparison to other POI-category graphs

In terms of transportation Limlomwongse Suksmith & Nitivattananon (2015) indicated that airports can play an important role in influencing the price but only for houses which are located close enough to witness the negative effects an airport produces. Debrezion, Pels & Rietveld (2006) specified that train stations can positively affect the price up to 25% when fairly close by while, for negative effects the distance should almost be zero. Still, the problem with these kinds of points of interest is that the effect on real estate prices can be big but only for houses which are in the immediate vicinity of such POI. This can be confirmed by table 5, with train stations and airports not having a significant impact on the prediction model accuracy and furthermore being of little to no importance in the combined model in figure 15. It should however be taken into account that, while not appearing at the top of the importance list, the impact of transportation POI on a smaller scale of housing samples will be greater.

Lingering in the environment of points of interest which, on a small scale, can have a big impact on price we have junkyards, adult entertainment, stadiums & arenas and prisons. As mentioned in the literature these kinds of points can have major effects on prices for real estate in the immediate vicinity. The results follow the same path as mentioned with the transportation facilities with low improvements of predictive accuracy and little to no importance to the model. Again, these results do not mean that the distance to previously mentioned POI are of no significance. Since, removing the effect of the distance to any of these points actually lowers the predictive quality of the model, meaning that a good combination of features is needed instead of one or two singular ones.

One of the bigger POI introduced in the literature is education. Rivas et al. (2019) and Wada & Zahirovic-Herbert (2013) showed that distance to universities and elementary school respectively have a big influence on house price evaluation and thus on our prediction accuracy. Additionally, colleges and universities have a smaller impact due to lower quantities and centralization. Elementary schools score very high in our accuracy improvement table 5, as Wada & Zahirovic-Herbert (2013) already predicted. Furthermore, elementary schools prove to be a very important factor when taking locational features into account, as shown in figure 16. Not only due to smaller distances being very efficient for families with school aged children, but because of probable reduced crime rates as well.

These factors combined with the idea of elementary schools being scattered around every city, makes this one of the most interesting points of interest with regard to house price evaluation. Another point worth noticing is the kind of feature that is important here. Namely, distance to elementary schools within the 2000 meter radius (SUBelementaryschoolsblw2000) being more important than the average rating of these schools (SUBelementaryschoolsblw2000AVG). The conclusion here being that the preference for education lies with how many elementary schools are close by and how closeby the nearest school is. Whilst this may be the case for grade schools, the importance for colleges and universities shows the contrary. Presumably, due to these kinds of education occuring in smaller numbers, the average rating of the facility plays a more important role than the distance to the closest one. The features the model selected connected to this information are the quantity of elementary schools and the quality of universities within 2000m. Comparing figure 17 (price) with figures 18 and 19 we can clearly see that the locations with higher quantities of

elementary schools, as well as locations with universities of higher quality correlate.

