# Supervised Machine Learning: Regression

Final Peer Assignment

## Objective of the analysis

A goal of this project is focused on Linear Regression models to predict prices of houses in Kansas City (Washington, USA). Simple linear, with polynomial features and ridge regressions were created to analyze which model suites best to a dataset.

## Dataset description

The dataset "KC_house_data" was downloaded from a Kaggle.com (https://www.kaggle.com/astronautelvis/kc-house-data), but it originates from the UCI Machine Learning Repository.
This dataset originally contains 21613 rows and 22 columns with following features:
**id** - Unique ID for each home sold
**date** - Date of the home sale
**price** - Price of each home sold
**bedrooms** - Number of bedrooms
**bathrooms** - Number of bathrooms, where .5 accounts for a room with a toilet but no shower
**sqft_living** - Square footage of the apartment interior living space
**sqft_lot** - Square footage of the land space
**floors** - Number of floors
**waterfront** - A dummy variable for whether the apartment was overlooking the waterfront or not
**view** - An index from 0 to 4 of how good the view of the property was
**condition** - An index from 1 to 5 on the condition of the apartment,
**grade** - An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high-quality level of construction and design.
**sqft_above** - The square footage of the interior housing space that is above ground level
**sqft_basement** - The square footage of the interior housing space that is below ground level
**yr_built** - The year the house was initially built
**yr_renovated** - The year of the house's last renovation
**zipcode** - What zipcode area the house is in
**lat** - Lattitude
**long** - Longitude
**sqft_living15** - The square footage of interior housing living space for the nearest 15 neighbors
**sqft_lot15** - The square footage of the land lots of the nearest 15 neighbors

## Data exploration analysis

The first step in data exploration is to gather initial information with methods: .shape, .columns, .dtypes, and .info(). This basic action shows how big is our data set, which types of data columns contain, whether there are missed values or not. In this case, the data frame has 21613 rows and 22 columns, has no issues with missed values and incorrect data types. But for further data analysis and modelling columns "Unnamed: 0","id", "zipcode", "date", "lat" and "long" were dropped. Columns with units in the British imperial and United States customary systems of measurement (e.g. "**sqft_living**", "**sqft_lot**" and etc.) were converted to SI system ($m^2$). After initial analysis and manipulations with the dataset there 17 features with units in SI measurement system. Afterwards method .describe was used to get a basic statistics of the dataset.

0]:

| | price | bedrooms | bathrooms | area_qm.m | lot_qm | floors | waterfront | view | condition | grade | above_qm | basement_qm | house_age | yr_renovated | living15_qm | lot15_qm | sqft_living15t |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 221900.0 | 3 | 1.00 | 109.624675 | 524.897808 | 1.0 | 0 | 0 | 3 | 7 | 109.624675 | 0.000000 | 66 | 0 | 1340.0 | 524.897808 | 124.489038 |
| 1 | 538000.0 | 3 | 2.25 | 238.758826 | 672.796216 | 2.0 | 0 | 0 | 3 | 7 | 201.597919 | 37.160907 | 70 | 1991 | 1690.0 | 709.680416 | 157.004831 |
| 2 | 180000.0 | 2 | 1.00 | 71.534745 | 929.022668 | 1.0 | 0 | 0 | 3 | 6 | 71.534745 | 0.000000 | 88 | 0 | 2720.0 | 748.978075 | 252.694166 |
| 3 | 604000.0 | 4 | 3.00 | 182.088443 | 464.511334 | 1.0 | 0 | 0 | 5 | 7 | 97.547380 | 84.541063 | 56 | 0 | 1360.0 | 464.511334 | 126.347083 |
| 4 | 510000.0 | 3 | 2.00 | 156.075808 | 750.650316 | 1.0 | 0 | 0 | 3 | 8 | 156.075808 | 0.000000 | 34 | 0 | 1800.0 | 697.045708 | 167.224080 |

