

The Housing Project

Submitted by:

ABHISHEK PAI

**ACKNOWLEDGMENT**

The Dataset and related information were provided by Flip Robo as part internship phase. This entire file package included a dataset with is column understanding to help with project analysis. Furthermore, guidance was provided by staff mentors in Flip Robo through doubt-solving sessions and tickets. Few code-related help and structure-related understanding was been taken from sources available on the internet.

**INTRODUCTION**

* Business Problem Framing

Required 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.

* 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

There is not much research performed as the Data and related information was provided by the source itself, which was been taken into consideration based on the information given by Flip Robo.

* Motivation for the Problem Undertaken

The Project was assigned by flip Robo as part of the internship phase for better understanding the concept and getting the idea of the industry.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

After importing data various analyses were performed which had univariate, bivariate, and multivariate analysis.

Univariate analysis:Univariate analysis is the simplest form of analyzing data. It doesn't deal with causes or relationships and its major purpose is to describe; It takes data, summarizes that data, and finds patterns in the data.

Bivariate anslysis:Bivariate analysis is one of the simplest forms of quantitative analysis. It involves the analysis of two variables, to determine the empirical relationship between them. Bivariate analysis can help test simple hypotheses of association.

Multivariate analysis: Multivariate statistics is a subdivision of statistics encompassing the simultaneous observation and analysis of more than one outcome variable. Multivariate statistics concerns understanding the different aims and background of each of the different forms of multivariate analysis, and how they relate to each other.

Outliers and Skewness: Outlier Analysis is a process that involves identifying the anomalous observation in the dataset.Skewness measures the deviation of a random variable's given distribution from the normal distribution, which is symmetrical on both sides. Both were done to bring data in a good shape

* Data Sources and their formats

After loading the data, the information of data was been checked and a five-row sample was been observed.

Later after performing analysis on the dataset, the Outlier and skewness of data was been removed.

Feature engineering was been performed based on the requirement.

Some features were not necessary for ML and were dropped before model building

* Data Pre-processing Done

The entire data was in form of CSV and was a mixture of numbers, and objects. The output variable is information in a pattern of 1 and 0 each having its significant meaning. The output was based on the data which was provided by a source on the behavioral pattern of the entity. The object part was been converted and extracted to perform ML

* Hardware and Software Requirements and Tools Used

The system with a 16 core processor was been used,

The operating system was Windows 10,

Anaconda 3 was been used for performing ML

Libraries: import numpy as np

import pandas as pd

import seaborn as sns

matplotlib

matplotlib.pyplot as plt

plotly.express as px

sklearn.metrics import classification\_report

scipy.stats import skew

sklearn.preprocessing import power\_transform

sklearn.ensemble import RandomForestClassifier

sklearn.ensemble import XGBClassifier

sklearn.ensemble import GradientBoostingClassifier

sklearn.tree import DecisionTreeClassifier

sklearn.metrics import accuracy\_score

sklearn.preprocessing import power\_transform

sklearn import metrics

sklearn.metrics import roc\_auc\_score

sklearn.metrics import accuracy\_score,confusion\_matrix,f1\_score,mean\_squared\_error,roc\_curve,precision\_recall\_curve

sklearn.model\_selection import train\_test\_split

warnings

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Performing univariate bivariate and multivariate analysis helped a lot to understand the nature of data-based, also performing outliers check and skewness helped to shape data, based that the data was scaled and then principal component was performed.

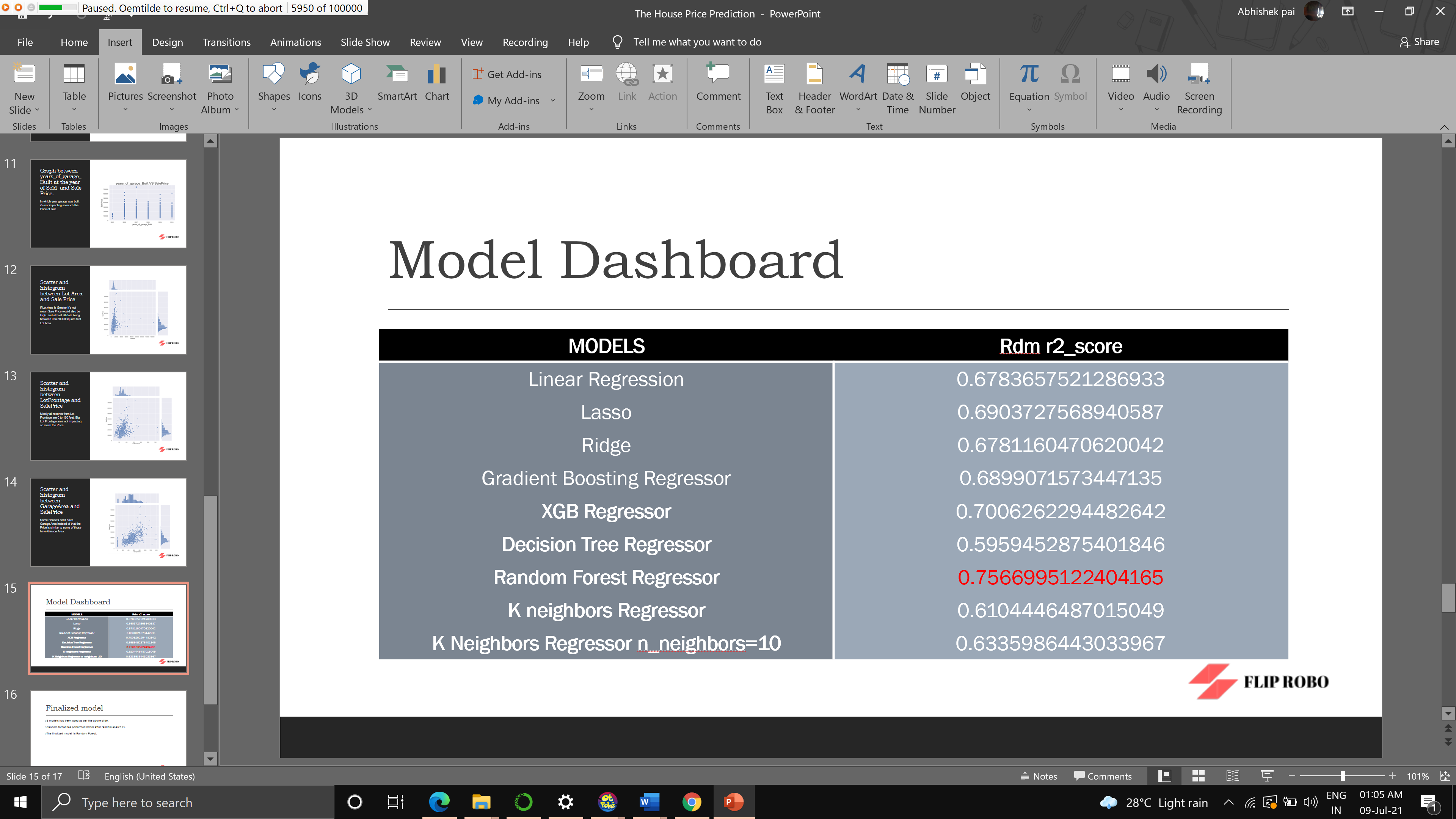
.

