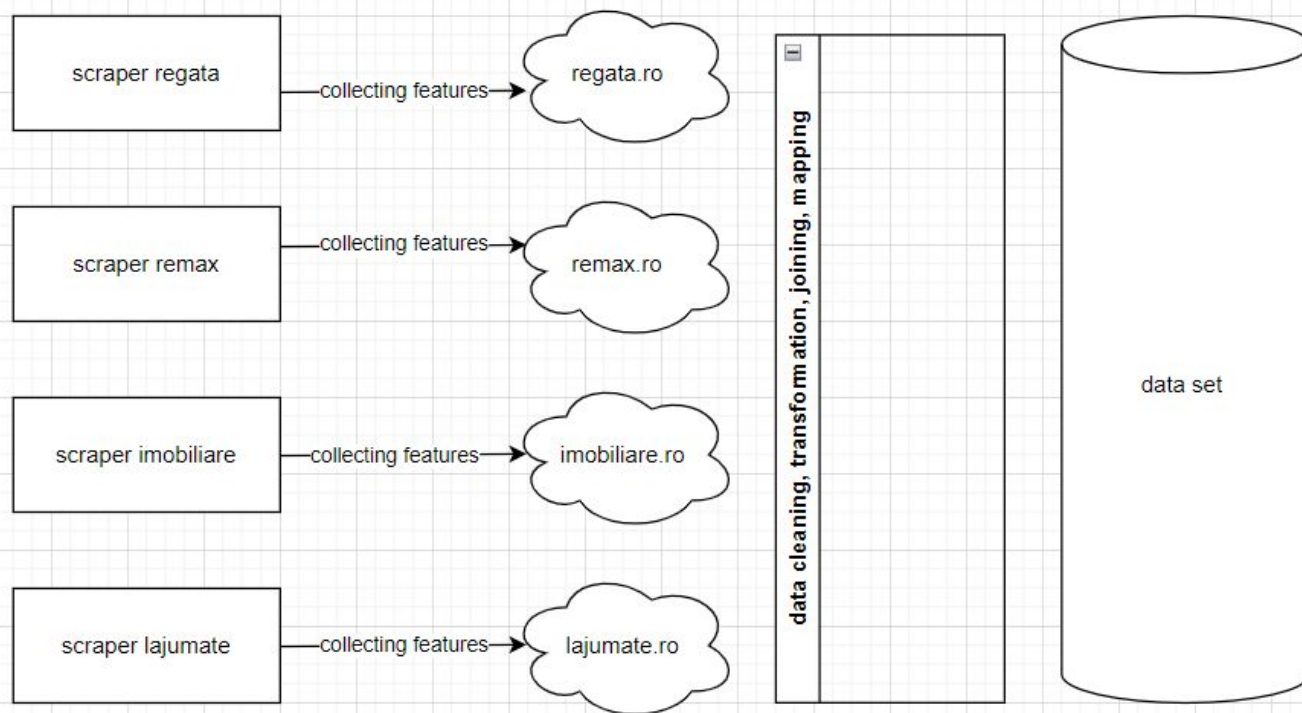


# Real estate data collection, mining and learning

Fulea Andrei, 511  
Burz Florin, 511

# Dataset

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



after data is collected, another python service is used to clean the data set into form ready to be fit by a m.l. algorithm

