

Housing Project

Submitted by:

DARSHAN M KATGERI

**INTRODUCTION**

* **Business Problem Framing**

House prices are increasing every year which has necessitated the modelling of house price prediction. These models constructed, help the customers to purchase a house suitable for their need

* **Conceptual Background of the Domain Problem**

Houses are one of the necessary needs of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies.

* **Review of Literature**

Dr. M. Thamarai, Dr. S P. Malarvizhi used two algorithms like decision tree regression and multiple linear regression were used in predicting the prices of the houses. Comparatively the performance of multiple linear regression is found to be better than the decision tree regression in predicting the house prices.

* **Motivation for the Problem Undertaken**

House prices are increasing every year which has necessitated the modelling of house price prediction. These models constructed, help the customers to purchase a house suitable for their need

**Analytical Problem Framing**

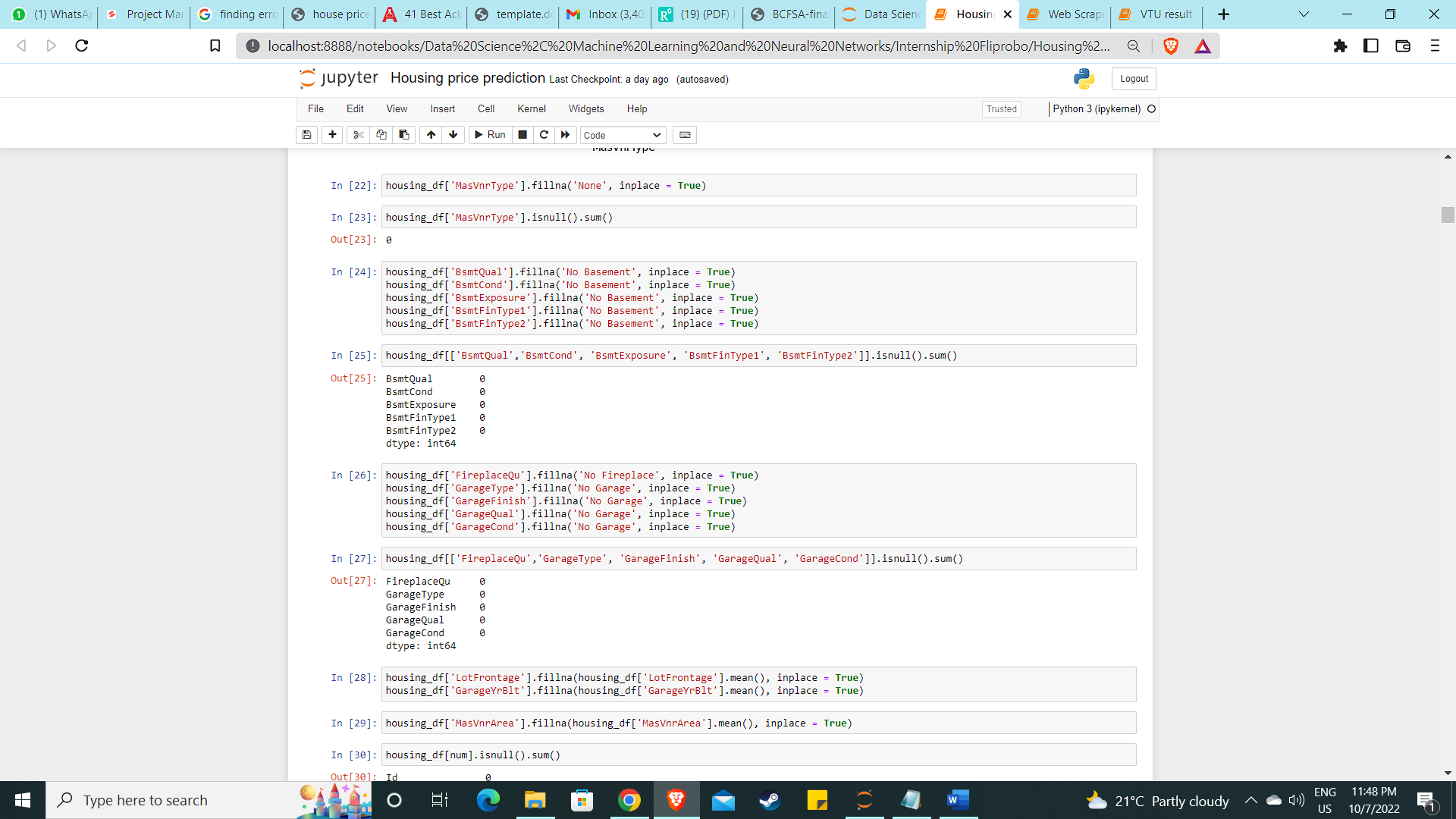
* **Data Sources and their formats**

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below.

The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:

What are the data sources, their origins, their formats and other details that you find necessary? They can be described here. Provide a proper data description. You can also add a snapshot of the data.

* **Data Pre-processing Done**



* **Hardware and Software Requirements and Tools Used**

Hardware requirements:

* Laptop or PC to analyse the data
* Software requirements: 4GB/8GB RAM (8GB or high preferred), minimum 1GB graphics card (any), 512GB HDD.
* Tools Used: Jupyter Notebook – Entire model
* Libraries used: NumPy (for numerical computations like imputations), Pandas (used reading the data, data cleaning, data manipulation, correlation, summary statistics, skewness, feature engineering) Seaborn (for plotting), Matplotlib (for plotting), scikit-learn, SciPy (for z score), stats models (for VIF to check multicollinearity), joblib (for pickling i.e. deploying the best model trained)

**Model Development and Evaluation**

* **Problem solving methods**

Approach followed during the course of predicting the house sales price

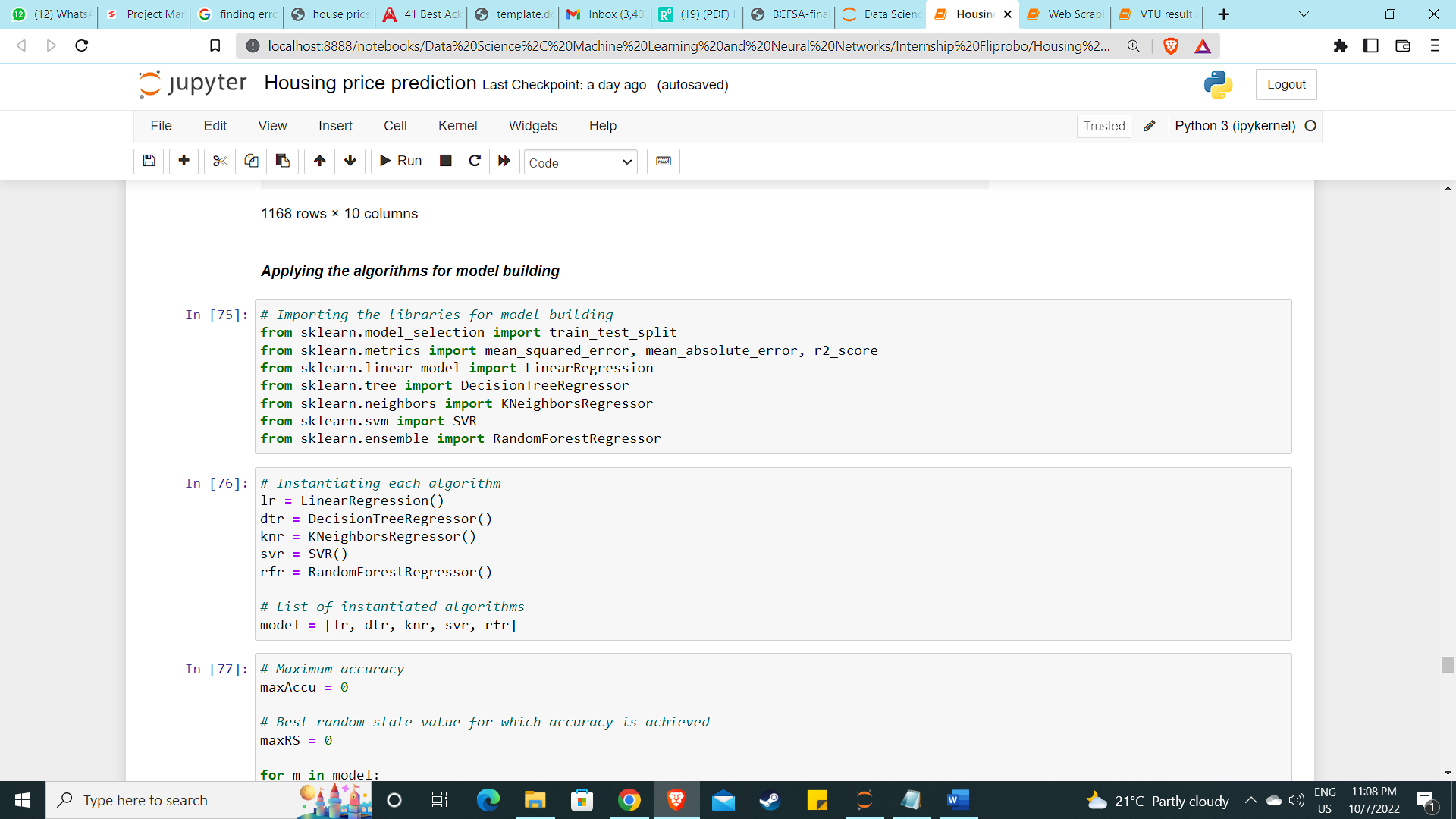
* Data given was imported as DataFrame. Next, the basic information like the datatypes, checking the count of non-null values in the data with info() function, then summary statistics of both numerical and categorical data which provides us the information about the count, frequency, unique, top values for the categorical data and mean, standard deviation, minimum, maximum, 25th percentile, 50th percentile (median), 75th percentile of the numerical data.
* Checking null values, imputing the null values, dropping the columns with null values almost close to the size to data, plots like countplot, barplot, regression plot, categorical plot were plotted to check the effectiveness of the input variable with the target variable and its behaviour.
* Next, correlation matrix was plotted through which we could identify the how the input variable is correlated with the target variable and also, we see the multicollinearity which is even checked by seeing the variance inflation factor.
* Next skewness was checked and removed with log and square transforming methods. Next outliers were identified.
* Later the data was split into features (X) and target (y).
* MinMaxScaler method was used for transforming and to reduce the dimensionality of the data PCA (principal component analysis) was used.
* Model Building – 5 models used for training along with cross validation and the best model was tuned with hyperparameters. And the two models were deployed one with tuned one and other without and comparison was done with predicted and original with plots.
* Lastly, the price was predicted for the test data provided which was cleaned before predicting the sales price of the house.
* **Algorithms used**

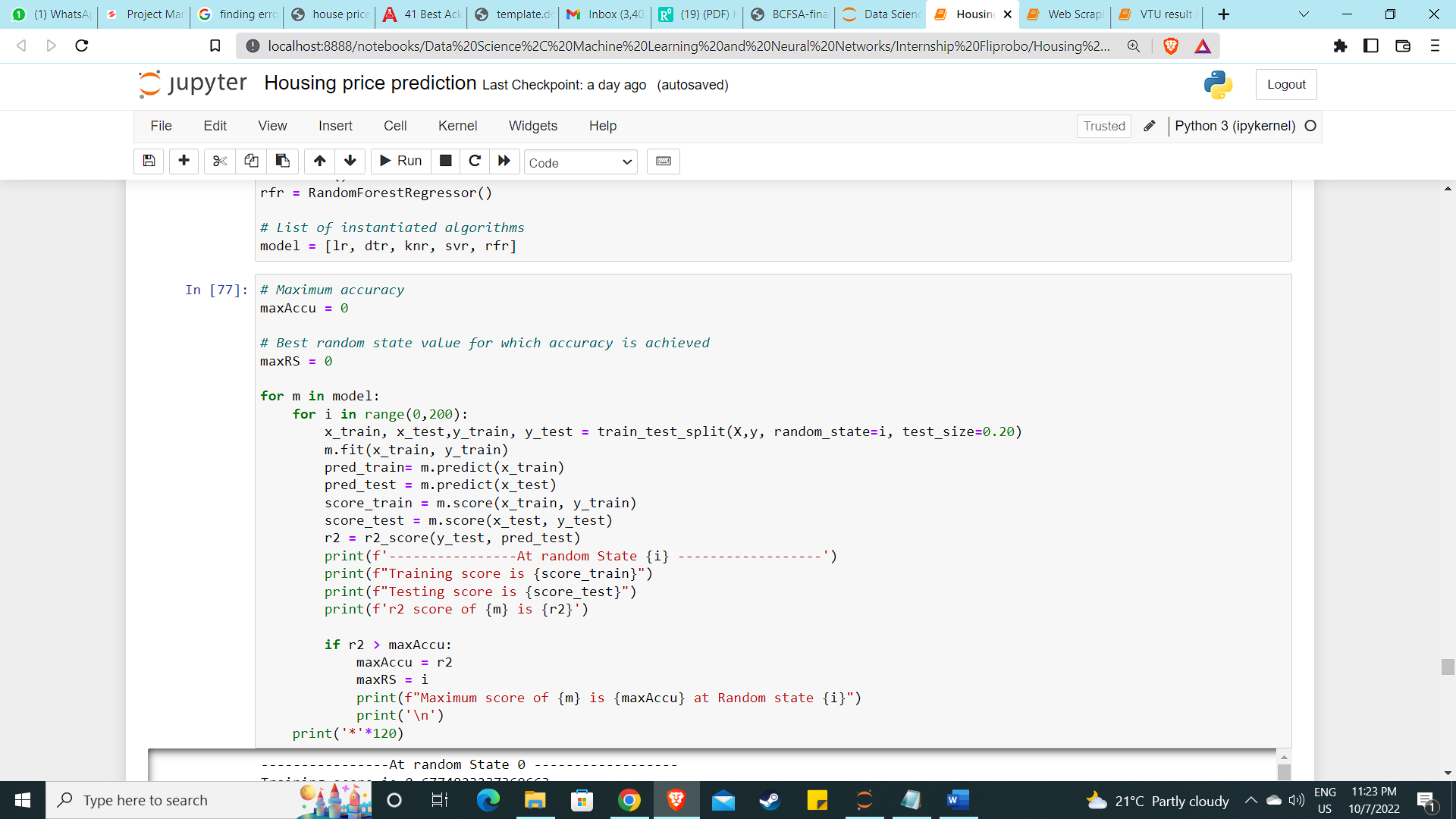
I have used total of 5 models for training purpose as listed below

* Linear Regression
* Decision Tree Regressor
* K Neighbour Regressor
* Support Vector Regressor
* Random Forest Regressor

All the models were run to get the best accuracy for best random state in a range of 0 to 200. Later cross validation was performed for all the models for cross validation range 2 to 15. The Best model was Random forest regressor which was tuned using gridsearchCV.

