Train In Data

Assumptions of Linear Models

Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_n X_{ni} + \varepsilon_i$$

- Y is the outcome variable
- X are the predictor variables
- β are the coefficients
- β_0 is the intercept
- ε_i is the difference between the predicted and the observed value of Y for the ith observation

Linear Model Assumptions

1. **Linearity**: The mean values of the outcome variable for each increment of the predictor(s) lie along a straight line. There is a linear relationship between predictors and target.

2. **Normally distributed errors**: the residuals (ε_i) are random, normally distributed with a mean of 0.

3. **Homoscedasticity**: At each level of the predictor variable(s), the variance of the residual terms should be constant.

4. **No perfect multicollinearity**: There should be no perfect linear relationship between two or more of the predictors.

When the assumptions are met

- The coefficients and parameters of the regression equation are said to be unbiased.

- The model is a good fit for the data.

When the assumptions are not met

We can’t fully trust the predictions of the model.

Some issues could be:

- Outliers
- Lack of homoscedasticity
- The variables are too skewed

What can we do?

Transforming the data is useful to correct the problems with outliers and homoscedasticity.

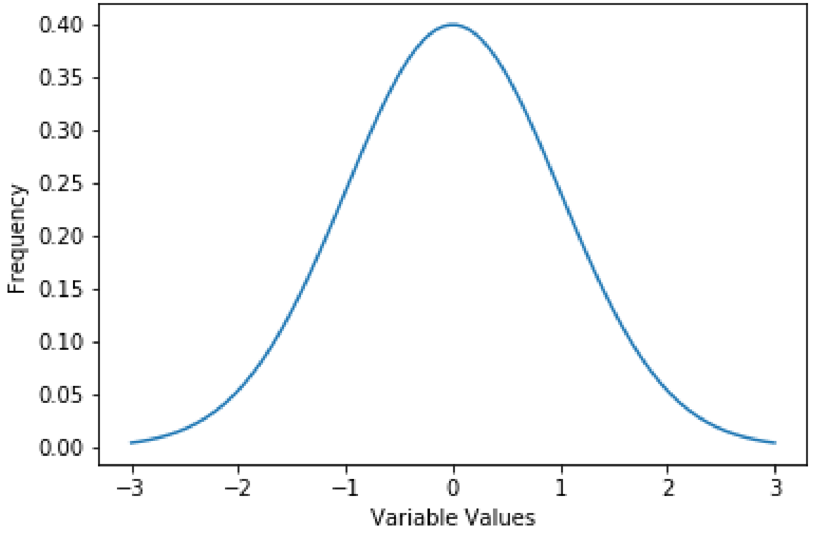
- Mathematical transformations
- Discretisation
- Remove or censor outliers

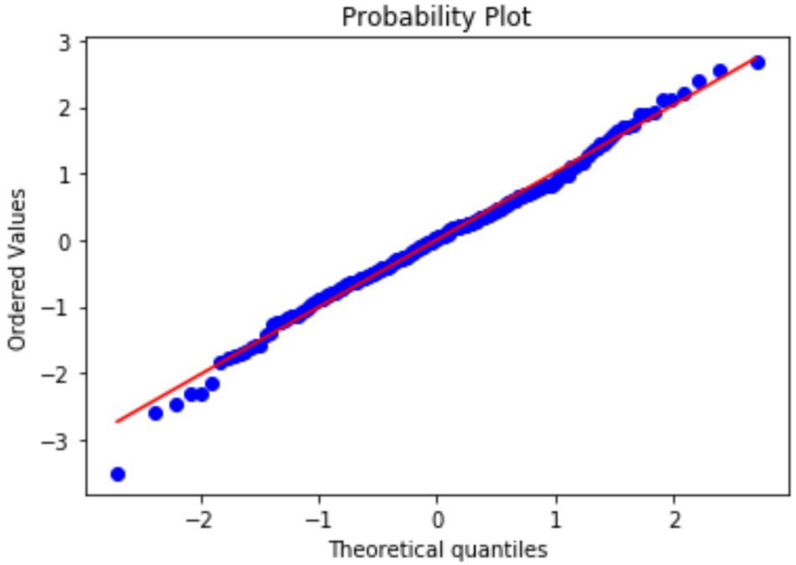
Evaluate model performance

Errors ~ N(0,σ)

