Correlation & Regression

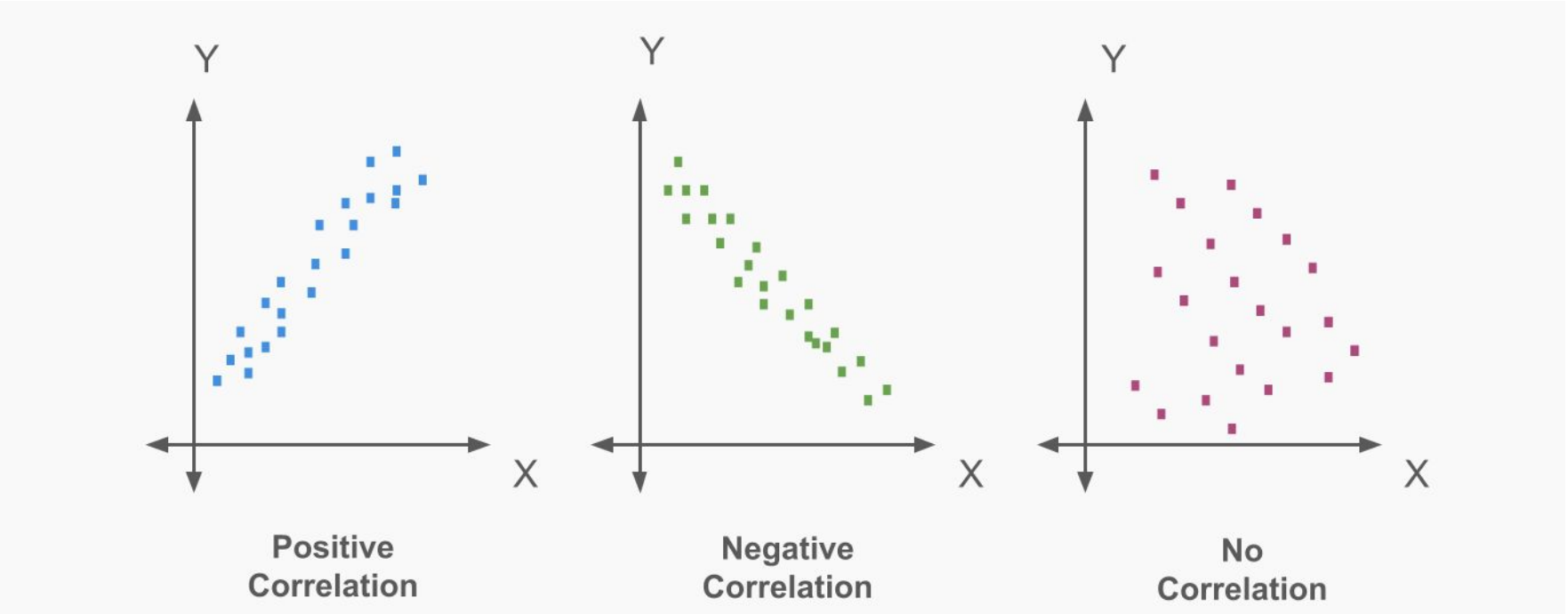
Correlation

- Correlation measures the *strength* and *direction* of a linear relationship between two numerical variables.
- Value ranges from -1 to +1:
  - +1: Perfect positive relationship
  - 0: No linear relationship
  - –1: Perfect negative relationship
- Only measures a *linear* correlation of variables

