

**NAME OF THE PROJECT**

**House Price Prediction Project**

Submitted by:

**Neha Sharma**

**INTRODUCTION**

**Problem Statement**

Houses are one of the necessary need 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. Our problem is related to one such housing company.

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:

• Which variables are important to predict the price of variable?

• How do these variables describe the price of the house?

**Business Goal**

Here we need to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

**Review of Literature**

[**LOT FRONTAGE**](https://www.lawinsider.com/dictionary/lot-frontage) means the horizontal distance between the side lot lines measured along the front lot lines. Where the front lot line is not a straight line or where the side lot lines are not parallel, the lot frontage is to be measured by a line parallel to the chord of the lot frontage and a distance from the front lot line equal to the required depth of the front yard.

**Lot Area:** A lot area is the total area of a property, including the yard up to the boundaries (property line), while the floor area is the area inside the building that is occupiable, up to and including the exterior walls.

**Alley:** a narrow passageway between or behind buildings.

**Motivation for the Problem Undertaken**

The purpose of this research is to explore the factors that influence the house prices and what step should be taken on which feature to maximize the profit.

**Analytical Problem Framing**

**Data Sources and their formats**

The dataset used in this project comes from the client database. The dataset is in CSV format and divided into train and test data set. In train data set there are 1168 entries represents aggregate information about 81 features of house where test data set contain 292 rows.

Following are the description of each feature present in the data set.

**MSSubClass:** Identifies the type of dwelling involved in the sale.

20 1-STORY 1946 & NEWER ALL STYLES

30 1-STORY 1945 & OLDER

40 1-STORY W/FINISHED ATTIC ALL AGES

45 1-1/2 STORY - UNFINISHED ALL AGES

50 1-1/2 STORY FINISHED ALL AGES

60 2-STORY 1946 & NEWER

70 2-STORY 1945 & OLDER

75 2-1/2 STORY ALL AGES

80 SPLIT OR MULTI-LEVEL

85 SPLIT FOYER

90 DUPLEX - ALL STYLES AND AGES

120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER

150 1-1/2 STORY PUD - ALL AGES

160 2-STORY PUD - 1946 & NEWER

180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER

190 2 FAMILY CONVERSION - ALL STYLES AND AGES

**MSZoning:** Identifies the general zoning classification of the sale.

A Agriculture

C Commercial

FV Floating Village Residential

I Industrial

RH Residential High Density

RL Residential Low Density

RP Residential Low Density Park

RM Residential Medium Density

**LotFrontage:** Linear feet of street connected to property

**LotArea:** Lot size in square feet

**Street:** Type of road access to property

Grvl Gravel

Pave Paved

**Alley:** Type of alley access to property

Grvl Gravel

Pave Paved

NA No alley access

**LotShape:** General shape of property

Reg Regular

IR1 Slightly irregular

IR2 Moderately Irregular

IR3 Irregular

**LandContour:** Flatness of the property

Lvl Near Flat/Level

Bnk Banked - Quick and significant rise from street grade to building

HLS Hillside - Significant slope from side to side

Low Depression

**Utilities:** Type of utilities available

AllPub All public Utilities (E,G,W,& S)

NoSewr Electricity, Gas, and Water (Septic Tank)

NoSeWa Electricity and Gas Only

ELO Electricity only

**LotConfig:** Lot configuration

Inside Inside lot

Corner Corner lot

CulDSac Cul-de-sac

FR2 Frontage on 2 sides of property

FR3 Frontage on 3 sides of property

**LandSlope:** Slope of property

Gtl Gentle slope

Mod Moderate Slope

Sev Severe Slope

**Neighborhood:** Physical locations within Ames city limits

Blmngtn Bloomington Heights

Blueste Bluestem

BrDale Briardale

BrkSide Brookside

ClearCr Clear Creek

CollgCr College Creek

Crawfor Crawford

Edwards Edwards

Gilbert Gilbert

IDOTRR Iowa DOT and Rail Road

MeadowV Meadow Village

Mitchel Mitchell

Names North Ames

NoRidge Northridge

NPkVill Northpark Villa

NridgHt Northridge Heights

NWAmes Northwest Ames

OldTown Old Town

SWISU South & West of Iowa State University

Sawyer Sawyer

SawyerW Sawyer West

Somerst Somerset

StoneBr Stone Brook

Timber Timberland

Veenker Veenker

**Condition1:** Proximity to various conditions

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

**Condition2:** Proximity to various conditions (if more than one is present)

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

**BldgType:** Type of dwelling

1Fam Single-family Detached

2FmCon Two-family Conversion; originally built as one-family dwelling

Duplx Duplex

TwnhsE Townhouse End Unit

TwnhsI Townhouse Inside Unit

**HouseStyle:** Style of dwelling

1Story One story

1.5Fin One and one-half story: 2nd level finished

1.5Unf One and one-half story: 2nd level unfinished

2Story Two story

2.5Fin Two and one-half story: 2nd level finished

2.5Unf Two and one-half story: 2nd level unfinished

SFoyer Split Foyer

SLvl Split Level

**OverallQual:** Rates the overall material and finish of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

**OverallCond:** Rates the overall condition of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

**YearBuilt:** Original construction date

**YearRemodAdd:** Remodel date (same as construction date if no remodeling or additions)

**RoofStyle:** Type of roof

Flat Flat

Gable Gable

Gambrel Gabrel (Barn)

Hip Hip

Mansard Mansard

Shed Shed

**RoofMatl**: Roof material

ClyTile Clay or Tile

CompShg Standard (Composite) Shingle

Membran Membrane

Metal Metal

Roll Roll

Tar&Grv Gravel & Tar

WdShake Wood Shakes

WdShngl Wood Shingles

**Exterior1st:** Exterior covering on house

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

**Exterior2nd:** Exterior covering on house (if more than one material)

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

**MasVnrType:** Masonry veneer type

BrkCmn Brick Common

BrkFace Brick Face

CBlock Cinder Block

None None

Stone Stone

**MasVnrArea:** Masonry veneer area in square feet

**ExterQual:** Evaluates the quality of the material on the exterior

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

**ExterCond:** Evaluates the present condition of the material on the exterior

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

**Foundation:** Type of foundation

BrkTil Brick & Tile

CBlock Cinder Block

PConc Poured Contrete

Slab Slab

Stone Stone

Wood Wood

**BsmtQual:** Evaluates the height of the basement

Ex Excellent (100+ inches)

Gd Good (90-99 inches)

TA Typical (80-89 inches)

Fa Fair (70-79 inches)

Po Poor (<70 inches

NA No Basement

**BsmtCond:** Evaluates the general condition of the basement

Ex Excellent

Gd Good

TA Typical - slight dampness allowed

Fa Fair - dampness or some cracking or settling

Po Poor - Severe cracking, settling, or wetness

NA No Basement

**BsmtExposure:** Refers to walkout or garden level walls

Gd Good Exposure

Av Average Exposure (split levels or foyers typically score average or above)

Mn Mimimum Exposure

No No Exposure

NA No Basement

**BsmtFinType1:** Rating of basement finished area

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

NA No Basement

**BsmtFinSF1:** Type 1 finished square feet

**BsmtFinType2:** Rating of basement finished area (if multiple types)

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

NA No Basement

**BsmtFinSF2:** Type 2 finished square feet

**BsmtUnfSF:** Unfinished square feet of basement area

**TotalBsmtSF:** Total square feet of basement area

**Heating:** Type of heating

Floor Floor Furnace

GasA Gas forced warm air furnace

GasW Gas hot water or steam heat

Grav Gravity furnace

OthW Hot water or steam heat other than gas

Wall Wall furnace

**HeatingQC:** Heating quality and condition

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

**CentralAir:** Central air conditioning

N No

Y Yes

**Electrical:** Electrical system

SBrkr Standard Circuit Breakers & Romex

FuseA Fuse Box over 60 AMP and all Romex wiring (Average)

FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)

FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)

Mix Mixed

**1stFlrSF:** First Floor square feet

**2ndFlrSF:** Second floor square feet

**LowQualFinSF**: Low quality finished square feet (all floors)

**GrLivArea:** Above grade (ground) living area square feet

**BsmtFullBath:** Basement full bathrooms

**BsmtHalfBath:** Basement half bathrooms

**FullBath:** Full bathrooms above grade

**HalfBath:** Half baths above grade

**Bedroom:** Bedrooms above grade (does NOT include basement bedrooms)

**Kitchen:** Kitchens above grade

**KitchenQual:** Kitchen quality

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

**TotRmsAbvGrd:** Total rooms above grade (does not include bathrooms)

**Functional:** Home functionality (Assume typical unless deductions are warranted)

Typ Typical Functionality

Min1 Minor Deductions 1

Min2 Minor Deductions 2

Mod Moderate Deductions

Maj1 Major Deductions 1

Maj2 Major Deductions 2

Sev Severely Damaged

Sal Salvage only

**Fireplaces:** Number of fireplaces

**FireplaceQu:** Fireplace quality

Ex Excellent - Exceptional Masonry Fireplace

Gd Good - Masonry Fireplace in main level

TA Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement

Fa Fair - Prefabricated Fireplace in basement

Po Poor - Ben Franklin Stove

NA No Fireplace

**GarageType:** Garage location

2Types More than one type of garage

Attchd Attached to home

Basment Basement Garage

BuiltIn Built-In (Garage part of house - typically has room above garage)

CarPort Car Port

Detchd Detached from home

NA No Garage

**GarageYrBlt:** Year garage was built

**GarageFinish:** Interior finish of the garage

Fin Finished

RFn Rough Finished

Unf Unfinished

NA No Garage

**GarageCars:** Size of garage in car capacity

**GarageArea:** Size of garage in square feet

**GarageQual:** Garage quality

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

**GarageCond:** Garage condition

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

**PavedDrive:** Paved driveway

Y Paved

P Partial Pavement

N Dirt/Gravel

**WoodDeckSF:** Wood deck area in square feet

**OpenPorchSF:** Open porch area in square feet

**EnclosedPorch:** Enclosed porch area in square feet

**3SsnPorch:** Three season porch area in square feet

**ScreenPorch:** Screen porch area in square feet

**PoolArea:** Pool area in square feet

**PoolQC:** Pool quality

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

NA No Pool

**Fence:** Fence quality

GdPrv Good Privacy

MnPrv Minimum Privacy

GdWo Good Wood

MnWw Minimum Wood/Wire

NA No Fence

**MiscFeature:** Miscellaneous feature not covered in other categories

Elev Elevator

Gar2 2nd Garage (if not described in garage section)

Othr Other

Shed Shed (over 100 SF)

TenC Tennis Court

NA None

**MiscVal:** $Value of miscellaneous feature

**MoSold:** Month Sold (MM)

**YrSold:** Year Sold (YYYY)

**SaleType:** Type of sale

WD Warranty Deed - Conventional

CWD Warranty Deed - Cash

VWD Warranty Deed - VA Loan

New Home just constructed and sold

COD Court Officer Deed/Estate

Con Contract 15% Down payment regular terms

ConLw Contract Low Down payment and low interest

ConLI Contract Low Interest

ConLD Contract Low Down

Oth Other

**SaleCondition:** Condition of sale

Normal Normal Sale

Abnorml Abnormal Sale - trade, foreclosure, short sale

AdjLand Adjoining Land Purchase

Alloca Allocation - two linked properties with separate deeds, typically condo with a garage unit

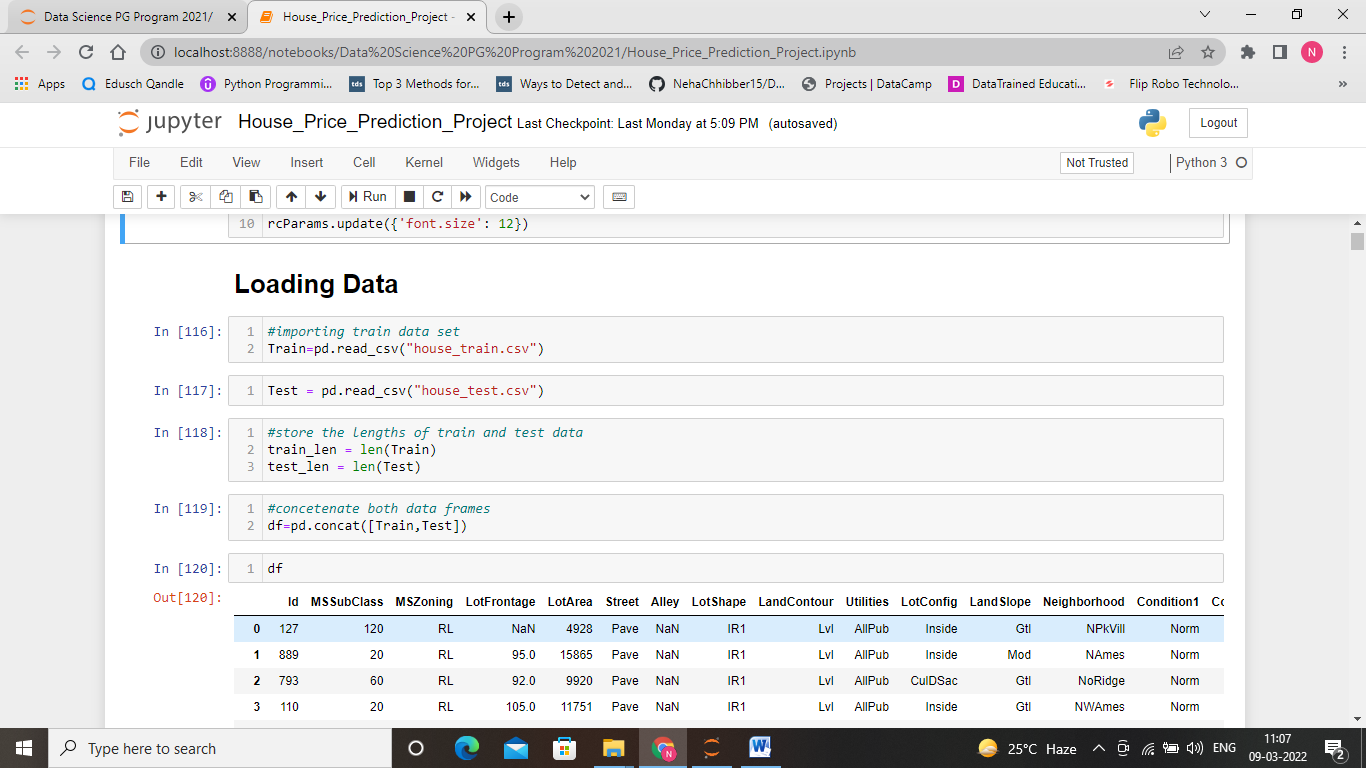
Family Sale between family members

Partial Home was not completed when last assessed (associated with New Homes)

**Preparing Data Set**

First, we need to install different libraries that will help us to apply different operations on the data set. Imported pandas for applying operations on dataframe, numpy for mathematical operations, seaborn and matplotlib for visualization and scikit learn for building model.

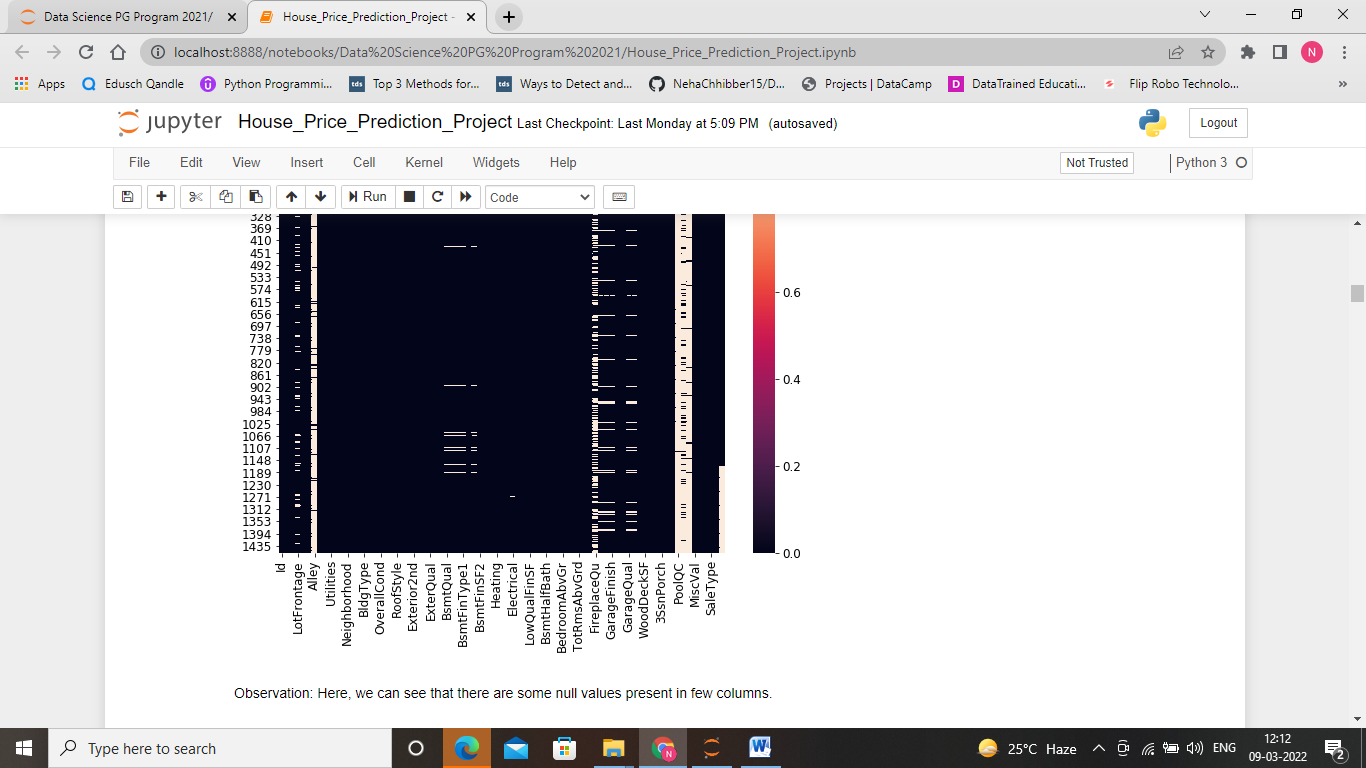
We have concatenated train and test data so that preprocessing steps should be applied on both the data set.



