
# Group Coursework Submission Form

## Specialist Masters Programme

| | |
|---|---|
| **Please list all names of group members:**<br>(Surname, first name)<br>1. Nesvik, Even<br>2. Windsand, Gunnar<br>3. Viermyr, Mikkel | 4. Soh, Avery<br>5. Amirgaliyeva, Dina<br>**GROUP NUMBER:** **8** |
| **MSc in:**<br>Business Analytics | |
| **Module Code:**<br>SMM641 | |
| **Module Title:**<br>Revenue Management and Pricing | |
| **Lecturer:**<br>Oben Ceryan | **Submission Date:**<br>29/03/19 |

**Declaration:**

By submitting this work, we declare that this work is entirely our own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work we acknowledge that we have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. We also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

**Marker's Comments (if not being marked on-line):**

**Deduction for Late Submission:** ____ **Final Mark:** ____ %

Revenue Management and Pricing

SMM 641

# A data driven approach to pricing - Airbnb case study


CASS Business School.

Group 8

Deadline: 17.04.19 12:00PM

# 1 OVERVIEW - EXECUTIVE SUMMARY

Through this case, we explore open data on Airbnb listing on Oslo – Norway. The main purpose is to be able to uncover the underlying aspects listing prices on Airbnb and be able to utilize this in order to predict/recommend prices based on their individual features.

The methodologies used are machine learning techniques like random forest regression which is further explained under part 3. We also utilize simple pricing methods like price sensitivity, demand modelling and willingness to pay.

From the pricing modelling, we can explain a decent part of the variation in price, and we are able to systematically recommend prices based on their features. We choose to go with a smaller part of the originally mined data, as the initial data was not satisfactory enough, however, the results are not affected accordingly

In the demand modelling, we had to make some crucial assumptions as the data does not give us any indication of this measure. We base the calculation on listing time, and number of reviews as a popularity index and average number of renting days in Oslo (3 nights). From this, we are able to deduce an estimated demand in days per listing. we further convert this to a linear function and fit an appropriate model (OLS).

The optimization model is fed the original listing price, recommended price from price-model, estimated demand and data query (data is sorted by neighborhood). The model takes advantage of global price change impacts and leverages that back to the listing in order to adjust a new estimated demand, and finally yields a decision whether the listing should change or keep the listing price.

# 2 INTRODUCTION

## 2.1 ABOUT AIRBNB

Instead of copying other travel sites like Travelcity and Expedia, Airbnb takes a unique approach toward lodging. As one of the recent mastodons in the scaring economy, Airbnb offers someone's home as a place to stay instead of a hotel.

Founded in 2008 Airbnb exists to create a world where anyone can belong anywhere - Airbnb

Airbnb Uniquely leverages technology to economically empower millions of people around the world to unlock and monetize their spaces. They accommodation marketplace provides access to 6+ million unique places to stay in more than 81.000 cities and 191 countries (Airbnb, 2018)

## 2.2 PROBLEM STATEMENT

As a part of SMM641 (Revenue Management and Pricing), we aim to help airbnb hosts (both current and new hosts) to price their listings such that they maximize revenue. Utilize open-source data from Airbnb, create a price recommender engine based on a demand function from segments/classes of hosts that share similar attributes.

### 2.2.1 Overall Business Goals

1 Explore and utilize the data available at airbnb in a pricing environment

2 Explore the pricing at airbnb and uncover what are the significant factors of a given listing.

3 Model demand, based on given data

4 Build a decision model based on the above criteria's

In this project, we aim to deliver a functional model that allows airbnb host to optimally price their listings. While we strive for accuracy, our core and underlying objective remains to produce a baseline optimal price recommender engine that may be easily improved in the future should one intend to add more variable/data or new features. We will base our model with logical reasoning backed with economics theorum *(Price elasticity, Price and demand relationship)*.

#### 2.2.2 Analytics Goals

1 Good coding practice - aim to make the codes readable for different(non-analytical) users. This document should explain the coding outcome from each Python Notebook file linked

2 Estimating the underlying value (Be it the baseline price of a Demand function)

3 Utilise Machine Learning tools and pricing theories taught during the course. To our best abilities we intend to be mindful of statistical assumptions.

