|  |  |  |
| --- | --- | --- |
| D:\Admission\Admissions_1011\wwwroot\hdrlogo.gif | Department of Computer Engineering  D. J. Sanghvi College of Engineering  University of Mumbai  A. Y. 2018-2019 | E:\new logo.JPG |

**PROJECT TITLE**

*House Price Prediction System*

**MINI PROJECT**

By

**Mansi Jhaveri 60004178009**

**Forum Shah 60004178018**

Guide:

**Name of the Guide**

Designation

**CERTIFICATE**

This is to certify that the project entitled **“House Price Prediction System”** is a bonafide work of **“Mansi Jhaveri” (60004178009)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of B.E. in Computer Engineering

**(Name and sign)**

**Guide**

**(Name and sign) (Name and sign)**

**Head of Department Principal**

**Mini Project Report Approval.**

This mini project report entitled (**House Price Prediction System**) by (**Author**) is approved for the degree of **B.E. in Computer Engineering.**

Examiners

1.---------------------------------------------

2.---------------------------------------------

Date:

Place:

Declaration

I/We declare that this written submission represents my/our ideas in my/our own words and where others' ideas or words have been included, I/We have adequately cited and referenced the original sources. I/We also declare that I/We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my/our submission. I/We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

-----------------------------------------

(Signature)

Mansi Jhaveri (60004178009)

Forum Shah (60004178018)

-----------------------------------------

(Name of students and Roll Nos.)

Date:

Table of Contents

Table of Contents ii

1. Abstract 1

2. Introduction 2

3. Literature Survey 3

4. Proposed System 4

4.1 Scope of the project 4

4.2 System Design 4

4.3 Implementation Details 4

4.4 Result Analysis 4

4.5 Conclusion and Future Scope 4

4.6 References 4

# 1. Abstract

The phenomenon of the falling or rising of the house prices has attracted interest from the researcher as well as many other interested parties. House prices increase every year, so there is a need for a system to predict house prices in the future. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house. There are three factors that influence the price of a house which include physical conditions, concept and location. This research aims to predict house prices in Mumbai city with Machine learning algorithms.The previous research works used various regression techniques to address the question of the changes house price. This work considers the issue of changing house price as a classification problem and applies machine learning techniques to predict whether house prices will rise or fall. This work applies various feature selection techniques such as variance influence factor, Information value, principle component analysis and data transformation techniques such as outlier and missing value treatment as well as box-cox transformation techniques. The goal of the project is to create a system that can provide users with location-specific predictions and trends.

# 2. Introduction

The aim of this document is to gather and analyze and give an in-depth insight of the complete House price prediction system by defining the problem statement in detail. Nevertheless, accurate prediction of house prices has been always a fascination for the buyers, sellers and for the bankers also. The detailed requirements of the House price prediction system are provided in this document. The purpose of the document is to collect and analyze all assorted ideas that have come up to define the system, its requirements with respect to consumers. Also, we shall predict and sort out how we hope this product will be used in order to gain a better understanding of the project, outline concepts that may be developed later, and document ideas that are being considered, but may be discarded as the product develops. The diversity of features makes it challenging to estimate an adequate market price. Apart from providing a summary of the important features of the house, the house description is also a means of raising curiosity in the reader, or in other words to persuade the person. It is possible that there are certain word sequences in the natural language text that seduce potential buyers more than others. Therefore, there might be a relation between the language used in the description and the price of the property. This comparison does not focus primarily on the house characteristics, but on all words within the description.

# 3. Literature Survey

The previous research papers have used different machine learning algorithms to develop this project which include :- Naïve Bayes algorithm, Random Forest and Support Vector Regression. In this project, regression algorithm is used to implement it. The previous papers used hedonic pricing model to estimate house prices in the past decade. The project uses the dataset that consists of 81 attributes and 1460 entries. The past papers used Naïve Bayes algorithm which is a supervised machine-learning algorithm that uses the Bayes’ Theorem, which assumes that features are statistically independent. The theorem relies on the naive assumption that input variables are independent of each other, i.e. there is no way to know anything about other variables when given an additional variable. Regardless of this assumption, it has proven itself to be a classifier with good results. Naive Bayes Classifiers rely on the Bayes’ Theorem, which is based on conditional probability or in simple terms, the likelihood that an event (A) will happen given that another event (B) has already happened. Essentially, the theorem allows a hypothesis to be updated each time new evidence is introduced. Naive Bayes Classifiers rely on the Bayes’ Theorem, which is based on conditional probability or in simple terms, the likelihood that an event (A) will happen given that another event (B) has already happened. Essentially, the theorem allows a hypothesis to be updated each time new evidence is introduced. Support vector machines are linear discriminant functions (classifier) with the maximum margin is the best. The margin is defined as the width that the boundary could be increased by, before hitting a data point .Random Forests are ensemble classifiers constructed from of a set of Decision Trees, with the output of the classifier being the mode of the output of the Decision Trees. Random Forests combine the “bagging” idea of Breiman with the idea of random selection of features. The algorithm for inducing a Random Forest was developed by Leo Breiman and Adele Cutler. The artificial neural networks use neurons or perceptrons as the basic units. These perceptrons use a vector of realvalued inputs. These inputs are always having a linear combination between themselves. The Output is 1 is the function is more than if the result is greater than a threshold value and Output is 0, otherwise. Regression algorithm is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

# 4. Proposed System

This system will handle all the prediction queries from the user. The dataset used for this system consists of 81 attributes and 1460 entries. The system has used regression algorithm. This project predicts the efficient house pricing for real estate customers with respect to their budgets and priorities. This system give us a good prediction on the price of the house based on other variables.

## 4.1 Scope of the project

This project can provide users with location-specific predictions and trends on housing market. Obtaining predictions through the system is quick for the users, even though the prediction mechanism may involve computationally intensive tasks. The project has the power to handle prediction tasks from the user. It is built upon a platform which has the necessary computing power and implementation libraries installed.

## 4.2 System Design

## 4.3 Implementation Details

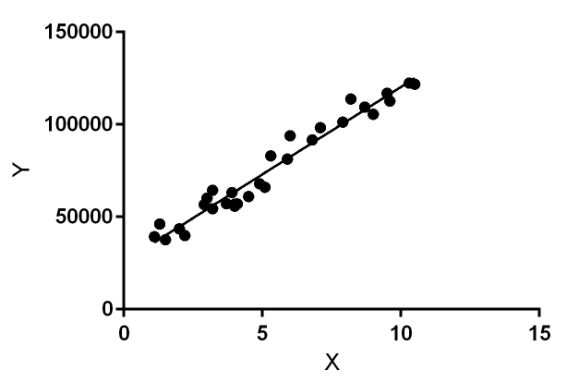
Technologies used:

* Python for backend
* Flask framework for integration of frontend and backend
* JavaScript frame work with CSS and HTML for front end
* Graphs for plotting

Machine Learning Algorithm:

**Linear Regression**

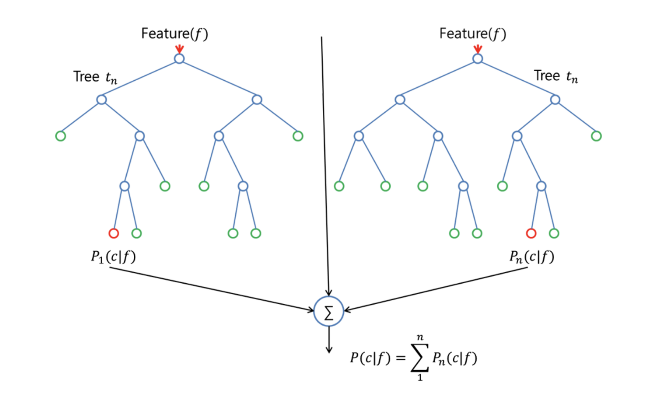
**Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