**Exploratory Data Analysis**

**Let’s do some exploratory data analysis on the given data set.**

There are null values present in few of the columns. Let us visualize by using heat map.



We need to treat these null values.

# ****Data Visualization****

# **On separating data set into numerical and categorical features we get 37 numerical and 44 categorical features.**

# **Let us visualize each numerical feature and explore more about each feature.**

# D:\Flip Robo Technolyges\Im1.png **Here, we can see that data present in the numerical feature are more skewed. So, we need to normalize it.**

# 

# In descriptive table we can see the average value of each feature.

# Bivariate Analysis

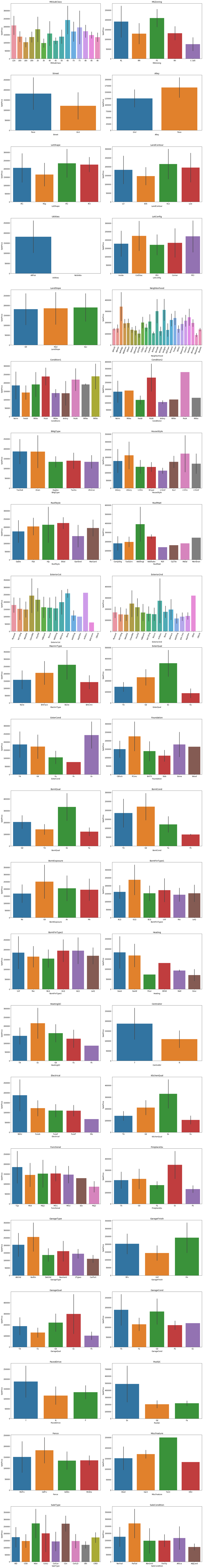
# Let us plot the scatter plot for each numerical variable against sales.

# D:\Flip Robo Technolyges\im2.png

**Observation:**

1. Lot frontage and lot area is not impacting much on sales prices.
2. As the year built shows the direct relation with sale price. Newly constructed houses have more sale price as compare to old houses.
3. Average condition houses are more in number and have more sale prices.
4. Recent modification made in the houses increases the sale price.
5. Masonry veneer area in square feet increases the sale price at certain level.
6. More the rating of basement finished area more will be the sales prices.
7. If the basement area is unfurnished than the sale prices are also affecting adversely.
8. More the basement area in square feet more will be the sale prices.
9. Sale price of 1st floor are more as compare to second floor. Moreover, 1st floor is coming up with more area as compare to second floor.
10. Low quality finished square feet is not impacting much on sale price.
11. More the number of full baths more is the sale price. Maximum three baths are available. Majority of the houses have 1 half bath and 2 full baths.
12. Most of the houses having 4 bedrooms have higher prices.
13. More the total number of rooms above grade (does not include bathrooms) more will be the prices.
14. More the garage area more will be the prices.
15. More the open porch area more will be the sale prices.
16. Pool area is not affecting much on sale prices.
17. We can see that in Month of September sales is more as compare to other months.

Let us observe categorical feature as well.



**Observations**

1. Here, we can see that prices of 2-STORY 1946 & NEWER are more as compare to other.

2. Sale price of houses in zone of Floating Village Residential and Residential Low Density are more as compare to others.

3. Sale price of houses having paved streets are more as compare to gravel street.

4. Sale price of Moderately Irregular lotshape is more as compare to others.

5. Sale Price of Hillside property is more as compare to other property.

6. Sale Price of houses having all public Utilities (E,G,W,& S) are more as compare to others.

7. Sale Price of Frontage on 3 sides of property and Cul-de-sac are more as compare to others.

8. Land slope does not affect the sale prices much.

9. Sale Prices of houses in the neighbourhood of Northridge, Stone Brook, Northridge Heights are more as compare to others.

10. Sale price of house near positive off-site feature--park, greenbelt, etc., Within 200' of North-South Railroad, Adjacent to postive off-site feature are more as compare to others.

11. Sale Prices of Townhouse End Unit and Single-family Detached are more as compare to others.

12. Sale Prices of Two story and Two and one-half story: 2nd level finished are more as compare to others.

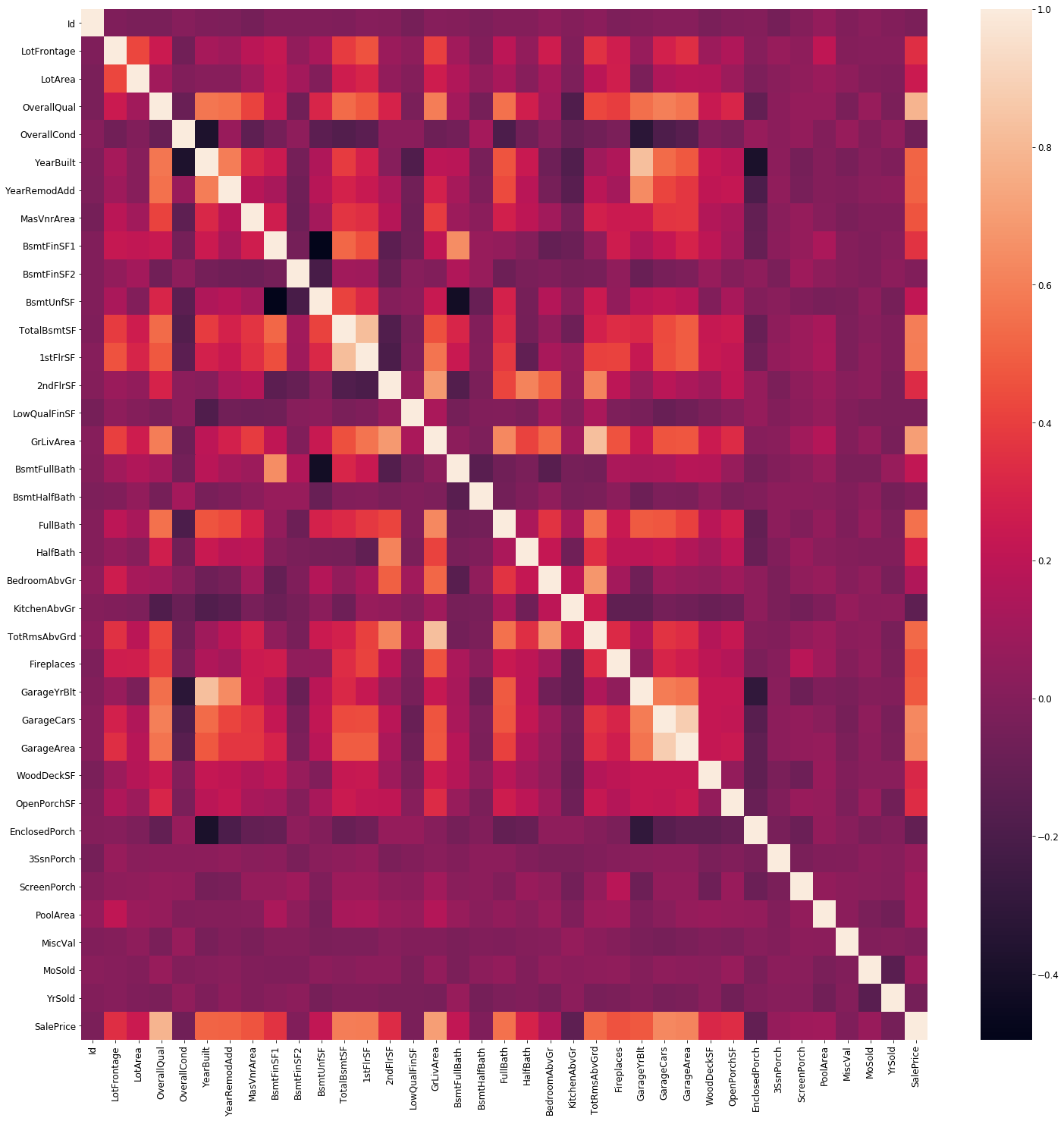
13.Sale Price of houses having roofstyle as Flat, hip and shed are more as compare to others.

14. Sale Price of houses having roofmaterial of Wood Shingles is more as compare to other.

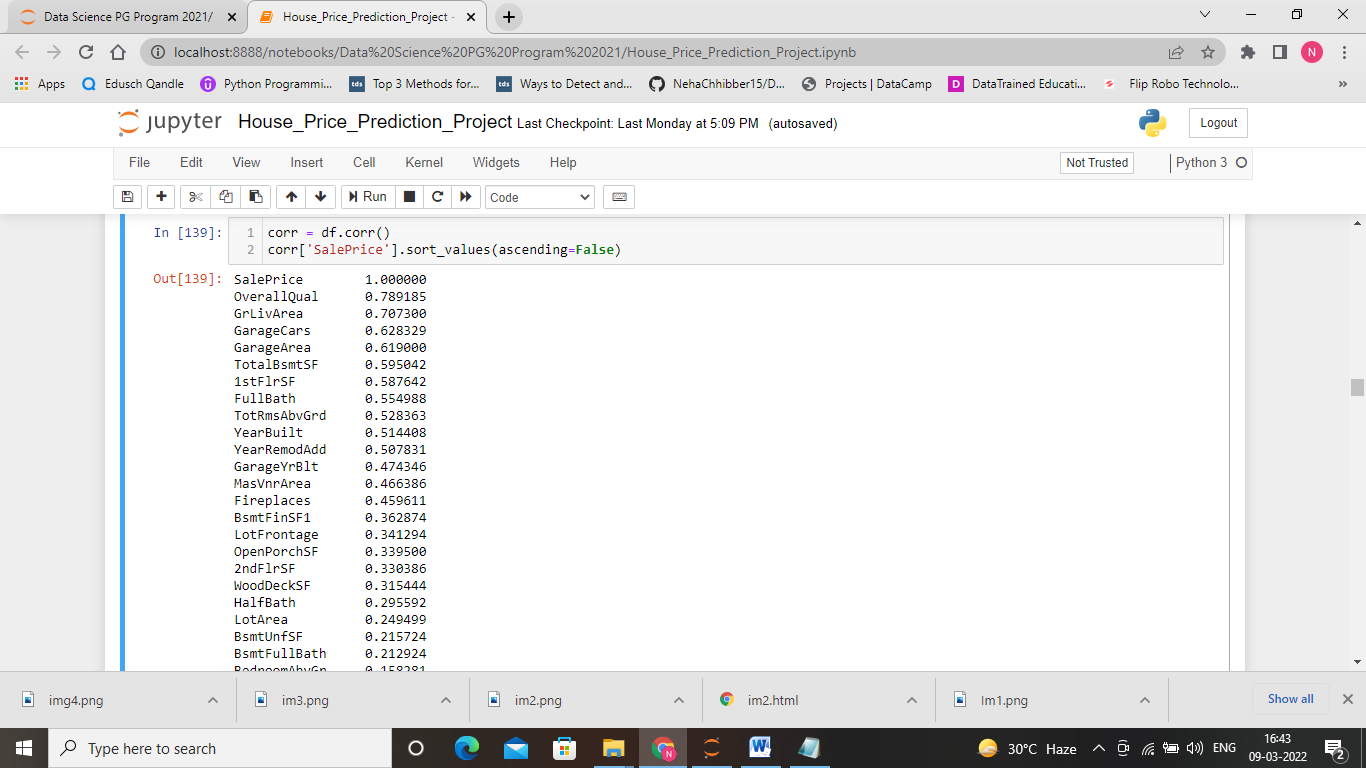
**Correlation**

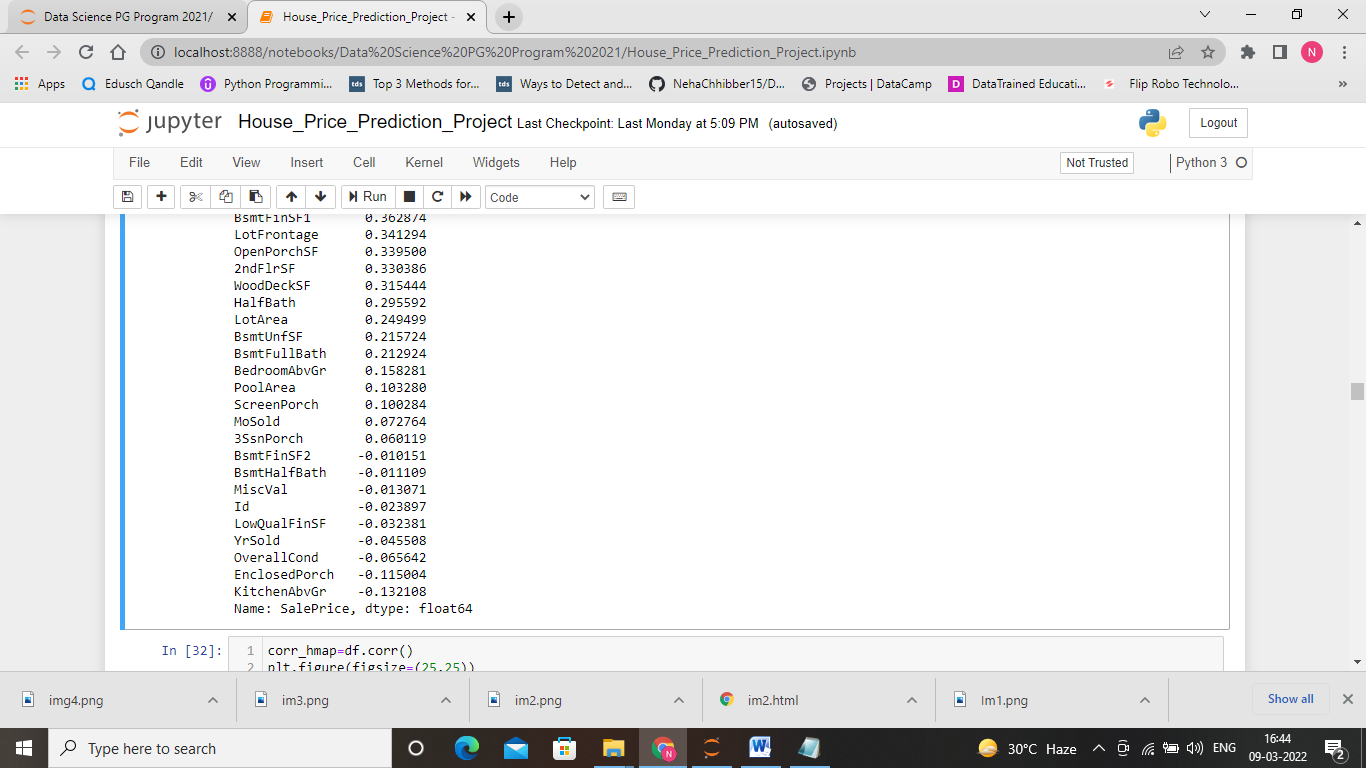
We are going to create now a correlation matrix to quantify and summarize the relationships between the variables.This correlation matrix is closely related with covariance matrix, in fact it is a rescaled version of the covariance matrix, computed from standardize features.

It is a square matrix (with the same number of columns and rows) that contains the Person’s correlation coefficient. Now, let us find the correlation of each feature by using heatmap.



We also need to find the correlation of each feature with sale price. So that we can find out which features are more important to predict our target variable.

