**ML LAB - 2**

**AYESHA SHARIFF – 21BDA18**

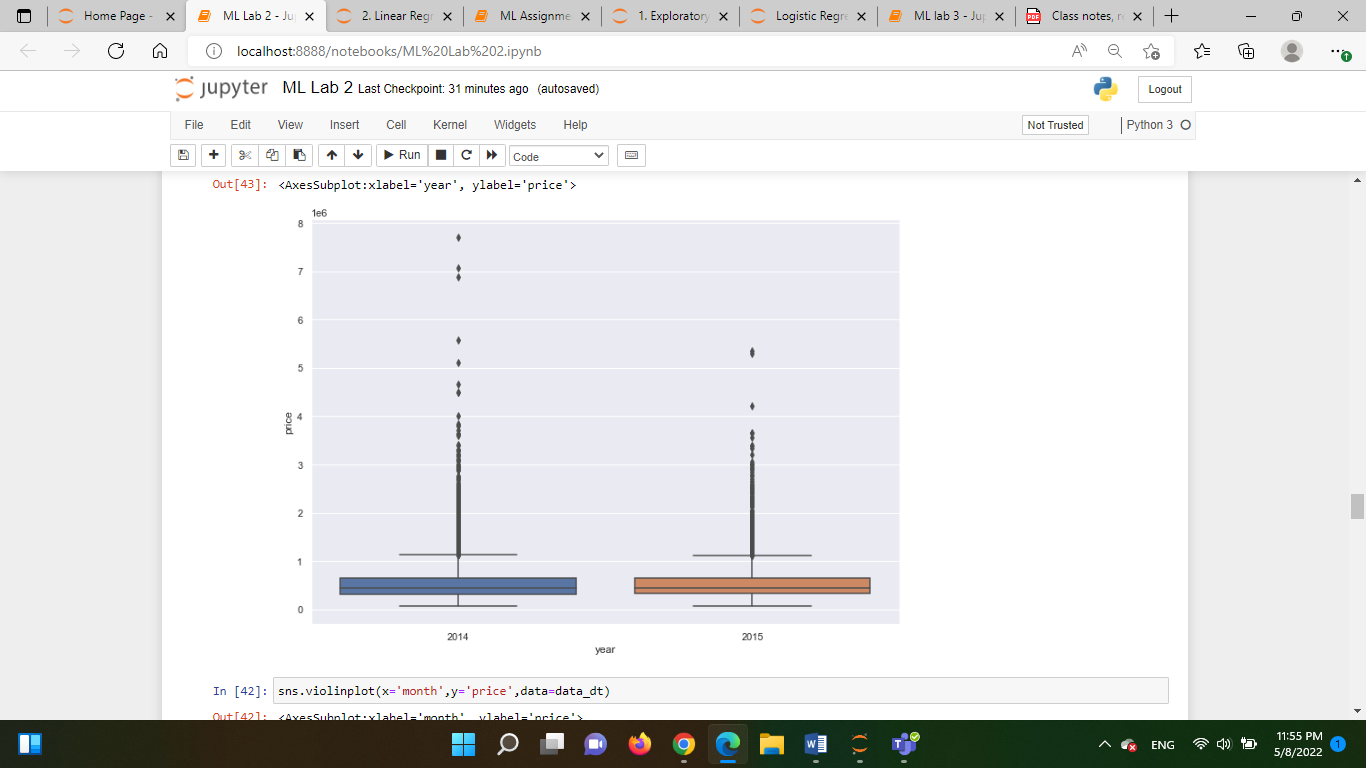
1. Document 5-6 key insights from EDA and support each point with a visualization.