1]:
```
df.describe()
```

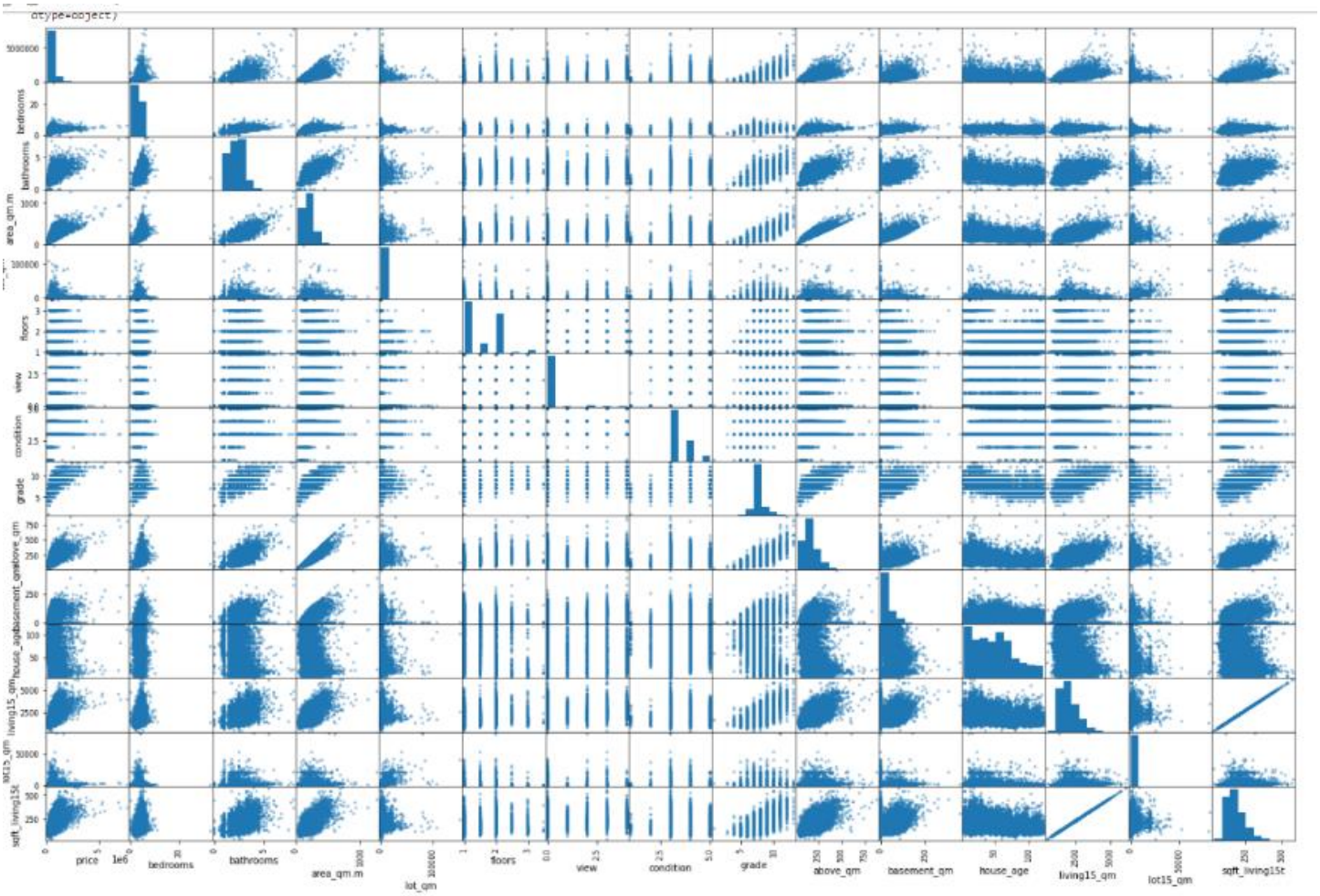
1]:

| | price | bedrooms | bathrooms | area_qm.m | lot_qm | floors | waterfront | view | condition | grade | above_qm | basement_qm | house_age | yr_renovated | living15_qm | lot15_qm | sqft_living15t |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 2.161300e+04 | 21613.000000 | 21613.000000 | 21613.000000 | 21613.000000 | 21613.000000 | 21613.000000 | 21613.000000 | 21613.000000 | 21613.000000 | 21613.000000 | 21613.000000 | 21613.000000 | 21613.000000 | 21613.000000 | 21613.000000 | 21613.000000 |
| mean | 5.400881e+05 | 3.370842 | 2.114757 | 193.227400 | 1403.471532 | 1.494309 | 0.007542 | 0.234303 | 3.409430 | 7.656873 | 166.145549 | 27.081851 | 49.994864 | 84.402258 | 1986.552492 | 1186.218474 | 184.555230 |
| std | 3.671272e+05 | 0.930062 | 0.770163 | 85.325241 | 3848.059412 | 0.539989 | 0.086517 | 0.766318 | 0.650743 | 1.175459 | 76.931529 | 41.116225 | 29.373411 | 401.679240 | 685.391304 | 2536.620181 | 63.674406 |
| min | 7.500000e+04 | 0.000000 | 0.000000 | 26.941657 | 48.309179 | 1.000000 | 0.000000 | 0.000000 | 1.000000 | 1.000000 | 26.941657 | 0.000000 | 6.000000 | 0.000000 | 399.000000 | 60.479376 | 37.068004 |
| 25% | 3.219500e+05 | 3.000000 | 1.750000 | 132.571535 | 468.227425 | 1.000000 | 0.000000 | 0.000000 | 3.000000 | 7.000000 | 110.553698 | 0.000000 | 24.000000 | 0.000000 | 1490.000000 | 473.801561 | 138.424378 |
| 50% | 4.500000e+05 | 3.000000 | 2.250000 | 177.443330 | 707.729469 | 1.500000 | 0.000000 | 0.000000 | 3.000000 | 7.000000 | 144.927536 | 0.000000 | 46.000000 | 0.000000 | 1840.000000 | 707.915273 | 170.940171 |
| 75% | 6.450000e+05 | 4.000000 | 2.500000 | 236.900780 | 992.939428 | 2.000000 | 0.000000 | 0.000000 | 4.000000 | 8.000000 | 205.314010 | 52.025269 | 70.000000 | 0.000000 | 2360.000000 | 936.733556 | 219.249350 |
| max | 7.700000e+06 | 33.000000 | 8.000000 | 1257.896693 | 153414.994426 | 3.500000 | 1.000000 | 4.000000 | 5.000000 | 13.000000 | 874.210331 | 447.788926 | 121.000000 | 2015.000000 | 6210.000000 | 80936.454849 | 576.923077 |

