***ASSIGNMENT 2***

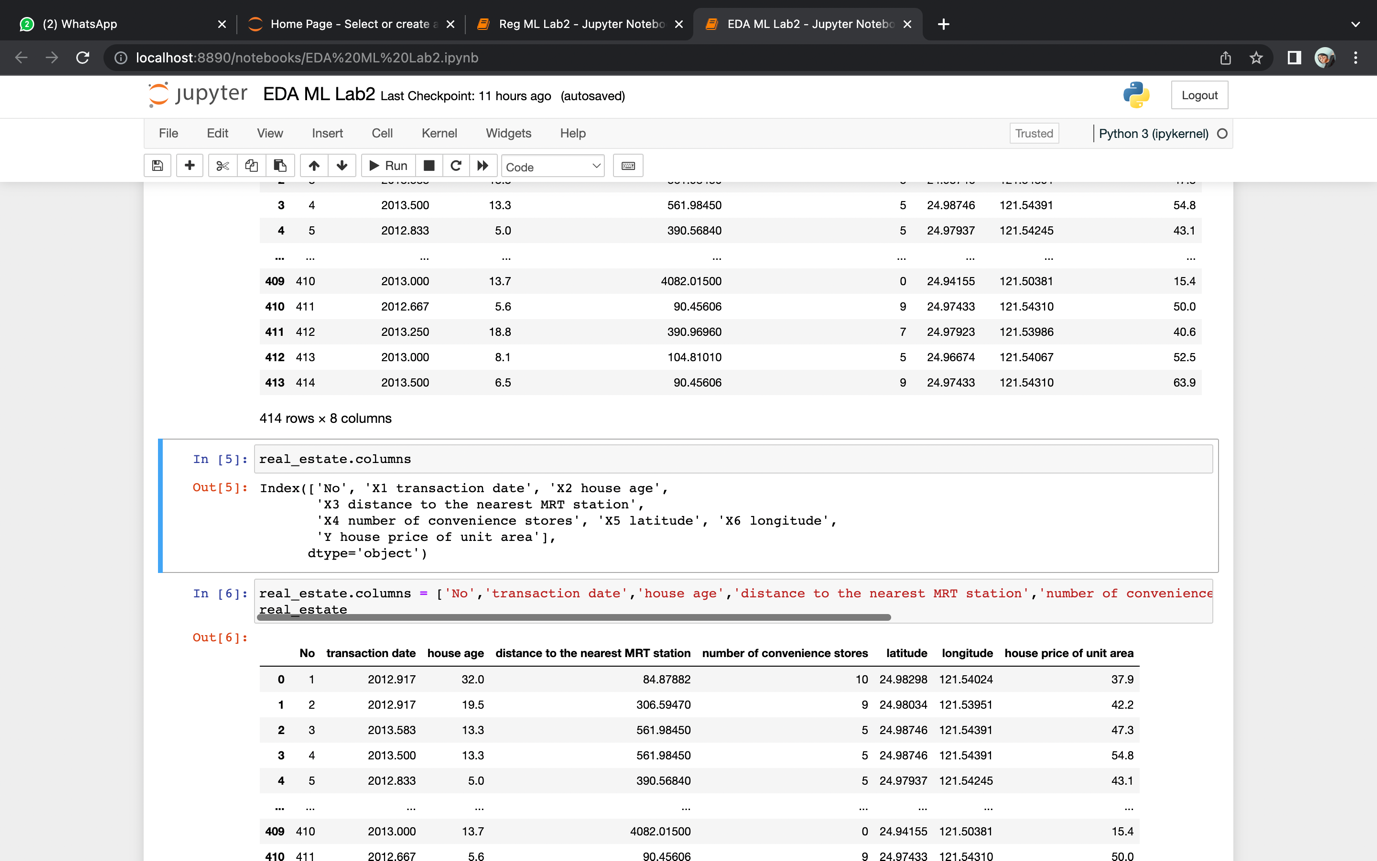
***ASHLY THAMPI***

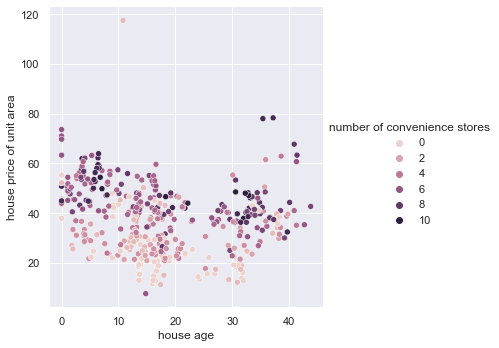
***21BDA19***

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

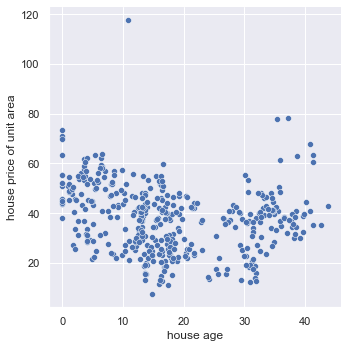
* The dataset is related to real estate data which consists of 414 rows × 8 columns.

