Agenda:

1. Dataset description
2. Analysis objective
3. Plan for analysis
4. EDA
5. Models used for training and testing
6. Summary and Results

1. Dataset description

Data was collected from Polish Amazon-like site – olx.pl
Data was collected using web crawler, contains only ads regarding rent of flats or rooms in a City Kraków in Poland. It consists of 5349 entries and 15 variables:

```
RangeIndex: 5349 entries, 0 to 5348
Data columns (total 15 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   promoted                            5349 non-null   object
 1   web_page                            5349 non-null   object
 2   market_type                         5349 non-null   object
 3   total_price                         5349 non-null   float64
 4   floor                               5349 non-null   object
 5   if_furnished                        5349 non-null   object
 6   building_type                       5349 non-null   object
 7   total_sqm                           5349 non-null   float64
 8   no_of_rooms                         5349 non-null   int32
 9   district                            5349 non-null   object
 10  accommodation                       5349 non-null   object
 11  is_street_mentioned                 5349 non-null   object
 12  is_distance_mentioned               5349 non-null   object
 13  is_type_of_communication_mentioned  5349 non-null   object
 14  is_shop_mentioned                   5349 non-null   object
dtypes: float64(2), int32(1), object(12)
memory usage: 606.1+ KB
```

promoted - Whether ad is promoted by a submitter (promoted ads have better positioning)
web_page - specifies web page on which ad was submitted
market_type - determines whether flat/room is being owned by company or a private person
total_price - total price for renting for month (expressed in PLN, which is EUR/PLN~4,5)
floor - floor on which flat/room is located
if_funrnished - determines whether flat/room is furnished
building_type - type of building in which flat/room is located
total_sqm - total sqm of flat/room
no_of_rooms - number of rooms in a flat
district - district in Kraków in which flat/room is located
accommodation - categorical variable : room or flat
is_street_mentioned - categorical variable : yes – if precise address is mentioned in ad description
is_distance_mentioned - categorical variable : yes – if distance to bus/tram stops/shops/restaurants are mentioned in ad description
is_type_of_communication_mentioned - categorical variable : yes – if nearest bus/tram/train stops are mentioned in ad description
is_shop_mentioned - categorical variable : yes – if nearest shop/shops are mentioned in ad description

2. Analysis objective

Objective of this analysis is to find best model (supervised machine learning or Neural Network) which will best predict price for monthly rent