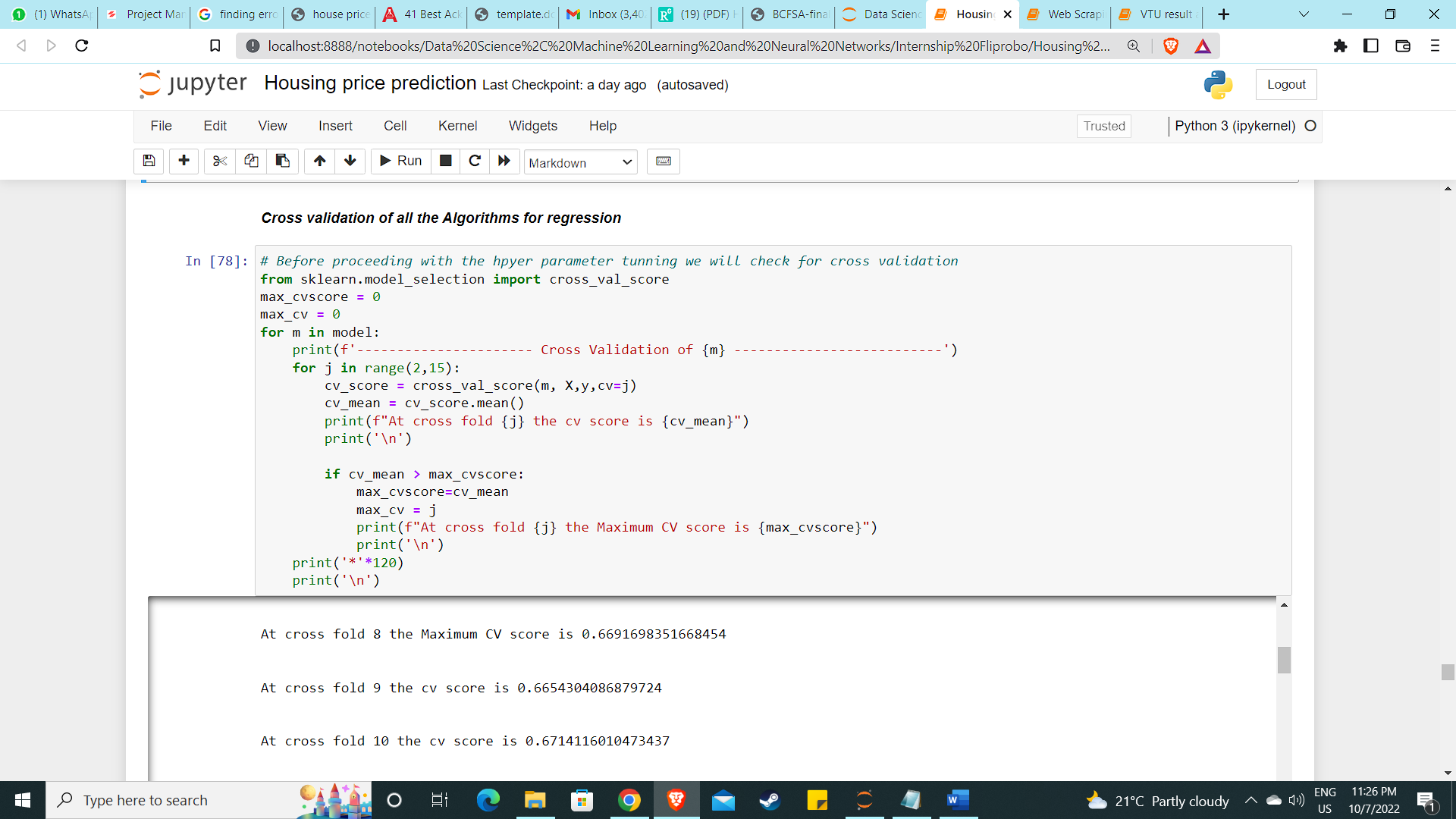
* **Run and evaluate selected models**



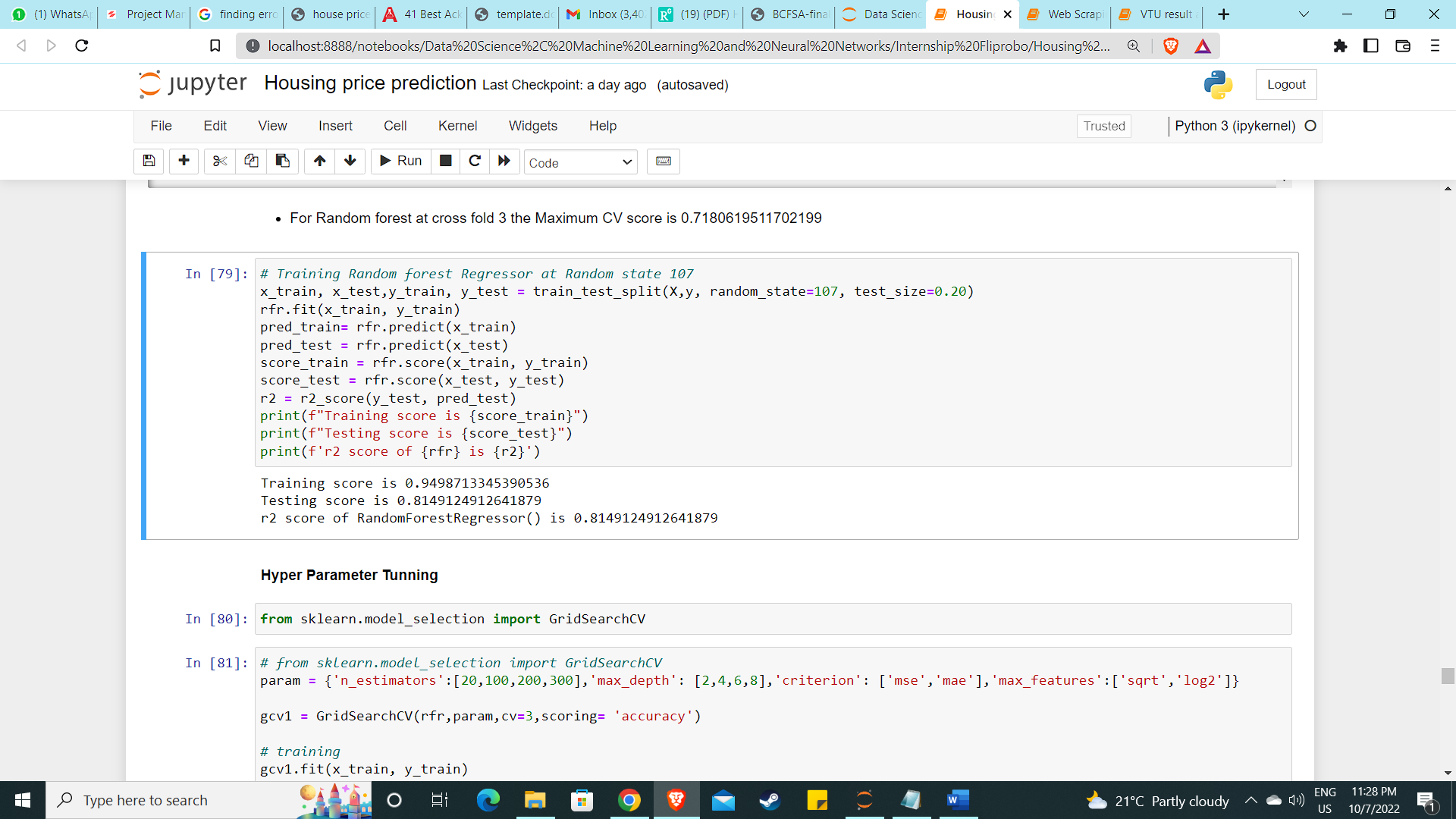


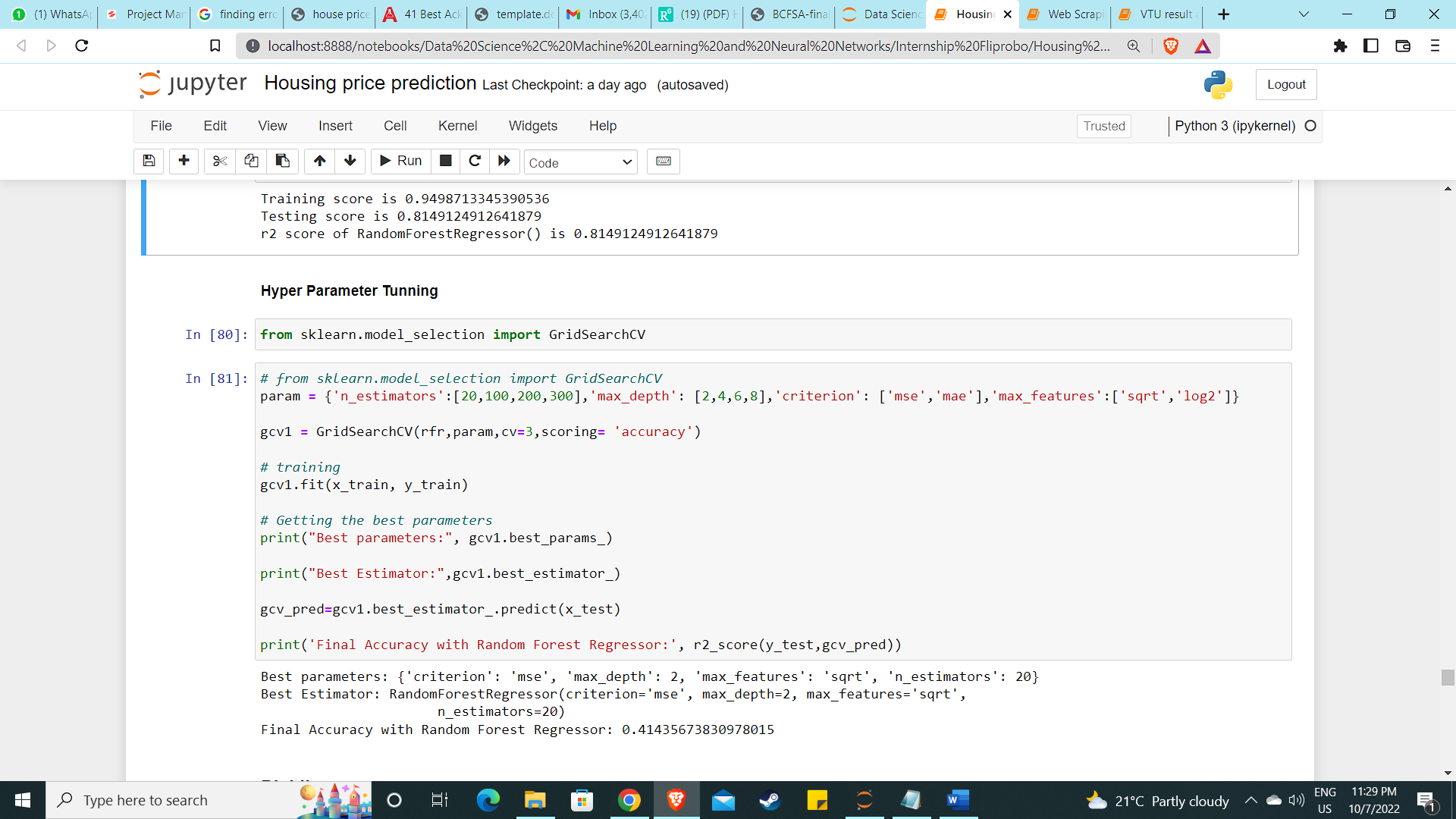
Random Forest Regressor

* Maximum score of RandomForestRegressor() is 0.8134658396095735 at Random state 107
* Training score is 0.9533843304262911
* Testing score is 0.8134658396095735
* r2 score of RandomForestRegressor() is 0.8134658396095735

