Airbnb Prices Prediction

Artificial Intelligence Assignment 2

December 2021

Matteo Giardini

Table of Contents

esade

1

Executive Summary

esade

Executive Summary

Problem

A Dutch Asset Management firm is looking to **invest €50 million in the Amsterdam real estate market.** More specifically they wish to:

- **Restore old buildings** and **list them on Airbnb**
- Distribute their investment across **60/70 buildings** and **reach an 8% ROI**

Solution

To develop an **investment strategy**, **two ML models** have been deployed:

- **Regression/Ensemble:** to understand which characteristics lead to higher prices on Airbnb
- **Clustering:** to divide the listings into different segments and narrow down the investment targets

Impact

5 features with high predictive power have been identified by the regressions/ensembles. These are:

- **name, capacity of the apartment, neighborhood, distance from center, n° of bathrooms**

3 types of buildings, arising from clustering, should be **leveraged to narrow down the scope of the search** of apartments to invest in

€50M investment to be spread over 60/70 buildings to reach 8% ROI

€50M
to be invested in Amsterdam real estate market

8%
Return on Investment

Investment Goal

60/70
buildings

Investment Strategy

airbnb

Goals

- Find out which characteristics of a building contribute to a price increase of the Airbnb listing
- Define a search strategy to decide which buildings will maximize the investment

esade

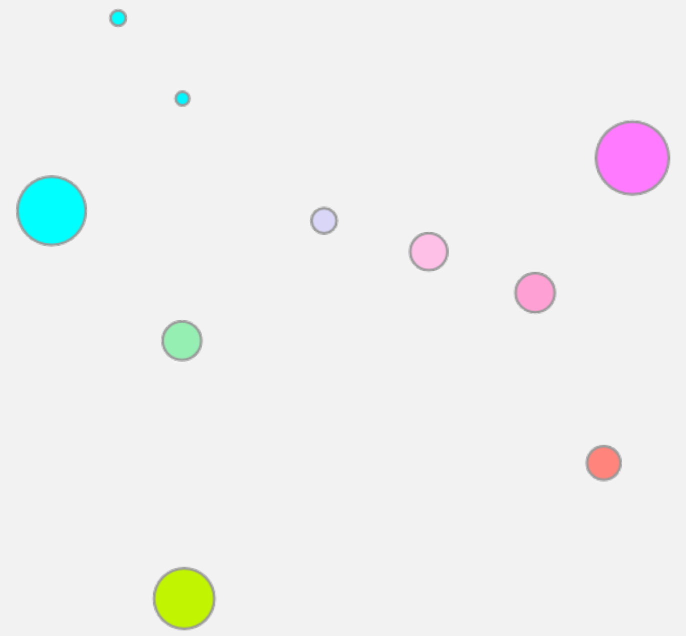
Two machine learning models to be deployed to fine-tune the investment targets

Ensemble

Ensemble	Mean
MAE 33.20	MAE 49.17 -48.11%
MSE 2,923.20	MSE 6,039.58 -106.61%
R-squared 0.52	R-squared 0.00

To understand which **characteristics drive up Airbnb prices**

K-Means Clustering



Divide the listings into different segments and narrow down the investment targets

esade

Search to target estates with 4 characteristics

Based on the information gathered by the regression and ensembles, the asset management firm should **develop a search strategy that targets all apartments with the following characteristics:**

- **Neighbourhoods** → Zuid, Bijlmer-Centrum, Noord-West, De Baarsjes-Oud West
- **Distance to Centre** → City Centre
- **Apartment Capacity** → More than 3 people
- **Number of Bathrooms** → Less than 2 bathrooms

Search strategy to be further refined and target 3 types of apartments

Based on the information given by the clusters, the asset management firm may refine its **search strategy that to target apartments that fall into 3 specific clusters,** as outlined in the table below:

	Characteristics	Neighborhood	Beds	Bathrooms	Price estimate
1	Apartment with view	De Baarsjes - Oud-West	2	1	~€130
2	Apartment with garden	Oud-Oost	1	1	~€122
3	Apartment in periphery	De Baarsjes - Oud-West	2	1	~€120

2

Answers To Exercises

Definition of the Business Problem

1 State the problem in business language. What do we want to improve?

“A Dutch Asset Management firm wishes to invest €50 million[1] in real estate in the Amsterdam area.
In order to reach 8% ROI[2], the firm wants to distribute its investments across 60/70 old buildings[1], which will be restored and listed on Airbnb.
For this, it is deemed necessary to find out which aspects of a property justify higher prices on Airbnb.“

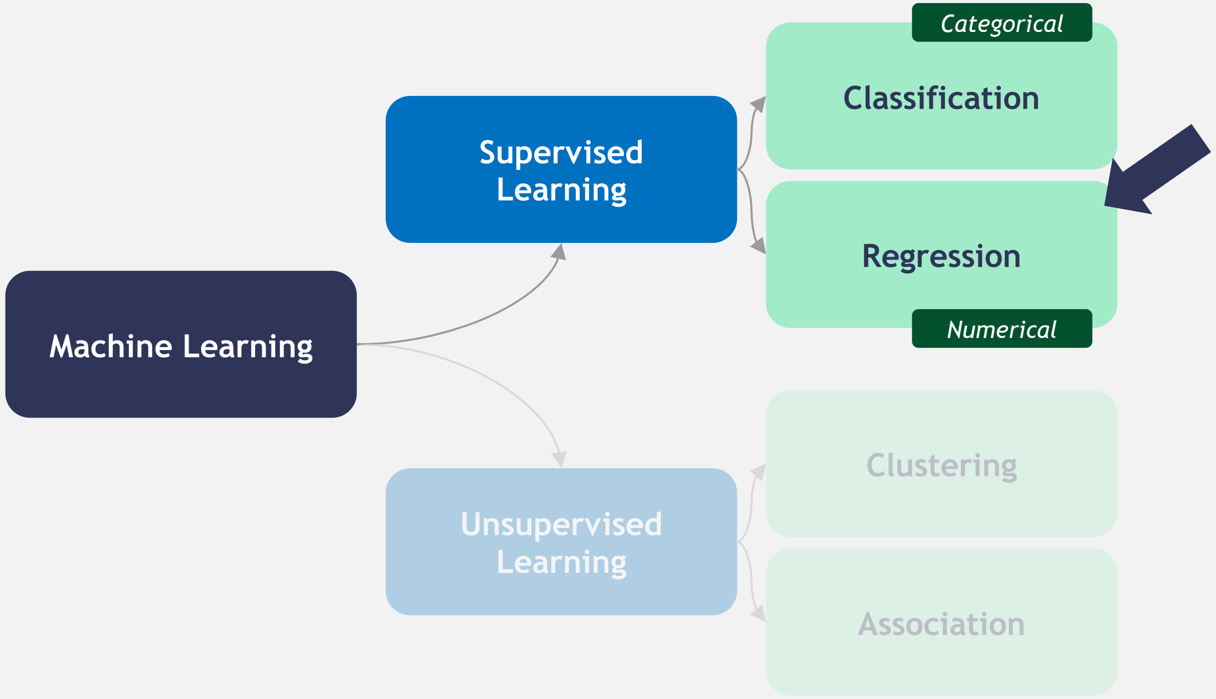
esade

1 Arbitrary numbers
2 https://www.millionacres.com/real-estate-basics/articles/whats-average-return-real-estate-investment/

This problem can be solved with Linear Regression

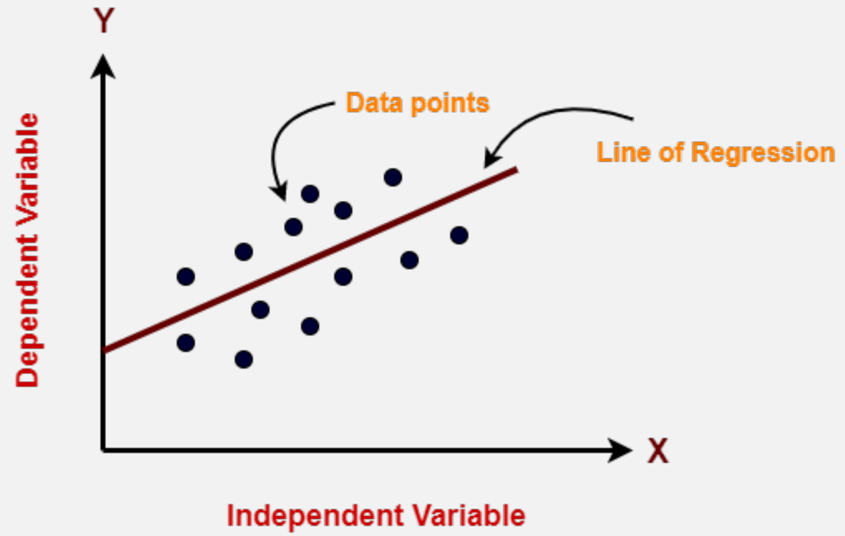
2 ML looks useful to solve this problem. Which kind of ML model do we want to train? Supervised or Unsupervised?
If Supervised, is it a Classification or a Regression problem? Why?

Why Supervised Learning?



Why Linear Regression?

Linear regression has the goal of making a **numerical prediction** based on the relationships of one or more **independent variables.**



Interpretability is key

3 Is the interpretability of the ML model important in this context? Why?

Interpretability of this Machine Learning Model is important in this specific context for **three reasons**:

- To properly evaluate and understand the **accuracy of the model**
- To fundamentally understand how to **improve the underlying algorithms** and finetune prediction
- To detect **which features contribute to the final prediction** to be able to make business decisions

esade

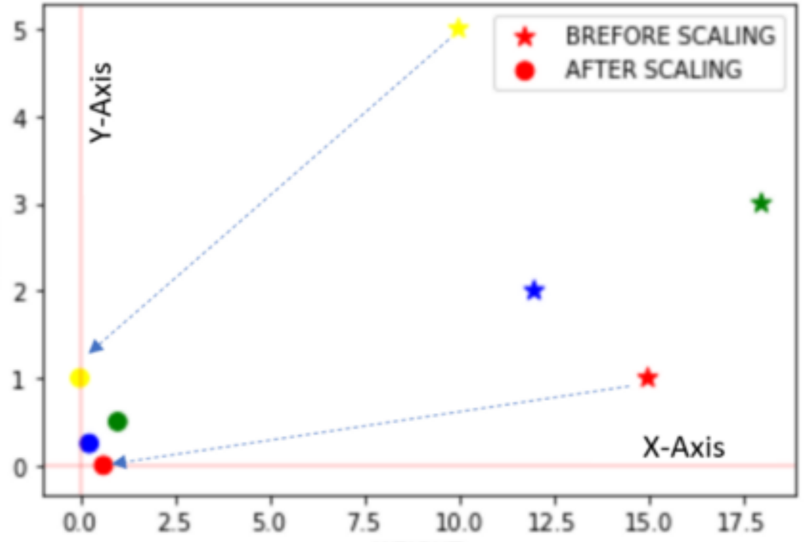
Data Preparation, rescaling of numeric features (0 to 1)

4 Sanity check: Does any variable have a strange distribution? Are there suspicious/corrupted values? Are there missing values or outliers?

Why scaling features?