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

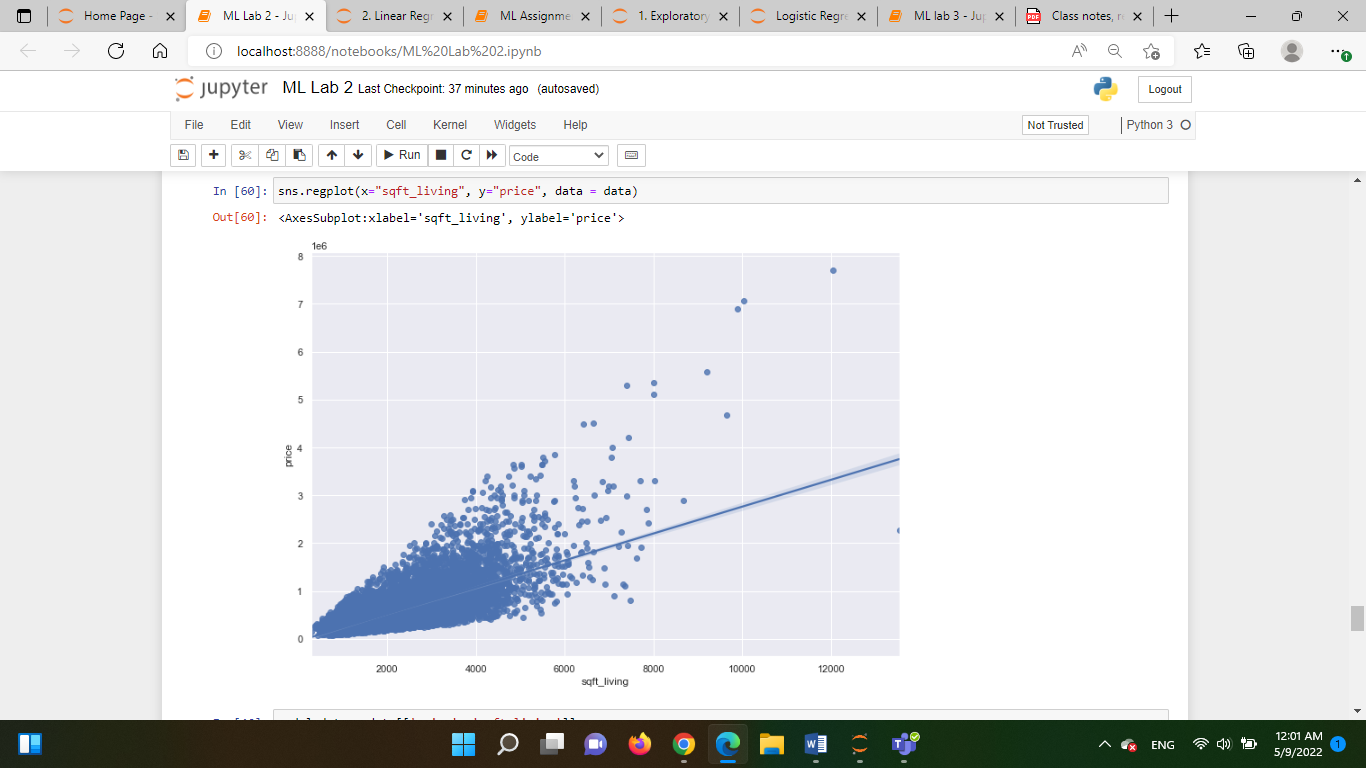
The following is the structure of the data set. Variable Name - Description

Id - House ID; date - Date house sold; price - House price; bedrooms - Number of bedrooms; bathrooms - Number of bathrooms; sqft\_living - Living room size; sqft\_lot - Lot size; floors - Number of floors; waterfront - Has access to waterway(0 = no; 1 = yes); view – View; condition - House condition(1 = bad; 5 = perfect); grade - House grade; sqft\_above - Above size; sqft\_basement - Basement size; yr\_built - Year when house was built; yr\_renovated - Year when house was renovated; zipcode - House zipcode; lat - House latitude; long - House longitude; sqft\_living15 - Living15; sqft\_lot15 - Lot15