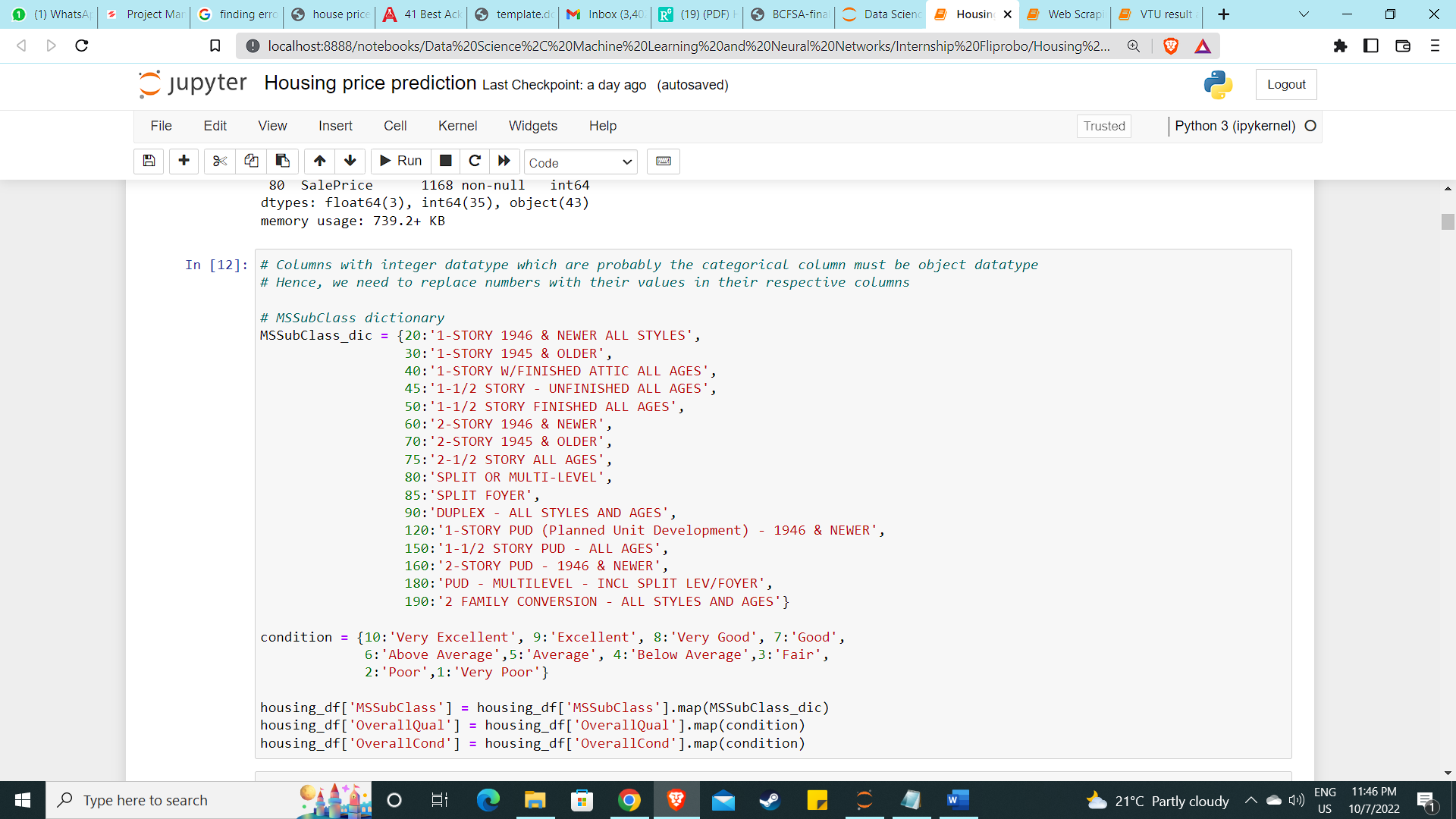


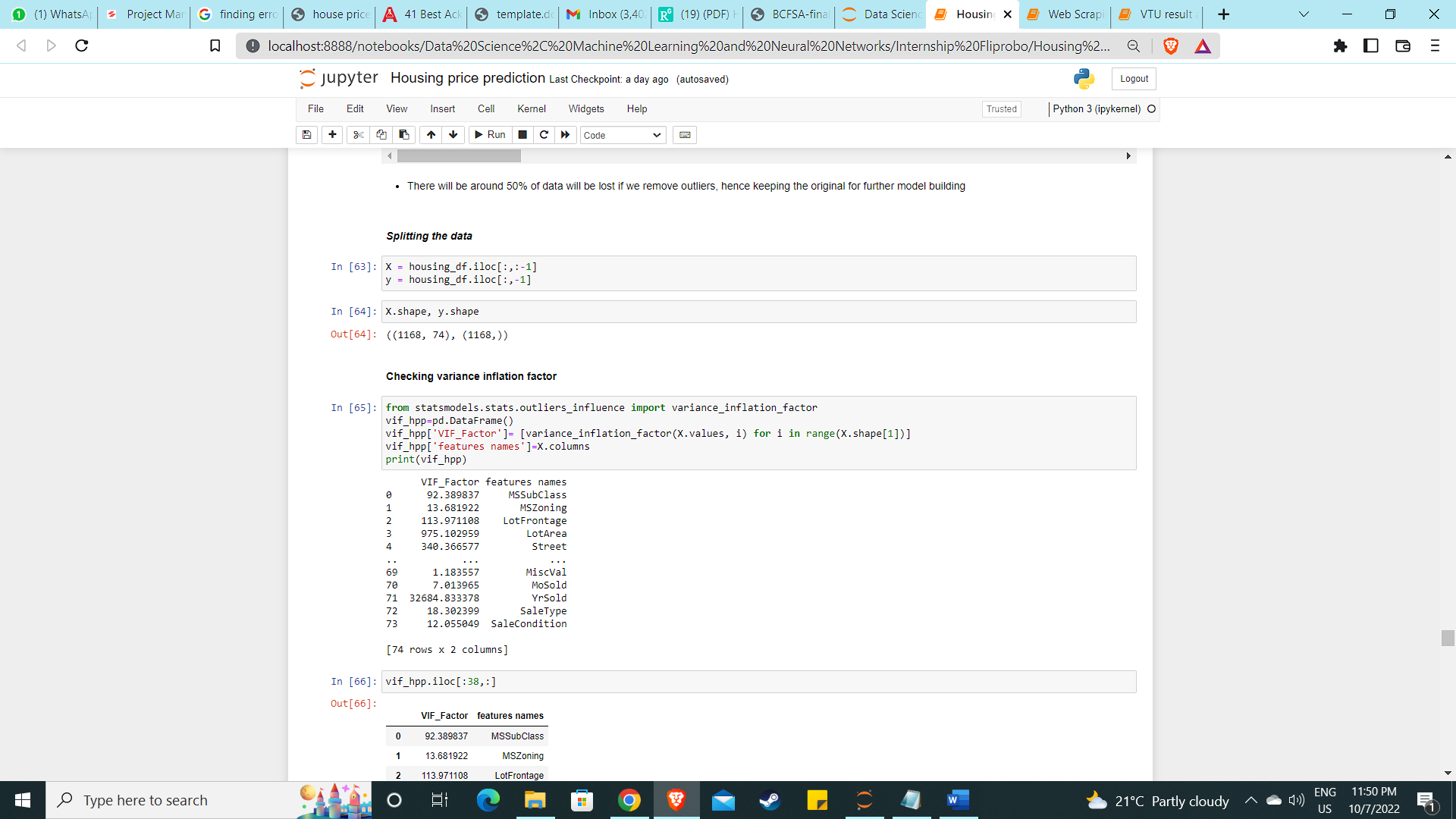
For Randomforest at cross fold 3 the Maximum CV score is 0.7180619511702199

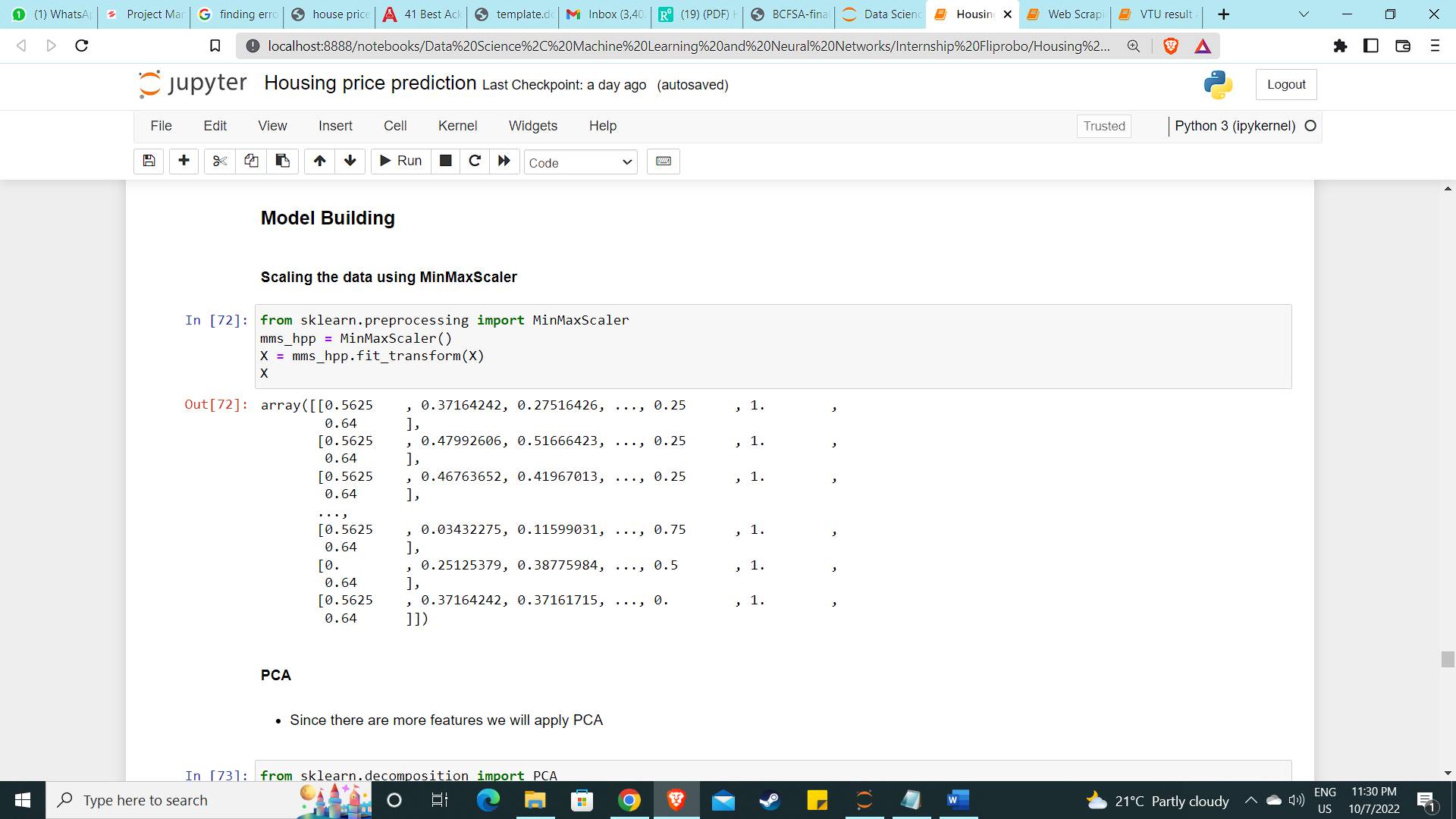




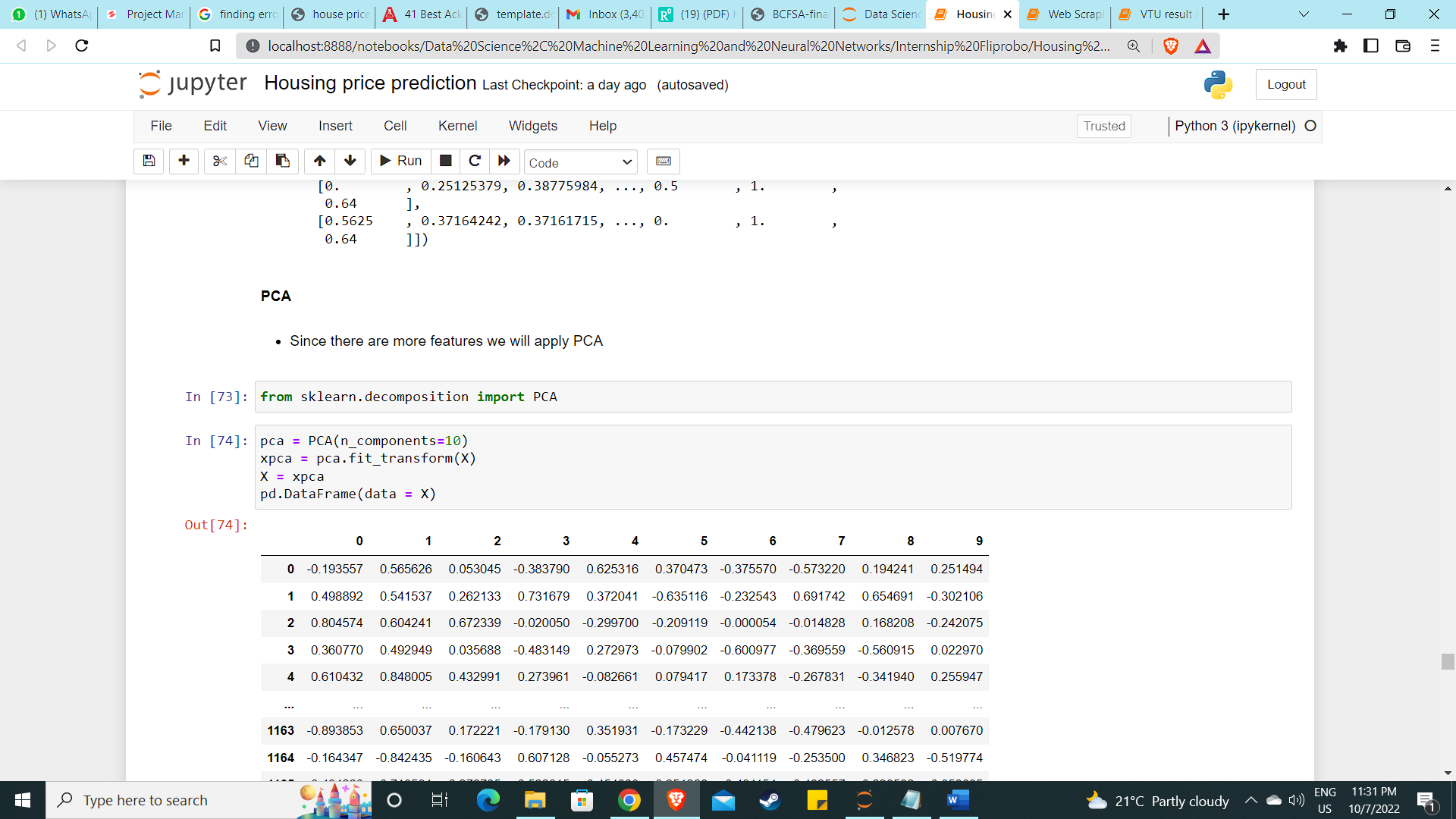
* **Key Metrics for success in solving problem under consideration**