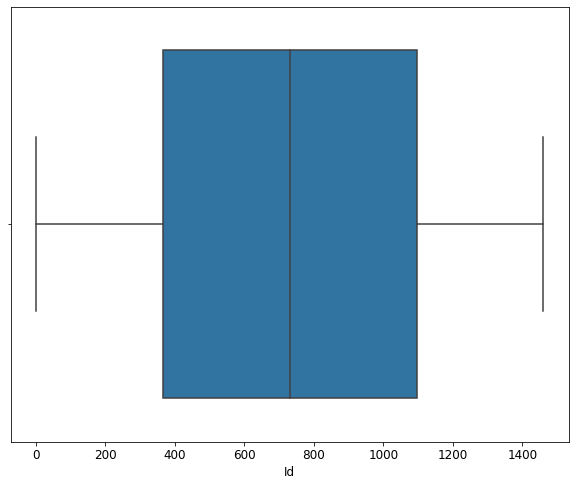


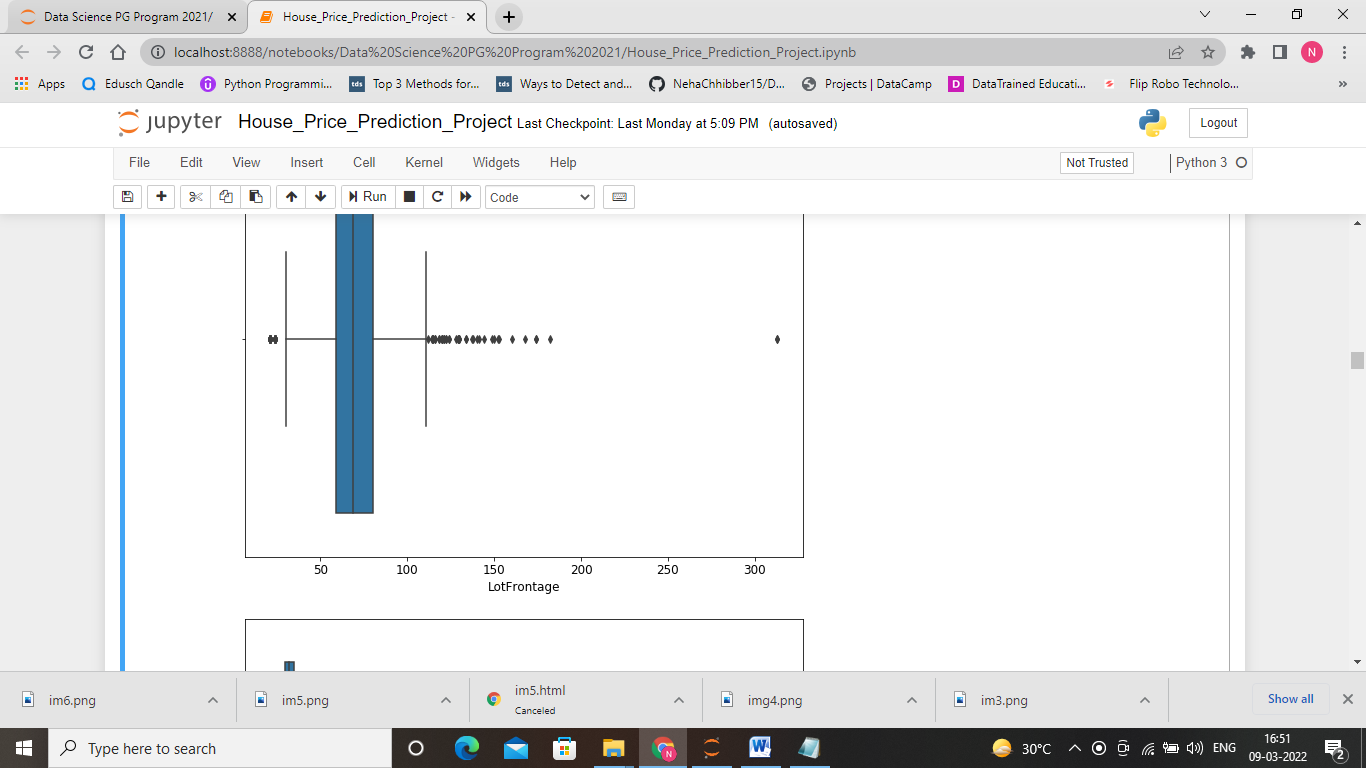


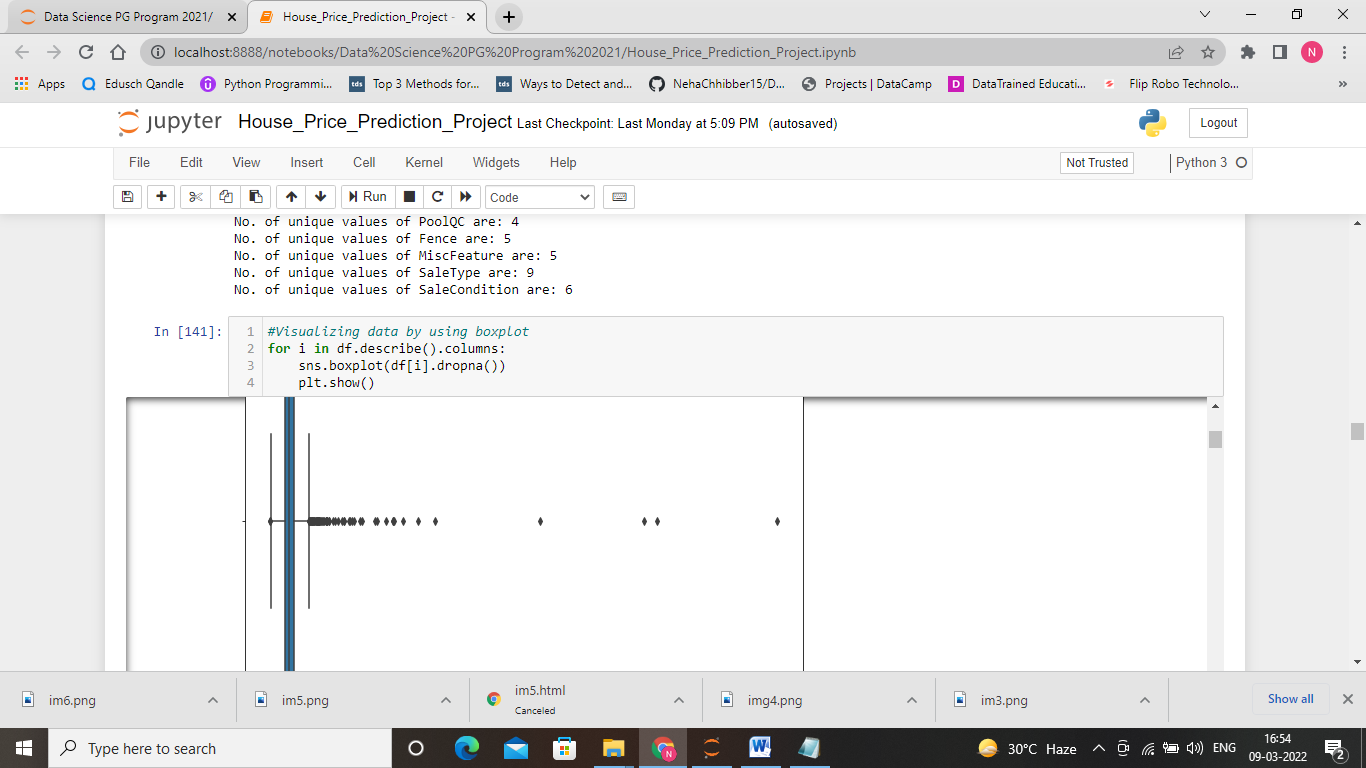
Here, we can see that OverallQual, Garage area, Garage cars are the most imaportant feature to predict sale price.

**Boxplot**

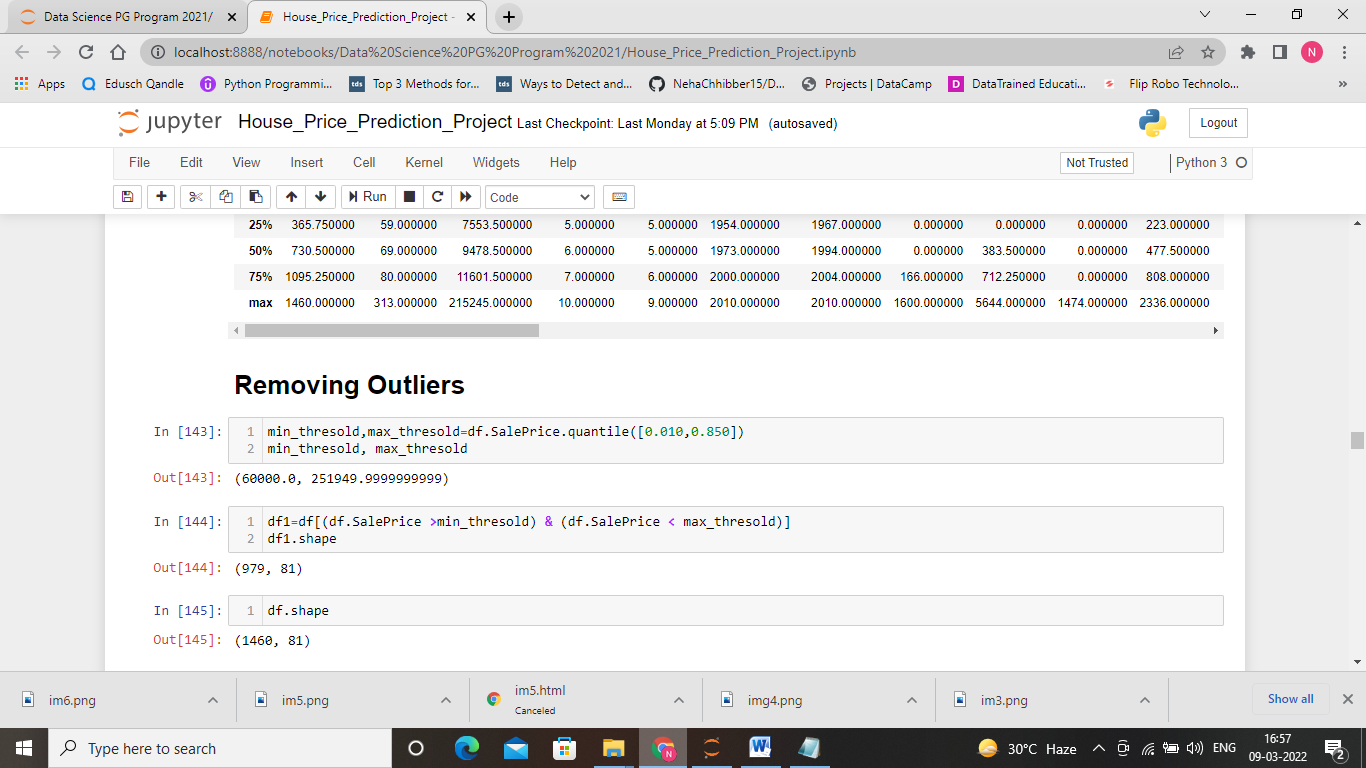
Let us find out whether there is any outlier present in the given dataset or not.







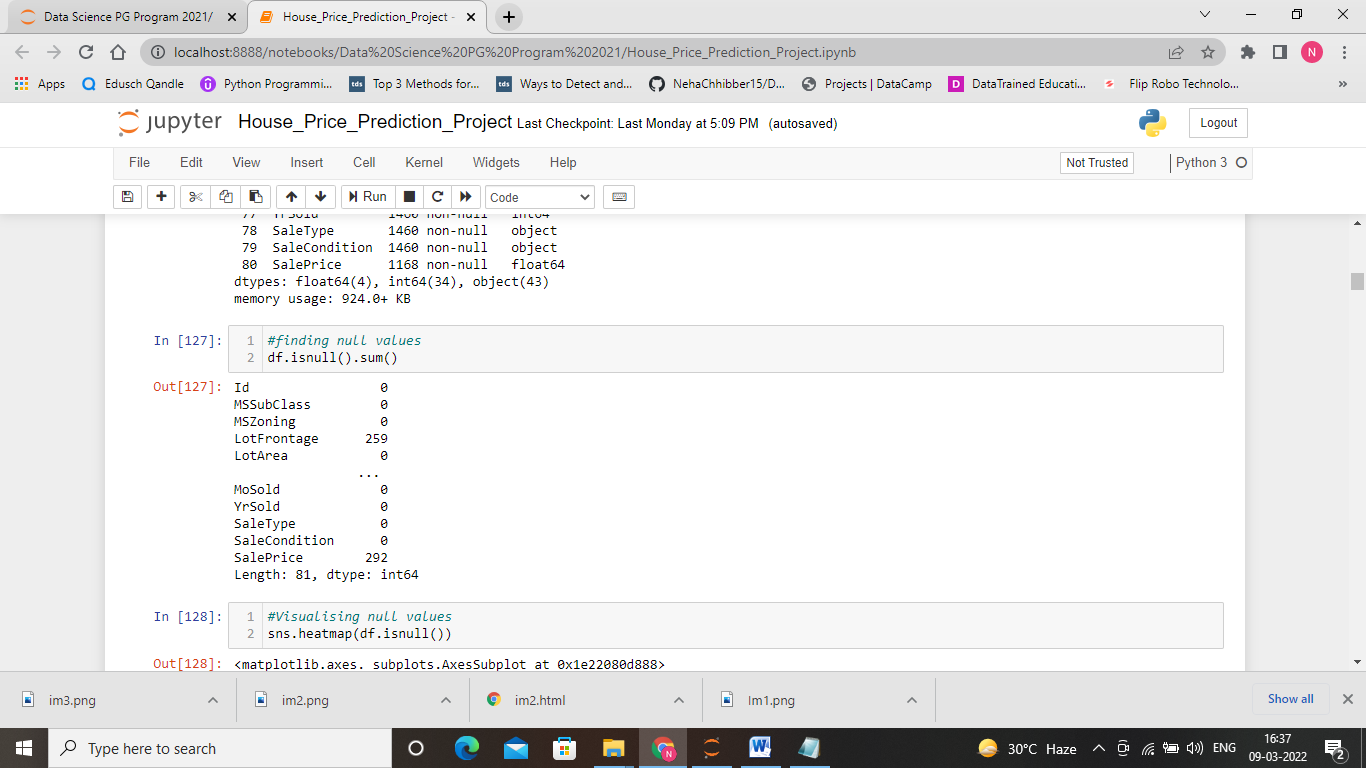
On observing the rest of the feature, we have observed that there are many outliers present. So, we need to remove or treat outliers. For that we have used quantile method by taking the range of 10 to 85% of quantile.