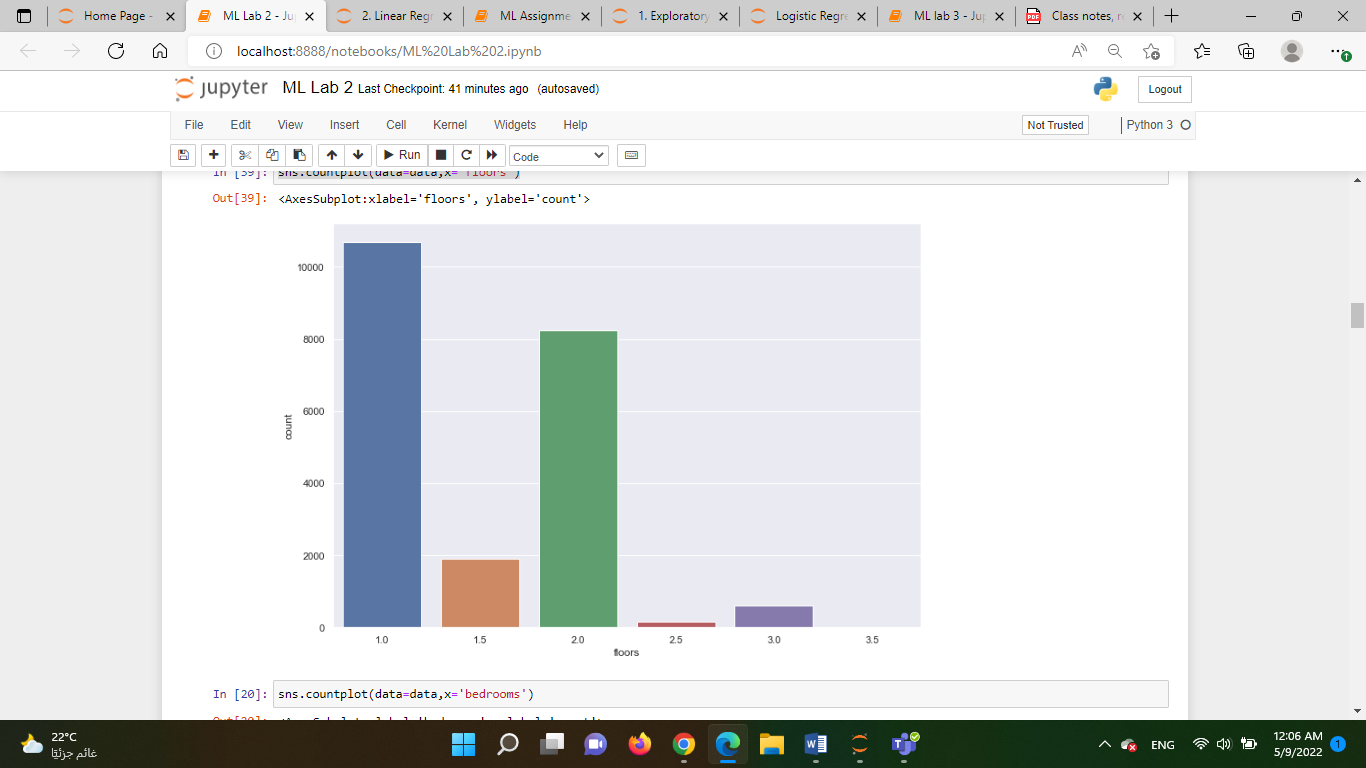
1. Most of the houses were sold in the year 2014.



