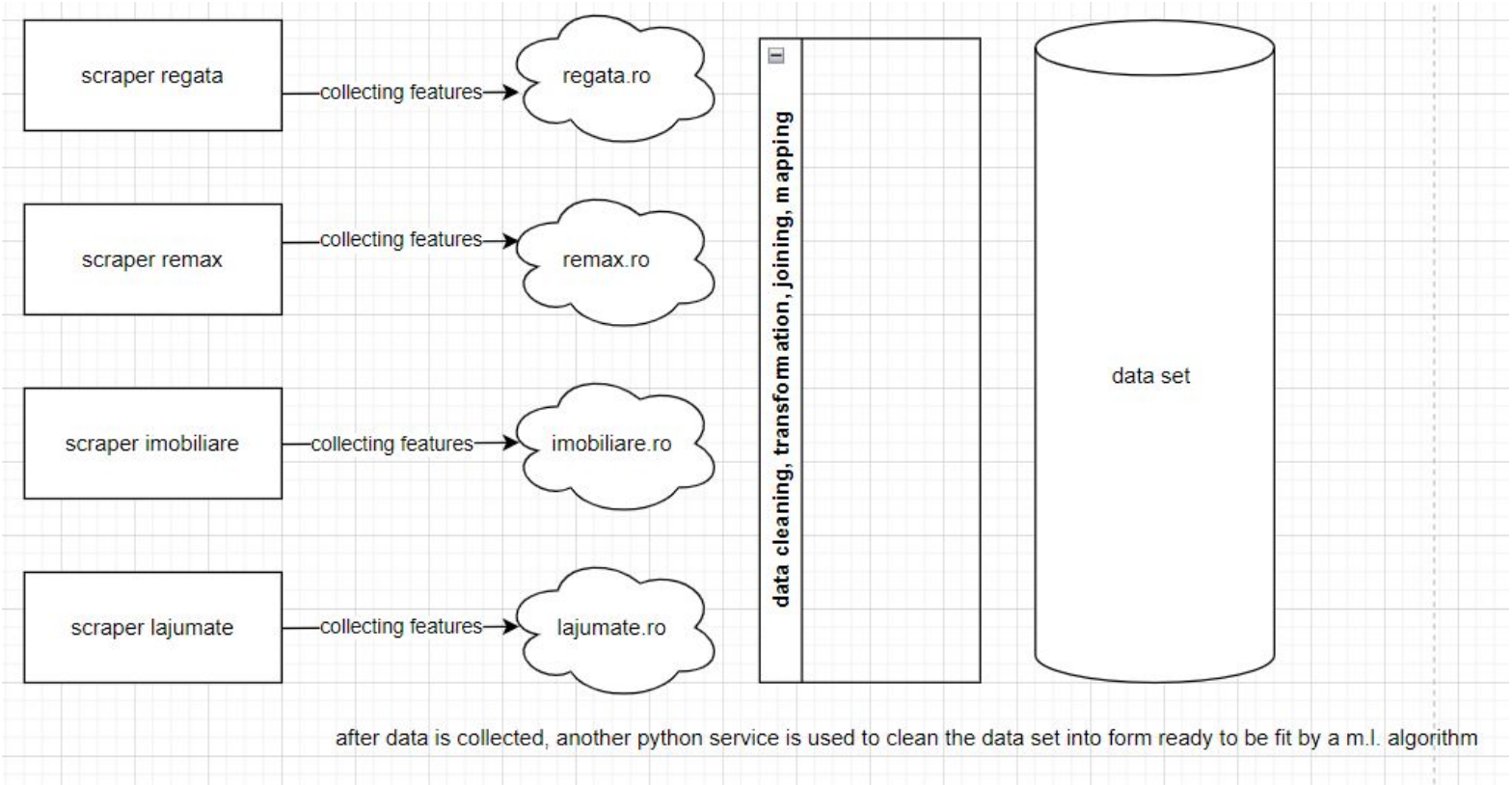
Real estate data collection, mining and learning

Fulea Andrei, 511
Burz Florin, 511

Dataset

- the dataset used is collected automatically from online sources with web scraping - ETL - .



