IBM Applied Data Science Capstone Project

Predicting house prices in Monza using Foursquare API

Alessandro Bonvini, February 12th 2021

Table of contents

1. Introduction: business problem

- **The task:** predicting house prices given a set of sold houses with a group of representative features.
- **Problem:** the square meter price of the houses depends also on the position of the house.
- **Project goal:** can the information brough by Foursquare API improve the performance of the regression models with respect of using only houses characteristics as features?
- **Interested audience:** houses search engines, real estate agencies …

2.1. Data: description (1/2)

- Hundreds of houses extracted from a real estate agency in Monza (Italy)
- Training dataset: 405 houses, testing dataset: 44 houses.

	PRICE	ADDRESS	ROOMS	METERS	BATHROOMS	FLOOR	FLOORS	YEAR	STATUS	TERRACE	GARDEN	GARAGE	ENERGY	NEIGHBORHOOD	GRADE
0	5317000	6 viale Cesare Battisti	5	542	3	GROUND	2	1900	RENOVATED	NO	YES	NO	A	1	EXPENSIVE
1	2970000	6 viale Cesare Battisti	4	295	3	MIDDLE	2	1900	RENOVATED	YES	NO	NO	A	1	EXPENSIVE
2	280000	3 via Ambrosini	3	115	2	MIDDLE	1	1980	GOOD	NO	NO	YES	E	1	NORMAL
3	1050000	16 via Carlo Porta	5	278	3	MIDDLE	2	1800	RENOVATED	YES	NO	YES	E	1	EXPENSIVE
4	690000	1 via Bellini	5	220	3	MIDDLE	1	1970	RENOVATED	YES	NO	YES	G	1	NORMAL
5	950000	14 via Sant'Andrea	3	272	3	GROUND	1	2020	NEW	NO	YES	YES	A3	1	NORMAL
6	450000	35 via Aliprandi Pinalla	3	145	1	LAST	1	1890	RENOVATED	NO	NO	YES	G	1	NORMAL
7	510000	9 via Ramazzotti	5	220	3	MIDDLE	1	1970	GOOD	NO	NO	YES	E	1	CHEAP
8	770000	via Donizetti	4	200	2	GROUND	1	2020	NEW	NO	YES	NO	A4	1	EXPENSIVE
9	650000	20 via Francesco Frisi	5	200	2	MIDDLE	1	1900	GOOD	NO	YES	NO	E	1	NORMAL

2.2. Data: description (2/2)

FEATURE	MEANING
PRICE	The price of the house, in Euros.
ADDRESS	The street of the house, in the following format: Number Street.
ROOMS	The number of rooms of the house. Bathrooms are not considered as rooms.
METERS	The commercial square meters of the house
BATHROOMS	The number of bathrooms of the house.
FLOOR	The main floor of the house (GROUND, MIDDLE, LAST, VILLA)
FLOORS	The number of floors of the house.
YEAR	The construction year of the house
STATUS	The current conditions of the house (BAD, GOOD, RENOVATED, NEW)
TERRACE	The house has a terrace large enough to be used for eating (YES/NO)
GARDEN	The house has a garden that can be used to let the kids play (YES/NO)
GARAGE	The house has a covered place to be used for parking cars (YES/NO)
ENERGY	The certified energy class of the house (from G to A4)
NEIGHBORHOOD	The neighborhood of the house
GRADE	The estate agent evaluation for the price (CHEAP, NORMAL, EXPENSIVE)

2.3. Data: preprocessing

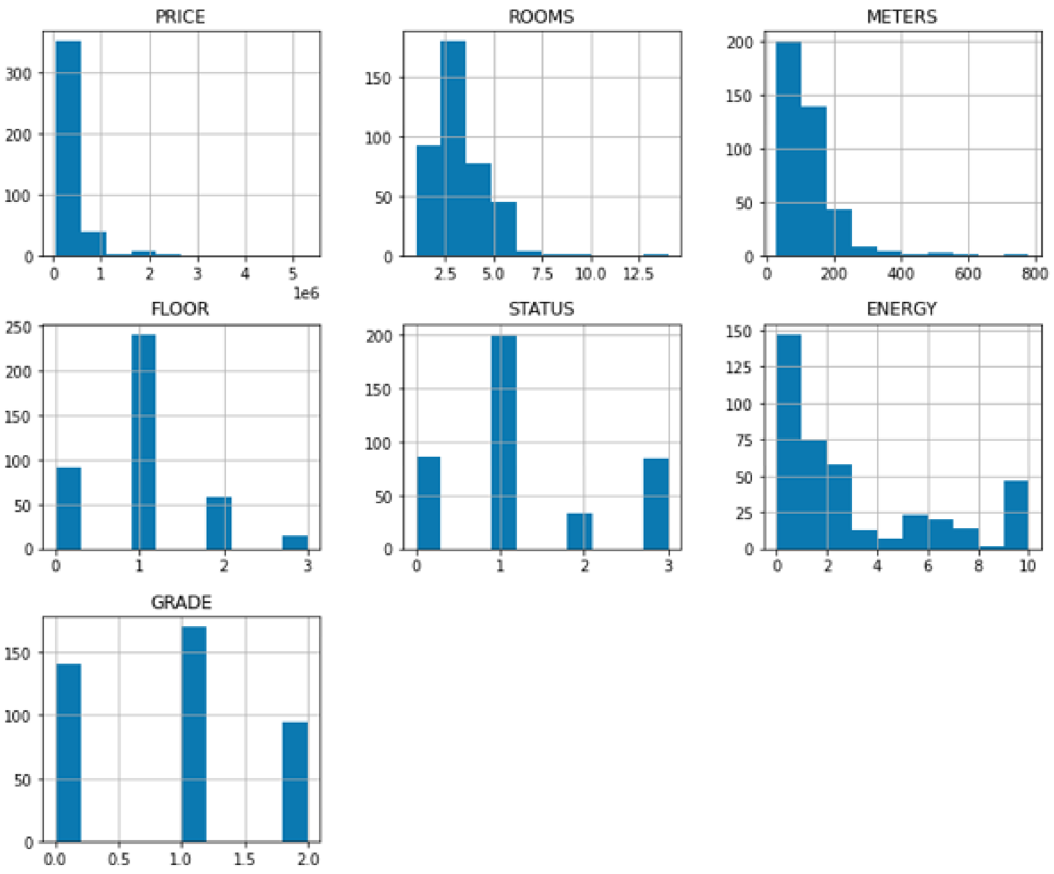
- Ordinal Encoding of categorial features

