Train In Data

Assumptions of Linear Models

Linear Model Assumptions

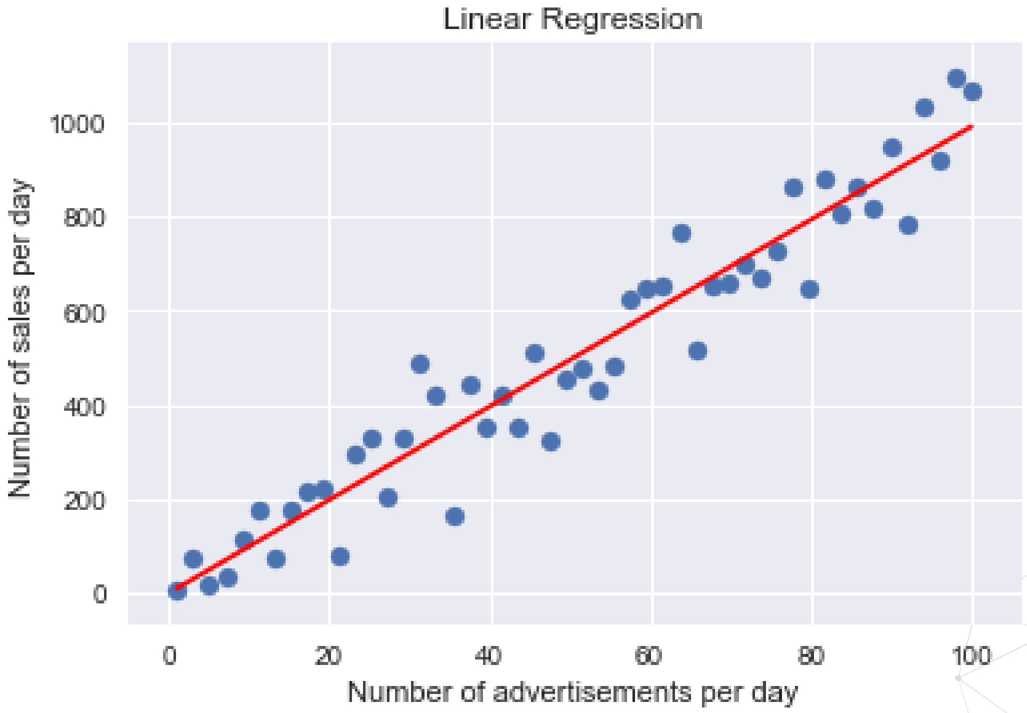
Linear models make the following assumptions about the independent variables (Xs)

- Linear relationship between the variables and the target
- Multivariate normality
- No or little co-linearity
- Homoscedasticity

Linear Relationship

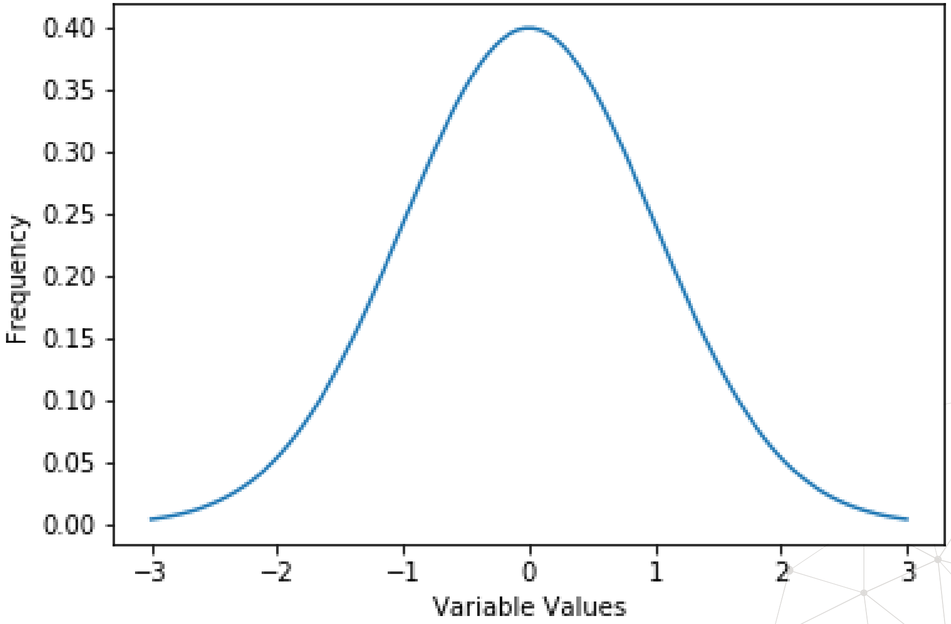
Y ≈ β0 + β1X1 + β2X2 + … + βnXn

- Linear relationship can be assessed with scatter plots
- Sometimes non-linear transformations of the variables and the target improve the linear relationship