* Testing of Identified Approaches (Algorithms)

|  |
| --- |
| Linear Regression |
| Lasso |
| Ridge |
| Gradient Boosting Regressor |
| **XGB Regressor** |
| **Decision Tree Regressor** |
| **Random Forest Regressor** |
| **K neighbors Regressor** |

.

* Run and Evaluate selected models

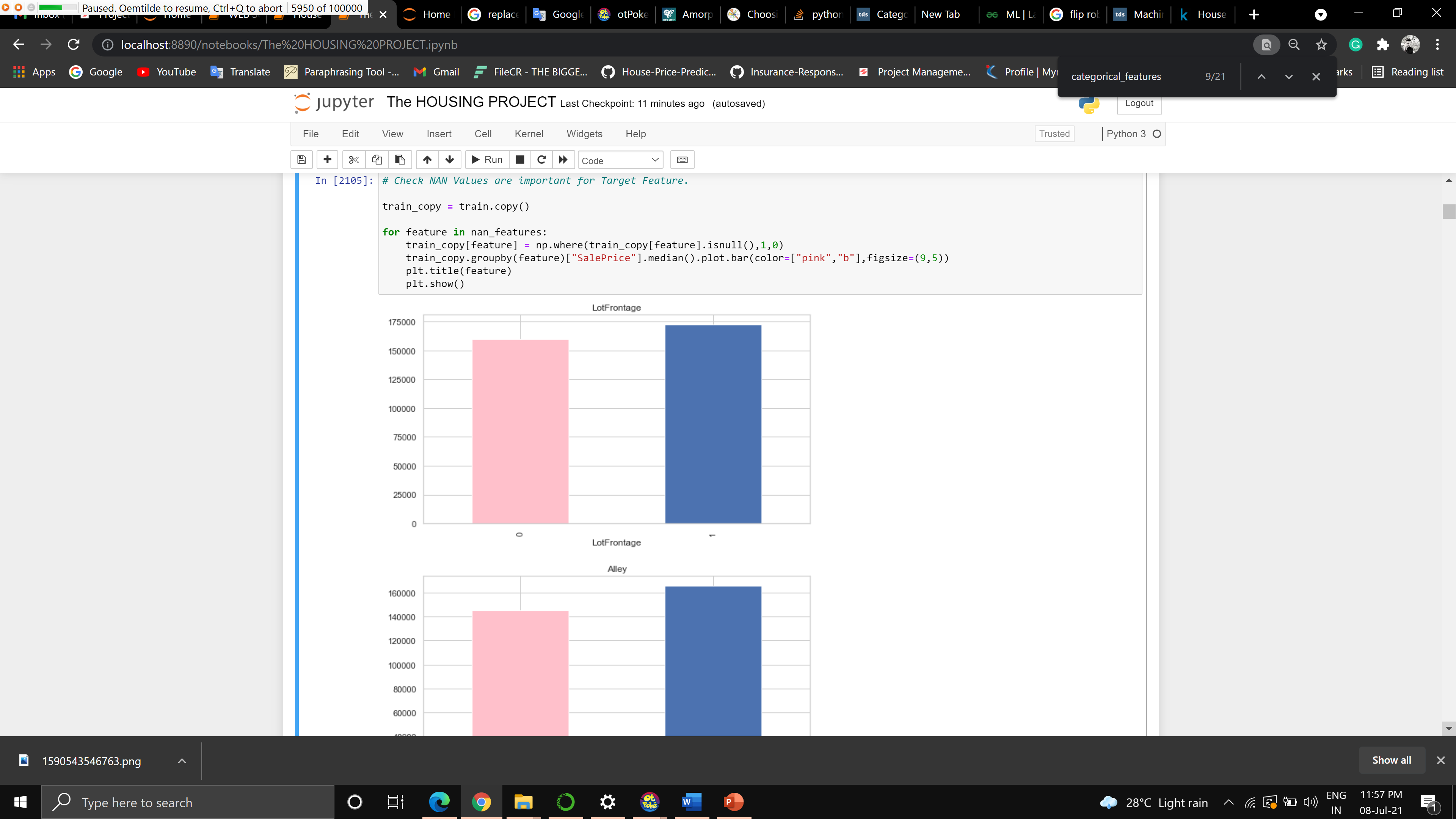


* Key Metrics for success in solving problem under consideration

**R2 score** is **used** to evaluate the performance of a regression model. It is the amount of the variation in the output dependent attribute which is predictable from the input independent variable.