ENCODED VALUE	FLOOR	STATUS	ENERGY	GRADE	TERRACE/ GARDEN/ GARAGE
0	GROUND	BAD	G	CHEAP	NO
1	MIDDLE	GOOD	F	NORMAL	YES
2	LAST	RENOVATED	E	EXPENSIVE	
3	VILLA	NEW	D		
4			C		
5			B		
6			A		
7			A1		
8			A2		
9			A3		
10			A4		

- Custom encoding of FLOOR feature: three columns GROUND, MIDDLE, HIGH depending on FLOOR and FLOORS

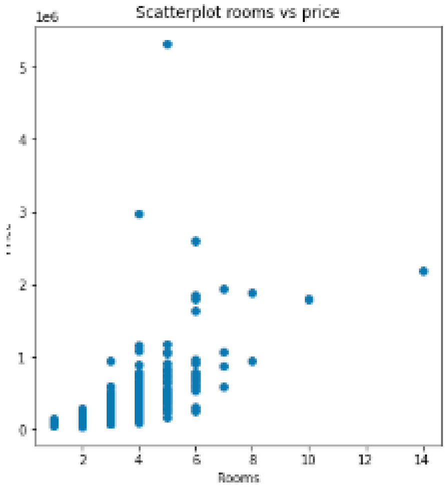
3.1. Methodology: exploratory data analysis (1/3)

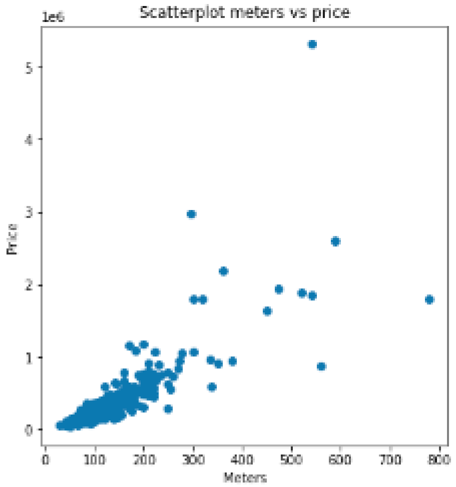
- Samples in training dataset are not equally distributed
- Some charateristics are more common than others (price < 1000000, rooms < 5, meters < 200 …)



