

Housing Price Prediction Project

Submitted by:

Jayashri Dandare

**ACKNOWLEDGMENT**

I would like to thank Flip Robo Technologies for providing me with the opportunity to work on this project from which I have learned a lot. I am also grateful to Miss Khushboo Garg for her constant guidance and support.

Some of the reference sources are as follows:

* Internet
* Medium.com
* StackOverflow

*TABLE OF CONTENTS*

[**ACKNOWLEDGMENT** 2](#_Toc71114959)

[**INTRODUCTION** 4](#_Toc71114960)

[BUSINESS PROBLEM FRAMING 4](#_Toc71114961)

[CONCEPTUAL BACKGROUND OF THE DOMAIN PROBLEM 4](#_Toc71114962)

[REVIEW OF LITERATURE 5](#_Toc71114963)

[MOTIVATION FOR THE PROBLEM UNDERTAKEN 5](#_Toc71114964)

[**ANALYTICAL PROBLEM FRAMING** 6](#_Toc71114965)

[MATHEMATICAL/ ANALYTICAL MODELING OF THE PROBLEM 6](#_Toc71114966)

[DATA SOURCES AND THEIR FORMATS 7](#_Toc71114967)

[DATA PREPROCESSING DONE 14](#_Toc71114968)

[DATA INPUTS- LOGIC- OUTPUT RELATIONSHIPS 19](#_Toc71114969)

[HARDWARE AND SOFTWARE REQUIREMENTS AND TOOLS USED 20](#_Toc71114970)

[**MODEL/S DEVELOPMENT AND EVALUATION** 26](#_Toc71114971)

[IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES (METHODS) 26](#_Toc71114972)

[TESTING OF IDENTIFIED APPROACHES (ALGORITHMS) 26](#_Toc71114973)

[RUN AND EVALUATE SELECTED MODELS 27](#_Toc71114974)

[KEY METRICS FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION 28](#_Toc71114975)

[VISUALIZATIONS 29](#_Toc71114976)

[INTERPRETATION OF THE RESULTS 59](#_Toc71114977)

[**CONCLUSION** 60](#_Toc71114978)

[KEY FINDINGS AND CONCLUSIONS OF THE STUDY 60](#_Toc71114979)

[LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE 61](#_Toc71114980)

[LIMITATIONS OF THIS WORK AND SCOPE FOR FUTURE WORK 62](#_Toc71114981)

**INTRODUCTION**

* **Business Problem Framing:**

This is a real estate problem where a US based housing company named Surprise Housing has decided to invest in Australian Market. Their agenda is to buy houses in Australia at prices below their actual value in the market and sell them at high prices to gain profit. To do this company uses data analytics to decide in which property they must invest.

Company has collected the data of previously sold houses in Australia and with the help of this data they want to know to the value of prospective properties to decide whether it will suitable to invest in the properties or not.

To know the value of properties company has provided data to us to do data analysis and to extract the information of attributes which are important to predict the price of the houses. They want a machine learning model which can predict the price of houses and also the significance of each important attribute in house prediction i.e, how and to what intensity each variable impacts the price of the house.

* **Conceptual Background of the Domain Problem:**

In real estate the value of property usually increases with time as seen in many countries. One of the causes for this is due to rising population.

The value of property also depends on the proximity of the property, its size its neighbourhood and audience for which the property is subjected to be sold. For example if audience is mainly concerned of commercial purpose. Then the property which is located in densely populated area will be sold very fast and at high prices compared to the one located at remote place.

The company is looking at prospective properties to buy houses to enter the market. We 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.

* **Review of Literature:**

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.

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.

Living cost in Australia for one person: $2,835 per month. Average living expenses for a couple: $4,118 per month. Average monthly living expenses for a family of 4: $5,378. Australia currently has the 16th highest cost of living in the world, with the USA and UK well behind at 21st and 33rd place respectively. Sydney and Melbourne are popular choices for expats moving to Australia.

House pricing in some of the top Australian cities:-

Sydney - median house price A$1,142,212

Adelaide- median house price A$542,947

Hobbart (smaller city)- median house price A$530,570.

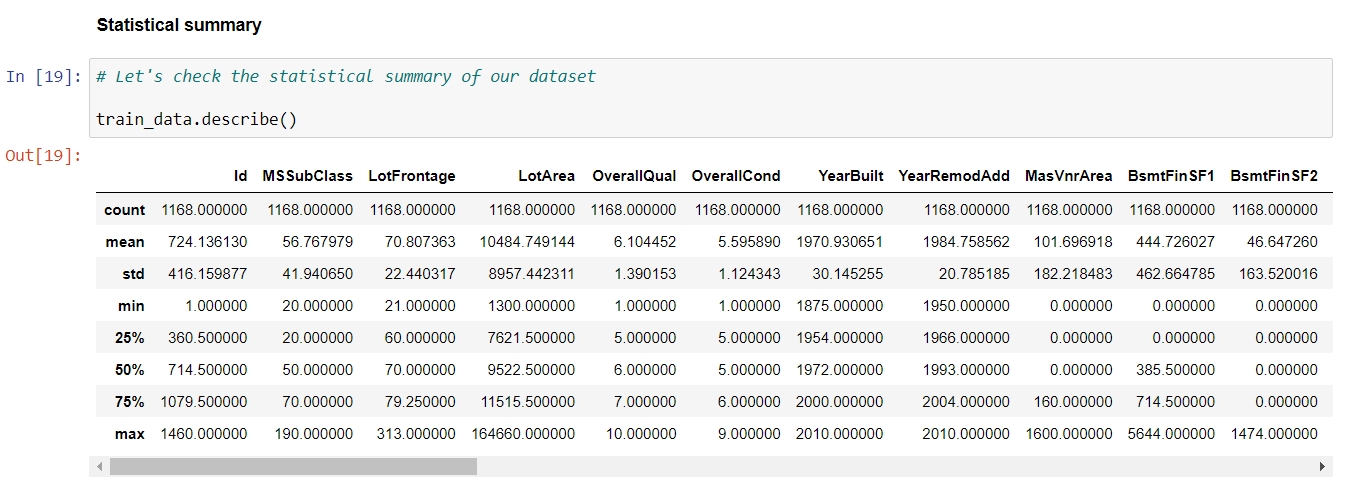
* **Motivation for the Problem Undertaken:**

To understand real world problems where Machine Learning and Data Analysis can be applied to help organizations in various domains to make better decisions with the help of which they can gain profit or can be escaped from any loss which otherwise could be possible without the study of data .One of such domain is Real Estate.

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.

**Analytical Problem Framing**

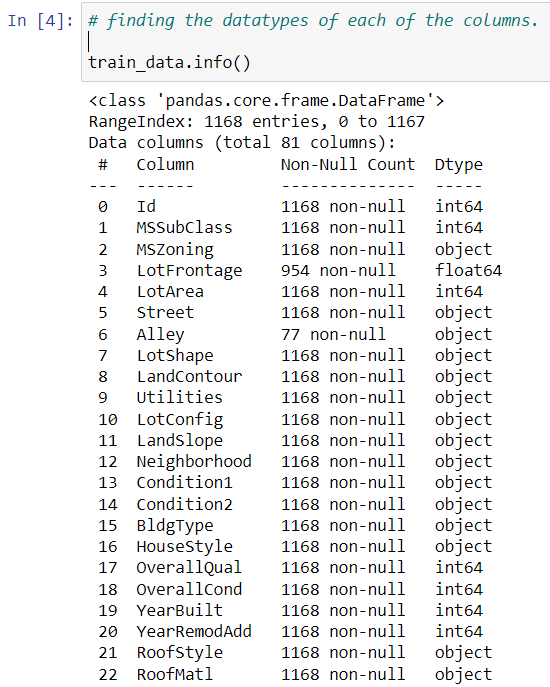
* **Mathematical/ Analytical Modeling of the Problem:**

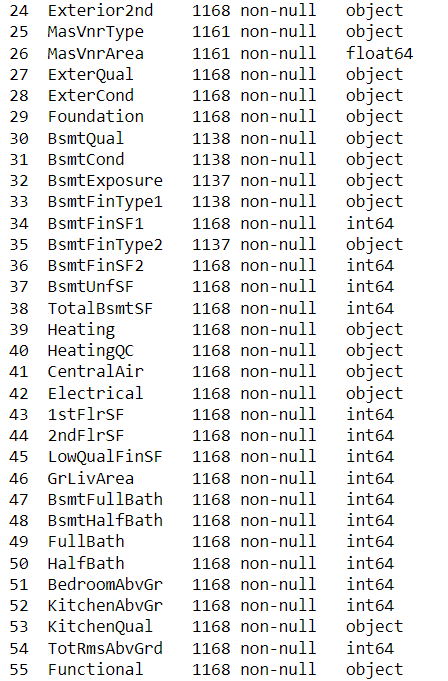
In this project we have performed various mathematical and statistical analysis such as we checked description or statistical summary of the data using describe, checked correlation using corr and also visualized it using heatmap.

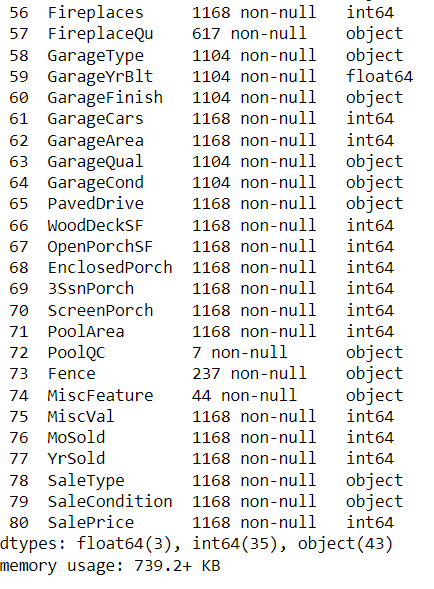
**We observed that :**

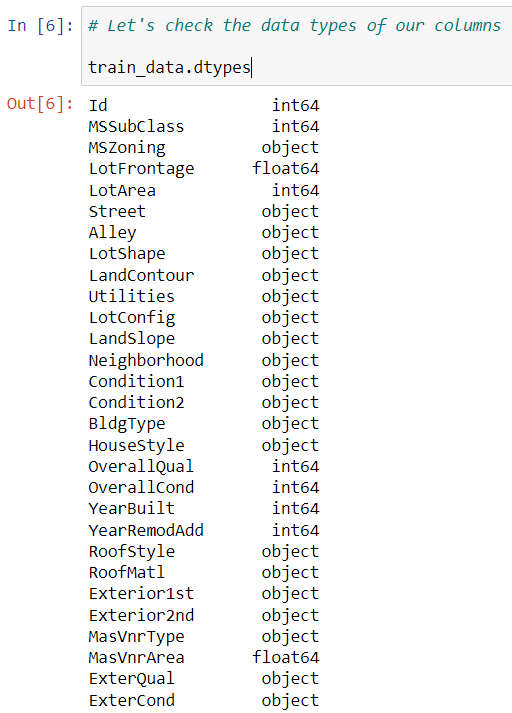
* Maximum standard deviation of 8957.44 is observed in LotArea column.
* Maximum SalePrice of a house observed is 755000 and minimum is 34900.
* In the columns Id, MSSubclass, LotArea, MasVnrArea, BsmtFinSF1, BsmtFinSF2, BsmtUnfsF, TotalBsmtSF, 1stFlrSF, 2ndFlrSF, LowQualFinSF, GrLivArea, BsmtFullBath, HalfBath, TotRmsAbvGrd, WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea, Miscval, salePrice mean is considerably greater than median so the columns are positively skewed.
* In the columns FullBath, BedroomAbvGr, Fireplaces, Garagecars, GarageArea, YrSold Median is greater than mean so the columns are negatively skewed.
* In the columns Id, MSSubClass, LotFrontage, LotArea, MasVnrArea, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, 1stFlrSF, 2ndFlrSF, LowQualFinSF, GrLivArea, BsmtHalfBath, BedroomAbvGr, ToRmsAbvGrd, GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea, MiscVal, SalePrice there is considerable difference between the 75 percentile and maximum so outliers are present.