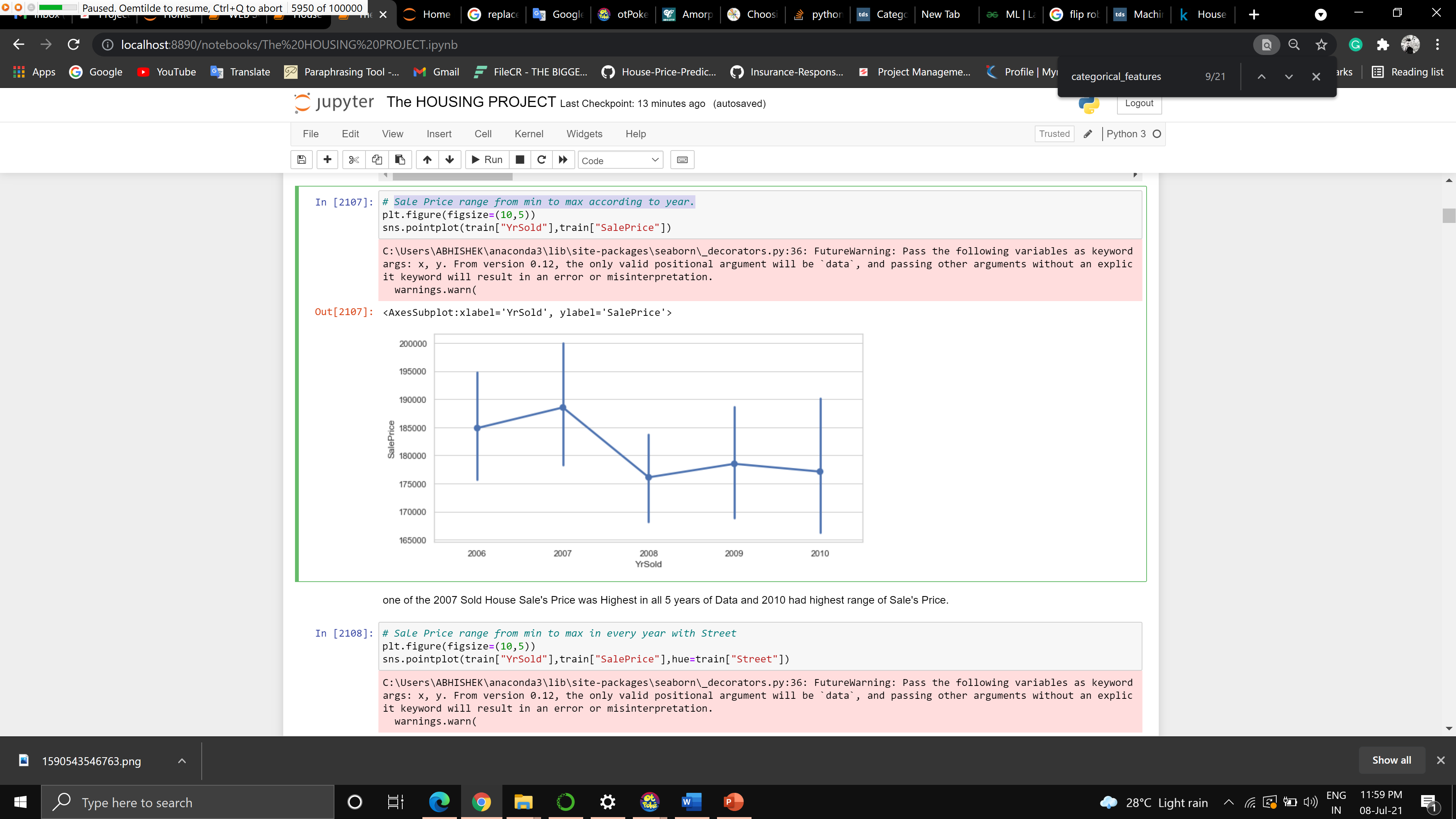
* Visualizations

NAN Values are important for Target Feature



Clearly visible that missing value has also Impact on Dependent variable, so we need to replace these nan value with something meaningful

Sale Price range from min to max according to year



One of the 2007 Sold House Sale's Price was Highest in all 5 years of Data and 2010 had highest range of Sale's Price.

* Interpretation of the Results

8 models has been used as per the above slide .

Random forest has performed better after random search cv.

The finalized model is Random Forest.

.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Houses with a good overall condition, with a greater living room which is above ground along with a greater basement area are expensive houses.

People are preferring those houses that has furnished basement with a greater area with a garage having capacity of storing more cars.

It seems that most of the people prefer to buy two story houses, with greater floor area.

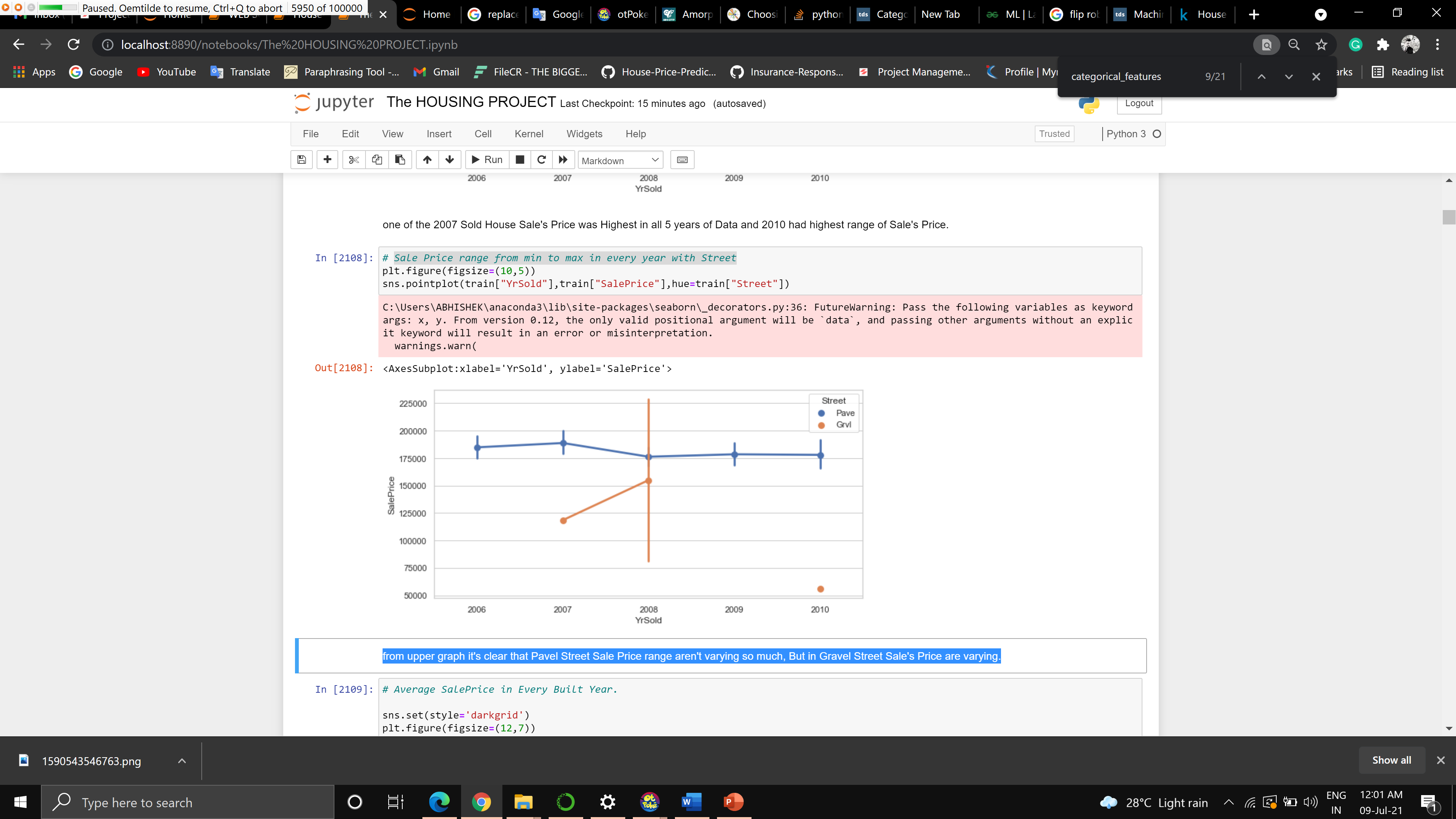
Big houses or a mansions with a greater area are even more expensive.