Data cleaning

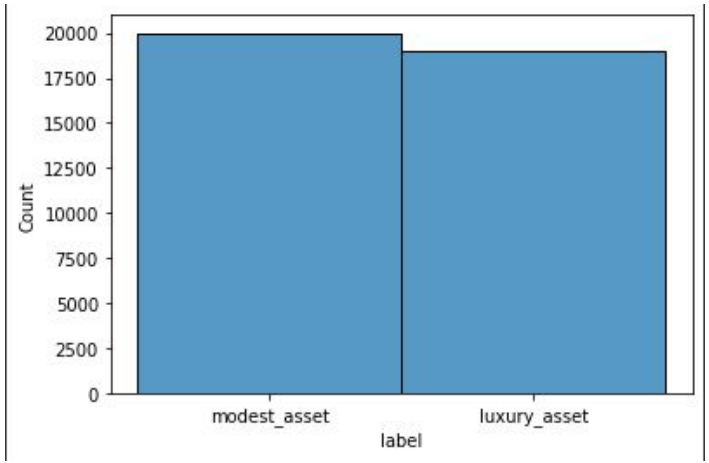
- in the initial form, the data is in object format read by pandas
- after the cleaning and transformation, we eliminate useless part of text and keep just relevant information + casting to data type

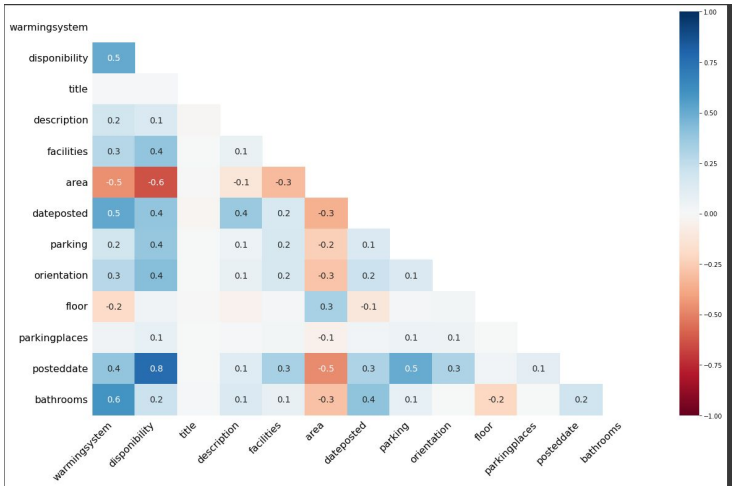
price	33.000 €
assetstate	Bună
warmingsystem	Calorifere, Termoficare
disponibility	Imediat
colector	remax
title	Garsoniera vanzare in bloc de apartamente Bucu...
description	\r\n\r\nPROPRIETATEA ESTE GREVATA DE SARCINA U...
facilities	NaN
compartimentation	Decomandat
rooms	1 camera
yearconstruction	1985.0
confort	1
area	Sectia Politie 20
pagenumber	1.0
dateposted	Acum o zi
town	bucuresti-ilfov
parking	NaN
assettype	apartamente
transactiontype	vanzare
orientation	vedere stradala
neighborhood	NaN
balcony	1 balcon
furnished	nespecificat
floor	1
parkingplaces	2.0
posteddate	23 August 2021
link	https://www.remax.ro/anunt/75391/garsoniera-de...
squaremetres	36.62 mp
bathrooms	1.0