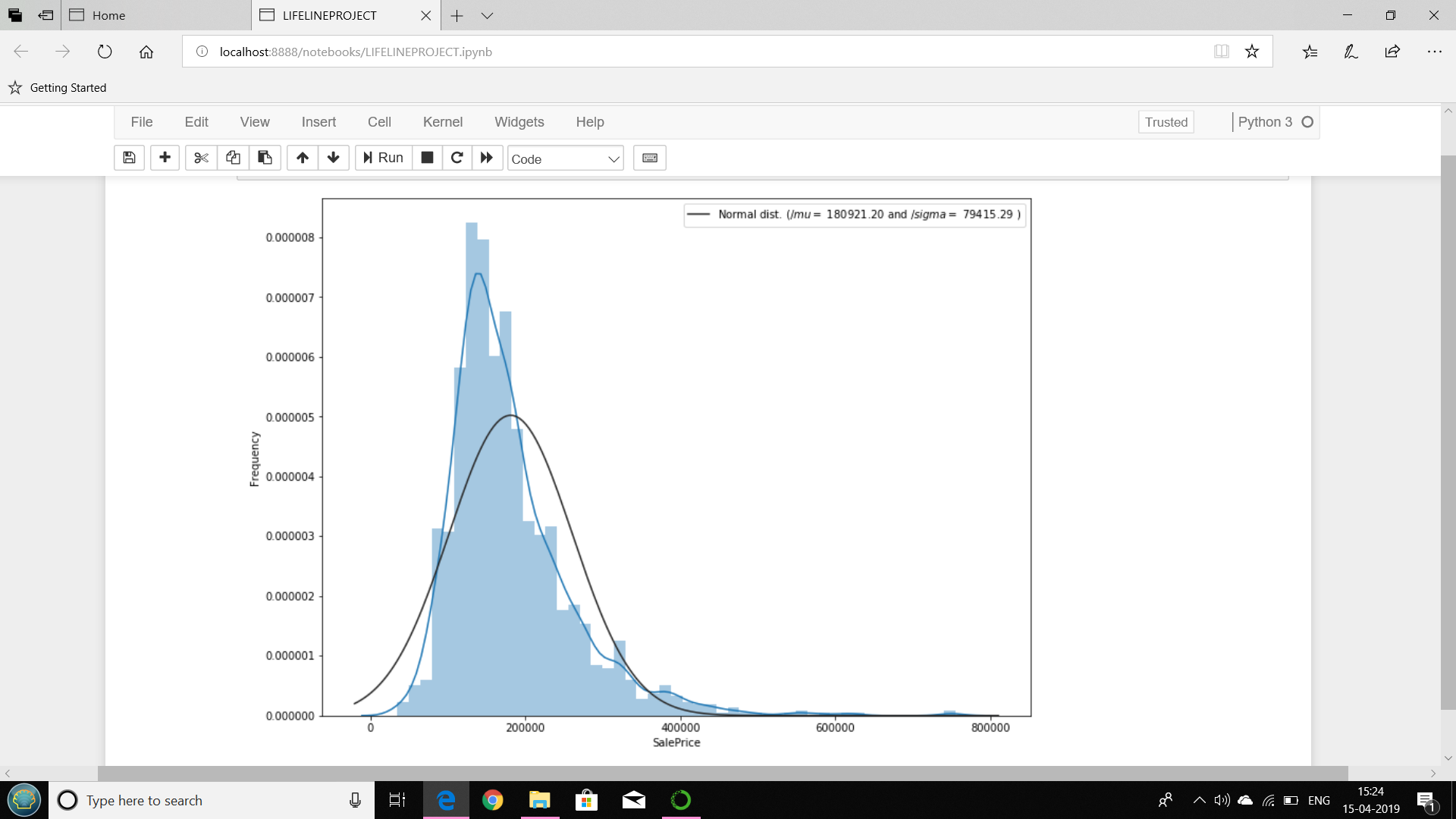
**Random Forest Regression:**

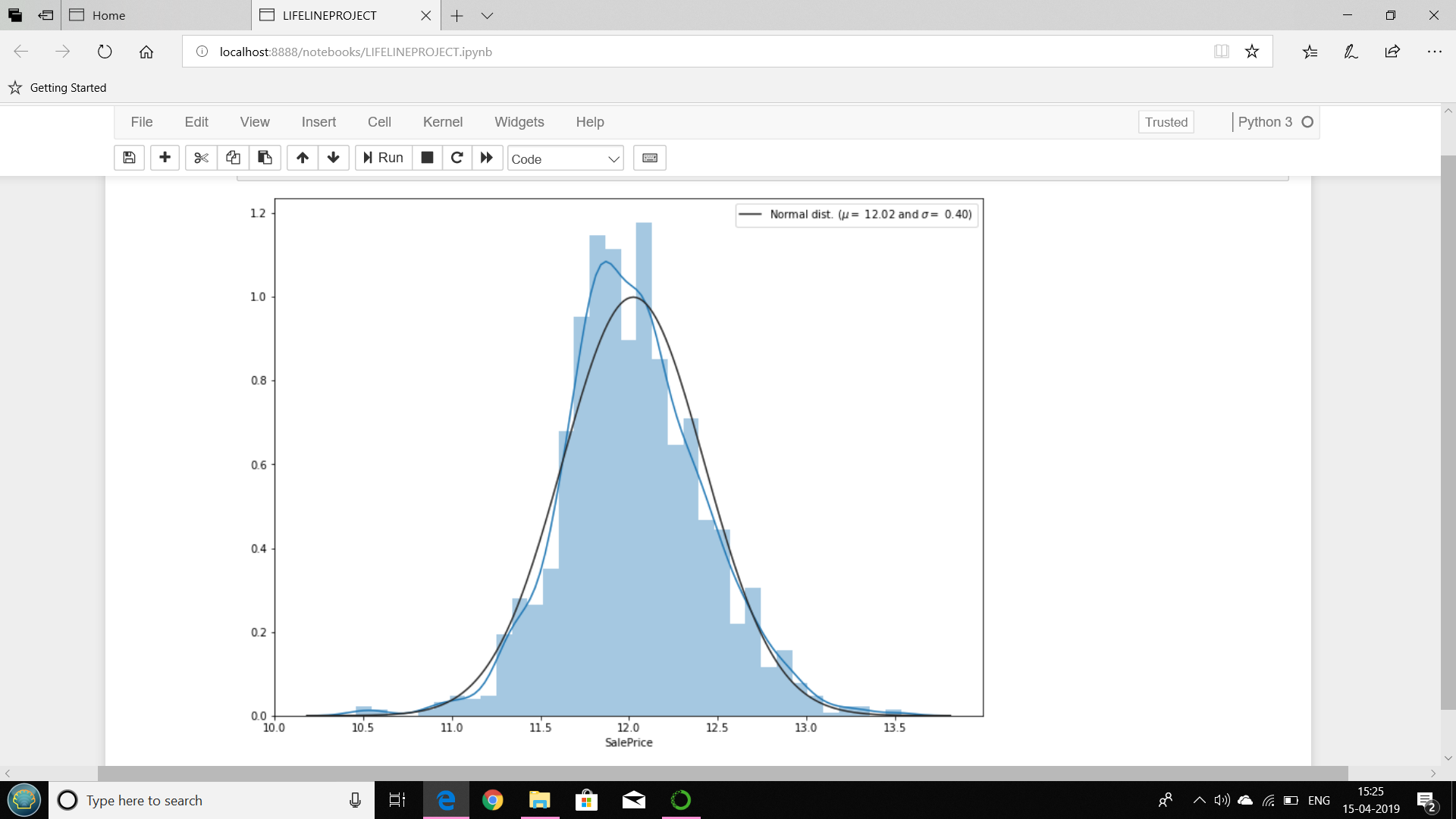
A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called **Bootstrap Aggregation**, commonly knownas **bagging**. What is bagging you may ask? Bagging, in the Random Forest method, involves training each decision tree on a different data sample where sampling is done with replacement.

[](https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&ved=2ahUKEwjZrMDajtLhAhWK6nMBHQjGB3wQjRx6BAgBEAU&url=http%3A%2F%2Fwww.aionlinecourse.com%2Ftutorial%2Fmachine-learning%2Frandom-forest-regression&psig=AOvVaw3cAO17SWAkNpNhYz_iAUPA&ust=1555418257362551)

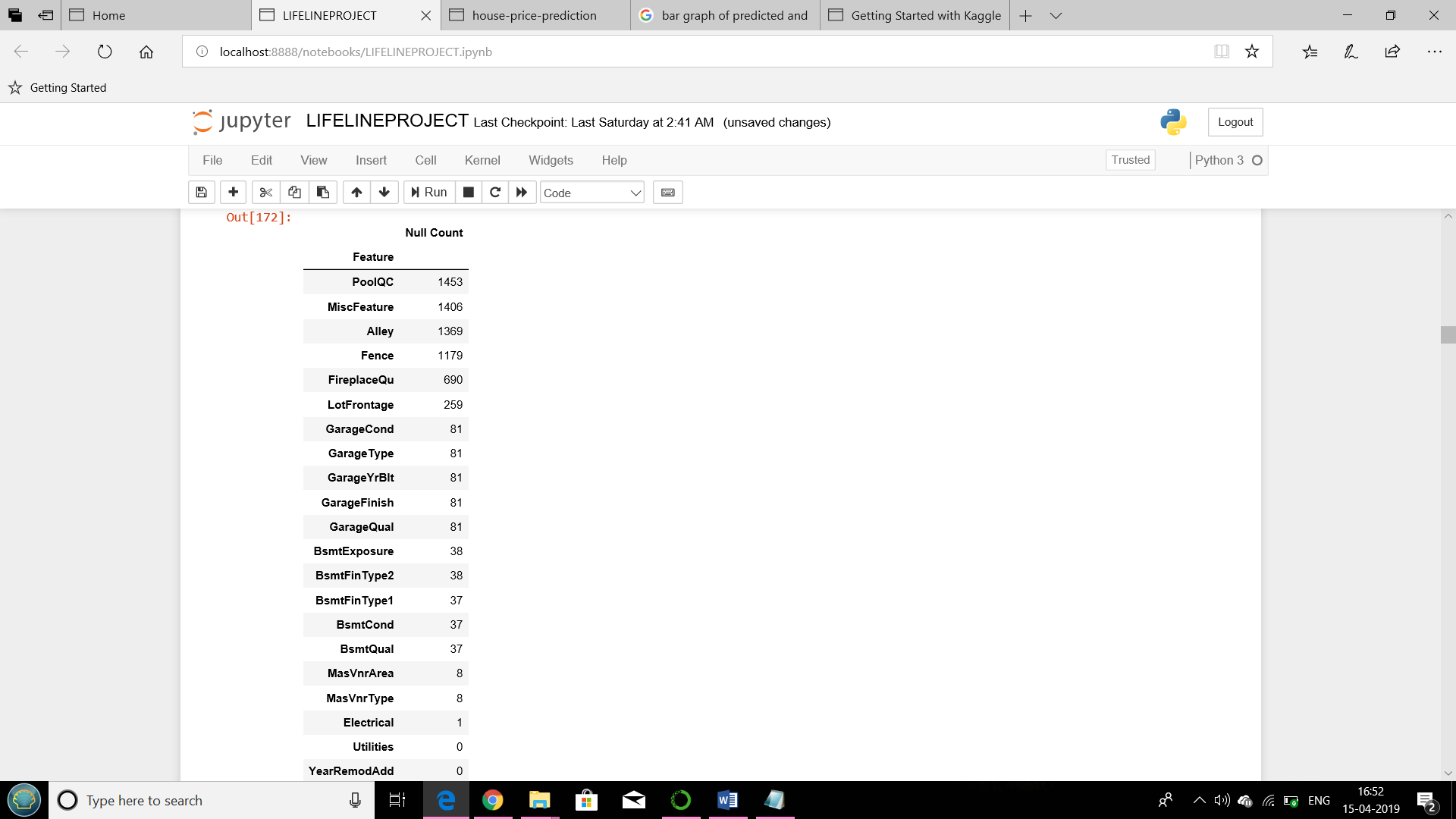
* **The project is using the above two mentioned machine learning algorithm for predicting the value of house.**
* **The accuracy rate using these two model has been shown in the screen shot which is available in result analysis section.**