3.1. Methodology: exploratory data analysis (2/3)

- Price correlated with square meters

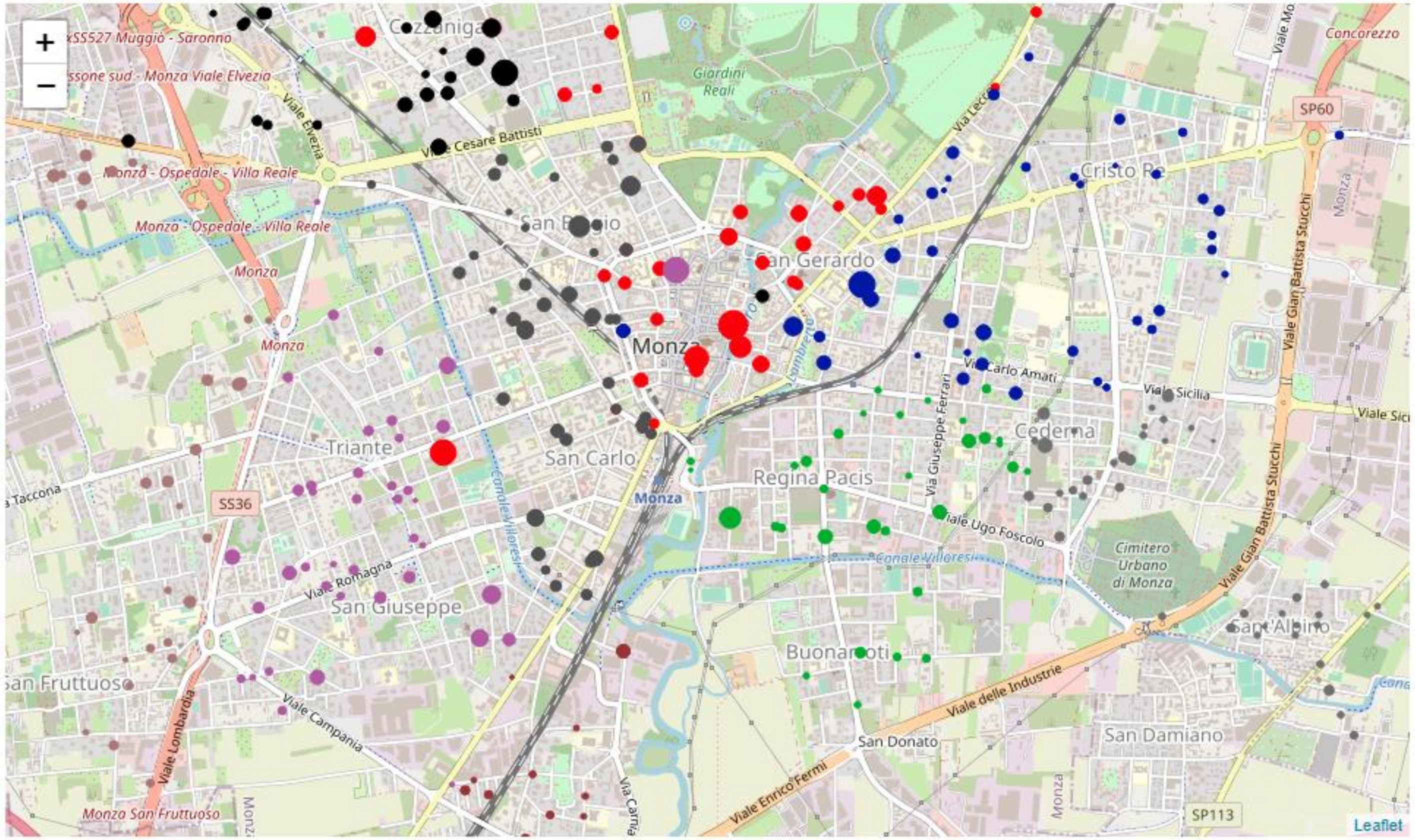




- House price = square meters * square meters price
- Add square meters price as column and use it as target variable
- Remove outliers: houses with price > 2900000

3.1. Methodology: exploratory data analysis (3/3)

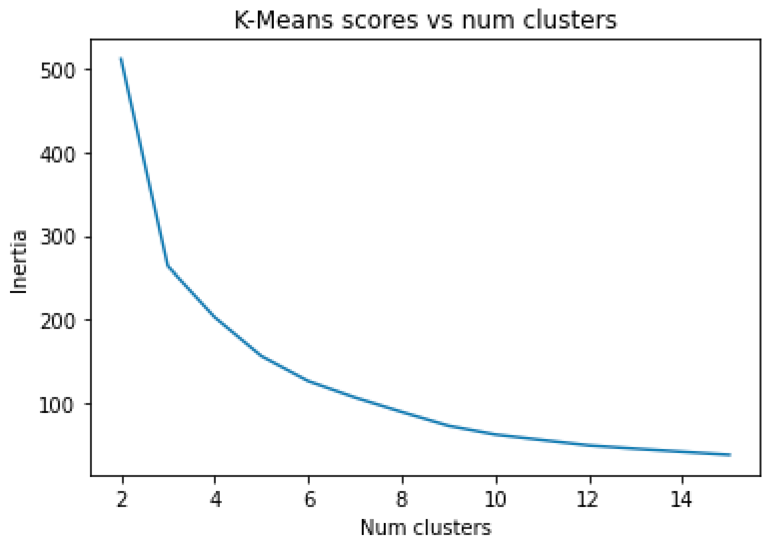
- Plot houses: circle proportional to meter price, color is neighborhood