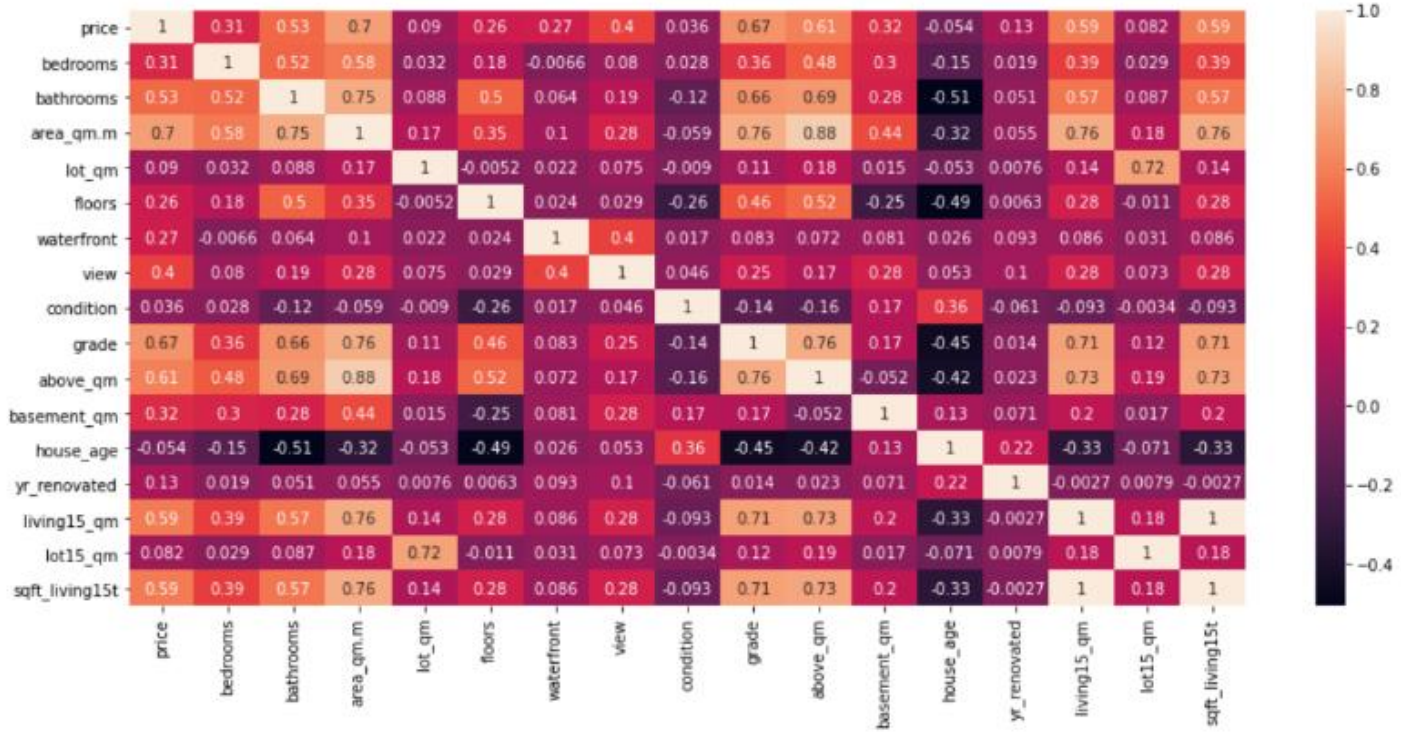
The next step was to see a correlation between features. For this purpose a method .corr( ) was applied. For better understanding, this matrix was visualized as a scatter matrix and heat map (see pictures below).

]:

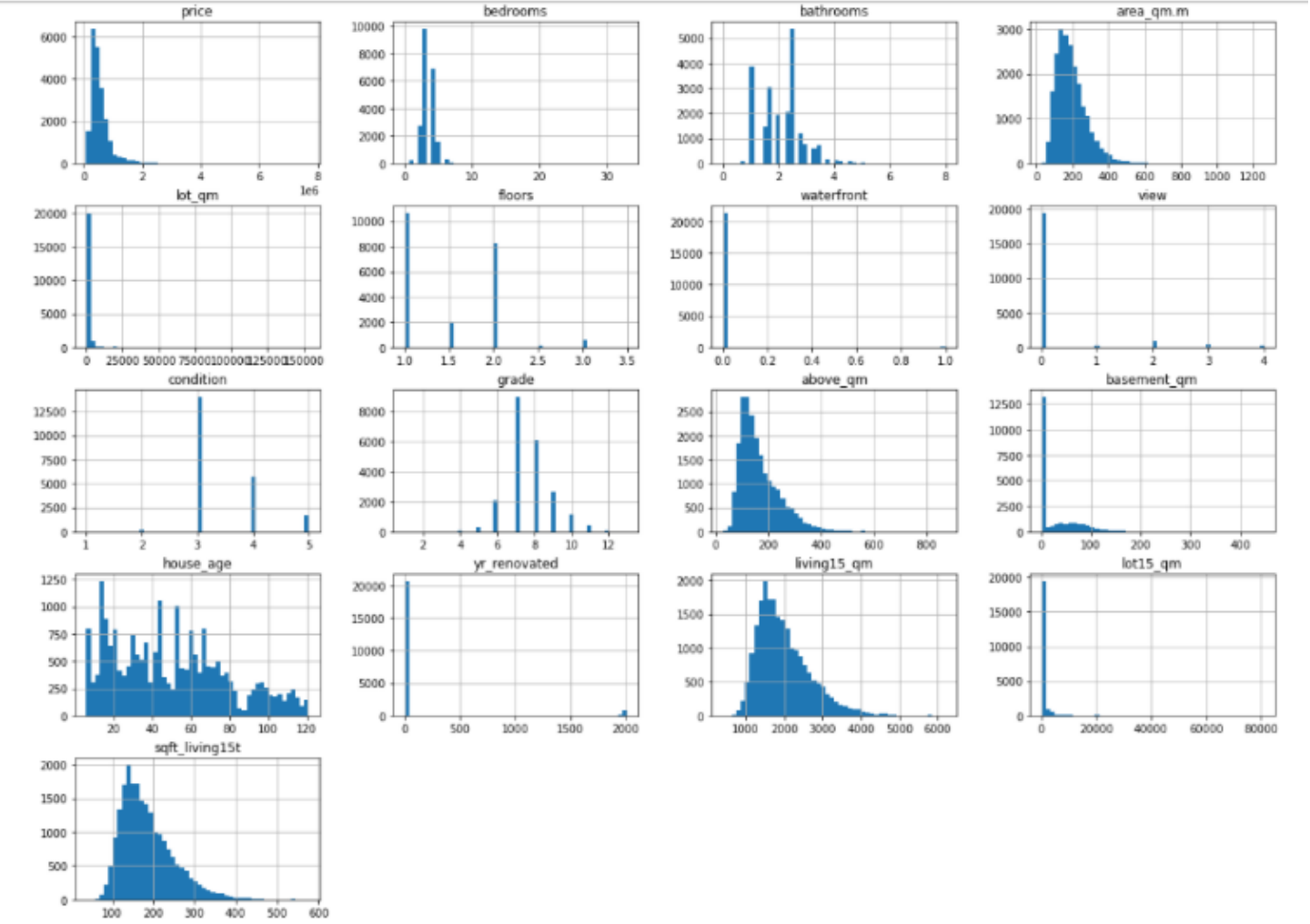
| | price | bedrooms | bathrooms | area_qm.m | lot_qm | floors | waterfront | view | condition | grade | above_qm | basement_qm | house_age | yr_renovated | living15_qm | lot15_qm | sqft_living15t |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| price | 1.000000 | 0.308350 | 0.525138 | 0.702035 | 0.089661 | 0.256794 | 0.266369 | 0.397293 | 0.036362 | 0.667434 | 0.605567 | 0.323816 | -0.054012 | 0.126434 | 0.585379 | 0.082447 | 0.585379 |
| bedrooms | 0.308350 | 1.000000 | 0.515884 | 0.576671 | 0.031703 | 0.175429 | -0.006582 | 0.079532 | 0.028472 | 0.356967 | 0.477600 | 0.303093 | -0.154178 | 0.018841 | 0.391638 | 0.029244 | 0.391638 |
| bathrooms | 0.525138 | 0.515884 | 1.000000 | 0.754665 | 0.087740 | 0.500653 | 0.063744 | 0.187737 | -0.124982 | 0.664983 | 0.685342 | 0.283770 | -0.506019 | 0.050739 | 0.568634 | 0.087175 | 0.568634 |
| area_qm.m | 0.702035 | 0.576671 | 0.754665 | 1.000000 | 0.172826 | 0.353949 | 0.103818 | 0.284611 | -0.058753 | 0.762704 | 0.876597 | 0.435043 | -0.318049 | 0.055363 | 0.756420 | 0.183286 | 0.756420 |
| lot_qm | 0.089661 | 0.031703 | 0.087740 | 0.172826 | 1.000000 | -0.005201 | 0.021604 | 0.074710 | -0.008958 | 0.113621 | 0.183512 | 0.015286 | -0.053080 | 0.007644 | 0.144608 | 0.718557 | 0.144608 |
| floors | 0.256794 | 0.175429 | 0.500653 | 0.353949 | -0.005201 | 1.000000 | 0.023698 | 0.029444 | -0.263768 | 0.458183 | 0.523885 | -0.245705 | -0.489319 | 0.006338 | 0.279885 | -0.011269 | 0.279885 |
| waterfront | 0.266369 | -0.006582 | 0.063744 | 0.103818 | 0.021604 | 0.023698 | 1.000000 | 0.401857 | 0.016653 | 0.082775 | 0.072075 | 0.080588 | 0.026161 | 0.092885 | 0.086463 | 0.030703 | 0.086463 |
| view | 0.397293 | 0.079532 | 0.187737 | 0.284611 | 0.074710 | 0.029444 | 0.401857 | 1.000000 | 0.045990 | 0.251321 | 0.167649 | 0.276947 | 0.053440 | 0.103917 | 0.280439 | 0.072575 | 0.280439 |
| condition | 0.036362 | 0.028472 | -0.124982 | -0.058753 | -0.008958 | -0.263768 | 0.016653 | 0.045990 | 1.000000 | -0.144674 | -0.158214 | 0.174105 | 0.361417 | -0.060618 | -0.092824 | -0.003406 | -0.092824 |
| grade | 0.667434 | 0.356967 | 0.664983 | 0.762704 | 0.113621 | 0.458183 | 0.082775 | 0.251321 | -0.144674 | 1.000000 | 0.755923 | 0.168392 | -0.446963 | 0.014414 | 0.713202 | 0.119248 | 0.713202 |
| above_qm | 0.605567 | 0.477600 | 0.685342 | 0.876597 | 0.183512 | 0.523885 | 0.072075 | 0.167649 | -0.158214 | 0.755923 | 1.000000 | -0.051943 | -0.423898 | 0.023285 | 0.731870 | 0.194050 | 0.731870 |
| basement_qm | 0.323816 | 0.303093 | 0.283770 | 0.435043 | 0.015286 | -0.245705 | 0.080588 | 0.276947 | 0.174105 | 0.168392 | -0.051943 | 1.000000 | 0.133124 | 0.071323 | 0.200355 | 0.017276 | 0.200355 |
| house_age | -0.054012 | -0.154178 | -0.506019 | -0.318049 | -0.053080 | -0.489319 | 0.026161 | 0.053440 | 0.361417 | -0.446963 | -0.423898 | 0.133124 | 1.000000 | 0.224874 | -0.326229 | -0.070958 | -0.326229 |
| yr_renovated | 0.126434 | 0.018841 | 0.050739 | 0.055363 | 0.007644 | 0.006338 | 0.092885 | 0.103917 | -0.060618 | 0.014414 | 0.023285 | 0.071323 | 0.224874 | 1.000000 | -0.002673 | 0.007854 | -0.002673 |
| living15_qm | 0.585379 | 0.391638 | 0.568634 | 0.756420 | 0.144608 | 0.279885 | 0.086463 | 0.280439 | -0.092824 | 0.713202 | 0.731870 | 0.200355 | -0.326229 | -0.002673 | 1.000000 | 0.183192 | 1.000000 |
| lot15_qm | 0.082447 | 0.029244 | 0.087175 | 0.183286 | 0.718557 | -0.011269 | 0.030703 | 0.072575 | -0.003406 | 0.119248 | 0.194050 | 0.017276 | -0.070958 | 0.007854 | 0.183192 | 1.000000 | 0.183192 |
| sqft_living15t | 0.585379 | 0.391638 | 0.568634 | 0.756420 | 0.144608 | 0.279885 | 0.086463 | 0.280439 | -0.092824 | 0.713202 | 0.731870 | 0.200355 | -0.326229 | -0.002673 | 1.000000 | 0.183192 | 1.000000 |