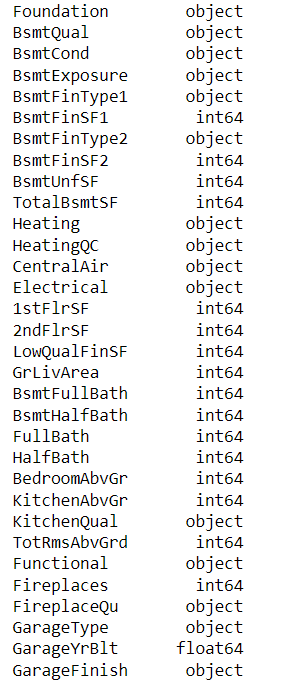
* **Data Sources and their formats:**
* The variable features of this problem statement are as :
* MSSubClass: Identifies the type of dwelling involved in the sale
* MSZoning: Identifies the general zoning classification of the sale
* LotFrontage: Linear feet of street connected to property
* LotArea: Lot size in square feet
* Street: Type of road access to property
* Alley: Type of alley access to property
* LotShape: General shape of property
* LandContour: Flatness of the property
* Utilities: Type of utilities available
* LotConfig: Lot configuration
* LandSlope: Slope of property
* Neighborhood: Physical locations within Ames city limits
* Condition1: Proximity to various conditions
* Condition2: Proximity to various conditions (if more than one is present)
* BldgType: Type of dwelling
* HouseStyle: Style of dwelling
* OverallQual: Rates the overall material and finish of the house
* OverallCond: Rates the overall condition of the house
* YearBuilt: Original construction date
* YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)
* RoofStyle: Type of roof
* RoofMatl: Roof material
* Exterior1st: Exterior covering on house
* Exterior2nd: Exterior covering on house (if more than one material)
* MasVnrType: Masonry veneer type
* MasVnrArea: Masonry veneer area in square feet
* ExterQual: Evaluates the quality of the material on the exterior
* ExterCond: Evaluates the present condition of the material on the exterior
* Foundation: Type of foundation
* BsmtQual: Evaluates the height of the basement
* BsmtCond: Evaluates the general condition of the basement
* BsmtExposure: Refers to walkout or garden level walls
* BsmtFinType1: Rating of basement finished area
* BsmtFinSF1: Type 1 finished square feet
* BsmtFinType2: Rating of basement finished area (if multiple types)
* BsmtFinSF2: Type 2 finished square feet
* BsmtUnfSF: Unfinished square feet of basement area
* TotalBsmtSF: Total square feet of basement area
* Heating: Type of heating
* HeatingQC: Heating quality and condition
* CentralAir: Central air conditioning
* Electrical: Electrical system
* 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
* TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
* Functional: Home functionality (Assume typical unless deductions are warranted)
* Fireplaces: Number of fireplaces
* FireplaceQu: Fireplace quality
* GarageType: Garage location
* GarageYrBlt: Year garage was built
* GarageFinish: Interior finish of the garage
* GarageCars: Size of garage in car capacity
* GarageArea: Size of garage in square feet
* GarageQual: Garage quality
* GarageCond: Garage condition
* PavedDrive: Paved driveway
* 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
* Fence: Fence quality
* MiscFeature: Miscellaneous feature not covered in other categories
* MiscVal: $Value of miscellaneous feature
* MoSold: Month Sold (MM)
* YrSold: Year Sold (YYYY)
* SaleType: Type of sale
* SaleCondition: Condition of sale

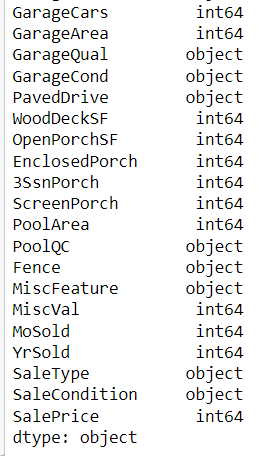






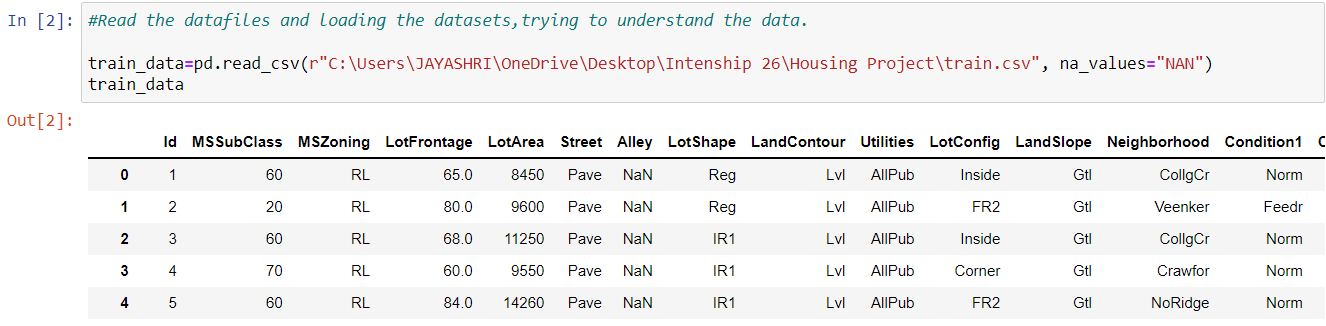




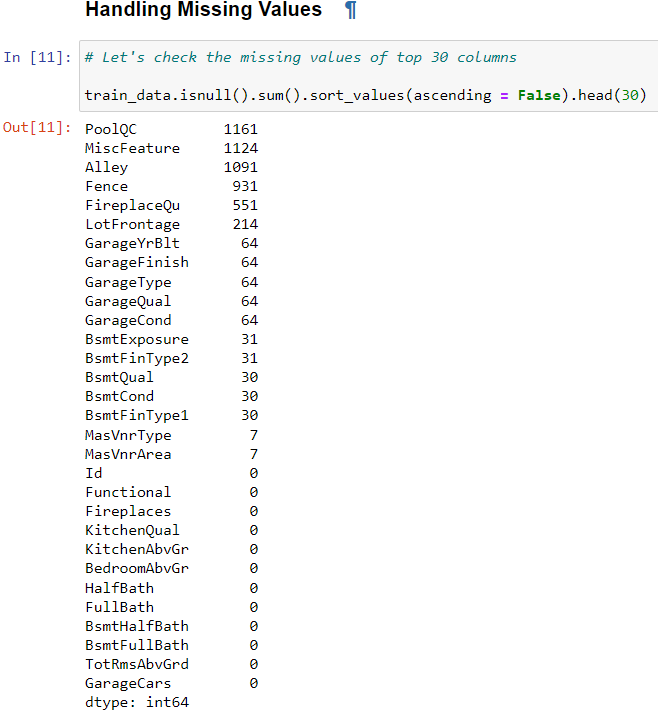


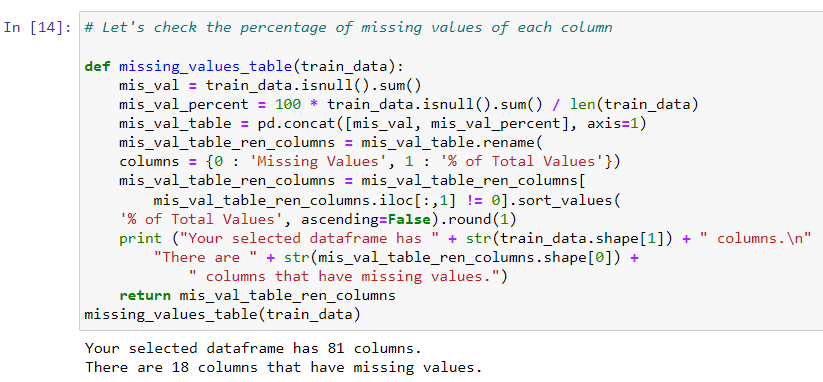
* **Data Preprocessing Done:**

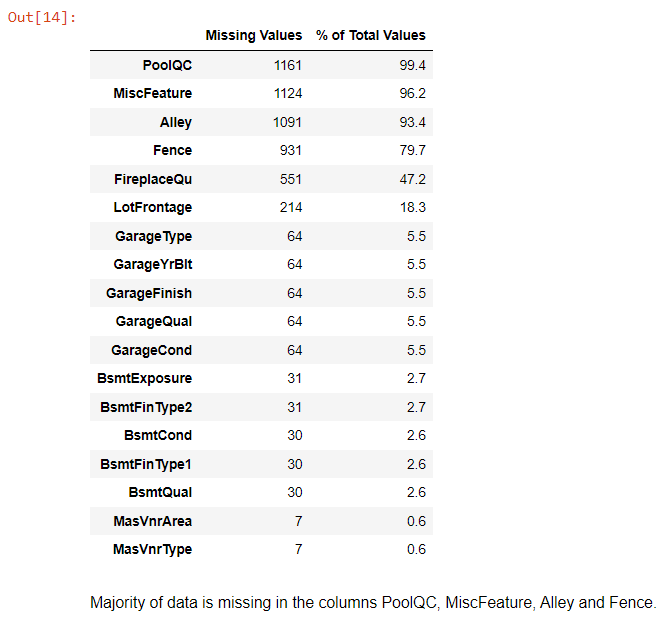
After loading all the required libraries, we loaded the data into our jupyter notebook.