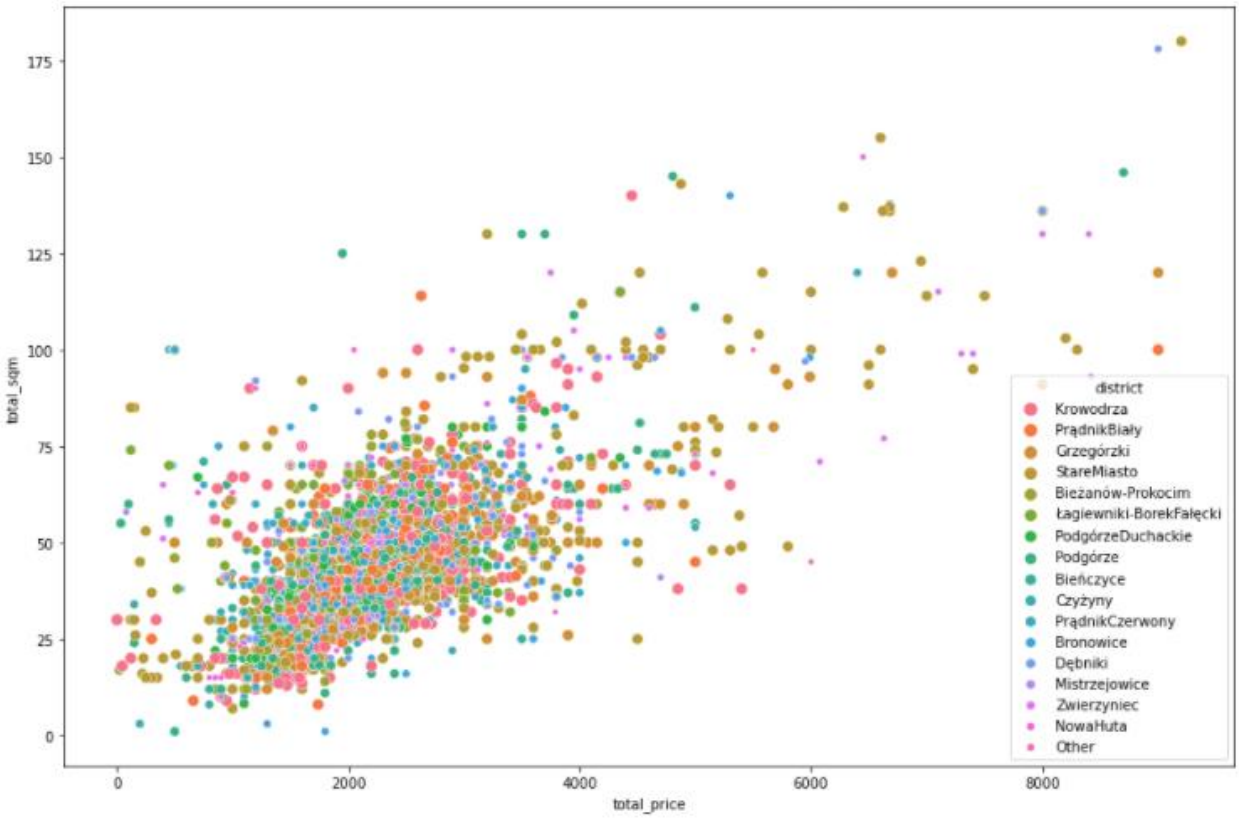
3. Plan for analysis

Analysis was performed in five steps

1. Data was cleaned
2. Two supervised machine learning models (XGBoost, Random Forest) were trained and evaluated using Randomized Grid Search CV
3. Three Unsupervised learning models (K-Means, Mean Shift, DBSCAN) were used for clustering to later improve fit of SL and NN models
4. Using Randomized Grid Search CV optimal Sequential Neural Network model was built , data used for training and testing this model was already clustered
5. Results were summarized

4. EDA

Relationship between price and total_sqm with respect to different districts:



