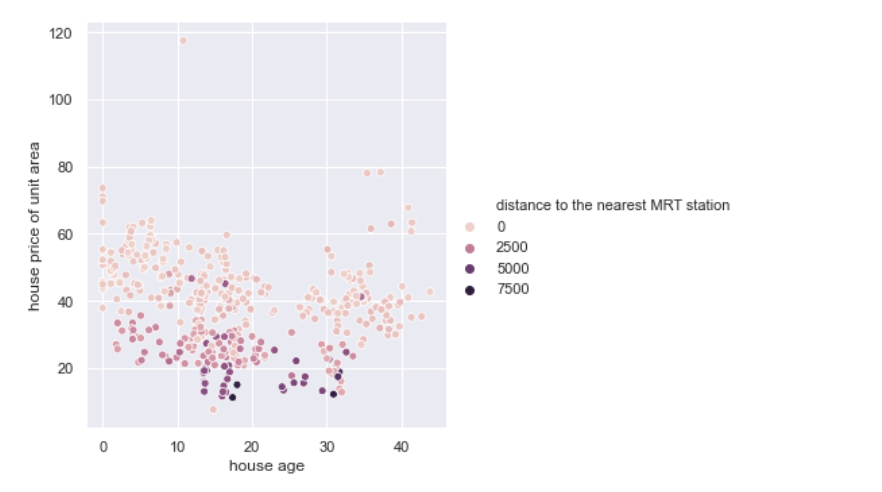
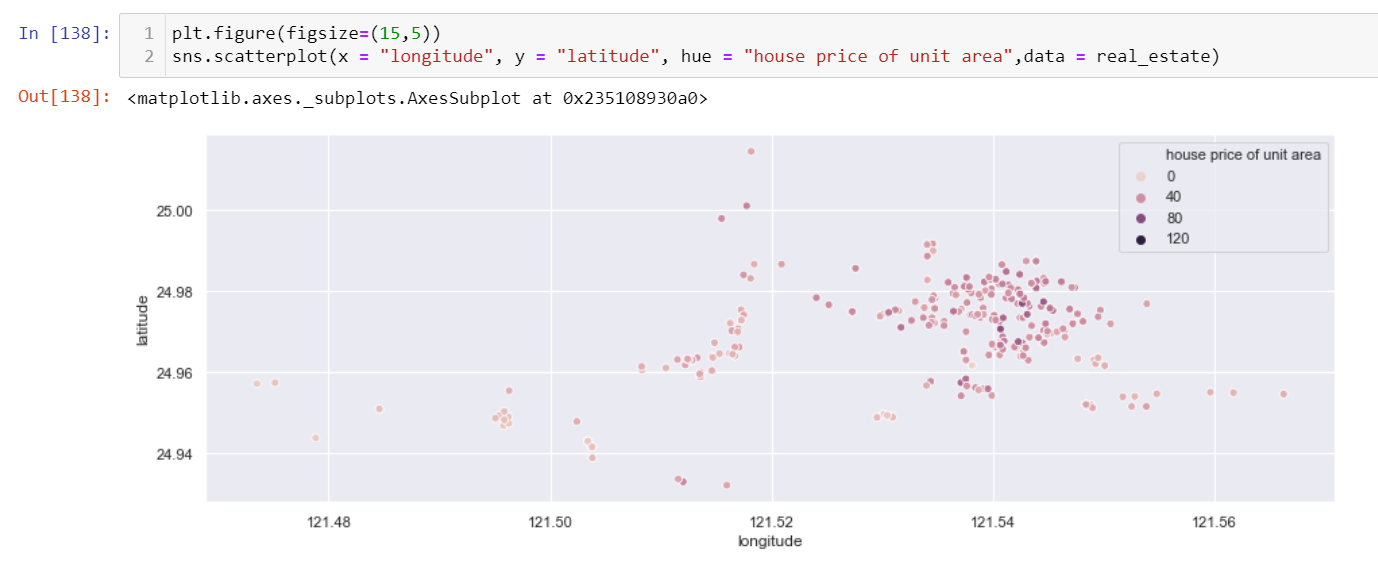
21BDA26

(B)

1. The distance to the nearest metro station is higher than the area where the house price of the unit area is less and vice versa.



1. Around certain latitudes clustered data spotting huse unit prices in the area.



2.

(i) What are the assumptions of linear regression?

* Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y.
* Independence: The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
* Homoscedasticity: The residuals have constant variance at every level of x.
* Normality: The residuals of the model are normally distributed

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

R Square/Adjusted R Square.

Mean Square Error (MSE)

Root Mean Square Error (RMSE)

Mean Absolute Error (MAE)

MAE is a straightforward metric that determines the absolute difference between actual and projected values.

MSE is a widely used and straightforward statistic that accounts for a small change in mean absolute error. Finding the squared difference between the actual and anticipated value is defined as a mean squared error.

The R Square score is a metric that measures the performance of your model, not the loss in terms of how many wells it performed. It is also known as Coefficient of Determination or sometimes also known as Goodness of fit.

(iii) Can R² be negative?

It is possible to get a negativ

e 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) What is dummy variable trap?

Using categorical data in Multiple Regression Models is a powerful method to include non-numeric data types into a regression model. Categorical data refers to data values which represent categories - data values with a fixed and unordered number of values. In a regression model, these values can be represented by [dummy variables](http://en.wikipedia.org/wiki/Dummy_variable_%28statistics%29) - variables containing values such as 1 or 0 representing the presence or absence of the categorical value.

The Dummy variable trap is a scenario where there are attributes that are highly correlated (Multicollinear) and one variable predicts the value of others. When we use *one-hot encoding* for handling the categorical data, then one dummy variable (attribute) can be predicted with the help of other dummy variables.

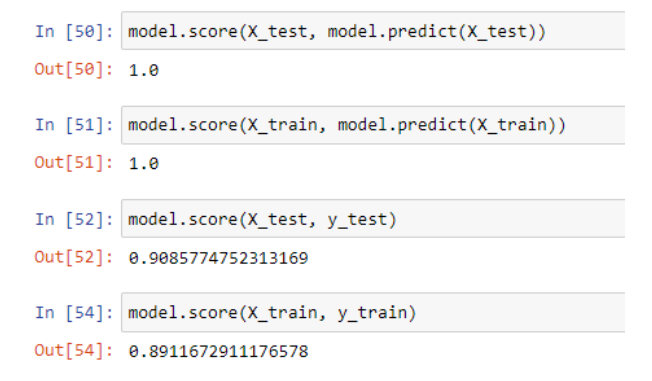
(v) Is One Hot Encoding different from Dummy Variables?

Encoding categorical variables can be done in two ways. Let's say there are n values in a category variable. It is converted into n variables using one-hot encoding, and n-1 variables using dummy encoding. If we have k categorical variables with n values each. Hot encoding produces kn variables, whereas dummy encoding produces kn-k variables.

(vi) How is polynomial regression different from linear regression?

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.

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



When you call score on classifiers like LogisticRegression, RandomForestClassifier, etc. the method computes the accuracy score of the input and its target value.

There is a model accuracy of around 0.9085, which means the model is highly accurate with the given test data. Test data usually accounts for a part of the data taken in order to peform tests and obtain accurate results.