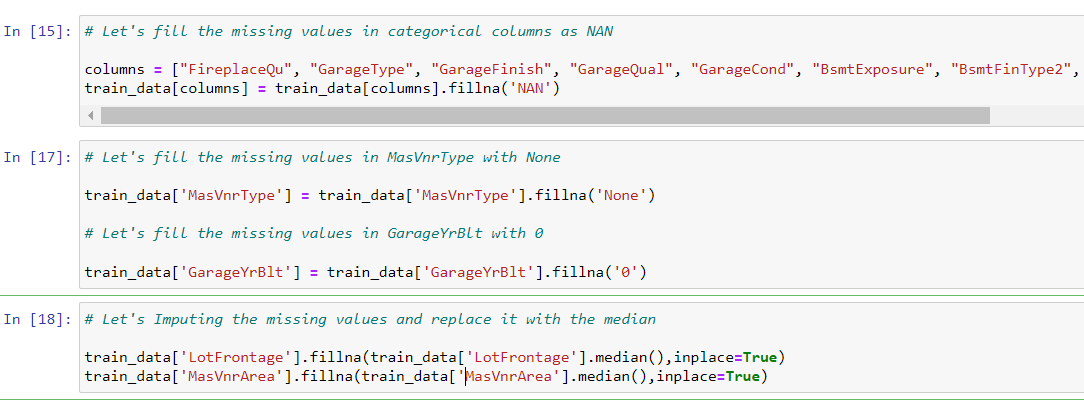


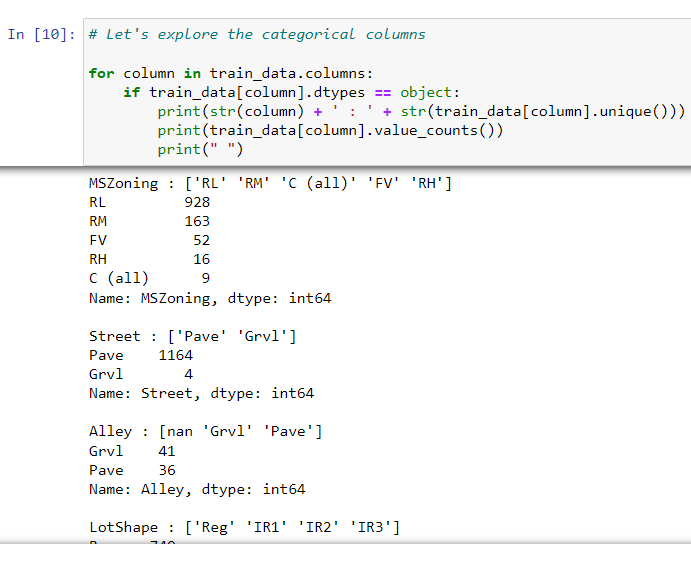
Feature Engineering has been used for cleaning of the data. Some unused columns have been deleted and even some columns have been bifurcated which was used in the prediction. We first done data cleaning. We first looked percentage of values missing in columns then we imputed missing values.





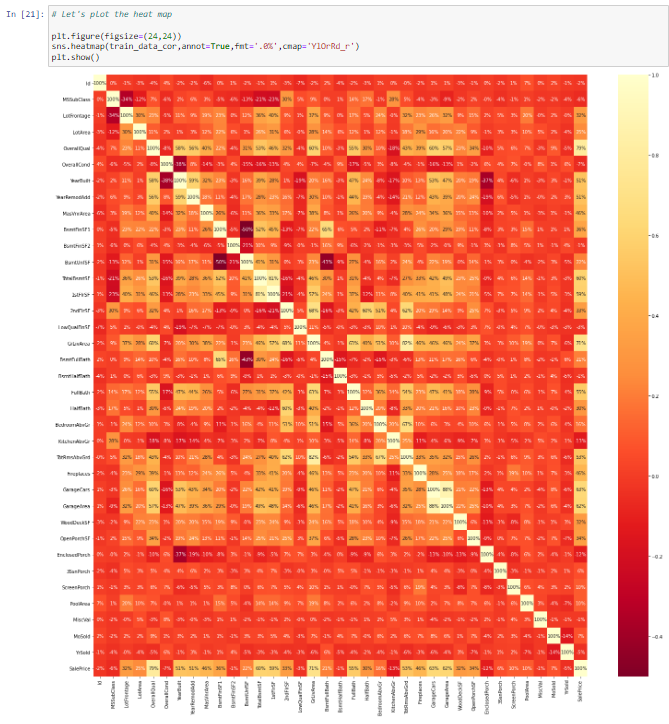






We observed that there is only one unique value present in Utilities so will be dropping this column. Then we encoded all the categorical columns into numerical columns using dummy variables

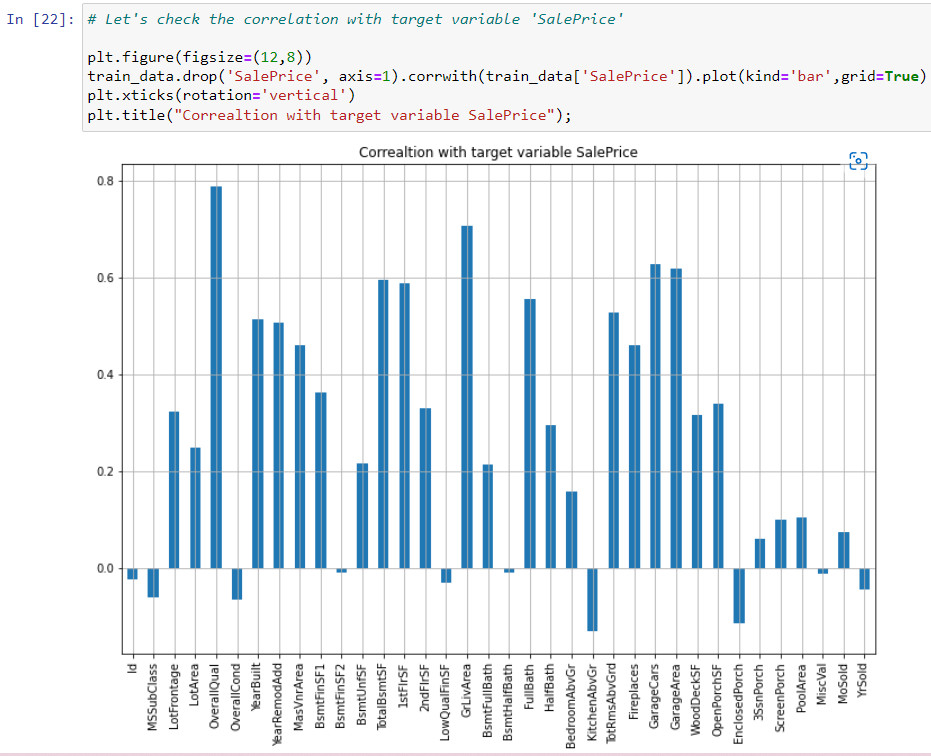
Then we checked the correlation with the help of heatmap.



#### We observed that:

* SalePrice is highly positively correlated with the columns OverallQual, YearBuilt, YearRemodAdd, TotalBsmtSF, 1stFlrSF, GrLivArea, FullBath, TotRmsAbvGrd, GarageCars, GarageArea.
* SalePrice is negatively correlated with OverallCond, KitchenAbvGr, Encloseporch, YrSold.
* We observe multicollinearity in between columns so we will be using Principal Component Analysis(PCA).
* No correlation has been observed between the column Id and other columns so we will be dropping this column.
* **Data Inputs- Logic- Output Relationships:**

Here we check the correlation between all our feature variables with target variable label



#### We observed that:

* The column OverallQual is most positively correlated with SalePrice.
* The column KitchenAbvGrd is most negatively correlated with SalePrice.

Set of assumptions related to the problem under consideration

By looking into the target variable label we assumed that it was a Regression type of problem.

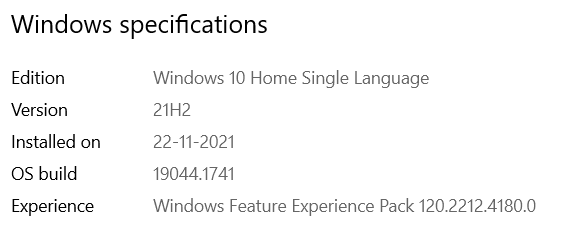
We observed multicollinearity in between columns so we assumed that

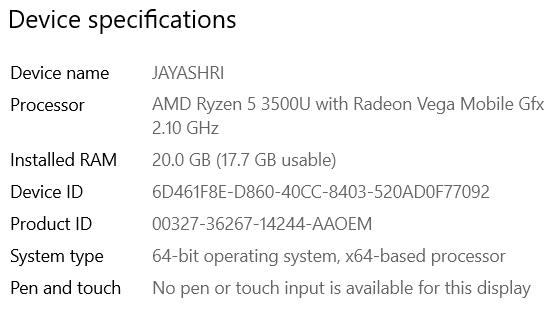
we will be using Principal Component Analysis (PCA).

We also observed that only one single unique value was present in Utilities column so we assumed that we will be dropping these columns.

**Hardware and Software Requirements and Tools Used**

* **Hardware:**





* **SOFTWARE:**
* Jupyter Notebook (Anaconda 3) – Python 3.7.6
* Microsoft Excel 2010
* **LIBRARIES:**
* The tools, libraries and packages we used for accomplishing this project are pandas, numpy, matplotlib, seaborn, scipy stats, sklearn.decomposition pca, sklearn standardscaler, GridSearchCV, joblib.



***From sklearn.preprocessing import StandardScaler***

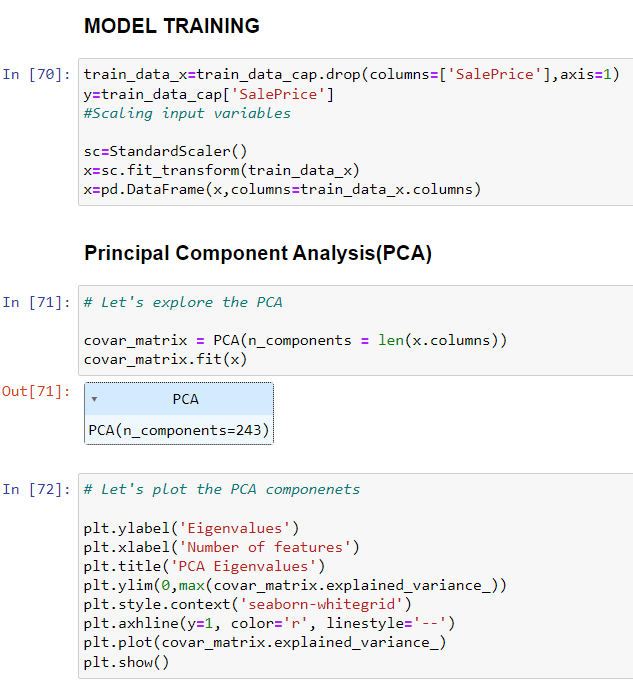
As these columns are different in scale, they are standardized to have common scale while building machine learning model. This is useful when you want to compare data that correspond to different units.