- Normality can be assessed with histograms and Q-Q plots

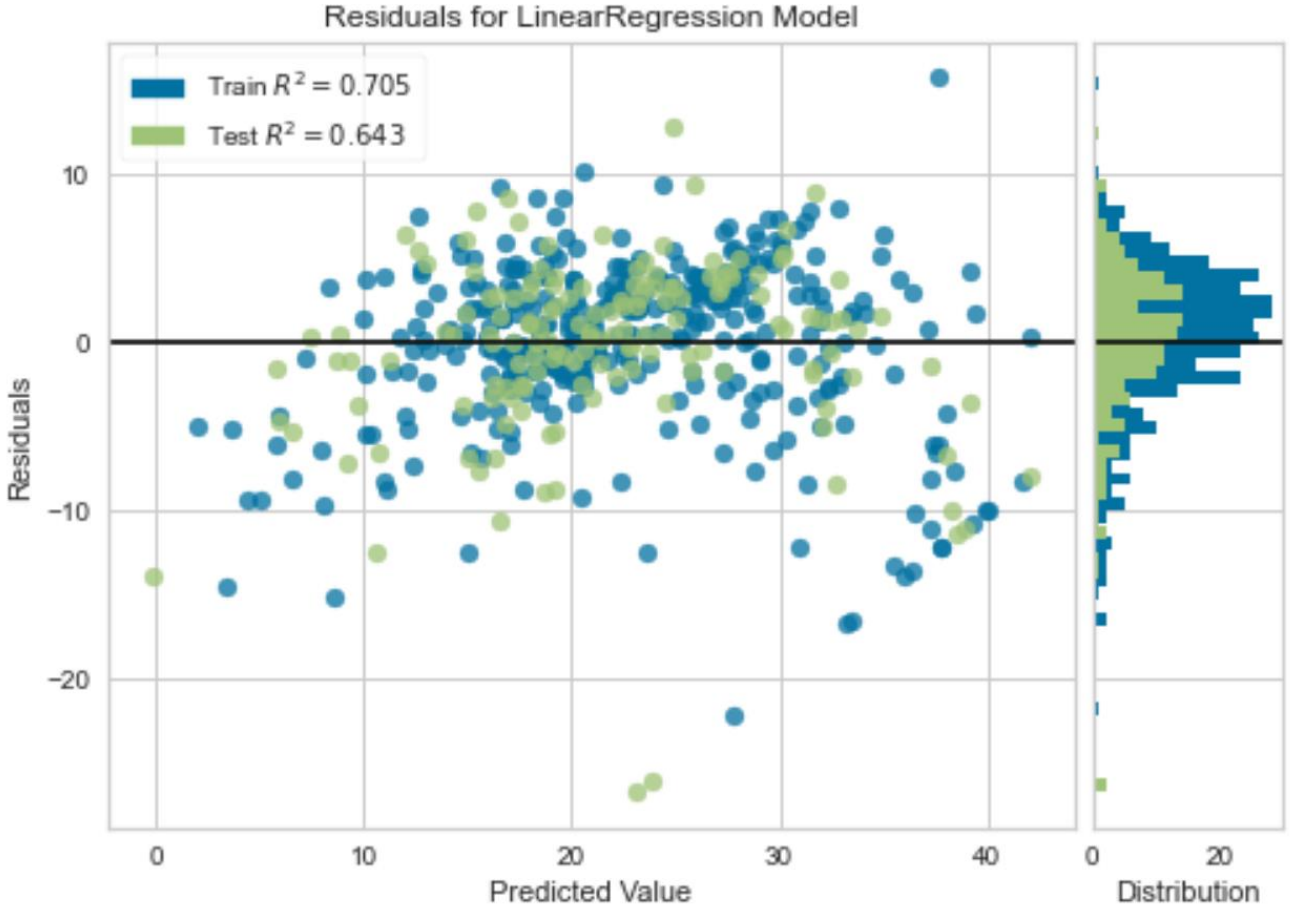
- Normality can be statistically tested, for example with the Kolmogorov-Smirnov test.