- Neighborhood division not correct

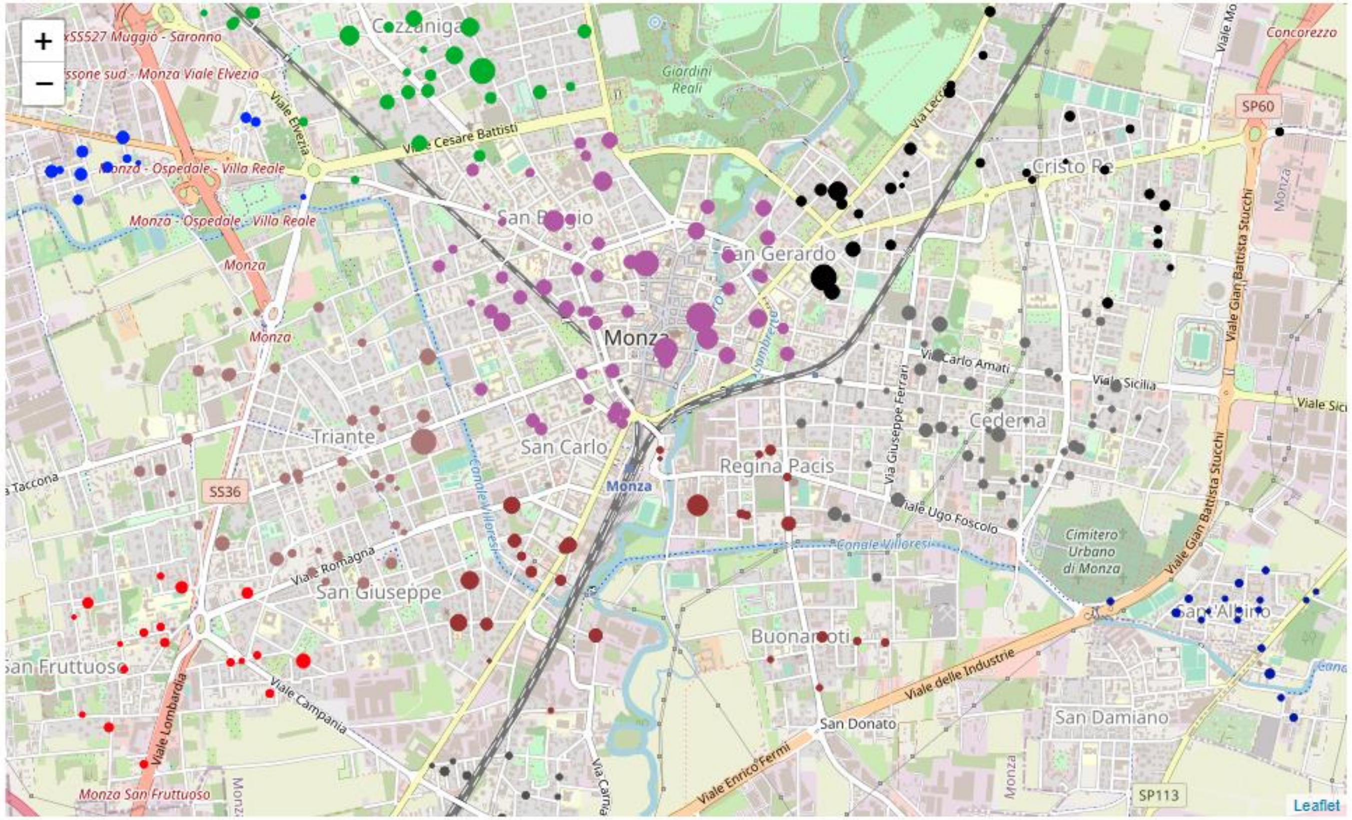
3.1. Methodology: K-Means neighboroods clustering (1/2)

- Perform K-Means clustering with coordinates to obain better neighbors
- Test from 2 to 15 clusters
- Choose 10 neighborhoods



3.1. Methodology: K-Means neighboroods clustering (2/2)

- K-Means clustering neighborhoods: better division



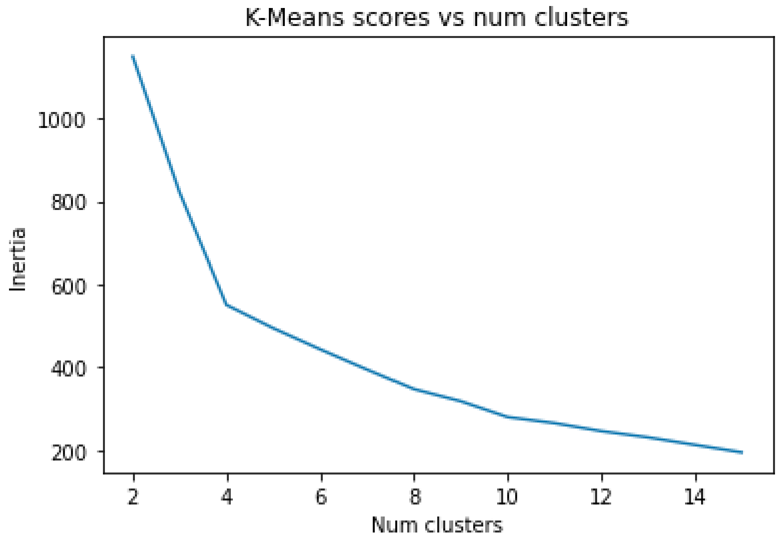
3.2. Methodology: retrieving venues with Foursquare API

- Does houses close to popular venues cost more?
- Retrieve top 30 trending venues in Monza:
  - Call explore endpoint
  - Set parameters sortByPopularity = 1, section = topPicks
- Retrieved venues become houses features:
  - 1 – the venue is present in the sorroundings of the house
  - 0 – the venus is not present in the sorroundings of the house

	LAT	LNG	Villa Reale	Piazza Trento e Trieste	Istituti Clinici Zucchi	Parco di Monza - Ingresso Alle Grazie	U2	Parco di Monza - Viale cavriga	Dori	Civico 1	La Rinascente	Duomo di Monza	Macellerie Monzesi	La Feltrinelli
0	45.60266	9.26639	0	0	0	0	0	1	0	0	0	0	0	0
1	45.58266	9.27903	0	1	1	0	0	0	1	0	1	1	0	1
2	45.59647	9.27031	1	0	0	0	0	1	0	0	0	0	0	0
3	45.59982	9.26604	0	0	0	0	0	1	0	0	0	0	0	0
4	45.58688	9.27912	0	1	1	0	0	0	1	0	0	1	0	0