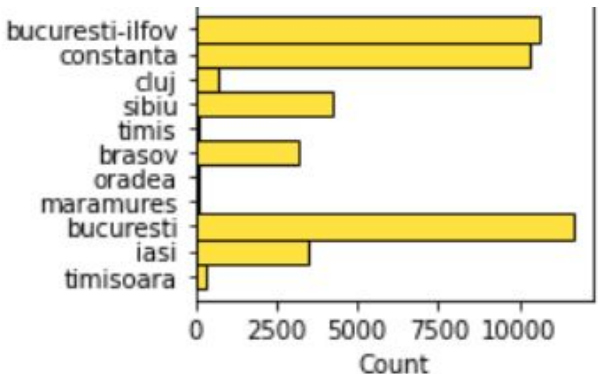
price	33000
assetstate	2
warmingsystem	Calorifere, Termoficare
disponibility	Imediat
colector	4
title	Garsoniera vanzare in bloc de apartamente Bucu...
description	\r\n\r\nPROPRIETATEA ESTE GREVATA DE SARCINA U...
facilities	NaN
compartimentation	7
rooms	1
yearconstruction	1985
confort	2
area	Sectia Politie 20
pagenumber	1
dateposted	Acum o zi
town	bucuresti-ilfov
parking	NaN
assettype	apartamente
transactiontype	vanzare
orientation	vedere stradala
neighborhood	NaN
balcony	1
furnished	4
floor	1
parkingplaces	2.0
posteddate	23 August 2021
link	https://www.remax.ro/anunt/75391/garsoniera-de...
squaremetres	36
bathrooms	1.0