Numeric features are **rescaled from 0 to 1:**

- To give **equal importance to each feature**
- To facilitate ML algorithm to **process the data**



__Price not to be rescaled__ as it is the prediction target

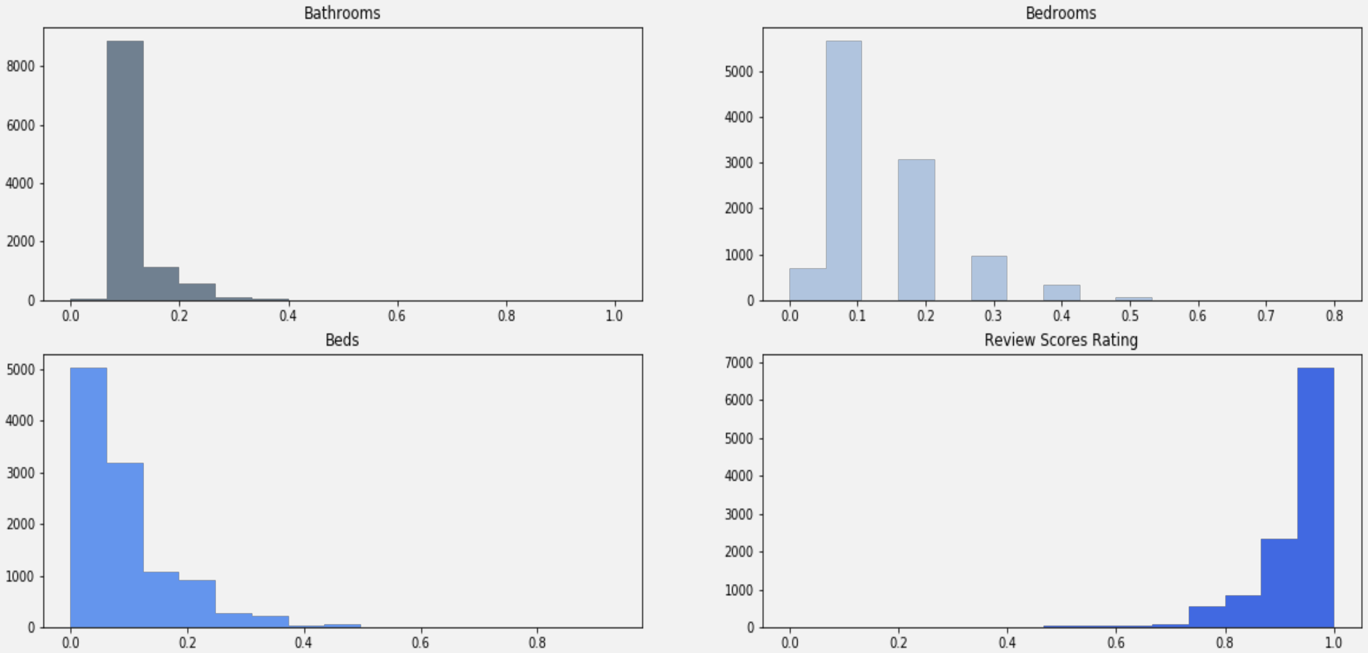
max_capacity(pax)	bathrooms	bedrooms	beds	review_scores_rating	price
0.066667	0.125	0.0	0.066667	0.9375	80
0.200000	0.125	0.2	0.066667	1.0000	129
0.200000	0.125	0.2	0.066667	1.0000	120
0.200000	0.125	0.2	0.066667	1.0000	111
0.333333	0.125	0.1	0.133333	0.9375	251
0.266667	0.125	0.3	0.133333	0.9375	150
0.200000	0.125	0.2	0.066667	0.9750	99
0.066667	0.125	0.1	0.000000	0.5000	55

Data Preparation, handling missing values

4 Sanity check: Does any variable have a strange distribution? Are there suspicious/corrupted values? Are there missing values or outliers?

Features	Missing Values
bathrooms	30
bedrooms	12
beds	10
review_scores_rating	2,624

Feature Distribution



! All missing values have been **substituted with the median** as the distributions are rather skewed. !

Data Preparation, filtering the data set

4 Sanity check: Does any variable have a strange distribution? Are there suspicious/corrupted values? Are there missing values or outliers?

To narrow down the scope of our predictions it is necessary to filter out several **property and room types.**
The Asset Management firm specifically aims at **acquiring old buildings:**

[property_type] For this, we are interested only in estate which can be restored (i.e., apartment, condominium)

[room_type] A private or shared room cannot be purchased by an external third-party

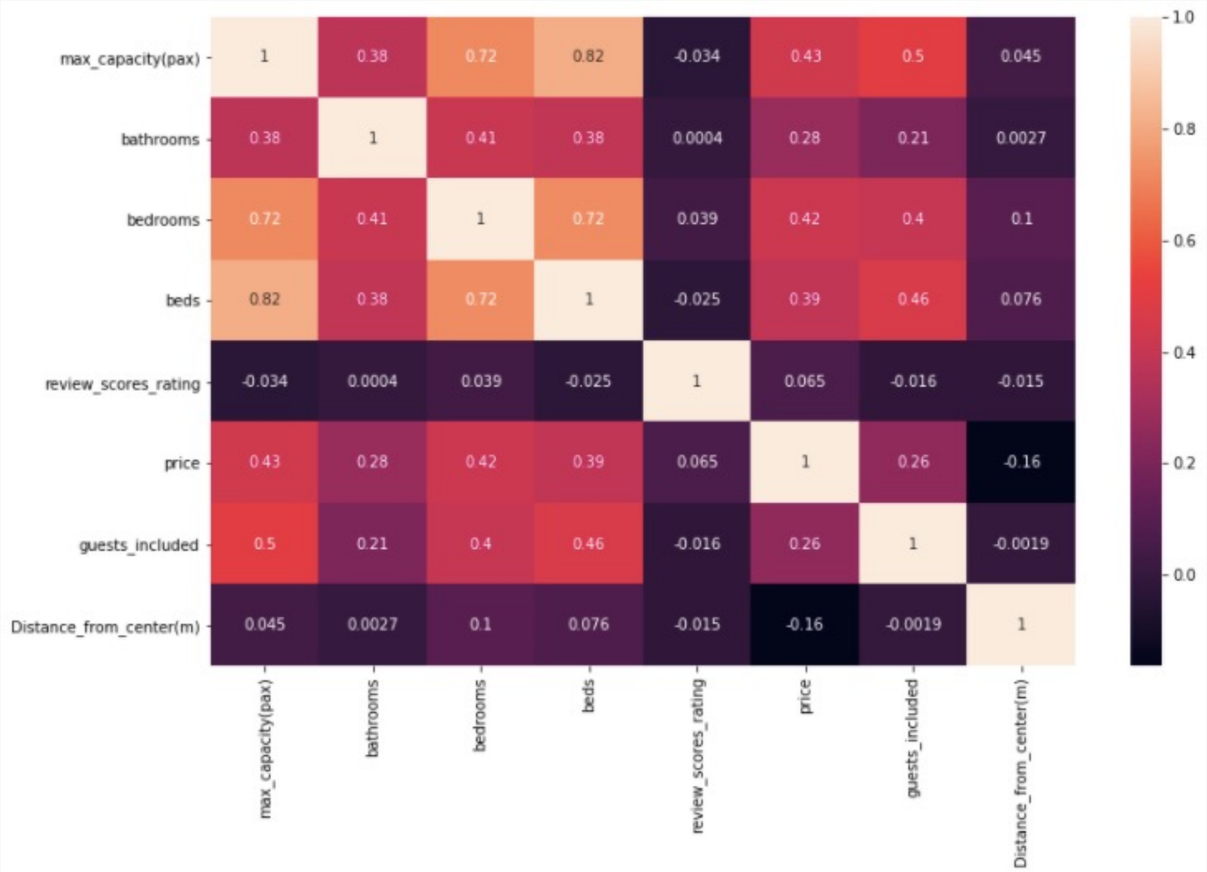
	Unique values	n° features removed[1]
property_type	['Apartment', 'House', '~~Bungalow~~', 'Bed & Breakfast', 'Condominium', 'Townhouse', 'Loft', '~~Boat~~', '~~Cabin~~', '~~Other~~', 'Villa', 'Chalet', '~~Camper/RV~~', 'Dorm', '~~Hut~~', '~~Tent~~', '~~Yurt~~', '~~Earth House~~']	534
room_type	['~~Private room~~', 'Entire home/apt', '~~Shared room~~']	2,577

esade

1 Total features after filtering: **10,738**

Data Preparation, feature removal based on correlation

4 Sanity check: Does any variable have a strange distribution? Are there suspicious/corrupted values? Are there missing values or outliers?



Feature removal

The following correlation matrix clearly displays **high correlation** between:

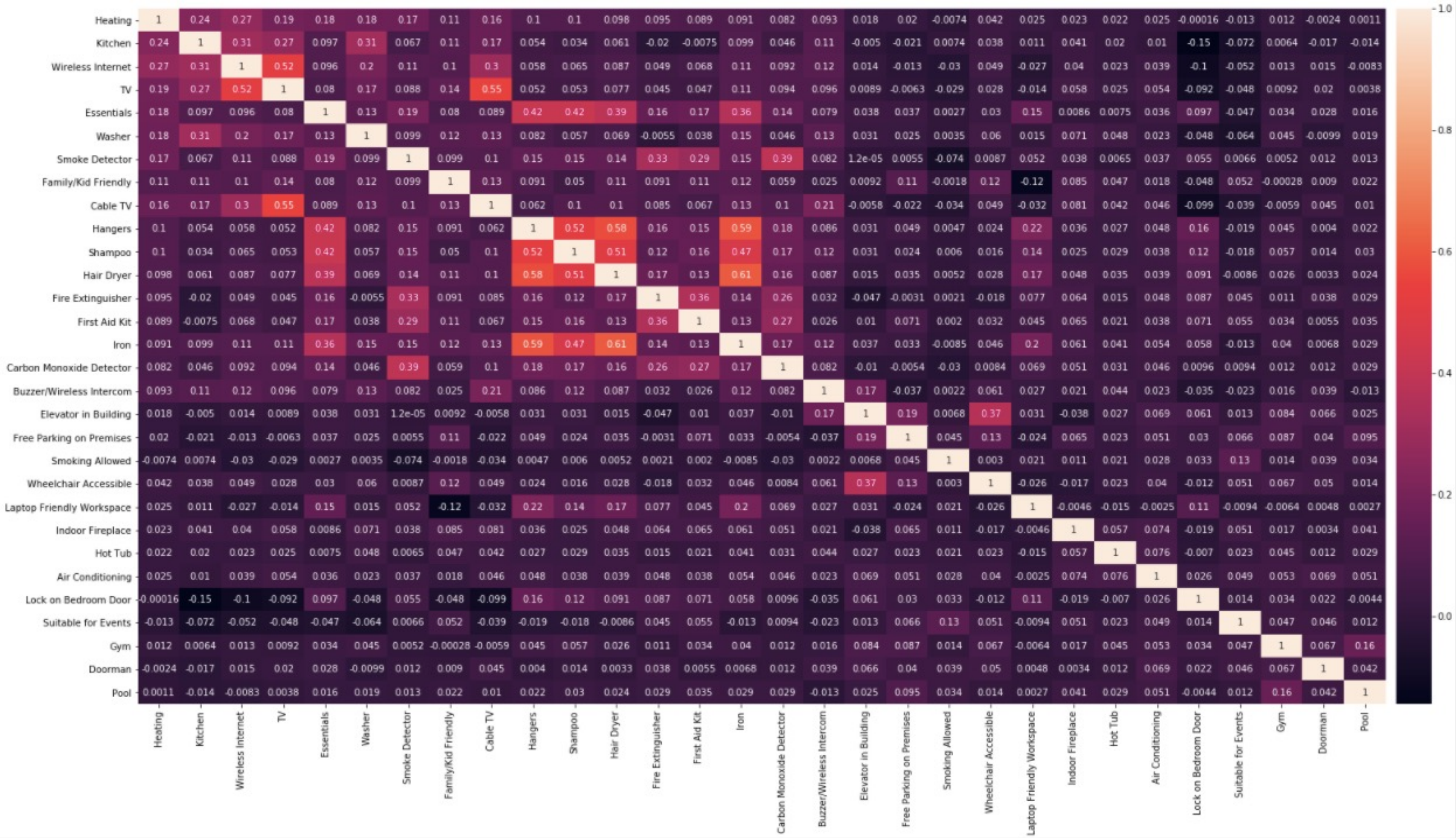
- *max_capacity* & *bedrooms*: 0.72
- *max_capacity* & *beds*: 0.82
- *bedrooms* & *beds*: 0.72

Thus, for now the feature *beds* may be removed.

! At later stages, removing one between *max_capacity* and *bedrooms* may also be appropriate

Data Preparation, feature removal based on correlation

4 Sanity check: Does any variable have a strange distribution? Are there suspicious/corrupted values? Are there missing values or outliers?