It seems that people are preferring the houses that has been recently remodeled or reconstructed.

* Learning Outcomes of the Study in respect of Data Science

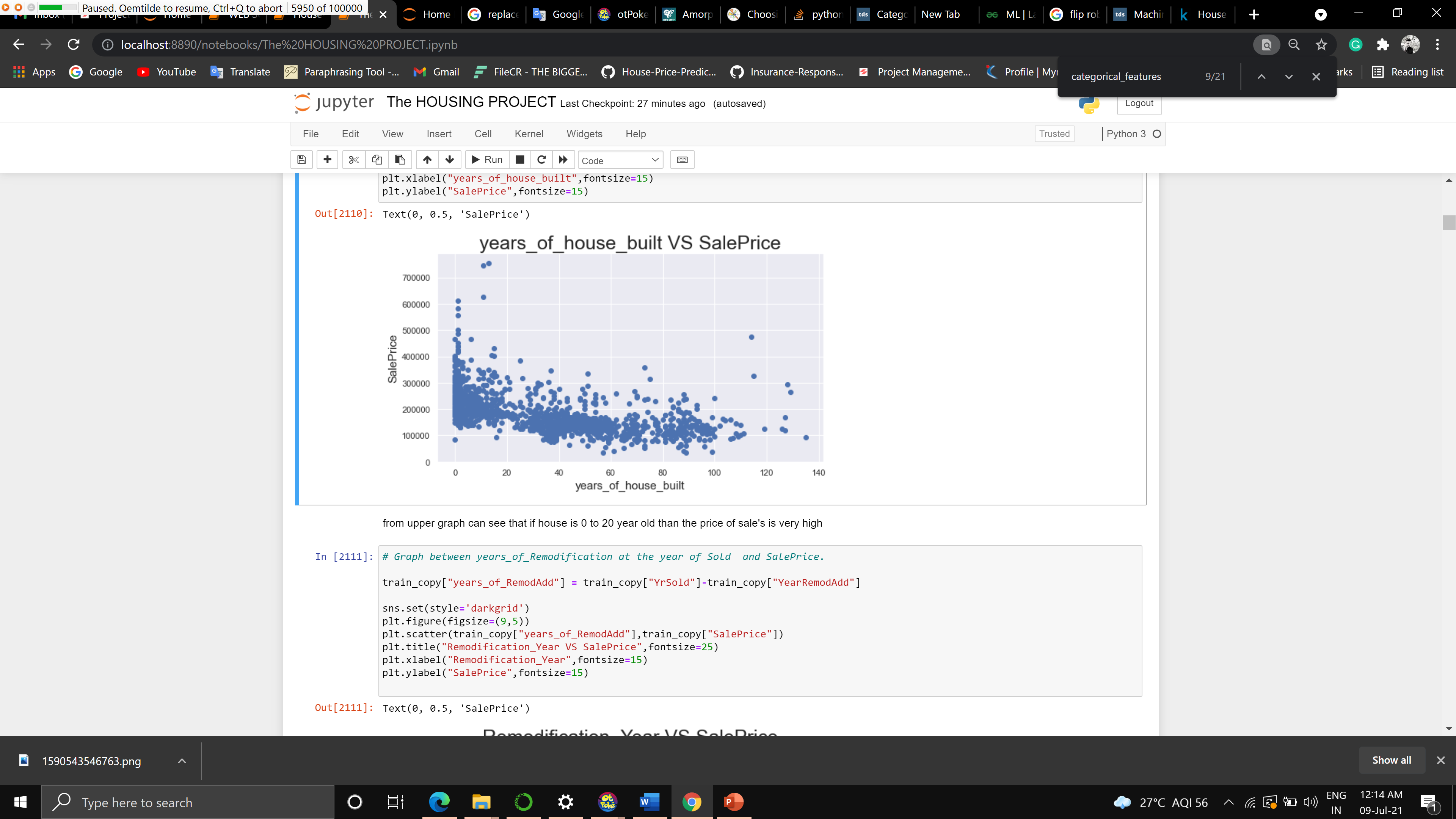
List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