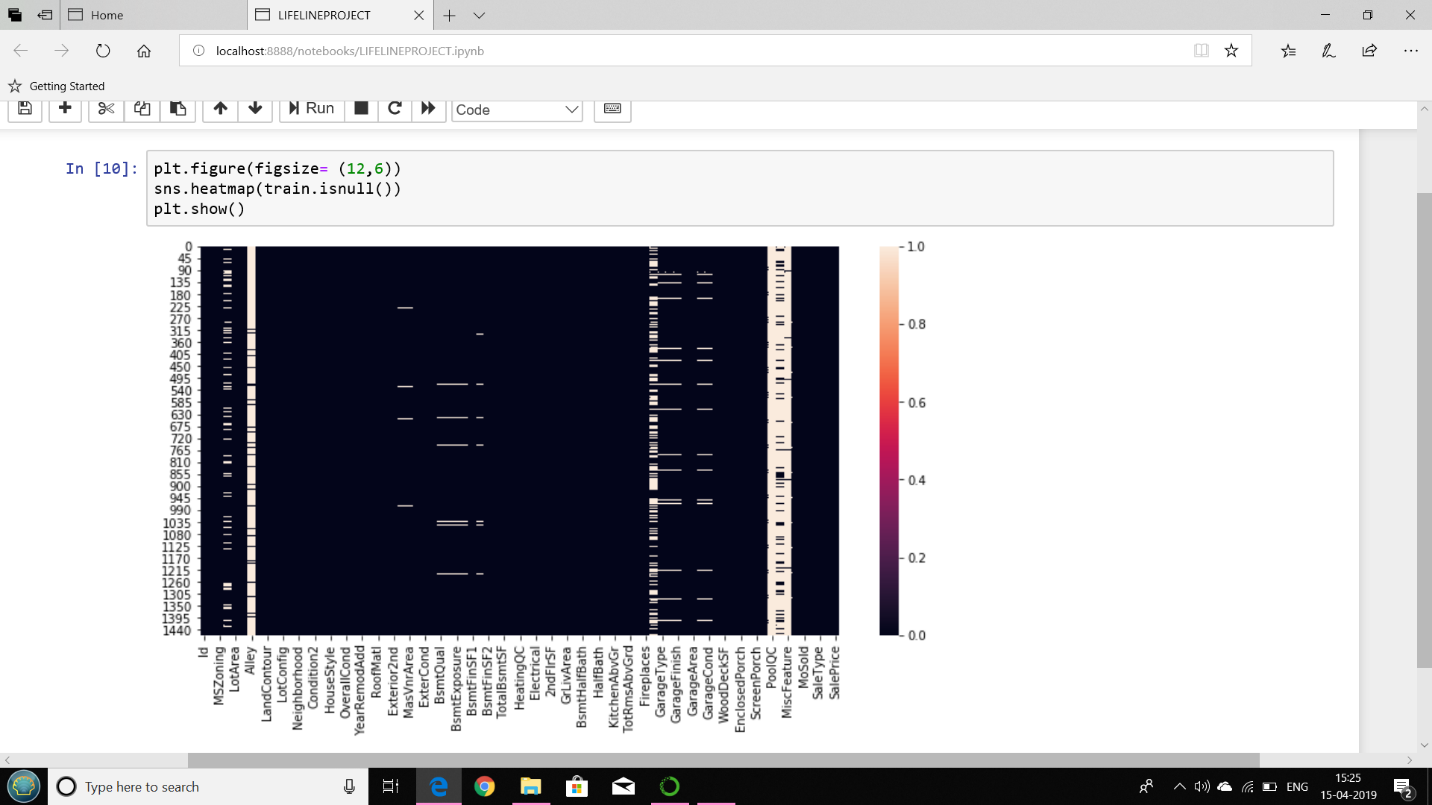
## 4.4 Result Analysis

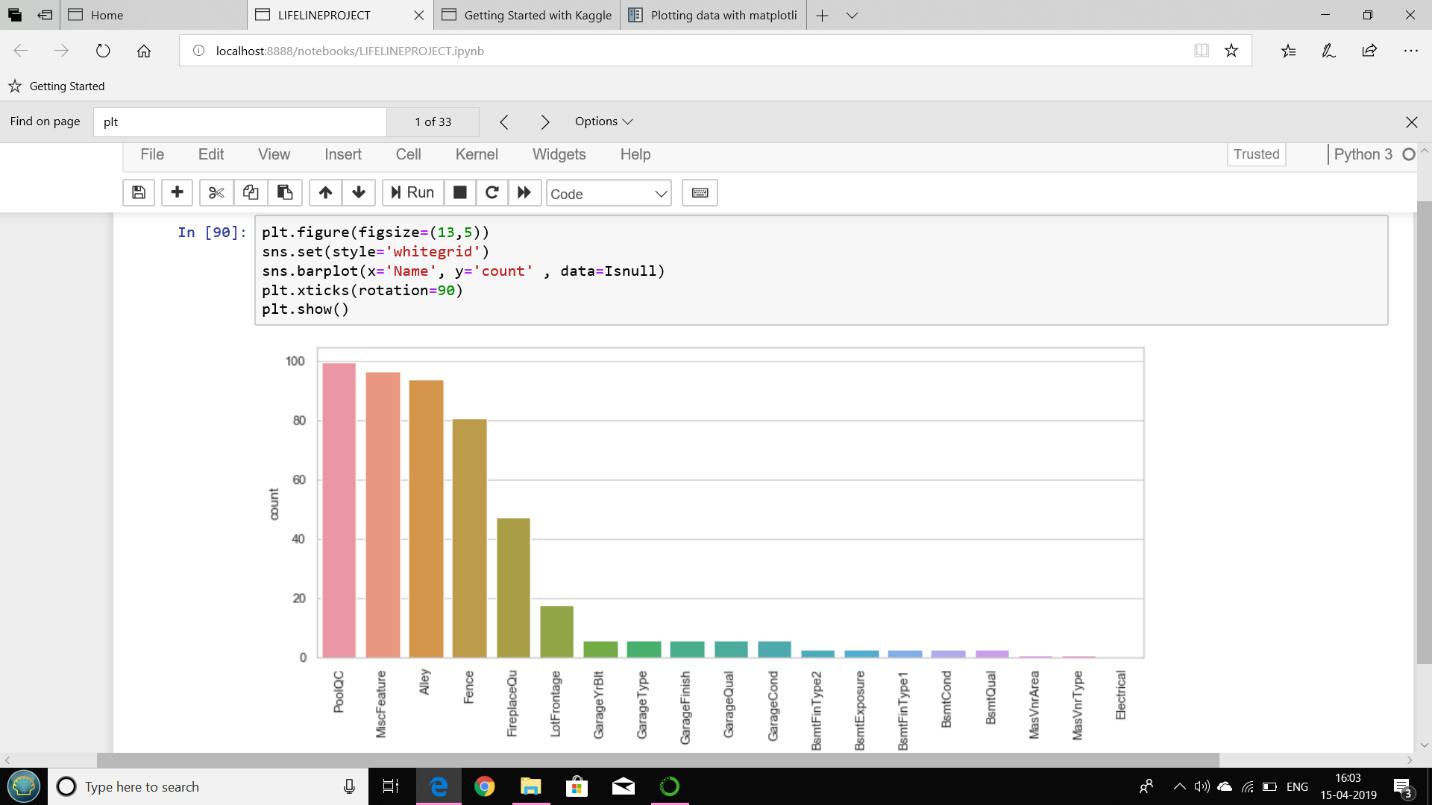
Considering frequency:

Not considering frequency :

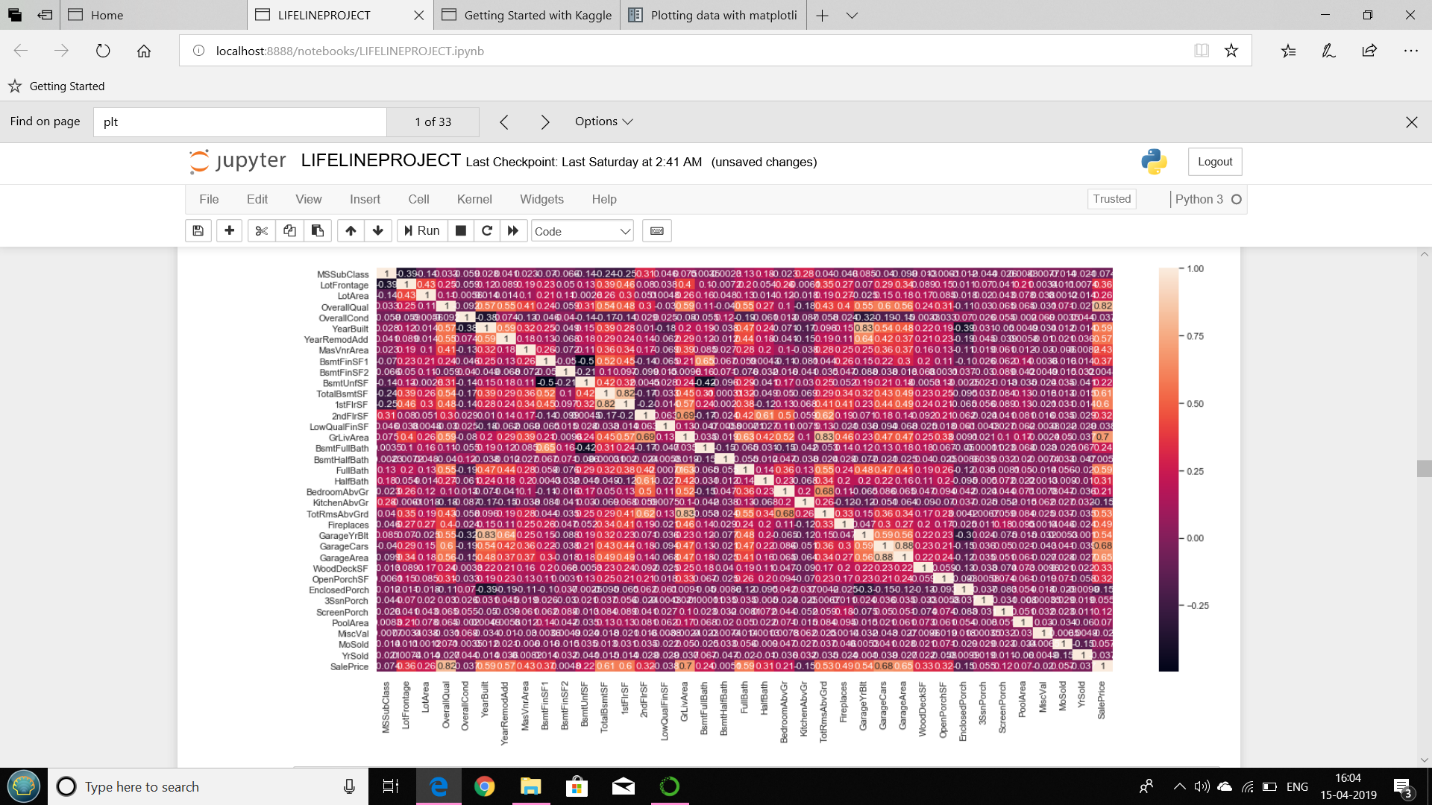
NULL VALUES REPRESENTATION:



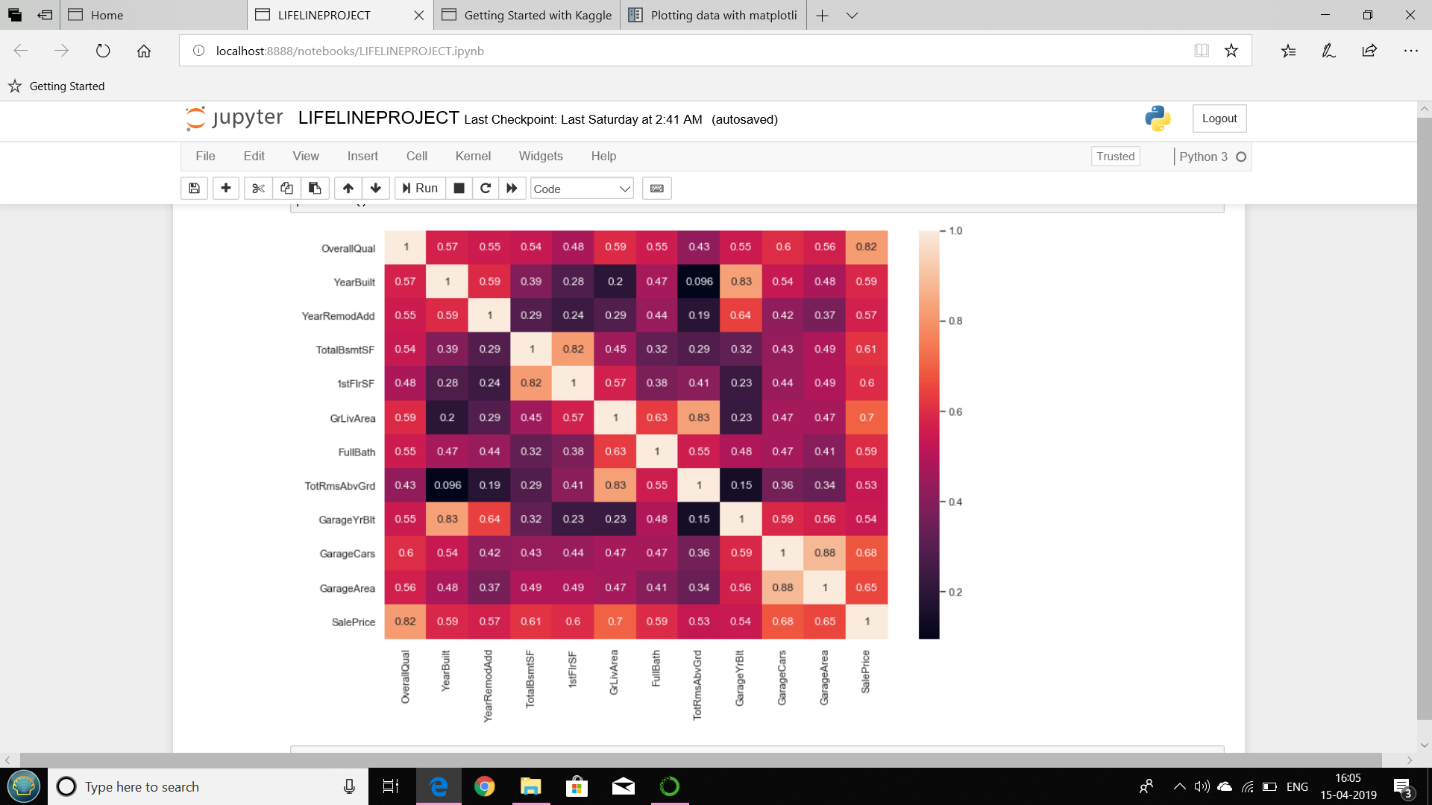


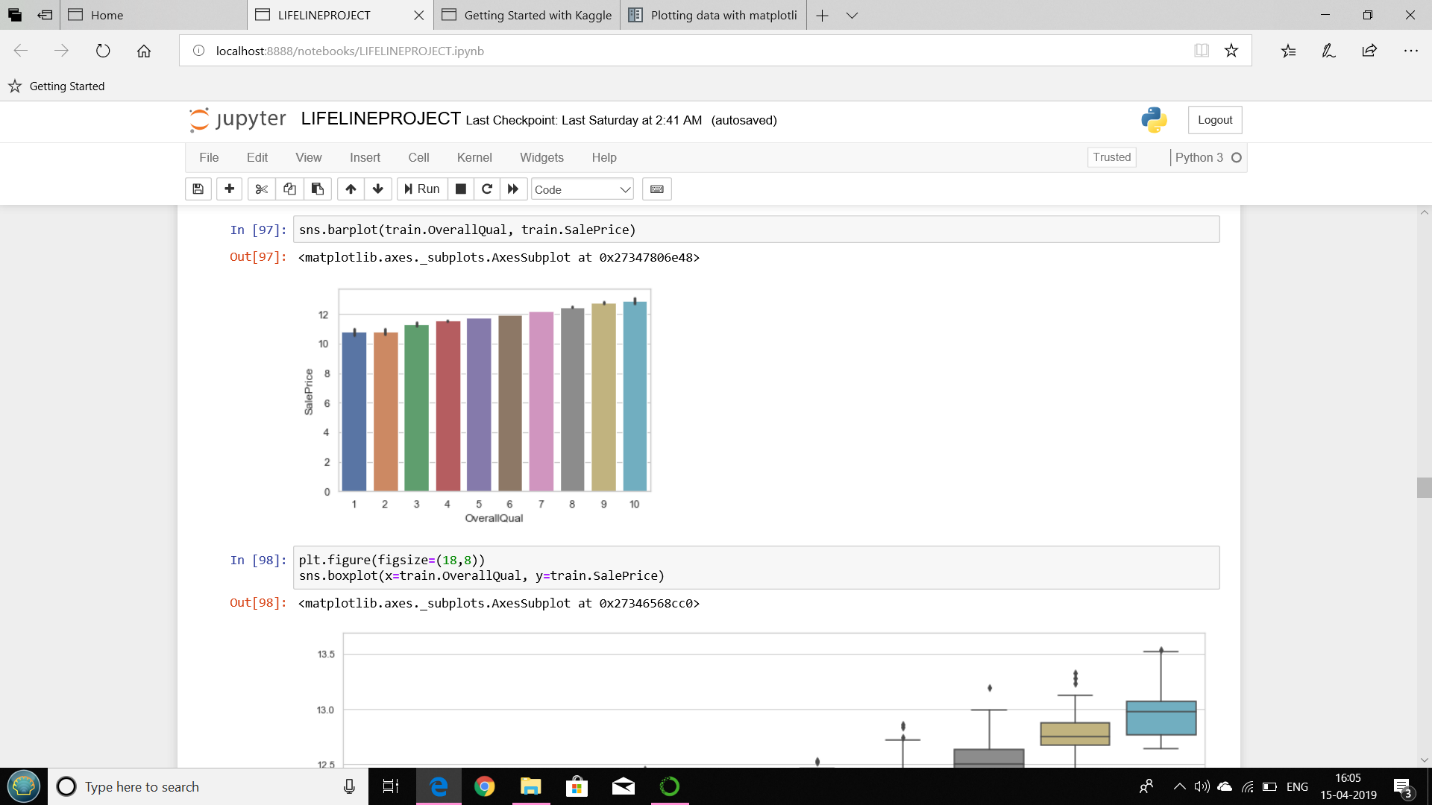


CORRELATION MATRIX(TRAINING DATASET):