Errors ~ N(0,σ)

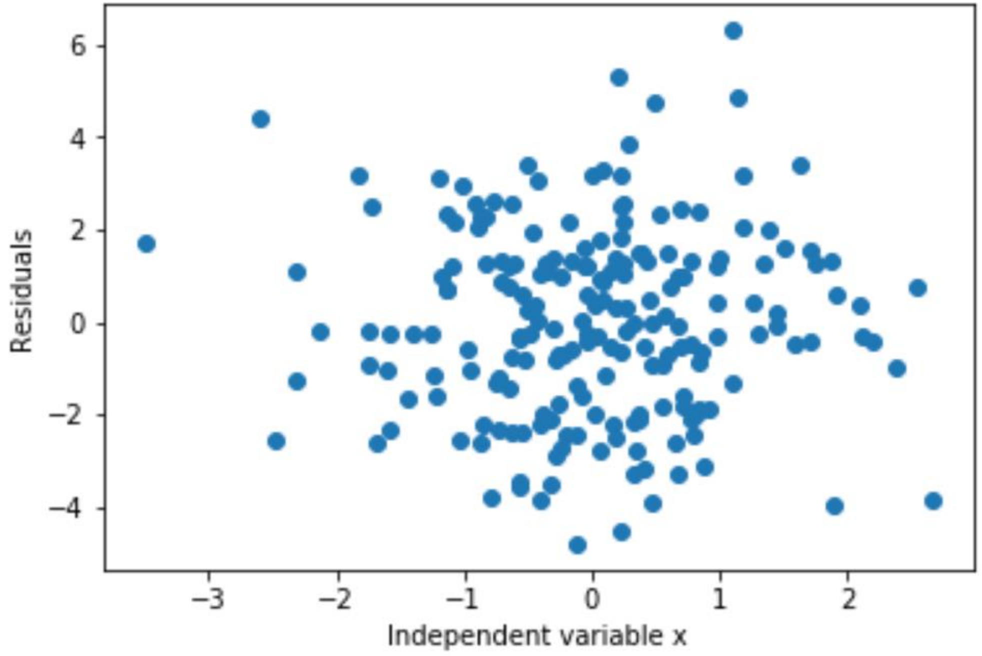
With Yellobrick, we can plot the histogram of the error term and its variance across the predicted values, to test for normality and homocedasticity at the same time.

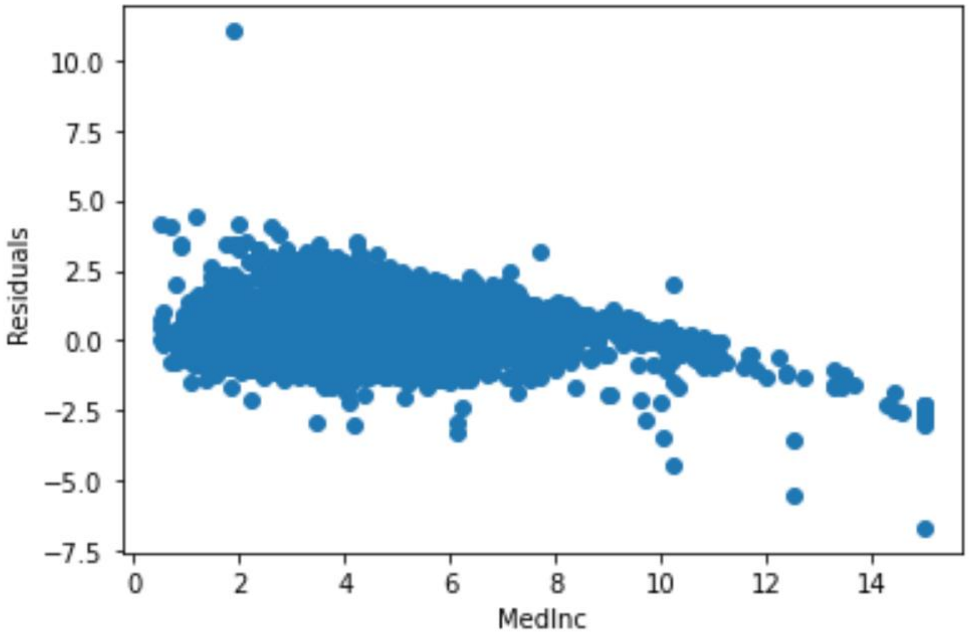


Homoscedasticity

- There are tests and plots to determine homoscedasticity.
  - Residuals plot
  - Levene's test
  - Barlett's test
  - Goldfeld-Quandt Test
- Visual inspection