## 2.3 ASSUMPTIONS

##### *2.3.1.1 Overall model*

An Airbnb listing's competitor is defined as other listings on Airbnb rather than hotels. The assumption is that prospective buyers have already decided to use Airbnb services and consequently are only looking at listings in the area and not weighing prices to hotels.

Partial rentals are not deducted, meaning that different parties cannot simultaneously book the same listing, even if their combined parties do not exceed that listing's capacity. One Airbnb booking is allowed per rental period - based on Airbnb's policy.

Sellers will only be considered as those people who already own a property and have owned the same property for a reasonable length. This allows for the assumption that the individual seller using our model will be able to procure the necessary data about their listing easily.

We also assume that the price is given in the market. The practical meaning of this is that the listings at airbnb reflect what the market is willing to pay for the respective objects. Price, is, in other words, a product of the features of the listing.

As we will see within our demand model, we have also assumed that listings are assumed to be available 365 days a year.

With regards to cost, we have not accounted for cost as we assume that listing within the same cluster should occur the same cost since they offere simillar amenities, same neighborhood or similar number of beds. Hence, we did not include cost but rather assumed that cost has already been considered as host initially priced their listings.

Host also price listings at one price for the whole year. This also means that we disregard seasonality and changing prices throughout the year. This however can be a feature we add in the future

We also make several assumptions within the construction of the demand function whereby we take "review" variables available in the dataset as proxy for demand. Furthermore, we assume that occupancy rate in Norway is akin to Her Scandanavian neighbours.

# 3 METHODOLOGY

## 3.1 MACHINE LEARNING

Machine learning (ML) is a category of algorithms that allow applications to increase their accuarcy in predicting outcomes without being explicitly programmed. The process invovled in ML are similar to that of data mining and predictive modeling. The goal is to apply a mathematical construct in order to detect and understand trends or structures in the data for future leverage. (SEA, 2016)

## 3.2 DECISION TREES

Decision trees (DT) are non-parametric supervised learning methods, used for both classification and regression (determined, labeld variable, or continious variable). The goal of DT is to create a model that predicts the value of a target variable by learning simple devision rules inferred form the data features. (Sciki-learn, 2017)

Let's look at a simple example. let's say we have data about a basket filled with apples and bananas, and we wanted the algorithm to determin whether we picked an apple or a banana. Intuitively, the algorithm would raise the question: is the object we are dealing with red or yellow? if red; we probably picked an apple.

## 3.3 RANDOM FOREST

A random forest is a supervised learning, meta estimator that fits a number of decision tree regression on various sub sample of the dataset. and uses averaging to improve the predictive accuarcy and control over-fitting. The sub-sample size is always the same as the orginial input

sample. We simply *bag* several Decision trees in order to increase our sercainty about the structure of the data. (Towards Data Science, 2018)

## 3.4 Advantages and Disadvantages of Decision Trees

### 3.4.1 Advantages

- Easy to view the relative importance it assigns to the input features
- Easy to use, as default hyperparameters often produce good results
- Harder to ovefit

### 3.4.2 Disadvantages

- Large number of trees can make the algorithm very slow and inefficient
- Require more computational resources
- less intuitive than for example decision trees

# 4 Analysis and Modeling

The structure we follow in this analysis is a standard data science/analysis process. The full process can be viewed in appendixes, under analysis and modeling. This is a process 10 step process designed to create structure in the analysis work.

Main goals with the modelling process it to:

1 For new hosts - Recommend a price and estimate the corresponding demand. Though that we can estimate the revenue. Of course, the more data we have, the greater accuracy we may get.

2 For already listed hosts - what is their recommended price and would a price change increase or decrease expected revenue. Within the detemined market of operation we identify the corresponding economic price elasticity.

#### 4.1.1 Price modelling

Full price modelling process can be viewed in `modelling.ipnb`

As mentioned, our model only contains price values higher than 200NOK and lower than 200NOK, which represents about 95% of the total data sample. Our goal with this model is not to reviece perfect predicted results, but more to be able to uncover a trend in the data that would indicate that there is a place for such a model. We therefor use MAE and MAPE as decision variable for optimal model and parameters.

$$MAE = \frac{\sum_{i=1}^{n} | y_i - x_i}{n} = \frac{\sum_{i=1^n} | e_i|}{n}$$

and

$$gMAPE(x) = argmin_{g \in g}\ \mathbb{E}[\frac{g(X) - Y}{Y} | X = x]$$

our lowest achieved MAE for the model is 194NOK, meaning that on average we have an error of $\pm$194 NOK per prediction.