Feature removal

The following correlation matrix clearly displays **moderate correlation** between:

- *TV* & *Cable TV*: 0.55
- *Shampoo* & *Hangers*: 0.52
- *Hangers* & *Iron*: 0.72

Thus, for now the features *TV*, *Shampoo and Hangers* may be removed.

esade

Remove review-related features to avoid target leakage

5 Is there any feature causing a Target leakage?

Definition of Target Leakage[1]	**Target leakage** occurs when a variable that is not a feature is used to predict the target. This occurs when the model is built, or trained, with information (known as the training dataset) that will not be available in unseen data

Features	Reason for removal
number_of_reviews reviews_per_month review_scores_rating	**Review-related features must be removed to avoid target leakage as they:** • are generated ex-post with respect to pricing decisions • will not be available for unlisted estates (i.e., potential investments) • are not directly related to price, thus not relevant for price predictions

Source 1: http://downloads.alteryx.com/betawh_xnext/MachineLearning/MLTargetLeakage.htm

(Preliminary) Feature Selection

6 Are all the features useful to predict the target variable?

To **minimize unnecessary efforts during data preparation** and to **avoid noise within the dataset**, several features have been removed early on.

Features	Reason for removal
summary	Superfluous as it contains information present in other features (e.g., location, bedrooms etc.)
host_response_rate, host_acceptance_rate	Features related to Airbnb host, completely disconnected to price (possible data leakage)
cleaning_fee	Added on top of price, should not influence price predictions
cancellation_policy,	Ex-post with respect to price decisions and not necessarily connected to price
‘24-Hour Check-in’, ‘Safety Card’, ‘Pets Allowed’, ‘Breakfast’	Features which may vary from host to host, and could change in case of new ownership

Splitting the dataset in train and test (80%-20%)

7 Perform the train-test split. Which percentages did you choose? Why?

Why 80-20 Split Ratio?

The **split ratio** utilized to divide the dataset into train and test is **80%-20%**, according to best practices

Such a ratio should guarantee that the two datasets are **suitable representations of the main dataset**

The main goal is therefore to **optimize model training** and **maximize prediction performance**

Note: Python code in Technical Appendix

Preliminary Feature Importance Analysis

8 Train a simple model and briefly analyse its metrics and the feature importance.

	Linear Regression	Mean
MAE	35.95	47.17 (-36.78%)
MSE	3,304.87	6,039.58 (-82.75%)
R-squared	0.45	0.00

Insights on most important predictors

By looking at the coefficients and p-values of the predictors:

Neighbourhood:

- Most neighbourhoods, apart from two (Gaasperdam - Driemond, De Aker - Nieuw Sloten) seem to have a **negative effect on price**

Max Capacity, Bathrooms, Bedrooms:

- All three features have **p-values extremely close to 0,** making them really **strong predictors**
- Bathrooms has the **highest coefficient** (>+320)
- Max capacity and bedrooms have **similar coefficients (~**+210)

Distance from centre

- p-value extremely close to 0
- Rather high negative coefficient (-245)
- The further from the centre, the less expensive the listing is

esade

Feature Engineering to increase predictive power

9 Feature engineering: Can new features be created by combining the original ones? If so, why do you think that the new ones can have a higher predictive power?

Feature Engineering on original dataset

1 avg_capacity = (beds + bedrooms + max_capacity) / 3

2 bathrooms_per_person = bathrooms / max_capacity

3 luxury_estate[1] = Hot Tub + Pool + Gym

4 distance_from centre_category →

City Centre	0.0-0.2
Periphery	0.2-0.4
Hinterland	0.4-0.6
Countryside	0.6-0.1

Why might these new feature have higher predictive power?

1. The new features might have **more predictive power** as they may add more value by giving an diverging redundant information coming from different features.
2. By segmenting *distance_to_centre* into 4 buckets the **complexity of the model may be reduced**

esade

1: luxury_estate set to true if 2 out of 3 features contained a 1 for the same listing

Feature selection to improve predictive performance

10 Taking into account the remaining variables, if you find that removing some more improves and simplifies your model, do feature selection.

Features removed and rationale

Predictors with low coefficients

The following features are removed as they have a **coefficient below 2 and high p-value:**

- *Kitchen*
- *Wireless internet*
- *Buzzer/Wireless Intercom*
- *Smoking Allowed*
- *Smoke Detector*

Redundant after Feature Engineering

The following features are removed as they are **already included in the new features:**

- *bedrooms*
- *beds*
- *bathrooms*
- *max_capacity(pax)*
- *Distance_from_centre(m)*
- *Hot Tub, Gym, Pool*

See previous slide on Feature Engineering

Coefficients and p-values of weak predictors

Predictor	Coefficient	p-value
Kitchen	0.4845	0.95
Wireless Internet	-0.4712	0.93
Buzzer/Wireless Intercom	0.5497	0.83
Smoking Allowed	0.5902	0.89
Smoke Detector	1.98	0.51

Already included in features generated through feature engineering:

- *avg_capacity*
- *bathrooms_per_person*
- *Distance_from_centre_categories*
- *luxury_estate*

esade

Outcome of Feature Selection

11 If applicable, did the new features or the removal of some improve the performance of your model?

Metric	Linear Regression	Mean	Linear Regression	Mean
MAE	35.95	47.17 (-36.78%)	33.79	47.41 (-40.30%)
MSE	3,304.87	6,039.58 (-82.75%)	2,425.35	4,853.19 (-100.00%)
R-squared	0.45	0.00	0.50	0.00

Insights

The performance metrics **improved only slightly:**

- MAE: -6%
- MSE: -36%
- R-Squared: +10%

esade

Comparison between different models (I/III)

12 Train another model with a different algorithm and compare their performance.

Comparison between (old) Linear Regression and (new) Ensemble (Boosted Trees)

Linear Regression	Mean	Ensemble NEW	Mean
MAE 33.79	MAE 47.41 -40.30%	MAE 33.20	MAE 49.17 -48.11%
MSE 2,425.35	MSE 4,853.19 -100.00%	MSE 2,923.20	MSE 6,039.58 -106.61%
R-squared 0.50	R-squared 0.00	R-squared 0.52	R-squared 0.00

Training an Ensemble further (slightly) improves MAE and R-squared but increases MSE.
↓ **MAE** -0.2% | ↑ **MSE** +17% | ↑ **R-squared** +4%

esade

5-fold cross validation in next slide

Comparison between different models (II/III)

12 Train another model with a different algorithm and compare their performance **[k-fold cross validation]**

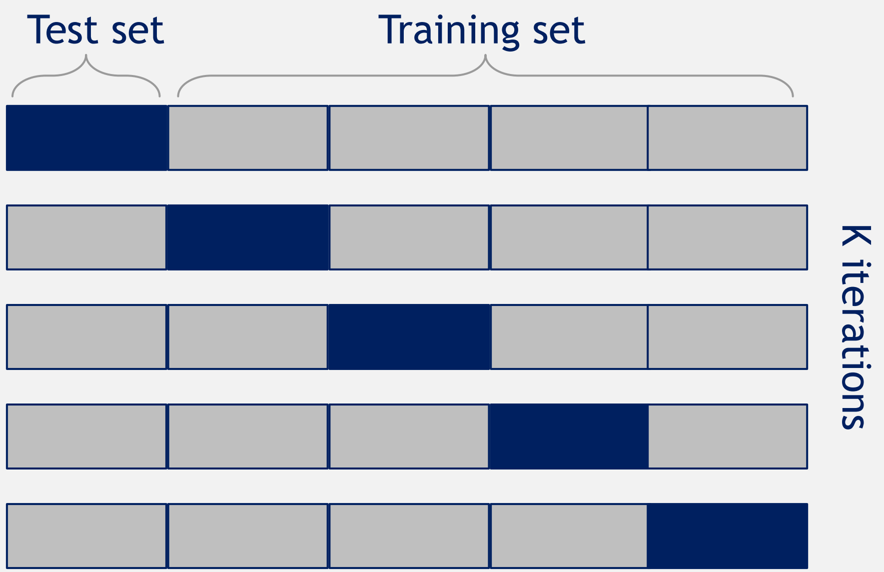
What is k-fold cross validation?

Problem:

When splitting the dataset between test and train, there is always a trade-off between the amount of data included in one dataset or the other.

Solution:

Through k resampling iterations, cross validation allows using the entire dataset both for testing and training.



Comparison between different models (III/III)

12 Train another model with a different algorithm and compare their performance **[k-fold cross validation]**

	Linear Regression	Mean
MAE	34.89	48.44 (-38.85%)
MSE	3,097.30	5,364.38 (-73.20%)
R-squared	0.42	0.00

Insights

i. Running a 5-fold cross validation on the dataset as an overall **negative effect on the performance metrics.**

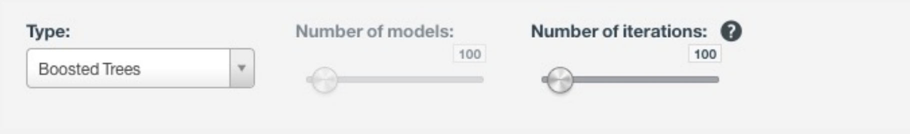
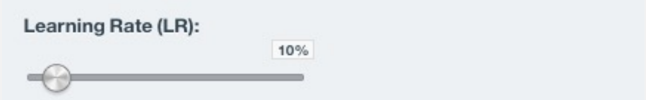
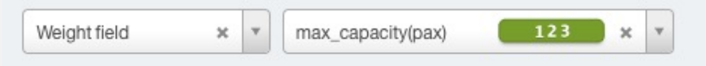
ii. The main goal of cross validation is to **evaluate the model's ability to make predictions with new data** that was not used during training.

iii. As the performance metrics of the model do not improve overall, the model has **low generalizability**.

esade

Finetuned Ensemble with Advanced Hyperparameters

13 Fine-tune your models and try to improve their performance.

Hyperparameters changed		Rationale
Model Type & Iterations		• Boosted Tree was selected over a simple decision tree to maintain high accuracy • 400 iterations to account for the largest # of scenarios possible
Boosting		• Early Holdout to perform the optimal number of iterations by holding out a portion of the dataset each time
Learning Rate		• Learning Rate is maintained at at 10% to avoid overfitting
Weights		• Additional weight is applied on max_capacity as it may be a strong predictor for price
Dataset Advanced Sampling		• A random sampling method is selected to reduce overfitting

All **performance metrics slightly worsen** as compared to previous model; **change is not significant.**

MAE	MSE	R-squared
⬇33.80	⬇3,211.05	⬇0.47

esade

What is overfitting?

14 Are your models overfitted? How did you check this? Why is an overfitted model useless?

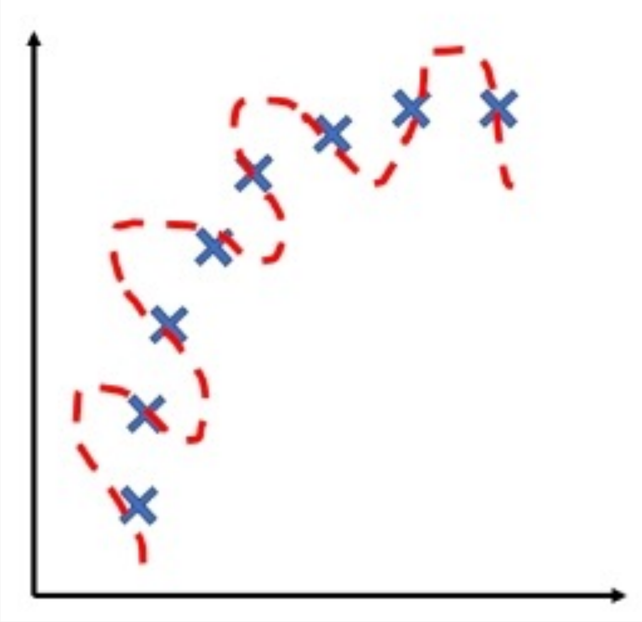
What is overfitting and how to check?

What:

An overfitted model is too specific (namely not general enough) and it does not predict well for unseen data.

How:

To check whether a model is overfitted it is necessary to compare the performance metrics of the evaluation of the test and training sets.



Linear regressions were overfitted, but not the ensemble

14 Are your models overfitted? How did you check this? Why is an overfitted model useless?

Model evaluation against training dataset

	Linear Regression	Mean
MAE	35.48	50.20 (-41.50%)
MSE	10,812.30	13,930.61 (-28.84%)
R-squared	0.22	0.00

Are the models overfitted?

The **linear regressions** performed initially **were overfitted** as:

- The **metrics** stemming from the **evaluation against the training set differ excessively** from those of the **test set**

See previous slides

What to do to avoid overfitting?