# Data cleaning

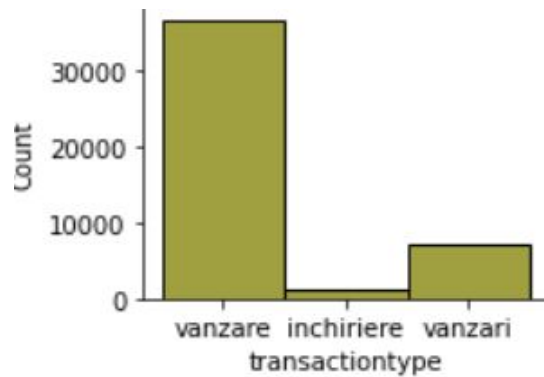
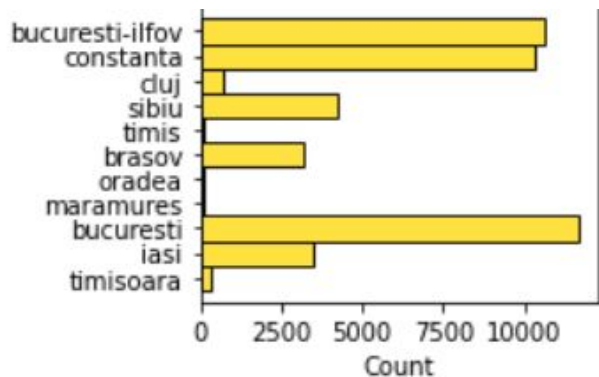
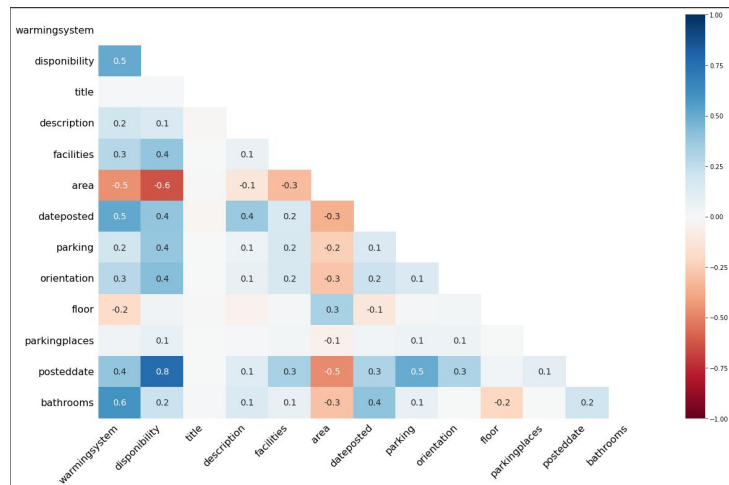
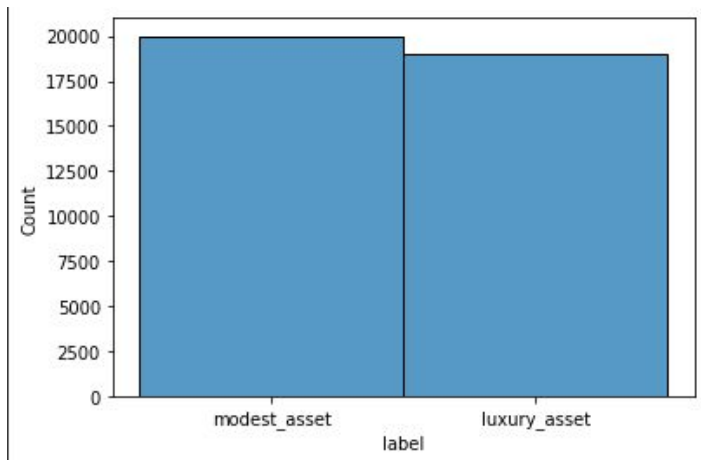
- in the initial form, the data is in object format read by pandas
- after the cleaning and transformation, we eliminate useless part of text and keep just relevant information + casting to data type

```
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
```

