**ASSIGNMENT**

LAB02 (B)

LILIYA PONNU SHAJI

21BDA38

(B) In a word file,

1.Document 5-6 key insights from EDA and support each point with a visualization.

2.Answer the following questions:

i.What are the assumptions of linear regression?

ii.How can we evaluate a Regression model? Define each metric and its interpretation.

iii.Can R squared be negative?

iv.What is dummy variable trap?

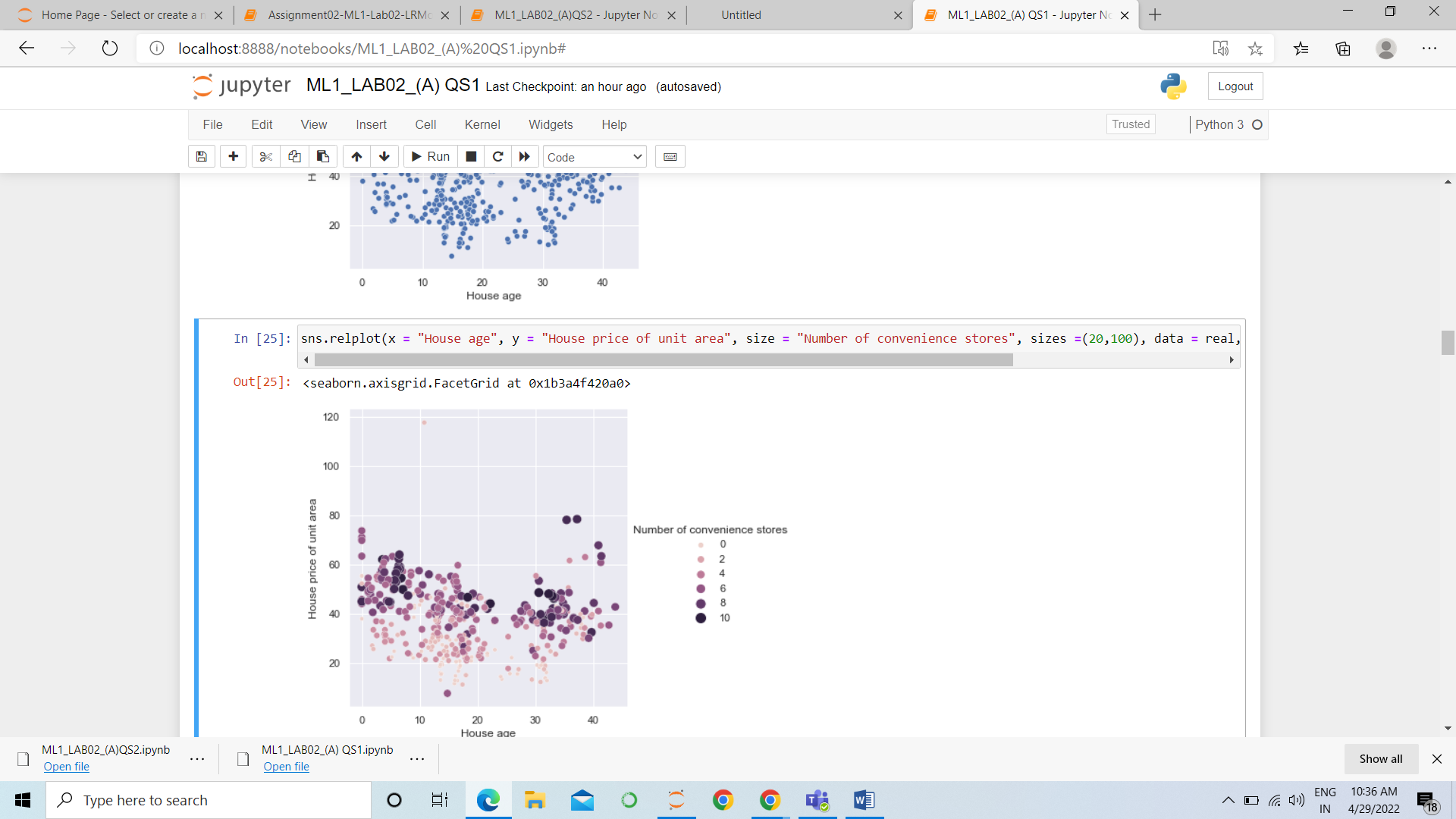
v.Is One Hot Encoding different from Dummy Variables?

vi.How is polynomial regression different from linear regression?