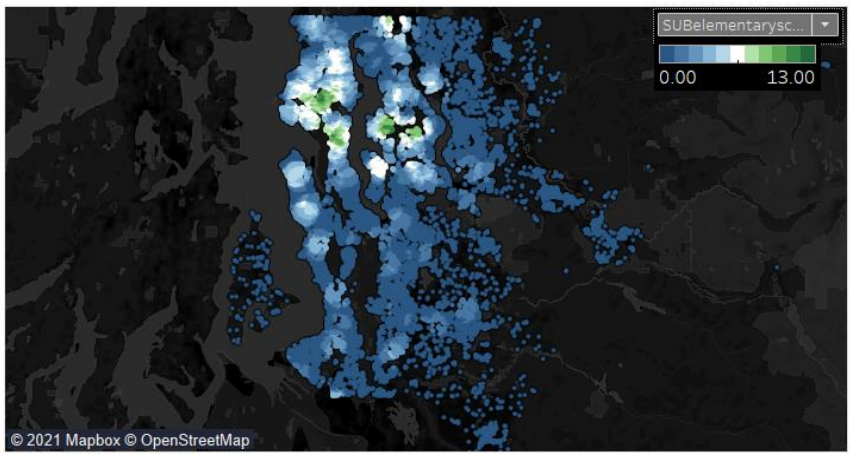


Figure 18: Elementary schools quantity graph

Figure 19: College and universities quality graph

Convenience & grocery stores and pubs are yet again two points of interest of very high importance as explored by Cerrato Caceres & Geoghegan (2017) and Chiang, Peng & Chang (2015). In this paper however, the effect of these points are not very clear. With mediocre improvements in predictive accuracy and low importance in our prediction model one might assume the importance of these POI to be limited, though this should be taken with a grain of salt. The cause may be a number of things such as a big effect on small scale (shorter distances) that fall into the background when examined globally or the difference in importance from convenience store density for lower-priced neighbourhoods to local living quality for higher-priced neighbourhoods is too advanced to be captured in our global model. Consequently, more comparisons of the economic impact of grocery and convenience stores in different urban cities, involving a wider range of density and social contexts, would offer fodder for thought and future research. Figure 20 shows no correlation between higher priced real estate and the number of convenience stores. This variable was still picked as an important feature to the model, which we presume is because they correlate with median prices.

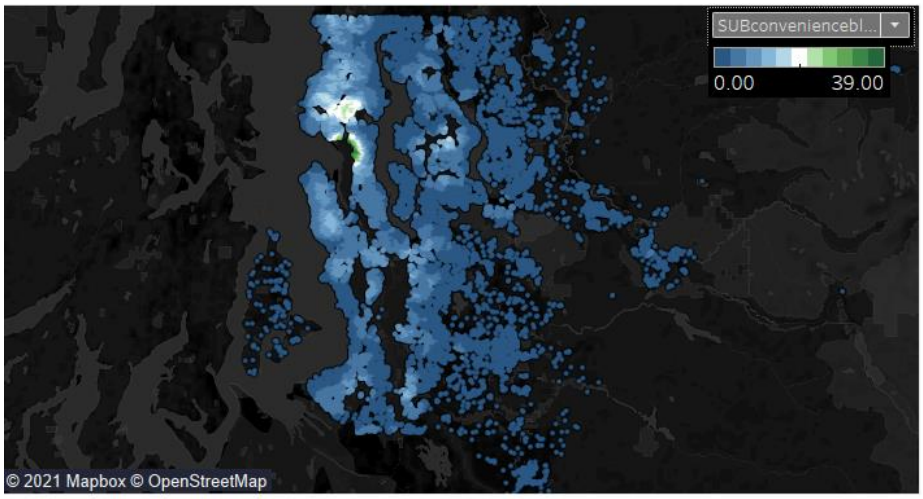


Figure 20: Quantity of convenience stores

Last but certainly not least is the parks variable. As Crompton (2015) indicated, a generalizable answer on the impact a park could have on the house price in its vicinity is not possible due to a number of factors. However, table 5 and figure 15 show that the influence a nearby park has on the price is vast. Topping of the table with the lowest MAPE value and thus being the most important singular feature, parks can be seen as a very important factor when determining house prices. More specifically, in the King County dataset and corresponding predictive model, the distance to the closest park is of more importance than the best average rating of all parks within a 2000 meter radius. Moreover, the magnitude of effect that a park can have on the valuation of housing prices may be due to the fact that parks often cover a great distance, thus potentially influencing more houses in its vicinity.

Our study makes clear that neighbourhood parks produce economic value. They offer important environmental functions, such as solar access and air movement, air pollutant removal, buffers between noise generators and receivers, and habitats to accommodate biodiversity (Jim & Chen, 2010). Government agencies and realtors should be aware of this effect and make all sorts of attempts to increase and improve the urban greenspace stock.

Figure 22 shows the final visualization of select variables annotated to the price plot. To repeat, the gradient would be blue for lower priced real estate and green for higher priced houses. First, prices are lower close to the airports. Then, locations with premium real-estate seem to have many elementary schools in the neighbourhood. And finally, another feature which seems to correlate with the premium pricing is closeness to parks.