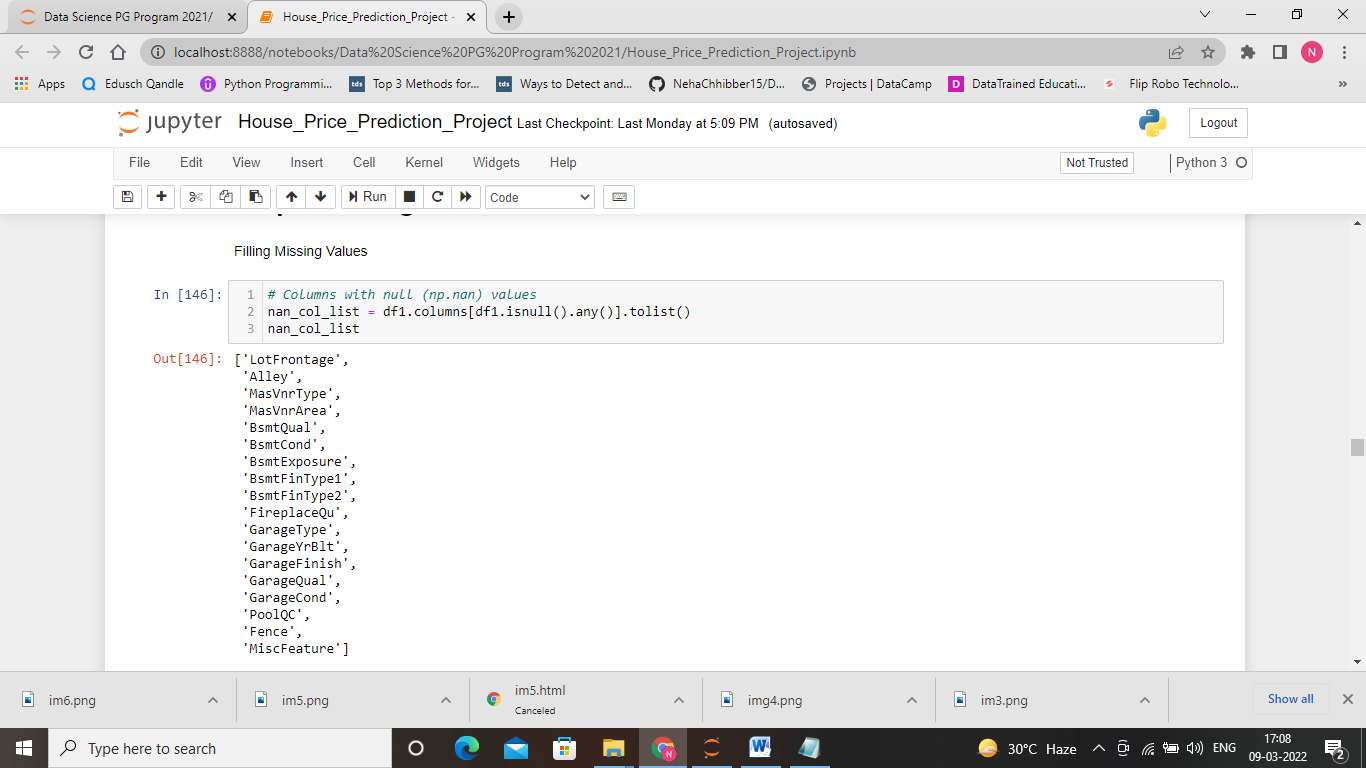
**Data Preprocessing**

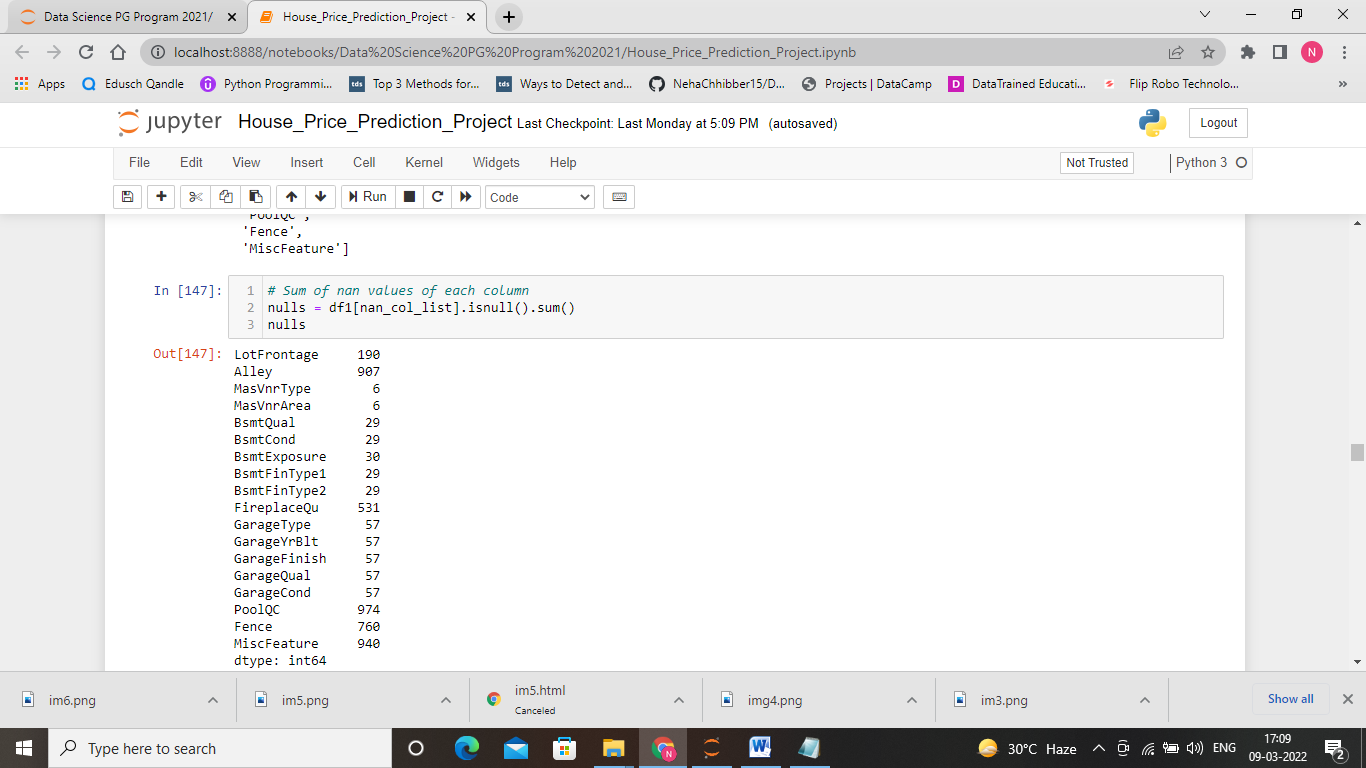
In the first section of the project, we will make an exploratory analysis of the dataset and provide some observations.

Here, we have checked whether any null values present in the given data set. For that we have checked df.info to know about missing values.



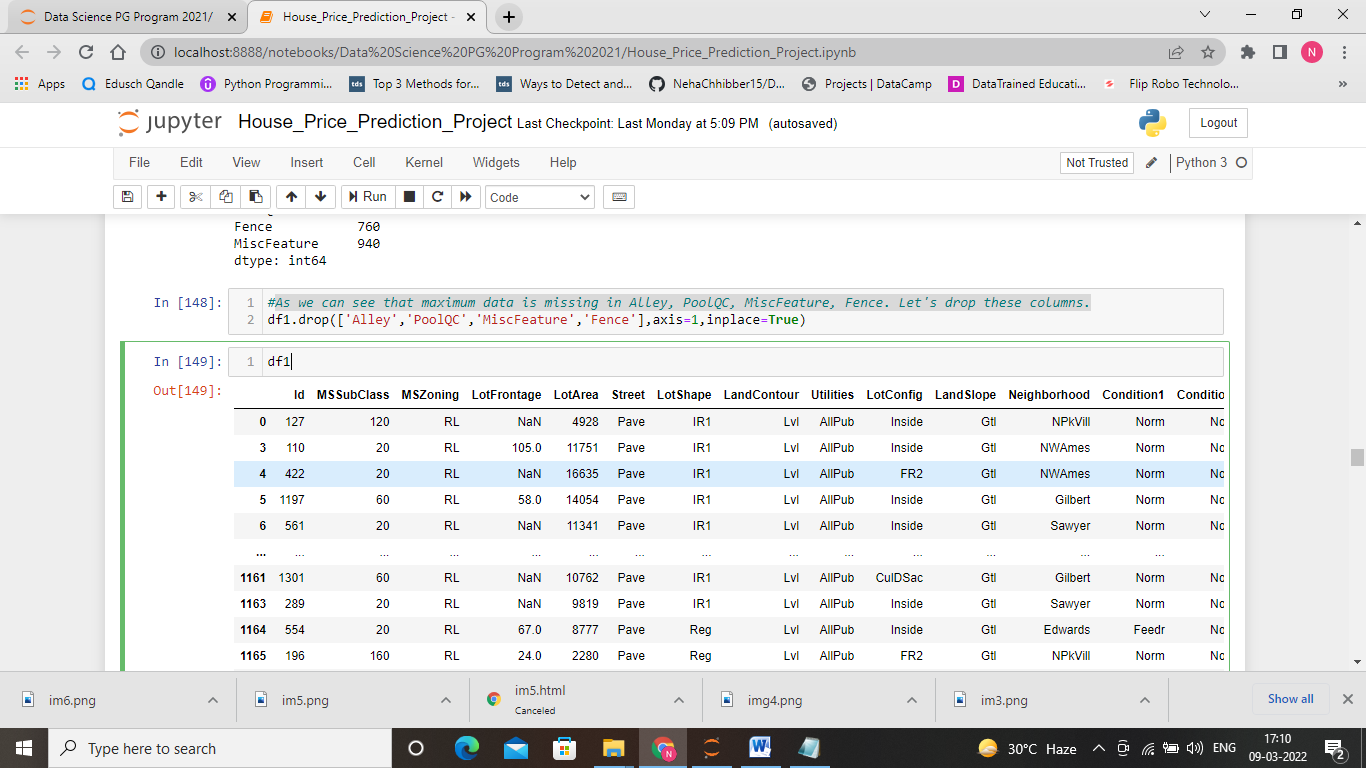
Here, we observed that there are null values in the given data set. Let us first find out which columns has null values.

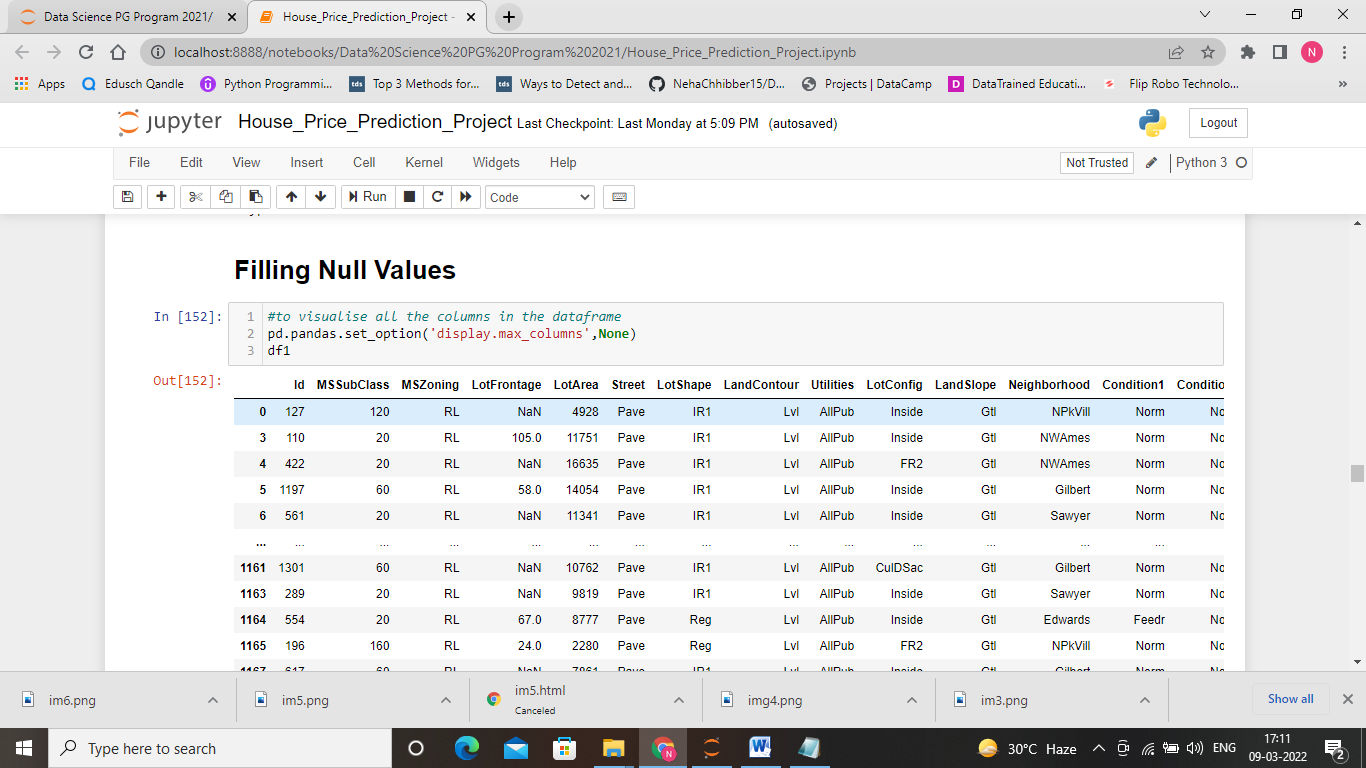


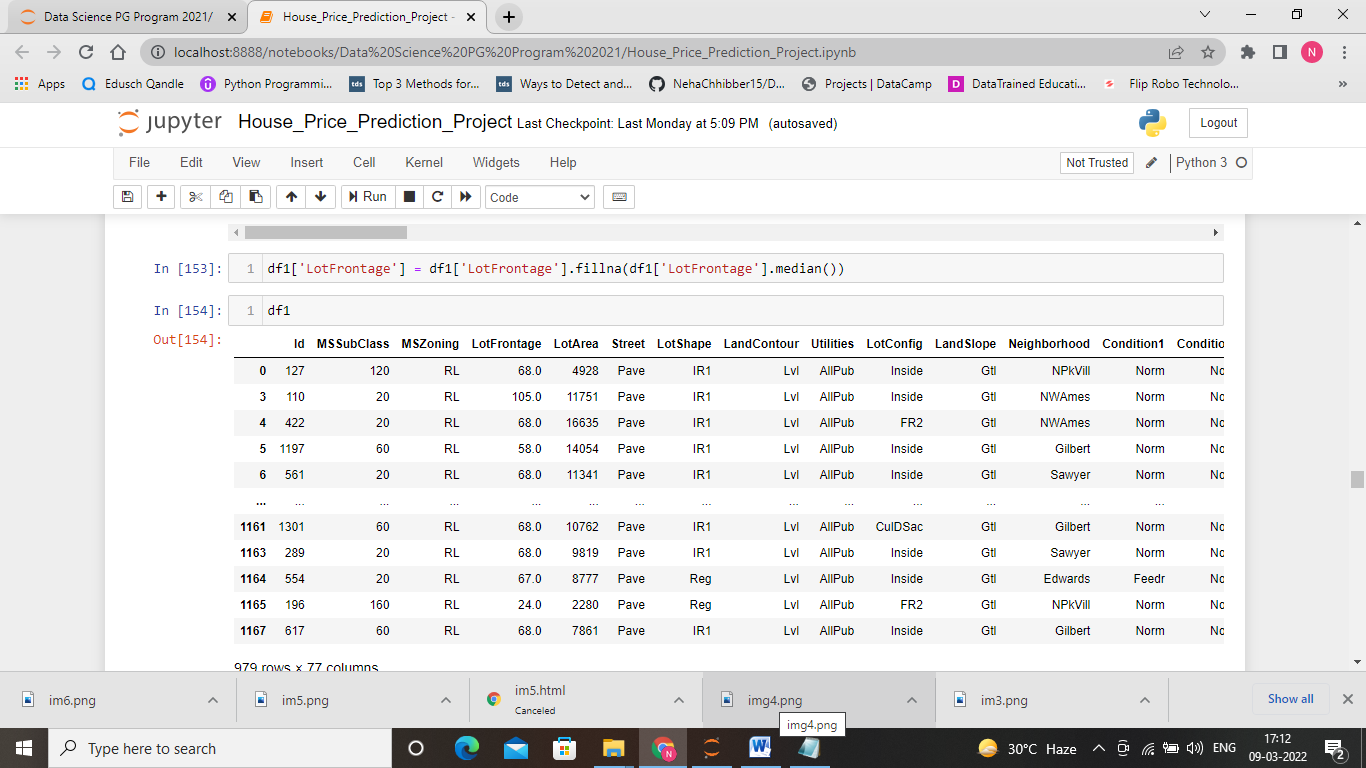


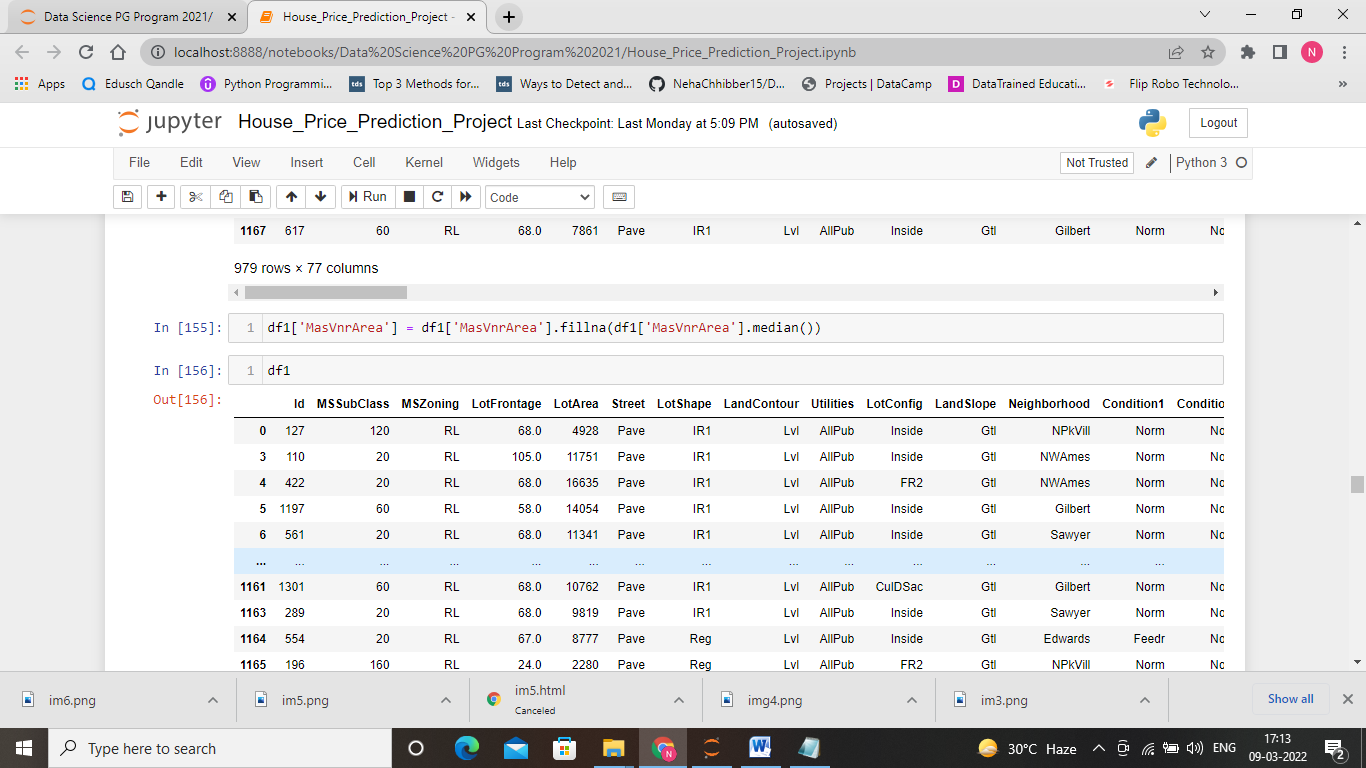
We can drop the columns that have 90% of null values.

As we can see that maximum data is missing in Alley, PoolQC, MiscFeature, Fence. Let's drop these columns.