Normality

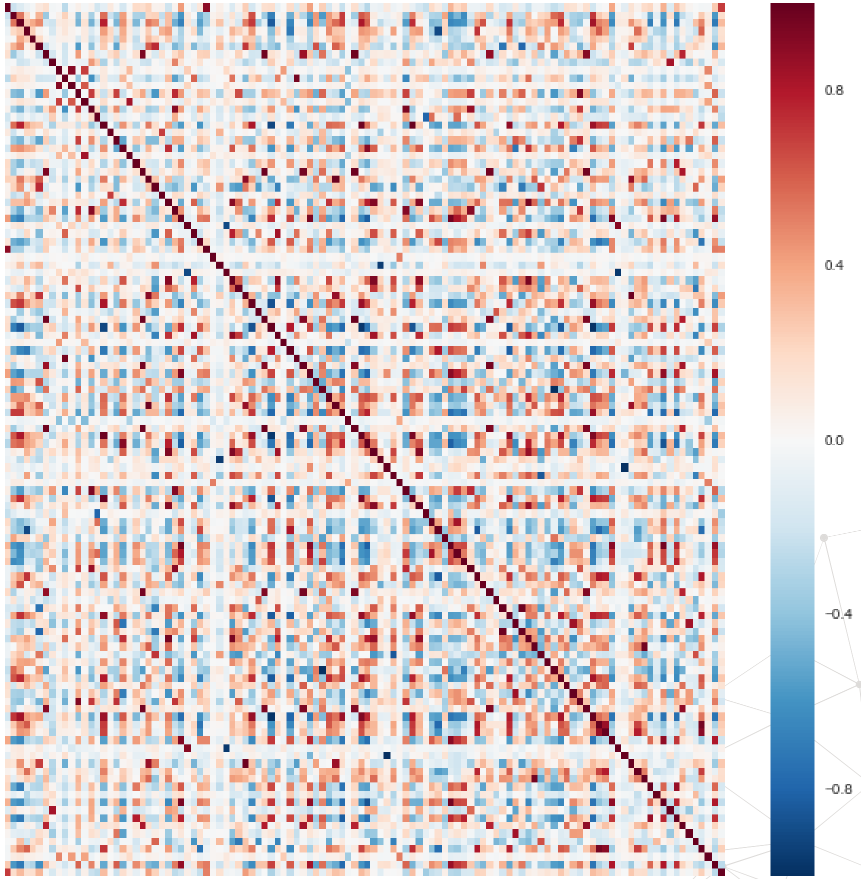
- Variables follow a Gaussian Distribution
- Normality can be assessed with histograms and Q-Q plots
- Normality can be statistically tested, for example with the Kolmogorov-Smirnov test.
- When the variable is not normally distributed a non-linear transformation (e.g., logarithm-transformation) may fix this issue.



No co-linearity

- 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)
  - Outside of the scope of this course
  - Check the course Feature Selection for Machine Learning



Homoscedasticity

- The independent variables have the same finite variance.

- Also known as homogeneity of variance.

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

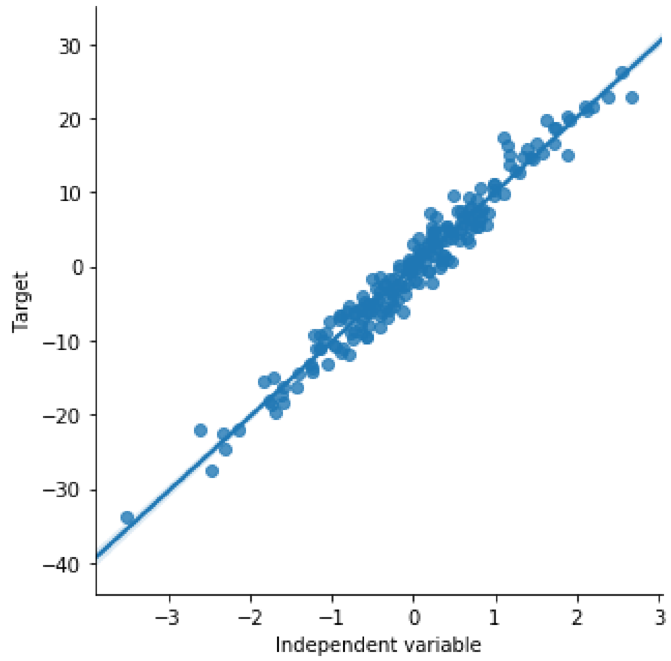
- Non-linear transformations and feature scaling can help improve homogeneity of variance