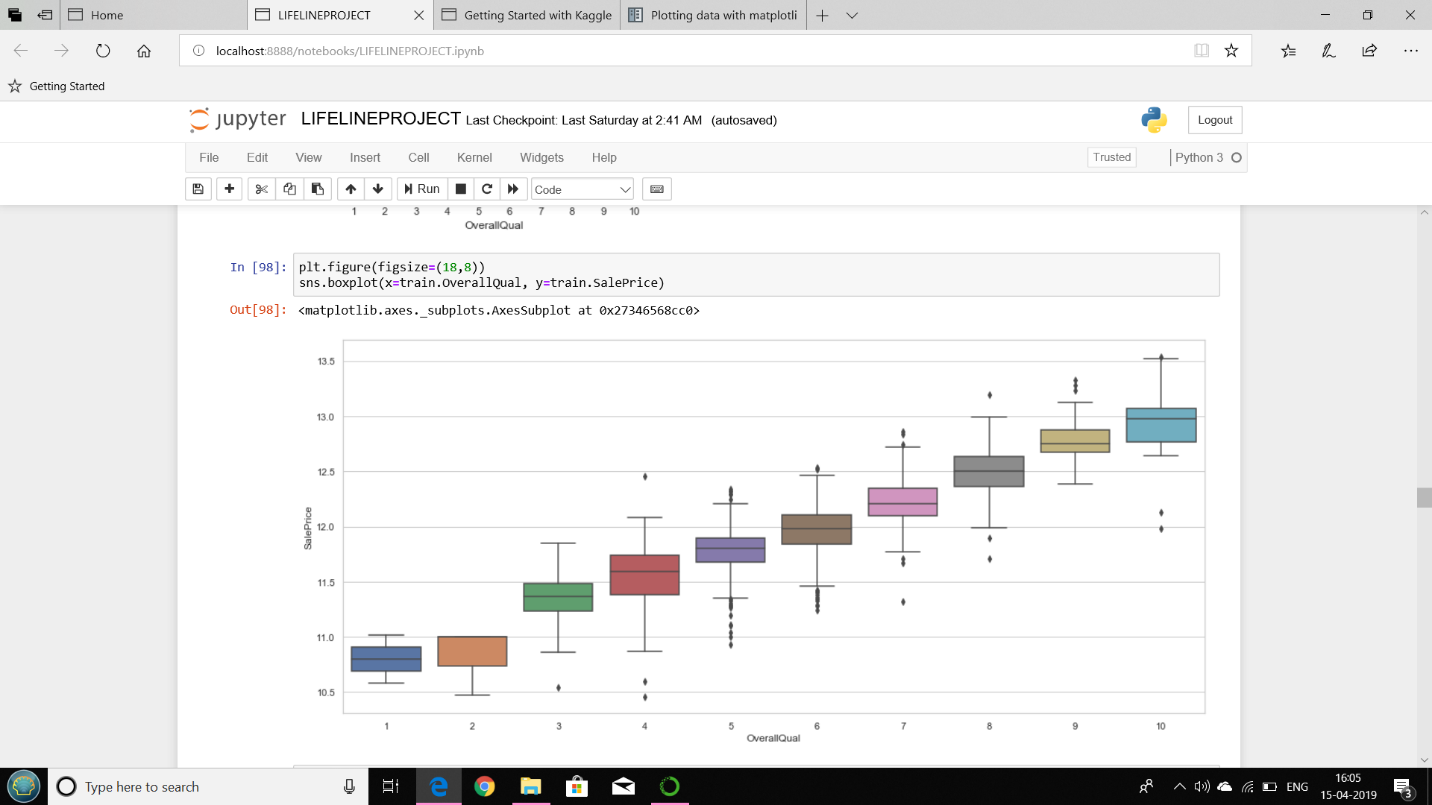
Considering top features for heatmap:



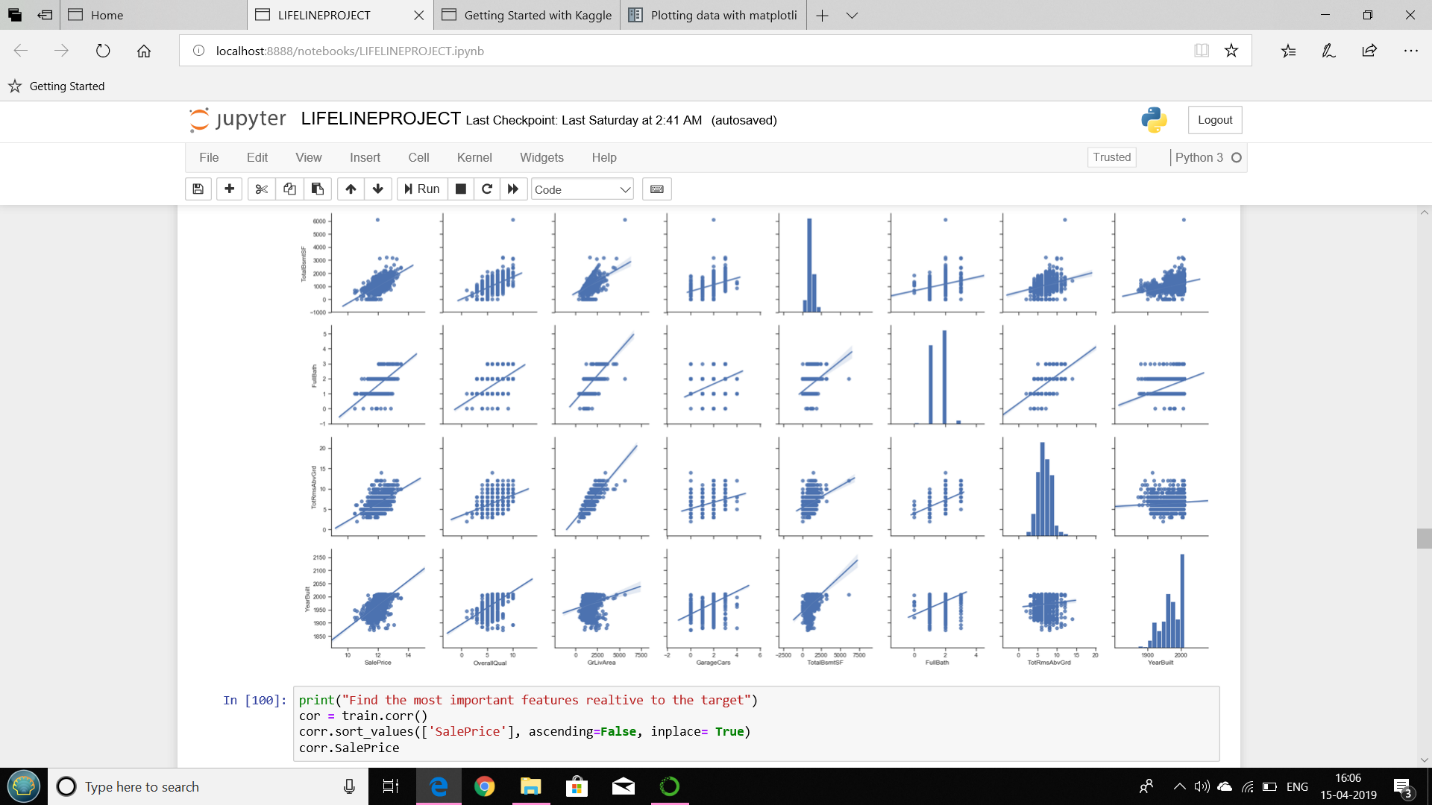
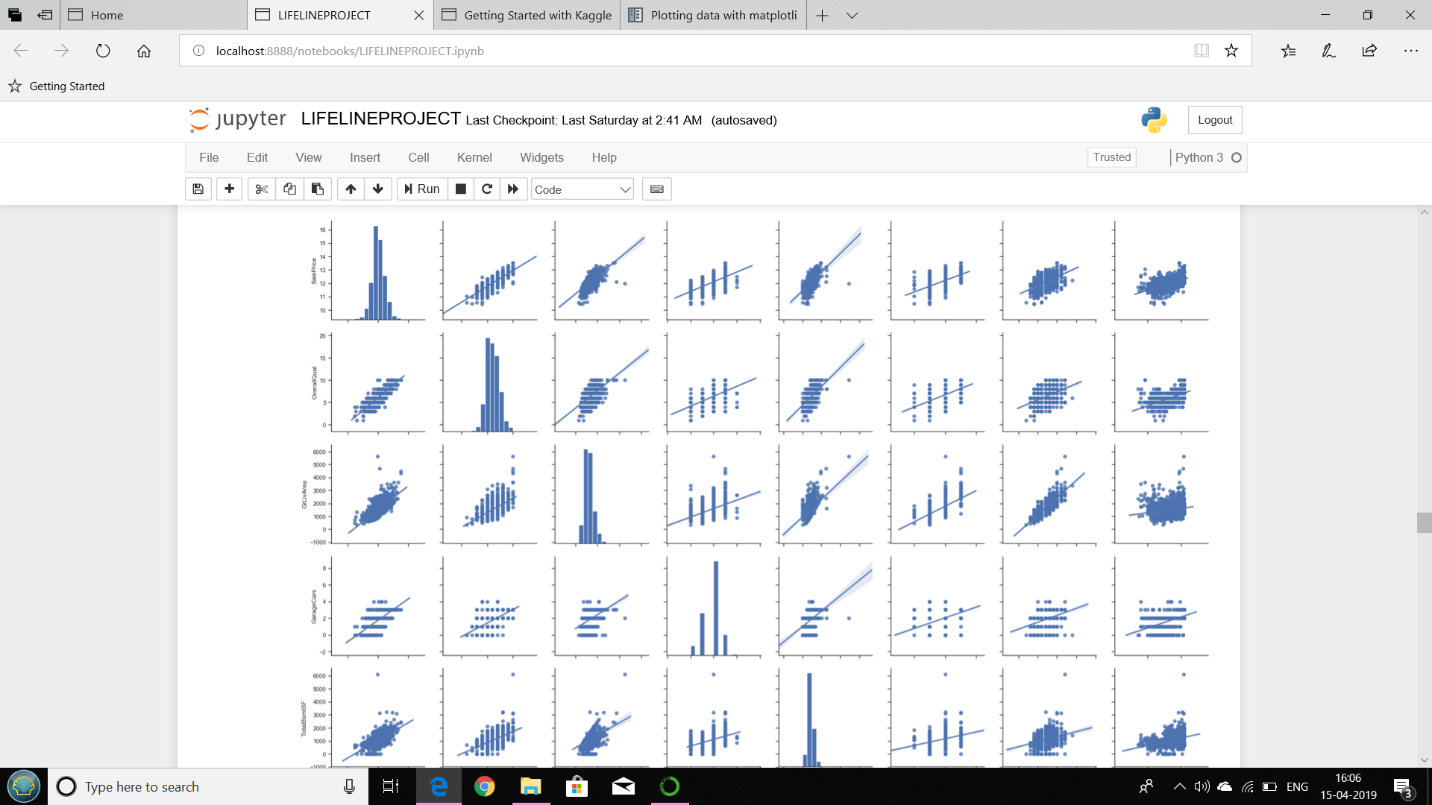