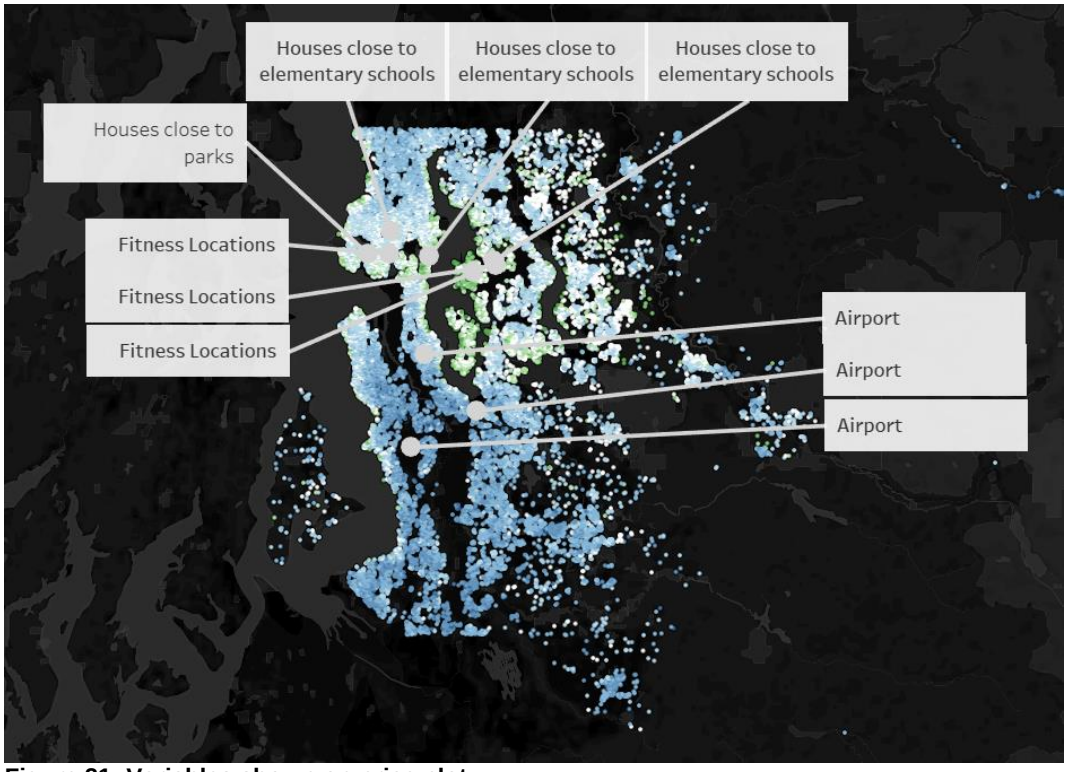


Figure 21: Variables shown on price plot

6.2 Validity

The data was divided into training and testing via a random 0.7-0.3 split. There is a time mismatch between the two datasets as the Yelp database contains data from 2020 while KC real-estate was sold in 2014/2015. Furthermore, we also believe that the construct is complete as the final dataset contained 21613 observations and 274 variables. When downloading the Yelp data the sampling used was 5000 coordinates per category which resulted in obtaining information on the POI even further away as well as having up to 2000 copies of POI per each category downloaded.

There are two main datasets: King County and Yelp. The KC dataset is of such high quality that it does not seem to be seen often in practice. As a result, the combination of both datasets consisted of a large number of variables. As observed in the data exploration, correlation between variables exist. If variables correlate it means that they are close to each other geographically (ie. bars and restaurants tend to be in the center). Therefore, because of the time mismatch it could be that our algorithm prioritizes bars when restaurants actually capture the price better. This could cause a decrease/increase in accuracy/errors when using the same algorithm on another location. We presume that the same variables would generalize well in another location as we tested a large number of variables (14 categories and 13 subcategories).

6.3 Further research and limitations

Using Google Maps (Gmaps) for scraping Yelp data was impossible, which resulted in crippled data. More specifically, to calculate the distances, the Euclidean distance was used. Had Gmaps data been available, not only would the information on Manhattan distance be available, but it would also be possible to observe the distance through time. This would be an even more superior measure than the Manhattan one. For example, how long it takes to get from point A to B via car, uber, bike or other transportation means. The con of using this distance through time is its priciness as each distance is a call to the database. This multiplies for each cell in the distance matrix which multiplies with the number of variables downloaded. We downloaded information on 5642 objects but now, with a proof of concept, we can lower that number and make GMaps more available. Gmaps is also not only more current in some areas but offers additional variables such as priciness, more qualitative information as well as multiple other variables (e.g., does this restaurant offer desserts).