<AxesSubplot:>



As it is shown above, our target feature “Price” has rather strong correlation with house area, grade and area of the interior housing space that is above ground level. The weakest correlation is with the area of land space, house age and condition.




The distribution of our features shows that features “price”, “area_qm.m”, “above_qm”, “living15_qm” and “sqft_living15t” have tails on the right. Other features are discrete (e.g. number of bathrooms, grade and so on) have discrete numbers. Feature “house_age” has interesting distribution: the main number of values lies between ~5 and ~ 80, with some peaks.

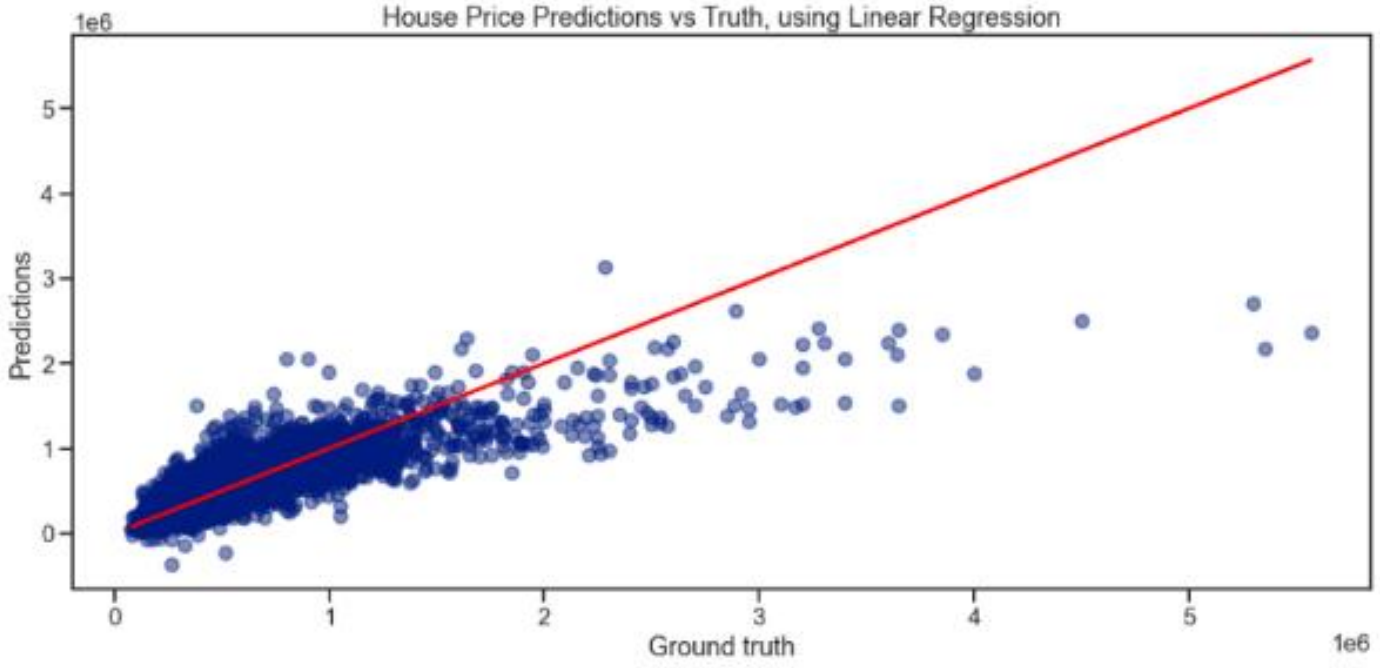
## Regressions

As a first step a basic Linear Regression with standard scaler. In this project feature “price” is our target, therefore it was set as y, other features were set X. The dataset was split to train and test subsets in ratio 70/30, with a random state 42.

This Linear Regression has following coefficients:

| Coefficients | | Interception |
|---|---|---|
| $a_1$ | -3.63641428e+04 | 541293.4336916765 |
| $a_2$ | 3.40609429e+04 | |
| $a_3$ | -4.11747147e+17 | |
| $a_4$ | -1.44647642e+03 | |
| $a_5$ | 1.45385607e+04 | |
| $a_6$ | 4.55195829e+04 | |
| $a_7$ | 3.36194745e+04 | |
| $a_8$ | 9.83797245e+03 | |
| $a_9$ | 1.38949634e+05 | |
| $a_{10}$ | 3.73655447e+17 | |
| $a_{11}$ | 1.98112250e+17 | |
| $a_{12}$ | 1.03568283e+05 | |
| $a_{13}$ | 5.17467041e+03 | |
| $a_{14}$ | 1.66243004e+16 | |
| $a_{15}$ | -1.34070317e+04 | |
| $a_{16}$ | -1.66243004e+16 | |

The $R^2$ score for this model equals to 0,6527 which is not a good result.