Train an ensemble which, by nature, is less likely to cause overfitting

- Based on the **Wisdom of the Crowds**[1] theory the more trees are added the **higher the generalizability**

Already done to answer question 12

1 https://towardsdatascience.com/ensemble-machine-learning-wisdom-of-the-crowd-56df1c24e2f5

Ensemble performance metrics interpretation

15 Interpret the metrics of the models. What represents each metric?

Ensemble	Technical Interpretation	Business Interpretation
MAE 33.20	• Mean of absolute difference of predicted and actual values	• How large of an error we can expect in the predicted prices <u>in absolute terms</u>
MSE 2,923.20	• Mean of squared difference of predicted and actual values • To adequately penalize large errors	• How well the model predicts prices with the given predictors
R-squared 0.52	• Comparison of $MSE_{\text{mean model}}$ with actual MSE_{model}	• How much better does the model predict prices as compared to the mean model

According to best practices, a **good R-squared should be above 0.70**, making this model rather **mediocre in predictive performance**

esade

MSE is the most relevant performance metric

16 Which metric would you pay more attention to if this was a real case? Why?

	Relevance	Rationale
MAE	Low	• Despite being easy to interpret, **MAE excessively emphasises large errors**, making it a rather misleading metric
MSE	High	• The business goal of this model is to **accurately predict price** • Thus, to maximize prediction accuracy, **MSE ought to be minimized** Root Mean Squared Error (RMSE) shall be considered
R-squared	Moderate	• R-squared is effectively **a consequence of MSE**, making it a secondary metric given our business problem

Features with the highest predictive power

17 In the best model, which are the features with the highest predictive power? Why do you think that this is the case?

Features	Importance (%)	Possible Explanation
name	41.39%	• The name encloses (all) relevant information about the offering which defines its attractiveness for guests and, in turn, the price they are willing to pay for
avg_capacity	23.16%	• The price grows with the number of people that the apartment can host
neighbourhood	9.07%	• The price increases if the apartment is located in more exclusive areas or more residential (and quite) neighbourhoods
Distance_from_center_category	5.36%	• The closer to the centre, the more expensive the apartment as it is more likely to be around highly demanded services, restaurants, shops etc.
bathrooms	4.57%	• Similarly to capacity, the more bathrooms in an apartment the more comfortable people feel and the higher the value of the building

! The top **5 most important features** account for **83.55%** of the predictive power of the model. !

Recommended strategy to maximize ROI

18 Based on all the information available, which business activities would you perform to maximize the return on the investment?

Based on the information gathered by the ML models, the asset management firm should **develop a search strategy that targets all apartments with the following characteristics**:

Neighbourhoods	Zuid, Bijlmer-Centrum, Noord-West, De Baarsjes-Oud West
Distance to Centre	City Centre
Apartment Capacity	More than 3 people
Number of Bathrooms	Less than 2 bathrooms

One new categorical and three new numerical variables

19 Which other relevant variables are we missing? Which additional variables would you like to have? Could the model be improved with external data?

Numerical
last_renovation

Number of years since last renovation of the building

Numerical
building_age

How old is the building where the apartment is located

Categorical
public_transport

Connection with transport
- **High** (metro + tram + bus)
- **Medium** (tram + bus)
- **Low** (only bus or none)

Numerical
sq_meters

How big is the apartment in terms of square meters

Integration with external data

- Given the mediocre performance metrics the model could definitely be **enriched with external data**
- Data provided by **real estate agencies** or by the **Amsterdam Geemente**[1] could be integrated

esade

1 Municipality

G-Means Clustering yields k = 19

20 How many clusters does G-Means propose? Which is the main difference between the K-means and the G-means algorithm?

G-Means Clustering Result

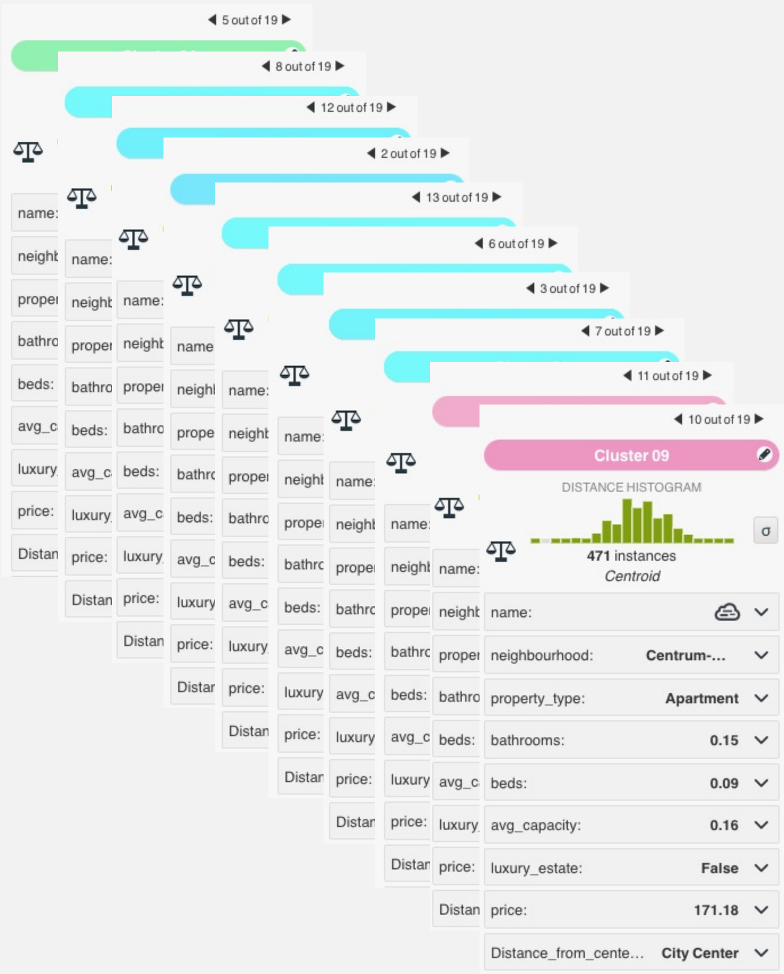
k = 19

Main differences between K-Means and G-Means

K-Means	G-Means
Number of clusters k must be pre-selected	(Optimal) number of clusters k is found by a statistical test
All features are automatically scaled	If data in the proximity of a cluster do not look Gaussian, the cluster is split
Useful when having a number of clusters in mind	Useful when not having a number of clusters in mind

Business analysis of top 10 clusters

21 Which are the main differences between them? Do these clusters make sense in business terms? Do they help you better interpret the dataset?



Features[1]	Interpretation
name	The clusters clearly divided the properties based on the name of the advert (e.g., canal house, apartment near Vondelpark)
avg_capacity	The average capacity for the main clusters seems to range from 3 to 6 people
neighbourhood	The clusters have divided the observations across the most common neighbourhoods being De Baarsjes – Oud-West and Centrum, both Oost & West)
Distance_from_center_category	Similarly to *neighbourhood* most clusters include apartments either in the Periphery or City Center
bathrooms	The majority of clusters have either one or two bathrooms

esade

1 Features with highest predictive power

Top 10 clusters include 83% of observations

22 How many data points contain the clusters, in total?

Cluster #	%	n° instances
00:	3.92%	417
01:	7.95%	845 ●
02:	6.07%	645 ●
03:	4.37%	465
04:	12.15%	1292 ●
05:	6.70%	712 ●
06:	5.89%	626 ●
07:	11.79%	1254 ●
08:	1.83%	195
09:	4.43%	471 ●
10:	5.45%	579 ●
11:	8.83%	939 ●
12:	7.53%	801 ●
13:	2.81%	299
14:	3.10%	330
15:	1.36%	145
16:	4.02%	427
17:	1.48%	157
18:	0.31%	33

● Top 10 clusters

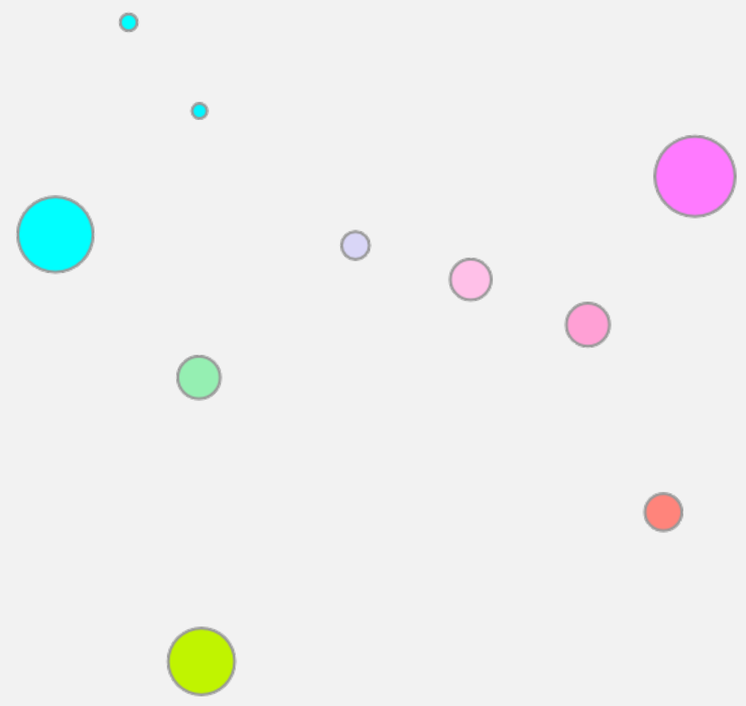
Top 10 clusters (out of 19) include **83% of instances** which amounts to:

8824
instances

K-Means clusters with k = 10

23 Try increasing/decreasing the number of clusters with K-means and choose your preferred clustering configuration. Why do you prefer this one?

K-Means Clustering Result



k = 10

Cluster Analysis

Advantages of k-means

K-Means Clustering with k=10 is better for 2 reasons:

- **Improved interpretability**
- **Lower Complexity**

Custer interpretation

The 3 main clusters outline clear patterns:

- Apartments with view in De Baarsjes – Oud-West, 2 beds, 1 bathroom, priced around €130
- Apartments with garden in Oud-Oost, 1 bed, 1 bathroom, priced around €122
- Apartments in De Baarsjes – Oud-West further from the centre, 2 beds, 1 bathroom, priced around €120

esade

Creating one ensemble for each cluster

24 Create a dataset for each of the clusters and train a model with each of them. Do the metrics of these specialized models improve with respect to the generalist one?

Cluster #	MAE	MSE	R-Squared
01:	39.63	5,084.89	0.22
02:	71.78	7,679.30	0.08
03:	171.28	28,818.60	0.53
04:	42.37	5,136.02	0.11
05:	55.39	5,604.83	0.33
07:	153.43	28,821.62	0.54
08:	39.47	5,071.39	0.18
09:	40.37	5,377.87	0.14
Tot AVG	**76.72**	**11,449.315**	**0.26**
Best Ensemble	**33.20**	**2,923.20**	**0.52**

- The **average performance metrics** of the models arising from the 10 clusters are **substantially lower** compared to the best ensemble previously trained
- However, **some clusters show better performance** and may be analysed individually

Notes: Clusters 00 and 06 have not been included in the table as they contain respectively 18 and 33 observation each, clearly not enough to run a robust model

Search strategy to target 3 specific clusters

25 Do your conclusions from Part 1 change now? Are you now able to propose more and better business strategies to your manager? If so, explain how you improved your recommendations.

Based on the information given by the clusters, the asset management firm may refine its **search strategy that to target apartments that fall into 3 specific clusters,** as outlined in the table below:

	Characteristics	Neighborhood	N° of beds	N° of bathrooms	Price estimate
1	Apartment with view	De Baarsjes - Oud-West	2	1	~€130
2	Apartment with garden	Oud-Oost	1	1	~€122
3	Apartment in periphery	De Baarsjes - Oud-West	2	1	~€120

esade

Why is clustering useful?

26 What do you think about clustering? Is it a useful technique?

Segmenting

Clustering allows to segment a population into more specific subsets

Deeper insights

For each cluster, deeper insights may be gained

High interpretability

Clustering is easy to interpret and very applicable

3

Appendix with Technical Procedures

Appendix: rescaling numeric features (0 to 1)

4 Sanity check: Does any variable have a strange distribution? Are there suspicious/corrupted values? Are there missing values or outliers?