Scatter plots

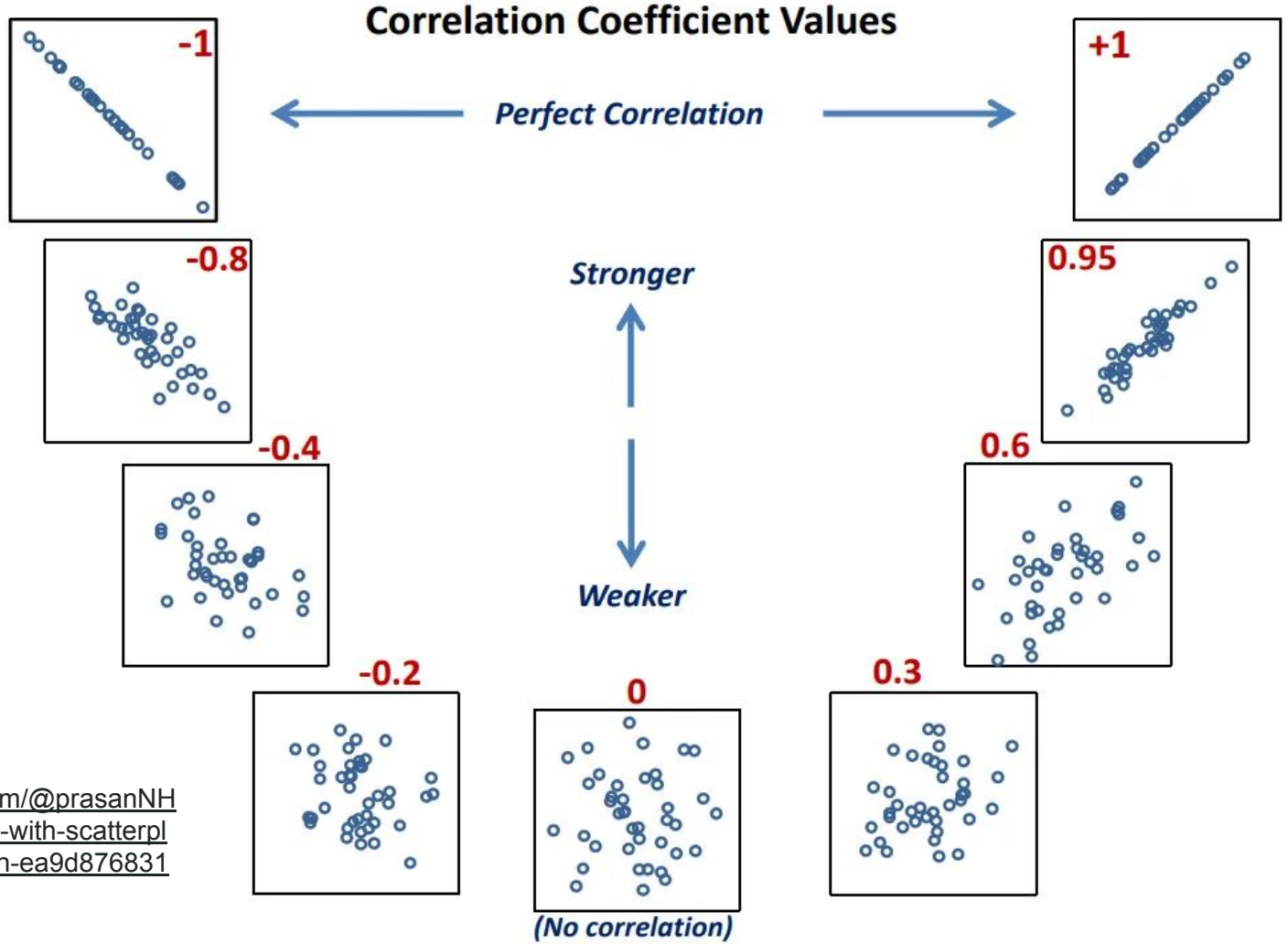
Plot values of one variable against values of another

Scatter plots



https://planyway.com/blog/how-to-make-a-scatter-plot

Scatte



https://medium.com/@prasanNH/unveiling-insights-with-scatterplots-and-correlation-ea9d876831a5