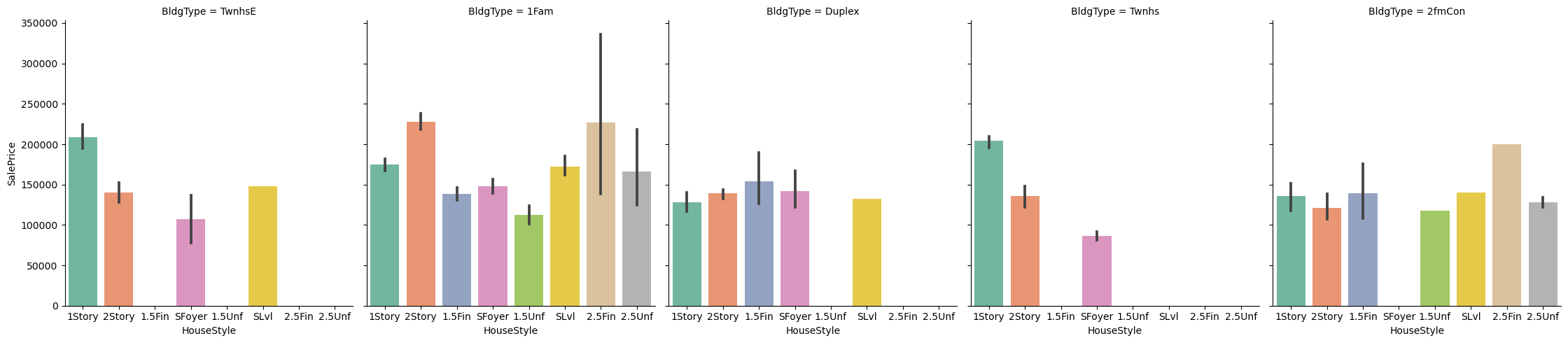
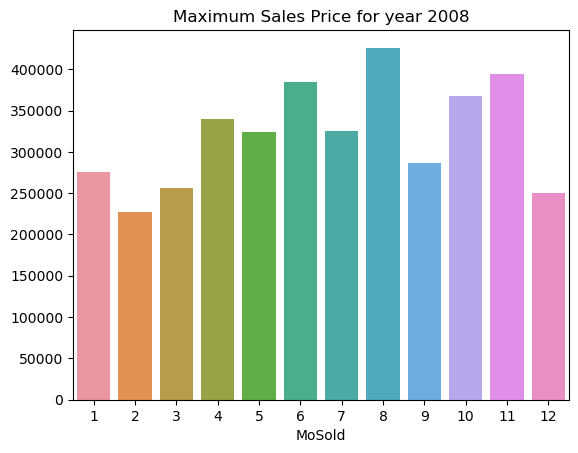
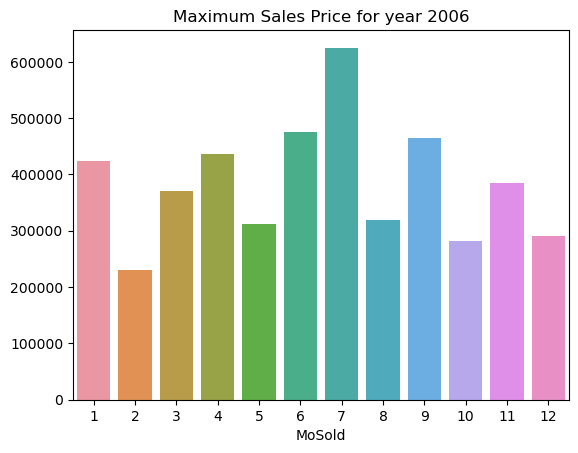
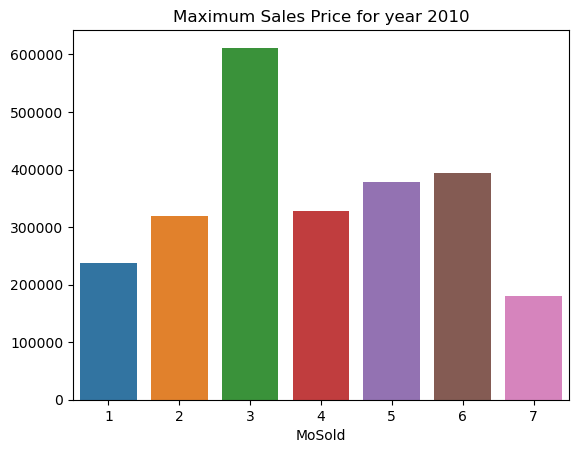
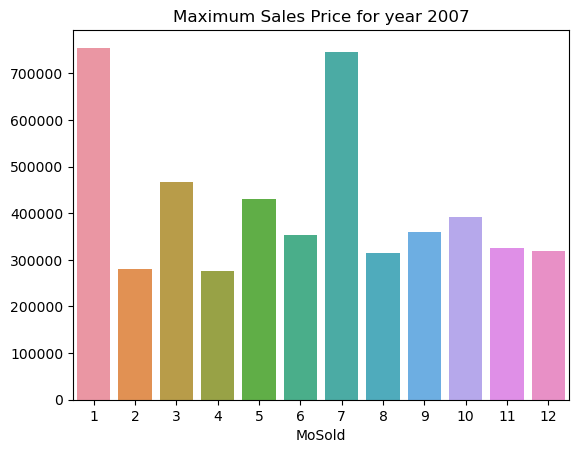
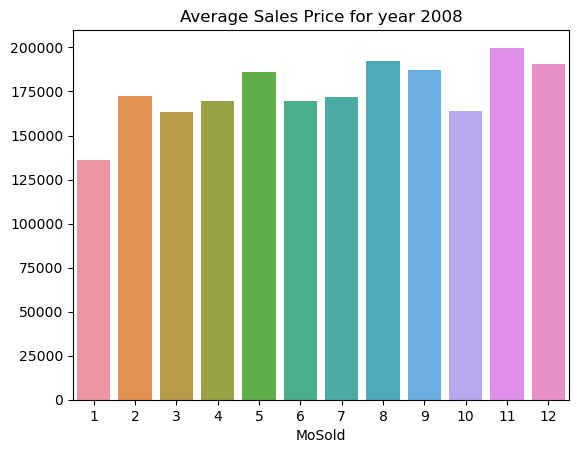
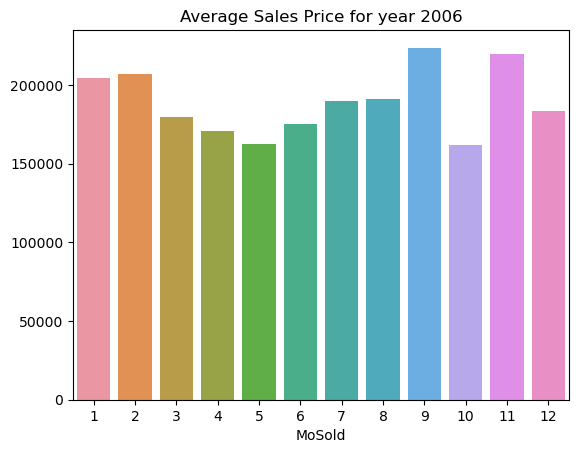
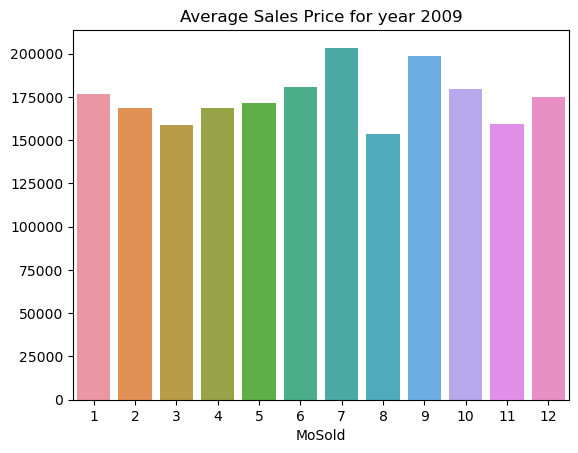
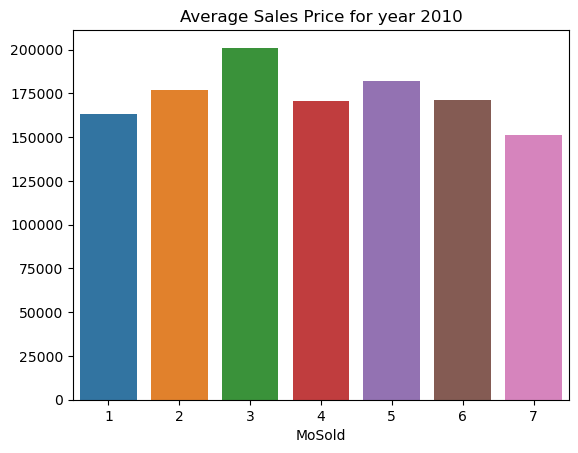
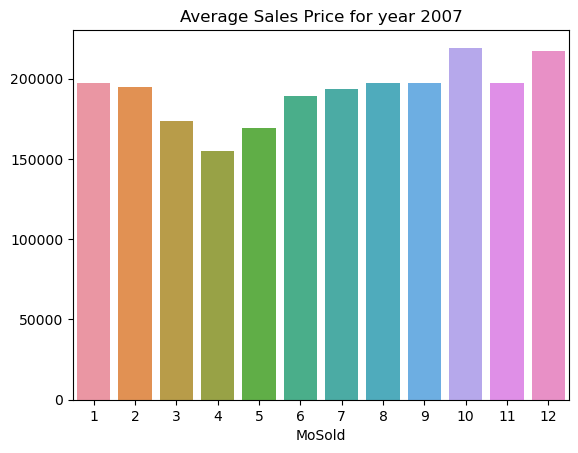
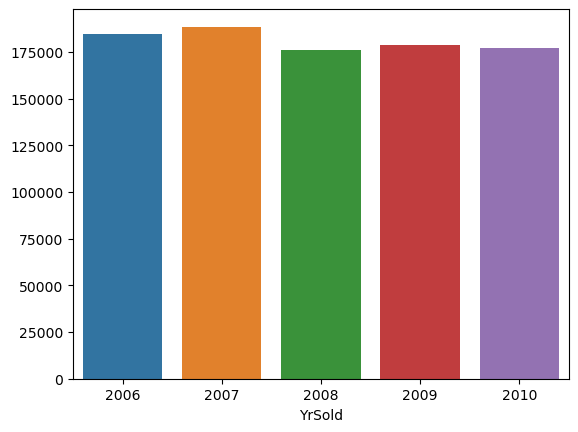
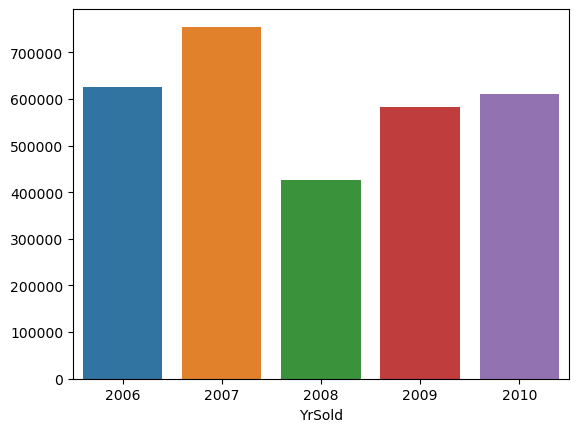
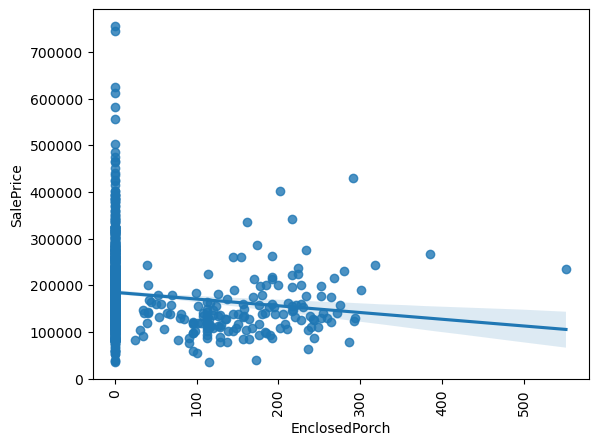
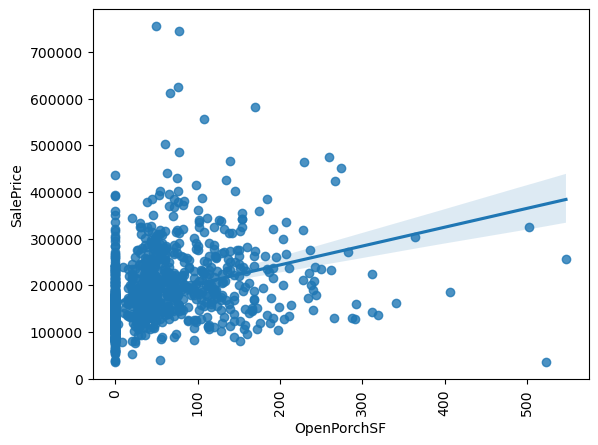
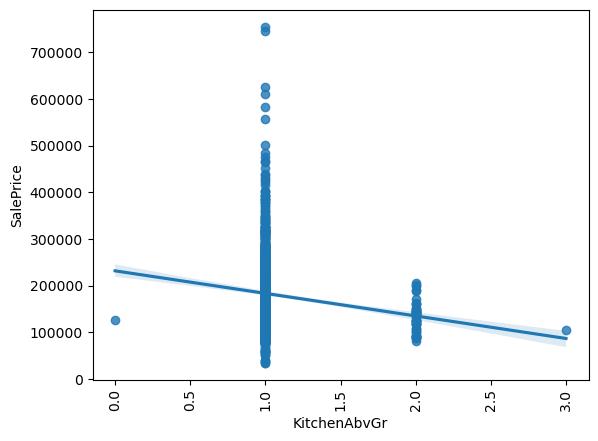
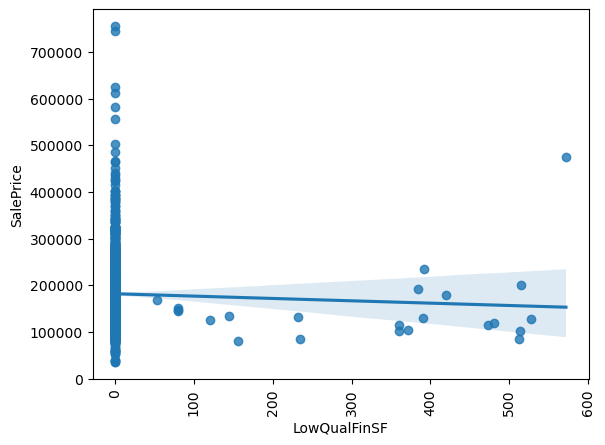
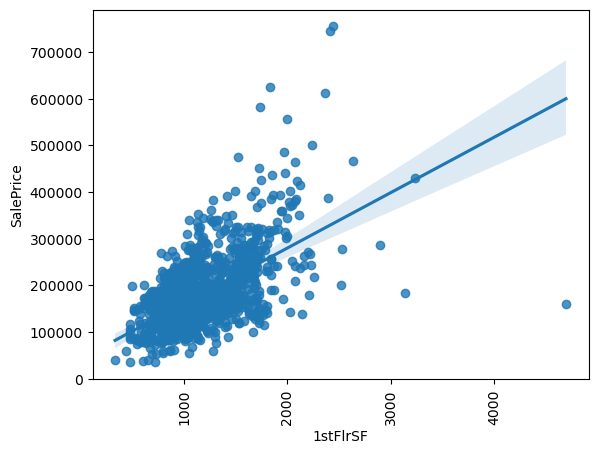
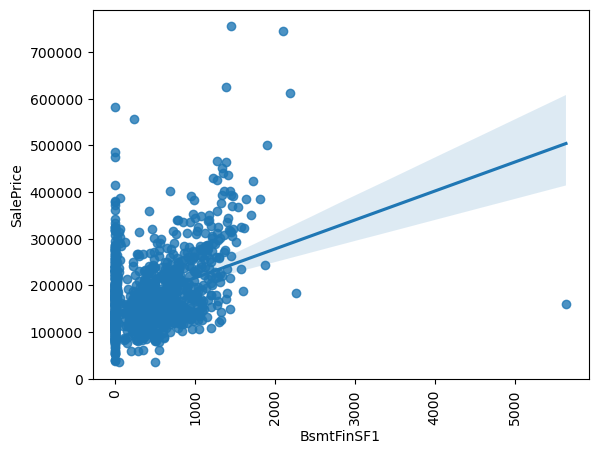
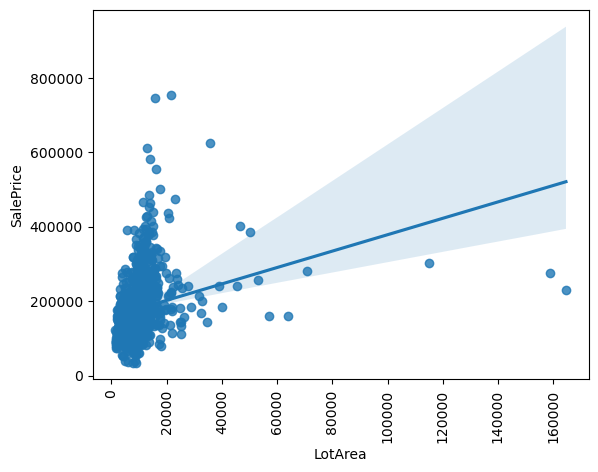
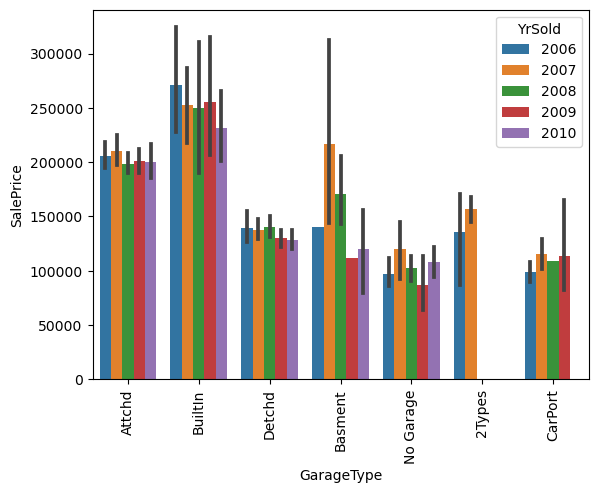
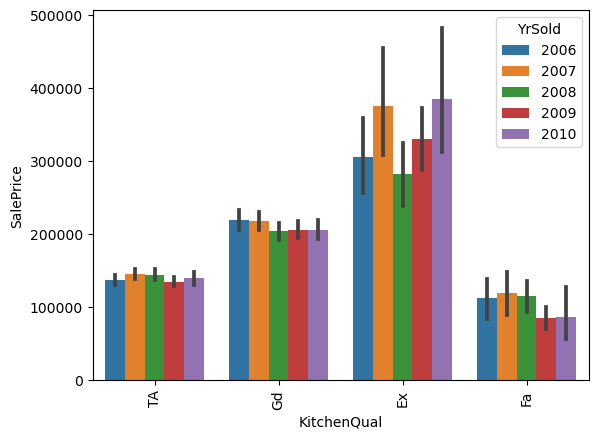
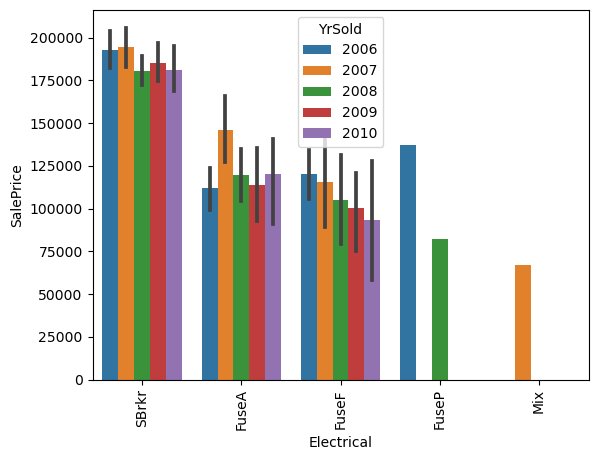
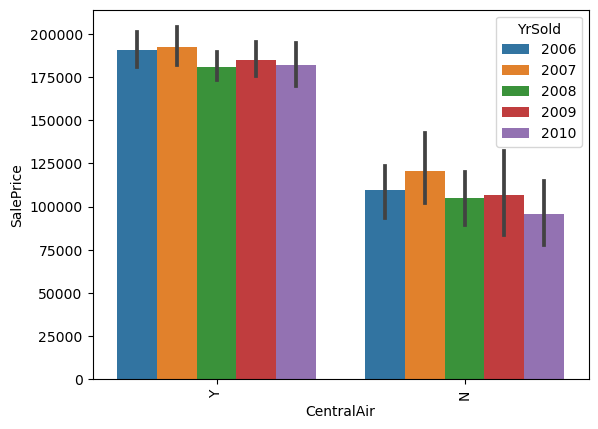
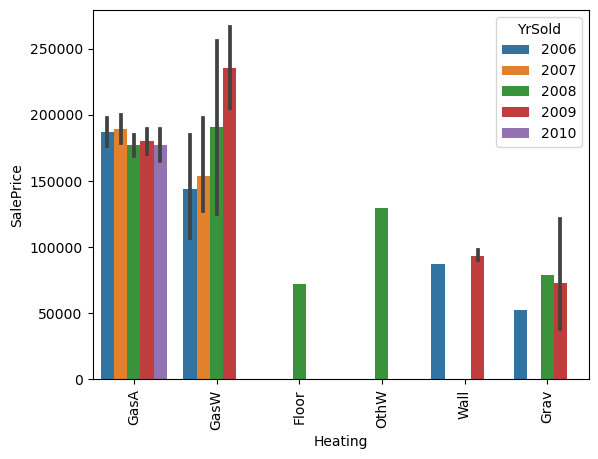
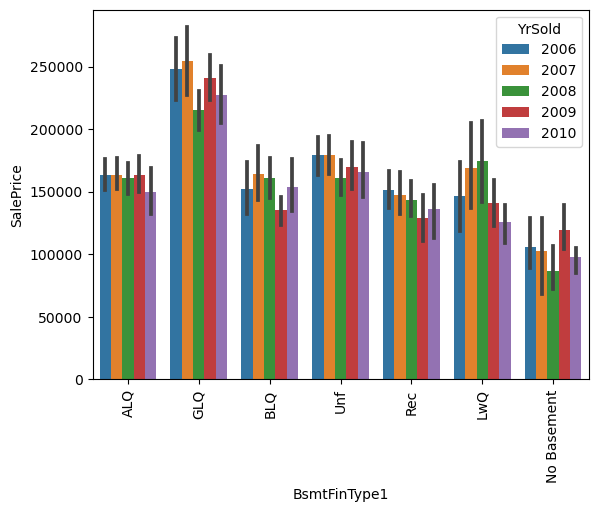
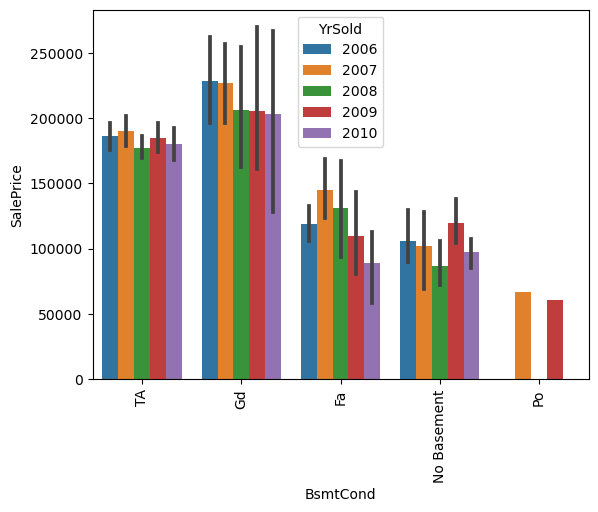
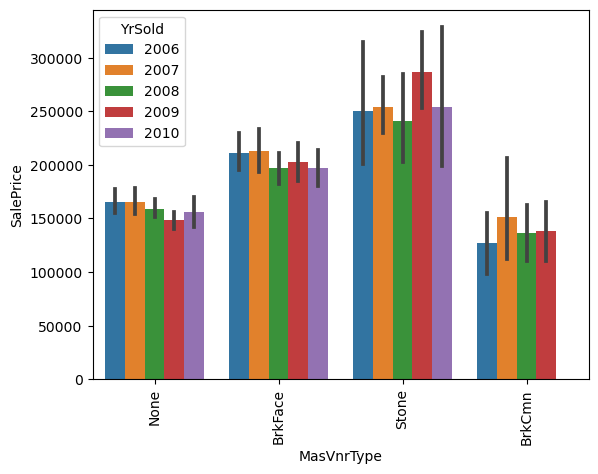
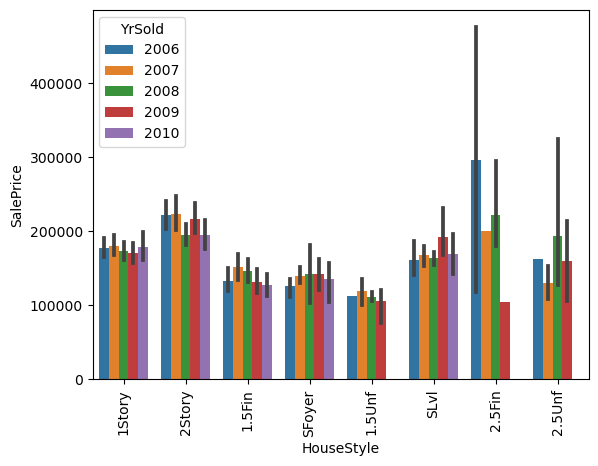
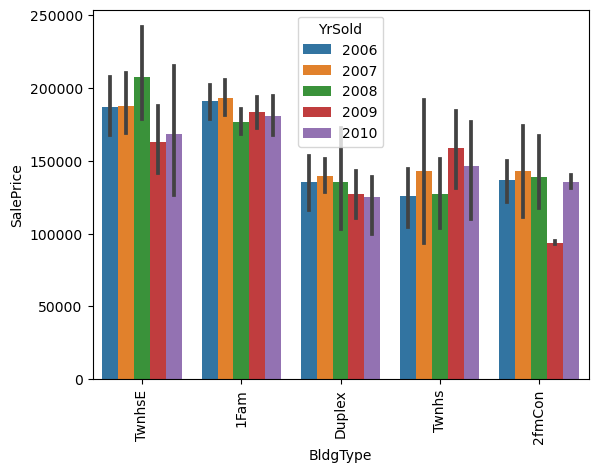
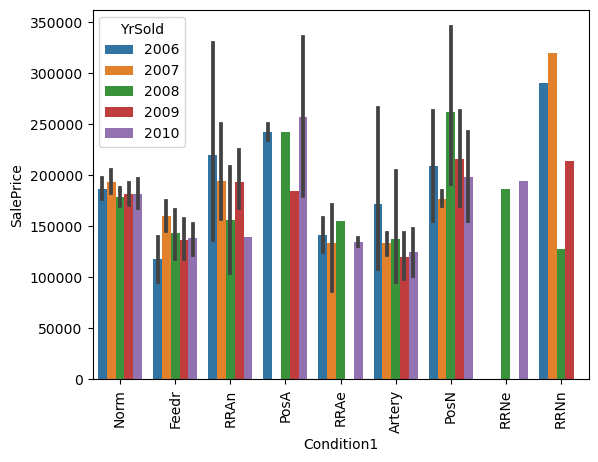
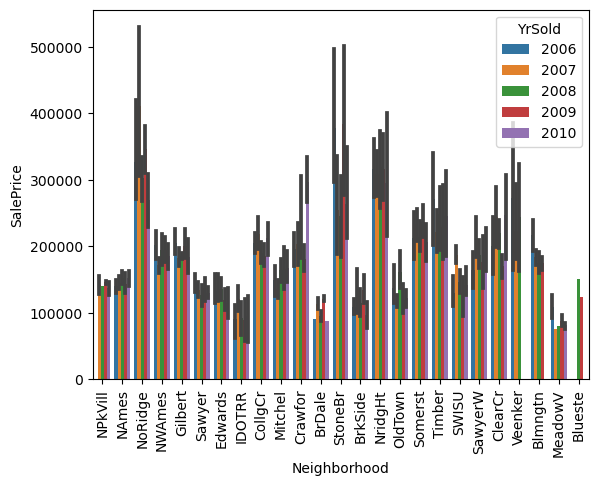