Homoscedasticity



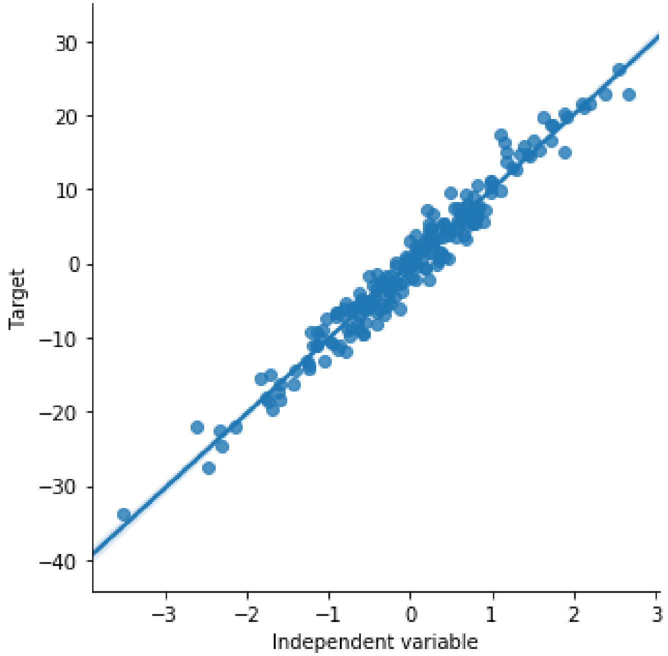


Homoscedasticity: the error term (that is, the “noise” in the relationship between the independent variables X and the dependent variable Y) is the same across all the independent variables.

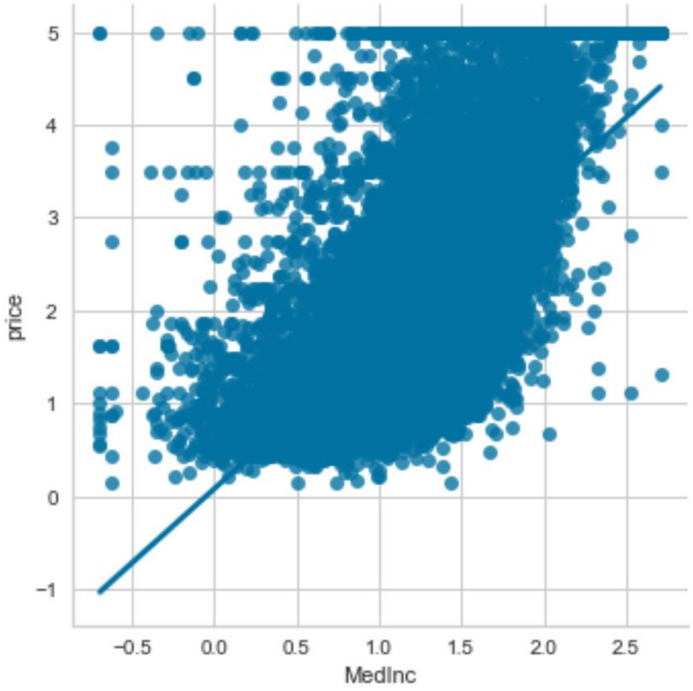
To identify homoscedasticity we need to plot the residuals vs each of the independent variables.