Sale Price range from min to max in every year with Street



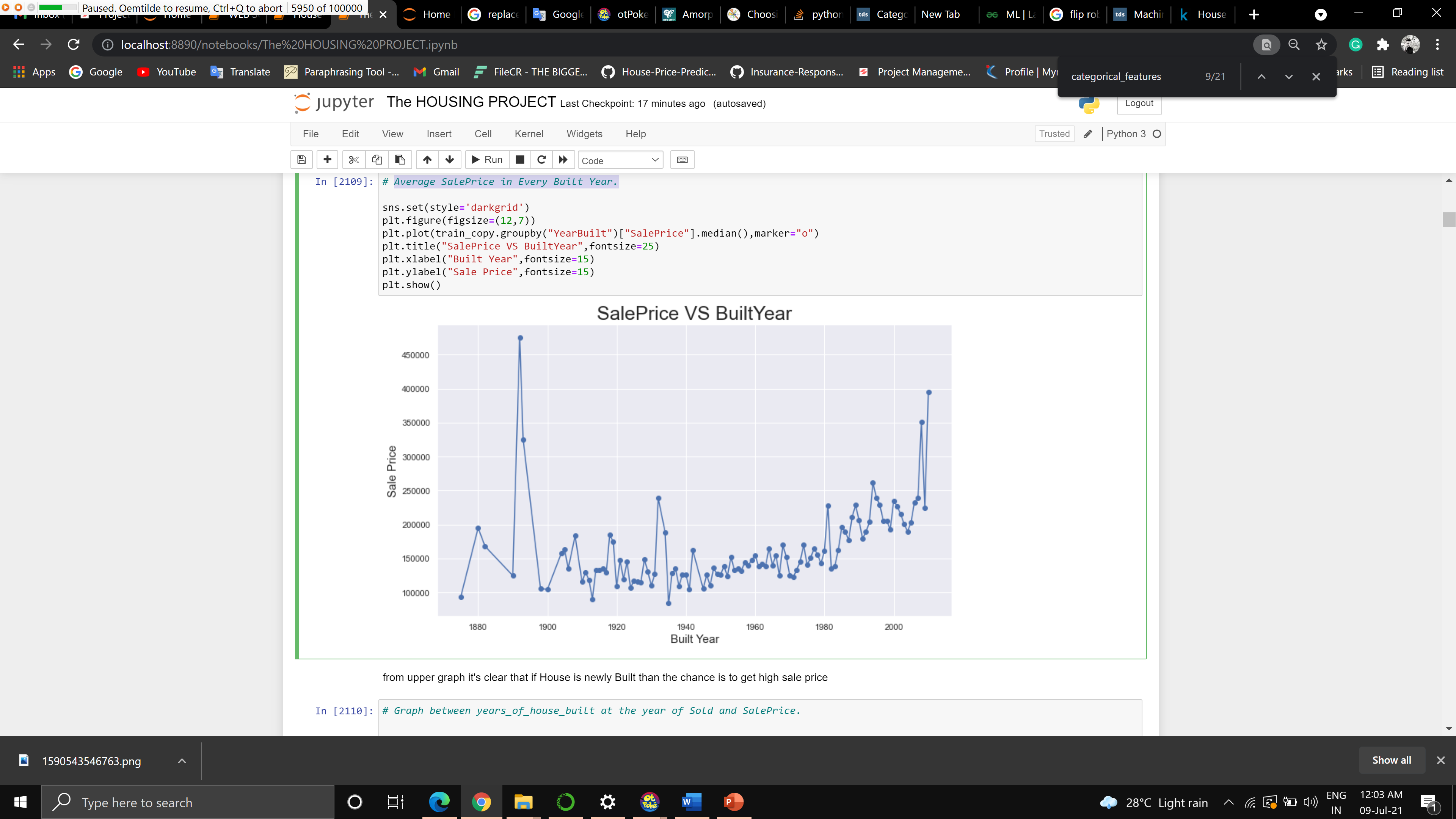
Pavel Street Sale Price range aren't varying so much, But in Gravel Street Sale's Price are varying.

Graph between years\_of\_house\_built at the year of Sold and Sale Price



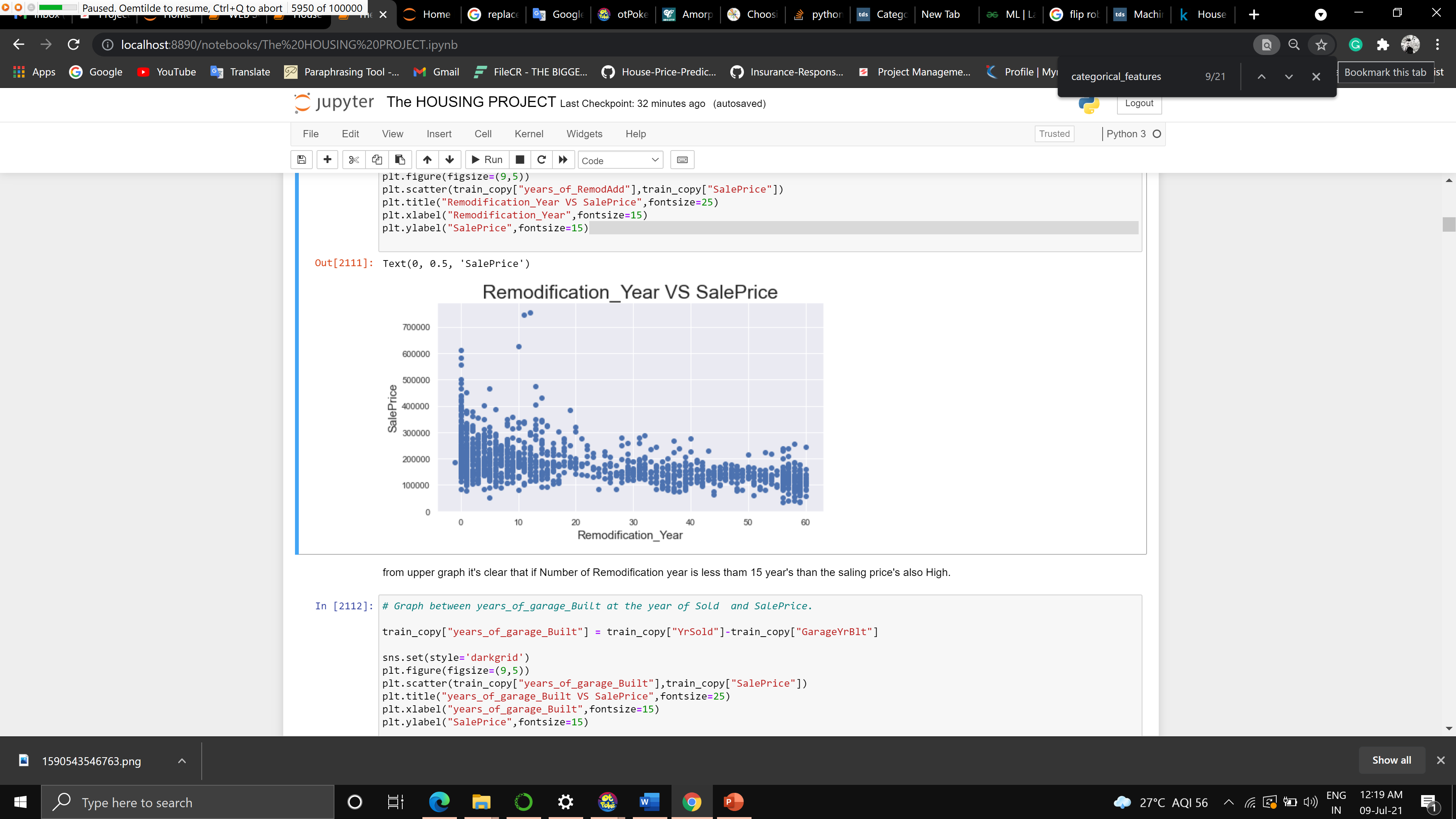
if house is 0- to 20-year-old than the price of sale's is very high