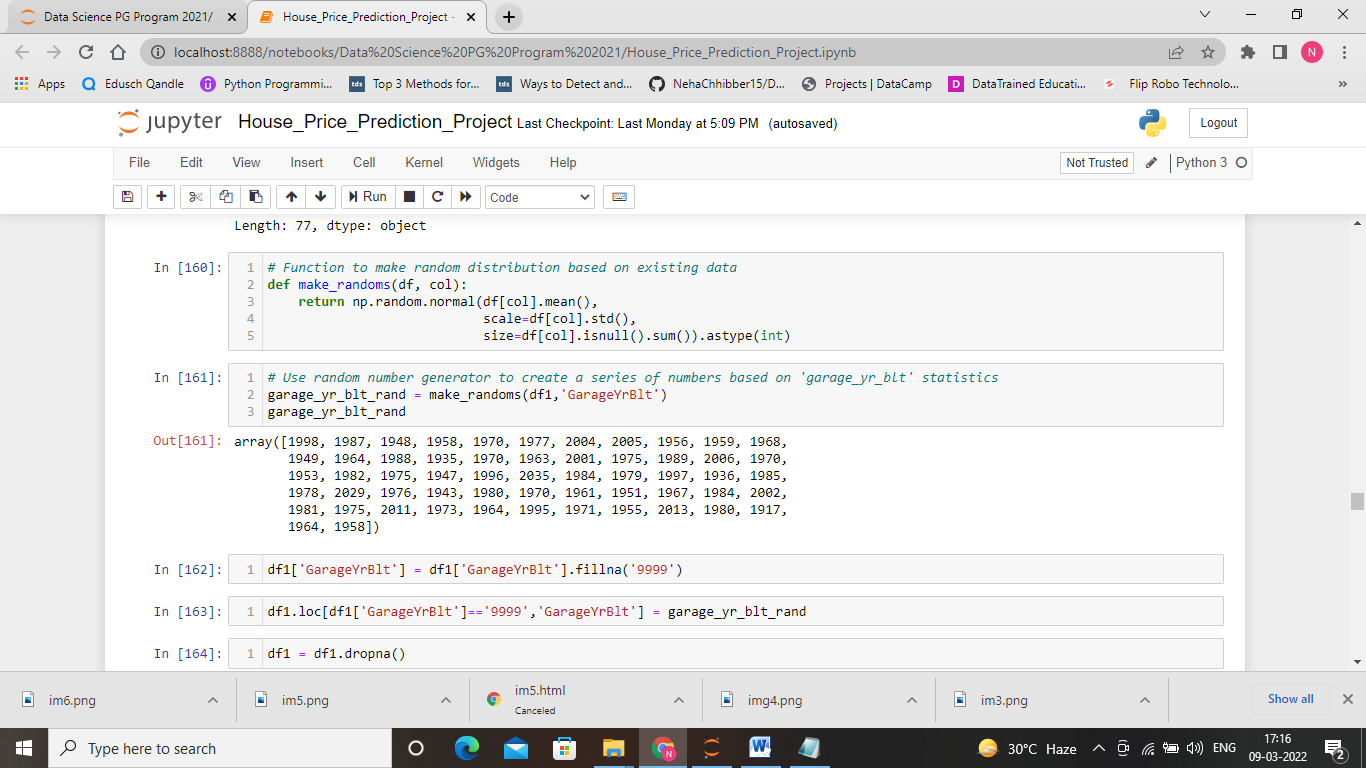




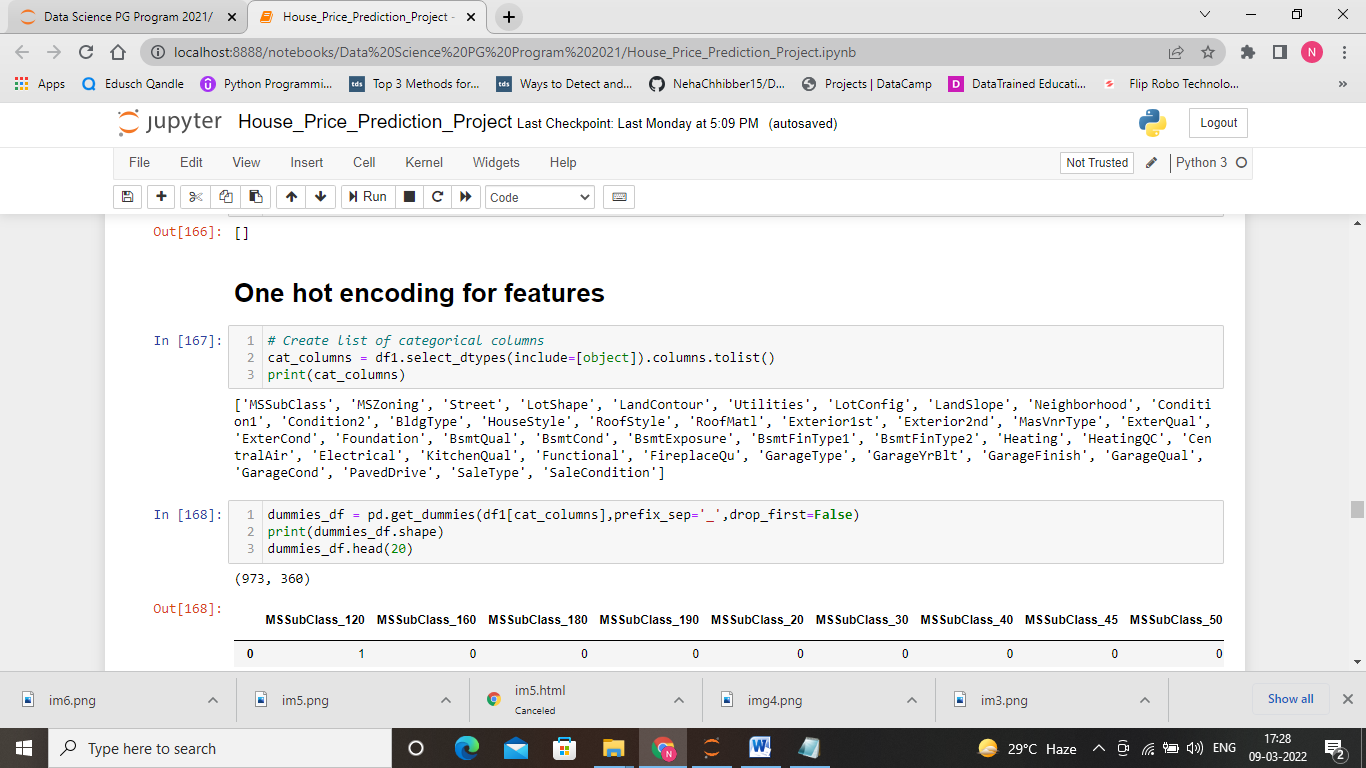




We have replace numeric value with median and categorical value with mode. We have used random function to replace year data.

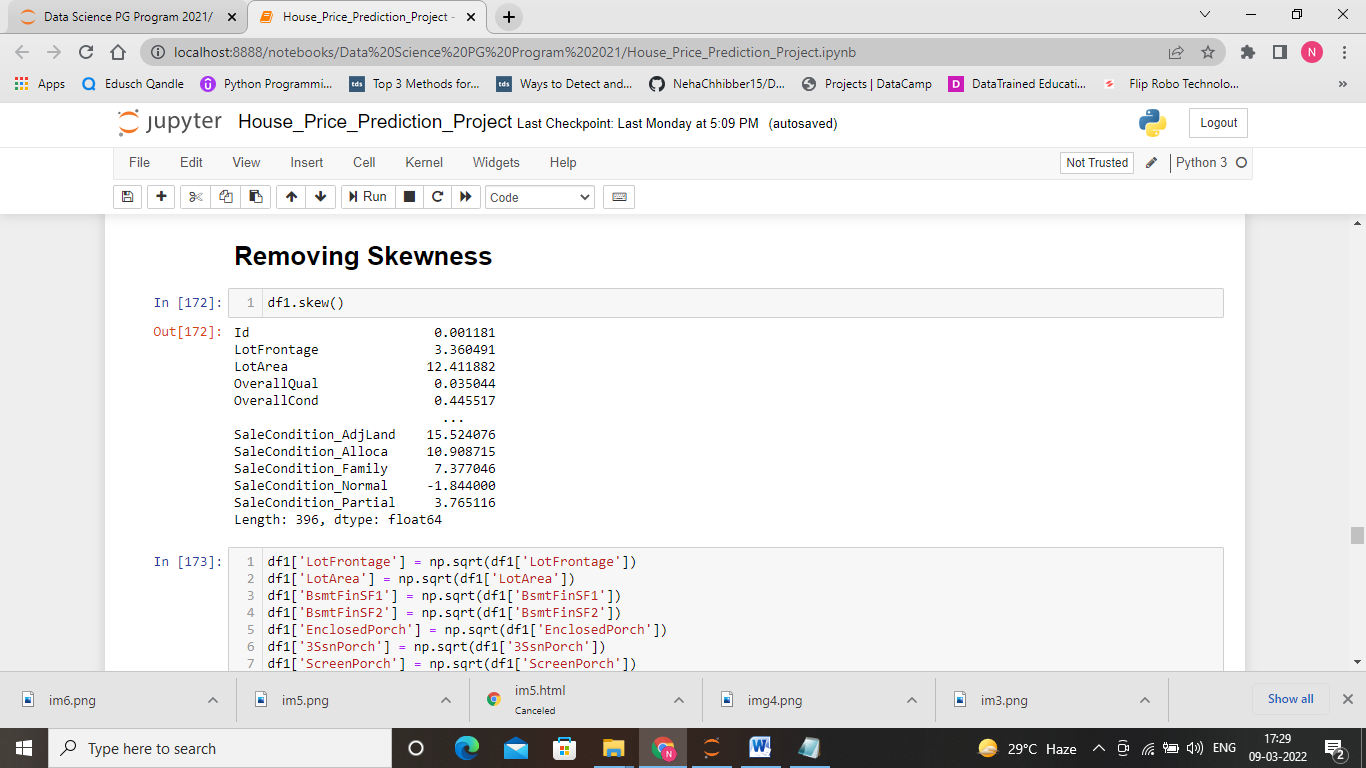


**One Hot Encoding for Features**



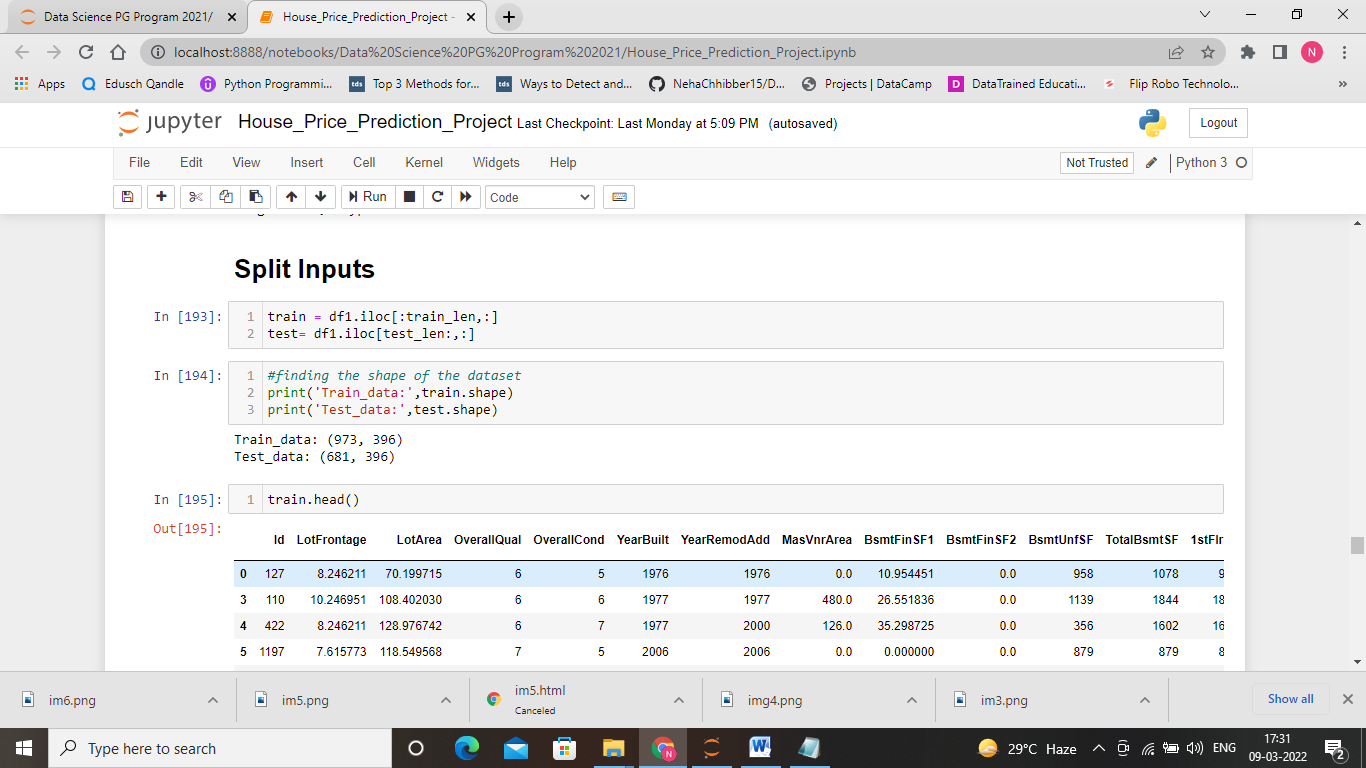
**Removing Skewness**

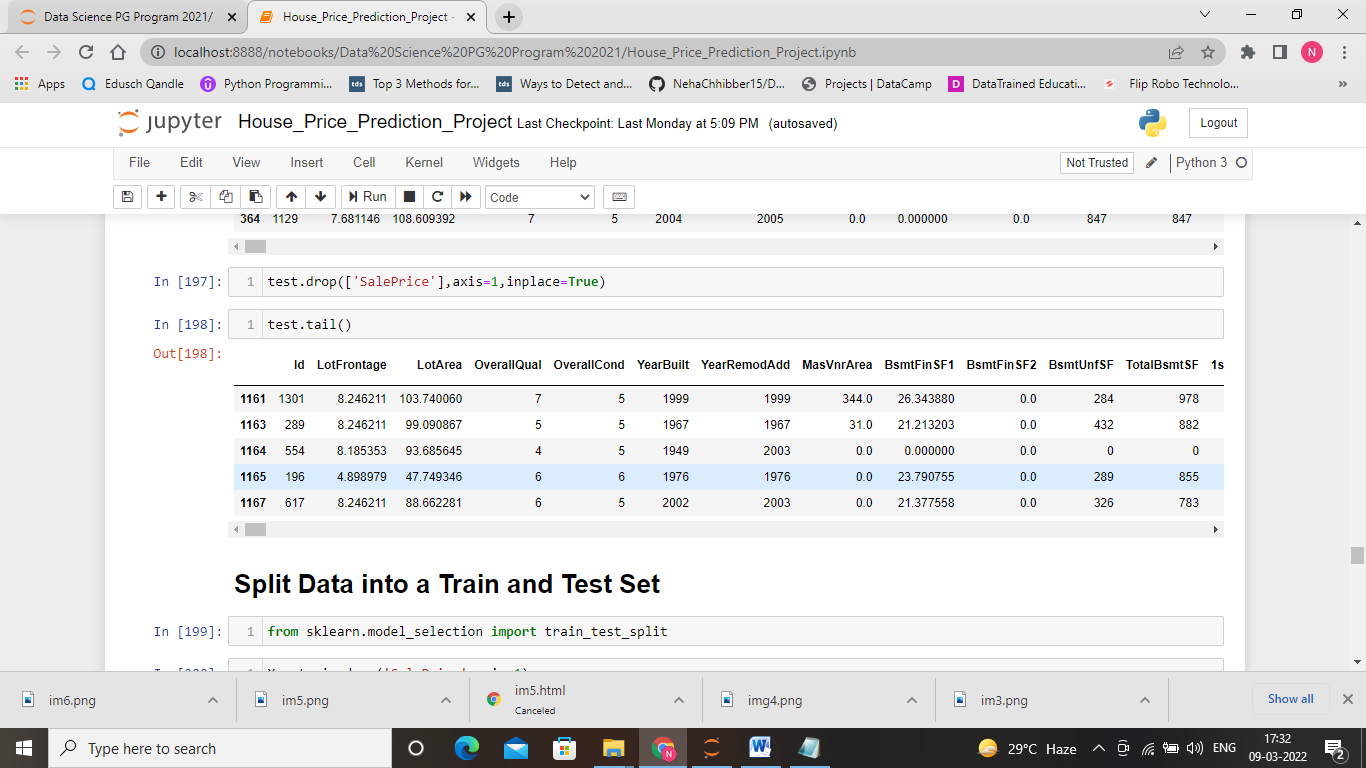
We need to remove the skewness so that modle should not be biased.



**Spliting Inputs**

Again we need to separate train and test data.





**Hardware and Software Requirements and Tools Used**

Here, we have used jupyter notebook to run the codes. Various libraries were used such as pandas, numpy, matplotlib, seaborn, scikit learn.