Another limitation of using Yelp Fusion is visualization. This is because R has a package called gmaps which is an extension of the ggplot package. Hence, this package has superior graphing capabilities to tableau because it allows for much more observations.

Another way of extending research is by adding new variable types. We could measure the distance to the closest object to a Yelp category. This variable could be interesting when combined with the airport variable because it would properly capture the effect, as opposed to the number of airports within a range we use currently. Variables like this, those that appear rare, could grow in importance with this treatment. Finally, the Yelp variables which we added do not feature ranges (e.g., all POI-categories between 50-100ms) but only below a certain distance. Adding this information would allow for an additional layer of information. Both of these can also be added to the GMaps option for further research.

General Conclusion

The goal of this thesis was to analyze whether points of interest have an influence on the price prediction capabilities of a random forest price prediction model. We approached this by firstly gathering the points of interests from Yelp and using this data to create information about each category of points of interest. And secondly we evaluated these categories after being introduced into our price prediction model.

Multiple distance features for every category were created to give us an indication of their importance based on a certain distance, ranging from 100m to 2km. As expected not every category contributed as much as some others but the analysis did show us the presence of certain price influencing categories.

Among these categories was parks which single handedly had the biggest impact on the accuracy. These points of interest are no new topic in the world of price prediction but compared to some research like Crompton (2015) parks do have a prominent influence on the price for their various purposes, economic and environmental (Jim & Chen, 2010). Next to parks we see that educational categories like universities and especially elementary schools also have an impact, different from each other yet both important, as seen in the literature Wada & Zahirovic-Herbert (2013), Rivas et al. (2019). The difference lies in the importance of the distance variables and average rating of the category within a certain distance. Where the distance to elementary schools is clearly more important, the opposite is seen for colleges and universities where the rating seems to be more important.

Sadly not every presumed insight from the literature can be found in our analysis due to our limitation to look on a smaller scale with a higher granularity. These categories include convenience and grocery stores, pubs, junkyards, stadiums and arenas and transportation facilities. All of which have been mentioned in previous research (Debrezion, Pels & Rietveld 2006, Ready, 2010, Brooks, Humphreys & Nowak, 2018) which showed a significant importance but this is where our model was not of high enough granularity to notice these big impact categories on small vicinities.

List of figures

List of tables

Sources

Articles

Brooks, T., Humphreys, B., & Nowak, A. (2018). Strip clubs, “secondary effects” and residential property prices. *Real Estate Economics*, 48(3), 850-885. doi: 10.1111/1540-6229.12236

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., & Grisel, O. et al. (2021). API design for machine learning software: experiences from the scikit-learn project. Retrieved 8 May 2021, from https://arxiv.org/abs/1309.0238

Cerrato Caceres, B., & Geoghegan, J. (2017). Effects of new grocery store development on inner-city neighborhood residential prices. *Agricultural And Resource Economics Review*, 46(1), 87-102. doi: 10.1017/age.2016.29

Chiang, Y., Peng, T., & Chang, C. (2015). The nonlinear effect of convenience stores on residential property prices: A case study of Taipei, Taiwan. *Habitat International*, 46, 82-90. doi: 10.1016/j.habitatint.2014.10.017

Chica-Olmo, J. (2007). Prediction of housing location price by a multivariate spatial method: Cokriging. *Journal Of Real Estate Research*, *29*(1), 91-114. doi: 10.1080/10835547.2007.12091188

Clapp, John M, et al. (2002). “Predicting spatial patterns of house prices using LPR and Bayesian smoothing.” *Real Estate Economics*, vol. 30, no. 4, 505–532. doi: 10.1111/1540-6229.00048

Crompton, J. (2005). The impact of parks on property values: Empirical evidence from the past two decades in the United States. *Managing Leisure*, 10(4), 203-218. doi: 10.1080/13606710500348060

Debrezion, G., Pels, E., & Rietveld, P. (2006). The impact of rail transport on real estate prices: An empirical analysis of the dutch housing market. *SSRN Electronic Journal*. doi: 10.2139/ssrn.895270