Evaluate model assumptions

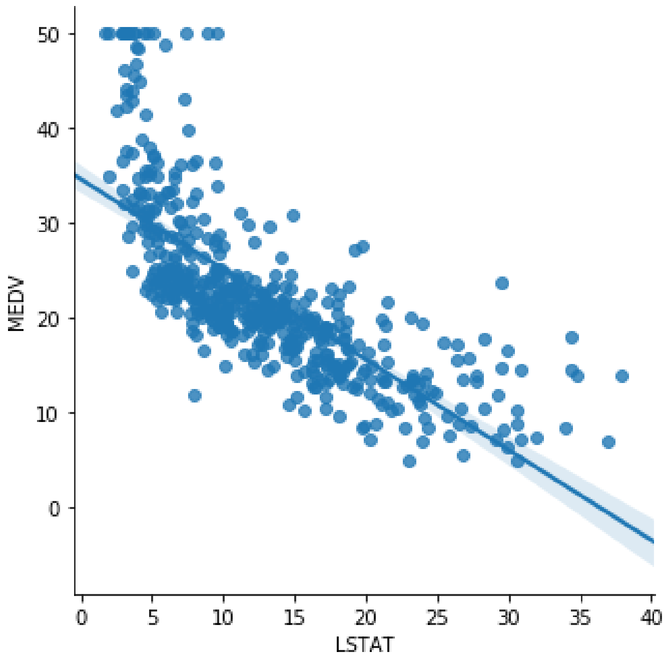
Compare model assumptions in simulated and real data