Average Sale Price in Every Built Year



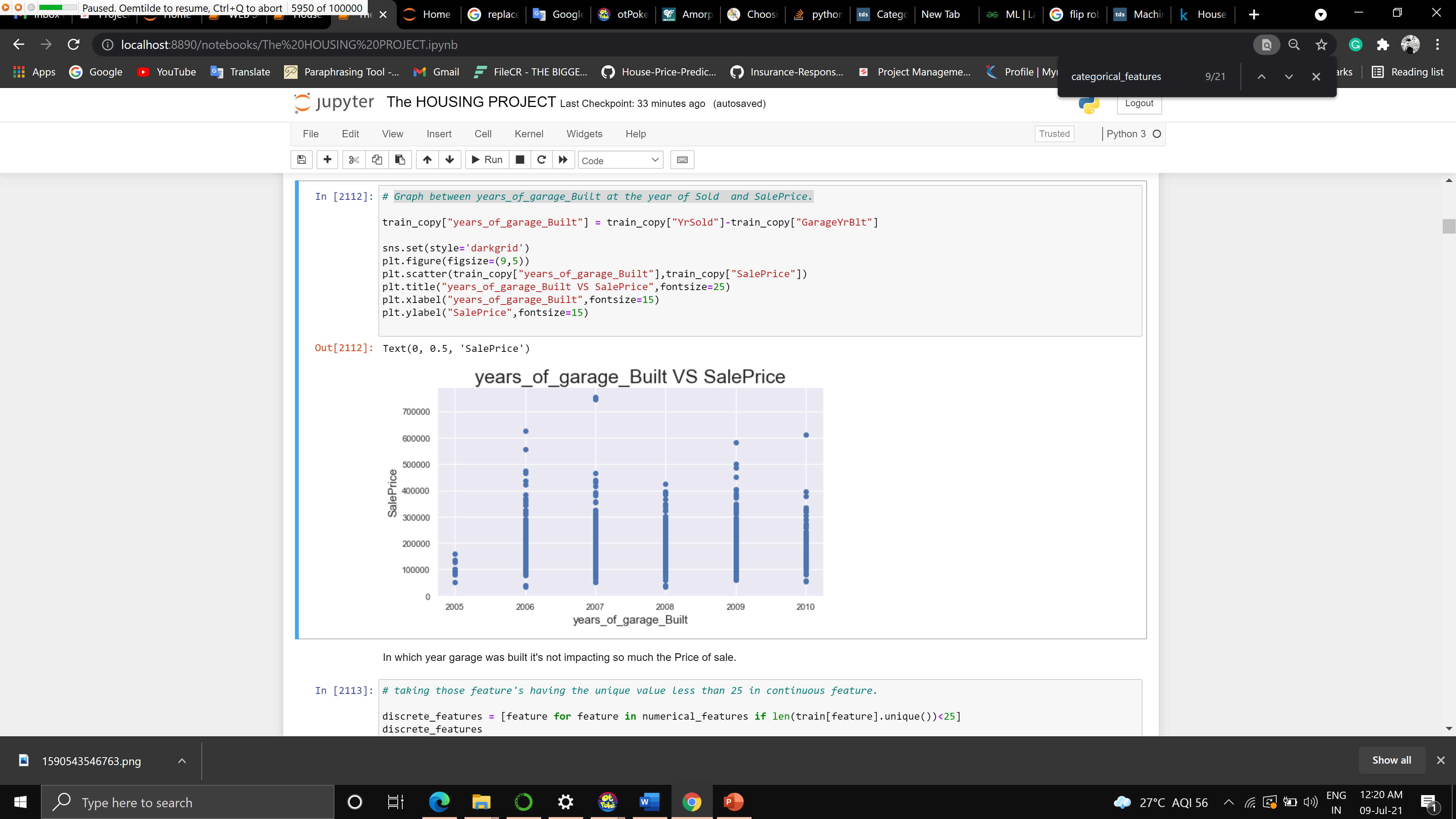
If House is newly Built than the chance is to get high sale price