After training and testing the model, the next natural question is which variables is important for the decision tree

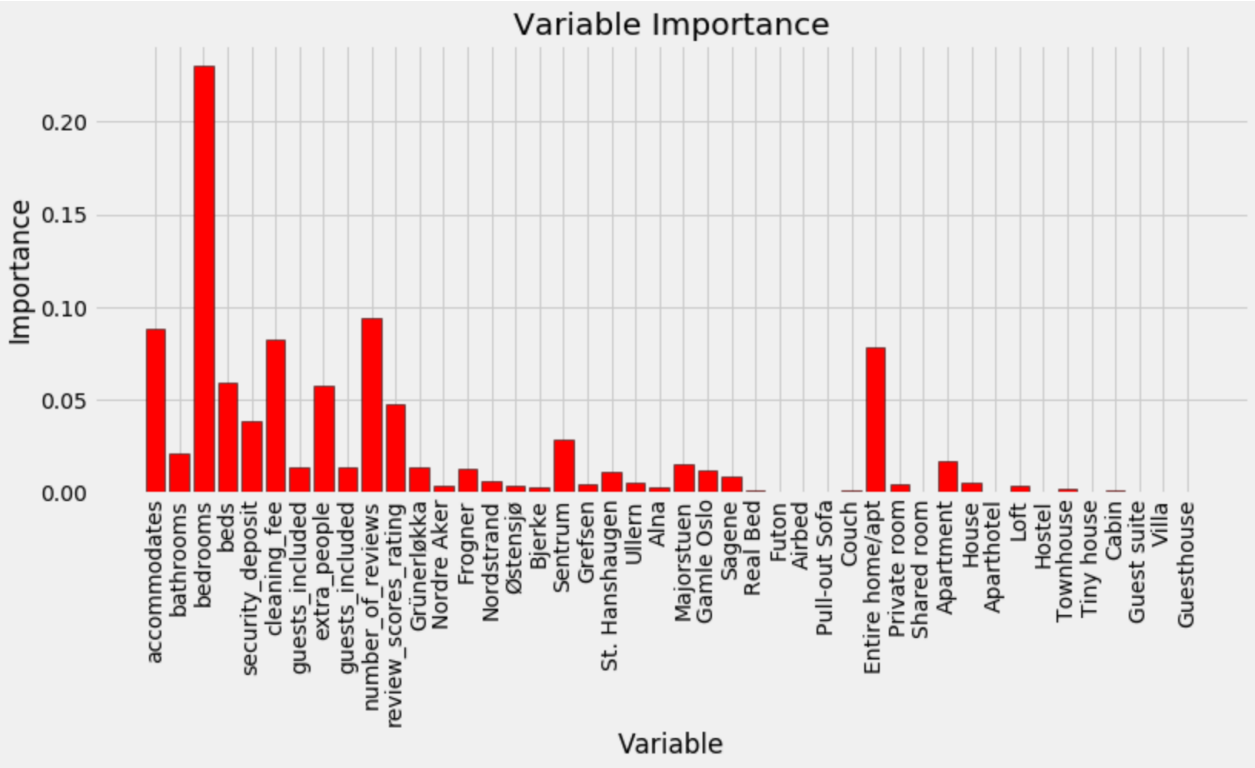


Figure1 - Variable Importance

Here we can see that number of bedrooms sticks out as the most important variable (explainability of about 25% of variation in price) for price, followd by accommodates, cleaning fee, number of reviews and entire home/apt wich are all around 10%.

An interesting observation is that the model seems to undervalue highly priced listings, where the error increases significantly in prices above 1200-1300NOK. This is possibly a problem of information, as the data - and therefor not the model - does not captures the quality of the individual listing in a sufficint way.