Linear Relationship – Scatter plots

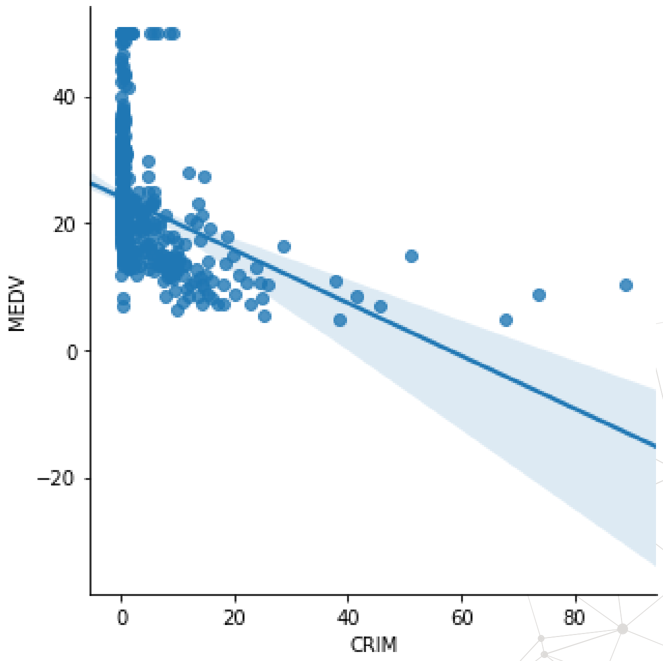
Expected – Simulated data



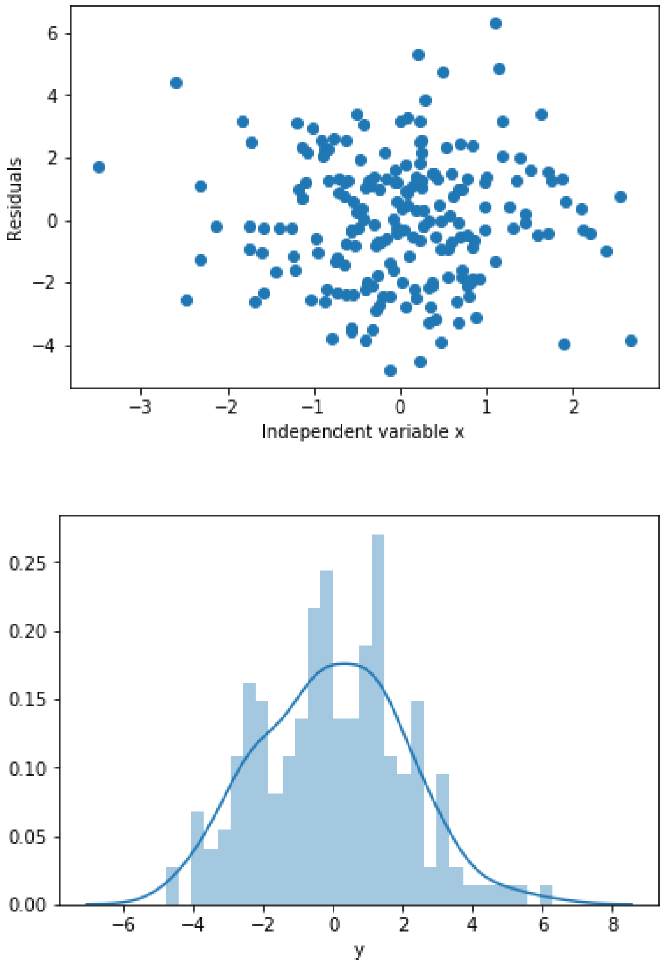
Somewhat linear relationship



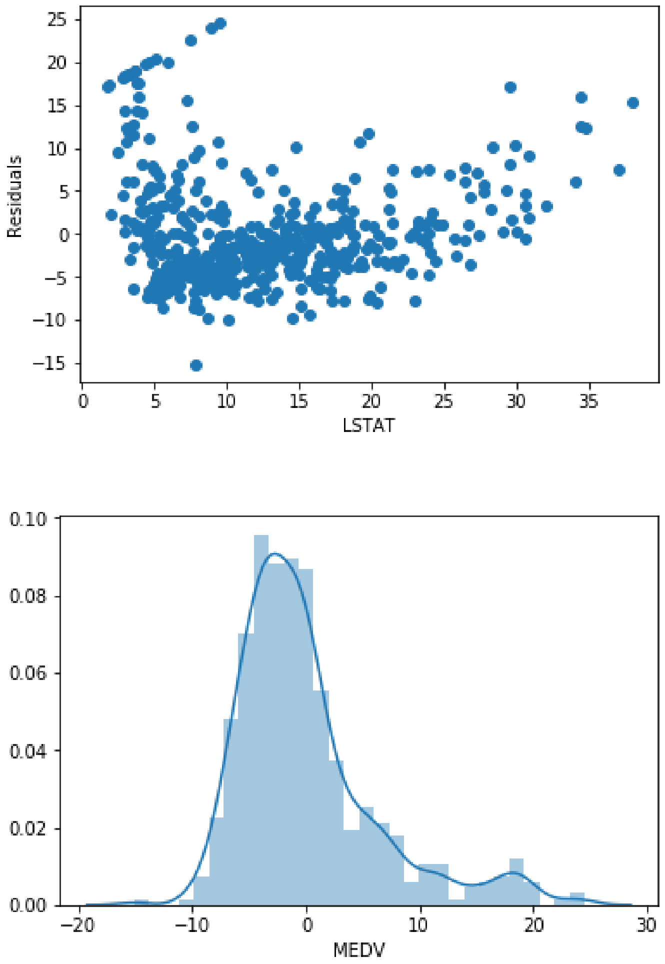
Non-linear relationship



Linear Relationship – Residual plots



Expected – Simulated data



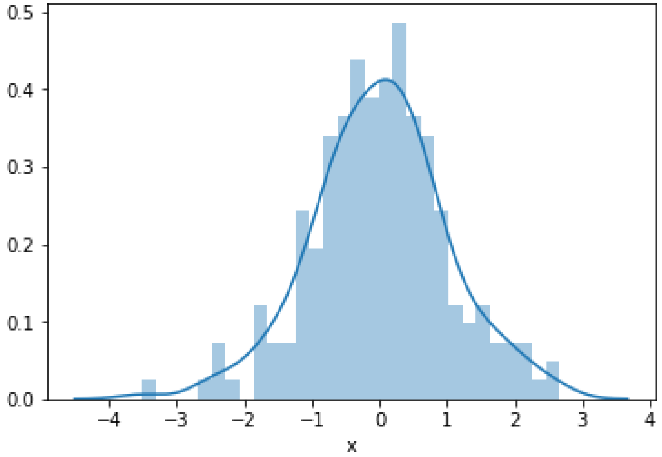
Somewhat linear relationship

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