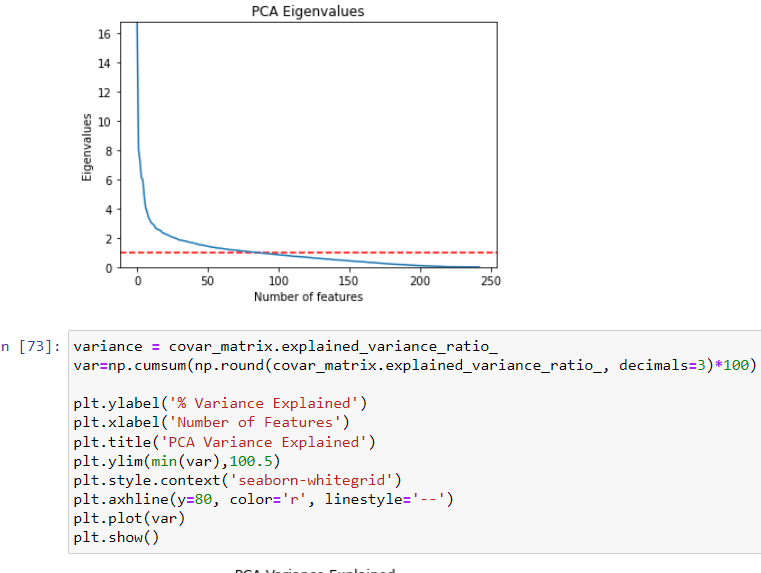
***from sklearn.preprocessing import Label Encoder***

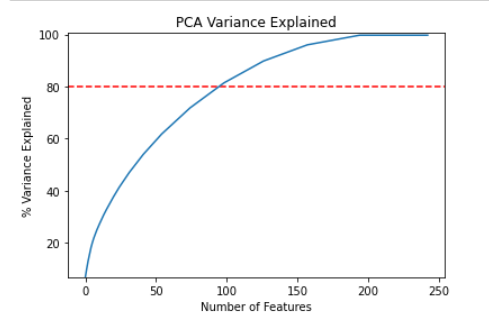
Label Encoder  and One Hot Encoder. These two encoders are parts of the SciKit Learn library in Python, and they are used to convert categorical data, or text data, into numbers, which our predictive models can better understand.

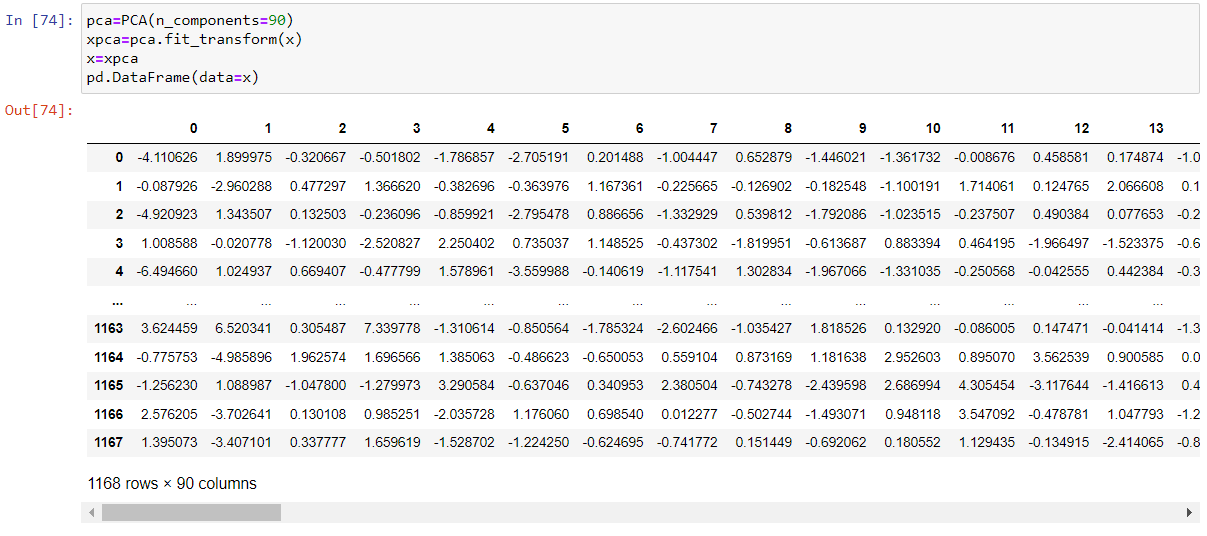
***from sklearn.model\_selection import train\_test\_split,cross\_val\_score***

* Train\_test\_split is a function in Sklearn model selection for splitting data arrays into two subsets: for training data and for testing data. With this function, you don't need to divide the dataset manually. By default, Sklearn train\_test\_split will make random partitions for the two subsets.
* Through pandas library we loaded our csv file ‘Data file’ into dataframe and performed data manipulation and analysis.
* With the help of numpy we worked with arrays.
* With the help of matplotlib and seaborn we did plot various graphs and figures and done data visualization.
* With scipy stats we treated outliers through winsorization technique.
* With sklearn.decomposition’s pca package we reduced the number of feature variables from 256 to 100 by plotting scrre plot with their Eigenvalues and chose the number of columns on the basis of their nodes.
* With sklearn’s standardscaler package we scaled all the feature variables onto single scale.









***from sklearn.linear\_model import LogisticRegression***

The library sklearn can be used to perform logistic regression in a few lines as shown using the LogisticRegression class. It also supports multiple features.

***from sklearn.tree import DecisionTreeClassifier***

Decision Tree is a white box type of ML algorithm. It shares internal decision-making logic, which is not available in the black box type of algorithms such as Neural Network. Its training time is faster compared to the neural network algorithm. The time complexity of decision trees is a function of the number of records and number of attributes in the given data.

***from sklearn.ensemble import RandomForestClassifier***

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max\_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

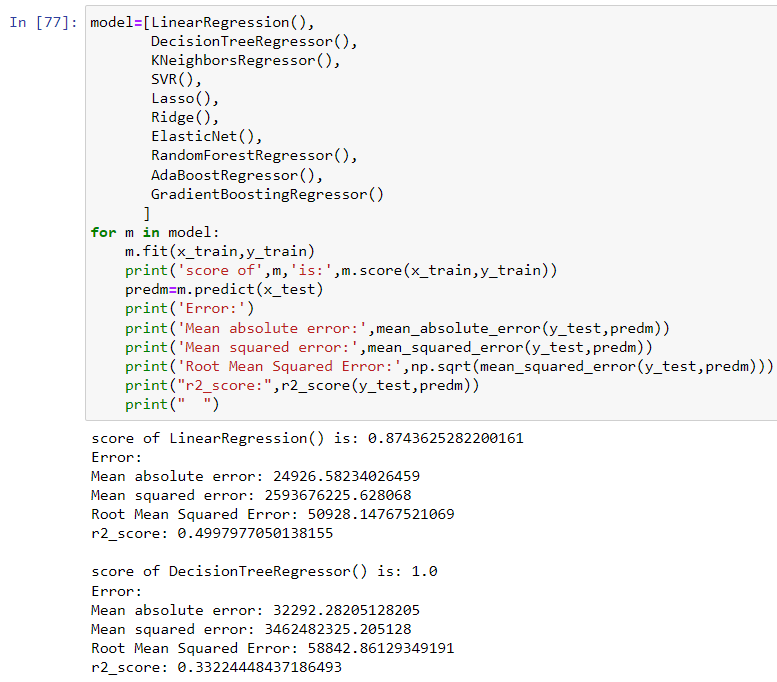
Through GridSearchCV we were able to find the right parameters for hyperparameter tuning. Through joblib we saved our model in csv format.

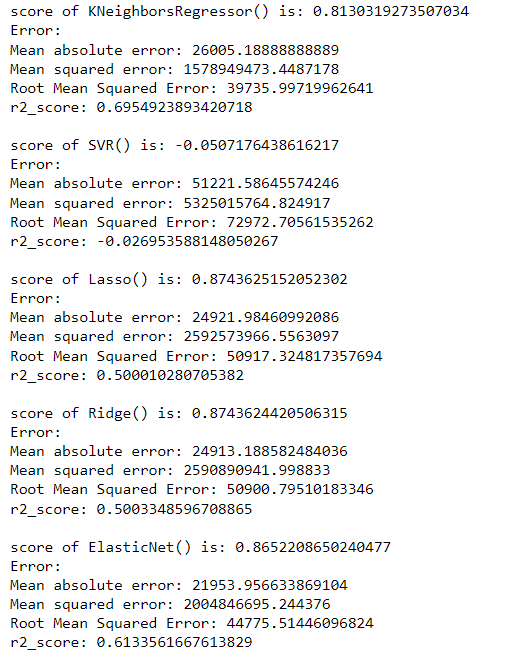
**Model/s Development and Evaluation**

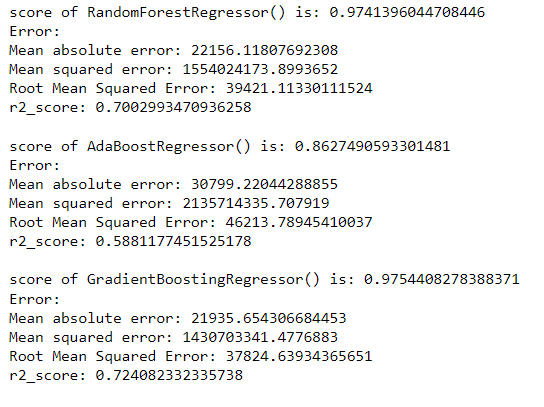
* **Identification of possible problem-solving approaches (methods):**
* We first converted all our categorical variables to numeric variables with the help of dummy variables to checkout and dropped the columns which we felt were unnecessary.
* We observed skewness in data so we tried to remove the skewness through treating outliers with winsorization technique.
* The data was improper scaled so we scaled the feature variables on a single scale using sklearn’s StandardScaler package.
* There were too many (256) feature variables in the data so we reduced it to 100 with the help of Principal Component Analysis(PCA) by plotting Eigenvalues and taking the number of nodes as our number of feature variables.
* **Testing of Identified Approaches (Algorithms):**

The algorithms we used for the training and testing are as follows:-

* Linear Regression
* Lasso
* Ridge
* Elastic Net
* SVR
* KNeighbors Regressor
* Decision Tree Regressor
* Random Forest Regressor
* Ada Boost Regressor
* Gradient Boosting Regressor
* **Run and Evaluate selected models:**