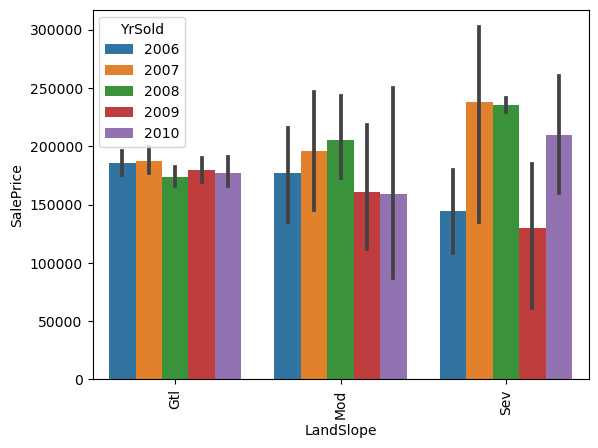
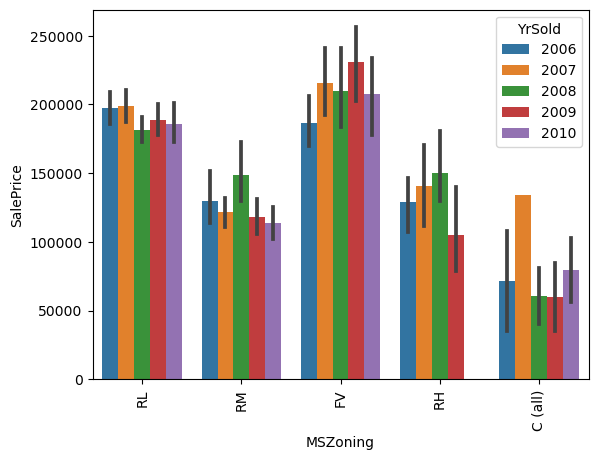
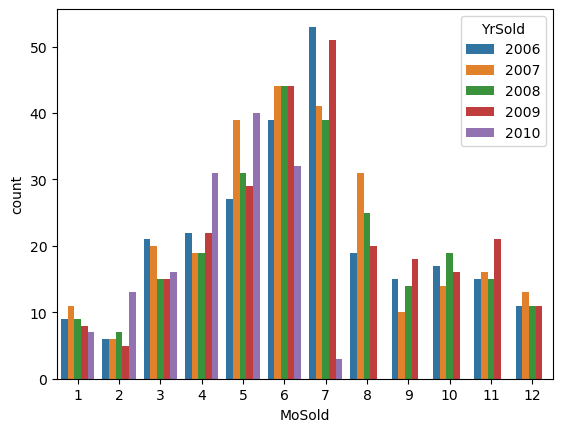
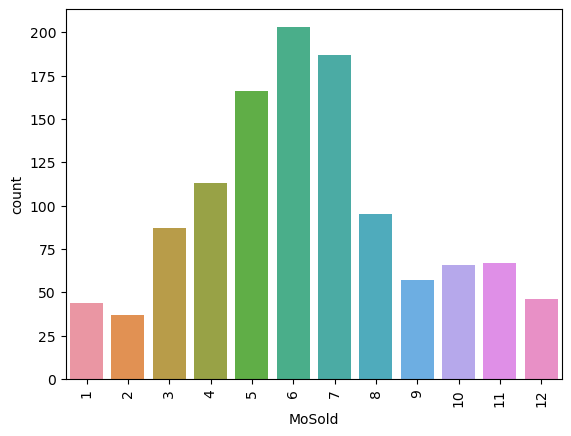
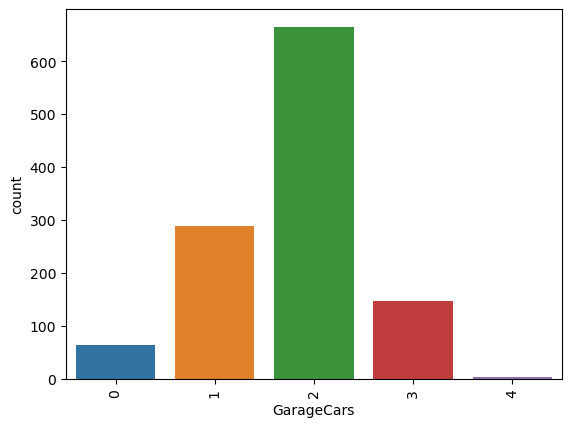
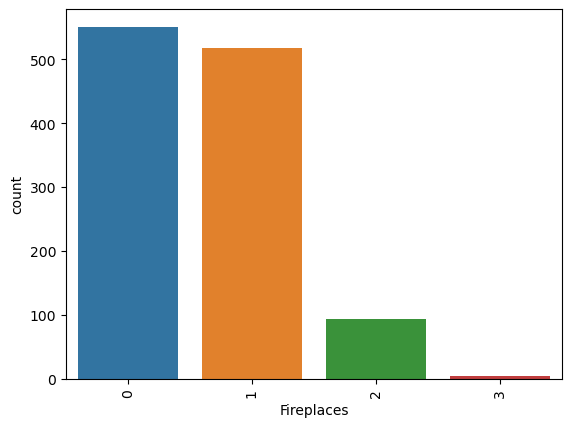
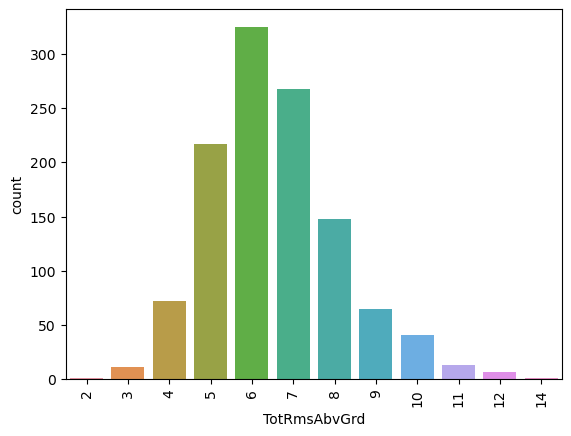
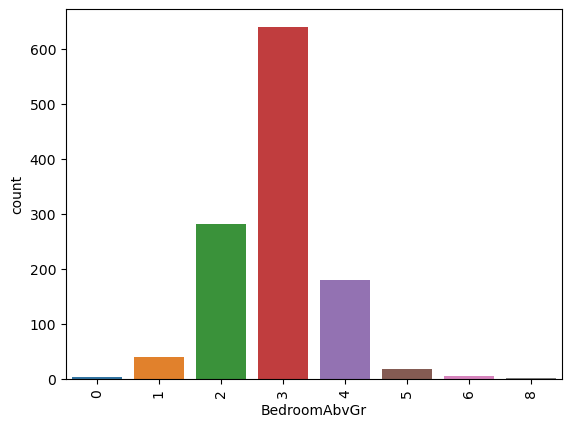
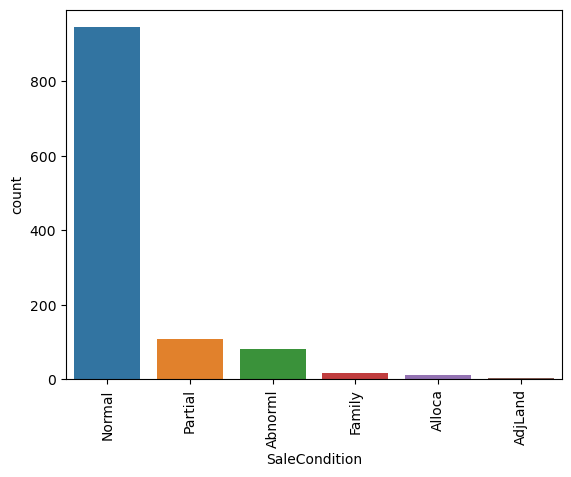
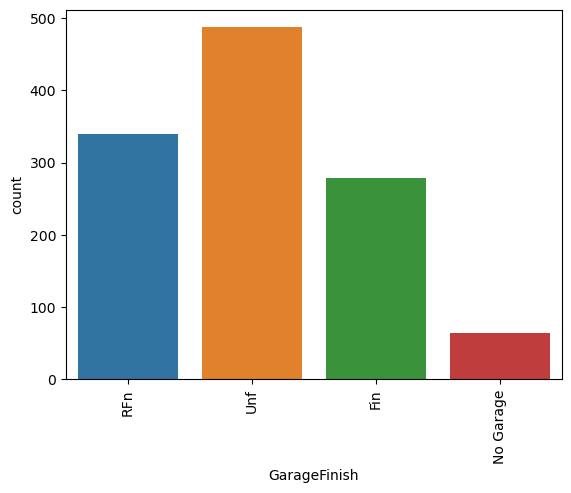
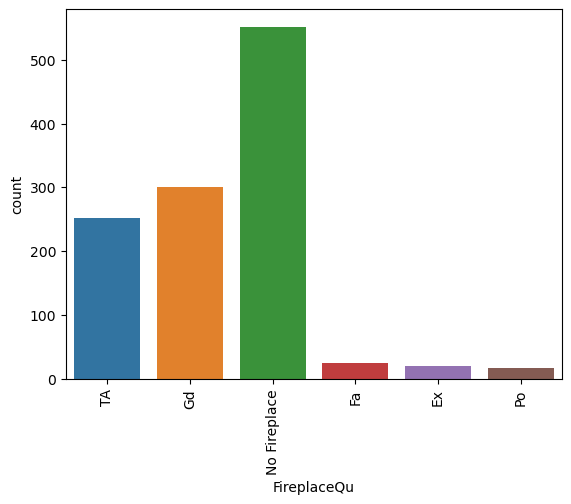
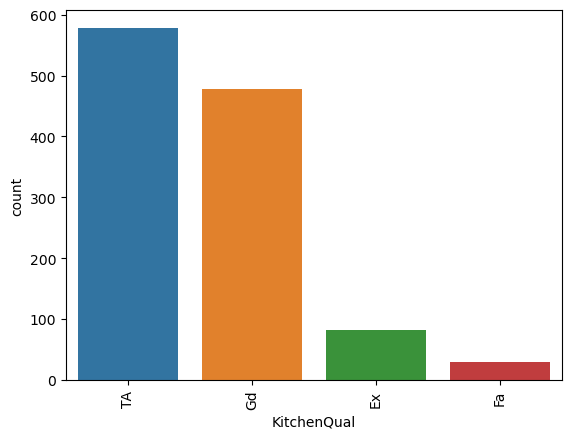
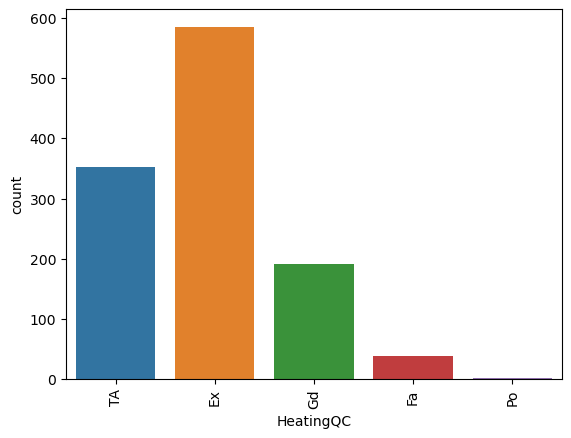
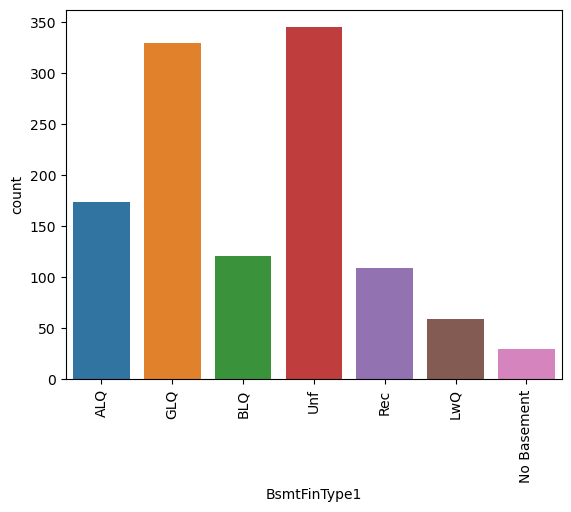
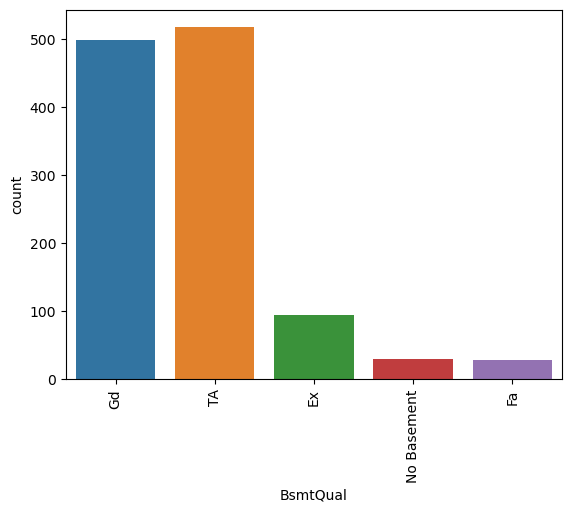
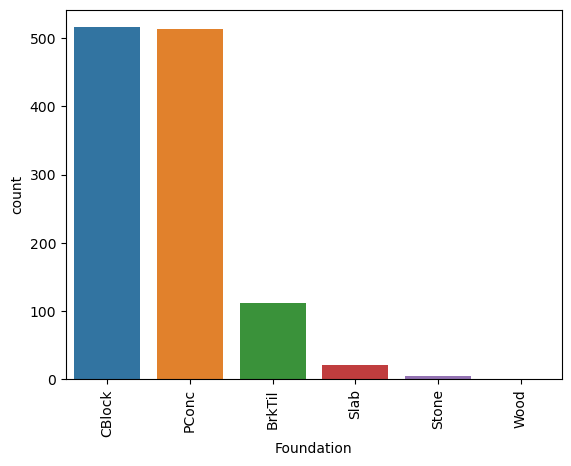
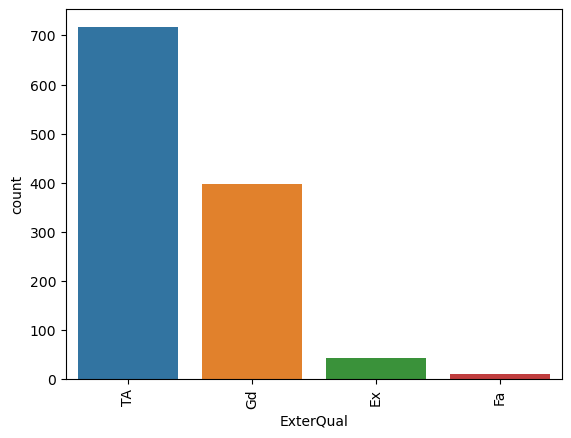
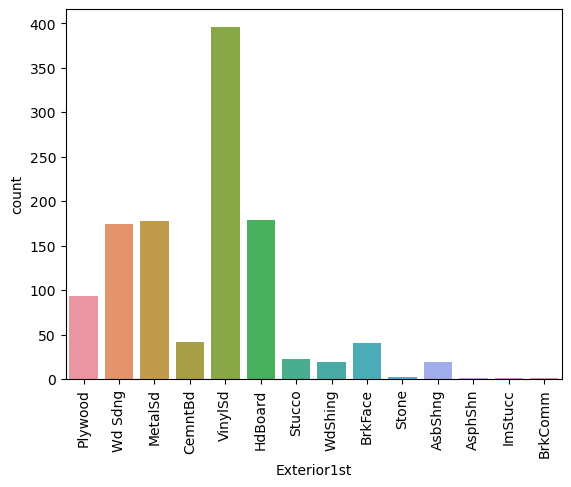
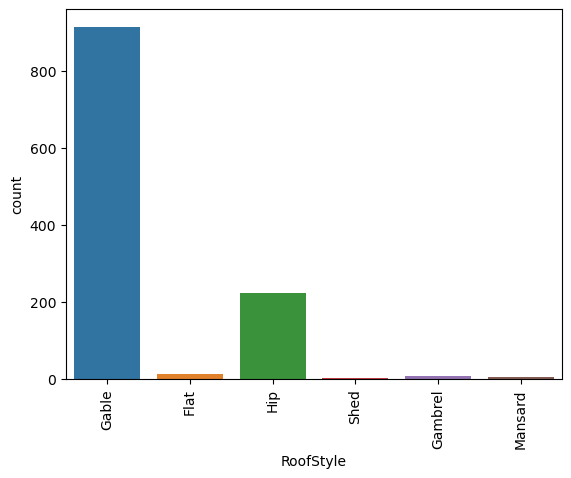
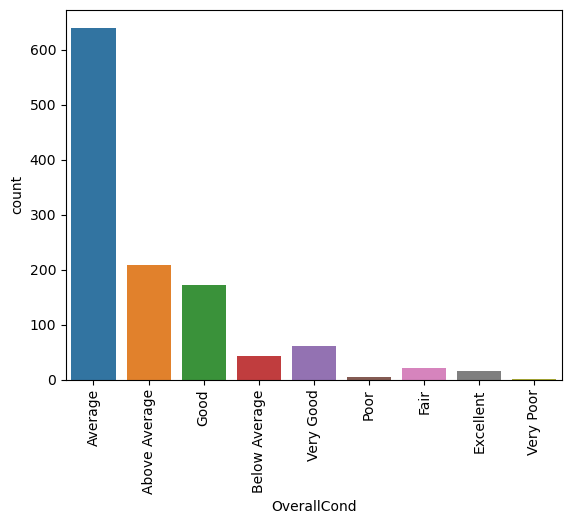
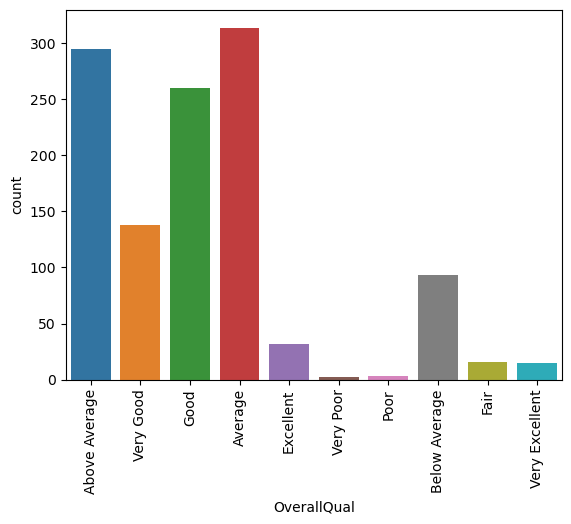
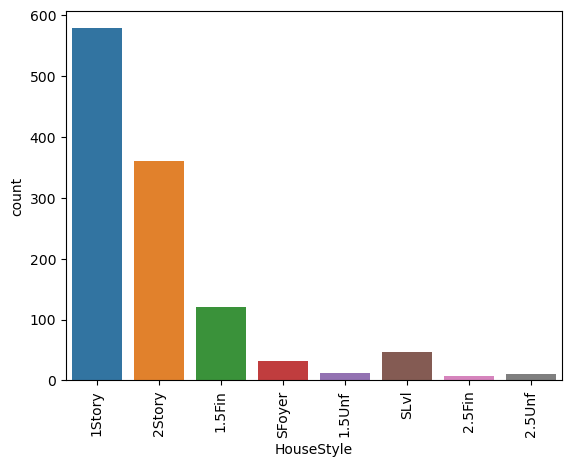
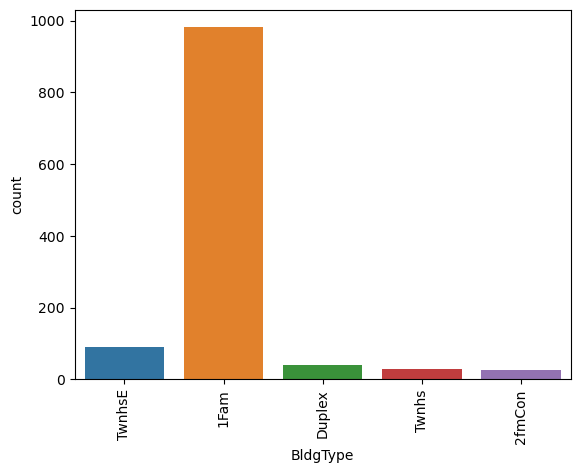
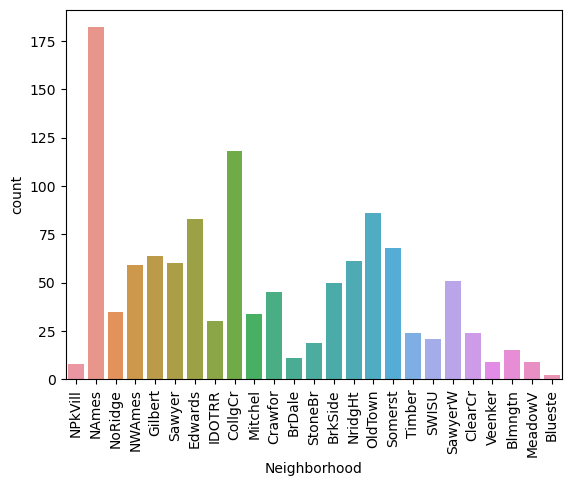
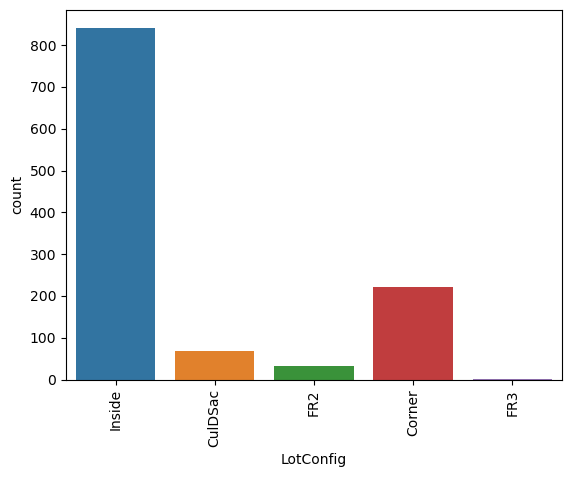
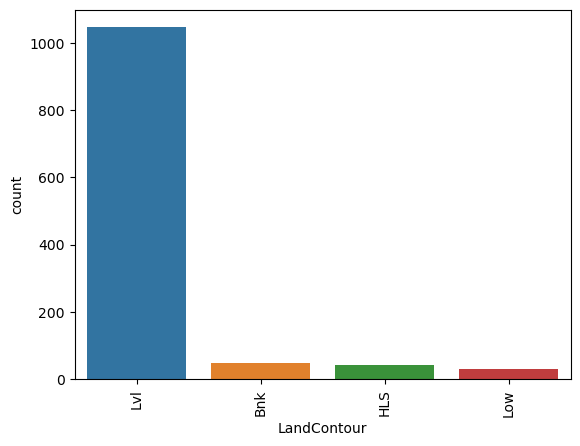
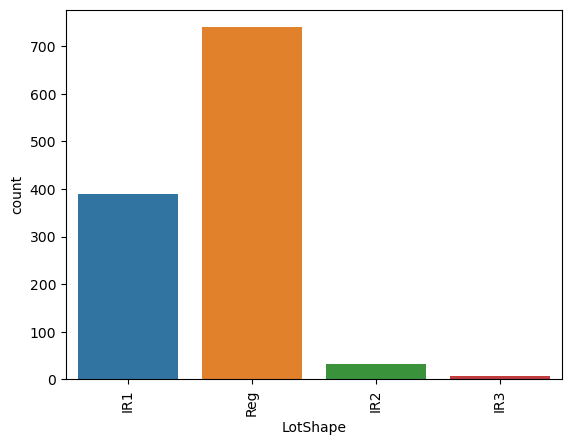
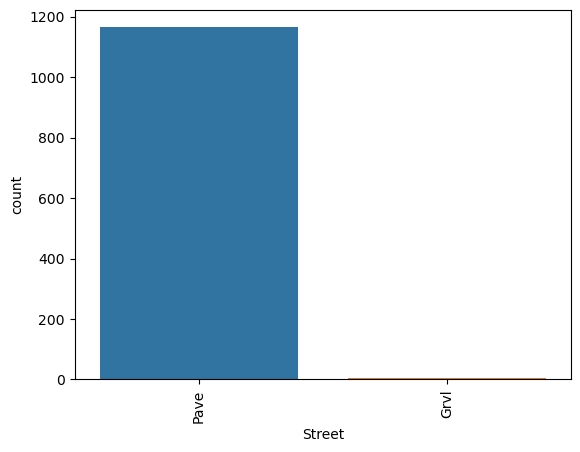
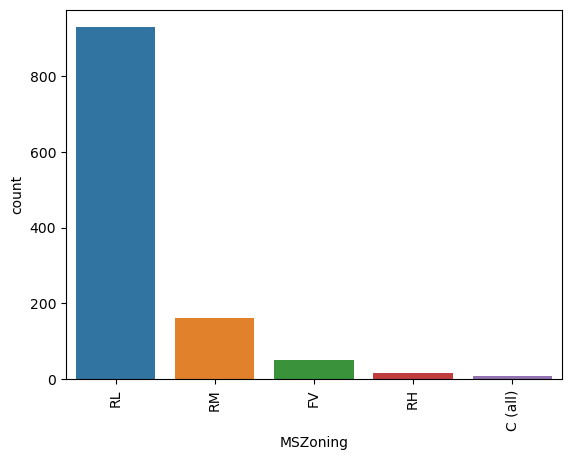
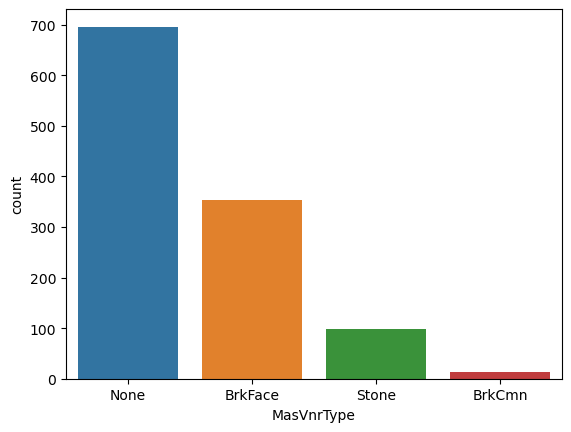
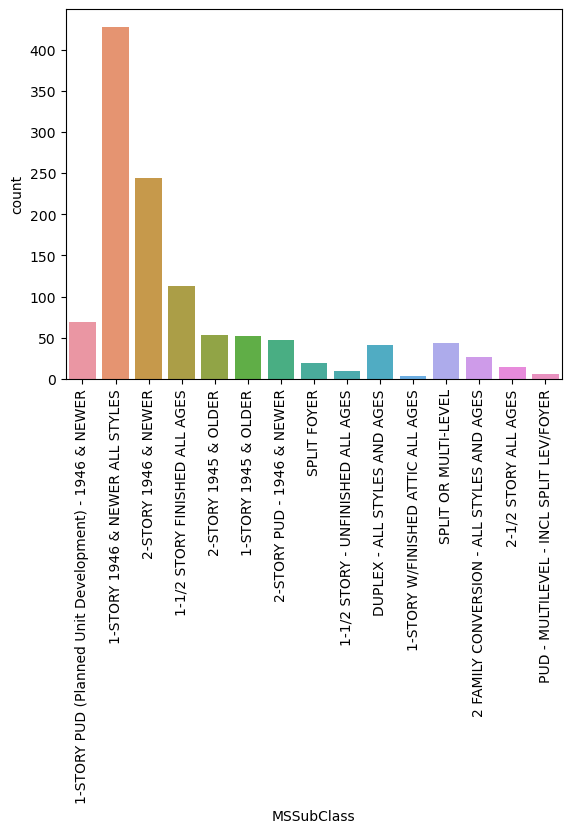


