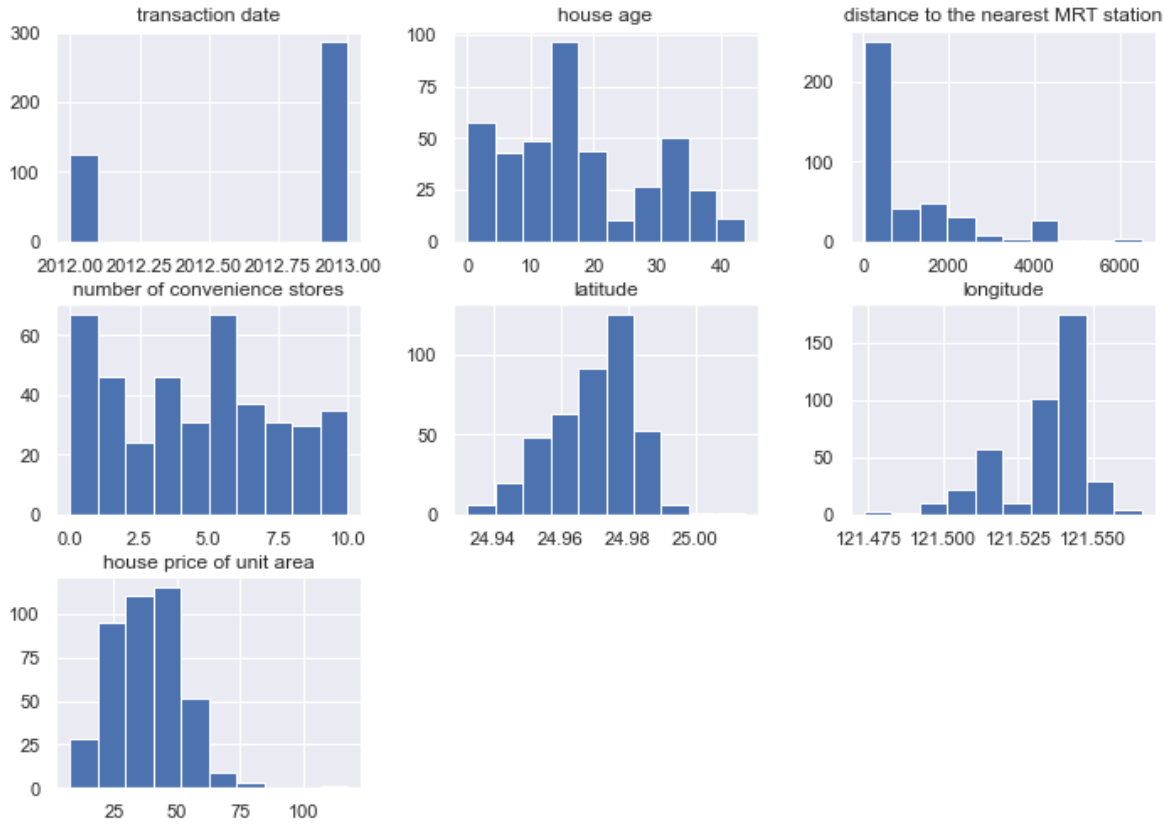
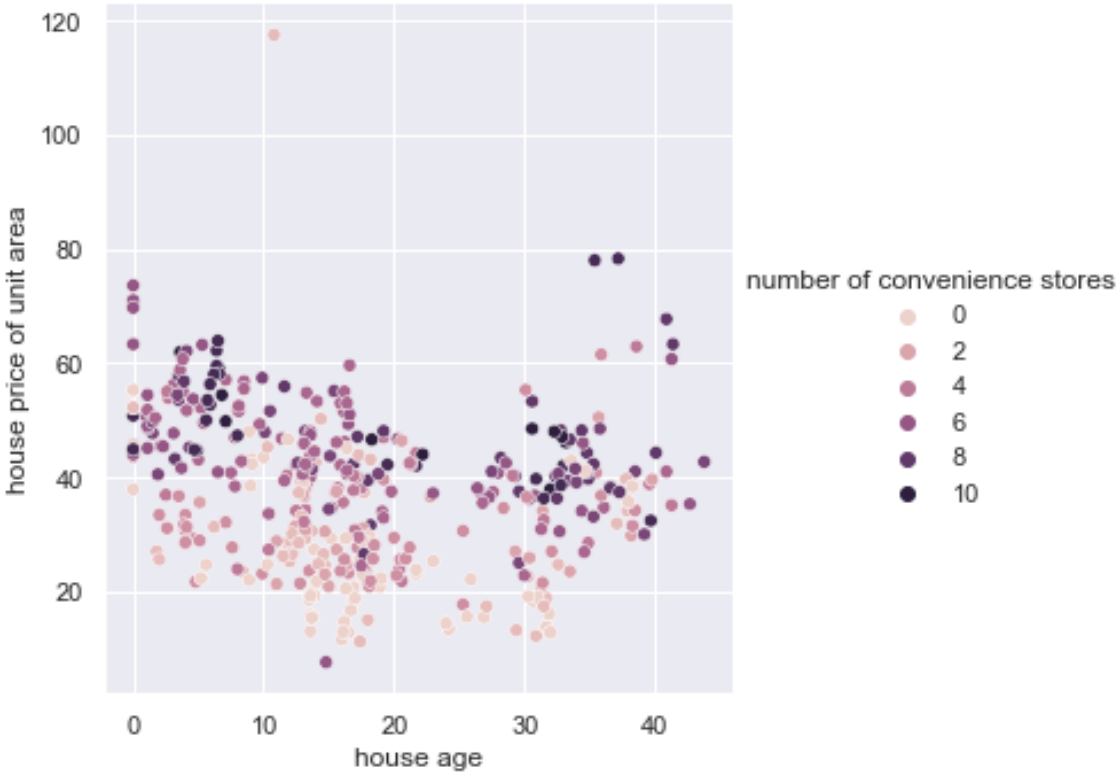
Assignment02-ML1-Lab02