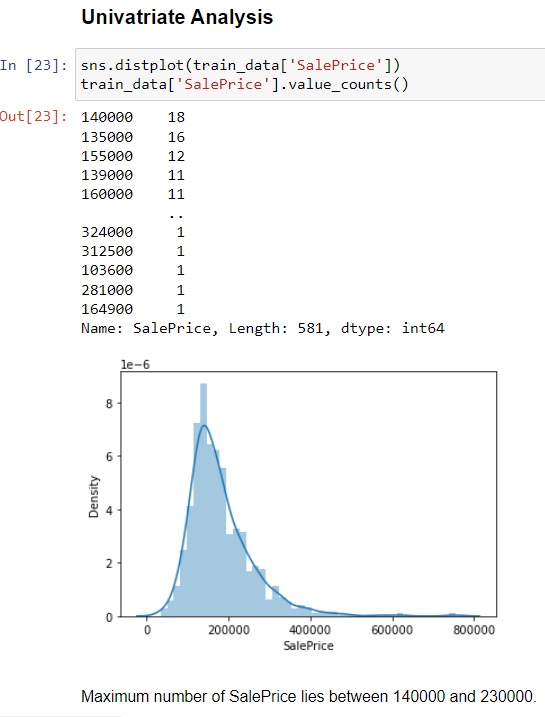




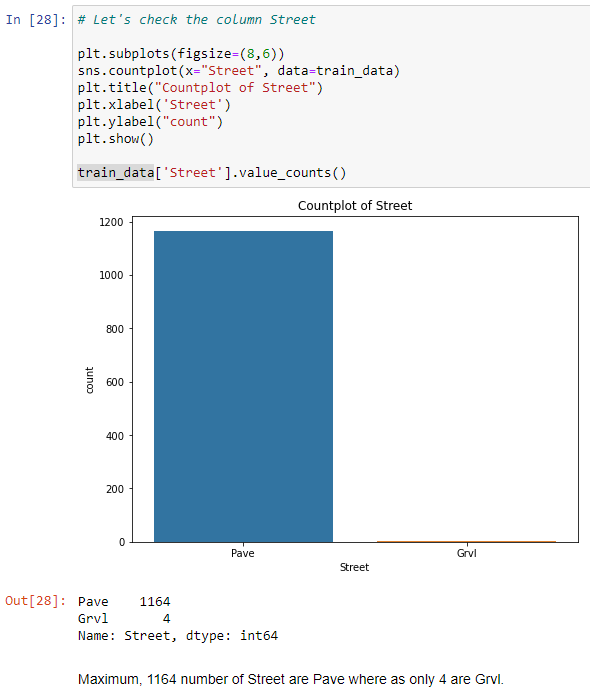
* **Key Metrics for success in solving problem under consideration:**

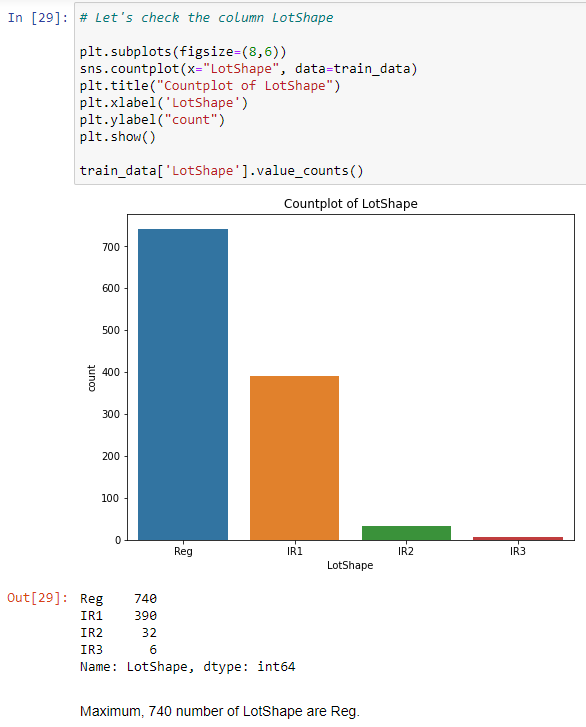
We used the metric Root Mean Squared Error by selecting the Ridge Regressor model which was giving us best(minimum) RMSE score.