Code snippet example to normalize (only) numeric features

```
from sklearn import preprocessing

to_normalize_cols = ['max_capacity(pax)',
                     'bathrooms',
                     'bedrooms',
                     'beds',
                     'guests_included',
                     'review_scores_rating',
                     'Distance_from_center(m)']
data_to_normalize = data[to_normalize_cols]
x = data_to_normalize.values #returns a numpy array
min_max_scaler = preprocessing.MinMaxScaler()
x_scaled = min_max_scaler.fit_transform(x)
data[to_normalize_cols] = pd.DataFrame(x_scaled)
```

Normalized features

max_capacity(pax)	bathrooms	bedrooms	beds	review_scores_rating	price
0.066667	0.125	0.0	0.066667	0.9375	80
0.200000	0.125	0.2	0.066667	1.0000	129
0.200000	0.125	0.2	0.066667	1.0000	120
0.200000	0.125	0.2	0.066667	1.0000	111
0.333333	0.125	0.1	0.133333	0.9375	251
0.266667	0.125	0.3	0.133333	0.9375	150
0.200000	0.125	0.2	0.066667	0.9750	99
0.066667	0.125	0.1	0.000000	0.5000	55

Appendix: handling missing values

4 Sanity check: Does any variable have a strange distribution? Are there suspicious/corrupted values? Are there missing values or outliers?

Missing values

```
data.isnull().sum()

name                     0
neighbourhood            0
property_type            0
room_type                0
max_capacity(pax)        0
bathrooms               30
bedrooms                12
beds                    10
price                    0
guests_included          0
review_scores_rating  2624
```

Code snippet to show feature distribution

```
plt.figure(figsize = (20,8))

ax1 = plt.subplot(2,2,1)
ax1.hist(data['bathrooms'], bins = 15, color = 'slategrey')
ax1.title.set_text('Bathrooms')

ax2 = plt.subplot(2,2,2)
ax2.hist(data['bedrooms'], bins = 15, color = 'lightsteelblue')
ax2.title.set_text('Bedrooms')

ax3 = plt.subplot(2,2,3)
ax3.hist(data['beds'], bins = 15, color = 'cornflowerblue')
ax3.title.set_text('Beds')

ax4 = plt.subplot(2,2,4)
ax4.hist(data['review_scores_rating'], bins = 15, color = 'royalblue')
ax4.title.set_text('Review Scores Rating')
```

Substitute missing values with median on BigML

Name:		Operation: Replace missing with	Field:
bathrooms	=	Median	bathrooms
bedrooms	=	Median	bedrooms
beds	=	Median	beds
review_scores_rating	=	Median	review_scores_rating

esade

Appendix: filtering the data set

4 Sanity check: Does any variable have a strange distribution? Are there suspicious/corrupted values? Are there missing values or outliers?

Code snippet to filter dataset

```
to_remove = ['Bungalow','Boat', 'Cabin', 'Other','Camper/RV', 'Hut', 'Tent', 'Yurt', 'Earth House']
data = data[~data['property_type'].isin(to_remove)]
```

```
remove = ['Private room', 'Shared room']
data = data[~data['room_type'].isin(remove)]
```

esade

Appendix: feature removal based on correlation

4 Sanity check: Does any variable have a strange distribution? Are there suspicious/corrupted values? Are there missing values or outliers?

Code snippet to plot correlation matrix (I/II)

```
corr_matrix_1 = data[columns_first].corr()
plt.subplots(figsize=(13,8))
sn.heatmap(corr_matrix_1, annot=True)
plt.show()
```

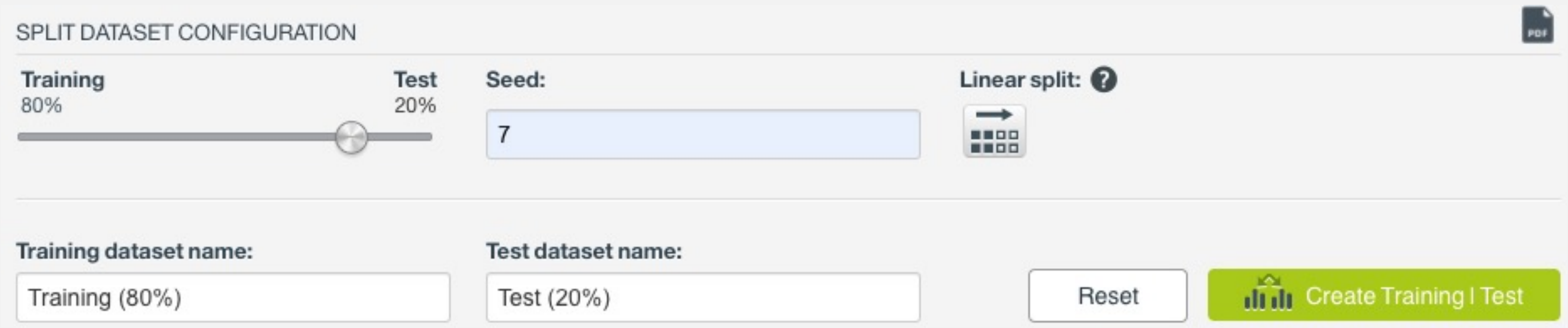
Code snippet to plot correlation matrix (II/II)

```
corr_matrix_2 = data[columns_second].corr()
plt.subplots(figsize=(30,14))
sn.heatmap(corr_matrix_2, annot=True)
plt.show()
```

Appendix: Train-Test Split

7 Perform the train-test split. Which percentages did you choose? Why?

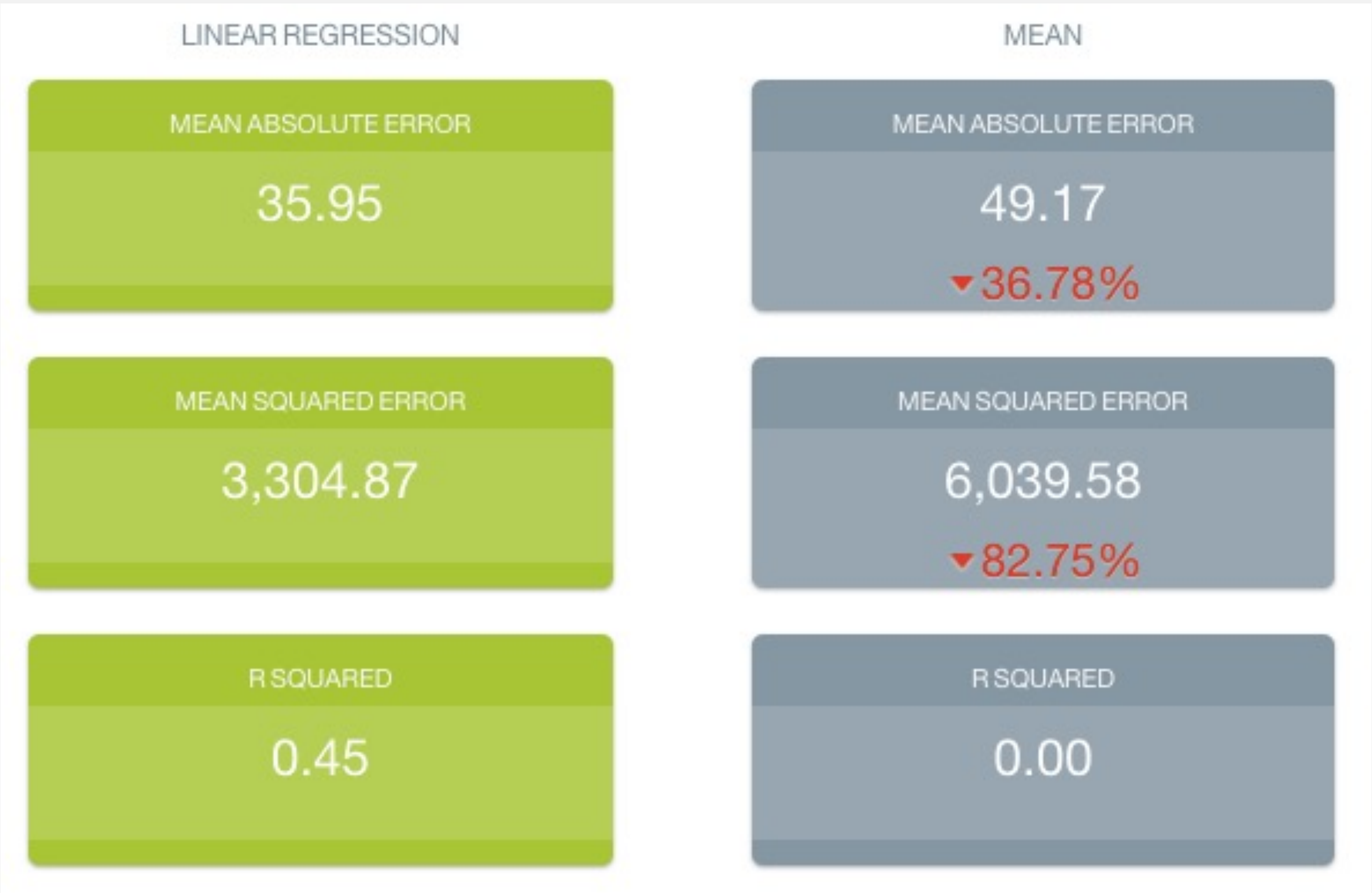
Splitting dataset into train and test



Appendix: First linear regression results

8 Train a simple model and briefly analyse its metrics and the feature importance.

Performance Metrics – Linear Regression Trial 1



Appendix: Feature Engineering

9 Feature engineering: Can new features be created by combining the original ones? If so, why do you think that the new ones can have a higher predictive power?

Code snippets for feature engineering

```
#avg_capacity = (beds + bedrooms + maxcapacity)/3
data['avg_capacity'] = (data['beds'] + data['bedrooms'] + data['max_capacity(pax)'])/3

#bathrooms_per_person = bathrooms / max_capacity
data['bathrooms_per_person'] = data['bathrooms']/data['max_capacity(pax)']
```

```
#luxury_estate = Hot Tub + Gym + Pool
is_luxury = ['Hot Tub', 'Pool', 'Gym']

for i, row in data[is_luxury].iterrows():
    if row.sum() > 1:
        data.at[i,'luxury_estate'] = True
    elif row.sum() < 2:
        data.at[i,'luxury_estate'] = False
```