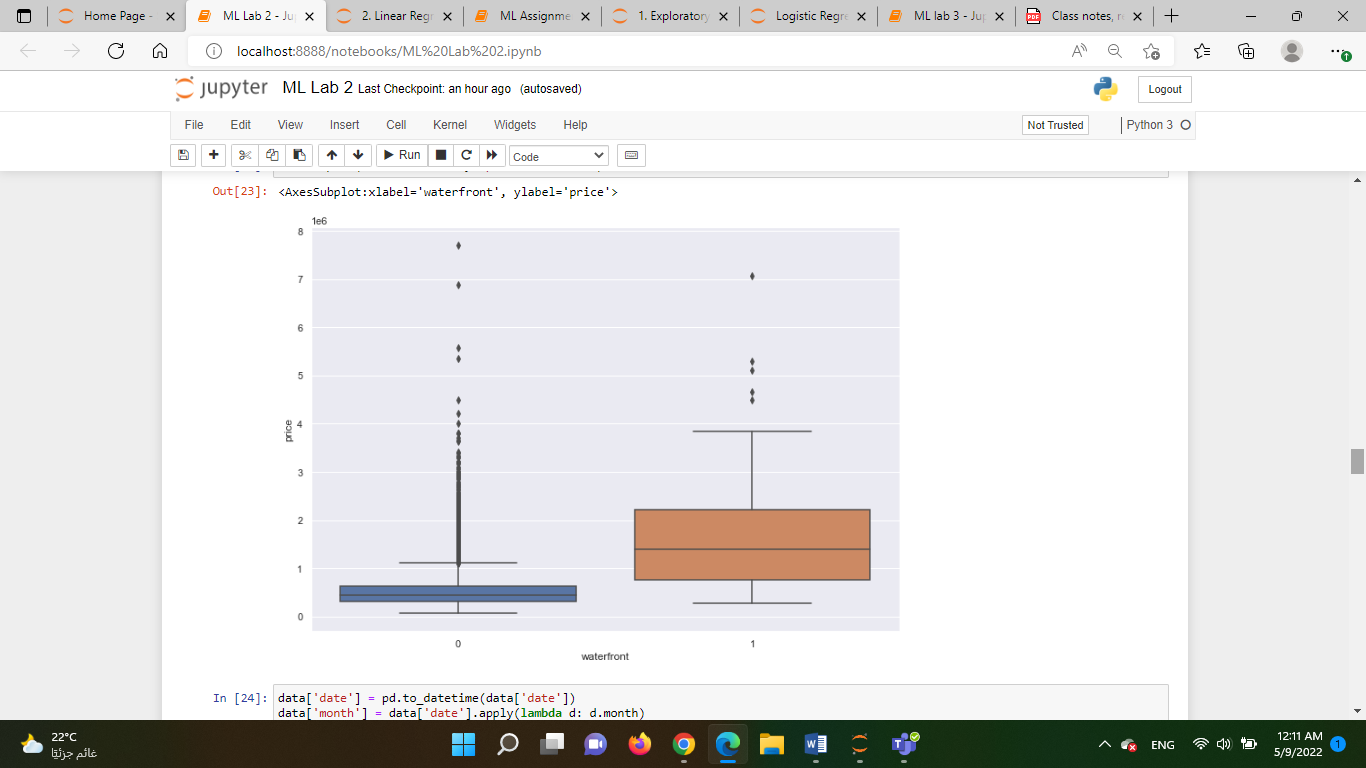
1. There is positive correlation between price and living room size.