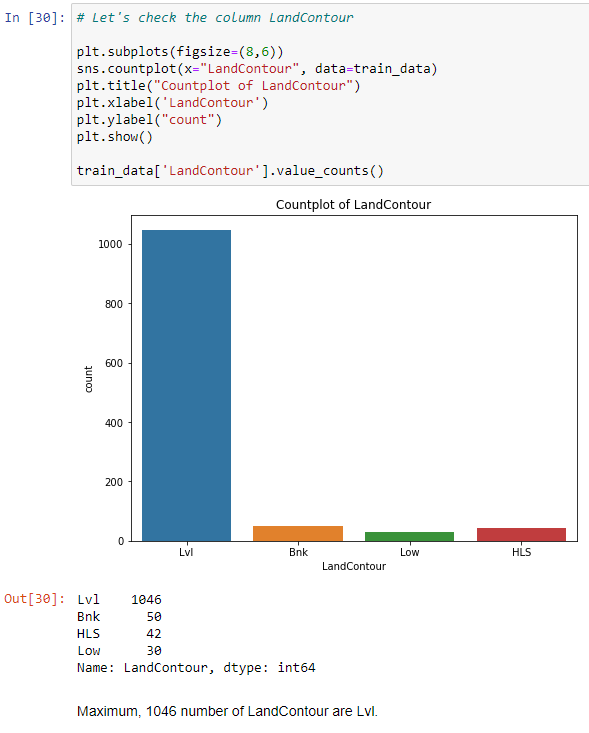
* **Visualizations:**



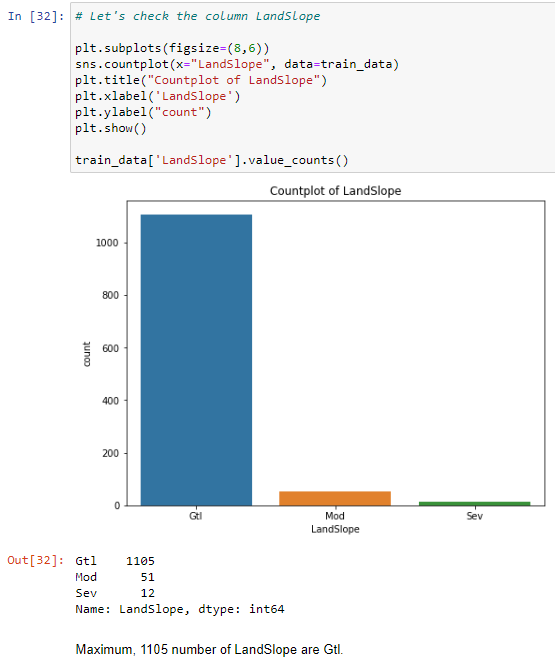


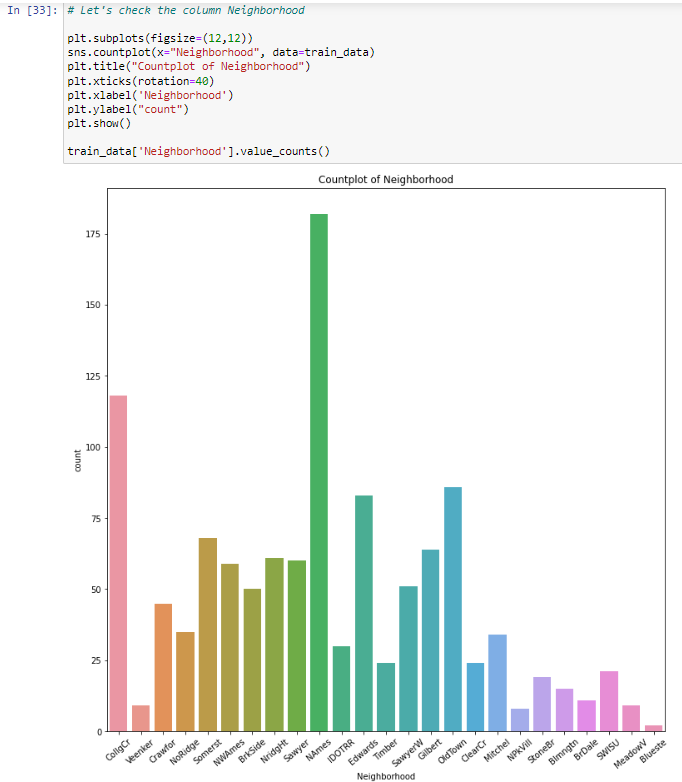




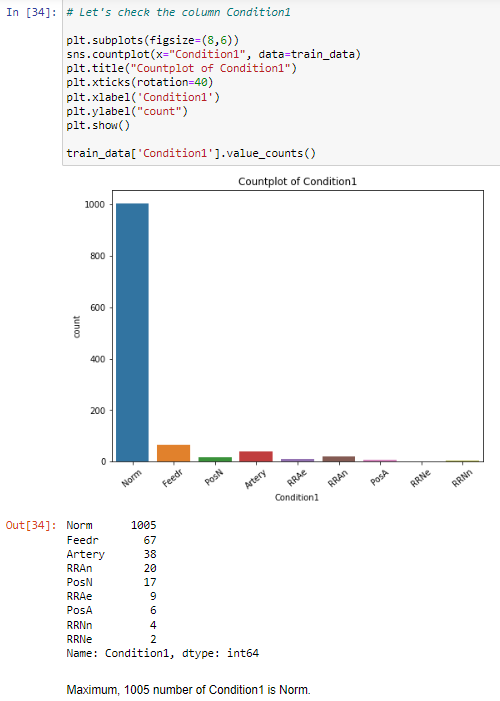


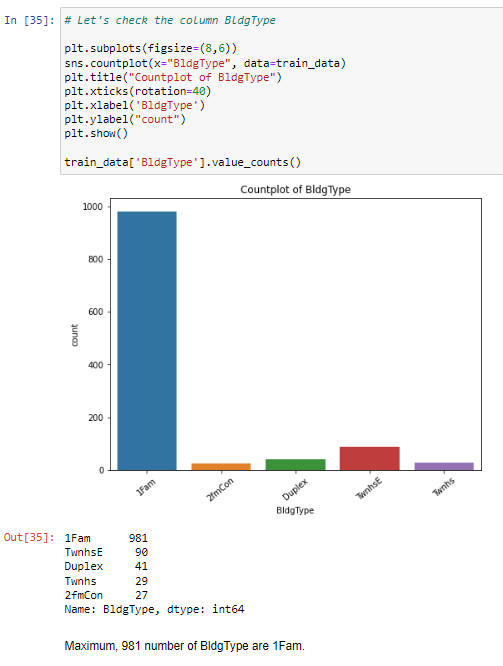


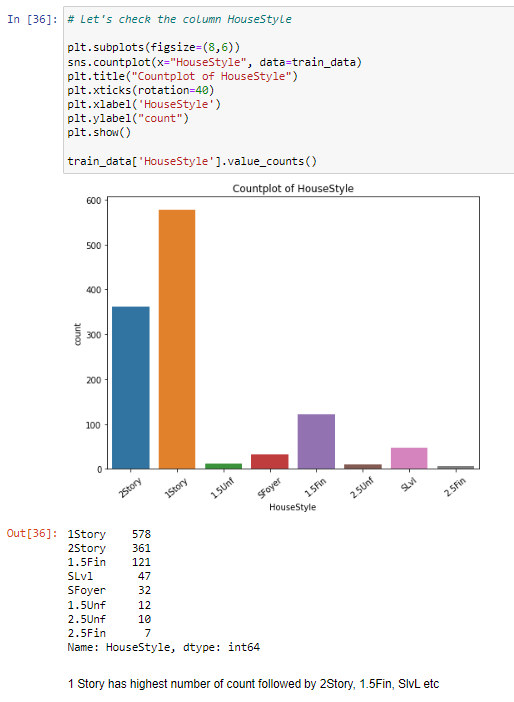


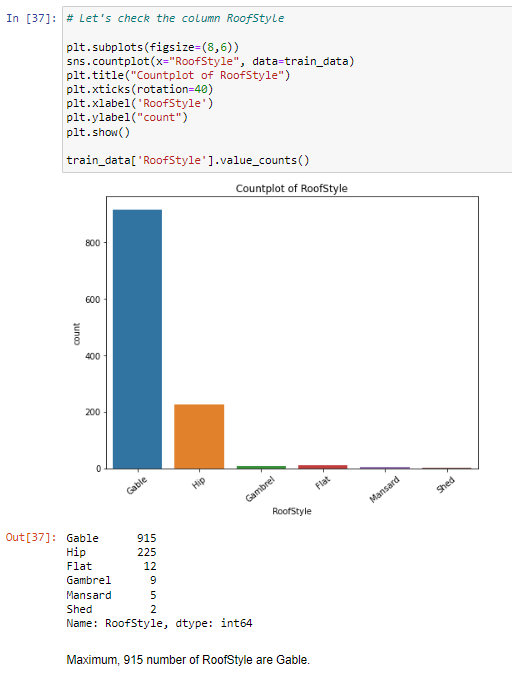


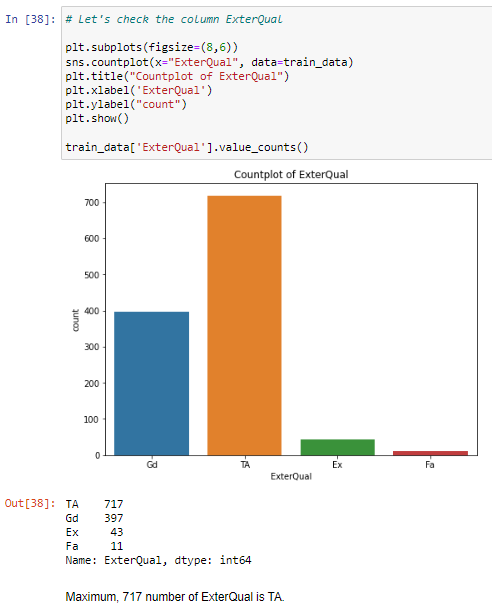


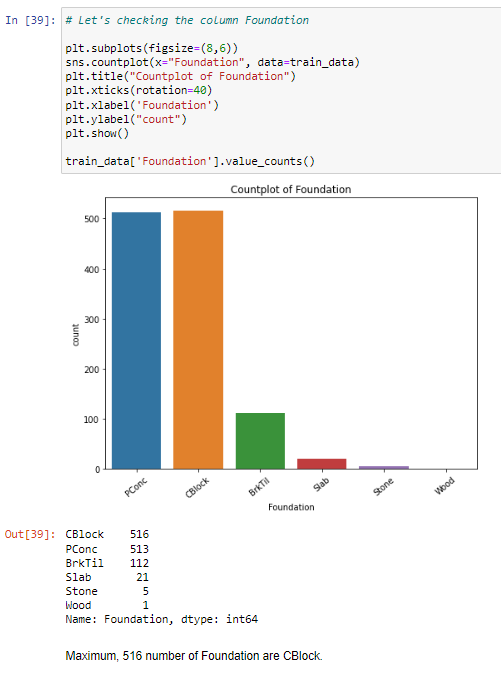




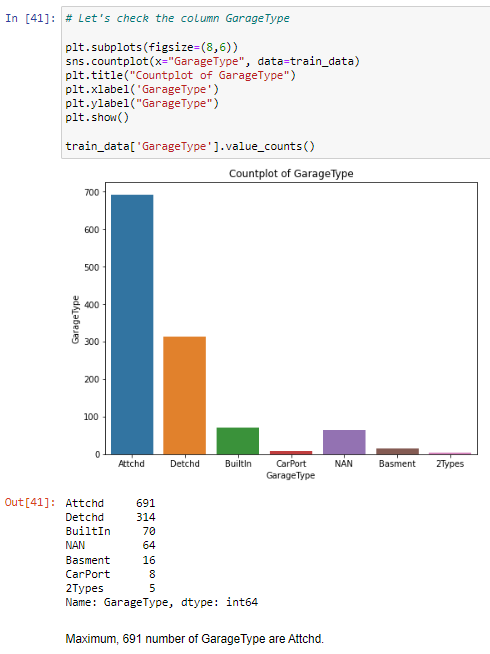


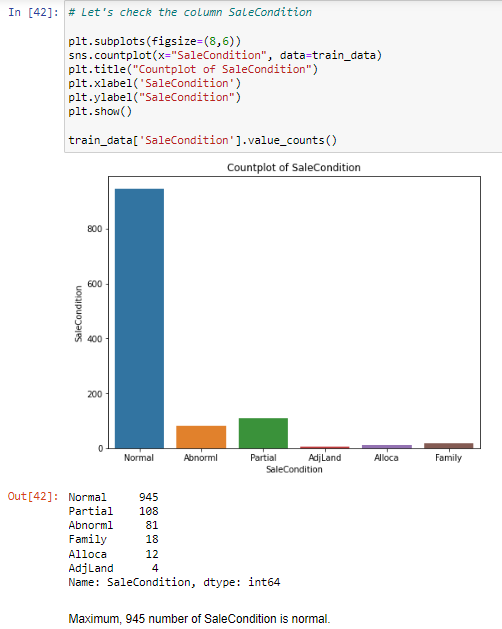


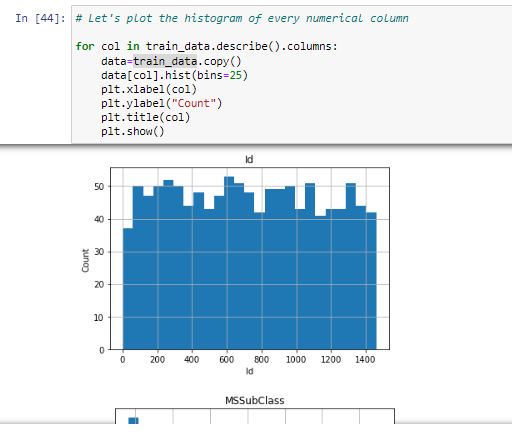


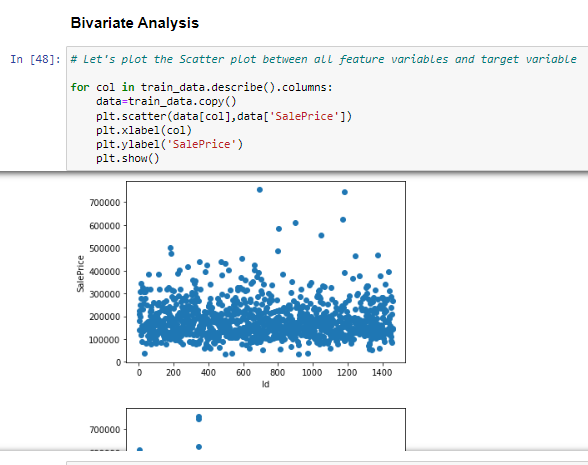


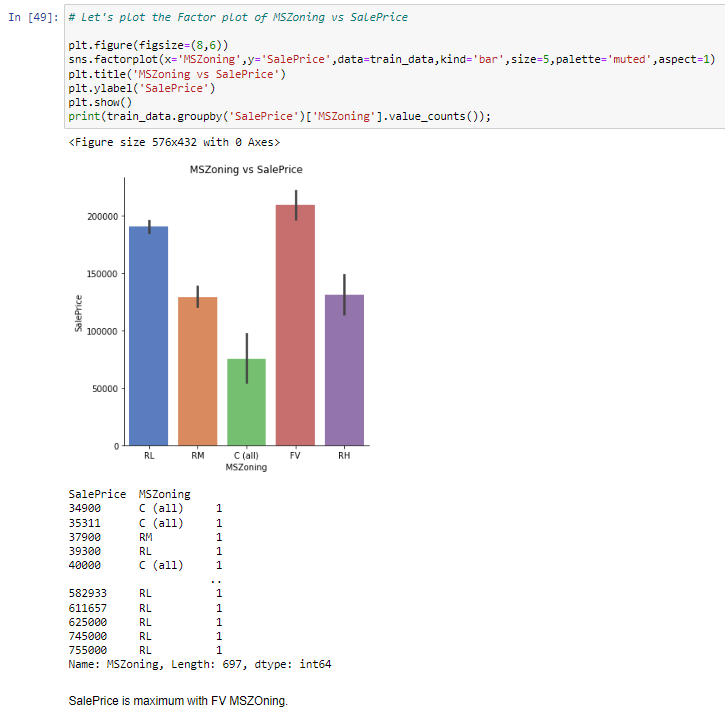


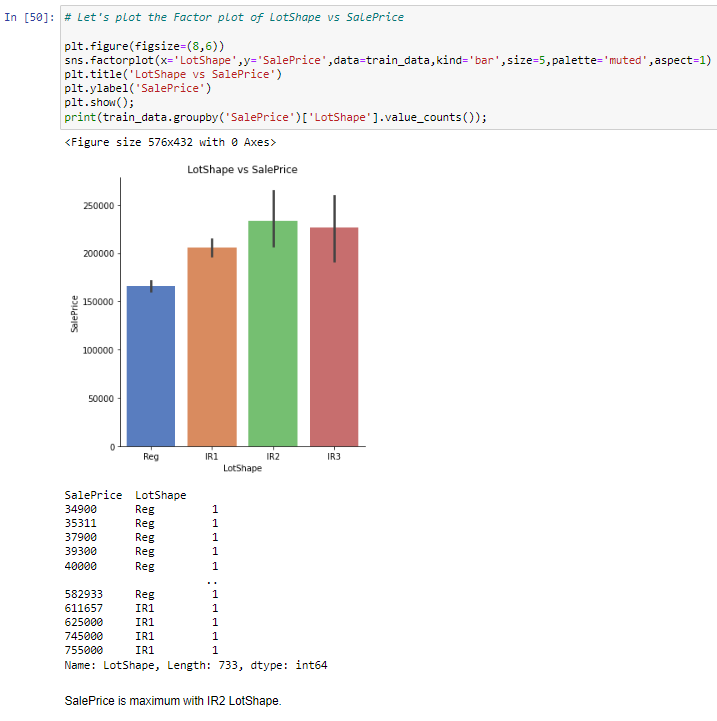


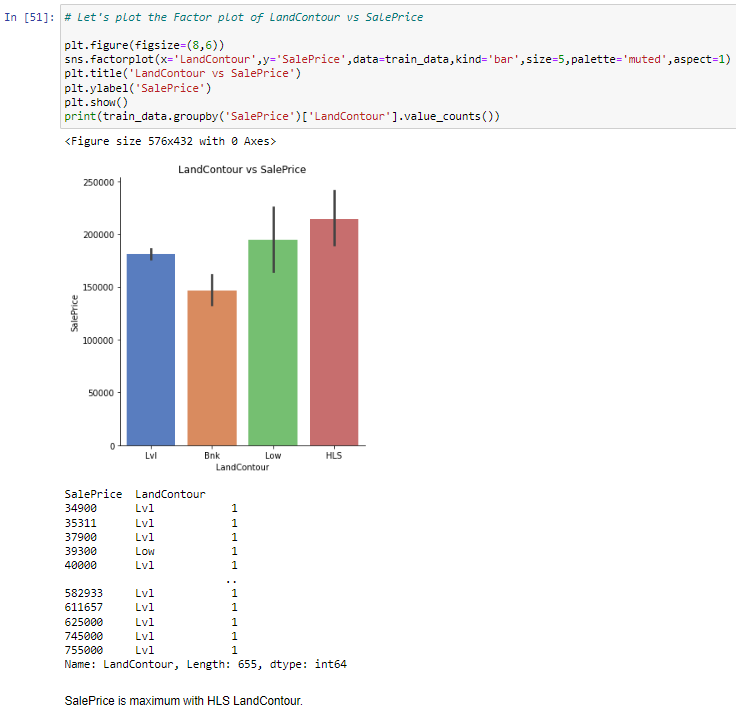




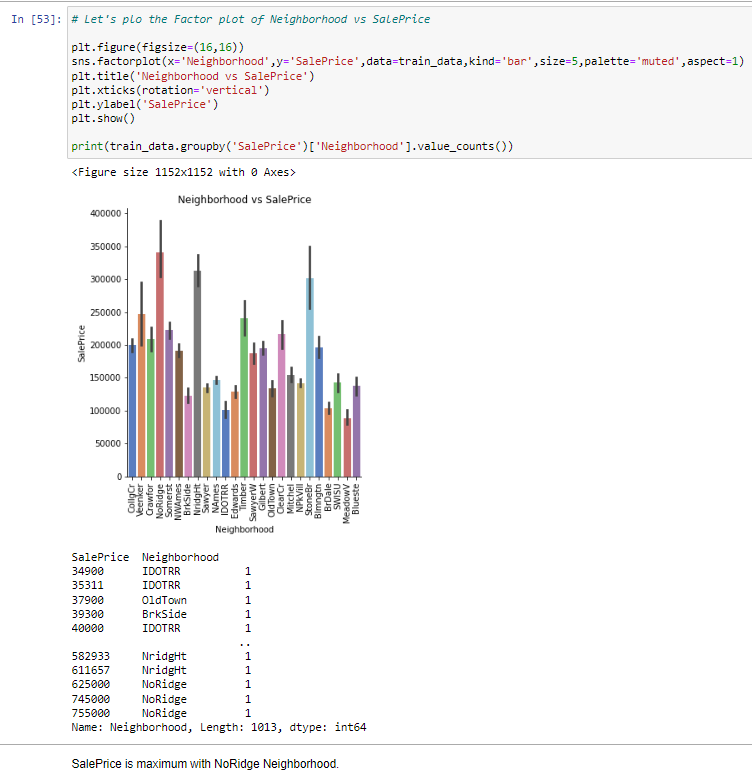


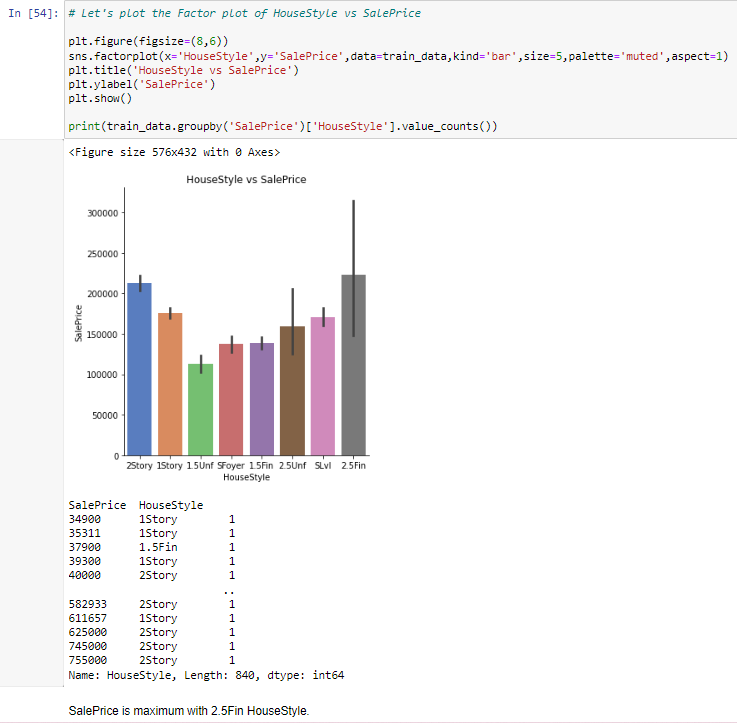


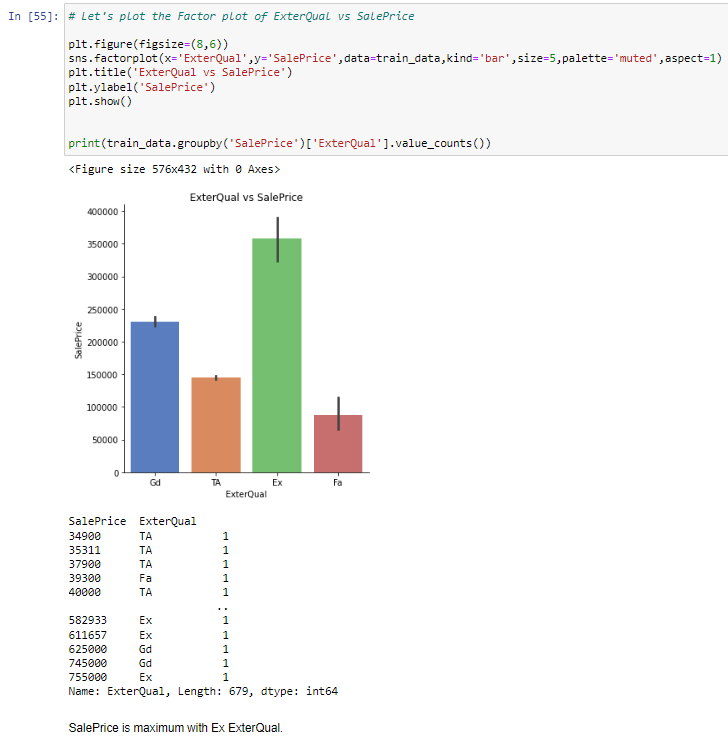


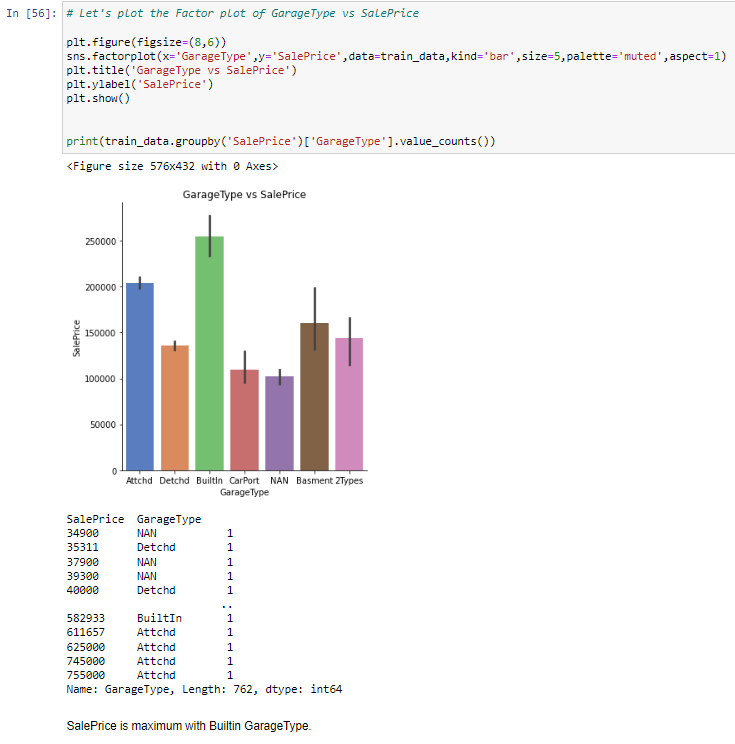


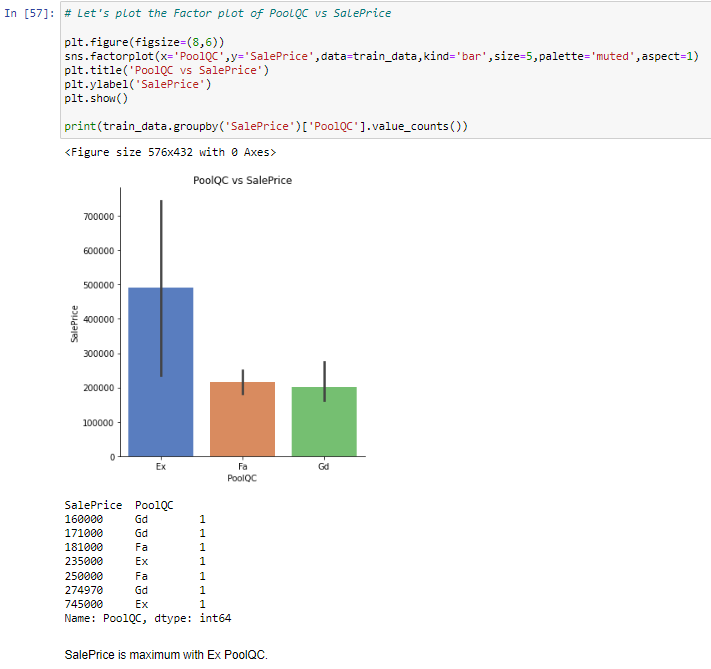


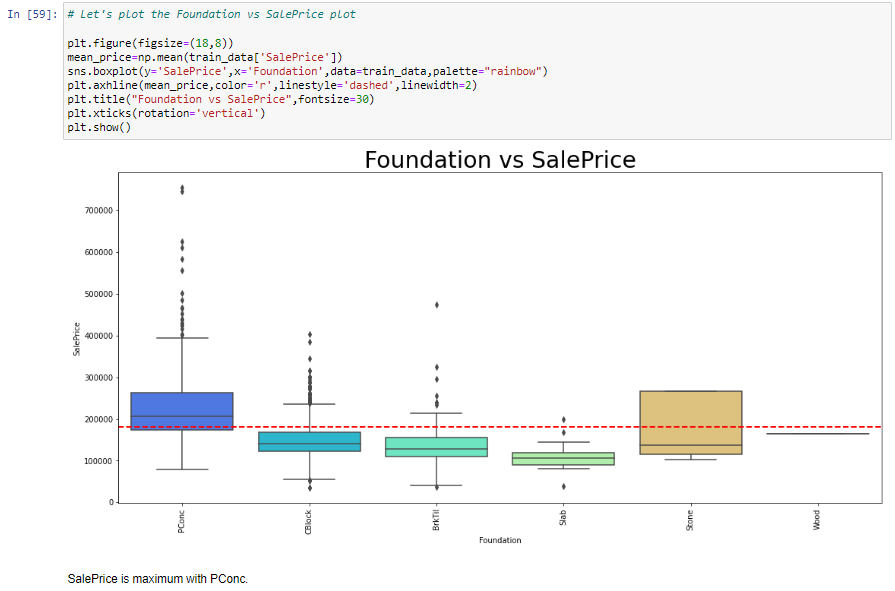


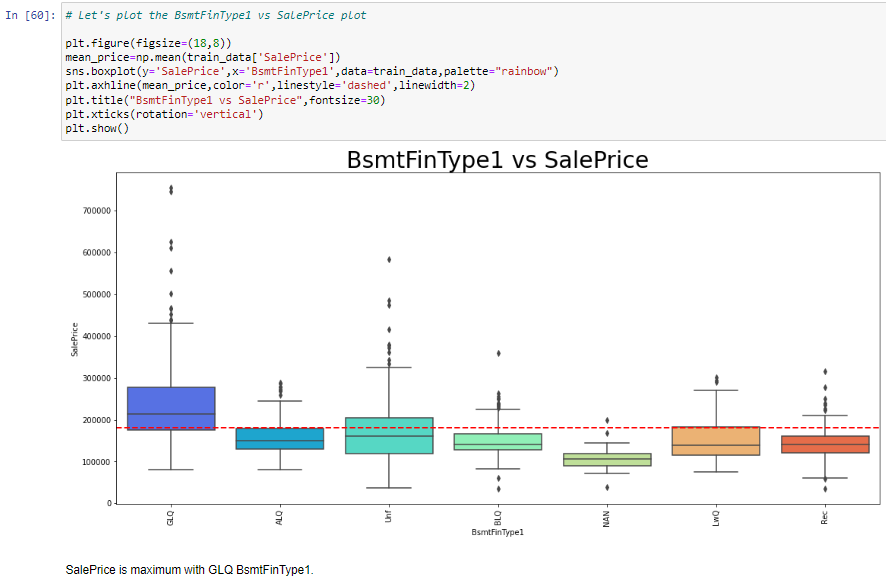


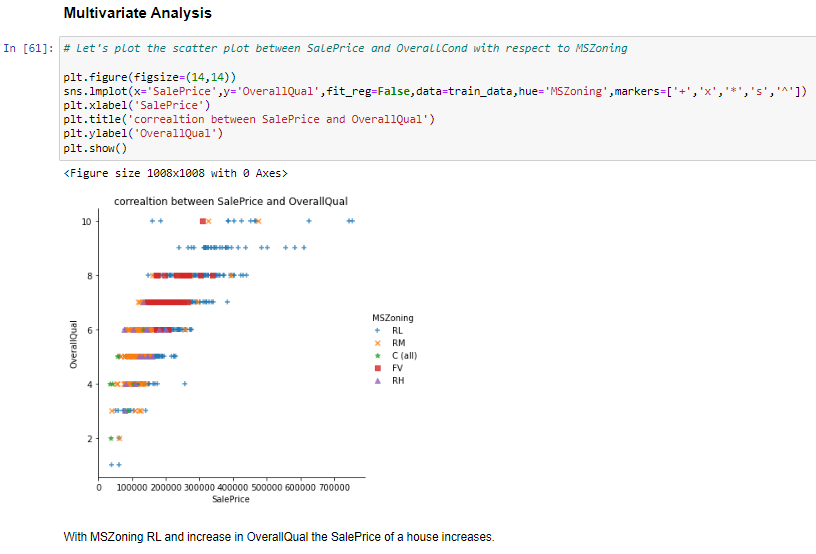


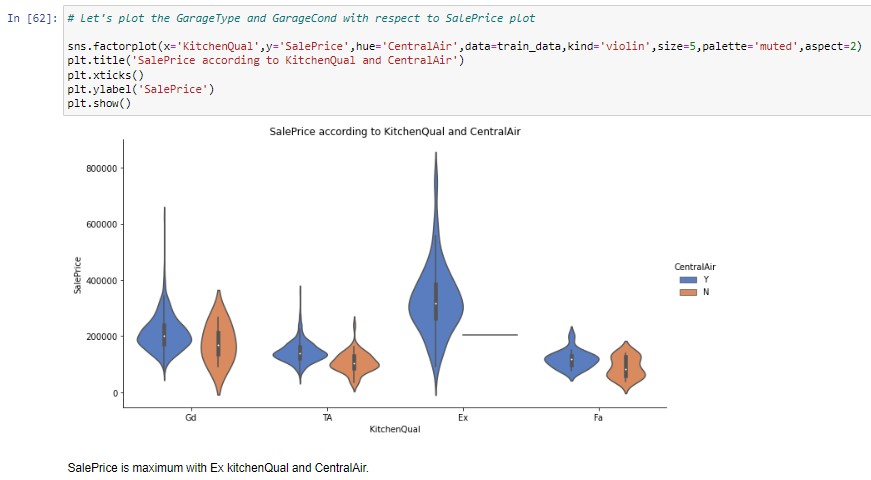


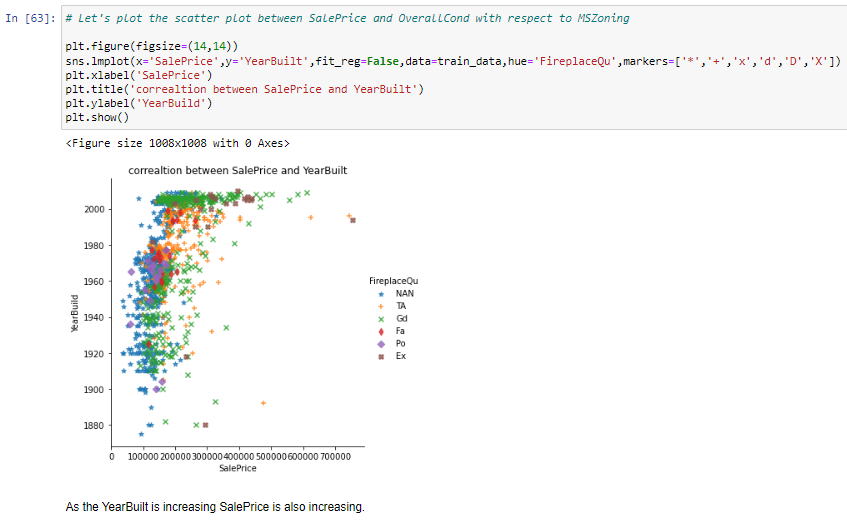


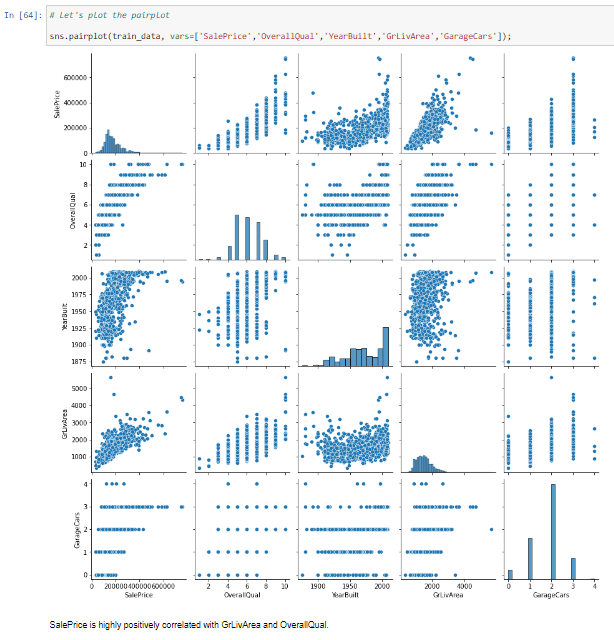