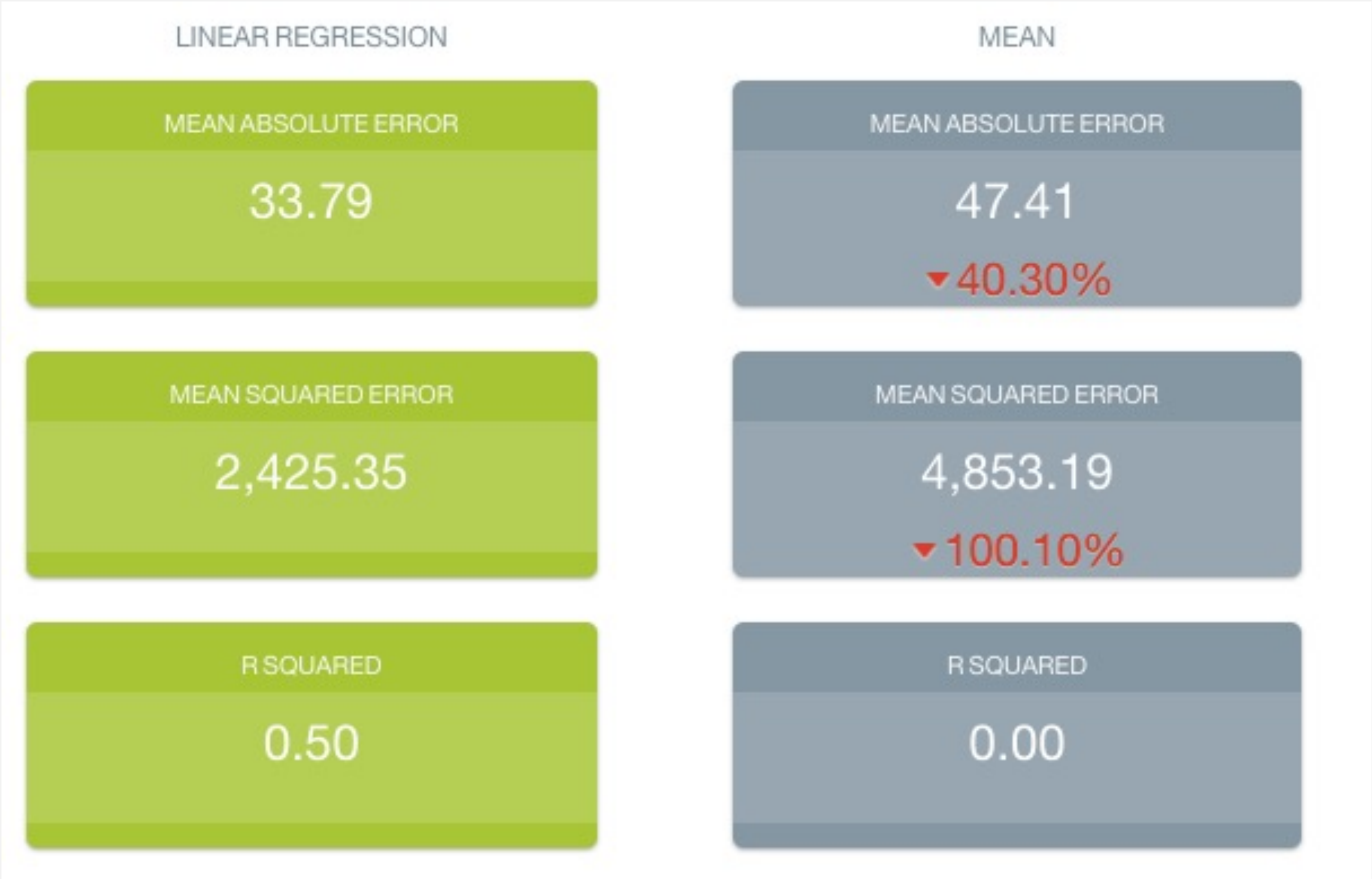
```
cut_labels_4 = ['City Center', 'Periphery', 'Hinterland', 'Countryside']
cut_bins = [0, 0.2, 0.4, 0.6, 1]
data['Distance_from_center_categories'] = pd.cut(data['Distance_from_center(m)'], bins=cut_bins, labels=cut_labels_4)
data['Distance_from_center_categories']
```

esade

Appendix: Outcome of feature selection

11 If applicable, did the new features or the removal of some improve the performance of your model?

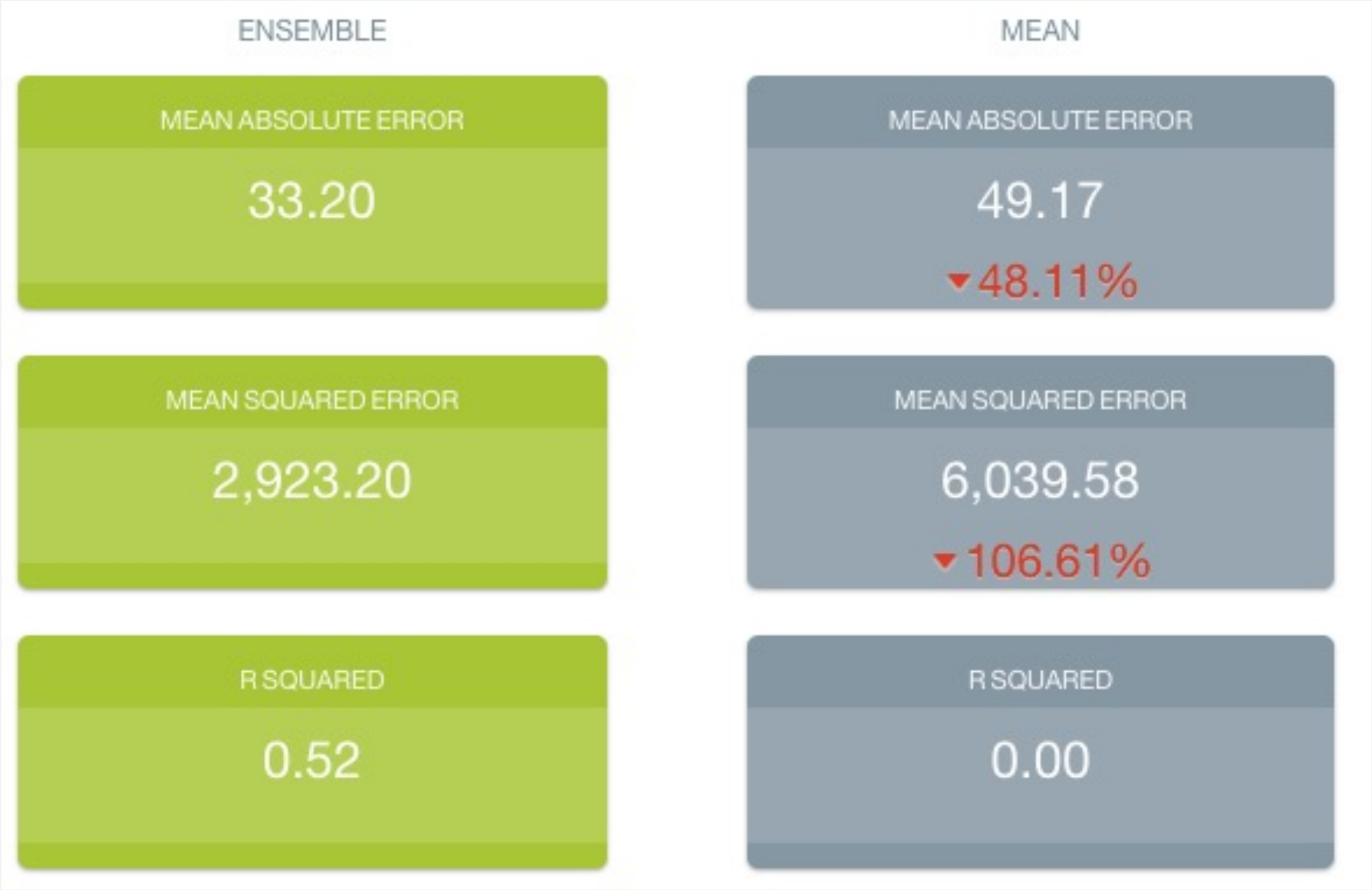
Performance Metrics – Linear Regression Trial 2



Appendix: Ensemble results

12 Train another model with a different algorithm and compare their performance.

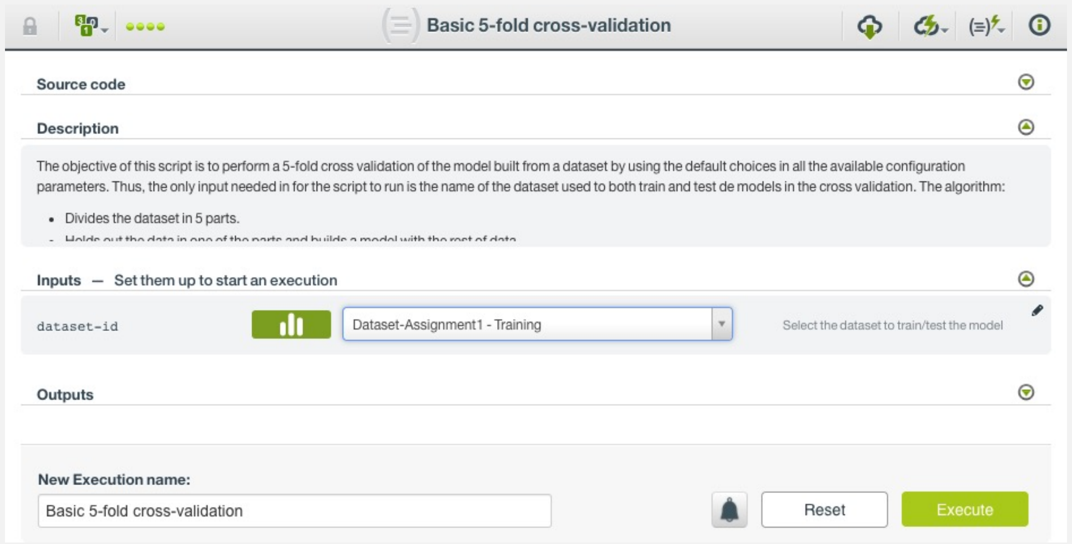
Performance Metrics – Ensemble Trial 1



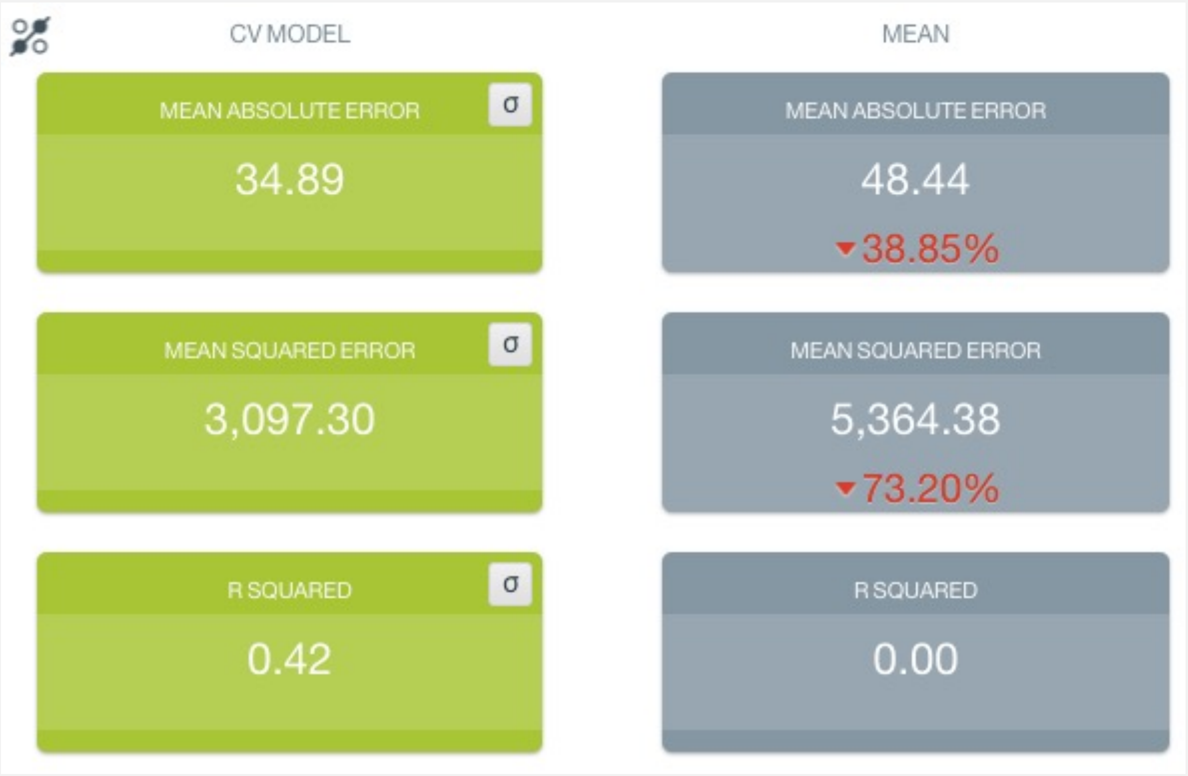
Appendix: 5-fold cross validation results