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

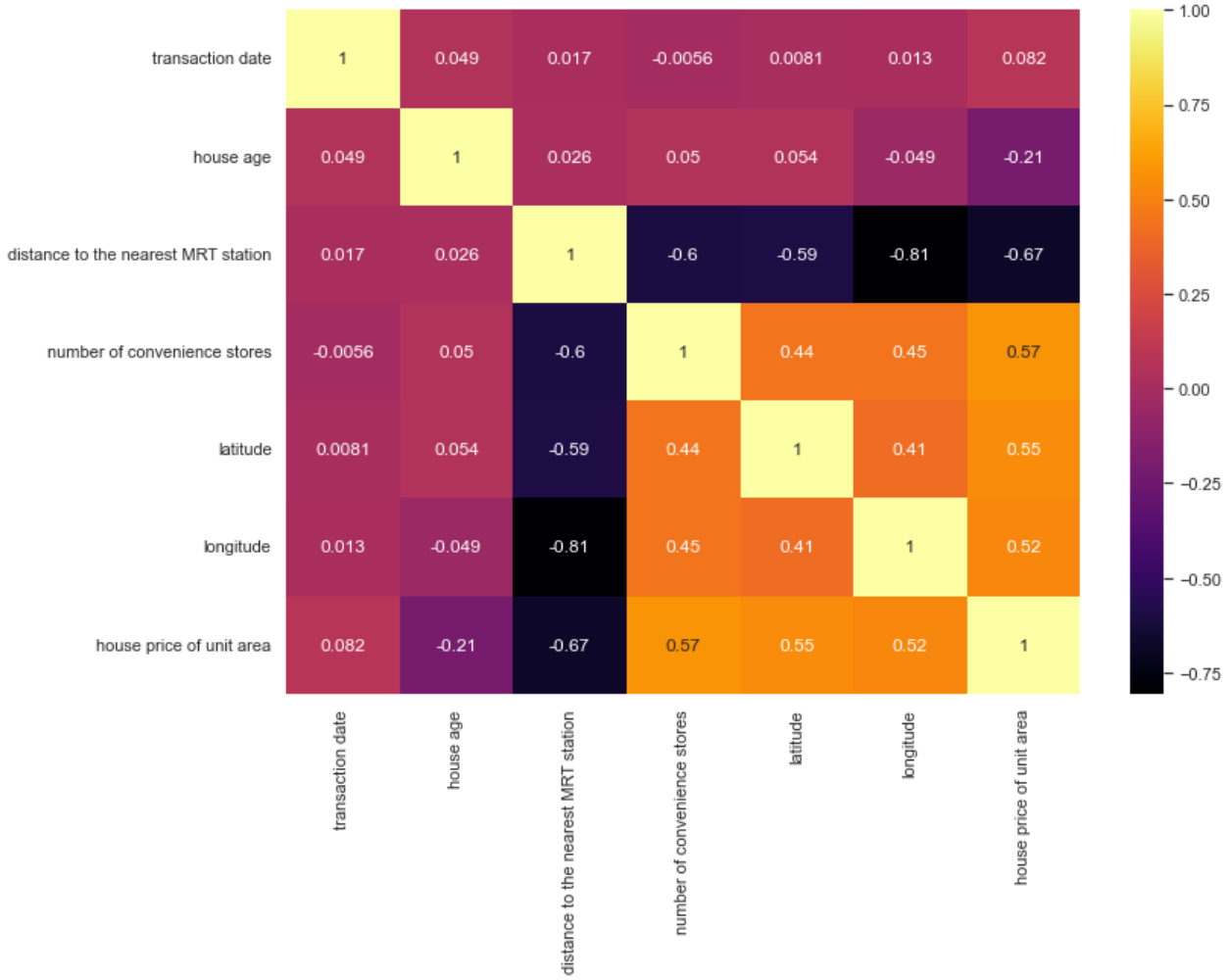
Values are normally distributed



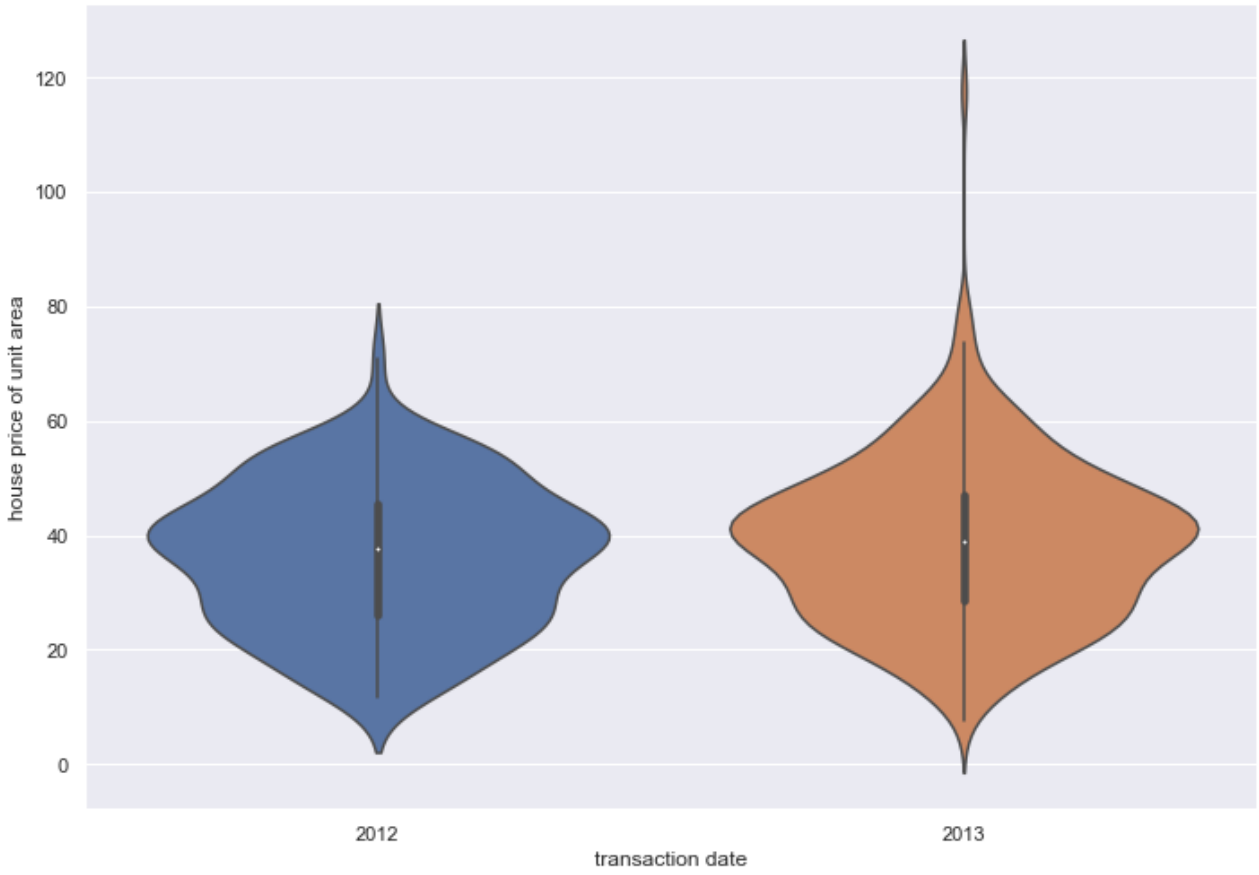
House Price depends on House age



Distance to the nearest MRT station is least correlated with other values



House Price increased in the year 2013



2. Answer the following questions:

- What are the assumptions of linear regression?

  There are four assumptions associated with a linear regression model:

  Linearity: The relationship between X and the mean of Y is linear.

  Homoscedasticity: The variance of residual is the same for any value of X.

  Independence: Observations are independent of each other.

  Normality: For any fixed value of X, Y is normally distributed.

- How can we evaluate a Regression model? Define each metric and its interpretation.
  - 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
- Can R squared be negative?
    - 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.
- What is dummy variable trap?
    - The Dummy variable trap is a scenario where there are attributes that are highly correlated (Multicollinear) and one variable predicts the value of others.
- Is One Hot Encoding different from Dummy Variables?
    - 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.
- 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
- 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.
    - The other best options are:
        - dtale
        - pandas profiling
        - Sweetviz
        - Autoviz
    - The one that I liked is dtale.