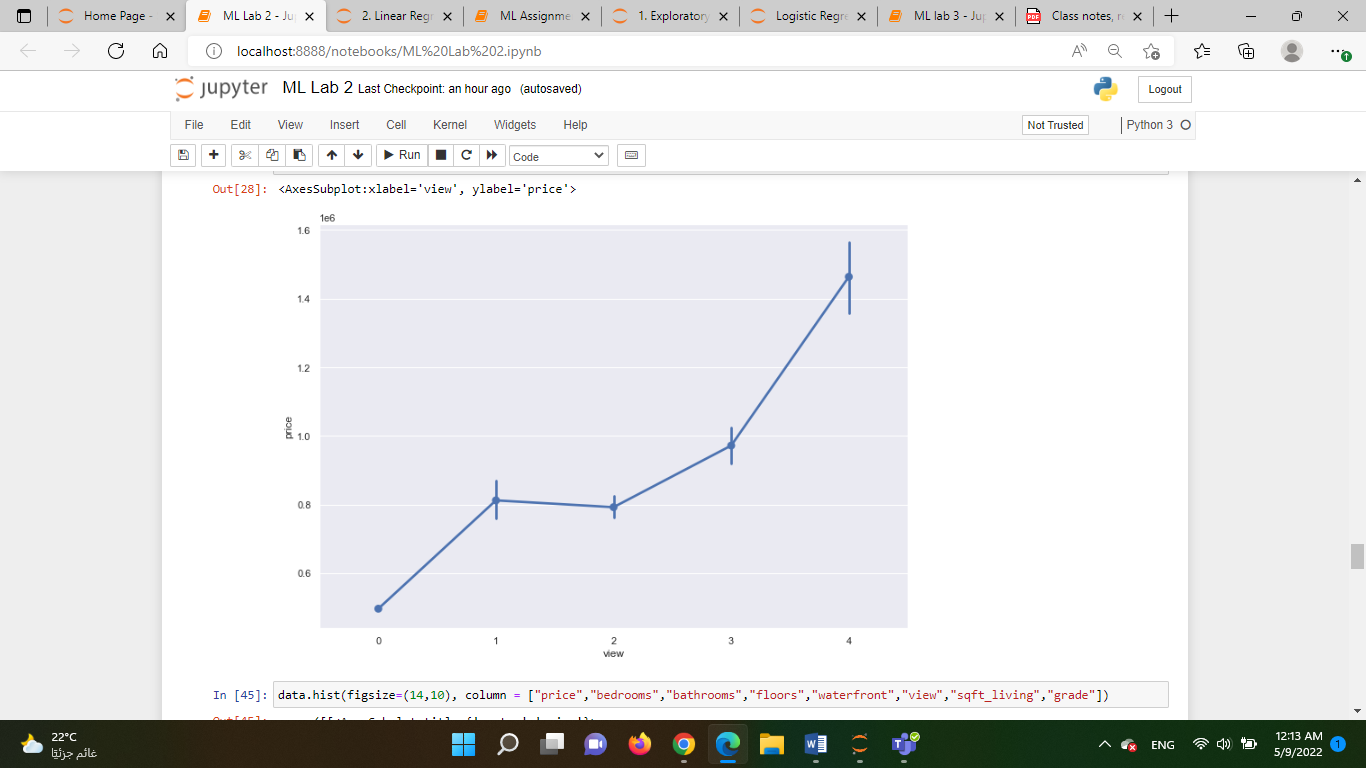
1. People prefer house having 1 floor more, followed by 2 and 1.5 floors.