## 4.2 Demand Modelling

Full demand modelling can be viewed in `demand_modelling_and_optimization.ipnb`

As the data does not hold any information regarding demand (number of nights rented, number of occupants etc) we have to make an estimate. hence, we use number of reviews and listing time as a proxy for demand. as there was no available data on average occupancy rate in norway, we use Denmark - Copenhagen (15,5%) and Sweeden - Stockholm (14.5%) as a hueristic target of yearly occupancy. Occupancy rate is also set to 365, meaning that all listings are available the whole year.

First we fill in the columns with 0 reviews to avoid `inf~` values (divided by zero), with random numbers between 1 and 4. these objects are listed very late 2018 or early 2019. The model assumes that about 50% of the people rented also place a review. The average length of stay per guest in Oslo is, according to the overview of the Airbnb Community published by Airbnb about 3 nights. Hence, we are gonna assume that each listing has 3 days as an average length of stay per booking.

To prevent unnatural high results, we also considered the maximum occupancy rate could not exceed 25 days, meaning even the busiest of listings will have several nights a month in which they go unrented.

#### 4.2.1 Demand Function

Next, we seperate the data into their respective neighborhoods, as we both assume, and get comfimred from our data exploration, that neighborhoods has an impact on price. By this we can can appriximate a linear demand curve for the individual neighborhood in Oslo. furthermore, we assume that *demand is a function of price*. By doing a quadratic transformation of the demand which consists of aggregate estimated monthly demand at each price point, we get close to a linear relationship between price and demand. Thus

$$\sqrt{Q_i} = \hat{\alpha} + \hat{\beta}_i X_i + \epsilon_i \qquad i = 1, \ldots, n$$

**OR**

$$Q_i = (\hat{\alpha} + \hat{\beta}_i X_i)^2 + \epsilon_i$$

## 4.3 Optimization/Decison Model

The optimazation model have mainly four factors

1 Old price ($P_1$) - initially set by listing $X_i$

2 new price($P_2$) - recommended price from price modelling - $P_i$

3 Old demand ($Q_1$) - demand as estimated from monthly reviews and listing time.

4 New demand ($Q_2$) - demand after leveraging the price elasticitiy from $P_1$ to $P_2$

The model takes the OLS regression results as specefied above, which calcualates the global (specified neighborhood) demand as a function of price. We then calcualte the price elasticity at $P_1, P_2$. Mathematical representation in appendix.

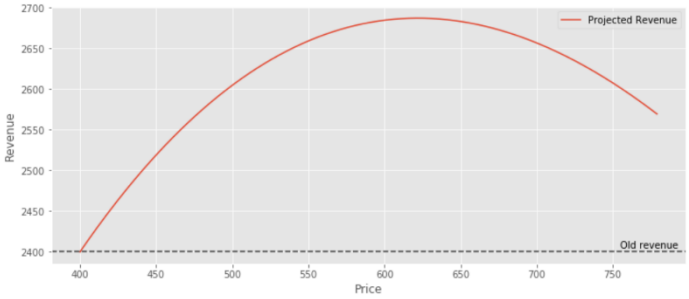
this yield change in demand based on change in price. for example: if $\Delta Q/\Delta P = 0.45$ and if the price changes by 1%, the quantity demanded will change by 0.45%. likewise, if price increases by 10% the demand will decrease by 4.5%.

We then leverage the results from the change in demand(%-change by lowering price) and elasticity, old price ($P_1$) and new_price ($P_2$) in order to calcuate $d_i$ which is the new estimated demand after price change at listing $X_i$. by definition, the change in demand will therefor depend on the angel of the slope in the given market.