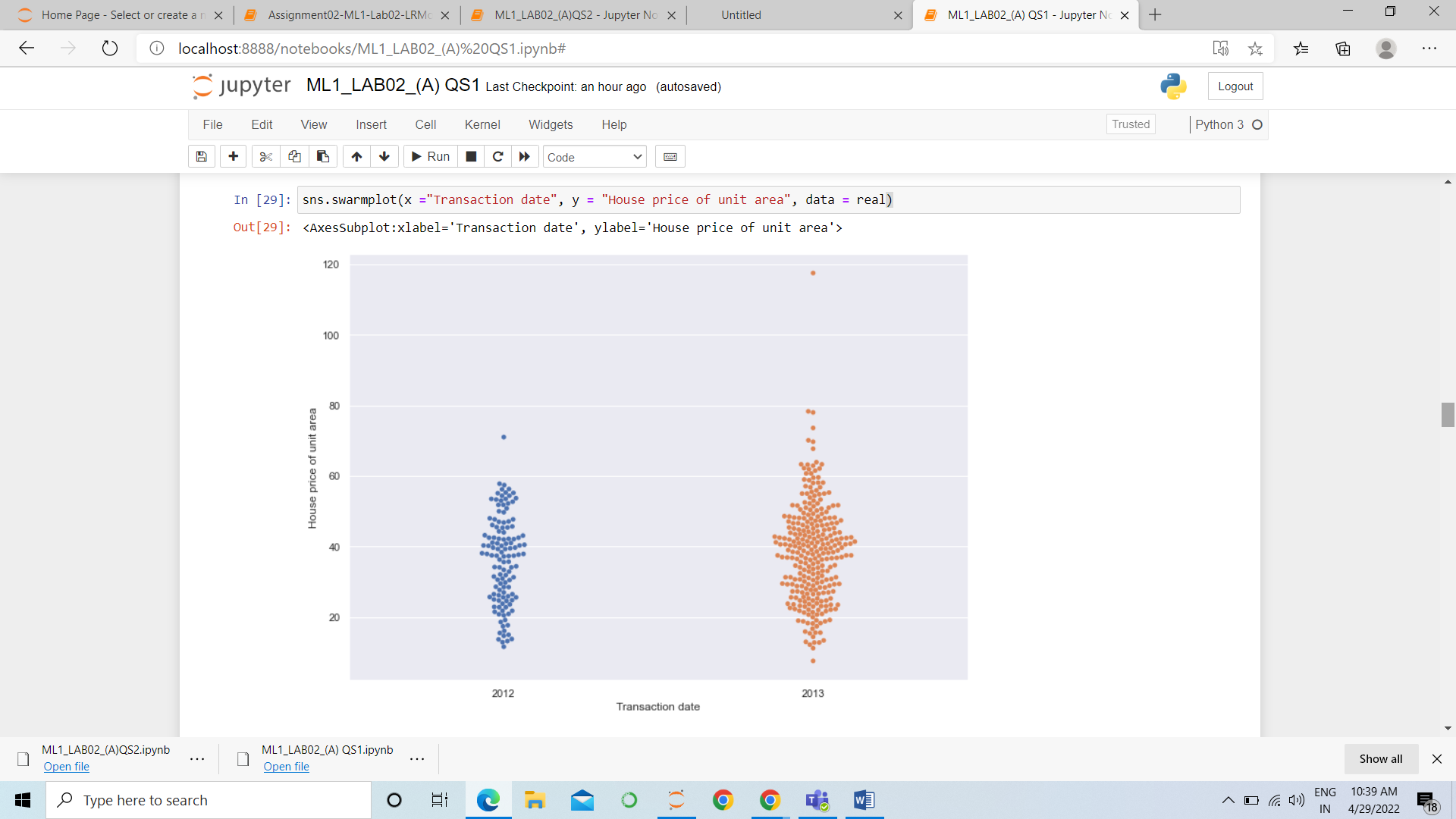
vii.Interpret the screenshot below from the notebook we discussed in class today:

Answers

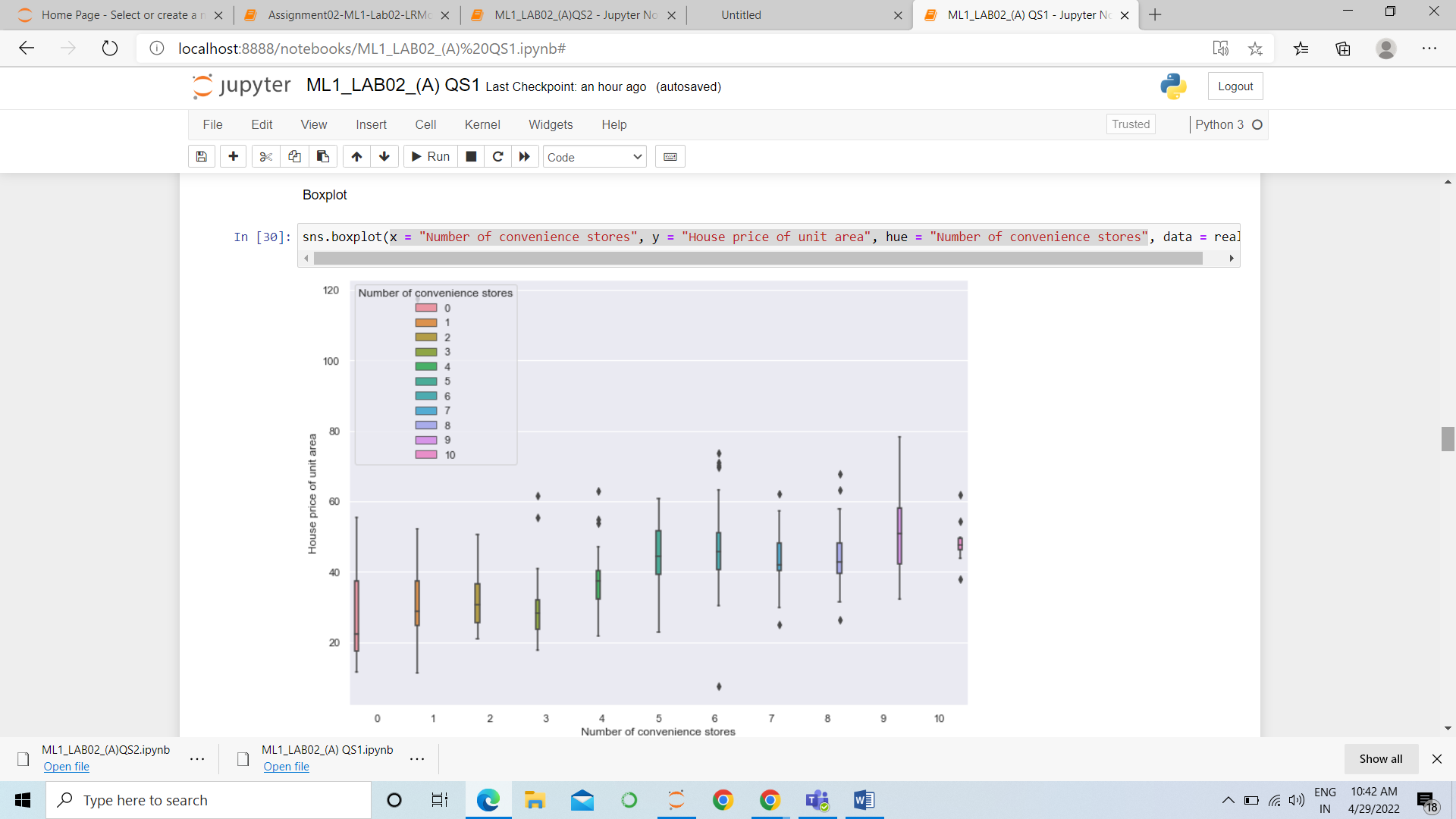
1.(a) House Price depends on House age



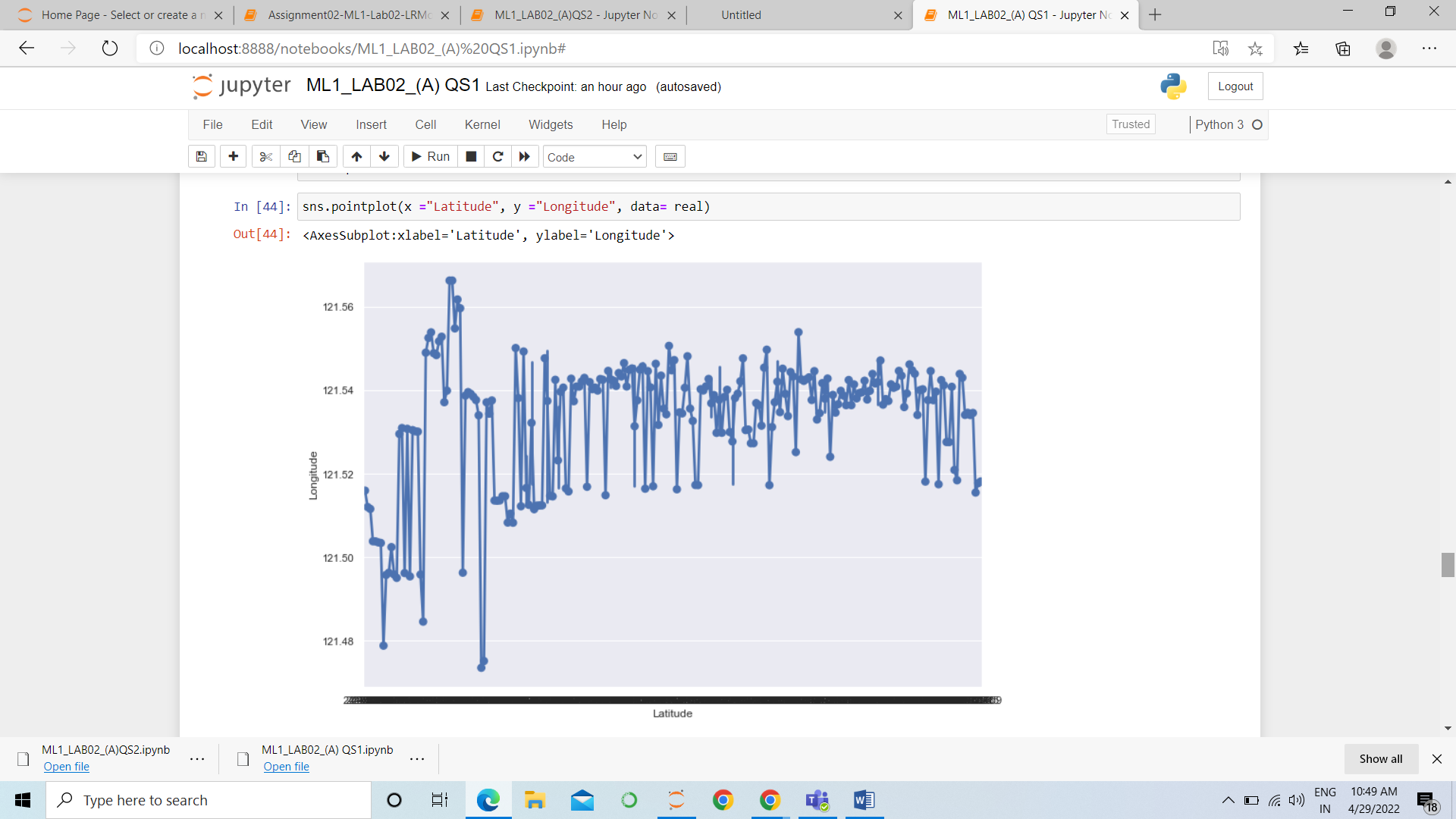
(b) From the figure below we can see that House price per unit area has been increased in the year 2013. Most of the data is concentrated at 2013.