Linear Relationship – Scatter plots

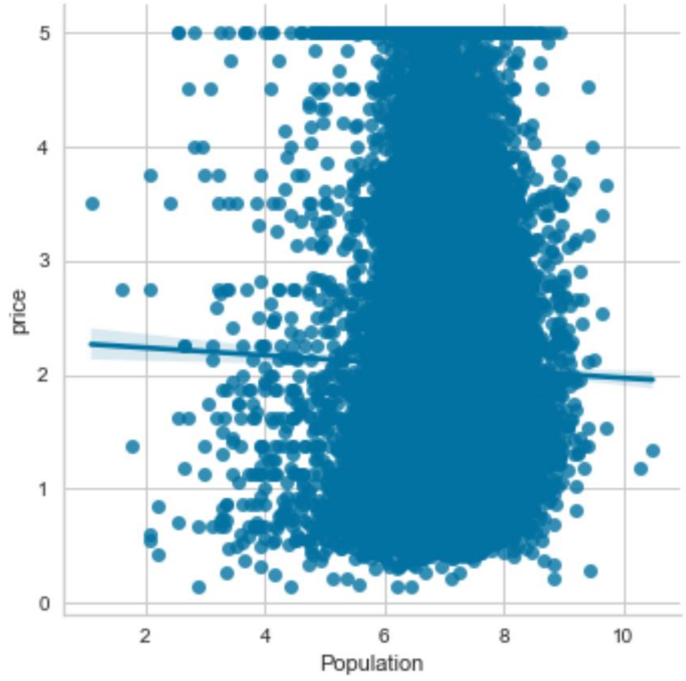
Expected – Simulated data



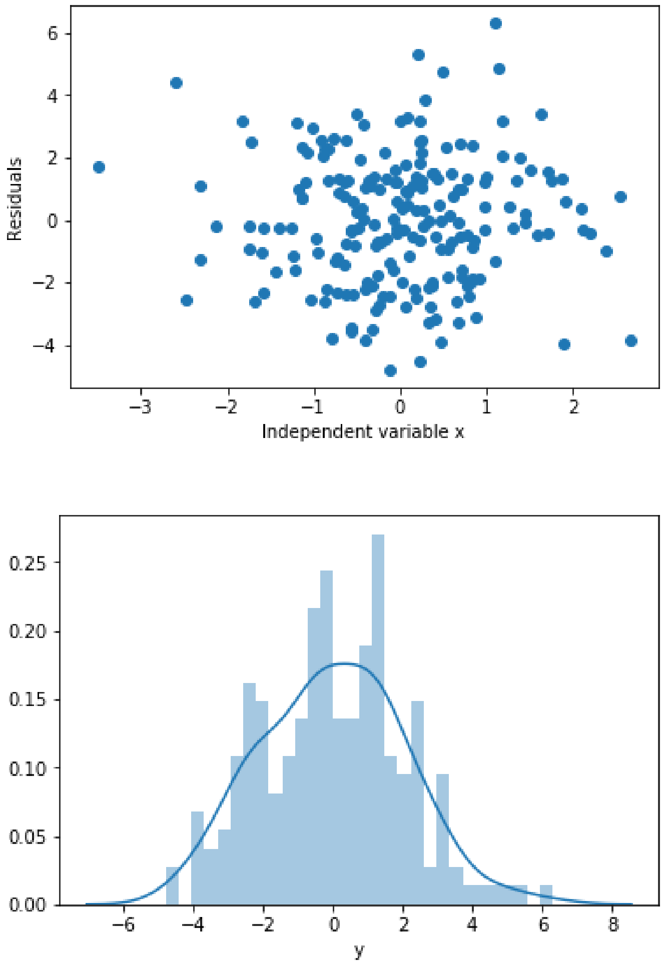
Somewhat linear relationship



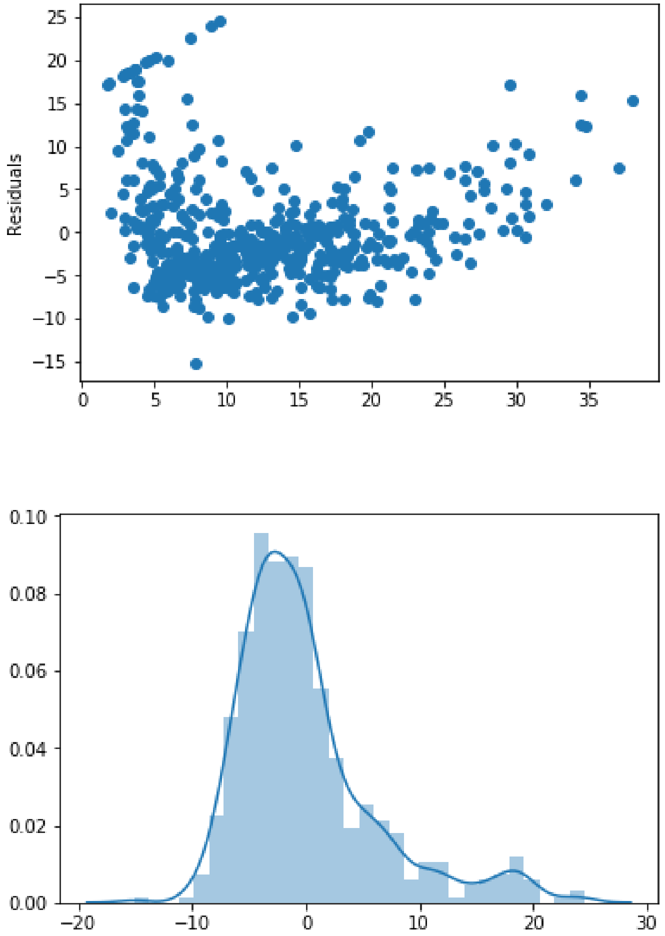
Non-linear relationship



Linear Relationship – Residual plots