1. Houses with waterfront has more price than without waterfront.

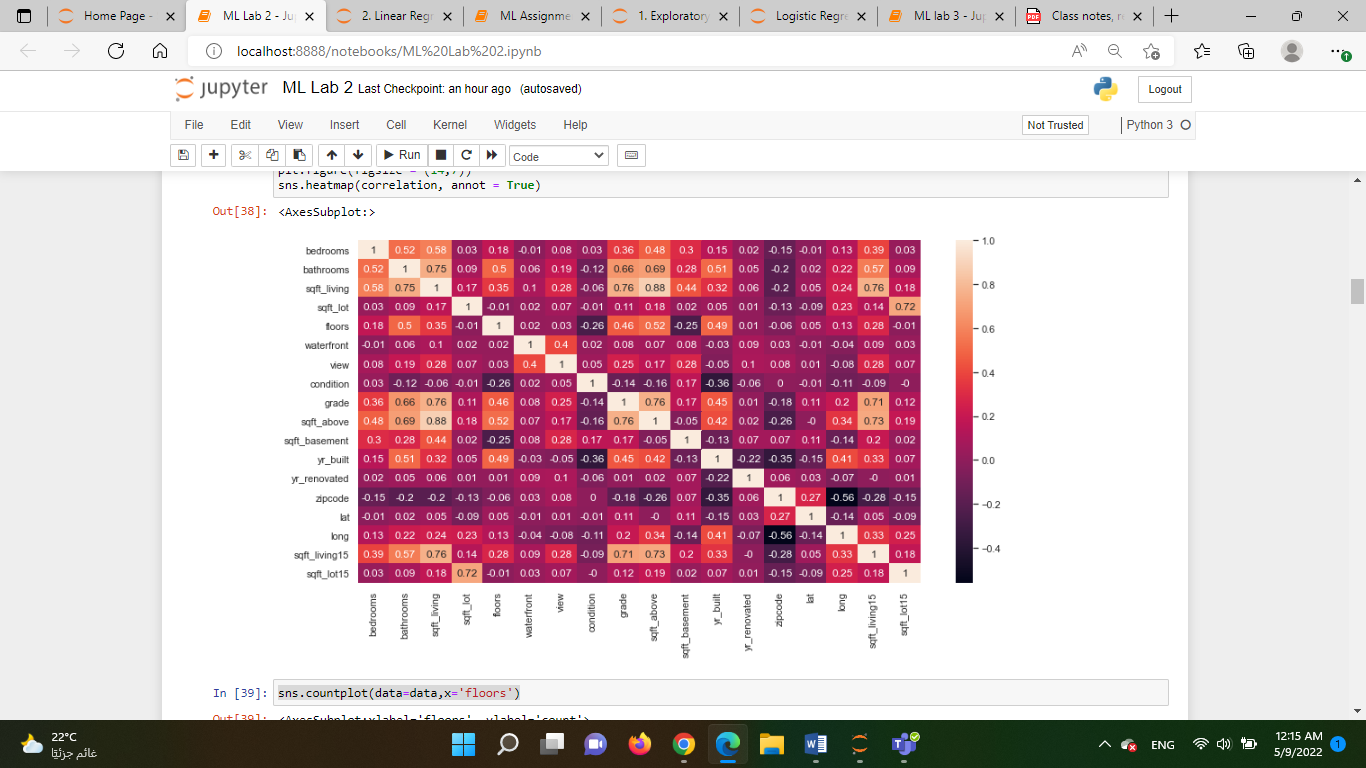


1. Houses with view has more demand so the prices are also high.



1. The correlation between variables ranges from light pink(high) to dark purple(low).

We can see that there is higher correlation between living room size and above size, living room size and house grade.



1. Answer the following questions:
2. What are the assumptions of linear regression?

* Independence of observations.
* No Hidden or Missing Variables.
* Relations between the independent and dependent variables must be linear.
* Residuals should follow a normal distribution.
* There is no or little multicollinearity. Multicollinearity is the phenomenon when a number of the explanatory variables are strongly correlated.
* Homoscedasticity. Homoscedasticity in a model means that the error is constant along the values of the dependent variable.
* All independent variables are uncorrelated with the error term.
* Observations of the error term are uncorrelated with each other.

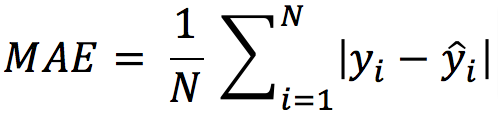
1. How can we evaluate a Regression model? Define each metric and its

interpretation.

Regression predicts a continuous dependent element in the presence of various independent elements. Linear regression tries to make a trend line that has the least difference between actual and predicted values. This difference is also known as residual.

MEAN ABSOLUTE ERROR (MAE):

It is the mean of the absolute difference between the actual value in the dataset and the value predicted by the model.



N = the count of the data points.

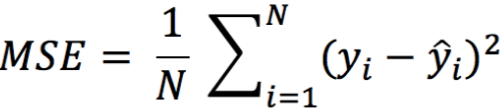
Y = the actual value in the dataset

Y cap = the model’s predicted value.

The absolute values are taken, and if it’s not then the negative and positive difference will cancel out each other. The smaller the MAE, the more accurate the model is. If MAE is zero it shows the model is perfect. If MAE is large then the model is not good.