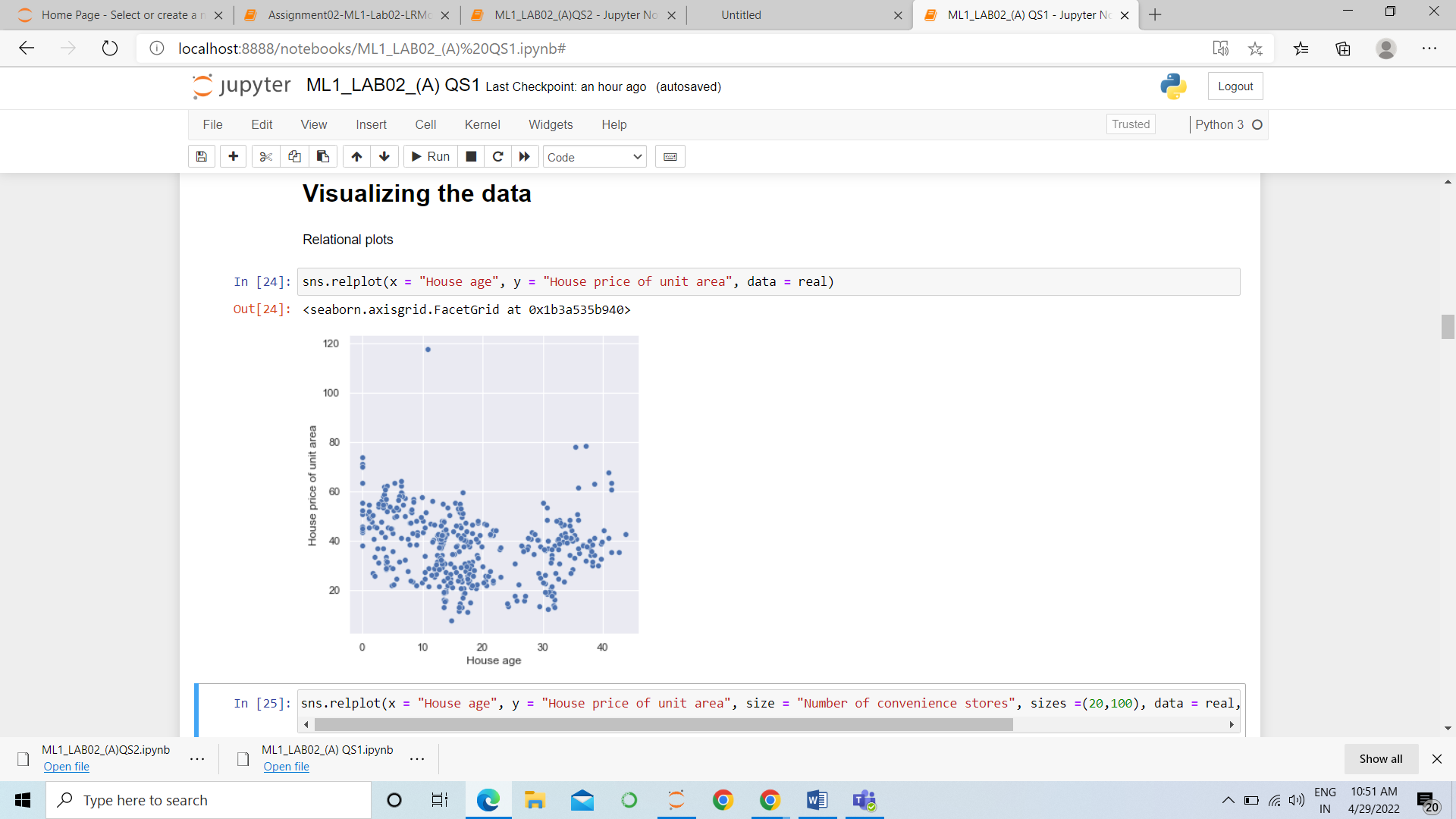
(c) From the plot below we can see that there are maximum of 10 convenience stores. It can be related to House price of unit area. We can see many outliers in the plot. The place where 9 convenient stores is having highest price.



(d) Latitude and longitude values are fluctuating.



(e) House age and House price of unit area are dependent. As the house age increases we can see that there is a highest house price.



# 2. (i)The Four Assumptions of Linear Regression

[**Linear regression**](https://www.statology.org/introduction-to-simple-linear-regression/) is a useful statistical method we can use to understand the relationship between two variables, x and y. However, before we conduct linear regression, we must first make sure that four assumptions are met:

**1. Linear relationship:** There exists a linear relationship between the independent variable, x, and the dependent variable, y.

**2. Independence:**The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.

**3. Homoscedasticity:**The residuals have constant variance at every level of x.

**4. Normality:**The residuals of the model are normally distributed.

(ii) MAE – Mean Absolute Error

The magnitude of difference between the prediction of

an observation and the true value of that observation

RMSE - Root of the Mean of the Square of Errors

It shows how far predictions fall from measured true

values using Euclidean distance

MSE – Mean Square Error

Average of the squared error that is used as the loss

function for least squares regression

R 2 – Coefficient of Determination

The proportion of the variation in the dependent

variable that is predictable from the independent

variable

Adjusted R 2

Modified version of R-squared that has been adjusted

for the number of predictors in the model.

iii. It is possible to get a negative R-square for equations that do

not contain a constant term. Because R-square is defined as

the proportion of variance explained by the fit, if the fit is

actually worse than just fitting a horizontal line then R-

square is negative.

iv. Dummy variable trap is a scenario where there are

attributes that are highly correlated (Multicollinear) and one

variable predicts the value of others.

v. One hot encoding can make values more than 2 while

labeling like [1,0,0], [0,1,0], while dummy variables are only

0 and 1.

vi. Polynomial regression is a form of Linear regression where only due to the Non-linear relationship between dependent and independent variables we add some polynomial terms to linear regression to convert it into Polynomial regression.