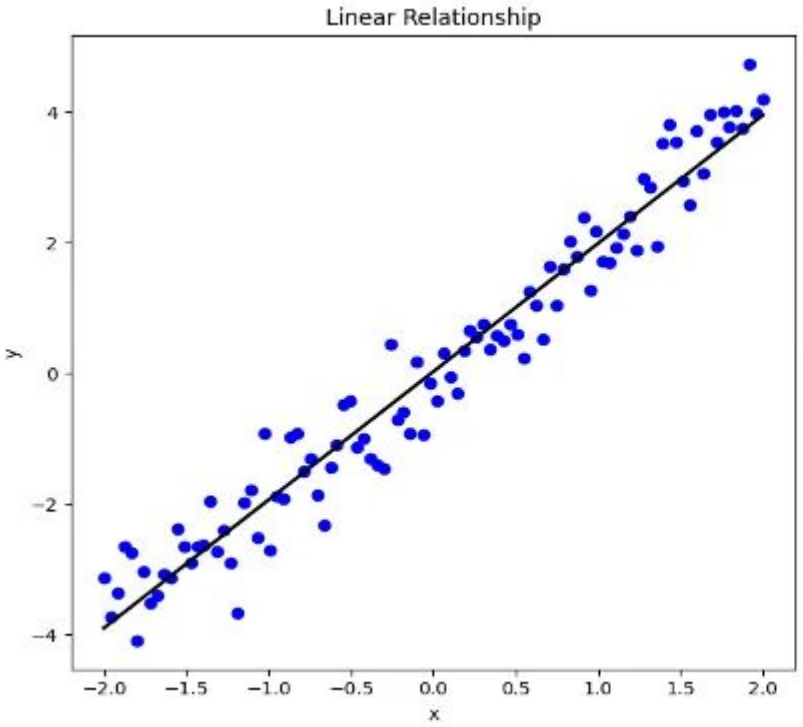
Scatter plots visualize correlation

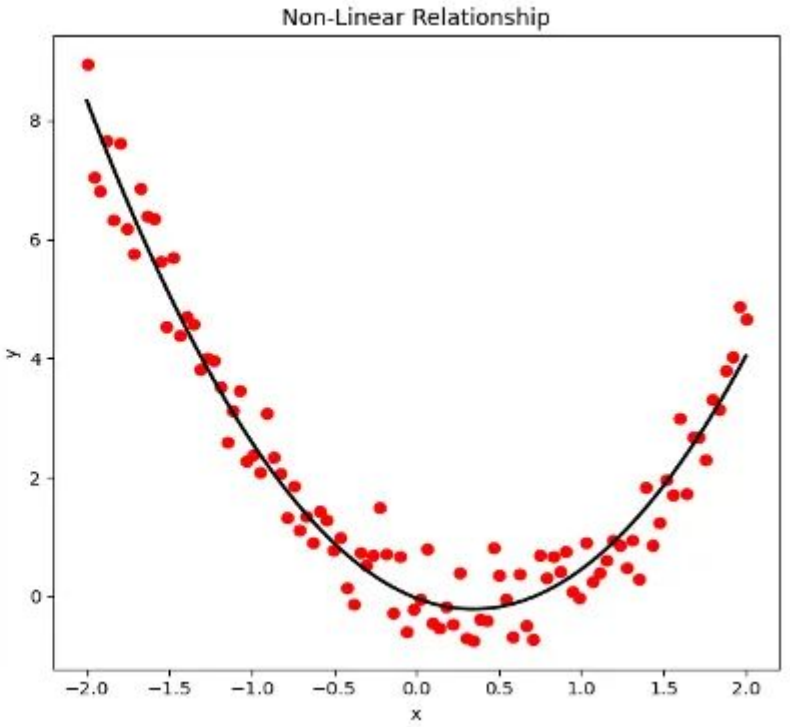
https://rpsychologist.com/correlation/

Scatter plots to check linear relationship

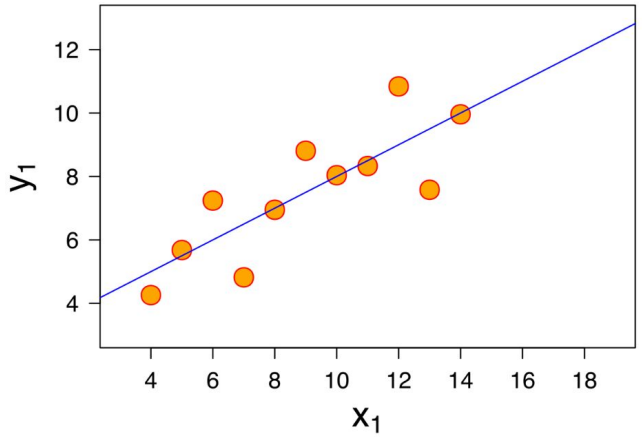
- Clusters or outliers
- Non-linear patterns