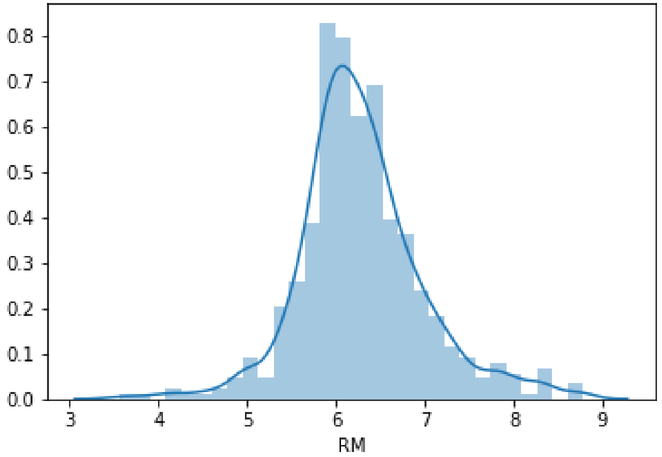
- Residuals are the difference between the predictions and the real value y.

Normality – Histograms

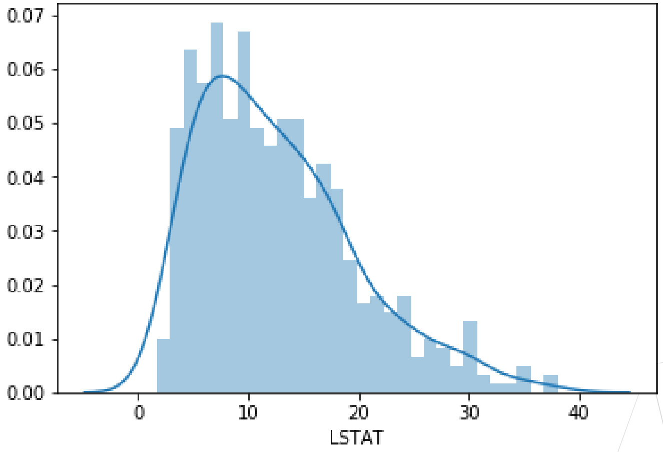
Expected – Simulated data



Somewhat linear relationship (RM)



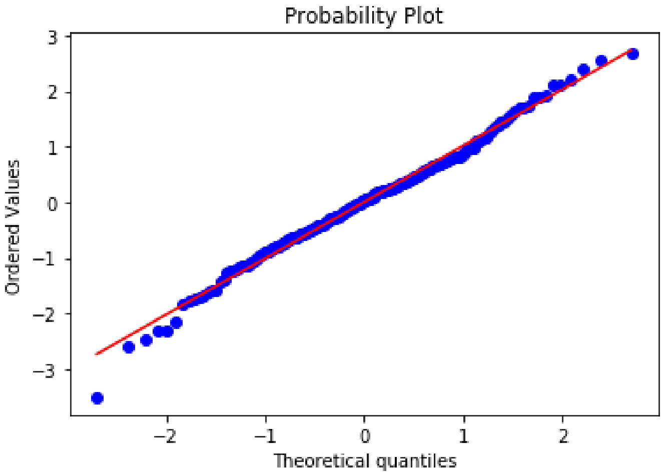
Non-linear relationship (LSTAT)



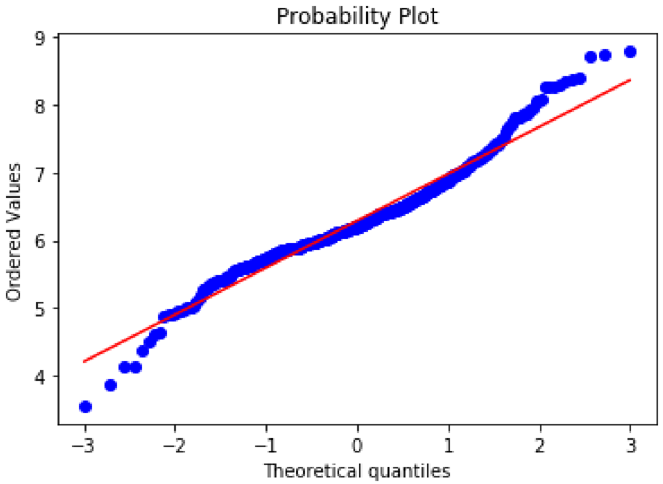
Gaussian distributions adopt a bell shape