## 4.4 EXAMPLES:

| ACTUAL PRICE | 400 |
|---|---|
| RECOMMENDED PRICE | 650 |
| ID | 21557233 |
| NAME | Private room in the real heart of Oslo |
| NUMBER_OF_REVIEWS | 123 |
| SQUARE_FEET | 0 |
| FIRST_REVIEW | 2017-11-04 |
| ACCOMMODATES | 2 |
| BATHROOMS | 1 |
| BEDROOMS | 1 |
| BEDS | 1 |
| SECURITY_DEPOSIT | 0 |
| CLEANING_FEE | 80 |
| GUESTS_INCLUDED | 1 |
| EXTRA_PEOPLE | 200 |
| LOCALTION: | GAMLE OSLO |

```
Unelastic 0.902165764471604
Change to new price
2572.845582781299
optimal price is
       demand      revenue  Optimal Price
222  4.320283  2687.215775            622
```




From this example we can see what the model outputs and some summaries considering the listing. This particular listing which accommodates 2 people, have 1 bed, 1 bath and is located in Gamle Oslo, has an and estimated demand at 9 days, but that changes to around 5 days after increasing the price. This is also close to optimal price for the listing, at 622NOK. Our new expected revenue is 2572NOK.

#### 4.4.1 Comments

One interesting notion from the model is the difference between global and local maximum. according to the estimated values the price that would yield highest profit (globally) is between 600 and 750NOK, under the assumption that anyone willing to pay 1500 is also willing to pay 750 for another night in the same neighborhood.

Another notion is the implication of using estimated montly demand as a constant, meaning that if we imput a monthly demand of 10 days a month, we say to the model that we garantee a occupation of 10 days at old price $P_2$.

#### 4.4.2 Critics of Model:

- We do not consider time as a factor. as airbnb as heavily priced after seasonality factors (holidays, weekends, etc.) the prices produce only represent a static, 1-year price.
- The model oes not adjust for increased popularity at airbnb - so early listings with few numbers of reviews gets undervalues and later listings (2015-2018) with a high number of reviews get over valued in terms of estimated monthly demand
- When users post a new listing, they have an option to fill in property attributes information. While on reality only a small number of attributes are mandatory to fill, the model will not work properly for the listing if the user does not fill in all possible ones as we include optional attributes in the regression model.

# 5 CONCLUSION

We have looked at the price at current state, and potantial profit by price-change given the market dynamics. By this the model was able to conclude wether the host should change the price according to the price recommended by the model.

We have uncovered that there are a few significant factors in the pricing decision. unfortionatly, the list of amenitites functioned mostly as noise, rather than valueable information in determning the quality of the listing, this may have damaged the results of the model, but a MAE of $\pm 194$ is all things considered very desent, which imples that there deffinitly room for such a price model. A model with an unatural low error would suggest that the market functioned perfectly, or that the model was overfitted.

# 6 APPENDIX

**Modelling process.**

1 Determine regired data

- As previosly mention we use open source data from insiderairbnb.com which have a total of 106 variables ranging from name and summary discription of listing, to latitude and longtidute. The data also contains amenities, which is a nested list containing specific feautures of the listing, for examples wifi, TV, Kitchen, etc.

2 Accessible format (preprocessing)

- From purely logical reasoning we deduct a number of variables that would have an impact on price. We also include amenities, which, with some preprocessing are converted to dummy variables. as a result we end up with a total of 142 variables. (detailes can be reviewd in `preprocessing_oslo.ipnb`)

3 Coerrect eventually missing data points/anomalies as required

- No convertion needed.
- after trail and error we realise that most of the anaomalies variabels just create noice in the model, we therefor choose to discard these variables in further analysis.
- We also choose to conventrate on prices below 2000 NOK and over 200 (Roughly £190, £19). For the purpose of this project, values beyond this range we consider anomalies. There are other alternative methods one may use and one of which would be to test for influenctial powers of these anomolies to possible include some (measure via Cook's Distance)

4 Prepare data for machiene learning model

- Split data into test and traning set

5 Baseline - if necessary

- No baseline is created.

6 Train Model

- We train the data on two models:
  - Random Forest regression

- k-nearest neighbors' regression

– They give similar results; however Random Forest regression yields a lower MAE (mean absolute error) than k-nearest neighbors, and is therefor the superior model.

7 Make predictions/recommendations Does the model work as intended?

– Predictions/recommendations are made in test set.

8 Compare results with actual values

– Comparing results

9 if not satisfactory - acquire more data and start from step 2

– results are satisfactory, however, extra data could be interesting to include, e.g., hotel price data and crime data. We can also set certain limitations to ensure our price recommendations are not outrageously high/ low.

10 Interpret and visual representation.

– Interpretation and visual representations created to get an understanding of the random forest model.