12 Train another model with a different algorithm and compare their performance **[k-fold cross validation]**

5-fold cross validation configuration



5-fold cross validation average results

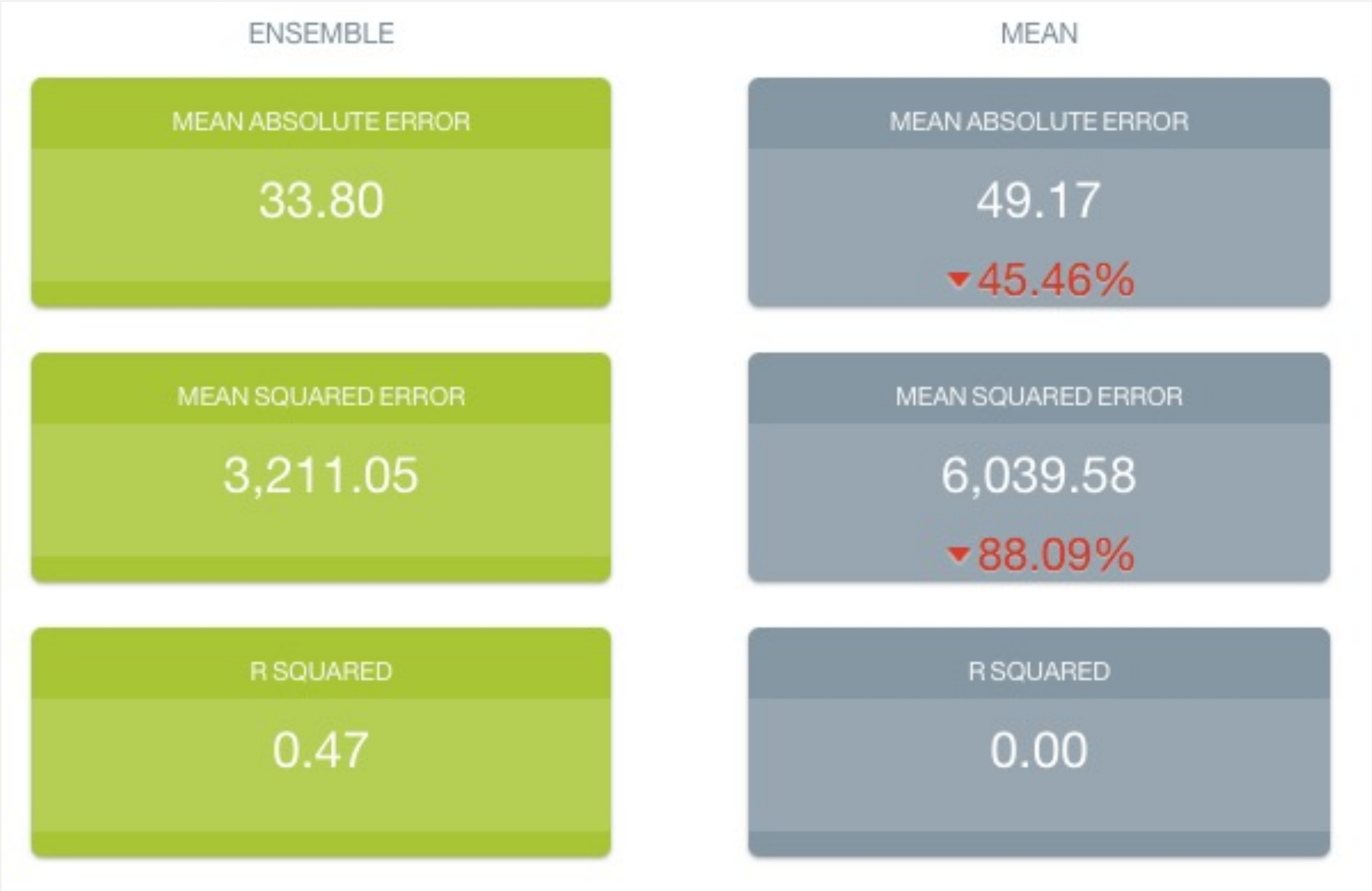


Appendix: Model finetuning with hyperparameters

13 Fine-tune your models and try to improve their performance.

Technical details on Advanced Hyperparameters on answer slide

Performance Metrics - Ensemble Trial 2



esade

Appendix: Evaluating model against test dataset

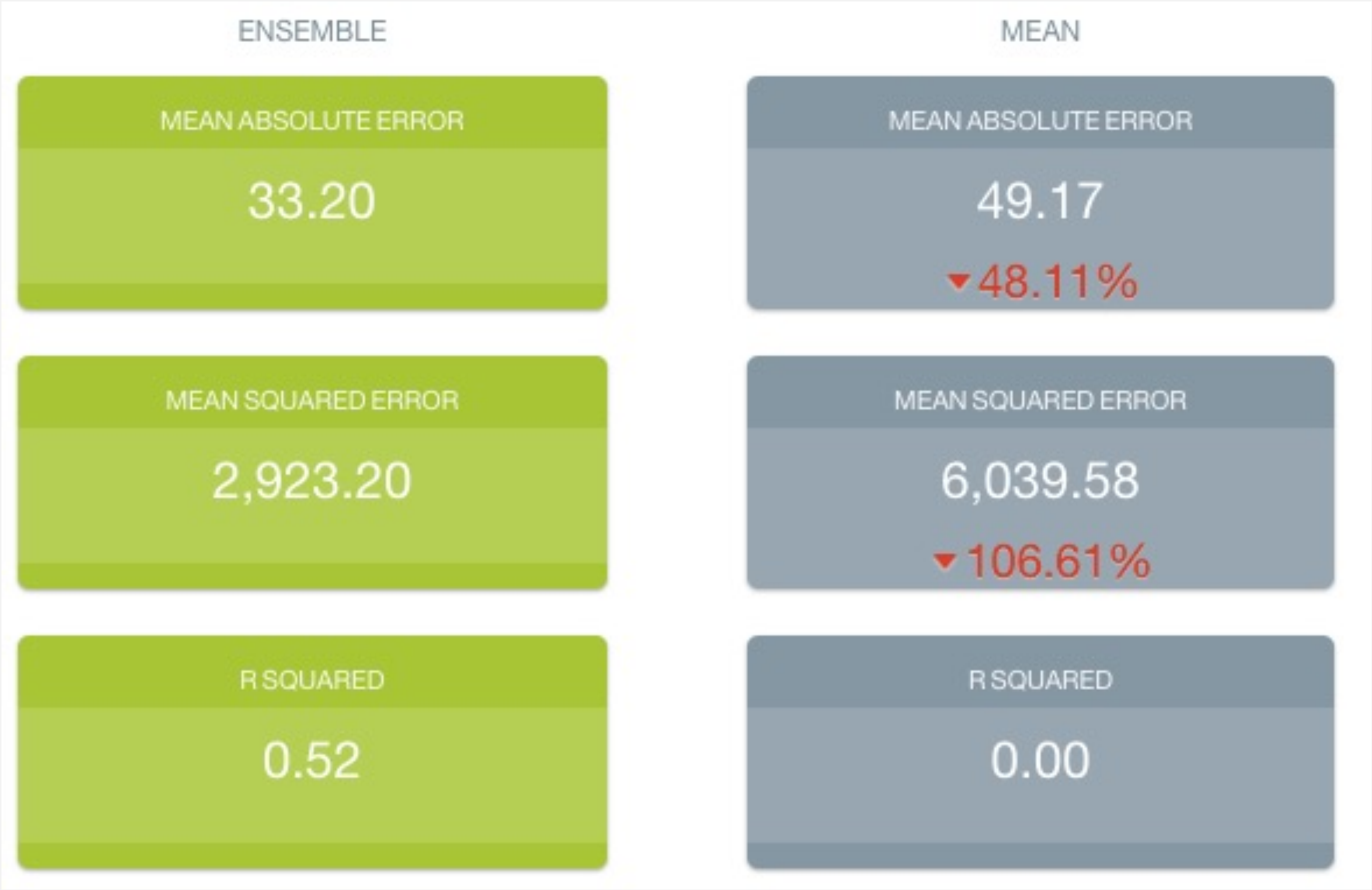
14 Are your models overfitted? How did you check this? Why is an overfitted model useless?

Performance Metrics – Ensemble Evaluation against test set

Appendix: Results of best ensemble

15 Interpret the metrics of the models. What represents each metric?

Performance Metrics – Ensemble Trial 1 (best results)



Appendix: Field importance of best ensemble

17 In the best model, which are the features with the highest predictive power? Why do you think that this is the case?

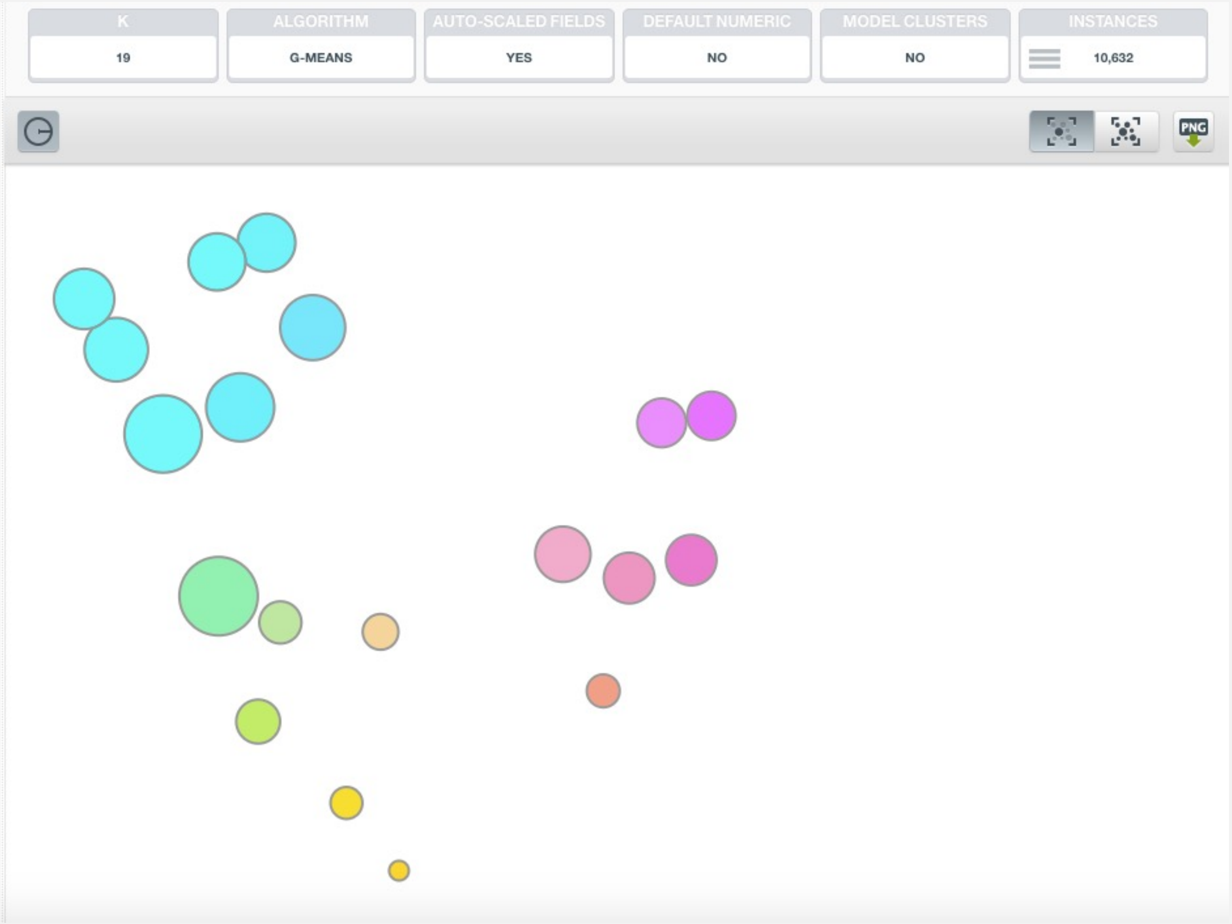
Field importance – Best model

```
Field importance:
    1. name: 41.39%
    2. avg_capacity: 23.16%
    3. neighbourhood: 9.07%
    4. Distance_from_center_categories: 5.36%
    5. bathrooms: 4.57%
    6. bathrooms_per_person: 2.95%
    7. Indoor Fireplace: 2.13%
    8. beds: 1.60%
    9. property_type: 1.50%
    10. First Aid Kit: 0.96%
    11. Fire Extinguisher: 0.93%
    12. Iron: 0.91%
    13. Carbon Monoxide Detector: 0.85%
    14. Elevator in Building: 0.79%
    15. Hair Dryer: 0.72%
    16. Air Conditioning: 0.67%
    17. Essentials: 0.63%
    18. luxury_estate: 0.55%
    19. Free Parking on Premises: 0.41%
    20. Wheelchair Accessible: 0.33%
    21. Suitable for Events: 0.19%
    22. Lock on Bedroom Door: 0.18%
    23. Heating: 0.13%
    24. Doorman: 0.04%
```