Normality – Q-Q plots

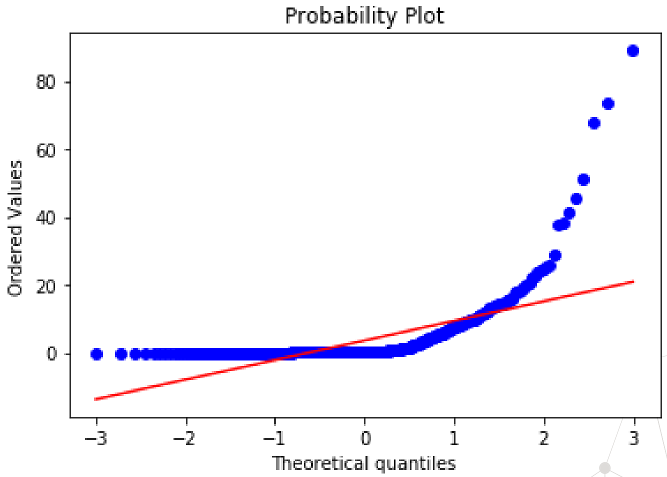
Expected – Simulated data



Somewhat linear relationship



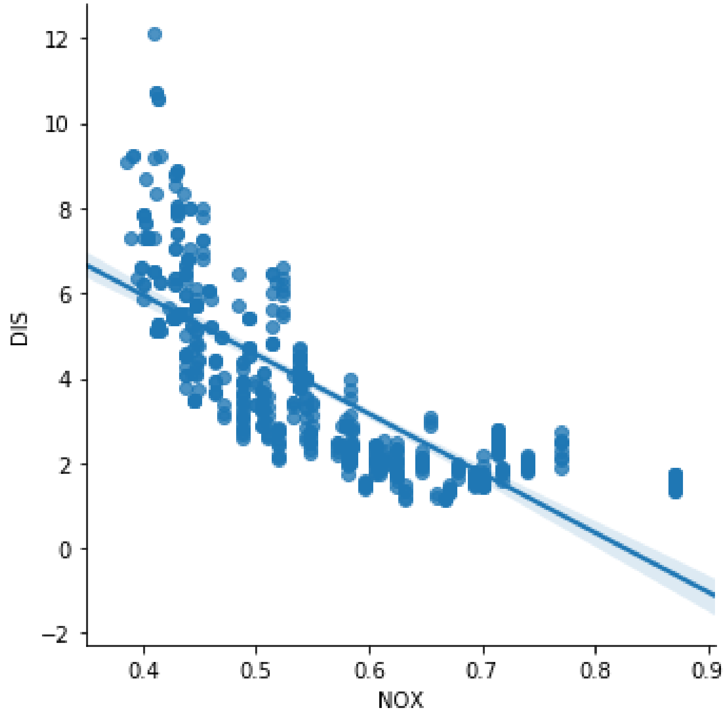
Non-linear relationship

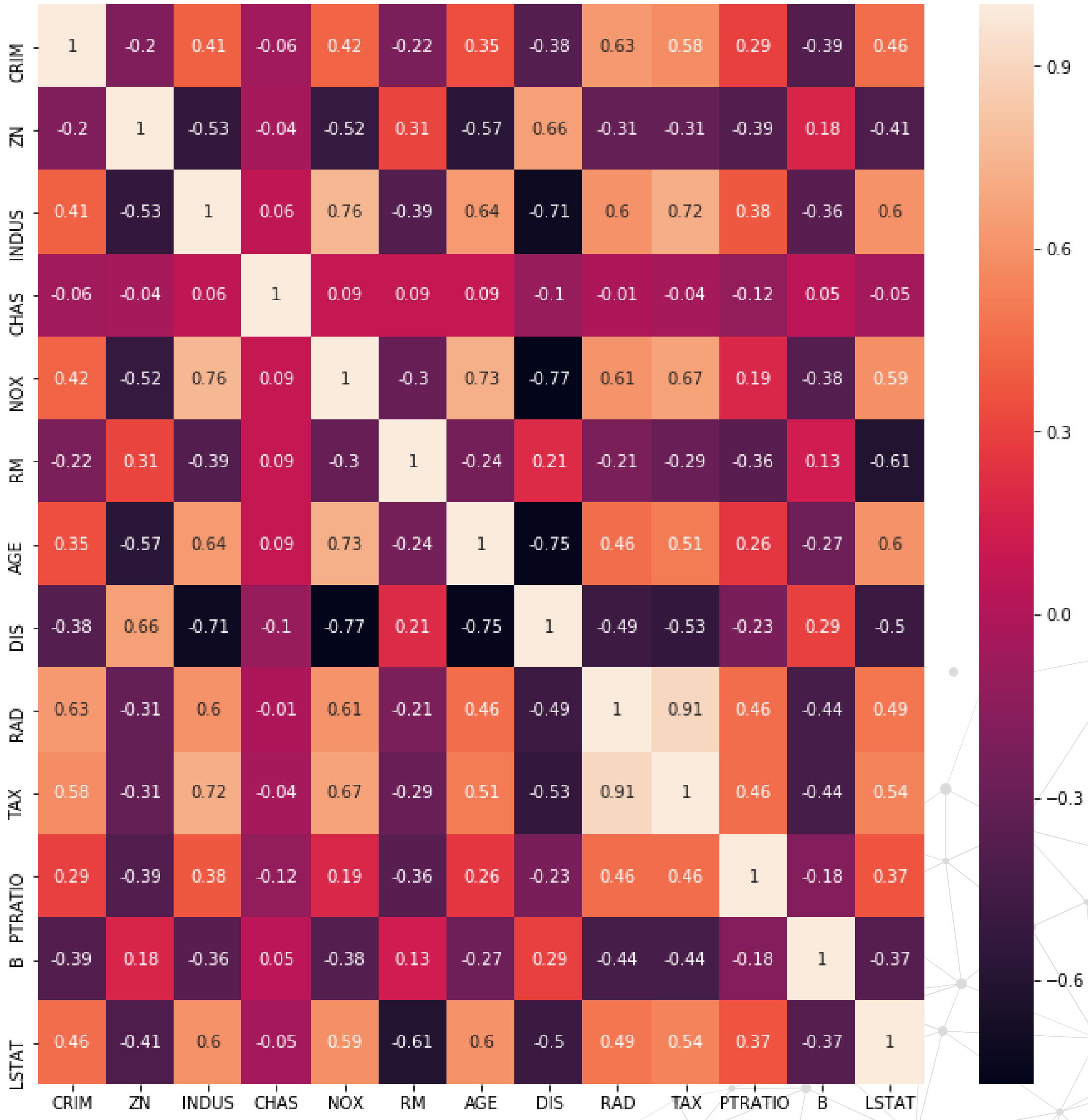


- Q-Q plots plot the variable quantiles in the y-axis and the expected quantiles of the normal distribution on the x-axis.
- If variable is normally distributed, the blue dots should fall on a 45 degree line