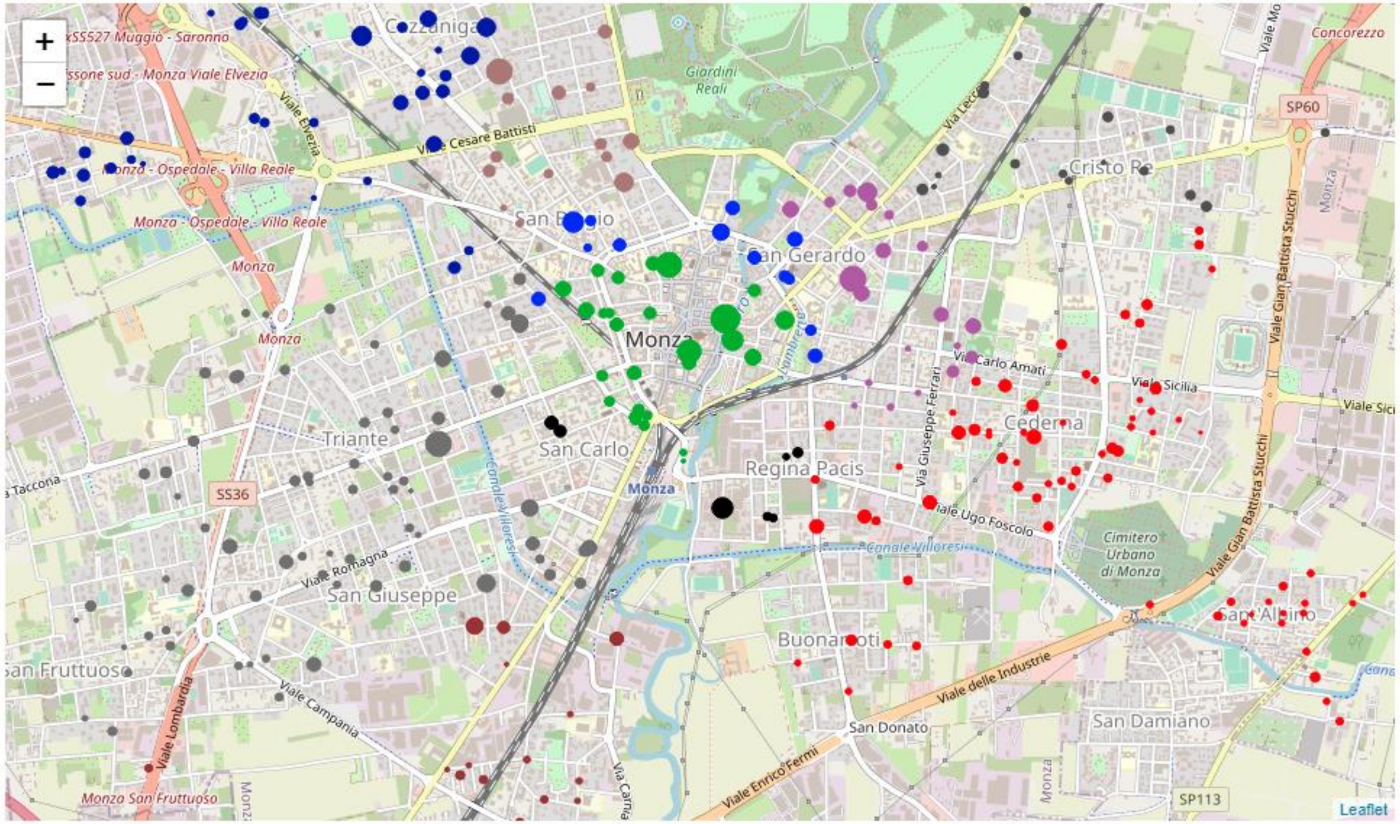
3.3. Methodology: K-Means neighboroods clustering with trending venues (1/2)

- How the new venues features impact on K-Means neighbors clustering?
- Try K-Means with venues features



3.3. Methodology: K-Means neighboroods clustering with trending venues (2/2)

- K-Means clustering with venues features
  - Divided houses «in the center» from houses «around the center»
  - Bigger clusters in peripherical neighborhoods



3.4. Methodology: regression (1/2)

- Four training datasets to try
- **Dataset #1**: only houses charateristics:

ROOMS	METERS	BATHROOMS	LAST	YEAR	STATUS	TERRACE	GARDEN	GARAGE	ENERGY

- **Dataset #2**: houses characteristics + K-Means cluster:

FEATURES OF DATASET 1	K-MEANS CLUSTER USING LAT AND LNG

- **Dataset #3**: houses characteristics + K-Means venues cluster:

FEATURES OF DATASET 1	K-MEANS CLUSTER USING TOP TRENDING VENUES

- **Dataset #4**: houses characteristics + venues features:

FEATURES OF DATASET 1	TOP TRENDING VENUES FEATURES

- All feature sets standardized with sklearn Standard Scaler

3.4. Methodology: regression (2/2)

- Four models to try

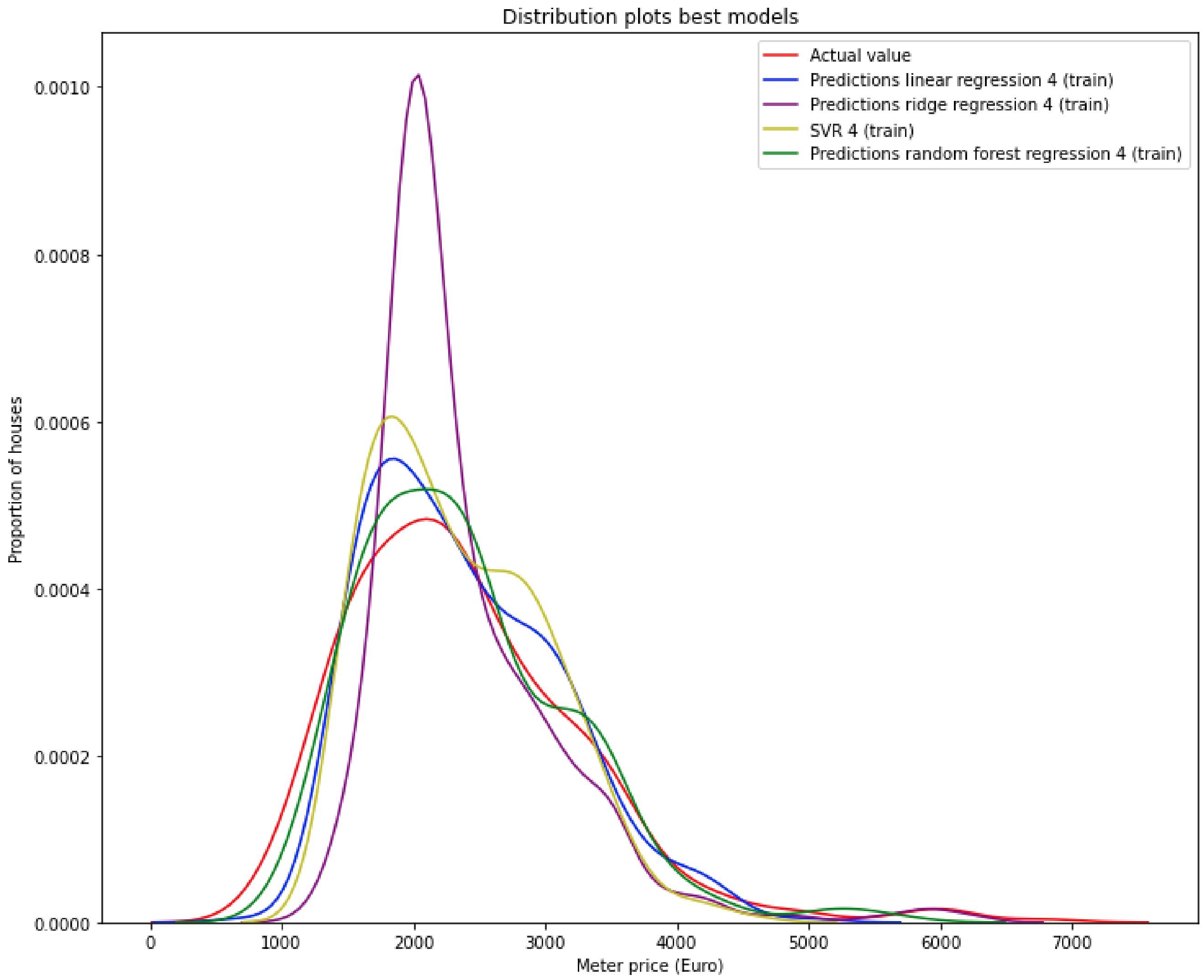
- **Multivariate Linear Regression**
- **Ridge Regression**
  - Feature transformation with 3° degree Polynomial Features
  - Alpha grid searching: 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, 100000
- **Support Vector Regression**
  - Kernel functions: linear, poly, rbf, sigmoid
  - C grid searching: 0.1, 1, 10, 100
- **Random Forest Regression**
  - Number of estimators grid searching: 5, 10, 50, 100, 200

- Using sklearn GridSearchCV with 5 folds cross-validation

4.1. Results: training results (1/3)

		R2 Score	RMSE
Multivariate Linear Regression	Dataset #1	0.4968	643.610
	Dataset #2	0.5029	639.662
	Dataset #3	0.5057	637.906
	Dataset #4	**0.6386**	**545.379**
Ridge Regression + Polynomial Features	Dataset #1	0.4464	675.064
	Dataset #2	0.4721	659.197
	Dataset #3	0.4689	661.225
	Dataset #4	**0.7745**	**430.790**
Support Vector Regression	Dataset #1	0.4703	660.295
	Dataset #2	0.4893	648.365
	Dataset #3	0.4832	652.202
	Dataset #4	**0.5951**	**577.315**
Random Forest Regression	Dataset #1	0.9178	260.069
	Dataset #2	0.9352	230.81
	Dataset #3	0.9307	238.76
	Dataset #4	**0.938**	**225.83**