**Price Elasticity**

change in price ($\Delta P$):

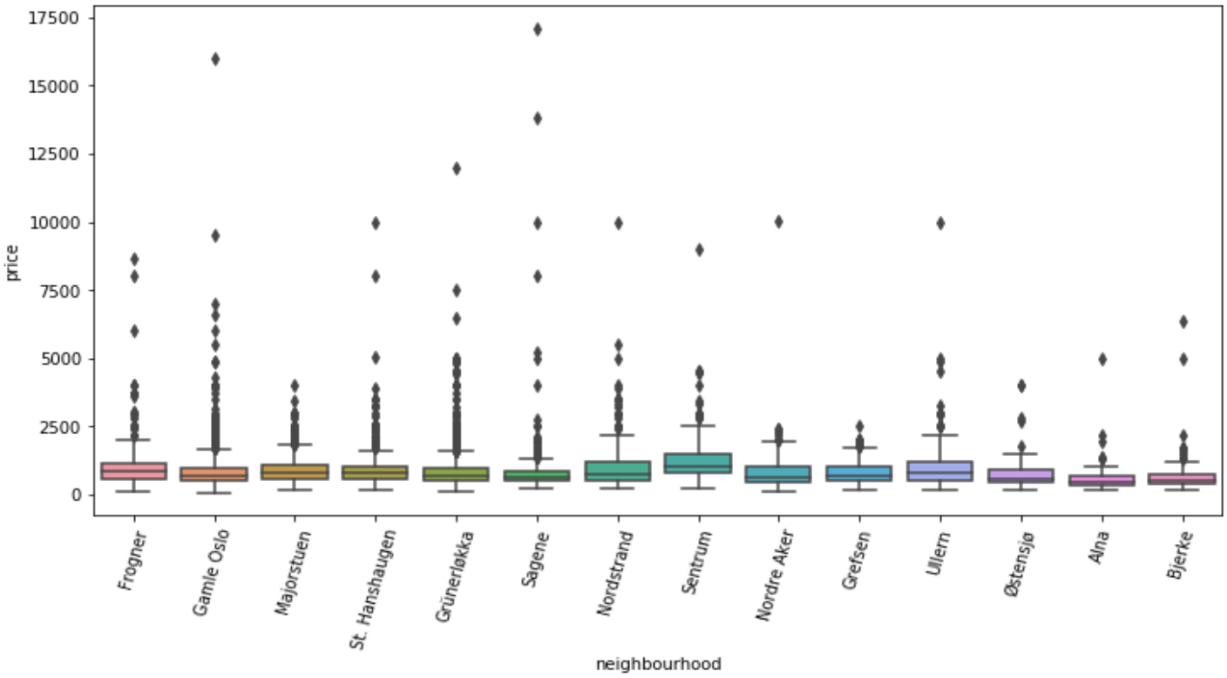
$$\frac{P_2 - P_1}{\frac{P_2 + P_1}{2}} \times 100$$

