In the name of Allah most
Gracious most merciful

-University of Bahri
College of Computer
Science and math

Name :Abdulsalam yagoub idris Alhellou

Title :project houses price .

I want predict price of KC city house for some features, when I predicted this houses price if any one need puy or sales I can get them their house price

1.data propagation

I read data by csv file ,I used head ()function to print first 5 rows and eatch columns in dataset,I used isna() to see if there missing data or not ,the data is clean,and I used describe function to see Statistical data. And use corr() to

see correlation between price and other features, I find some features it is not impact for the price, we can delete it.

1.1. Features selection

I have 16 columns in x , this features is alot I used SelectKBest to select best 7 features (there, are another class we can use to select features like SelectPercentile and SelectFromModel)

1.2 Dtata splitting

Explain in code

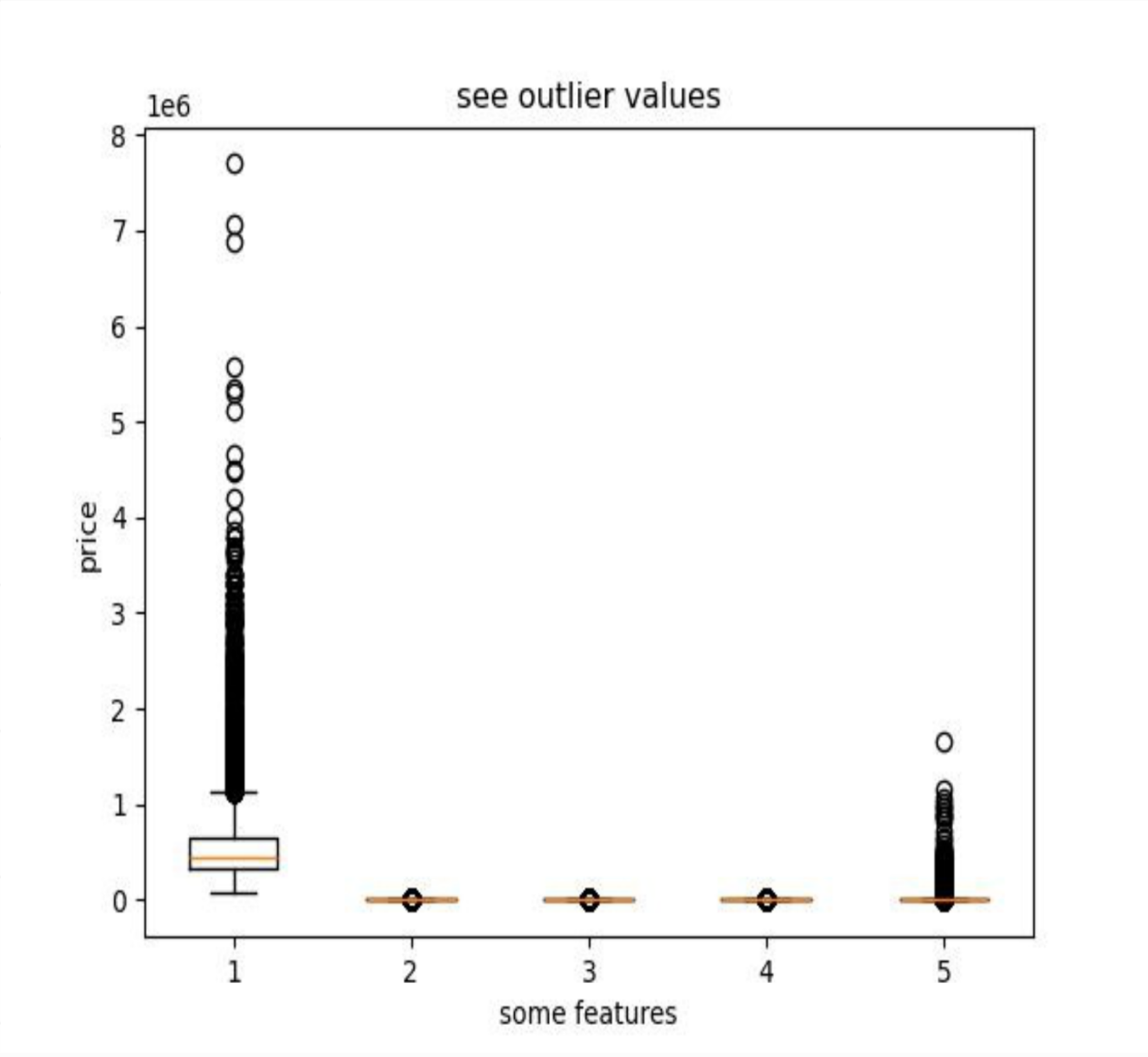
1. 3 Data scalling

Scale data in Small range, that make model work best(there are may type like StandardScaler and MinMaxScaler)