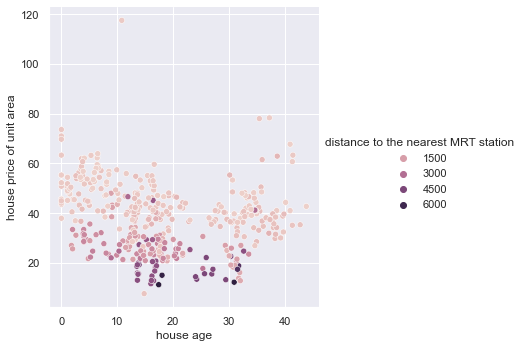
The columns are,

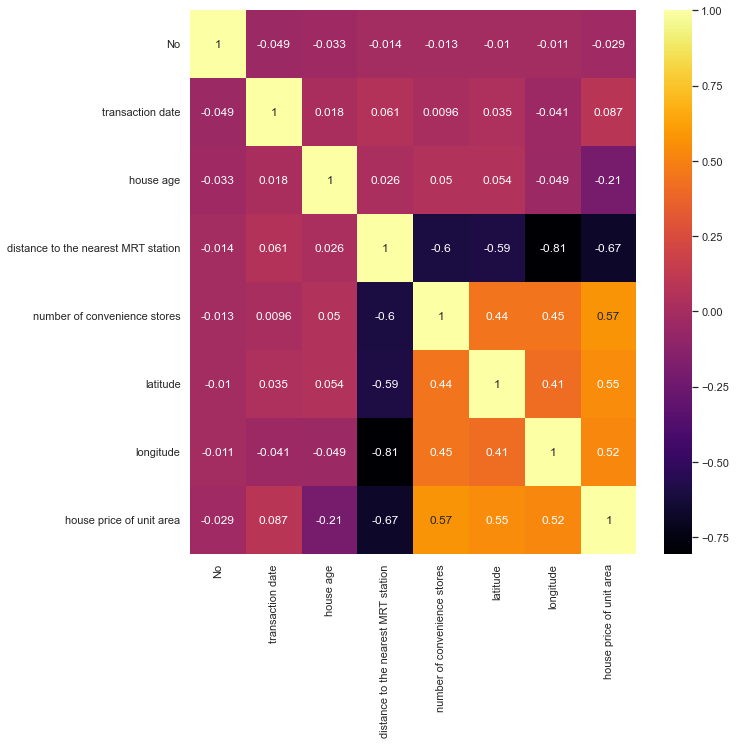




* There are more convenience stores where the house price of unit area is higher



* There is no much difference in the house price of unit area with the house age
* 
* The distance to the nearest metro station is higher than the area where the house price of the unit area is less



* Number of convenience stores and house price of unit area are highly correlated



* There is a slight increase in house price of unit area from 2012 to 2013

2. Answer the following questions:

**i. What are the assumptions of linear regression?**

* Linearity: The relationship between X and the mean of Y is linear.
* Homoscedasticity: The variance of residual is the same for any value of
* Independence: Observations are independent of each other.
* Normality: For any fixed value of X, Y is normally distributed.

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

In regression, there are three main measures for evaluating models:

* 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 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 R squared be negative?**

* R squared can be negative because it is not always the square of anything. R squared become negative only when the chosen model does not follow the data trend.

**iv.What is dummy variable trap?**

* The Dummy Variable Trap occurs when two or more dummy variables created by one-hot encoding are highly correlated (multi-collinear). This means that one variable can be predicted from the others, making it difficult to interpret predicted coefficient variables in regression models. In other words, the individual effect of the dummy variables on the prediction model can not be interpreted well because of multicollinearity.

**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 type of linear regression in which additional polynomial factors are added to linear regression to convert it to polynomial regression due to the non-linear relationship between the dependent and independent variables.
* Polynomial provides the best approximation of the relationship between the dependent and independent variable

**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.

**Bonus: We saw Sweetviz as an Automated EDA option. What are the other options? Try a few of them and share which one did you find the best.**

Other options for Sweetviz are:

1. dtale

2. pandas profiling

3. sweetviz

4. autoviz

DataPrep provides much more functionality than simple EDA. It can help you ingest more data sources and can help you get through large data sets faster.

In addition, the clean API in DataPrep can help you clean your data set without many hurdles.