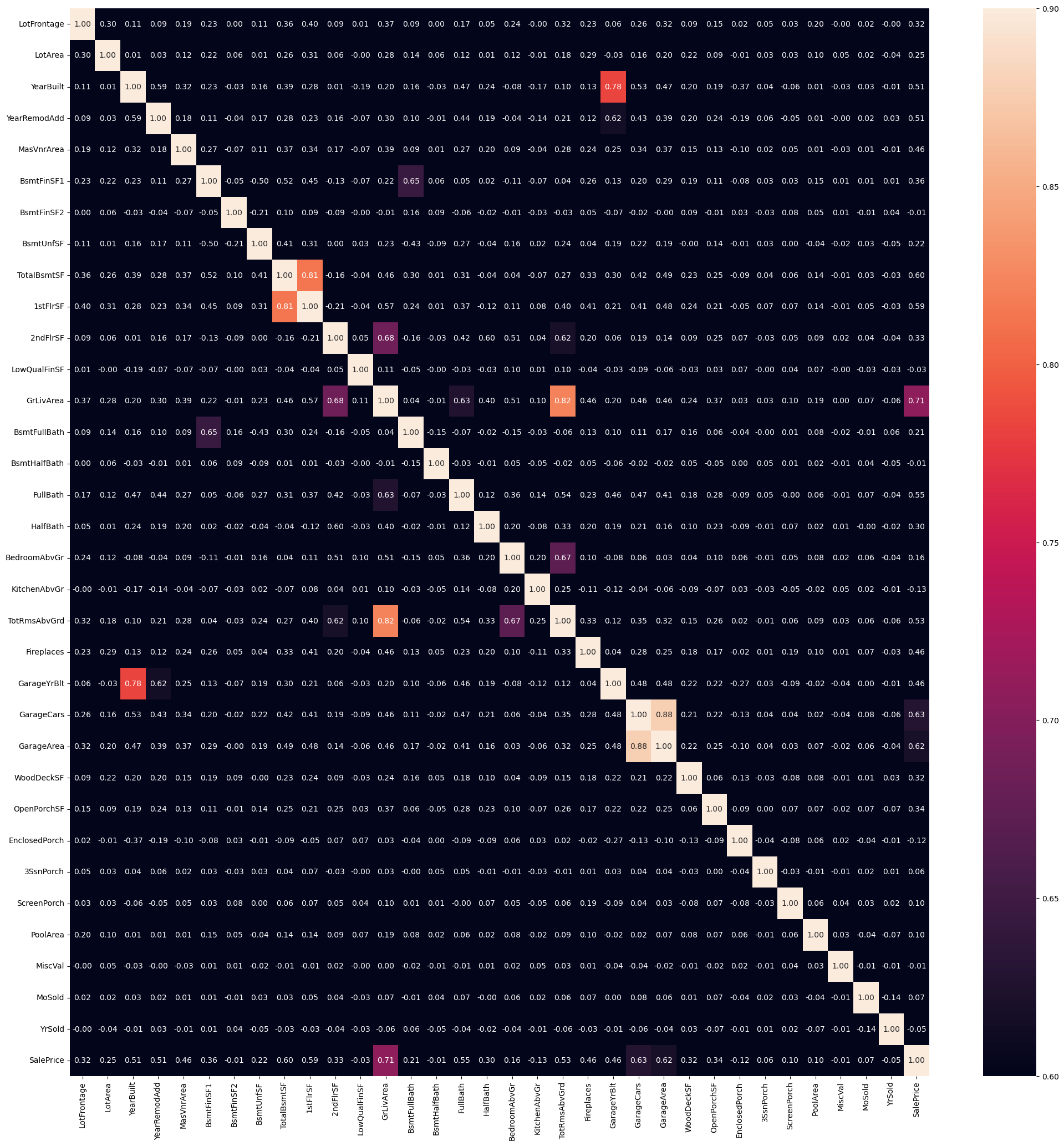
Scaling is used to transform the data to normal distribution.