Graph between years\_of\_Remodification at the year of Sold and Sale Price



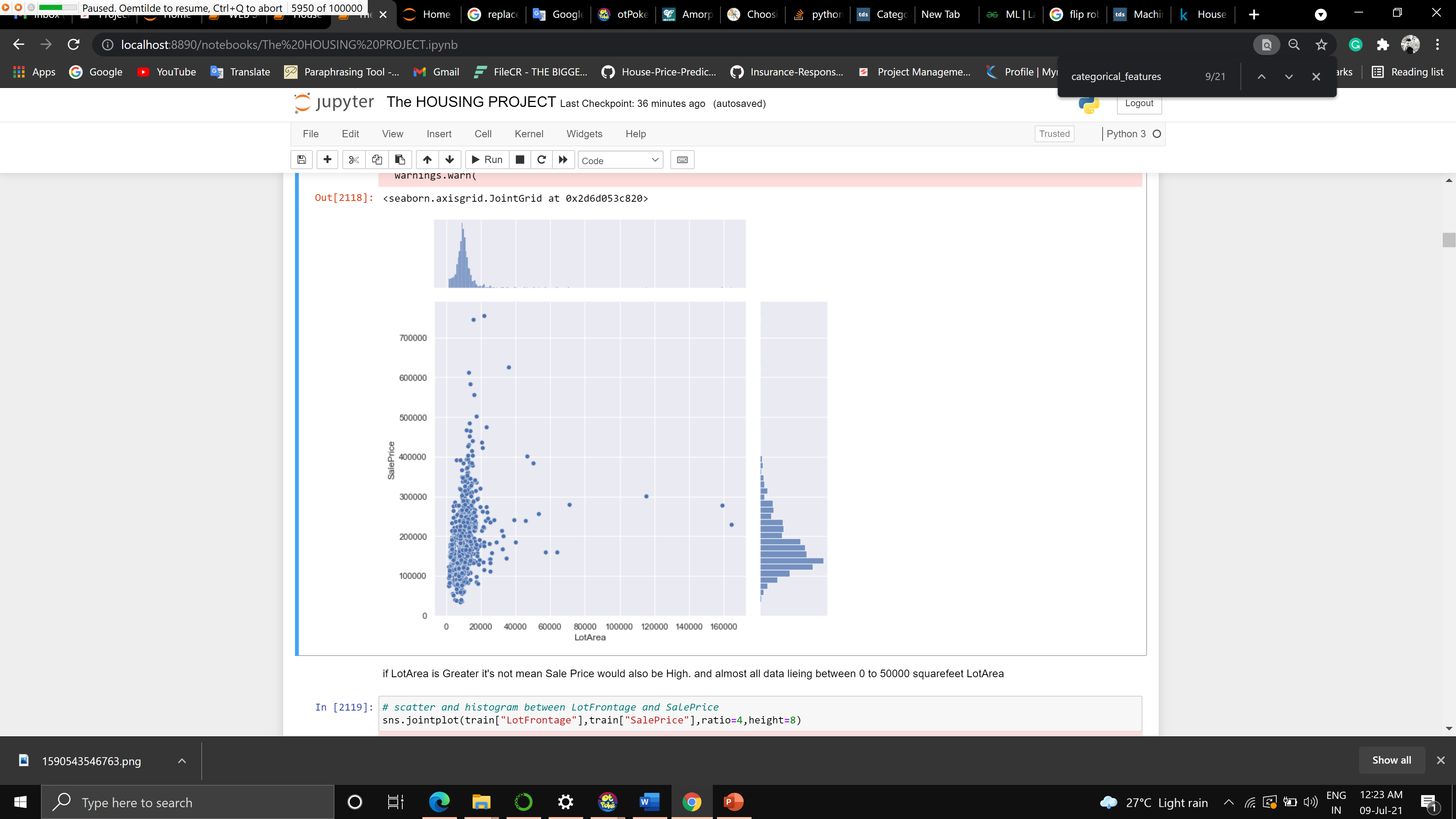
Number of Remodification year is less them 15 year's than the selling price's also High.

Graph between years\_of\_garage\_Built at the year of Sold and Sale Price



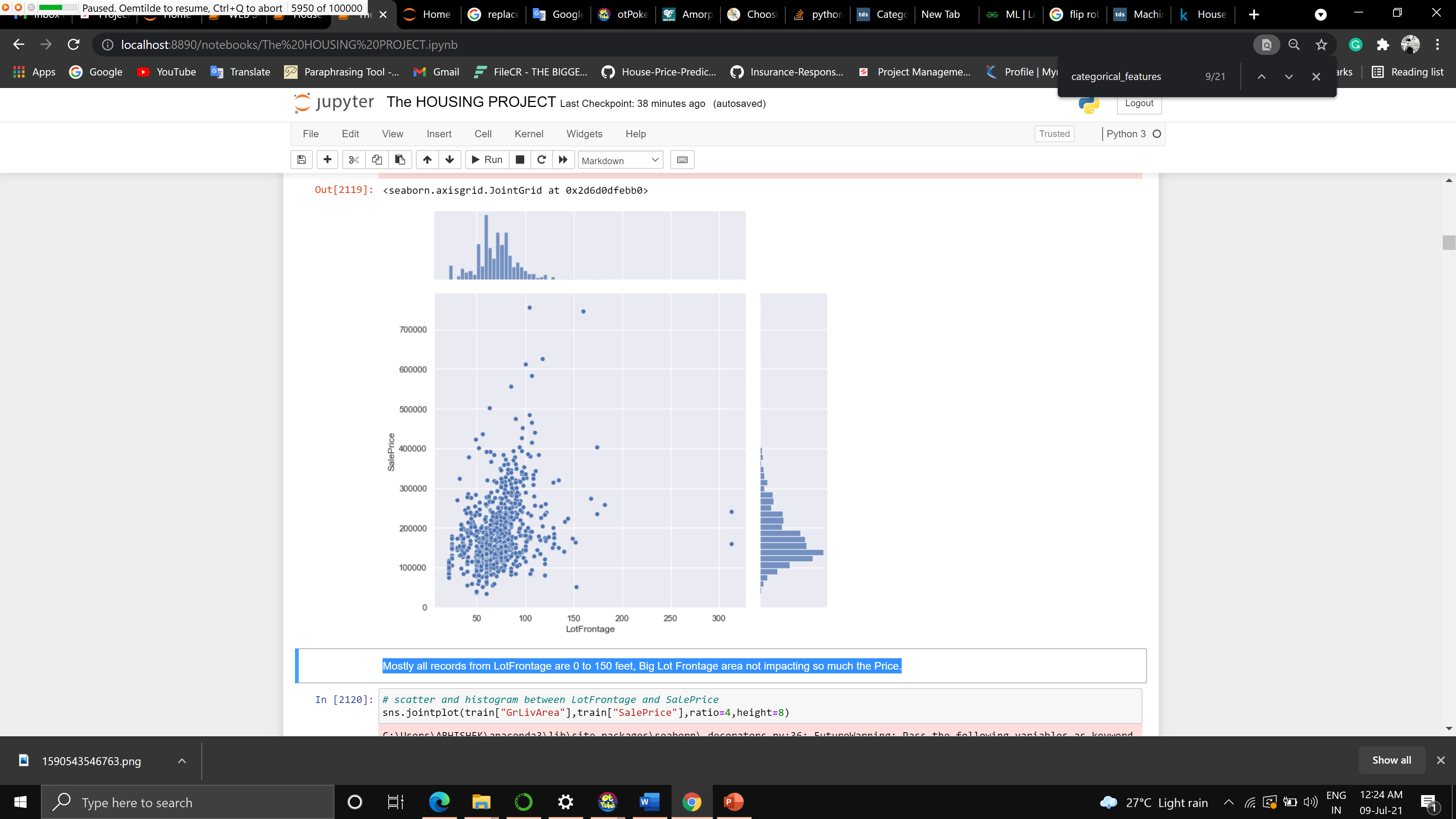
In which year garage was built it's not impacting so much the Price of sale.