Appendix: G-Means Clustering Results

20 How many clusters does G-Means propose? Which is the main difference between the K-means and the G-means algorithm?

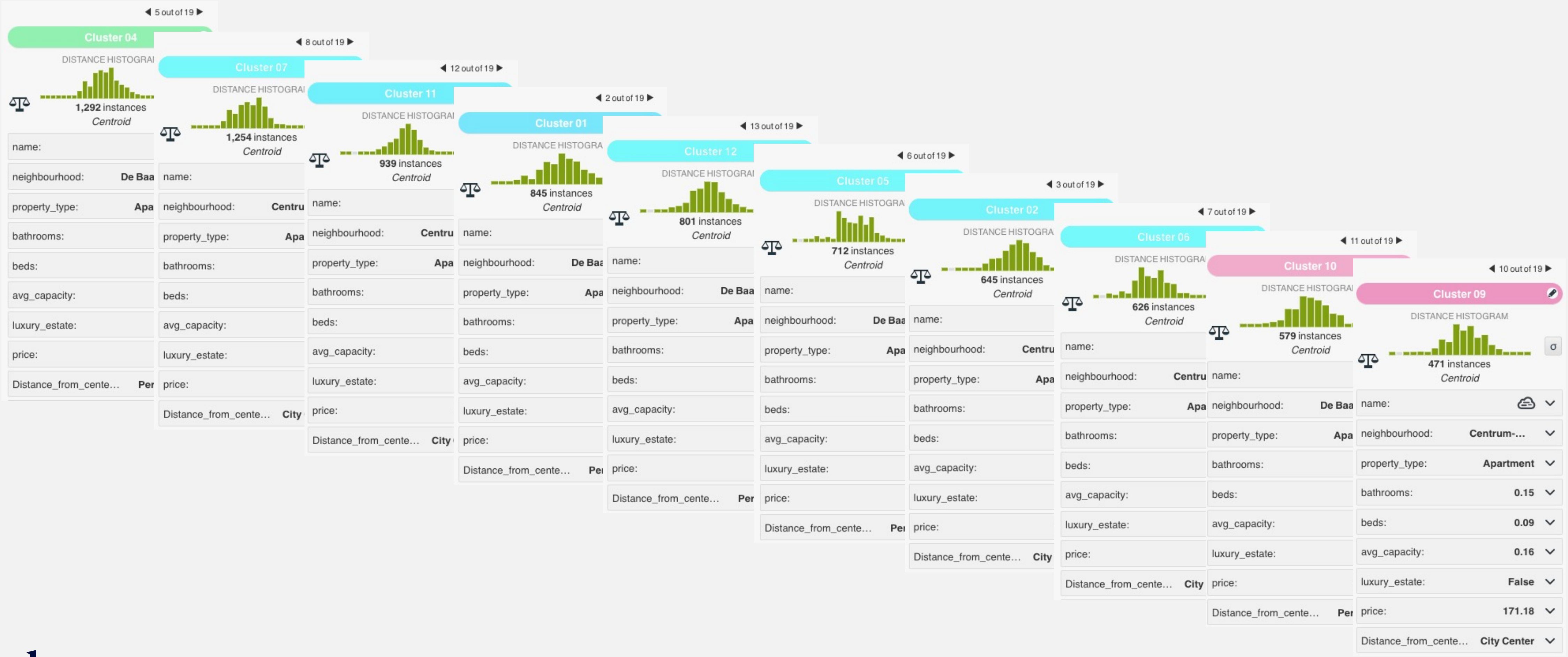
G-Means Clustering Results



Appendix: Top 10 G-Means clusters

21 Which are the main differences between them? Do these clusters make sense in business terms? Do they help you better interpret the dataset?

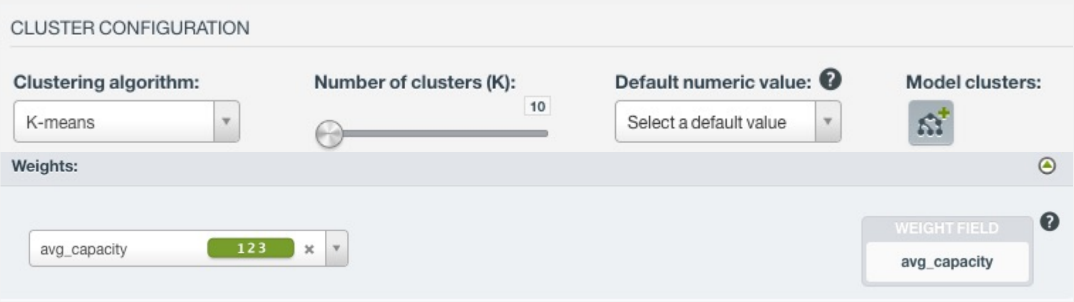
Top 10 clusters by number of instances - G-Means clustering



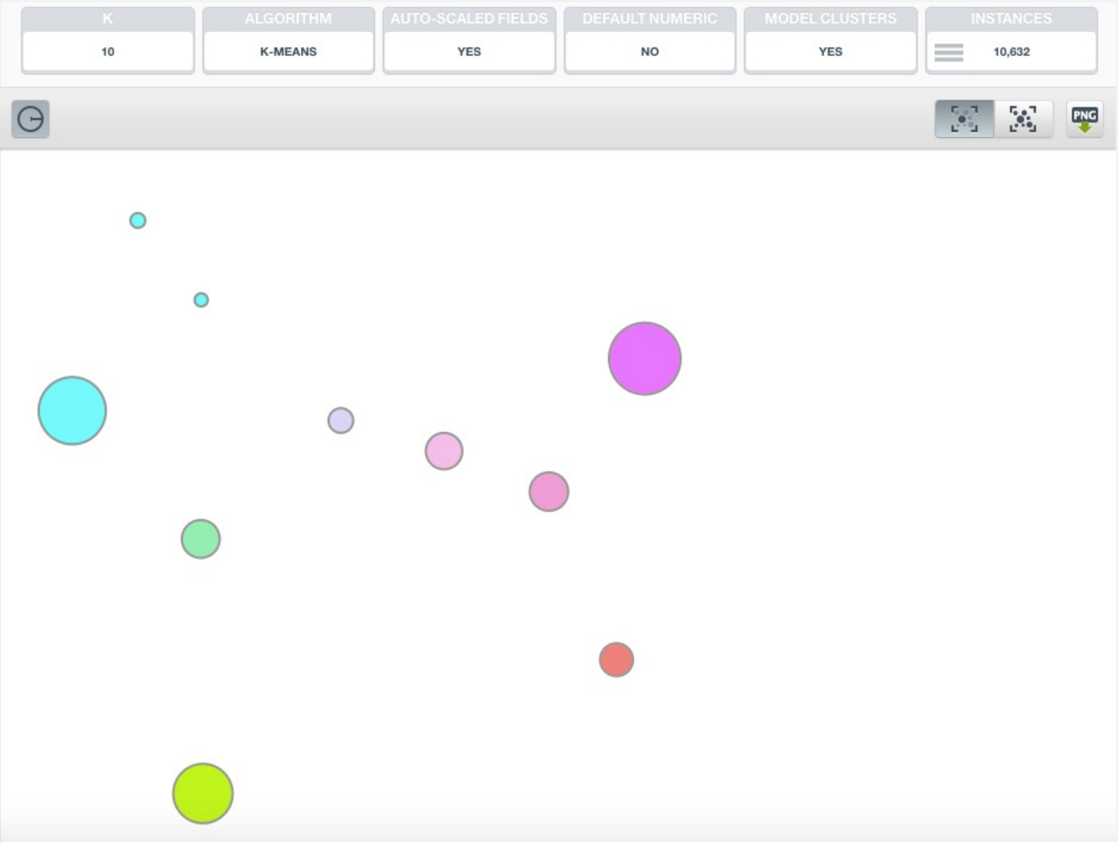
Appendix: K-Means Clustering Results

23 Try increasing/decreasing the number of clusters with K-means and choose your preferred clustering configuration. Why do you prefer this one?

K-Means Clustering configuration



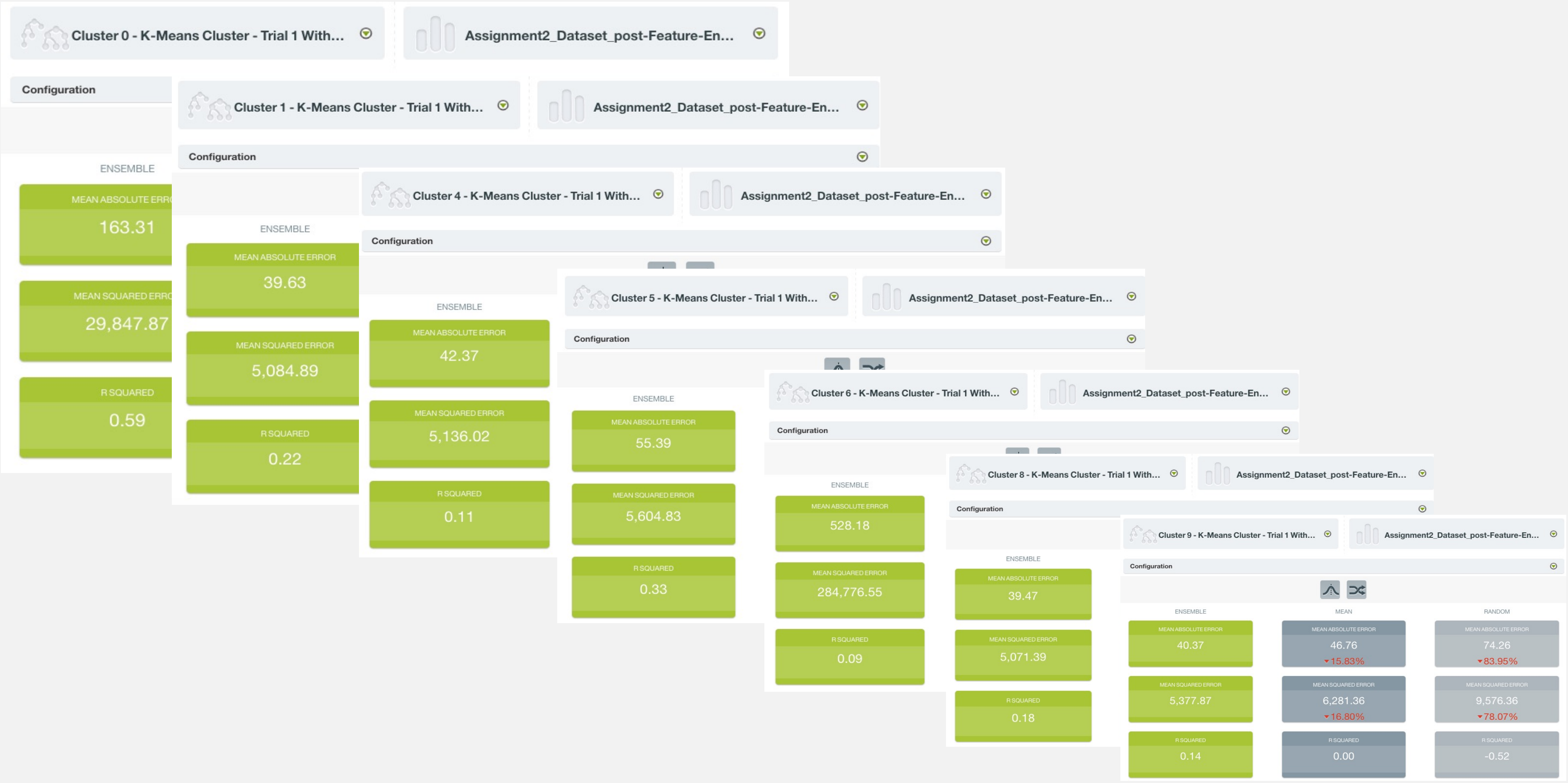
K-Means Clustering Results



Appendix: results of models for each cluster

24 Create a dataset for each of the clusters and train a model with each of them. Do the metrics of these specialized models improve with respect to the generalist one?

Performance metrics for models created from each k-means cluster



Thank you

Matteo Giardini

esade