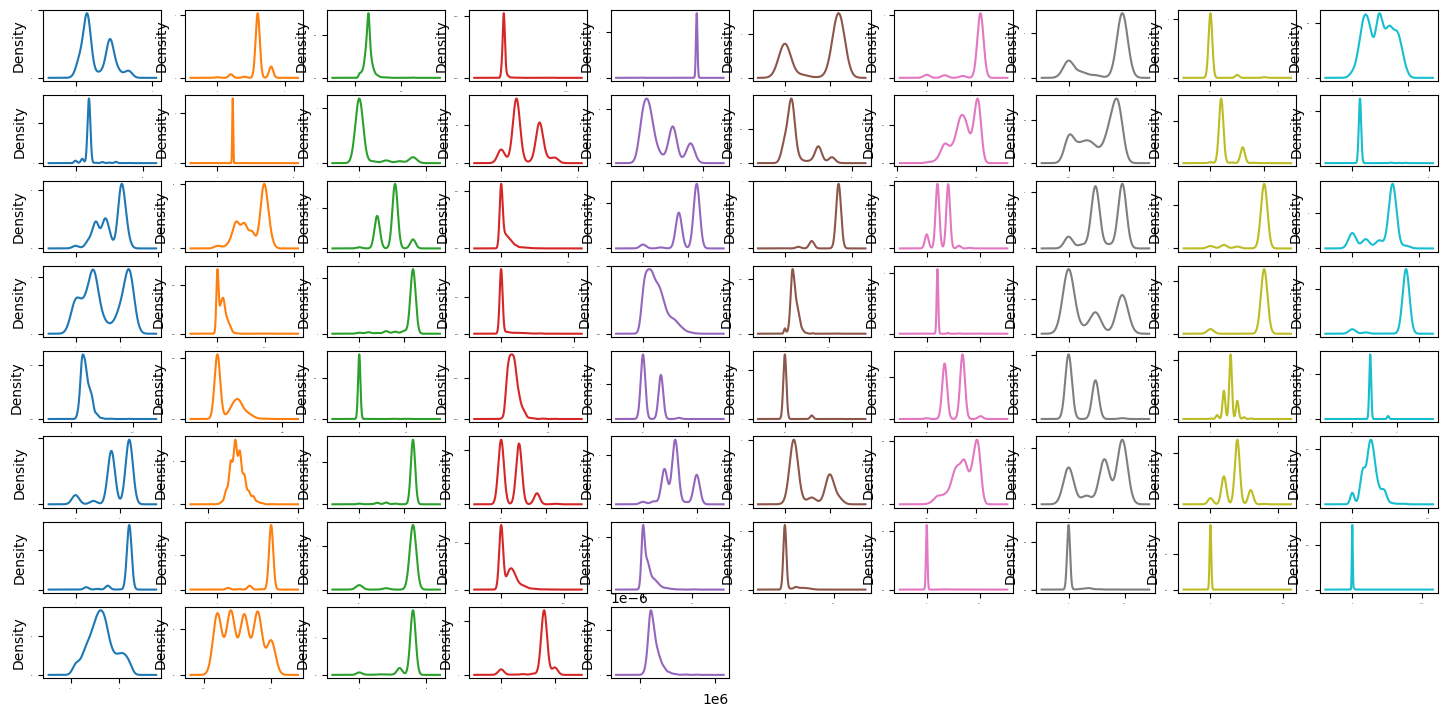
PCA (principal component analysis) is used since the number of features are to many, in order to reduce the dimensionality of the problem.