Scatter and histogram between Lot Area and Sale Price



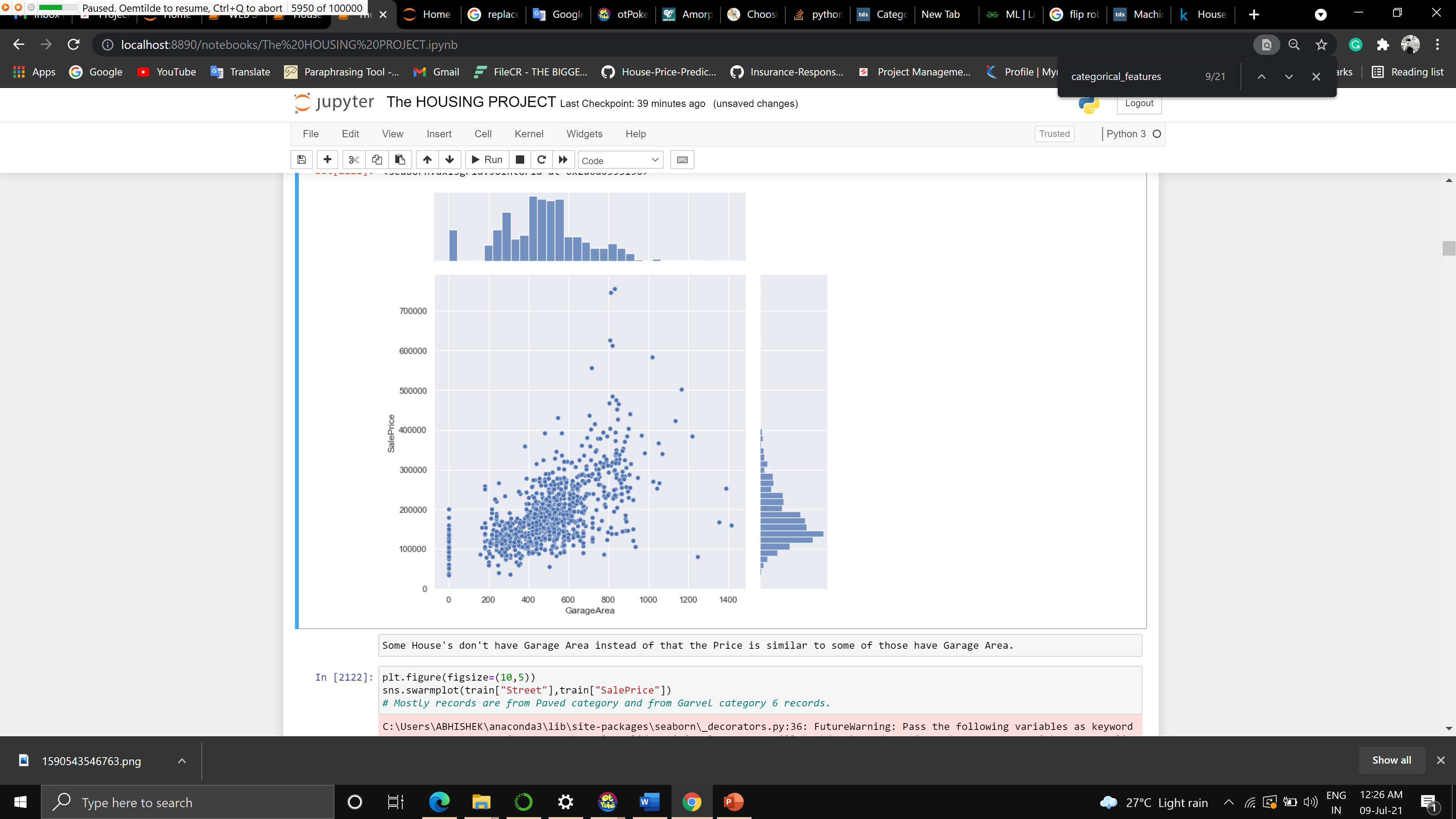
if Lot Area is Greater, it's not mean Sale Price would also be High. and almost all data lying between 0 to 50000 square feet Lot Area

Scatter and histogram between LotFrontage and SalePrice



Mostly all records from Lot Frontage are 0 to 150 feet, Big Lot Frontage area not impacting so much the Price.

Scatter and histogram between Garage Area and Sale Price



Some House's don't have Garage Area instead of that the Price is similar to some of those have Garage Area.

* Limitations of this work and Scope for Future Work

There is a lot of scope, more tweaks in model can help to get better results.