Counts of ads from different districs and their average prices

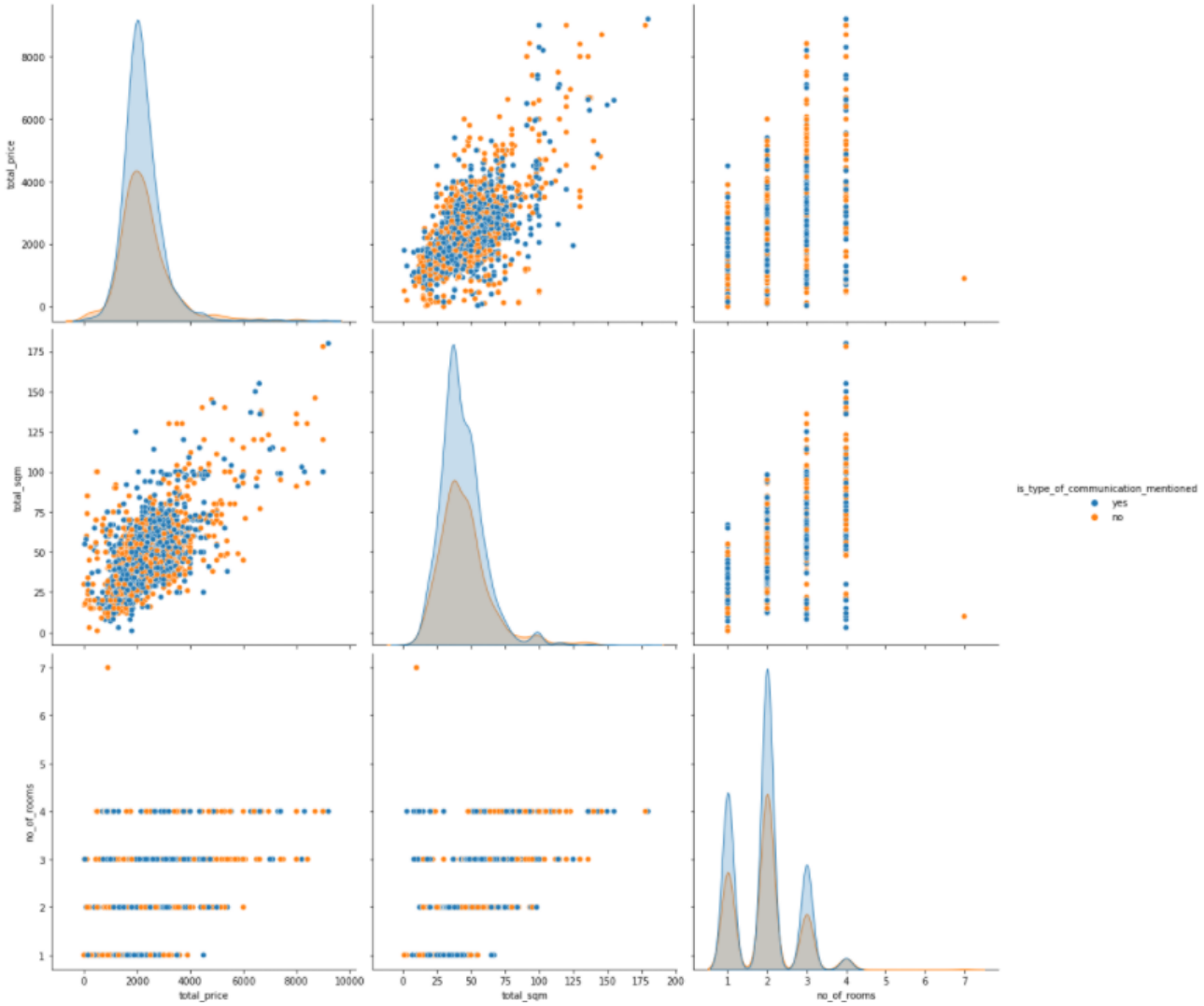
```
data_krk.district.value_counts()
✓ 0.4s
StareMiasto                662
Krowodrza                  562
PrądnikCzerwony            469
Grzegórzki                 462
Podgórze                   462
Dębniki                    428
PrądnikBiały               393
PodgórzeDuchackie          296
Bronowice                  280
Bieżanów-Prokocim          265
Łagiewniki-BorekFałęcki    251
Czyżyny                    183
Bieńczyce                  160
Mistrzejowice              137
NowaHuta                   127
Zwierzyniec                102
Other                       72
Name: district, dtype: int64
```

district	accommodation	total_price
Zwierzyniec	whole flat	3420.45
StareMiasto	whole flat	2548.08
Grzegórzki	whole flat	2524.53
Bronowice	whole flat	2326.08
Dębniki	whole flat	2320.36
Podgórze	whole flat	2277.66
Other	whole flat	2219.82
Krowodrza	whole flat	2206.00
PrądnikBiały	whole flat	2185.68
PrądnikCzerwony	whole flat	2081.07
Mistrzejowice	whole flat	2063.14
Łagiewniki-BorekFałęcki	whole flat	2058.56
Czyżyny	whole flat	2048.85
PodgórzeDuchackie	whole flat	2014.20
NowaHuta	whole flat	2007.07
Bieńczyce	whole flat	1970.70
Bieżanów-Prokocim	whole flat	1949.77
PodgórzeDuchackie	room	1546.67
Bieżanów-Prokocim	room	1540.00
Łagiewniki-BorekFałęcki	room	1490.00
Bronowice	room	1435.12
PrądnikBiały	room	1361.50
Krowodrza	room	1337.60
StareMiasto	room	1285.33
Bieńczyce	room	1275.00
Podgórze	room	1270.20
Dębniki	room	1224.58
PrądnikCzerwony	room	1145.07
Grzegórzki	room	1064.00
Czyżyny	room	1000.00
Other	room	900.00
NowaHuta	room	884.00
Zwierzyniec	room	400.00