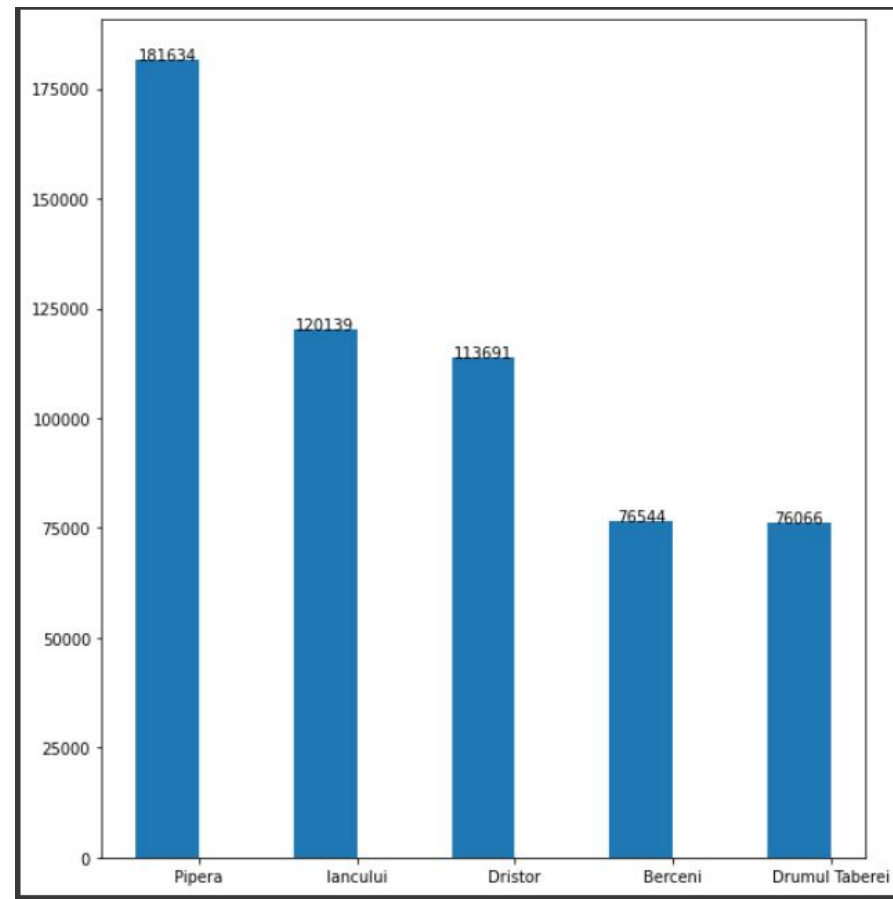
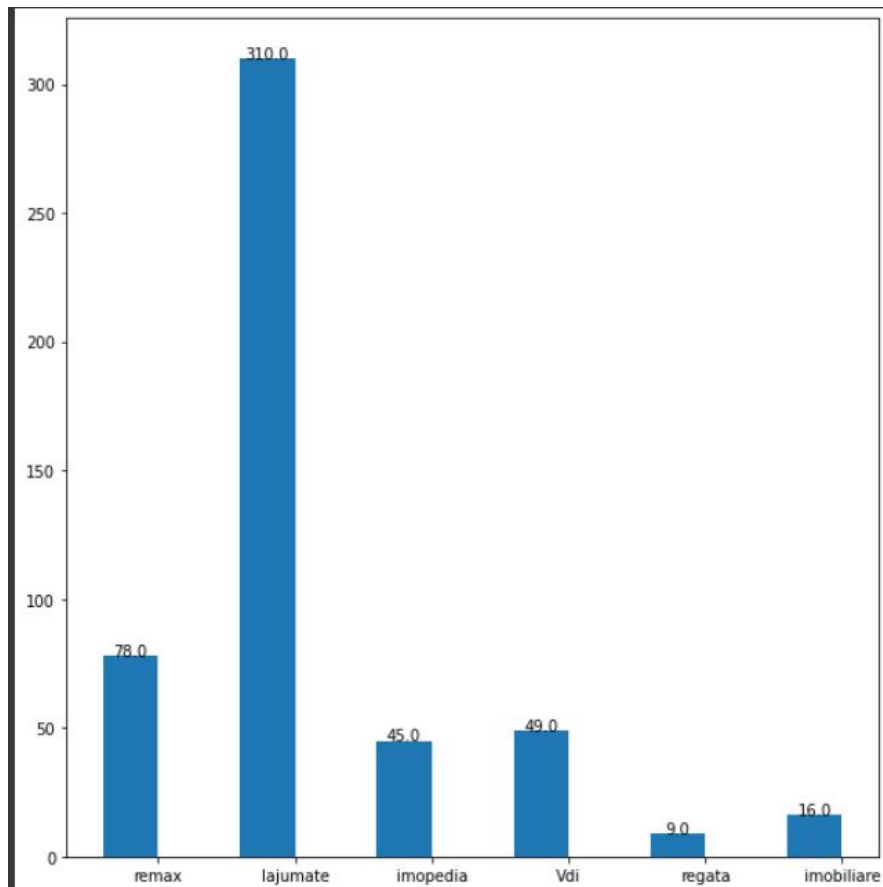
```
price                33000
assetstate            2
warmingsystem        Calorifere, Termoficare
disponibility        Imediat
colector             4
title                Garsoniera vanzare in bloc de apartamente Bucu...
description           \r\n\r\nPROPRIETATEA ESTE GREVATA DE SARCINA U...
facilities            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

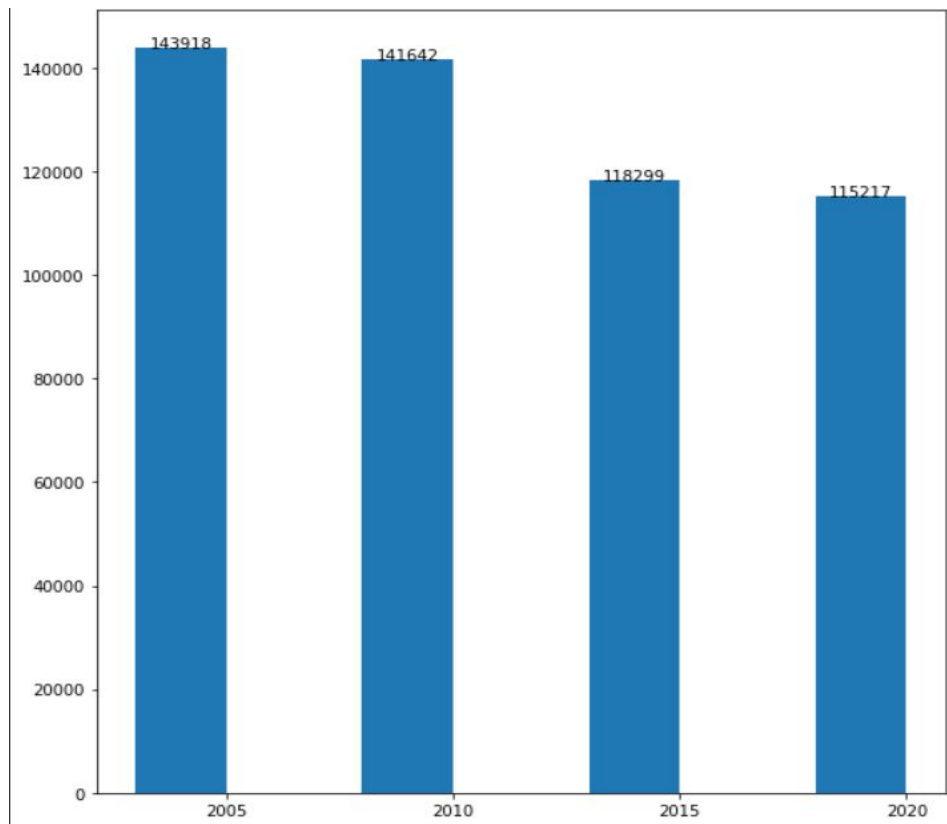