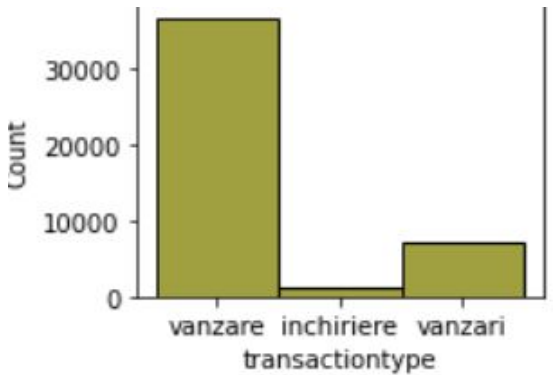
Exploratory Data Analysis

- static variables

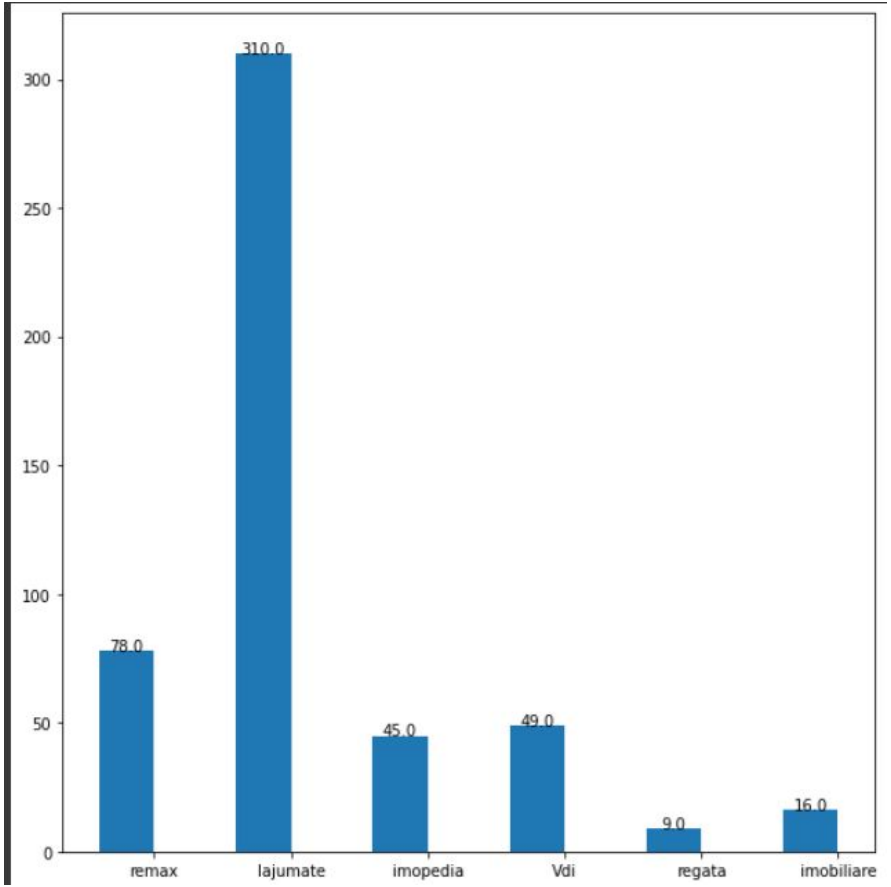


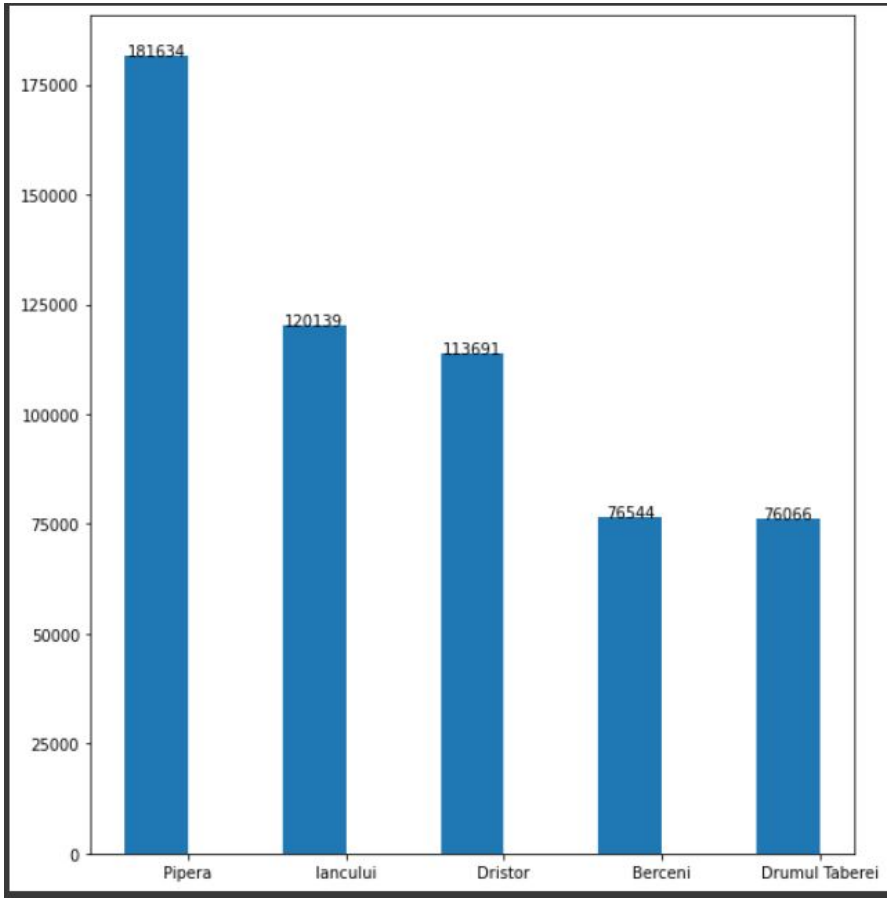




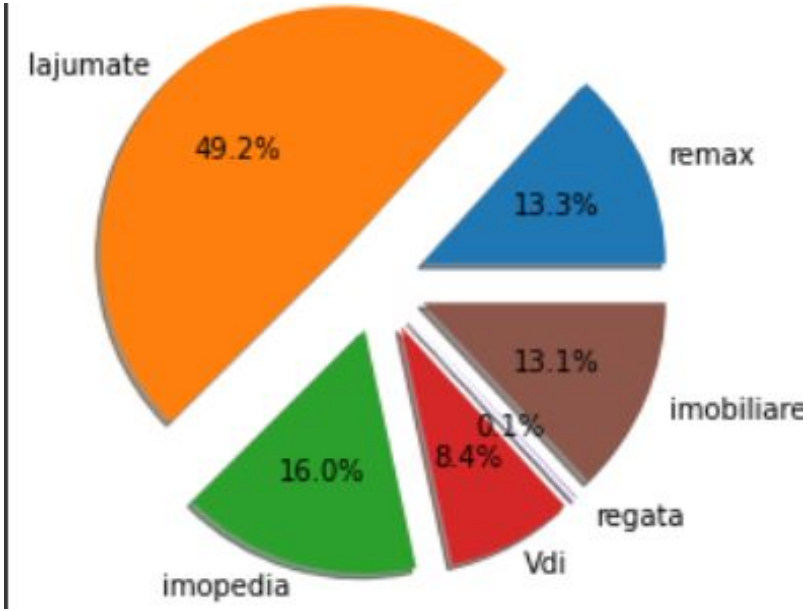


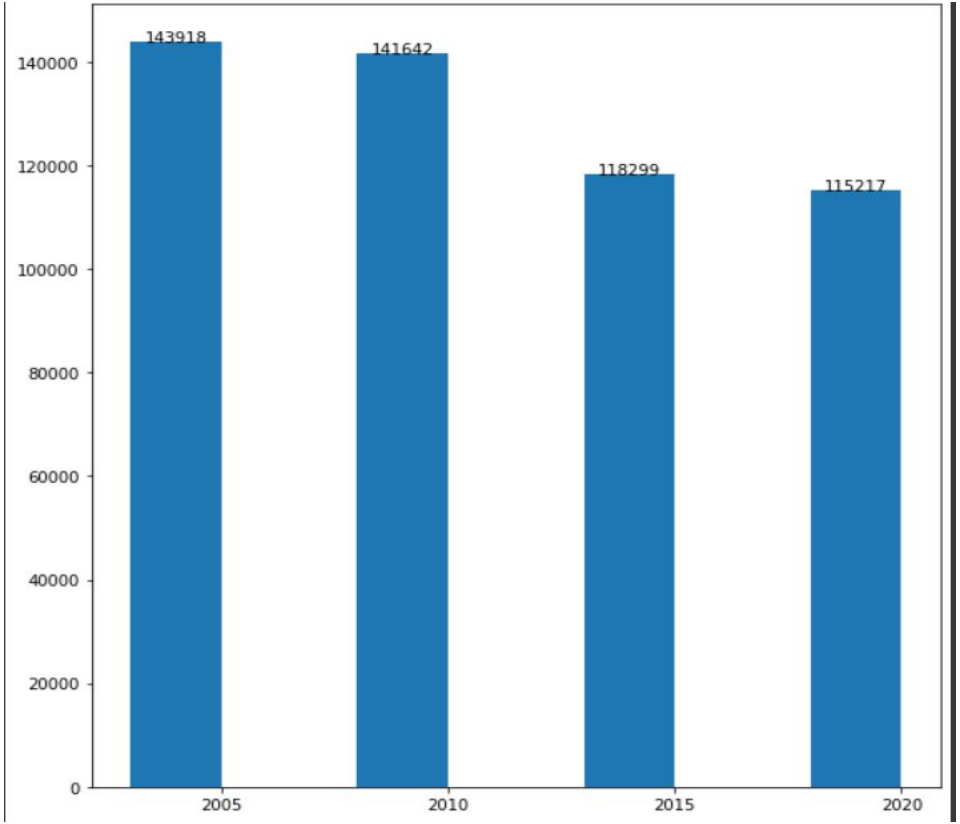
Distribution of medium price with areas and max page collector from each website





Distribution of medium price with year of construction and number of samples from each website





Regression problem - price prediction

- the task is to predict the price of an asset
- We construct the X and Y, y being df[‘price’] and X

```
x = national_real_estate_data_CLEANED[['rooms','yearconstruction','confort','furnished','squaremetres','compartimentation']]
```

- Data was splited with 0.3
- initial results:

```
x_train.shape, x_test.shape, y_train.shape, y_test.shape
```

```
((31510, 6), (13505, 6), (31510,), (13505,))
```