Multi Co-linearity

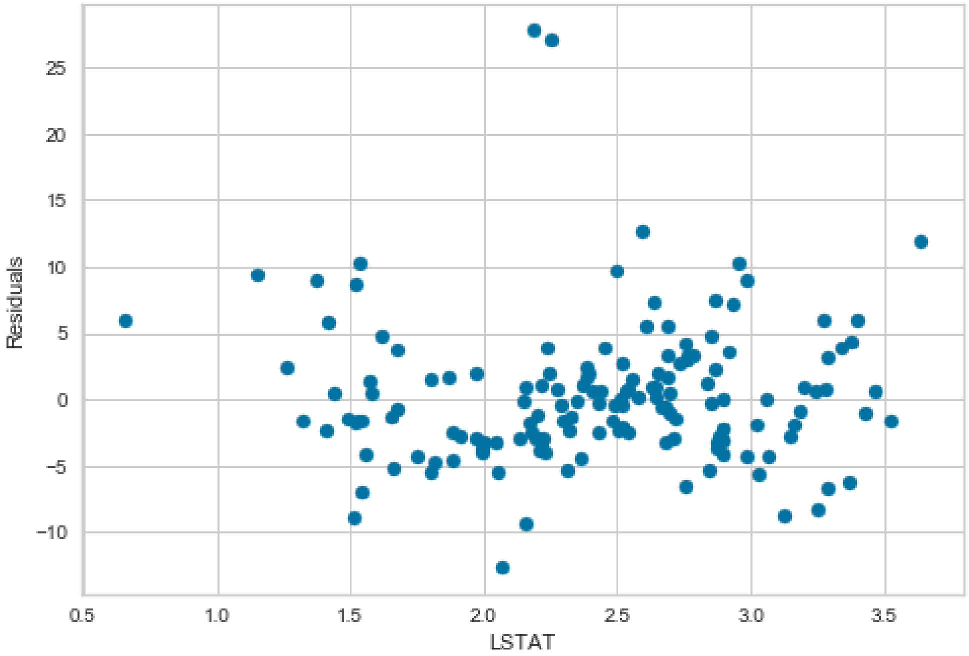
Evaluated by correlation

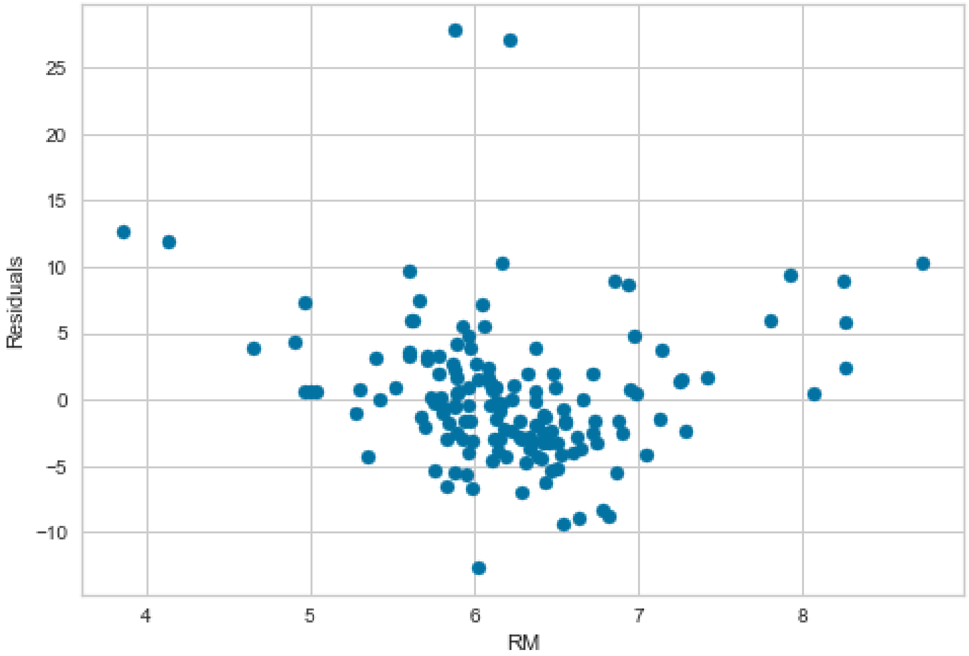




	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
CRIM	1	-0.2	0.41	-0.06	0.42	-0.22	0.35	-0.38	0.63	0.58	0.29	-0.39	0.46
ZN	-0.2	1	-0.53	-0.04	-0.52	0.31	-0.57	0.66	-0.31	-0.31	-0.39	0.18	-0.41
INDUS	0.41	-0.53	1	0.06	0.76	-0.39	0.64	-0.71	0.6	0.72	0.38	-0.36	0.6
CHAS	-0.06	-0.04	0.06	1	0.09	0.09	0.09	-0.1	-0.01	-0.04	-0.12	0.05	-0.05
NOX	0.42	-0.52	0.76	0.09	1	-0.3	0.73	-0.77	0.61	0.67	0.19	-0.38	0.59
RM	-0.22	0.31	-0.39	0.09	-0.3	1	-0.24	0.21	-0.21	-0.29	-0.36	0.13	-0.61
AGE	0.35	-0.57	0.64	0.09	0.73	-0.24	1	-0.75	0.46	0.51	0.26	-0.27	0.6
DIS	-0.38	0.66	-0.71	-0.1	-0.77	0.21	-0.75	1	-0.49	-0.53	-0.23	0.29	-0.5
RAD	0.63	-0.31	0.6	-0.01	0.61	-0.21	0.46	-0.49	1	0.91	0.46	-0.44	0.49
TAX	0.58	-0.31	0.72	-0.04	0.67	-0.29	0.51	-0.53	0.91	1	0.46	-0.44	0.54
PTRATIO	0.29	-0.39	0.38	-0.12	0.19	-0.36	0.26	-0.23	0.46	0.46	1	-0.18	0.37
B	-0.39	0.18	-0.36	0.05	-0.38	0.13	-0.27	0.29	-0.44	-0.44	-0.18	1	-0.37
LSTAT	0.46	-0.41	0.6	-0.05	0.59	-0.61	0.6	-0.5	0.49	0.54	0.37	-0.37	1

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.

The distributions should be similar.

Accompanying Jupyter Notebook

- Read the accompanying Jupyter Notebook
- Full demonstration of the linear assumptions and the influence of non-linear transformations

Train In Data

THANK YOU

www.trainindata.com