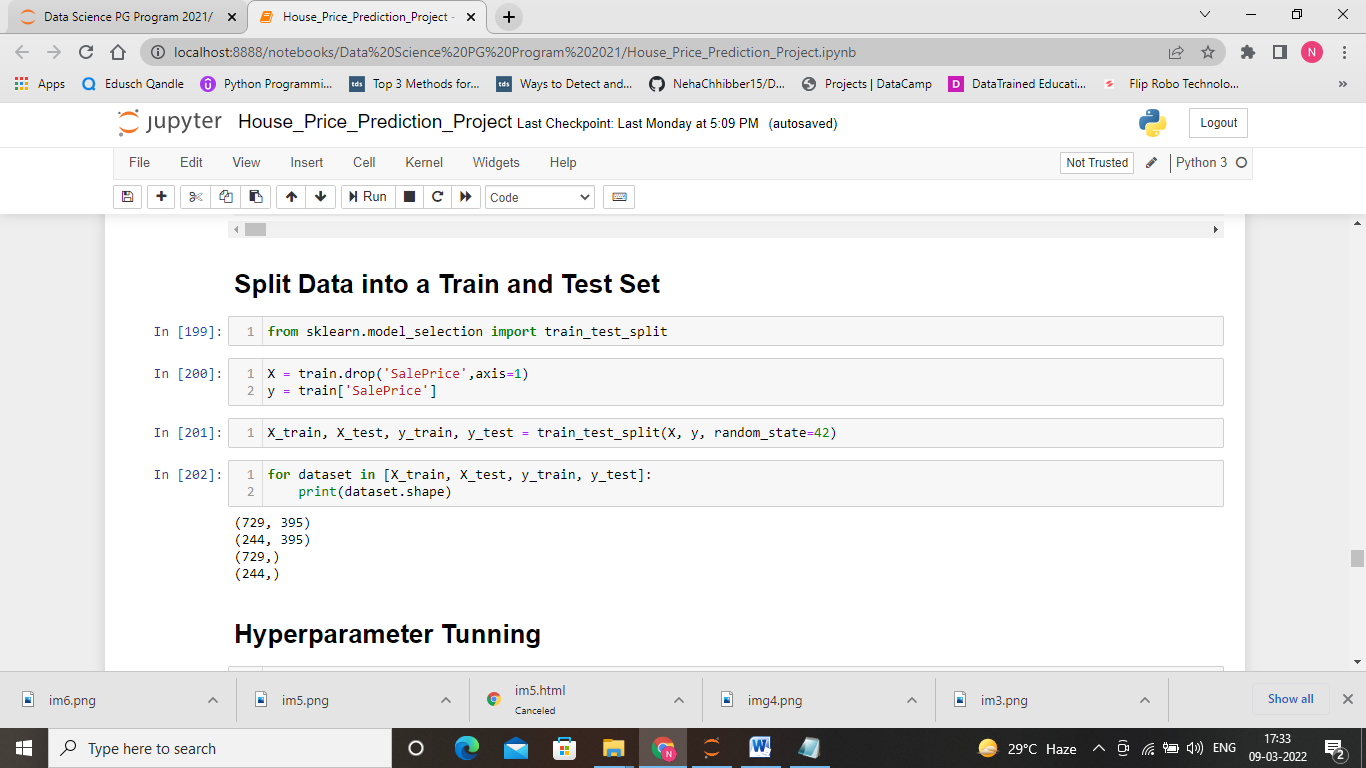
**Model/s Development and Evaluation**

**Identification of possible problem-solving approaches (methods)**

In this second section of the project, we will develop the tools and techniques necessary for a model to make a prediction. Being able to make accurate evaluations of each model’s performance through the use of these tools and techniques helps to reinforce greatly the confidence in the predictions.

**Training and Testing**

For this section we are taking the dataset and split the data into training and testing subsets. Typically, the data is also shuffled into a random order when creating the training and testing subsets to remove any bias in the ordering of the dataset.



You may ask now:What is the benefit to splitting a dataset into some ratio of training and testing subsets for a learning algorithm?

It is useful to evaluate our model once it is trained. We want to know if it has learned properly from a training split of the data. There can be 3 different situations:

1) The model didn´t learn well on the data, and can’t predict even the outcomes of the training set, this is called underfitting and it is caused because a high bias.

2) The model learn too well the training data, up to the point that it memorized it and is not able to generalize on new data, this is called overfitting, it is caused because high variance.

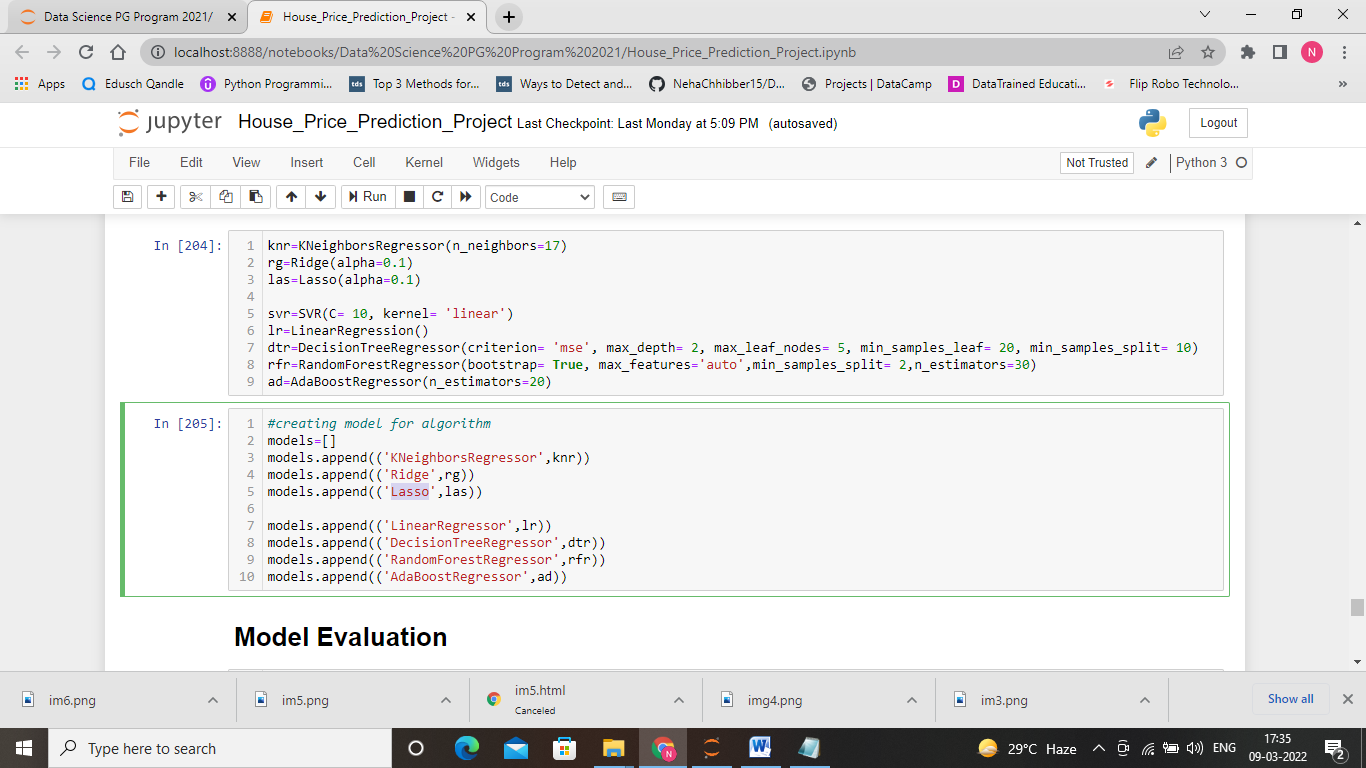
3) The model just had the right balance between bias and variance, it learned well and is able predict correctly the outcomes on new data.

**Testing of Identified Approaches (Algorithms)**

Listing down all the algorithms used for the training and testing.

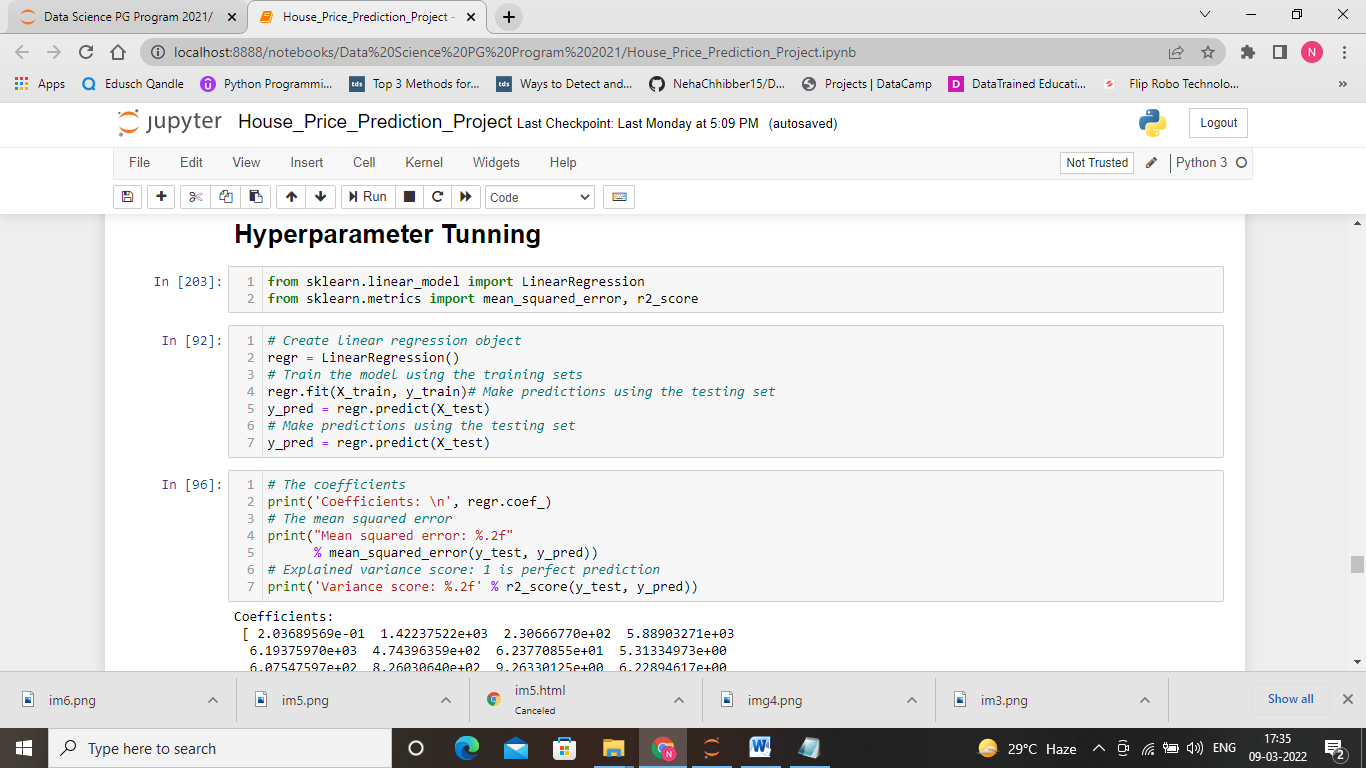
1. KNeighbors Regressor
2. Decision Tree Regressor
3. Random Forest Regressor
4. AdaBoost Regressor
5. Ridge
6. Lasso

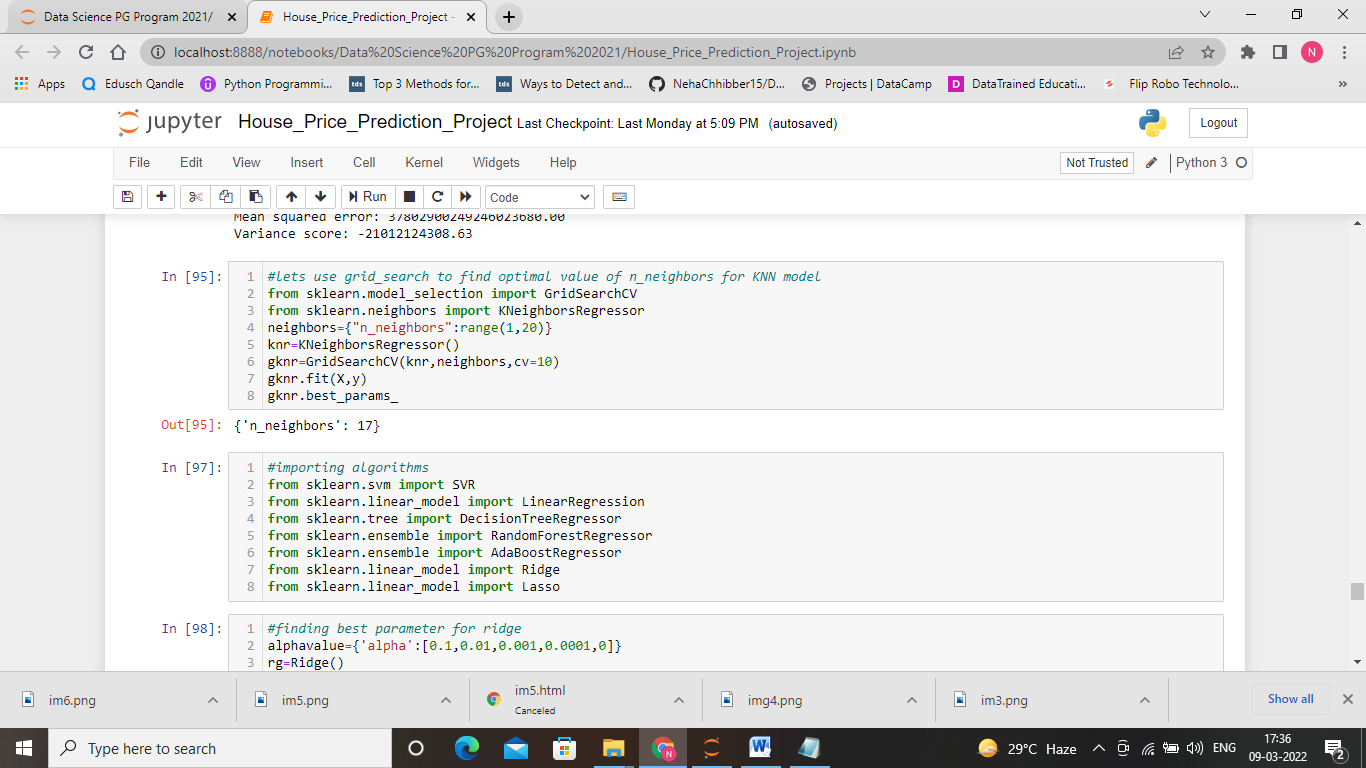
Here, we have installed all the algorithms from Sklearn library.