Expected – Simulated data



Somewhat linear relationship

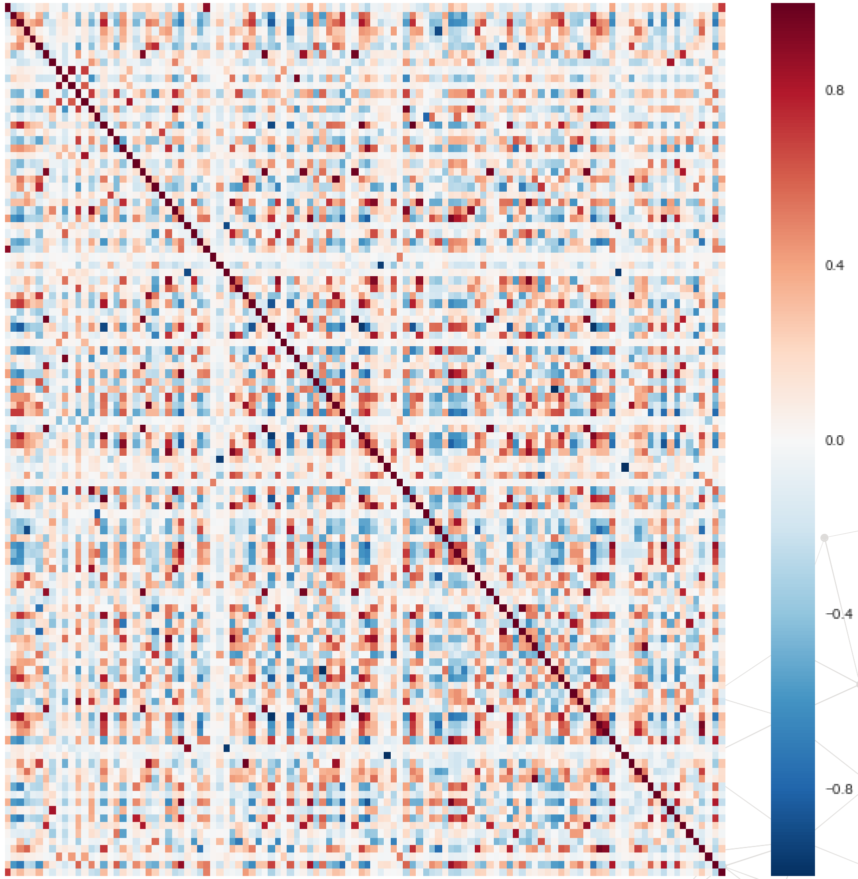
- If relationship between X and y is linear, residuals should be normally distributed and centred around 0

- Residuals are the difference between the predictions and the real value y.