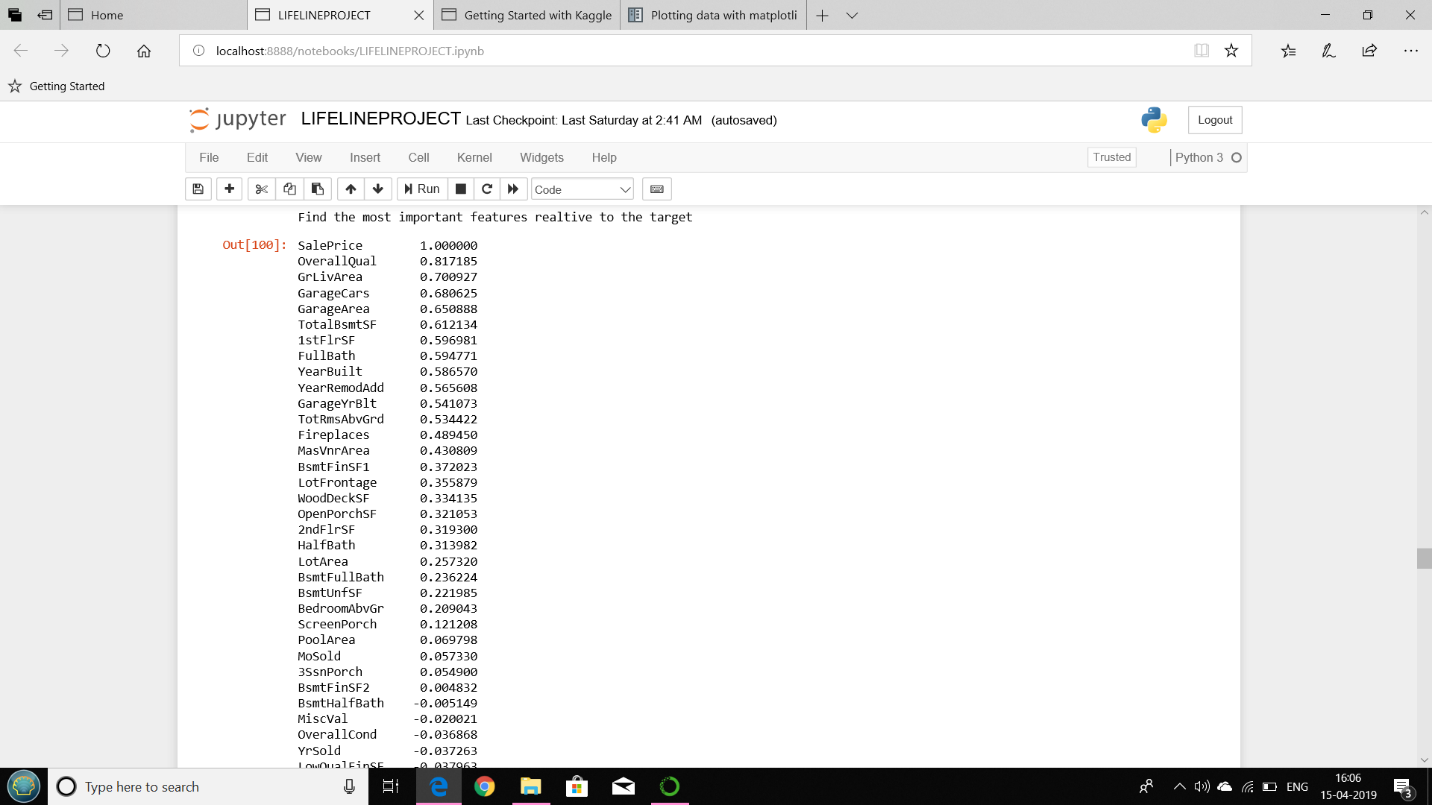
#### BOX GRAPH:

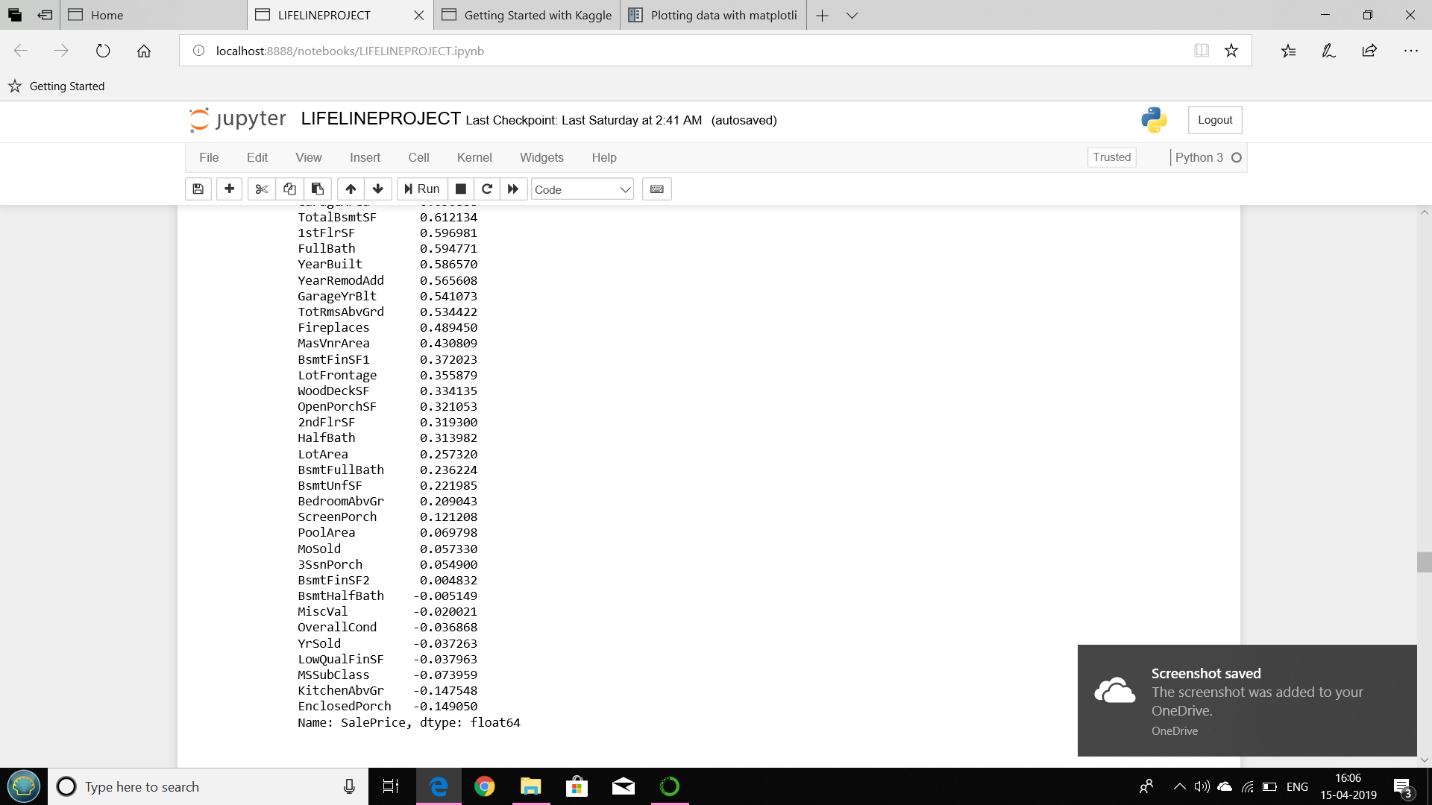
#### Relationship between categorical features and the target variable.The piece of code below shows how to plot a boxplot of the categorical variables SaleCondition w.r.t. the target variable.