* **Visualizations**





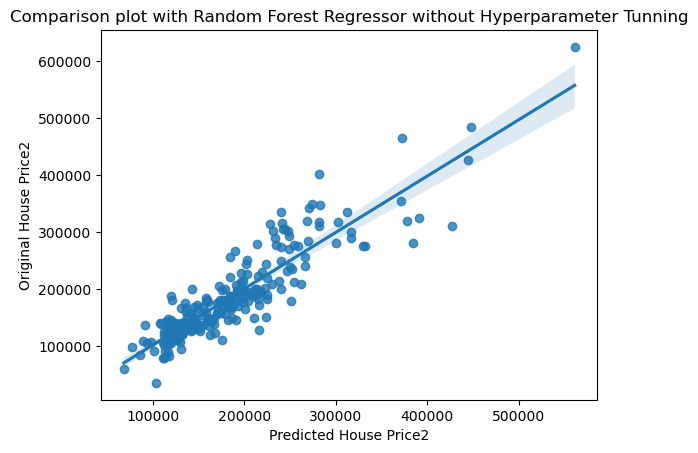
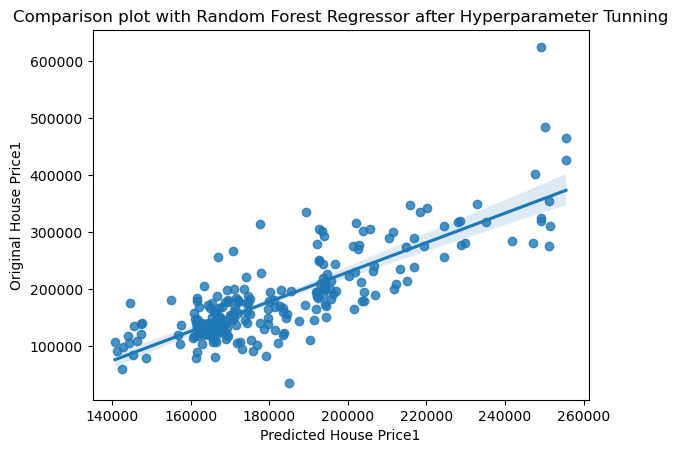
With the help of correlation matrix, we can identify the multicollinearity present in the data i.e., the features which are correlated to each other and then with variance inflation factor(vif) and correlation matrix we can decide on dropping the features which is less correlated between the two features.

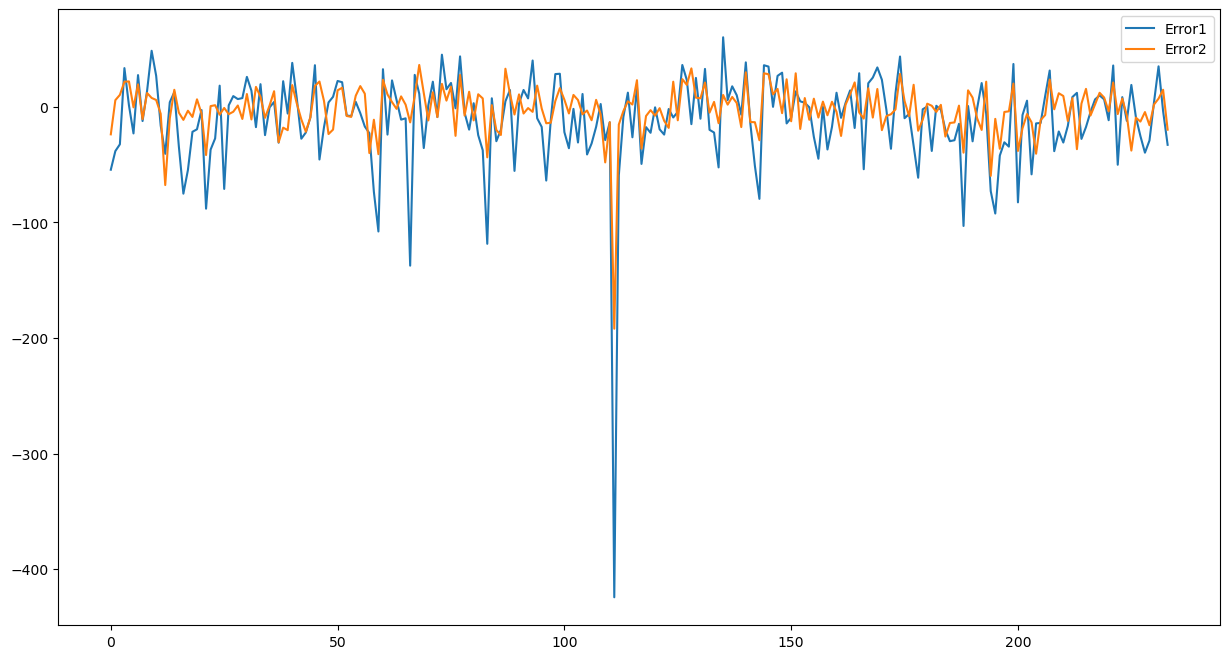
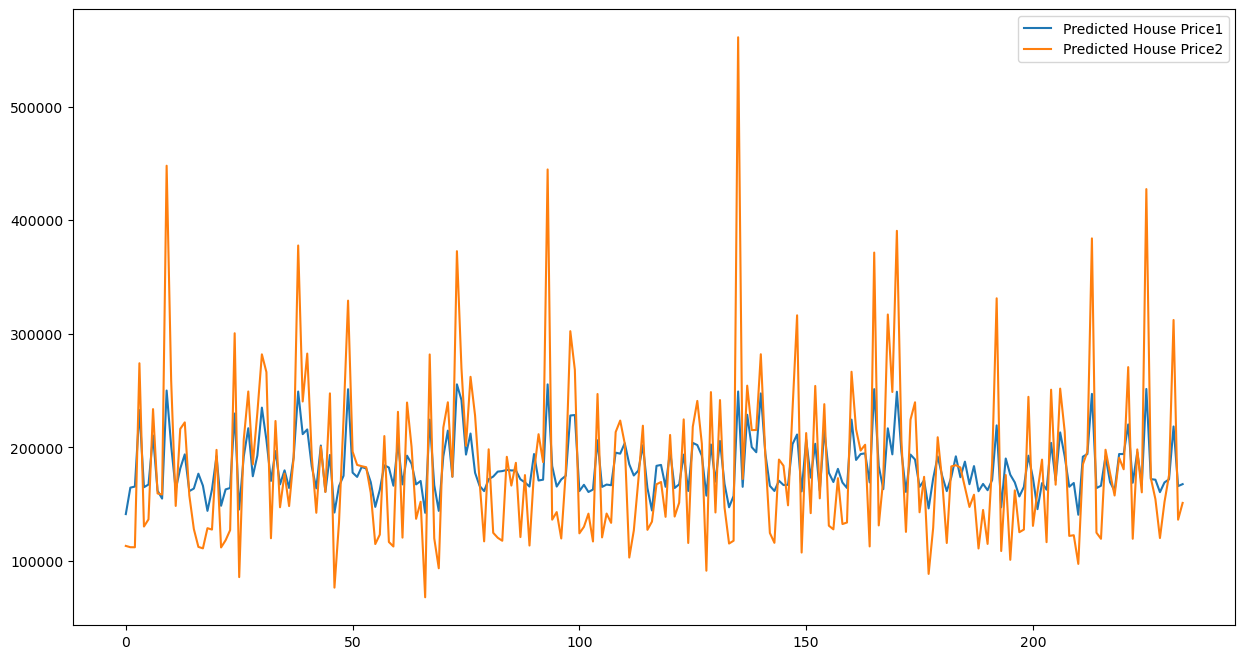
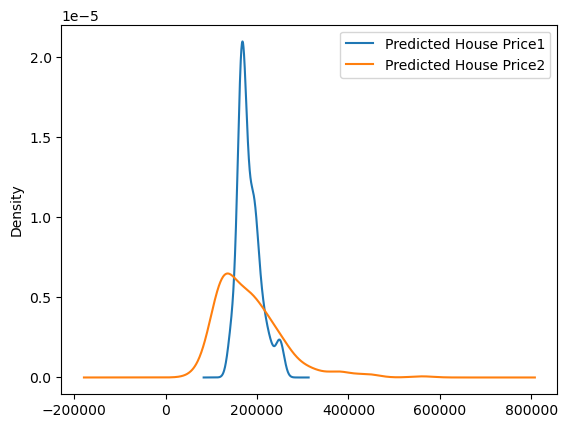
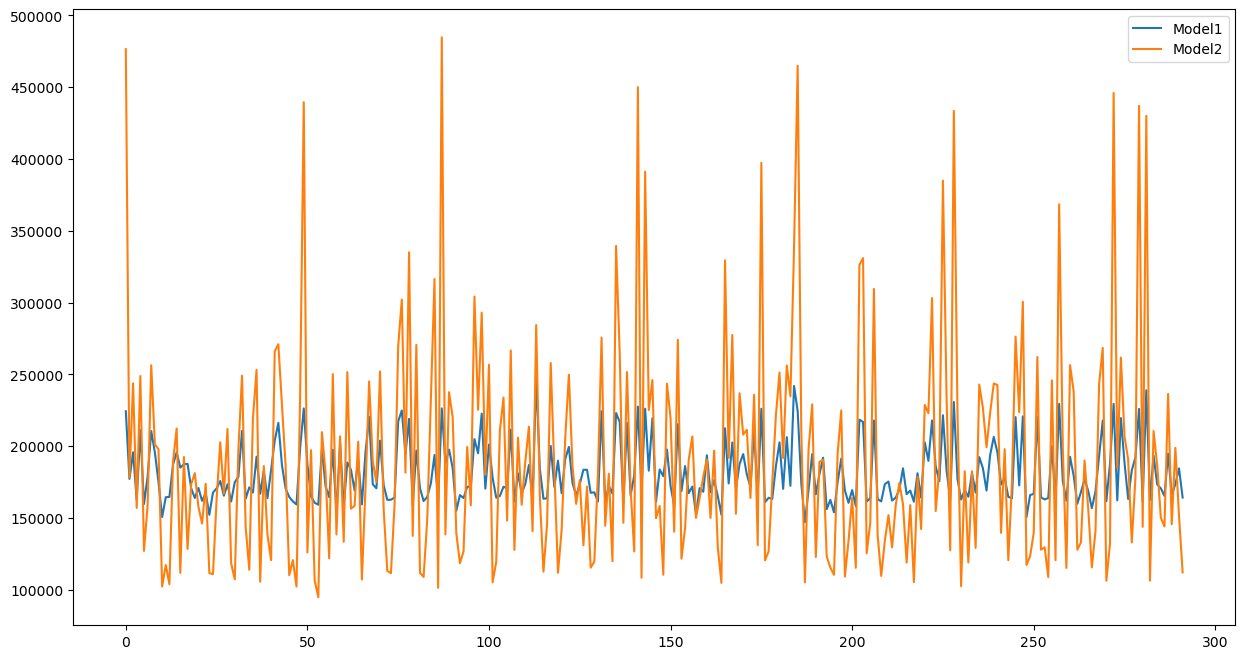
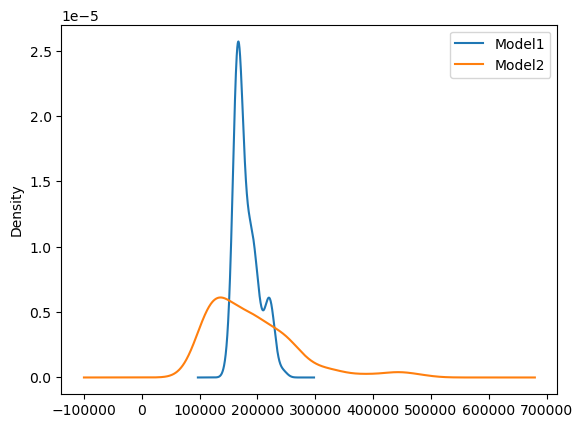


Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

If different platforms were used, mention that as well.

* **Interpretation of the Results**



    Give a summary of what results were interpreted from the visualizations, pre-processing and modelling.

**CONCLUSION**

* **Key Findings and Conclusions of the Study**
* Count plot provided the count of each categorical value present in column which provided how the data is distributed.
* These plots gave an idea how to replace the NaN values present in categorical features.
* Bar plots were plotted for all categorical variables with respect to the target variable (sales price) for all the years and one can see that neighbourhood, Excellent kitchen quality, building type, house style with Two and one-half story: 2nd level finished, good basement condition etc play important role in the sales price house.
* They later with the help of bar plot maximum sales and average sale each year and for every year sales every month. The Maximum sales happened to be in 2007.
* Next, a categorical plot was plotted for house style building type for sales price, the sales most commonly occurred for building type one family.
* After training, the random forest algorithm gave better results with accuracy of 81.34% for model without been tuned using hyperparameter tunning. With tunned model the final accuracy id 41.43%.
* The accuracy was less because the outliers were present in the data. The data was getting reduced to half if we remove outliers.
* Later after pickling the predicted and original sales price was plotted to check the comparison of both the models with and without tunning. The model without tunning gave better results as one can see through the error plot. The kde plot gave us the range how the data is getting predicted, the tunned model spread was lower with high density and whereas the one without tunning has good spread for predicted the sales price.
* Lastly, the sales price was predicted for the test data which was provided after applying data pre-processing, feature engineering as done for the training data.
* **Learning Outcomes of the Study in respect of Data Science**
* What I learned here is how one has to extensively look at the data and see its behaviour with help of EDA. This will help us identify the important features for predicting the target.
* **Limitations of this work and Scope for Future Work**
* The data can be more explored for feature engineering with more literature review.
* The Outliers present can be treated in a more efficient way to improve the accuracy of the model.
* We can apply more algorithms along with hyper tunning parameters to improve the accuracy.