Genesove, D., & Mayer, C. J. (1997). Equity and time to sale in the real estate market. *American Economic Review*, 87(3), 255–269. doi: 10.2307/2951345

Gibbons, S. (2004). The costs of urban property crime. *The Economic Journal*, 114(499), F441-F463. doi: 10.1111/j.1468-0297.2004.00254.x

Hong, J., Choi, H., & Woo-sung, K. (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. doi: 10.3846/ijspm.2020.11544

House Sales in King County, USA. (2021). Retrieved 1 October 2020, from https://www.kaggle.com/harlfoxem/housesalesprediction

Jim, C., & Chen, W. (2010). External effects of neighbourhood parks and landscape elements on high-rise residential value. *Land Use Policy*, 27(2), 662-670. doi: 10.1016/j.landusepol.2009.08.027

Khandelwal, V., Chaturvedi, A., & Gupta, C. (2020). Amazon EC2 spot price prediction using regression random forests. *IEEE Transactions On Cloud Computing*, *8*(1), 59-72. doi: 10.1109/tcc.2017.2780159

Kiel, K., & Zabel, J. (2008). Location, location, location: The 3L approach to house price determination. *Journal Of Housing Economics*, 17(2), 175-190. doi: 10.1016/j.jhe.2007.12.002

Limlomwongse Suksmith, P., & Nitivattananon, V. (2015). Aviation impacts on property values and management: The case of suvarnabhumi international airport. *IATSS Research*, 39(1), 58-71. doi: 10.1016/j.iatssr.2014.07.001

Ma, J., Cheng, J., Jiang, F., Chen, W., & Zhang, J. (2020). Analyzing driving factors of land values in urban scale based on big data and non-linear machine learning techniques. *Land Use Policy*, 94, 104537. doi: 10.1016/j.landusepol.2020.104537

McMillan, M., Reid, B., & Gillen, D. (1980). An extension of the hedonic approach for estimating the value of quiet. *Land Economics*, 56(3), 315. doi: 10.2307/3146034

Ready, R. (2010). Do landfills always depress nearby property values? *Journal Of Real Estate Research*, 32(3), 321-340. doi: 10.1080/10835547.2010.12091279

Rivas, R., Patil, D., Hristidis, V., Barr, J., & Srinivasan, N. (2019). The impact of colleges and hospitals to local real estate markets. *Journal Of Big Data*, *6*(1). doi: 10.1186/s40537-019-0174-7

Tu, C. (2005). How does a new sports stadium affect housing values? The case of FedEx field. *Land Economics*, *81*(3), 379-395. doi: 10.3368/le.81.3.379

Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., & Mueller, A. (2015). Scikit-learn. *Getmobile: Mobile Computing And Communications*, 19(1), 29-33. doi: 10.1145/2786984.2786995

Wada, R., & Zahirovic-Herbert, V. (2013). Distribution of demand for school quality: Evidence from quantile regression. Journal Of Housing Research, 22(1), 17-31. doi:10.1080/10835547.2013.12092070

Xiao, Y., Chen, X., Li, Q., Yu, X., Chen, J., & Guo, J. (2017). Exploring determinants of housing prices in Beijing: An enhanced hedonic regression with open access POI data. *ISPRS International Journal of Geo-Information*, 6(11), 358. doi:10.3390/ijgi6110358

Software:

Yelp Fusion API

Tableau Public 2021.1

Python packages:

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56)

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830

R libraries:

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.3. https://CRAN.R-project.org/package=dplyr

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Hadley Wickham (2020). httr: Tools for Working with URLs and HTTP. R package version 1.4.2. https://CRAN.R-project.org/package=httr

Jeroen Ooms (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:1403.2805 [stat.CO] URL https://arxiv.org/abs/1403.2805.

Robert J. Hijmans (2019). geosphere: Spherical Trigonometry. R package version 1.5-10. https://CRAN.R-project.org/package=geosphere

FACULTY OF BUSINESS AND ECONOMICS
Naamsestraat 69 bus 3500
3000 LEUVEN, België
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
feb.leuven@kuleuven.be
www.feb.kuleuven.be