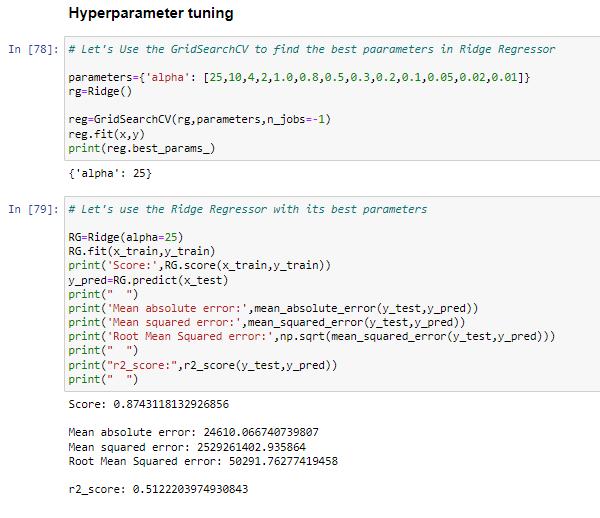








* **Interpretation of the Results:**
* From the visualization we interpreted that the target variable SalePrice was highly positively correlated with the columns GrLivArea, YearBuilt, OverallQual, GarageCars, GarageArea.
* From the preprocessing we interpreted that data was improper scaled.



From the modeling we interpreted that after hyperparameter tuning Ridge Regressor works best giving 87% score.

**CONCLUSION**

* **Key Findings and Conclusions of the Study:**

In this project we have tried to show how the house prices vary and what are the factors related to the changing of house prices.

The best score was achieved using the best parameters of Ridge Regressor through GridSearchCV though Lasso Regressor model performed well too.

* **Learning Outcomes of the Study in respect of Data Science:**

This project has demonstrated the importance of sampling effectively, modelling and predicting data.

Through data cleaning we were able to remove unnecessary columns and outliers from our dataset due to which our model would have suffered from overfitting or underfitting.

The few challenges while working on this project where:-

* + Improper scaling
  + Too many features
  + Missing values
  + Skewed data due to outliers

The data was improper scaled so we scaled it to a single scale using sklearns’s package StandardScaler.

There were too many(256) features present in the data so we applied Principal Component Analysis(PCA) and found out the Eigenvalues and on the basis of number of nodes we were able able to reduce our features upto 90 columns.

There were lot of missing values present in different columns which we imputed on the basis of our understanding.

The columns were skewed due to presence of outliers which we handled through winsorization technique.

* **Limitations of this work and Scope for Future Work**

While we couldn’t reach out goal in house price prediction without letting the model to overfit, we did end up creating a system that can with enough time and data get very close to that goal. As with any project there is room for improvement here. The very nature of this project allows for multiple algorithms to be integrated together as modules and their results can be combined to increase the accuracy of the final result. This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code