4.1. Results: training results (2/3)

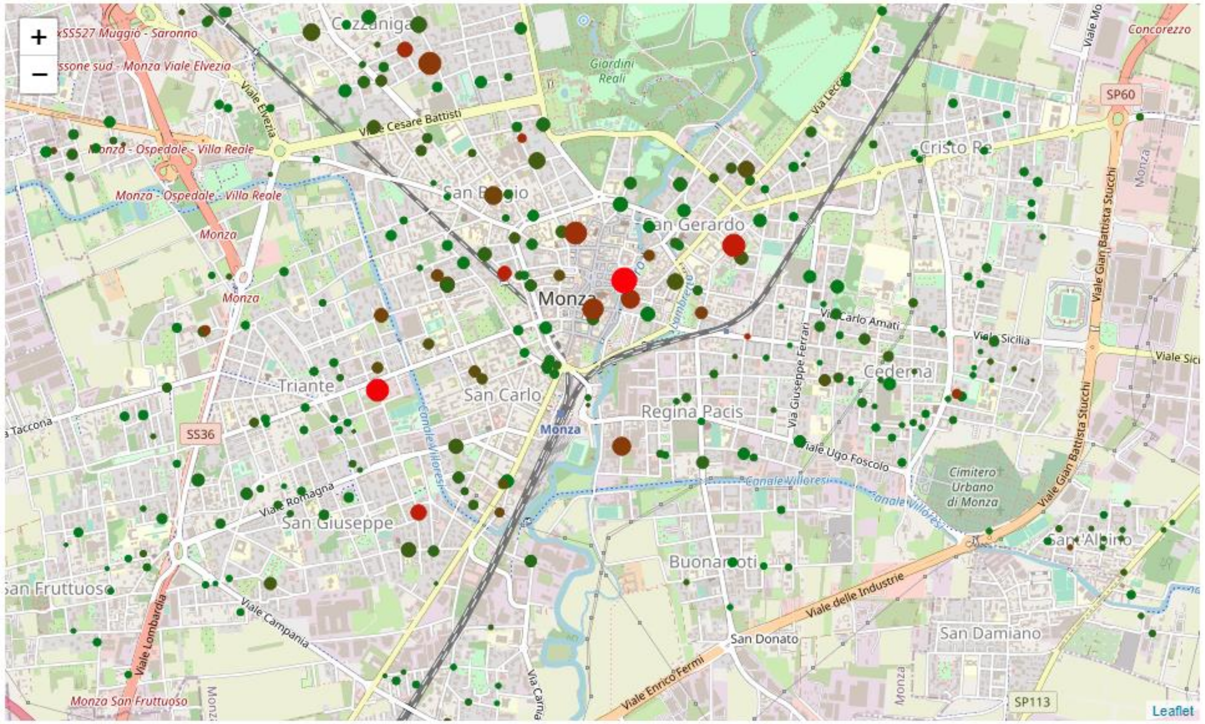


4.1. Results: training results (3/3)

- Worst predictions: particular cases

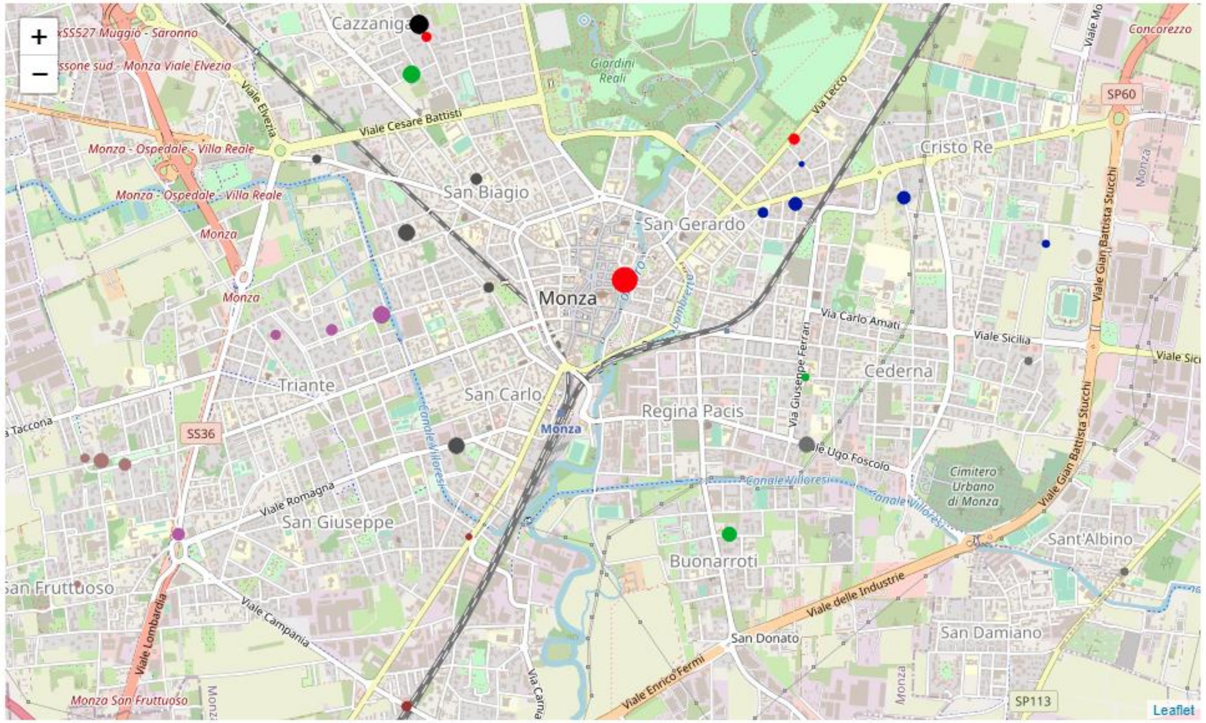
	PRICE	ADDRESS	ROOMS	METERS	BATHROOMS	FLOOR	FLOORS	GROUND	MIDDLE	LAST	YEAR	STATUS	TERRACE	GARDEN	GARAGE
14	1160000	2 piazza Garibaldi	4	171	2	2.0	1	0	0	1	2016	3.0	1.0	0.0	0.0
141	390000	8 via Asiago	5	180	3	2.0	2	0	1	1	2021	3.0	1.0	0.0	0.0