Average sqm per district:

district	total_sqm
Zwierzyniec	63.88
Dębniki	47.05
StareMiasto	46.90
NowaHuta	45.92
Łagiewniki-BorekFałęcki	45.85
Other	45.13
Bronowice	44.90
PrądnikCzerwony	43.72
PrądnikBiały	43.41
Grzegórzki	43.36
Krowodrza	43.06
PodgórzeDuchackie	43.02
Mistrzejowice	41.56
Podgórze	41.46
Bieżanów-Prokocim	40.64
Bieńczyce	40.39
Czyżyny	39.14

Pairplot between numerical variables with respect to whether type of communication is mentioned in an AD:



5. Models Training

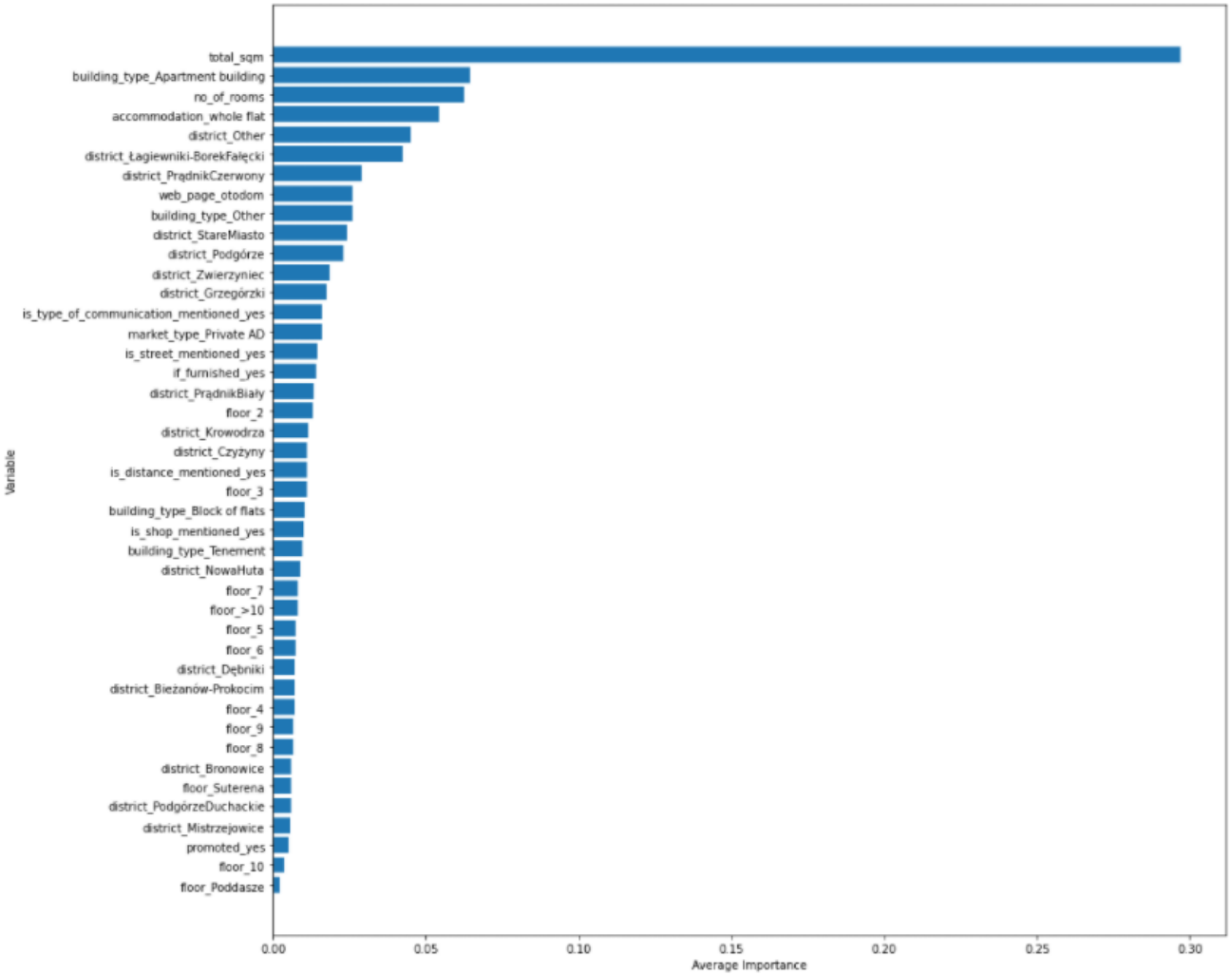
Section A:
SL models used:

-XGBRegressor
-RandomForestRegressor

Initial results after performing grid search and we can see XGBoost was clearly better:

	R2_train	MSE_train	R2_test	MSE_test
RandomForest	0.836326	0.1531	0.712496	0.36087
XGBRegressor	0.926784	0.068486	0.752124	0.311129

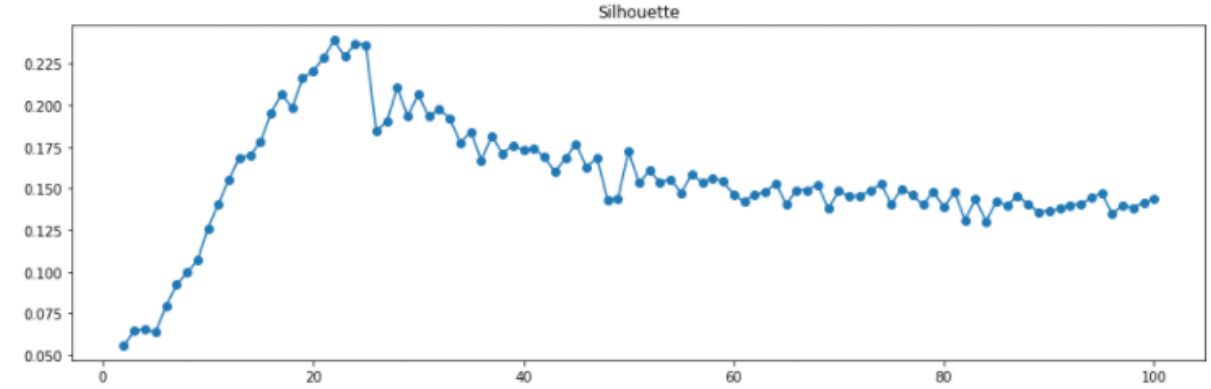
Features Importance:

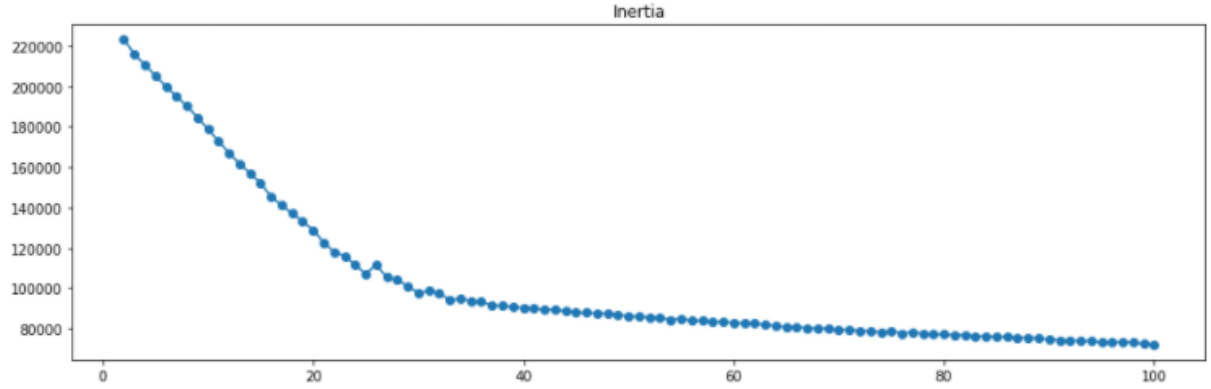


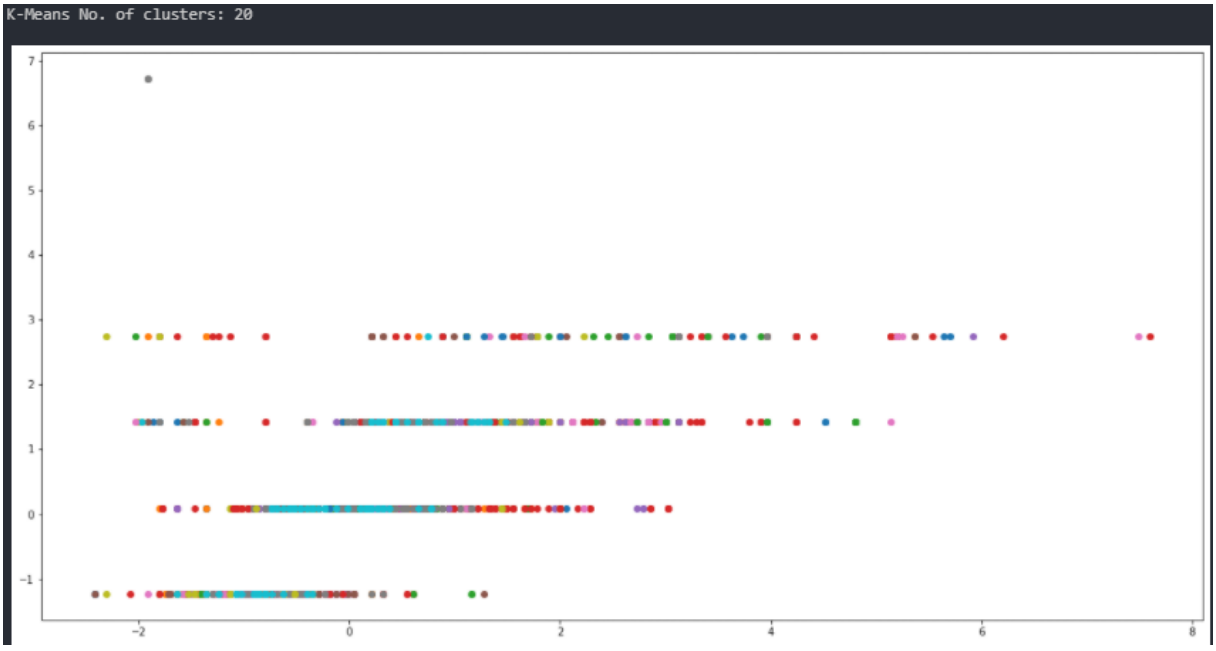
Section B:

UL models used:
-K-Means
-Mean Shift
-DBSCAN

1. K-Means

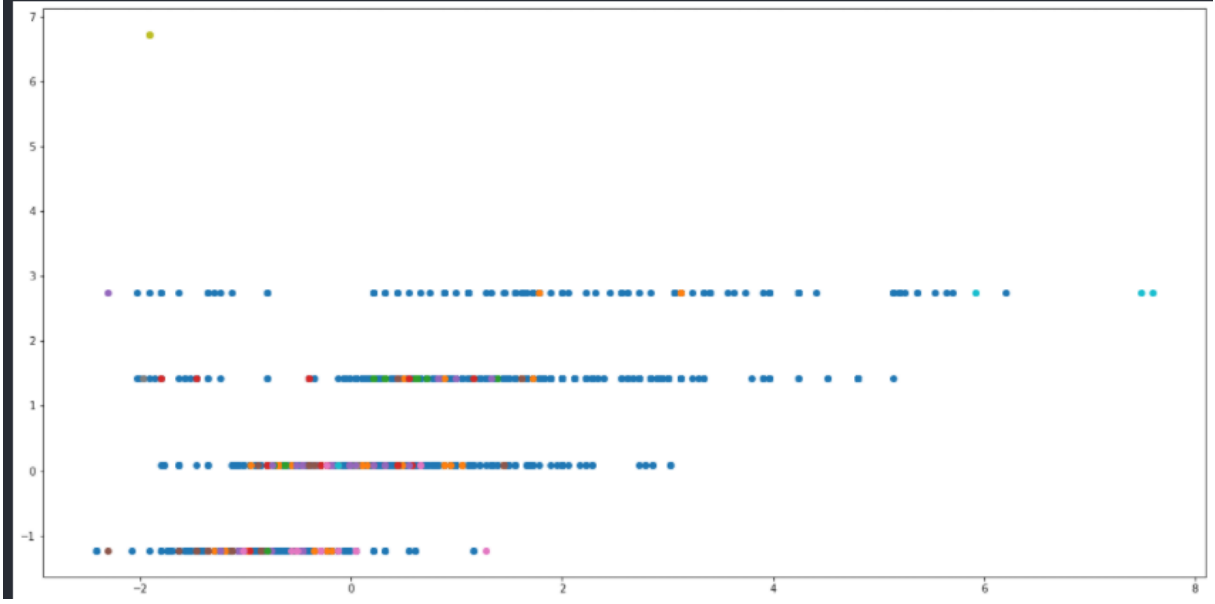






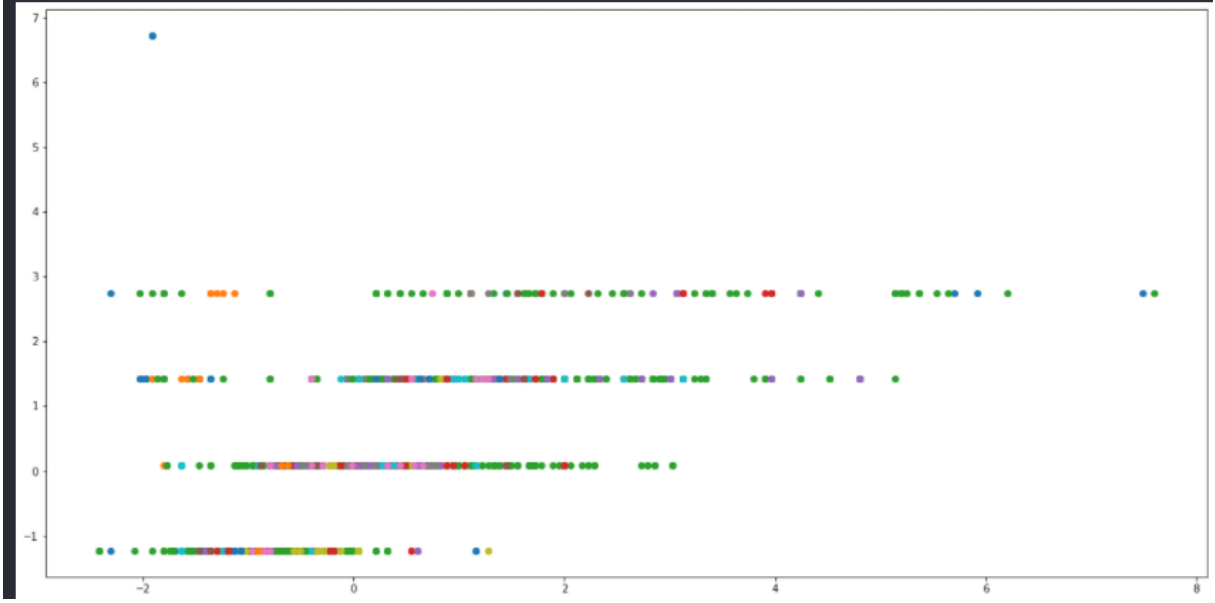
2. Mean Shift

```
Mean Shift No. of clusters: 16
```