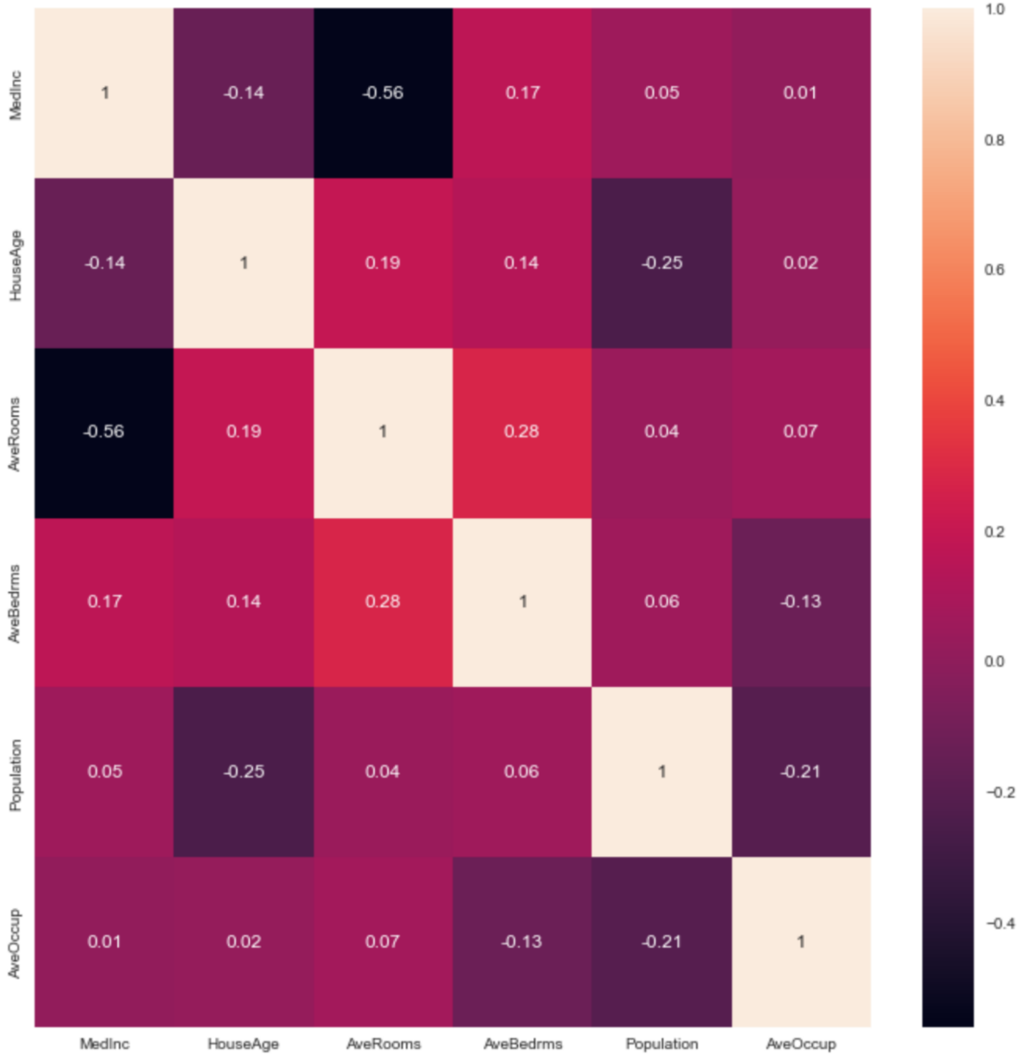
Multicolinearity

- Multicollinearity occurs when the independent variables are correlated with each other

- Multicollinearity can be assessed with a correlation matrix or the variance inflation factor (VIF)



Multi Co-linearity



	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup
MedInc	1	-0.14	-0.56	0.17	0.05	0.01
HouseAge	-0.14	1	0.19	0.14	-0.25	0.02
AveRooms	-0.56	0.19	1	0.28	0.04	0.07
AveBedrms	0.17	0.14	0.28	1	0.06	-0.13
Population	0.05	-0.25	0.04	0.06	1	-0.21
AveOccup	0.01	0.02	0.07	-0.13	-0.21	1

Accompanying Jupyter Notebook

- Read the accompanying Jupyter Notebook
- Demonstration of linear model assumptions and the effect of non-linear transformations

Appendix

More Linear Model assumptions

Linear Model Assumptions

1. **Variable types**: All predictor variables must be quantitative or categorical (with two categories), and the outcome variable must be quantitative, continuous and unbounded.
2. **Non-zero variance**: The predictors should have some variation in value (i.e., they do not have variances of 0).
3. **No perfect multicollinearity**: There should be no perfect linear relationship between two or more of the predictors. So, the predictor variables should not correlate too highly.
4. **Linearity**: The mean values of the outcome variable for each increment of the predictor(s) lie along a straight line. In plain English this means that it is assumed that the relationship we are modelling is a linear one.

Linear Model Assumptions

5. **Normally distributed errors**: It is assumed that the residuals in the model are random, normally distributed variables with a mean of 0.

6. **Homoscedasticity**: At each level of the predictor variable(s), the variance of the residual terms should be constant. Independent errors: For any two observations the residual terms should be uncorrelated (or independent)

7. **Independence**: all of the values of the outcome variable are independent

8. **Independent errors**: For any two observations the residual terms should be uncorrelated (or independent). This is sometimes described as a lack of autocorrelation.

THANK YOU

www.trainindata.com