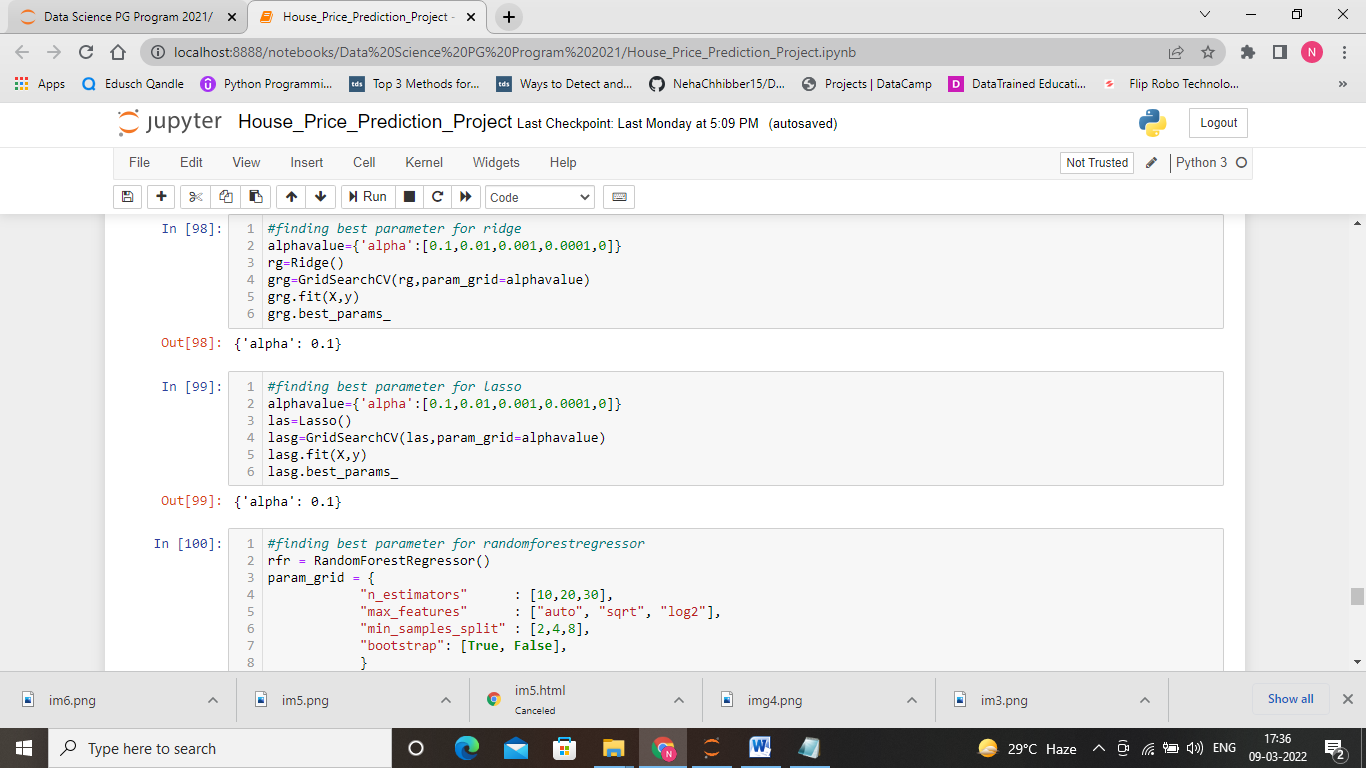


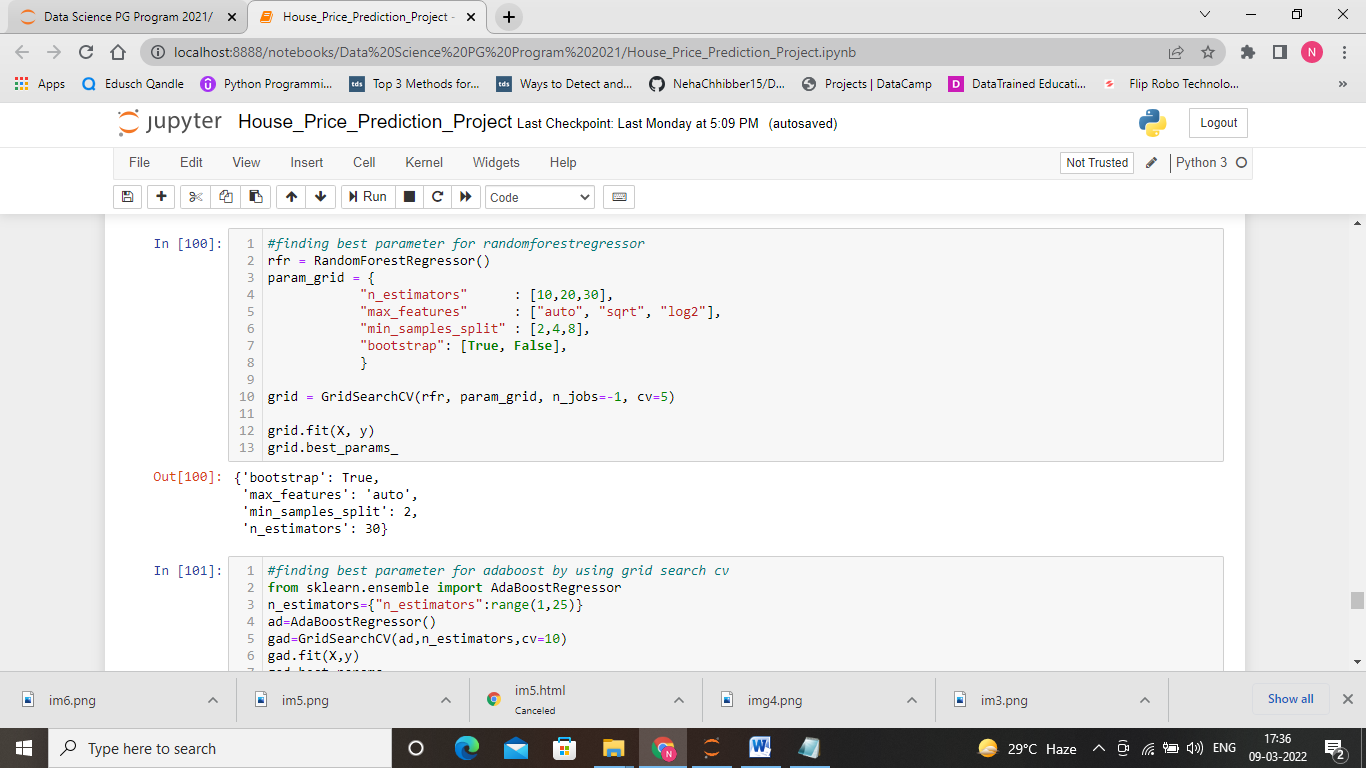
In this third section of the project, we’ll take a look at several models’ learning and testing performances on various subsets of training data.

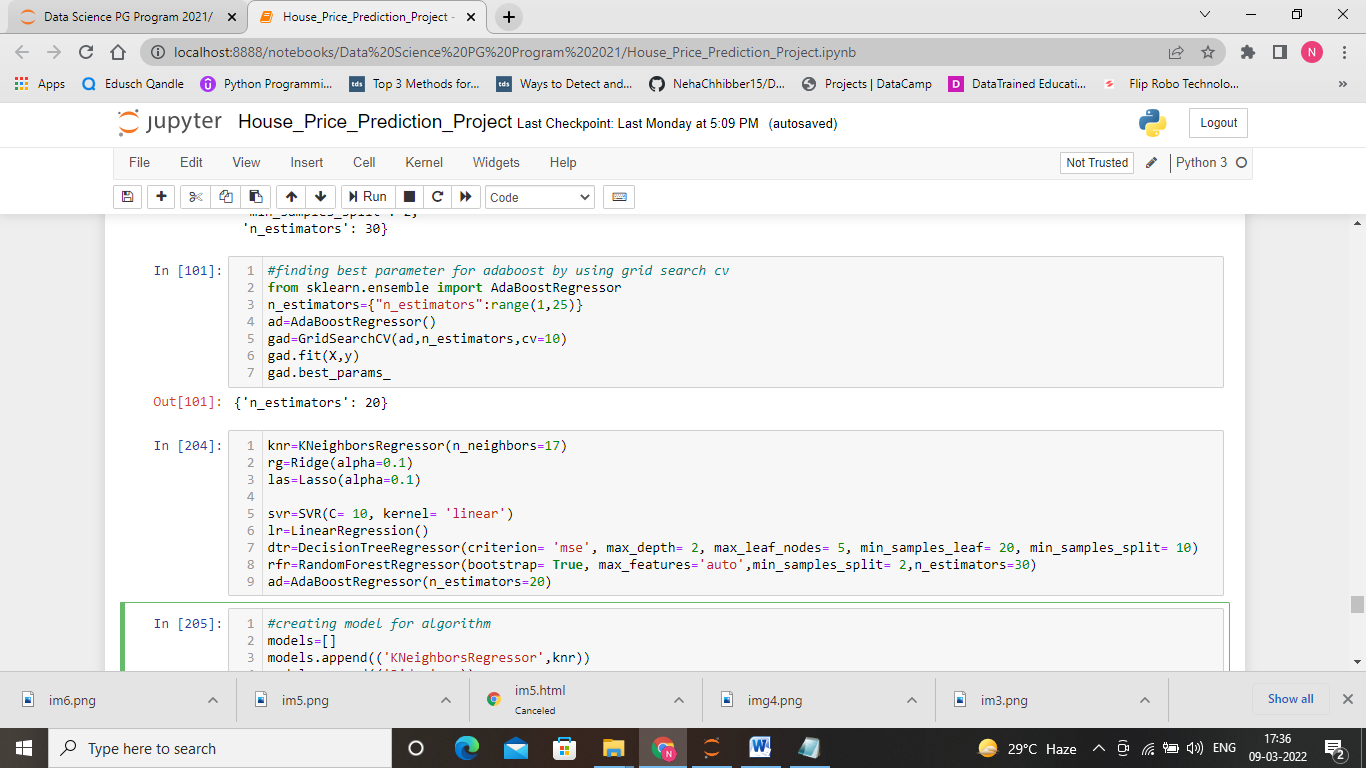
Before that we have to find the best parameters by doing hyperparameter tunning for each model.





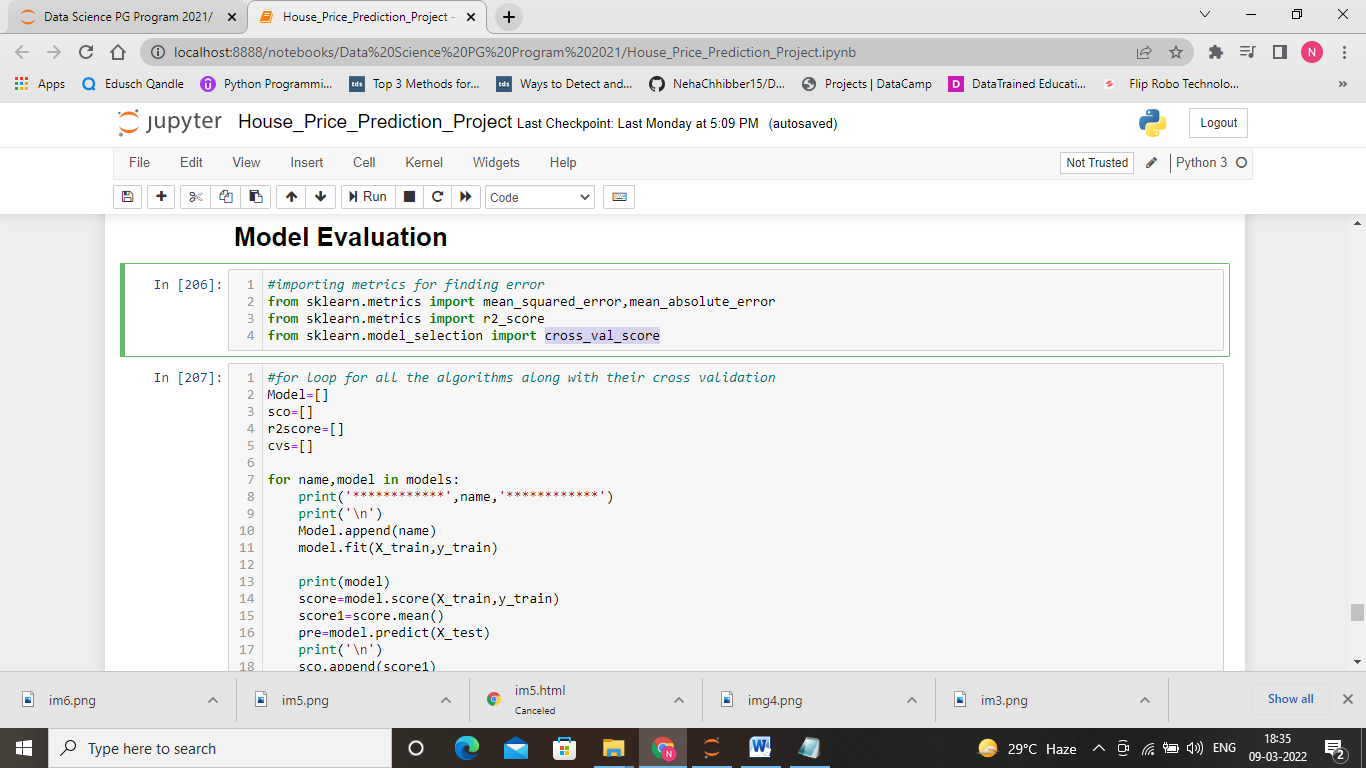


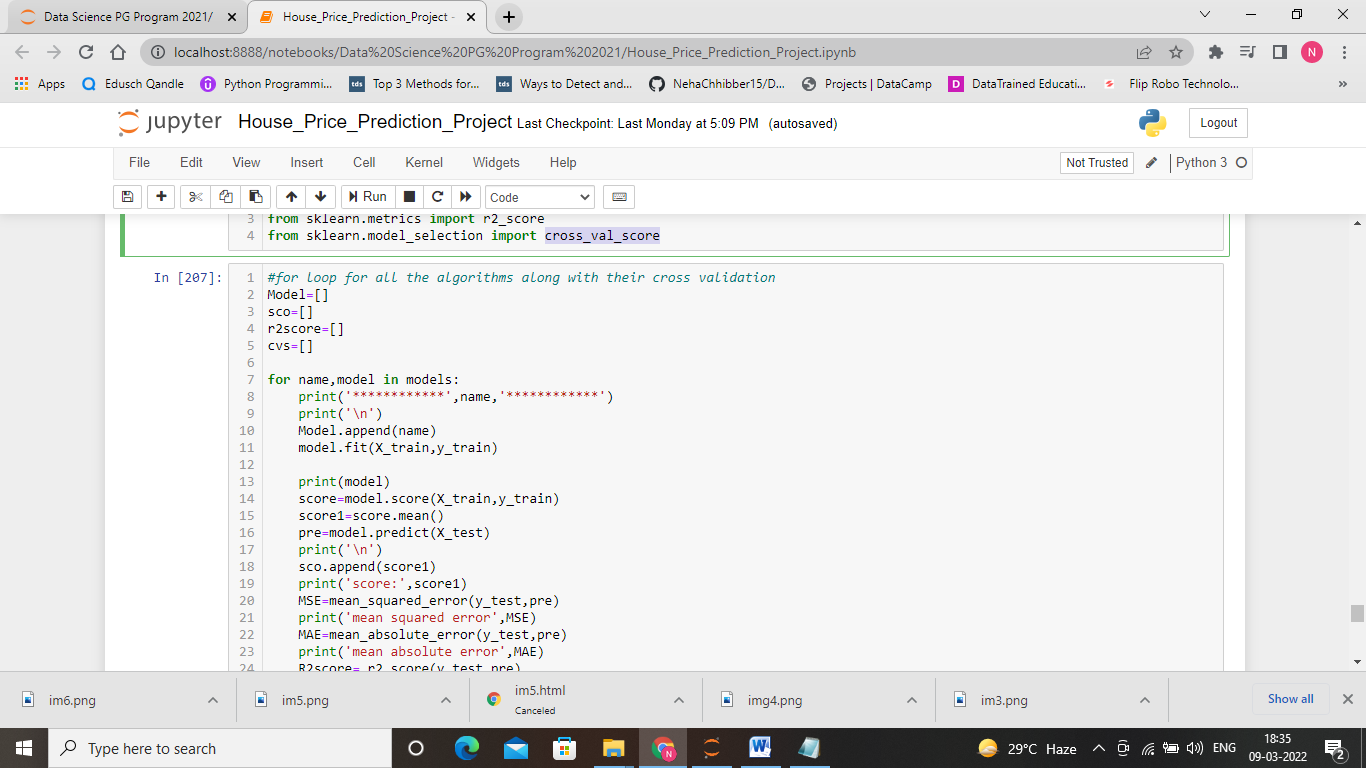


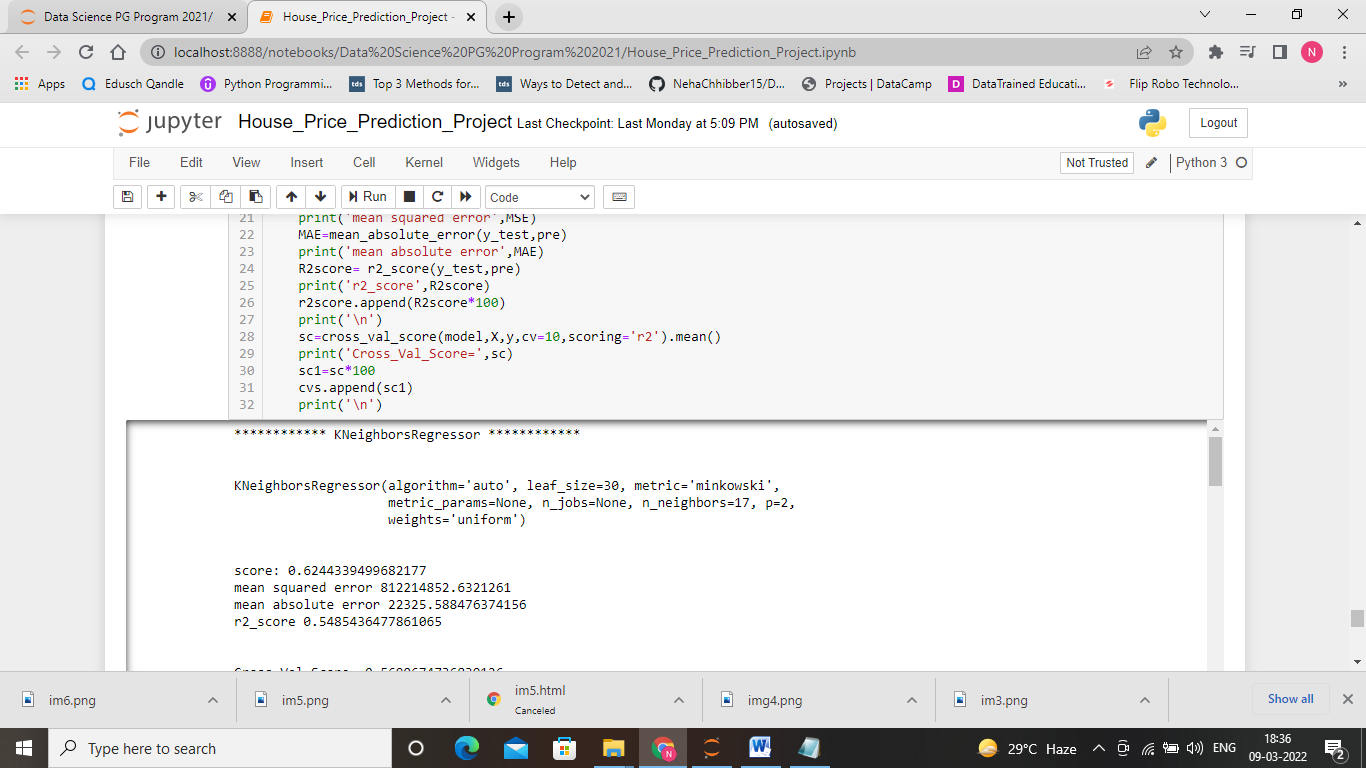


Let’s do the model evaluation on various parameters such as mean\_squared\_error, mean\_absolute\_error, r2\_score, cross\_val\_score.