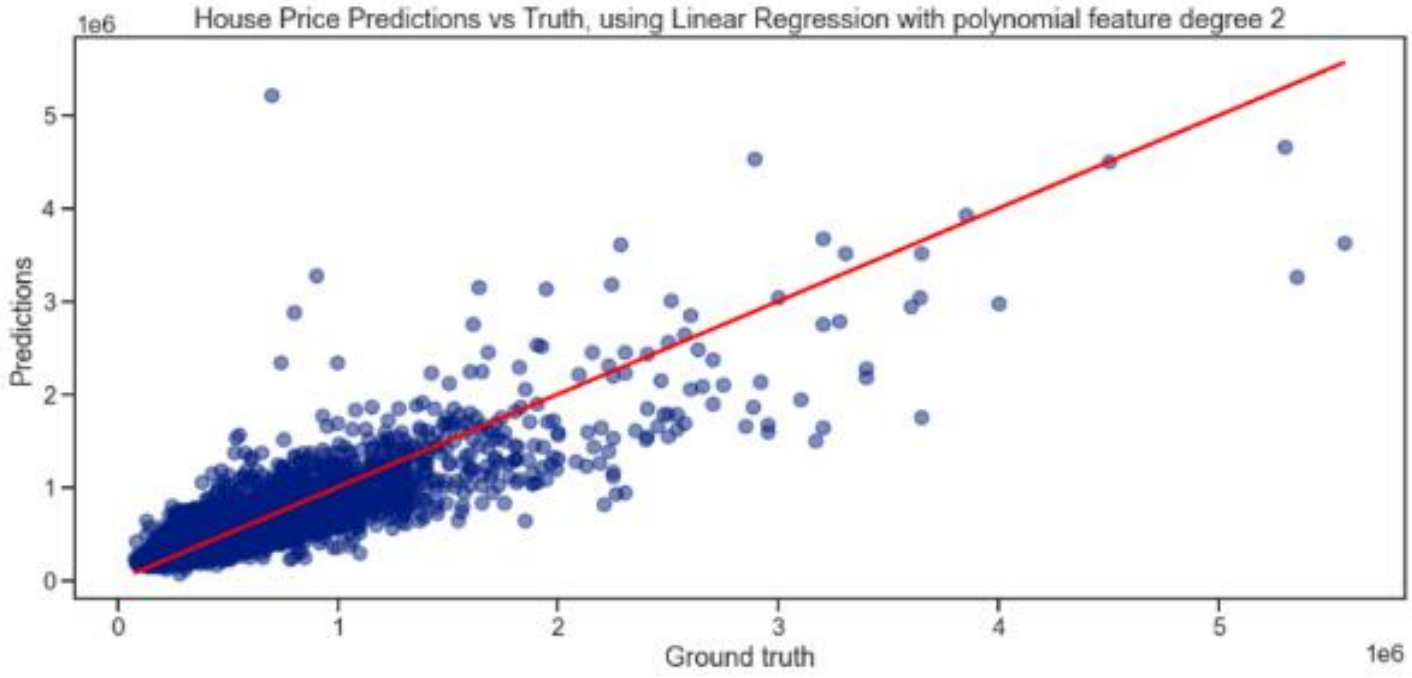

Linear Regression was also calculated with the use of cross-validation technique. The following parameters for KFold were used:

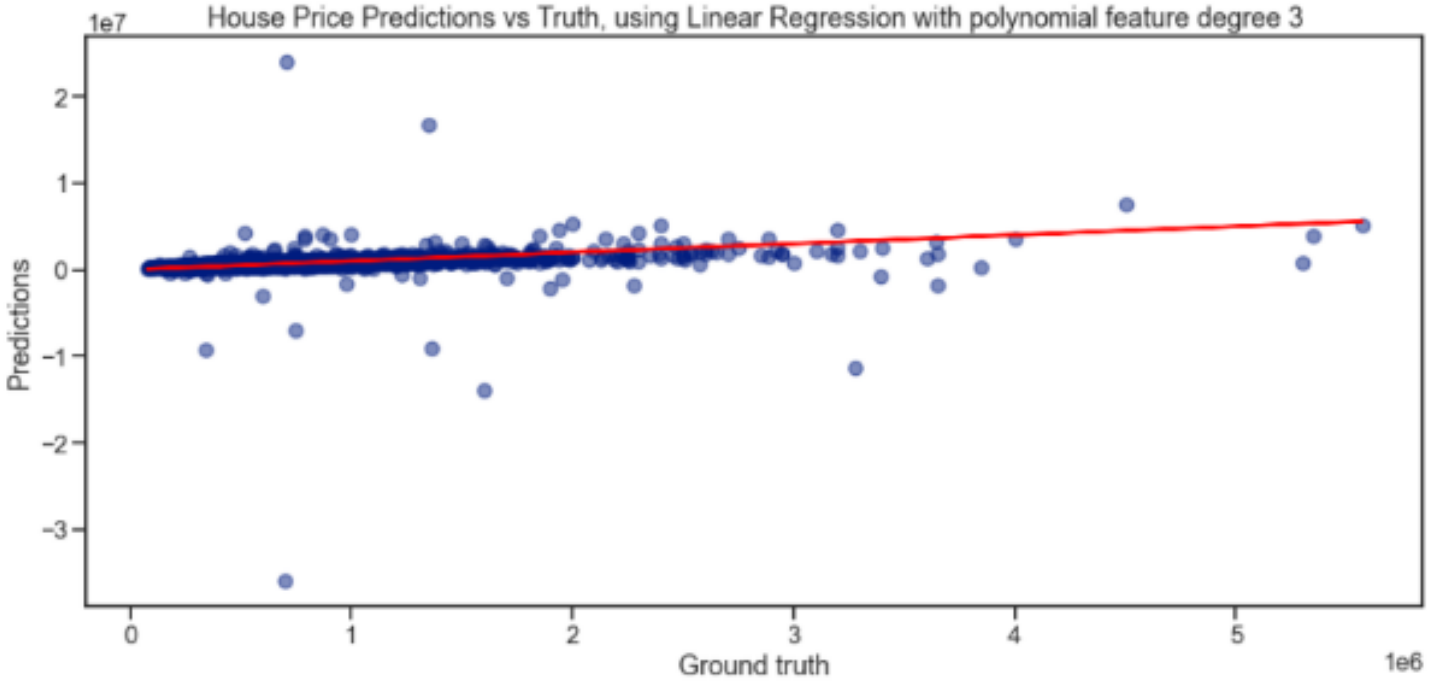
*kf = KFold (shuffle = True, random_state = 42, n_splits = 4)*