1.4 Data visualization (matplotlib library)

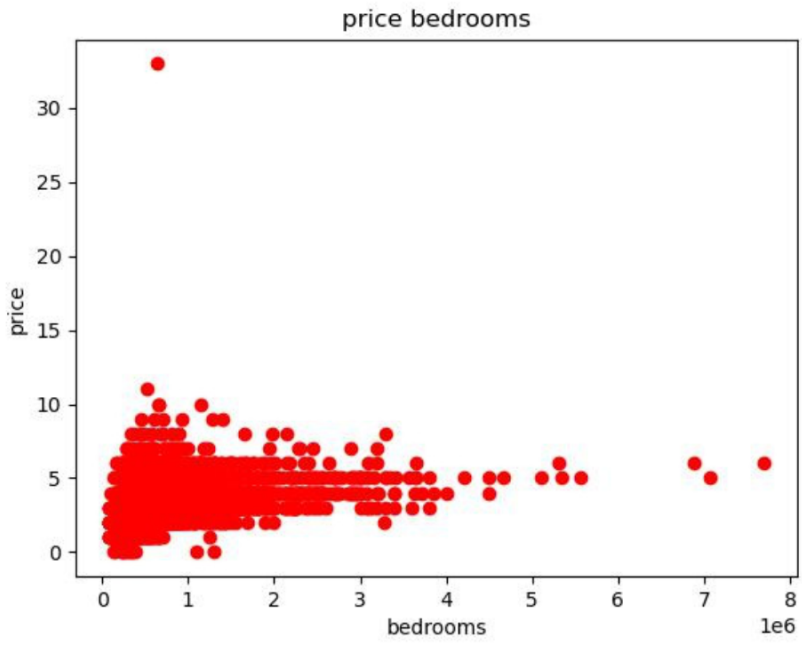
1.4.1 boxplot

I used boxplot to see data if it has outlier or not

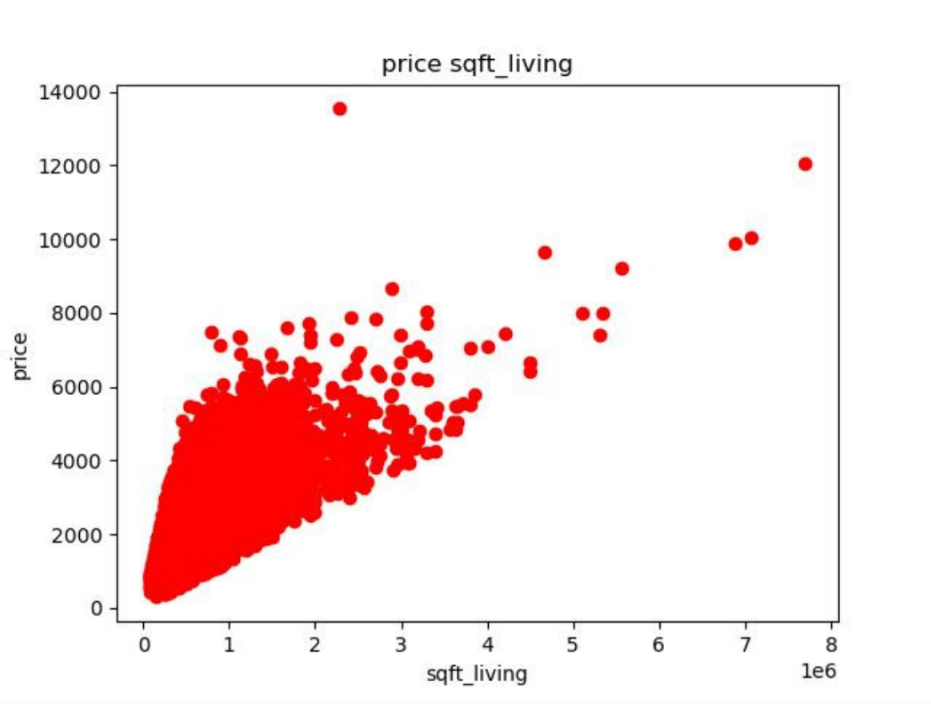


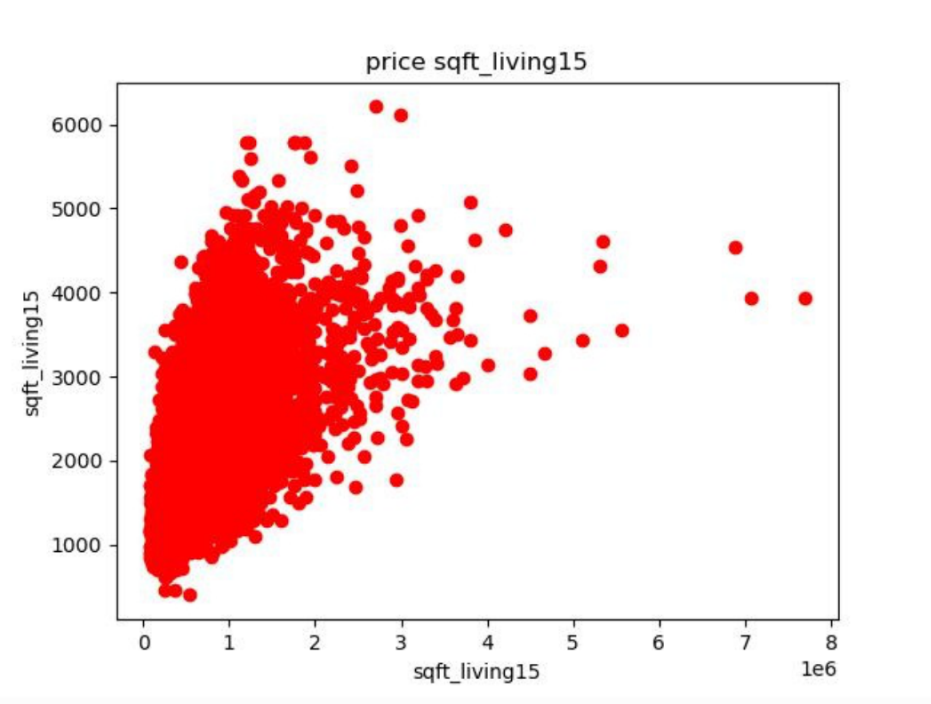
1.4.2 scatter plot

When I want show data after divided to x and y it give me error.I used some features from database like (sqft_living, sqft_lot)to draw scatter plot


price bedrooms
price
bedrooms
1e6
0
5
10
15
20
25
30
0
1
2
3
4
5
6
7
8


price bathrooms
price
bathrooms
1e6
0
1
2
3
4
5
6
7
8
0
1
2
3
4
5
6
7
8





The best model to train this database is model regression , because it can fit data and give us best solution

2. Training data

In this point we can use fit function to make

train between x_train, y_train

2.selecting model
I used linear regression but it isn't get me good results after that select RandomForestRegressor model to predict x_test values, it is best than LinearRegression in this database because r2_score more than 40%
3. Evaluate model
The model is good but the errors rate so higher , I used scalling data to reduce performance