3. DBSCAN

```
DBSCAN No. of clusters: 22
```



Section C:

Results SL with clustered data using XGBoost:

	No Unsupervised Learning	K-Means	Mean Shift	DBSCAN
R2_train	0.926784	0.925782	0.917275	0.926784
MSE_train	0.068486	0.069423	0.077381	0.068486
R2_test	0.752124	0.766278	0.750571	0.752124
MSE_test	0.311129	0.293364	0.313079	0.311129

To create Neural Network data clustered by K-Means model will be used

Section D:

After testing below parameters

```
NN_param_grid={'first_layer_init' : ['uniform', 'lecun_uniform', 'normal', 'zero', 'glorot_normal', 'glorot_uniform',
'he_normal', 'he_uniform'],
              'hidden_layer_init' : ['uniform', 'lecun_uniform', 'normal', 'zero', 'glorot_normal', 'glorot_uniform',
              'he_normal', 'he_uniform'],
              'last_layer_init' : ['uniform', 'lecun_uniform', 'normal', 'zero', 'glorot_normal', 'glorot_uniform',
              'he_normal', 'he_uniform'],

              'first_layer_density' : [10, 20, 30, 50, 100],
              'no_of_hidden_layers' : [1, 2, 3, 5, 10],
              'hidden_layer_density' : [10, 20, 30, 50, 100],

              'first_later_activation' : ['softmax', 'softplus', 'softsign', 'relu', 'tanh', 'sigmoid', 'hard_sigmoid',
              'linear'],
              'hidden_layer_activation' : ['softmax', 'softplus', 'softsign', 'relu', 'tanh', 'sigmoid', 'hard_sigmoid',
              'linear'],
              'last_layer_activation' : ['softmax', 'softplus', 'softsign', 'relu', 'tanh', 'sigmoid', 'hard_sigmoid', 'linear']
              ,

              'optimizer' : ['SGD', 'RMSprop', 'Adagrad', 'Adadelta', 'Adam', 'Adamax', 'Nadam']

             }

NN_results=train_test_model(model,X_km,Y_km,param_grid=NN_param_grid)

✓ 44m 45.3s                                                                    Python
```

Optimal NN model parameters were selected:

```
NN_results['Test_results']['Best_params']
✓ 0.5s
{'optimizer': 'SGD',
 'no_of_hidden_layers': 5,
 'last_layer_init': 'he_uniform',
 'last_layer_activation': 'tanh',
 'hidden_layer_init': 'lecun_uniform',
 'hidden_layer_density': 100,
 'hidden_layer_activation': 'tanh',
 'first_layer_init': 'uniform',
 'first_layer_density': 20,
 'first_later_activation': 'softplus'}
```

Which resulted in below R2:

```
NN_results['Test_results']['R2']
✓ 0.5s
0.42671311599670647
```

6. Summary and Results

During this analysis, three types of machine learning techniques were used in order to find optimal model which will correctly predict rent values for flats or rooms in a given district in Kraków.

Interesting findings are:

- most important parameter is total area of flat and type of building in which flat is located
- web page on which a person is looking for flat or room is important

Best model to perform this task of valuation is XGBoost with parameters:

```
XGBRegressor(n_estimators = 500, max_depth = 6, eta = 0.08, booster = 'gbtree')
```

There is a chance to improve this model, by:

- better segmentation (to tune DBSCAN or Mean Shift clustering models)
- to add more hyperparameters to Neural Network grid search