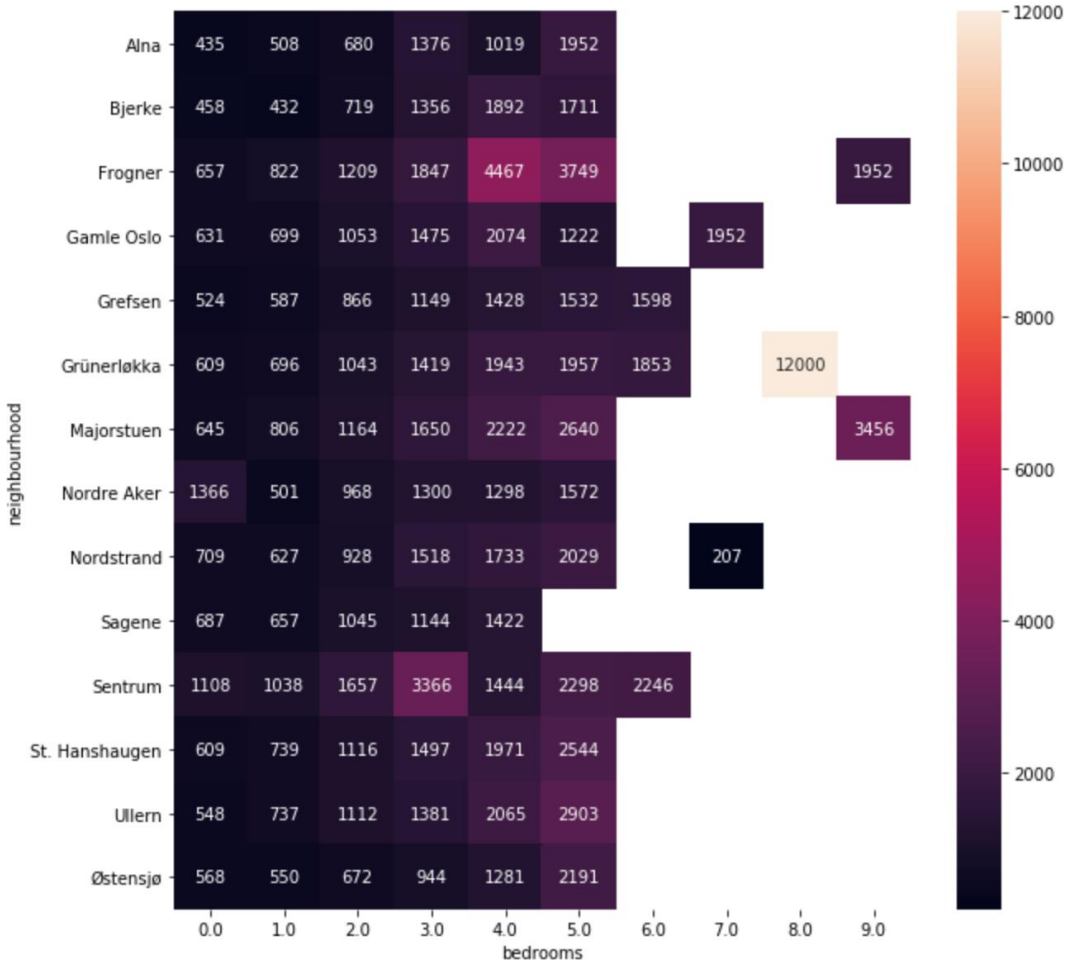
change in quantity ($\Delta D$):

$$\frac{Q_2 - Q_1}{\frac{Q_2 + Q_1}{2}} \times 100$$

Price Elasticity:

$$\frac{\Delta Q}{\Delta P}$$


price
neighbourhood
Frogner
Gamle Oslo
Majorstuen
St. Hanshaugen
Grünerløkka
Sagene
Nordstrand
Sentrum
Nordre Aker
Grefsen
Ullern
Østensjø
Alna
Bjerke


neighbourhood
bedrooms
Alna 435 508 680 1376 1019 1952
Bjerke 458 432 719 1356 1892 1711
Frogner 657 822 1209 1847 4467 3749 1952
Gamle Oslo 631 699 1053 1475 2074 1222 1952
Grefsen 524 587 866 1149 1428 1532 1598
Grünerløkka 609 696 1043 1419 1943 1957 1853 12000
Majorstuen 645 806 1164 1650 2222 2640 3456
Nordre Aker 1366 501 968 1300 1298 1572
Nordstrand 709 627 928 1518 1733 2029 207
Sagene 687 657 1045 1144 1422
Sentrum 1108 1038 1657 3366 1444 2298 2246
St. Hanshaugen 609 739 1116 1497 1971 2544
Ullern 548 737 1112 1381 2065 2903
Østensjø 568 550 672 944 1281 2191
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0
12000
10000
8000
6000
4000
2000

References:

Aribnb – About us available at: *https://press.airbnb.com/about-us/* accessed: 01.04.19

(SEA, *Machine learning – what is it?* Available at:

*https://searchenterpriseai.techtarget.com/definition/machine-learning-ML* accessed: 05.04.19

Scikit learn - *Random Forest regression available at: https://scikit-learn.org/stable/modules/tree.html*) accessed: 06.04.19

Towards data science – *Random Forest*. Available at: *https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd* accessed: 06.04.19

Scikit learn: Random Forest: *https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html* accessed: 03.04.19