- Delta prices distributions: higher meter prices, higher errors



4.2. Results: test set evaluation (1/3)

- Test dataset
  - 44 houses never seen during training
  - Same format of training dataset
  - Same preprocessing pipeline
  - Venues features retrieval with Foursquare

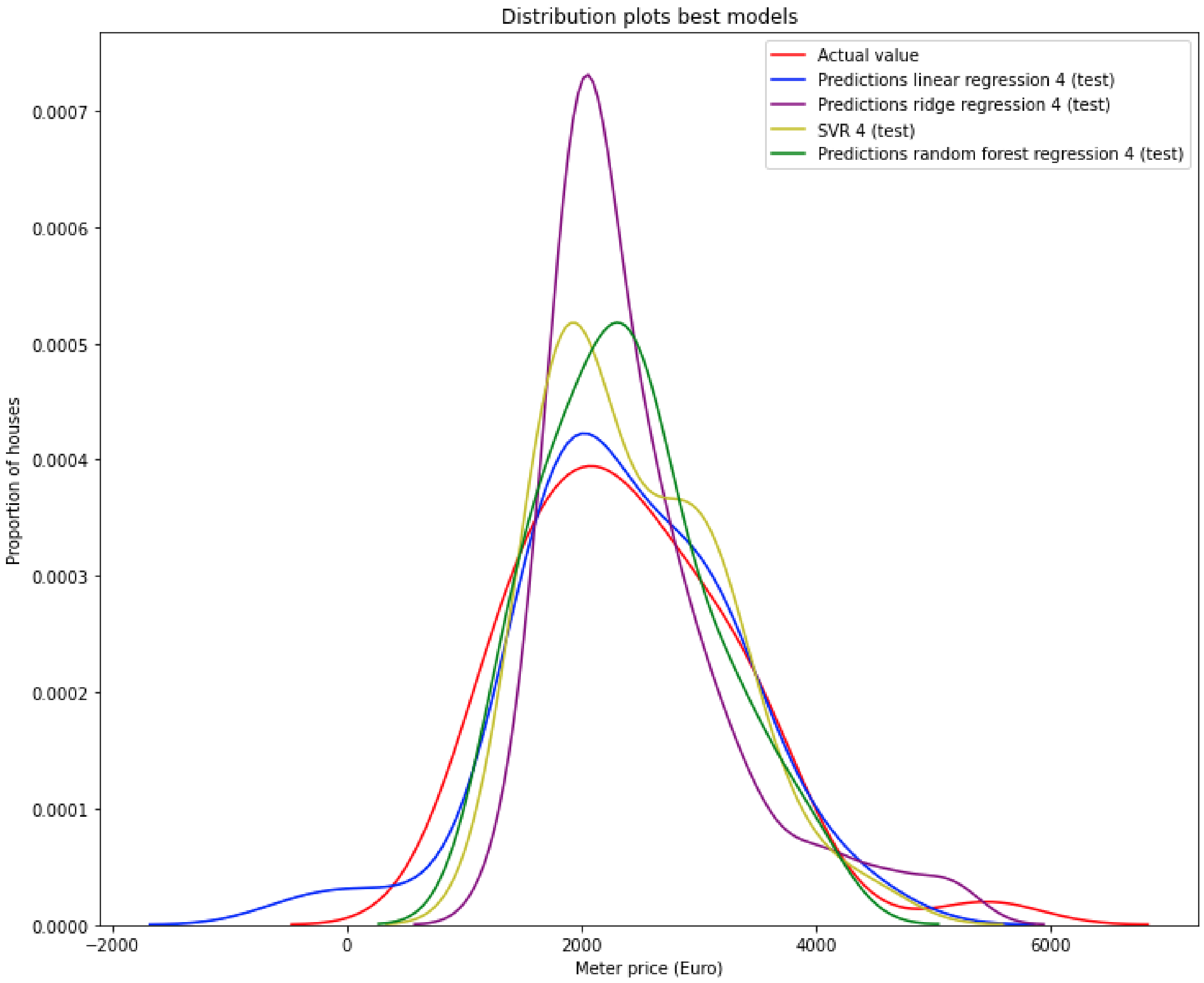


4.2. Results: test set evaluation (1/3)

- Tested only the best performing models: datasets #4

		R2 Score	RMSE
Multivariate Linear Regression	Dataset #4	0.3129	752.609
Ridge Regression + Polynomial Features	Dataset #4	0.4639	664.784
Support Vector Regression	Dataset #4	0.4427	667.761
Random Forest Regression	**Dataset #4**	**0.6975**	**499.34**

4.2. Results: test set evaluation (1/3)



5. Discussion

- Datasets with spatial information peforms better then dataset with only houses characteristics:
  - Venues features retrieved with Foursquare allowed to improve the predictions performances
  - Venus features big problem: they change over time
- Random Forest Regression outperformed other models training performances
  - Better captured training set distribution, maybe overfitted
  - Did not capture well enough prices > 5000
- Models did not generalize very well on test data
- Not enough samples in price range > 5000
- Very basic models tuning

6. Conclusion

- Improvements that can be made:
  - The training set can be enlarged. Include more samples with high square meter price
  - Improve spatial information. Find a way to stabilize the top trend venues features
  - Add other features
  - Test other models
  - Interesting to study the classification problem of predicting the GRADE class, this time given the price

THANKYOU