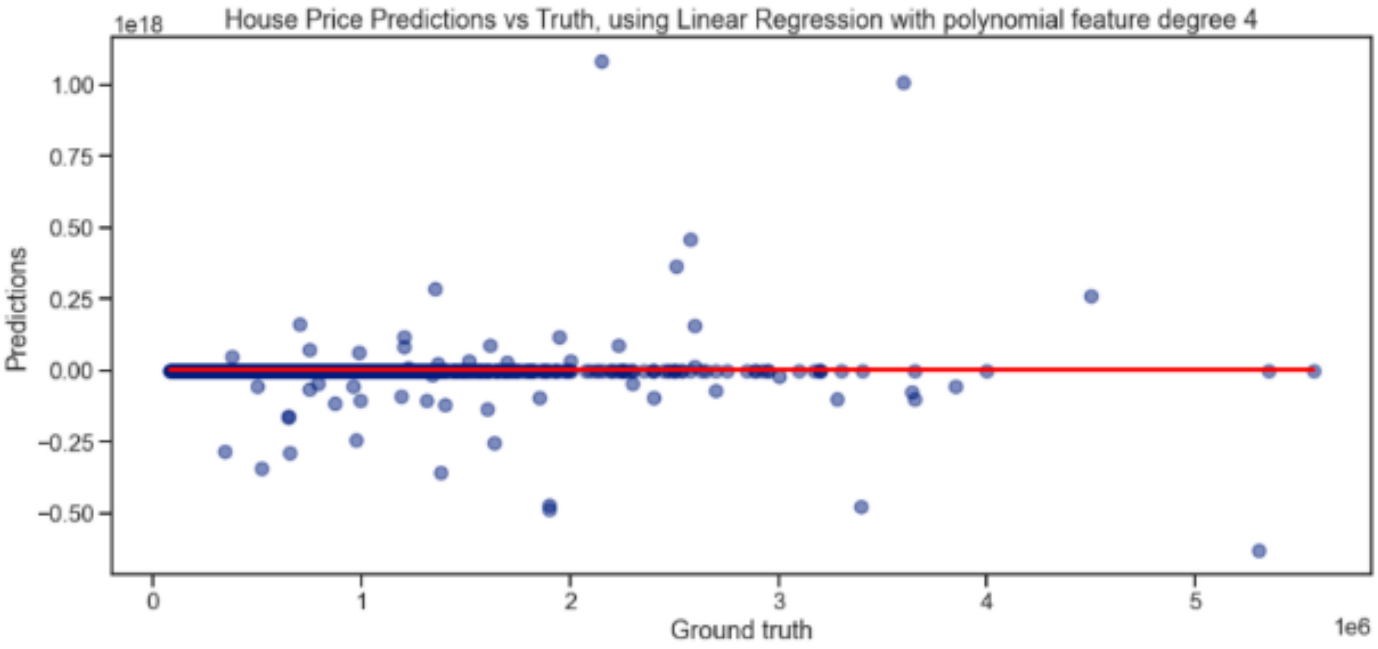
$R^2$ score is 0.6528, which is slightly better.

Linear regression with polynomial features 2nd, 3d and 4th degrees were used to built linear regressions. These models have following $R^2$ scores:

| Polynomial Feature Degree | $R^2$ score |
|---|---|
| 2 | 0.70517 |
| 3 | -2.5885 |
| 4 | -4.94109e+21 |