**Model Evaluation**

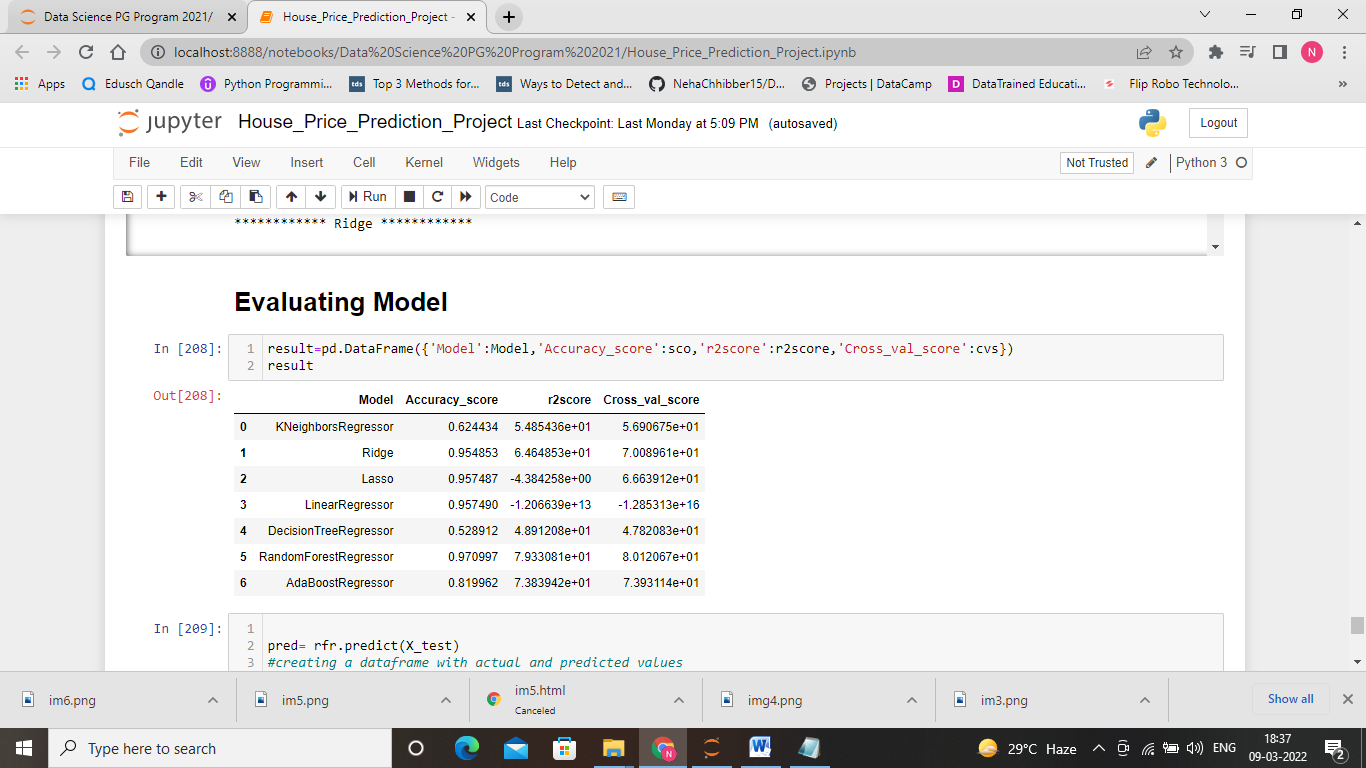






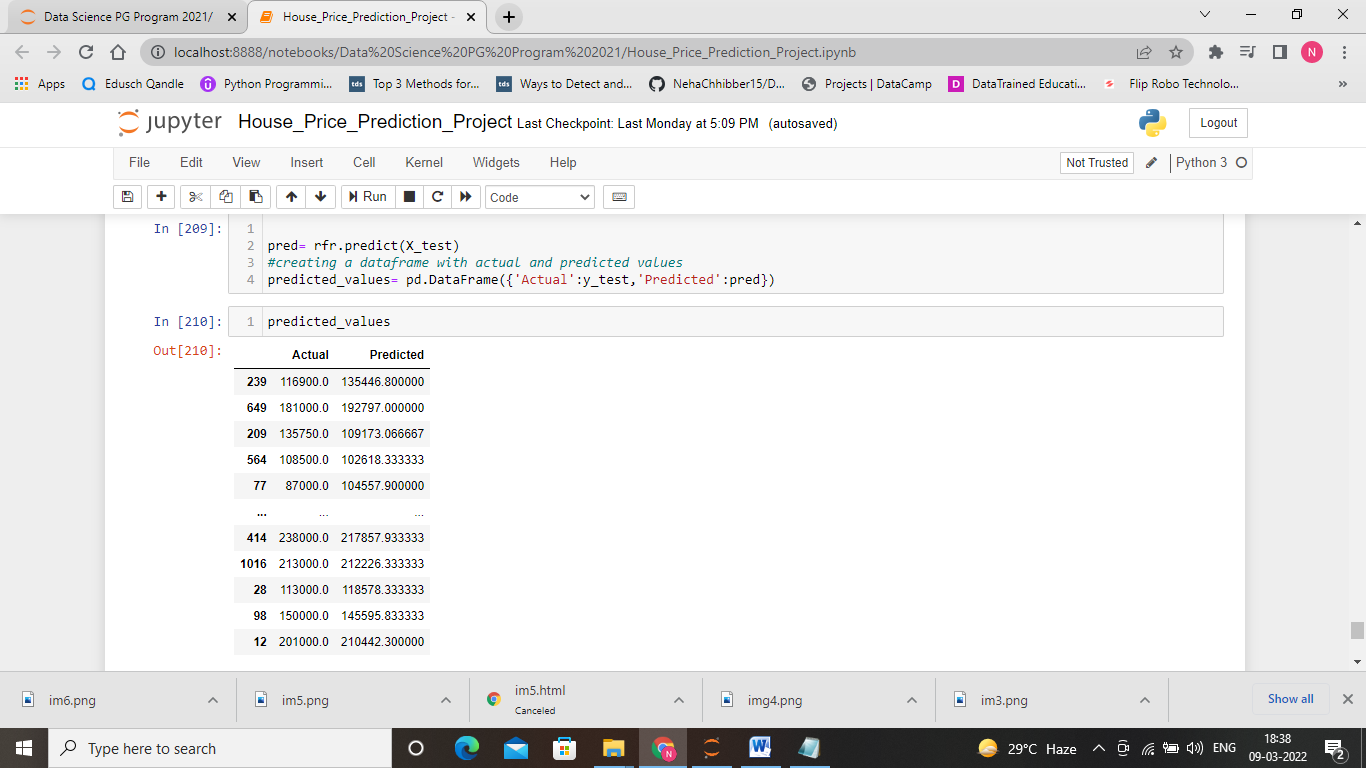
**Evaluating Model**

Let us compare the models and select the model with maximum accuracy, crossvalidation score.



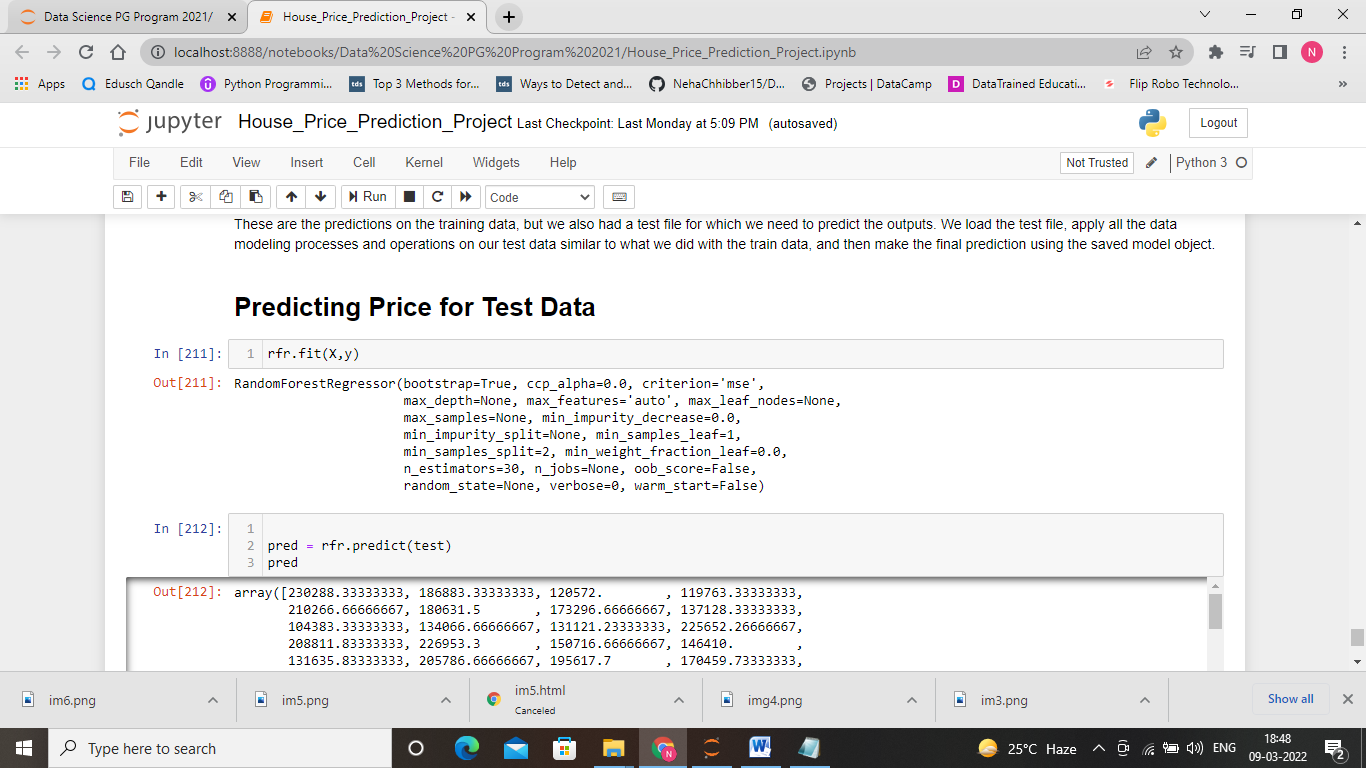
**Interpretation of the Results**

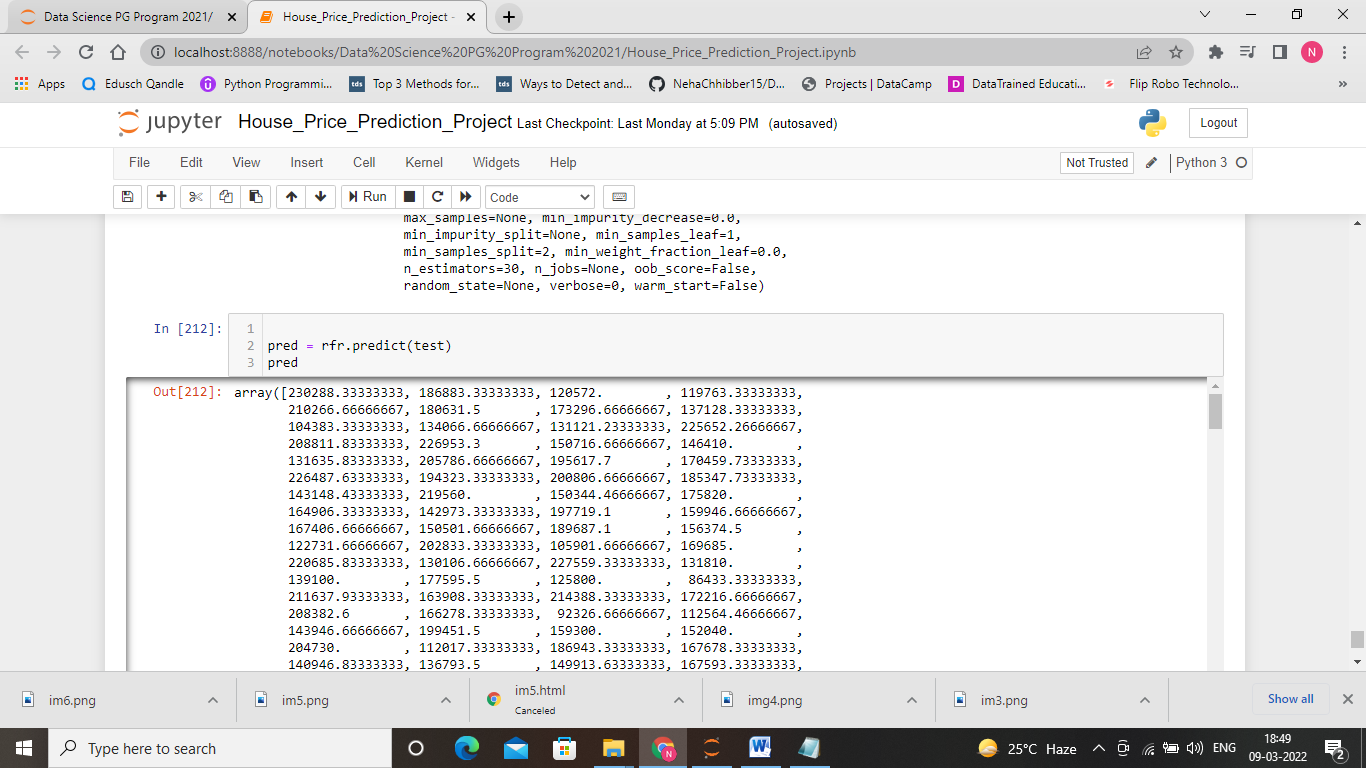
Here, we can see that random forest classifier gives best results amount all algorithms i.e., Accuracy\_score of approx. 97%, class\_val score of 80%.



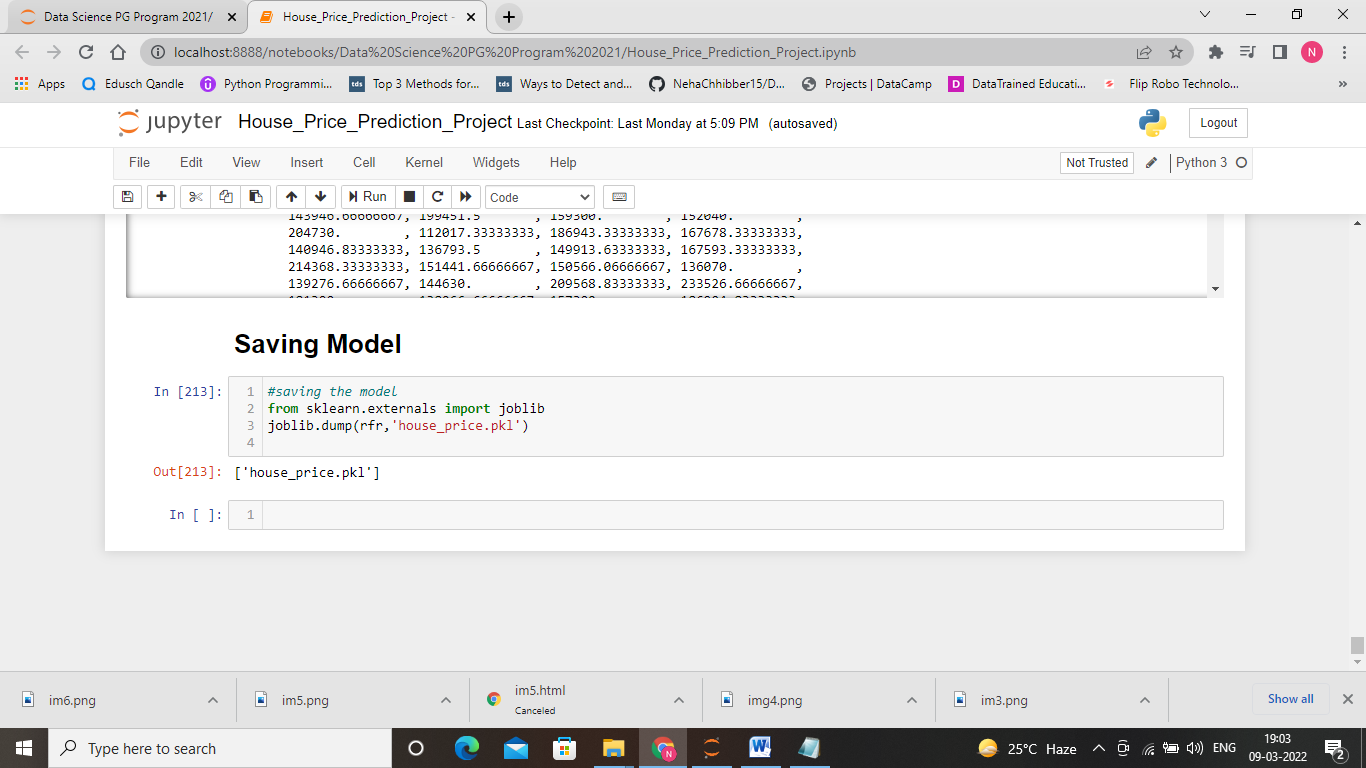
**Predicting Price for Test Data**

Let us predict price for test data set.





**Saving Model**



**CONCLUSION**

* Key Findings and Conclusions of the Study

Throughout this article we made a machine learning classification project and find out that random forest regressor is the most appropriate model among all above listed algorithms. In terms of visualization, we have seen that the given data has a lot of outlier, so to settle this we have removed outliers and scale the data by using standard scaler. We have observed that there are few of the factors which are very important for defining sale price of a house i.e. number of rooms, full or half baths, furnishing and area of garage, rooftop material used, neighbouring area, number of fire places available in a house, building year. All these factor are important for deciding the sales price of the house.