MEAN SQUARED ERROR (MSE):

This is the mean of the squared difference of the actual value in the dataset and the value predicted by the model.



N = the count of data points in the data.

Y = the actual value in the dataset.

Y cap= the model’s predicted value.

The MSE will be large if there are outliers in the dataset, this is not the case with MAE.

MSE focuses on larger errors, as when we are squaring the error the effect of large errors becomes more prominent.

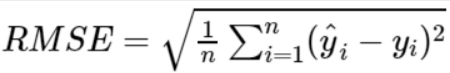
If the errors are low, lower than one, then it leads to underestimating the model’s error.

ROOT MEAN SQUARED ERROR:

It is the mean of root squared subtraction between the actual value in the dataset and the value predicted by the model.

It’s the same as MSE, we are just taking the root of it.

The smaller the value of root mean squared error, the more accurate the model is.

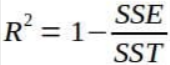


R Square:

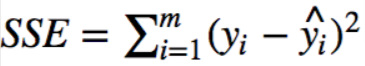
It estimates the ratio of the variance of the dependent element described by the target element.

It’s used for finding the accuracy of the model. It depicts the closeness of the data points to the trend line made by the model. This helps to make a link between the independent element and the target element.

R square is from zero to one. The nearer R square is to one, the more accurate the model.



SSE is the sum of the square difference of the residuals.



SST is the sum of the difference of the actual value of the data and the mean of all the actual values in data.

image

When we add more elements in the data then R square increases, for this we can use adjusted R square.

1. Can R squared be negative?

R squared isn't always the square of anything, therefore it can be negative without breaking any mathematical principles. Only when the chosen model does not follow the data trend does R squared become negative.

For equations without a constant term, it is possible to generate a negative R-square. Because R-square is defined as the proportion of variation explained by the fit, it is negative if the fit is worse than fitting a horizontal line.

1. What is dummy variable trap? In linear regression models, to create a model that can infer relationship between features (having categorical data) and the outcome, we use the dummy variable technique.

A “Dummy Variable” or “Indicator Variable” is an artificial variable created to represent an attribute with two or more distinct categories/levels.

The dummy variable trap is a scenario in which the independent variables become multicollinear after addition of dummy variables.

Multicollinearity is a phenomenon in which two or more variables are highly correlated. In simple words, it means value of one variable can be predicted from the values of other variable(s).

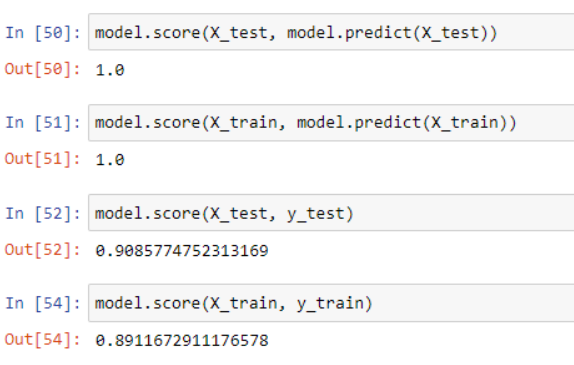
1. Is One Hot Encoding different from Dummy Variables?

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.

1. How is polynomial regression different from linear regression? Simple Linear Regression establishes the relationship between two variables using a straight line. It attempts to draw a line that comes closest to the data by finding the slope and intercept which define the line and minimize regression errors. Simple linear regression has only one x and one y variable.

Polynomial Regression is a one of the types of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E (y |x).

Polynomial Regression provides the best approximation of the relationship between the dependent and independent variable.

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

Here score shows the r square value where, R-Squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the

independent variable(s) in a regression model.

The regression predictions exactly fit the data if the r2 is 1.When the model fits the data worse than the poorest possible least-squares predictor, r2 values outside the range 0 to 1 occur.