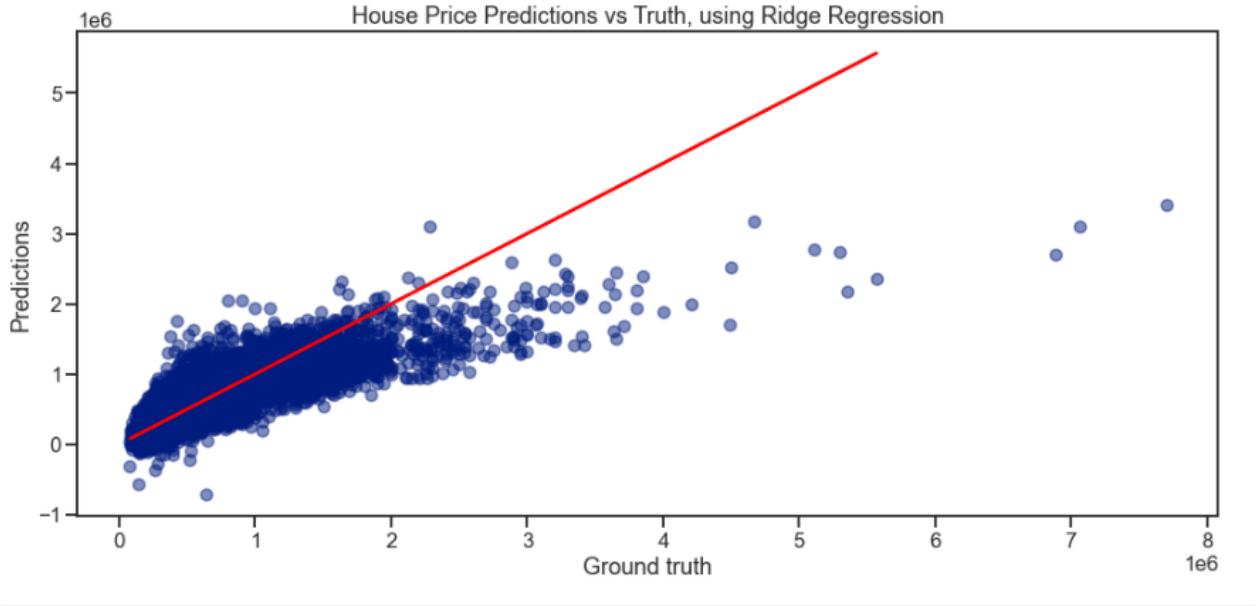



As it can been seen from the table above and pictures, polynomial degree 2 is optimal for our Linear Regression, but $R^2$ score is still relatively low, but of course better than simple Linear Regression. Therefore, Linear Regression with polynomial feature with cross - validation was performed only for the $2^d$ degree. In this case $R^2$ score equals to 0.729417.

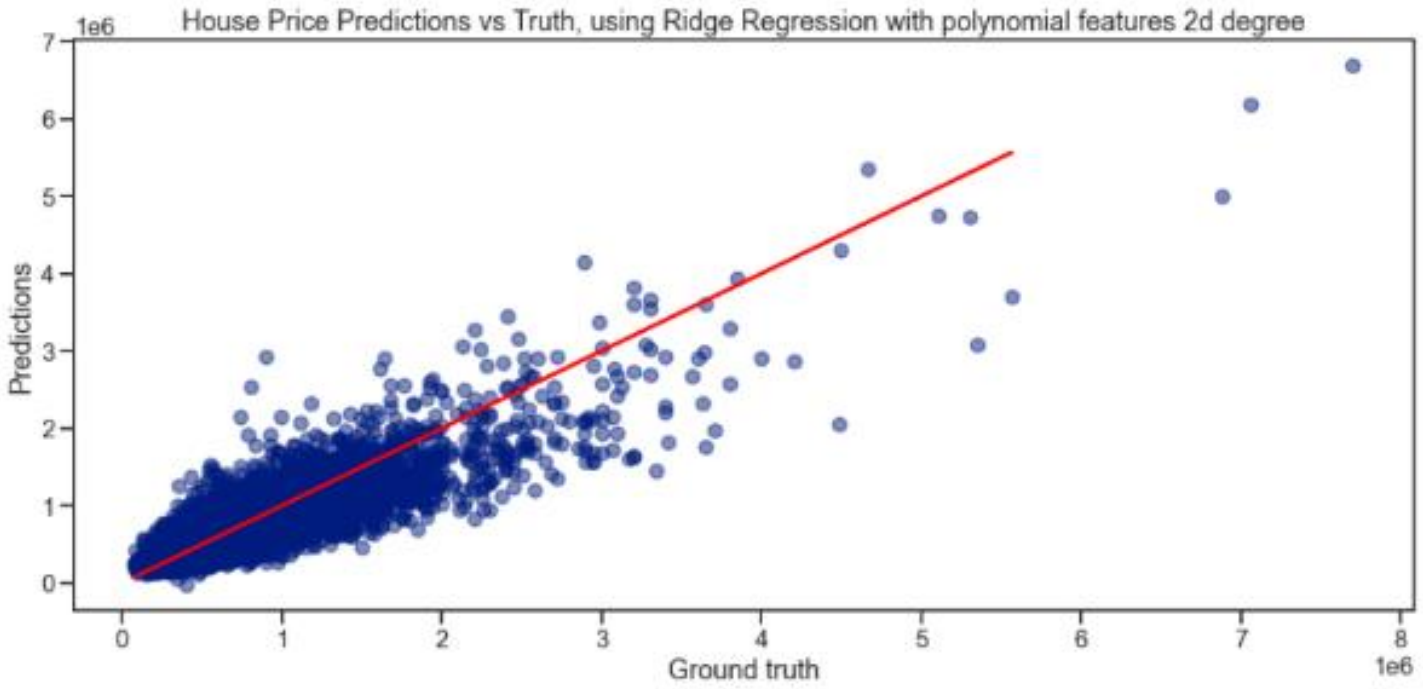
Ridge Regression was executed with Grid Search CV. Here pipeline and cross validation techniques were used. For Linear Regression optimal parameters, such as coefficients and alpha, were calculated. For alpha parameter following range was used:

*alphas = np.geomspace(0.1, 20, num = 1000)*

Here the best parameters are: $R^2\_score = 0.653731$ and *alpha = 20.0*.



Better results were obtained for Linear Regression with polynomial feature $2^d$ degree. In this case, best parameters are: *alpha = 0.1* and *R2_score = 0.7515105250174143*




## Conclusion

In this project different regressions were modelled. The best suited one is Linear Regression with polynomial feature $2^d$ degree. But still this model is not ideal for predictions based on given dataset. From my point of view, there is sense to try out other types of models, e.g. Random Forest Regression.