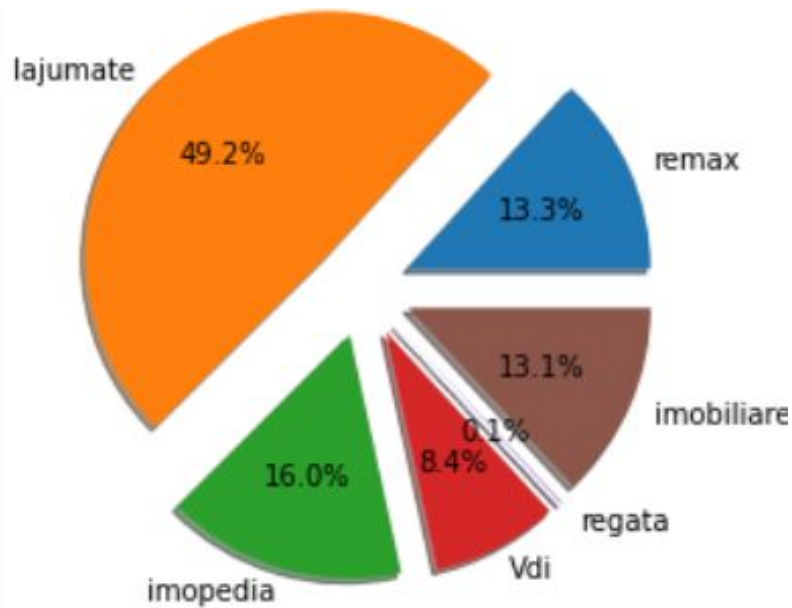
- static variables



# 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', 'compartmentation']]
```

- 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

# Preprocessing methods and hyper tunning

- 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	Artificial Neural Network	35766.347966	6.800477e+09	82465.005342	51.750149	0.000000
2	Polynomial 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

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