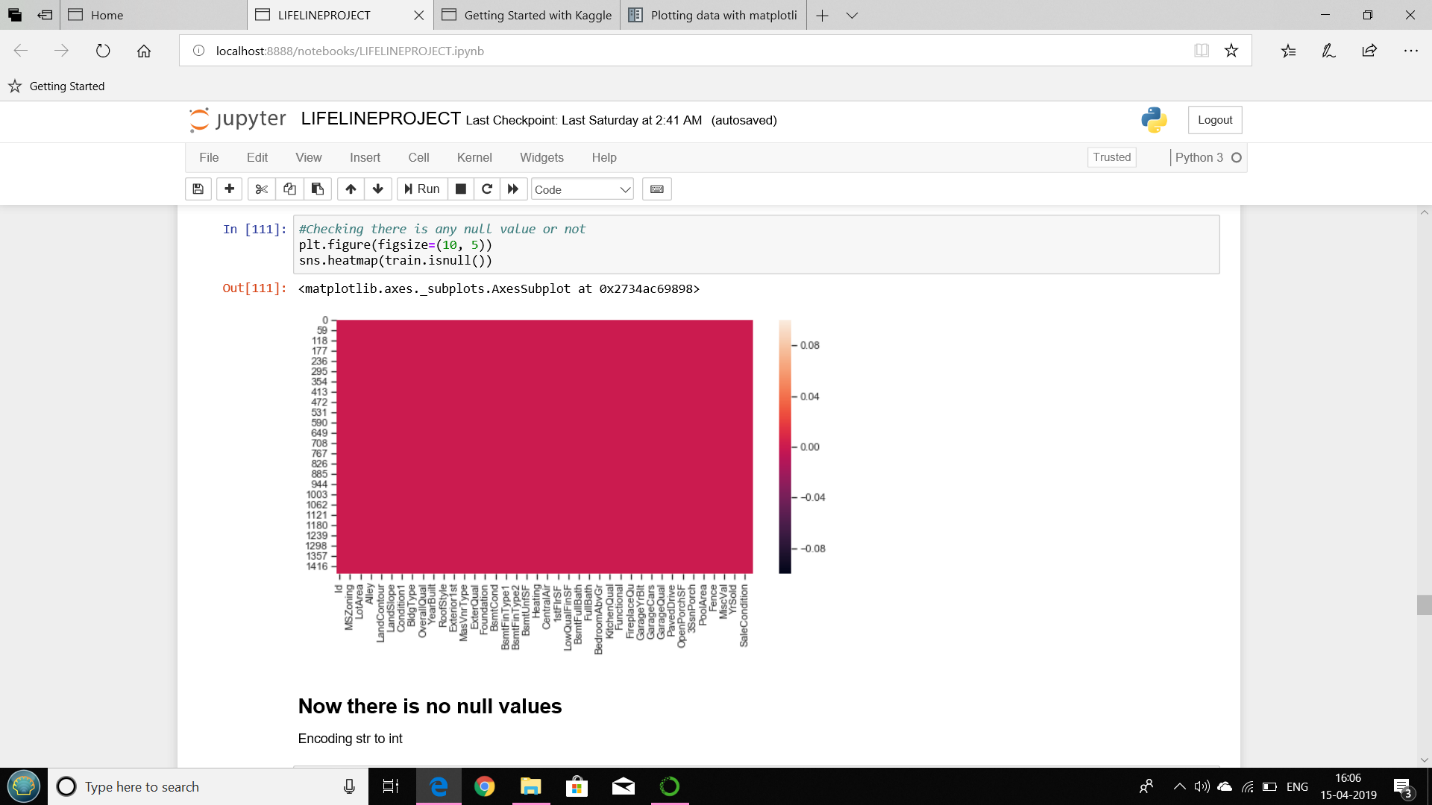
PLOTTING PAIRWISE:





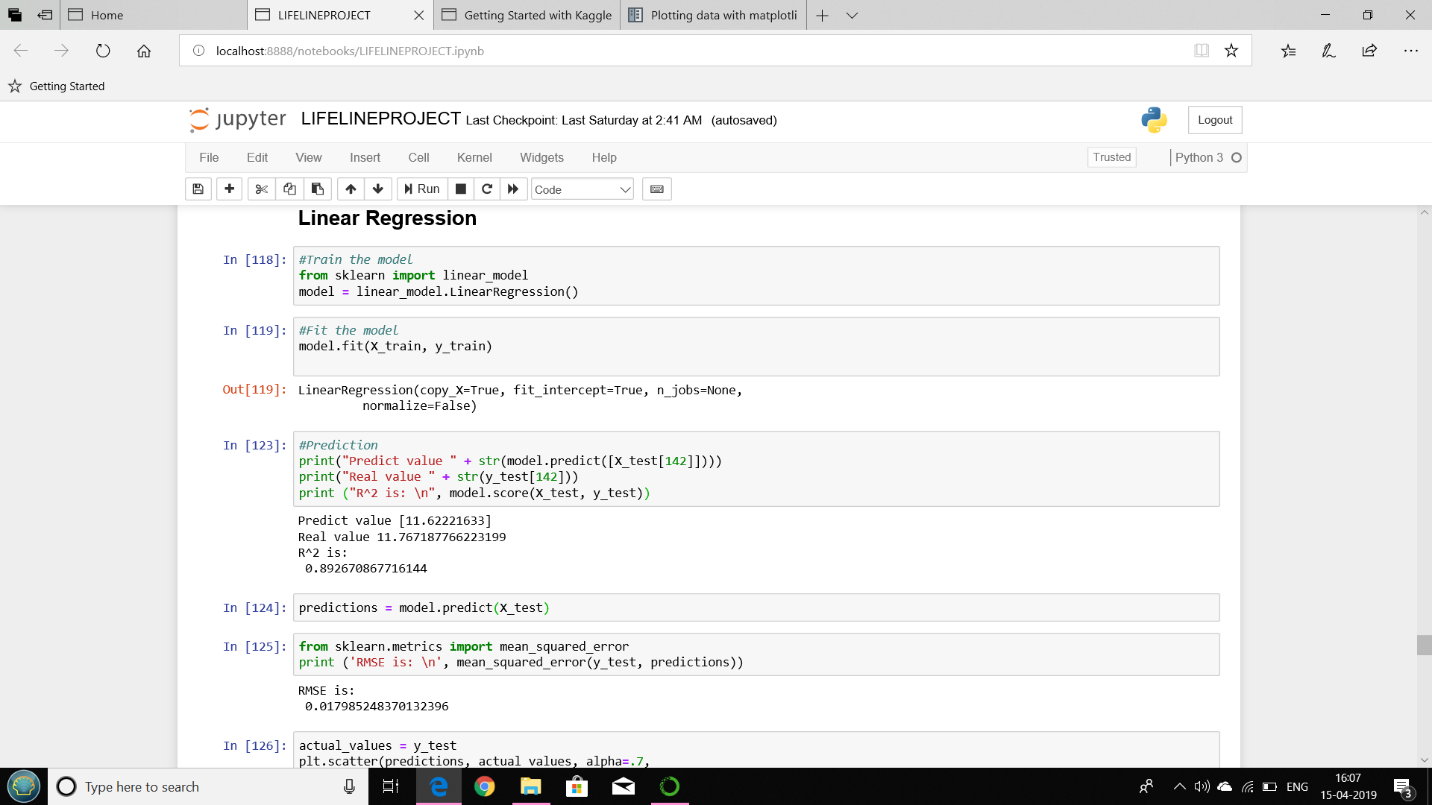


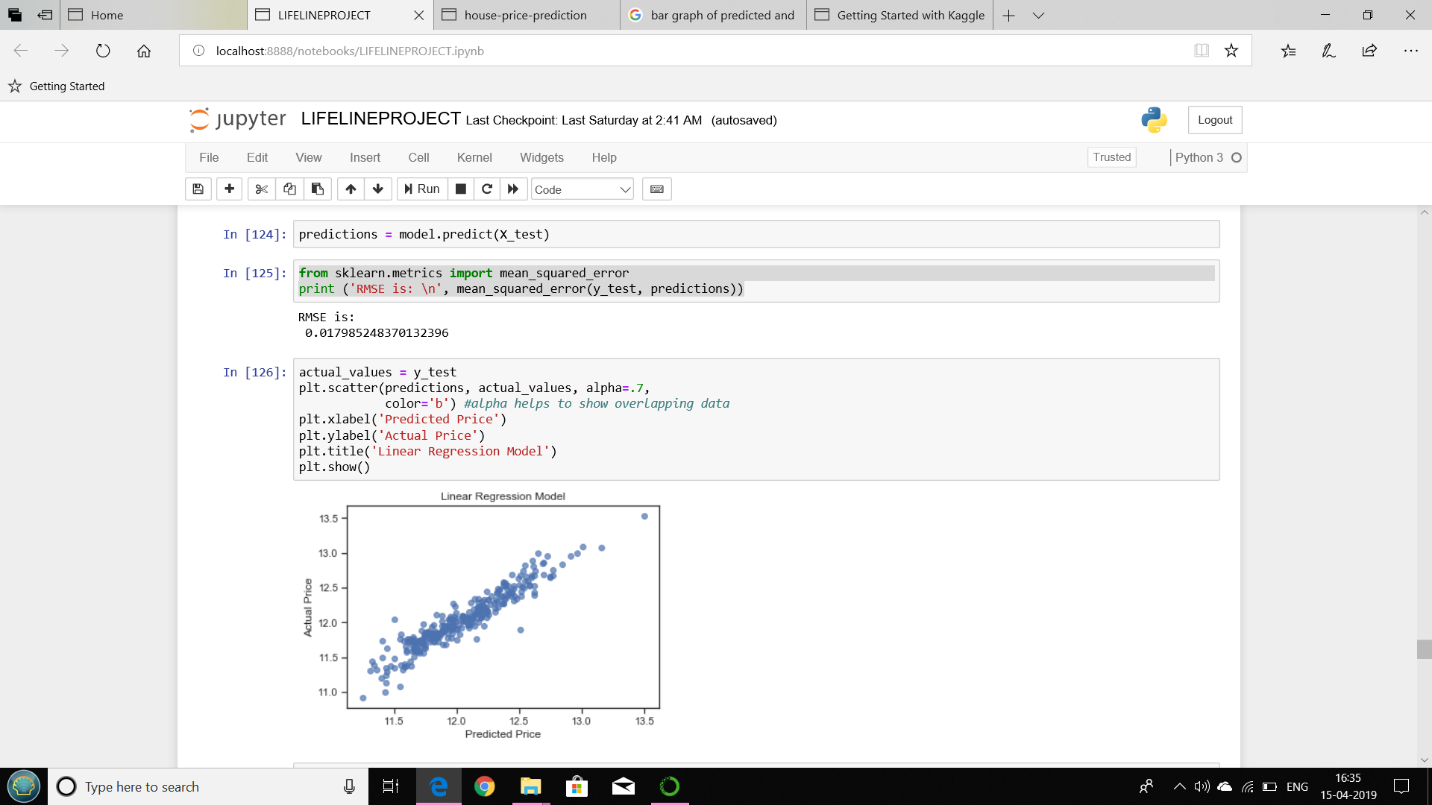
Eliminating null values:



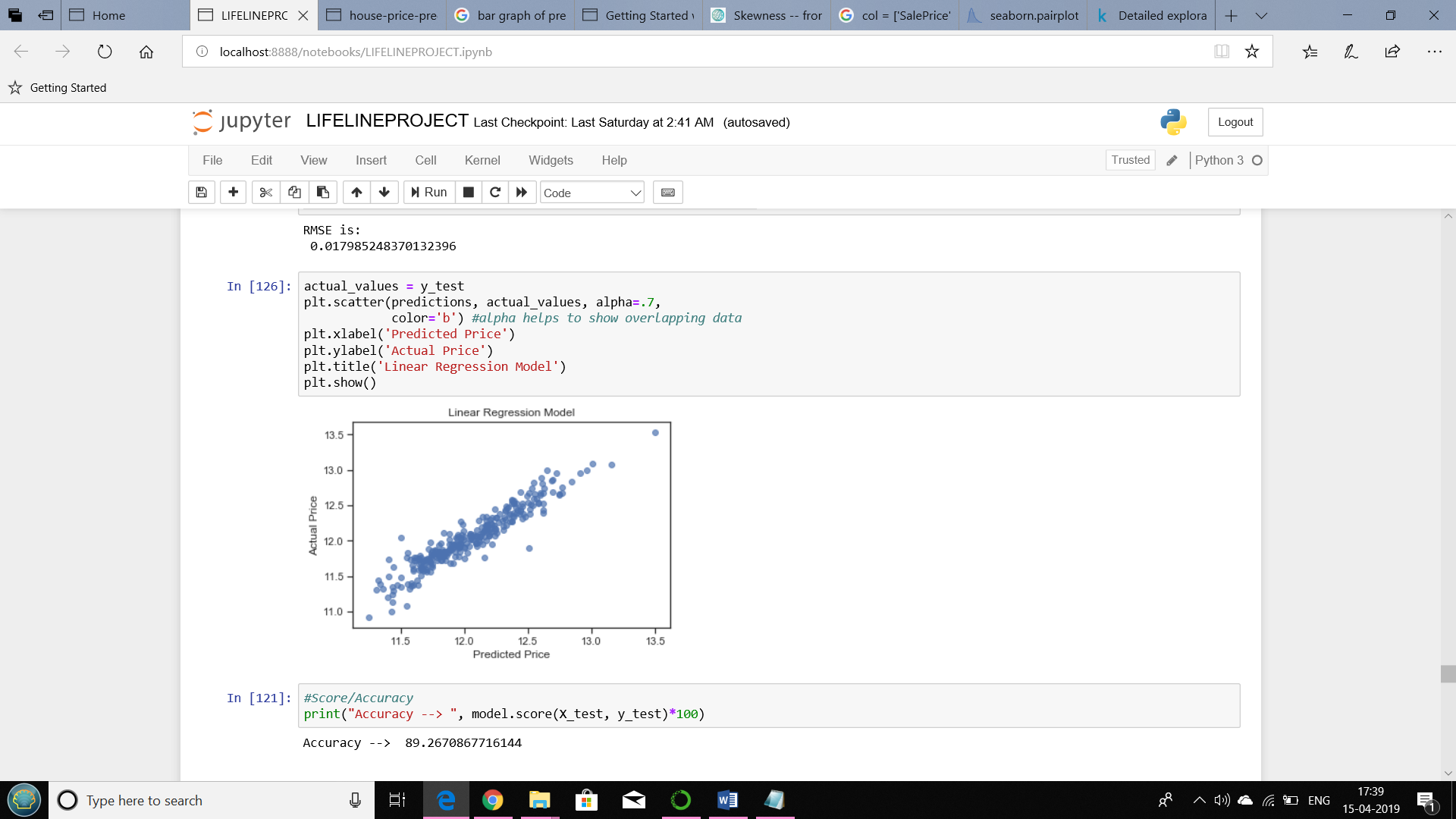
MODEL IMPLEMENTATION:

LINEAR REGRESSION RESULT:



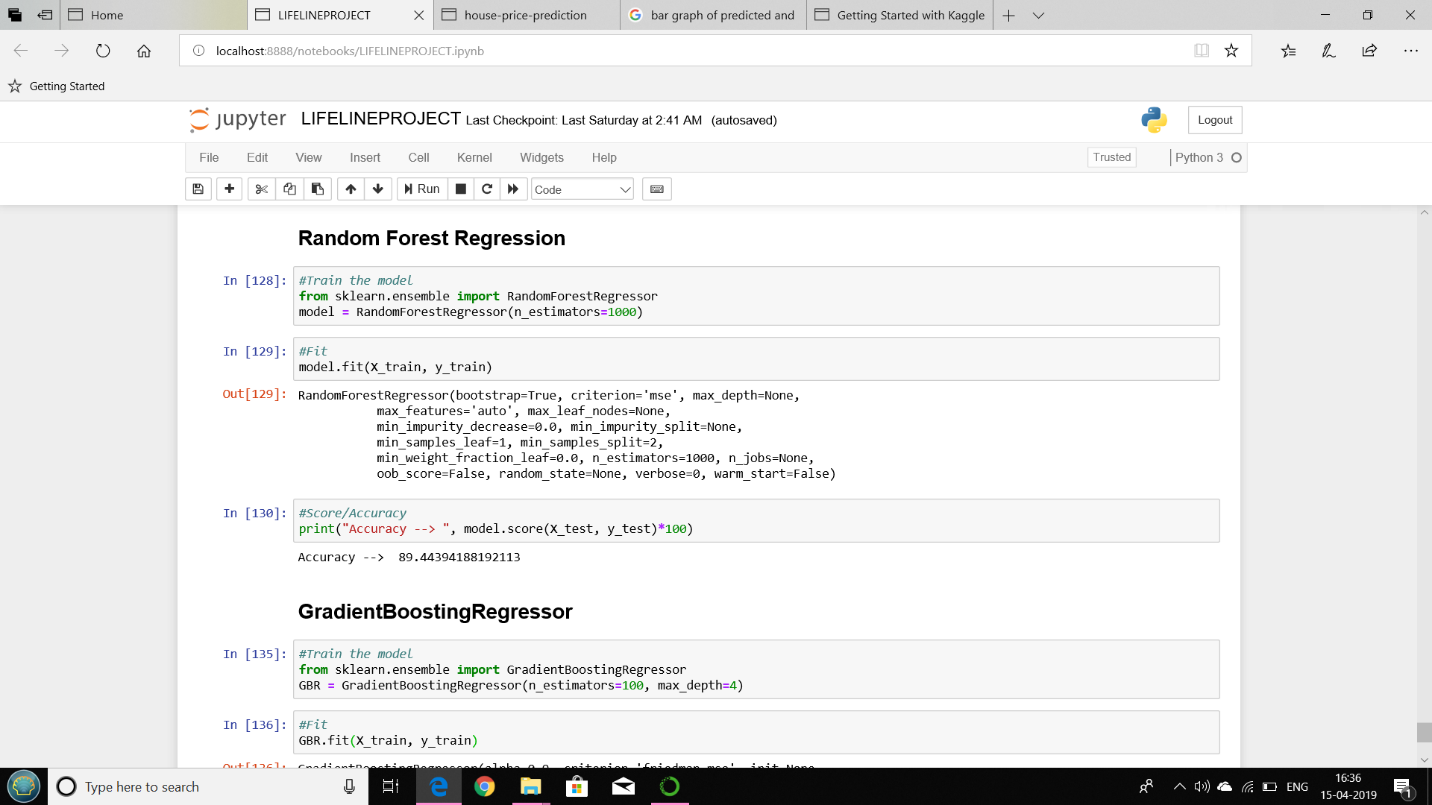


ACCURACY OF THE MODEL:



RANDOM FOREST REGRESSION:

ACCURACY OF THE MODEL:



## 4.5 Conclusion and Future Scope

We have managed to develop a system that provides users with a novel way to look at future housing price predictions. The regression methods have been explored and compared, before arriving at a pre-diction method. The model have been, so that future price predictions will tend towards more sensible values. We devised a way to utilise as much data as possible in our prediction system. The success of our approach to creating a system for generating predictions can be applied to other problem sets concerned with geographical variations in the prediction model.

Despite having produced a system that met our initial requirements, there are various improvements that can be made in the future.

The following factors can be considered to expand the project in the right direction in the future :-

•Consider more factors affecting housing prices :-

One main drawback of our prediction model is the lack of access to information. We would need to seek alternative sources of data besides the Land Registry, which can provide us with information on the area, the number of rooms, etc for each property. Economic factors such as the yearly inflation rate or GDP growth can also be considered.

•Consider heteroscedastic GP regression :-

Under the conventional GP regression method that we are using in our prediction method, the noise level is assumed to be uniform throughout the domain. How-ever, this assumption may not be true. By considering a heteroscedastic treatment of noise, we can treat noise as a random variable. This can potentially better account for the difference in data points across the years in our dataset. An alternative noise model can also be explored, by training the data against the logarithm of the prices.

•Optimise the prediction system through parallelised computations :-

A major concern with the prediction system is the loading time. Moreover, our dataset takes more than one dayto train. Rather than performing the computations sequentially, we can use multiple processors and parallelise the computations involved, which can potentially reduce the training time and prediction time. Another way to approach this problem is to look for alternative APIs that allow us to produce heat maps of similar quality but require fewer data points, or faster libraries to implement GPs.

•Add more functionalities into the application :-

We can provide options for user to select a borough or district to generate the heat maps, instead of entering a postcode and selecting a search radius. We can also look into ways to generate a heat map for the whole of initially in a zoomed-out view, only requesting for more data points when the user zooms in to display a clearer picture.

## 4.6 References

* Housing Price Prediction Using Neural Networks -Wan Teng Lim, Lipo Wang, Yaoli Wang, and Qing Chang
* Prediction of Real Estate Price Variation Based on Economic Parameters-LiLi and Kai-HsuanChu
* IJARCCE- House Price Forecasting

|  |  |
| --- | --- |
| |  | | --- | | * <https://pdfs.semanticscholar.org/782d/3fdf15f5ff99d5fb6acafb61ed8e1c60fab8.pdf> | |

* <https://pub.tik.ee.ethz.ch/students/2017-FS/BA-2017-06.pdf>
* <https://ijarcce.com/upload/2017/december-17/IJARCCE%2016.pdf>
* <https://nycdatascience.com/blog/student-works/housing-price-prediction-using-advanced-regression-analysis/>