	Model	MAE	MSE	RMSE	R2 Square	Cross Validation
0	Elastic Net Regression	45909.494407	1.393692e+10	118054.748756	19.226309	-0.197257
1	Artficial Neural Network	77615.584913	2.288715e+10	151284.983071	-32.646150	0.000000
2	Polynomail Regression	47169.806022	1.249788e+10	111793.897005	27.566545	0.000000
3	Robust Regression	49702.800140	3.629103e+11	602420.376733	-2003.305290	-606.203894
4	Ridge Regression	46001.150420	1.392967e+10	118024.028941	19.268341	-0.121264

Experiments

- with standardization applied

	Model	MAE	MSE	RMSE	R2 Square	Cross Validation
0	Elastic Net Regression	45273.962644	1.130314e+10	106316.231805	19.803440	-0.197257
1	Artficial Neural Network	35766.347966	6.800477e+09	82465.005342	51.750149	0.000000
2	Polynomail Regression	50864.469298	1.268010e+10	112605.962516	10.033802	0.000000
3	Robust Regression	69938.963973	4.299244e+11	655686.223668	-2950.343225	-549.799696
4	Ridge Regression	45353.652107	1.130304e+10	106315.775401	19.804128	-0.121264

- we can observe that neural net presents a better r2 square, so we modified the architecture by adding neurons on layers to improve ANN model

	Model	MAE	MSE	RMSE	R2 Square	Cross Validation
0	Artficial Neural Network	35038.359506	5.657220e+09	75214.495591	57.410266	0

Extending X with townId

	Model	MAE	MSE	RMSE	R2 Square	Cross Validation
0	Elastic Net Regression	45187.434903	1.049403e+10	1.024404e+05	19.500061	-0.450123
1	Artficial Neural Network	35287.994277	5.751872e+09	7.584109e+04	55.877250	0.000000
2	Polynomail Regression	45800.998701	9.307971e+09	9.647783e+04	28.598320	0.000000
3	Robust Regression	64494.858070	2.881867e+12	1.697606e+06	-22006.877258	-529.215522
4	Ridge Regression	45289.856102	1.049735e+10	1.024566e+05	19.474547	-0.403236

- we added 1 more feature, named townID. Starting from town as a string, we mapping using a world cities dataset that string with an ID
- because now the data are in range apart values, we transform X train and X test in standardized data
- changing standardization method from preprocessing.scale() into standardscaler

	Model	MAE	MSE	RMSE	R2 Square	Cross Validation
0	Elastic Net Regression	44625.216482	8.600103e+09	9.273674e+04	18.766281	-0.450123
1	Artficial Neural Network	31793.649879	4.561492e+09	6.753882e+04	56.913656	0.000000
2	Polynomail Regression	43758.945013	8.198176e+09	9.054378e+04	22.562752	0.000000
3	Robust Regression	79593.785742	1.912251e+12	1.382842e+06	-17962.490676	-434.650979
4	Ridge Regression	44745.877849	8.609961e+09	9.278988e+04	18.673165	-0.403236

Adding random forest regressor and gridsearch

	Model	MAE	MSE	RMSE	R2 Square	Cross Validation
0	Elastic Net Regression	44625.216482	8.600103e+09	9.273674e+04	18.766281	-0.450123
1	Artficial Neural Network	31793.649879	4.561492e+09	6.753882e+04	56.913656	0.000000
2	Polynomail Regression	43758.945013	8.198176e+09	9.054378e+04	22.562752	0.000000
3	Robust Regression	79593.785742	1.912251e+12	1.382842e+06	-17962.490676	-434.650979
4	Ridge Regression	44745.877849	8.609961e+09	9.278988e+04	18.673165	-0.403236
5	Random Forest Regressor	15280.705051	1.460016e+09	3.821015e+04	86.209175	0.000000

```
Fitting 2 folds for each of 288 candidates, totalling 576 fits
{'bootstrap': True, 'max_depth': 110, 'max_features': 3, 'min_samples_leaf': 3, 'min_samples_split': 8, 'n_estimators': 100}
None
```

```
param_grid = {
'bootstrap': [True],
'max_depth': [80, 90, 100, 110],
'max_features': [2, 3],
'min_samples_leaf': [3, 4, 5],
'min_samples_split': [8, 10, 12],
'n_estimators': [100, 200, 300, 1000]
}
```