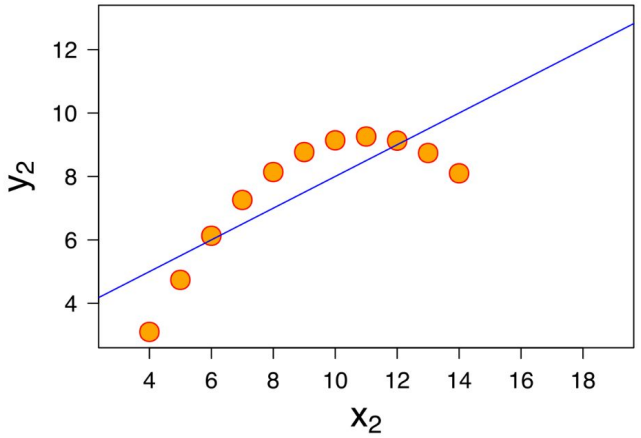
Check linear relationship

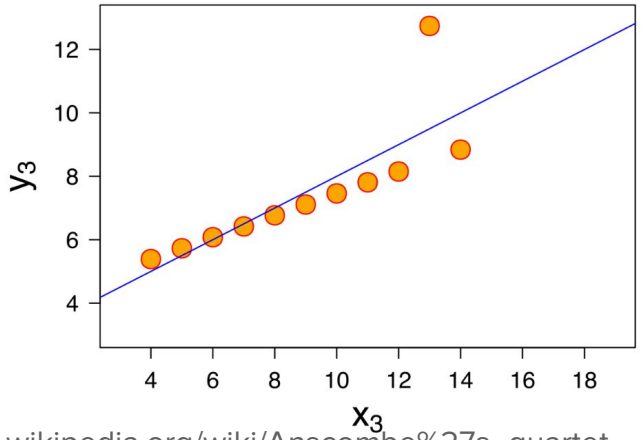


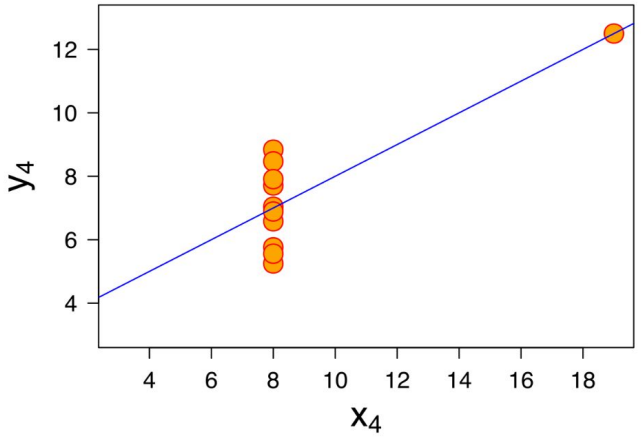


https://medium.com/@krypsa/understanding-linear-vs-nonlinear-relationships-in-data-science-45c05dd2d357



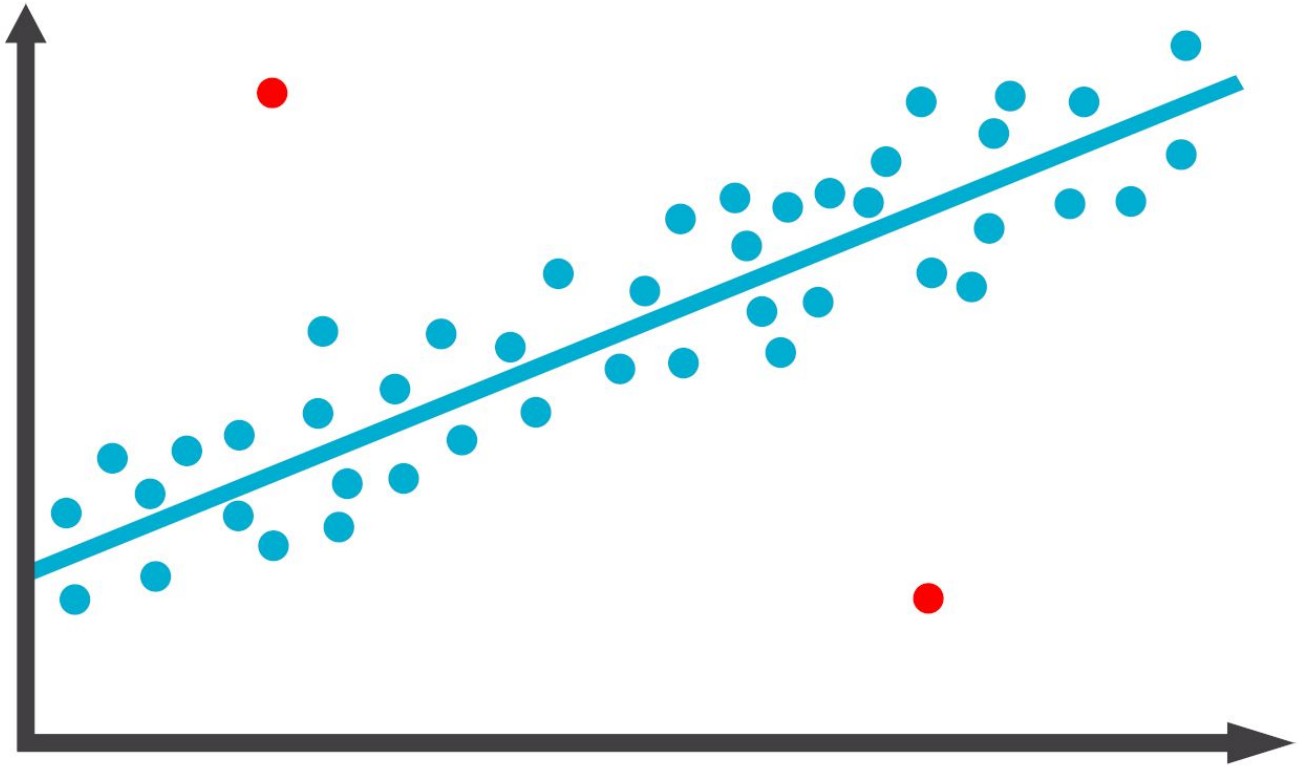


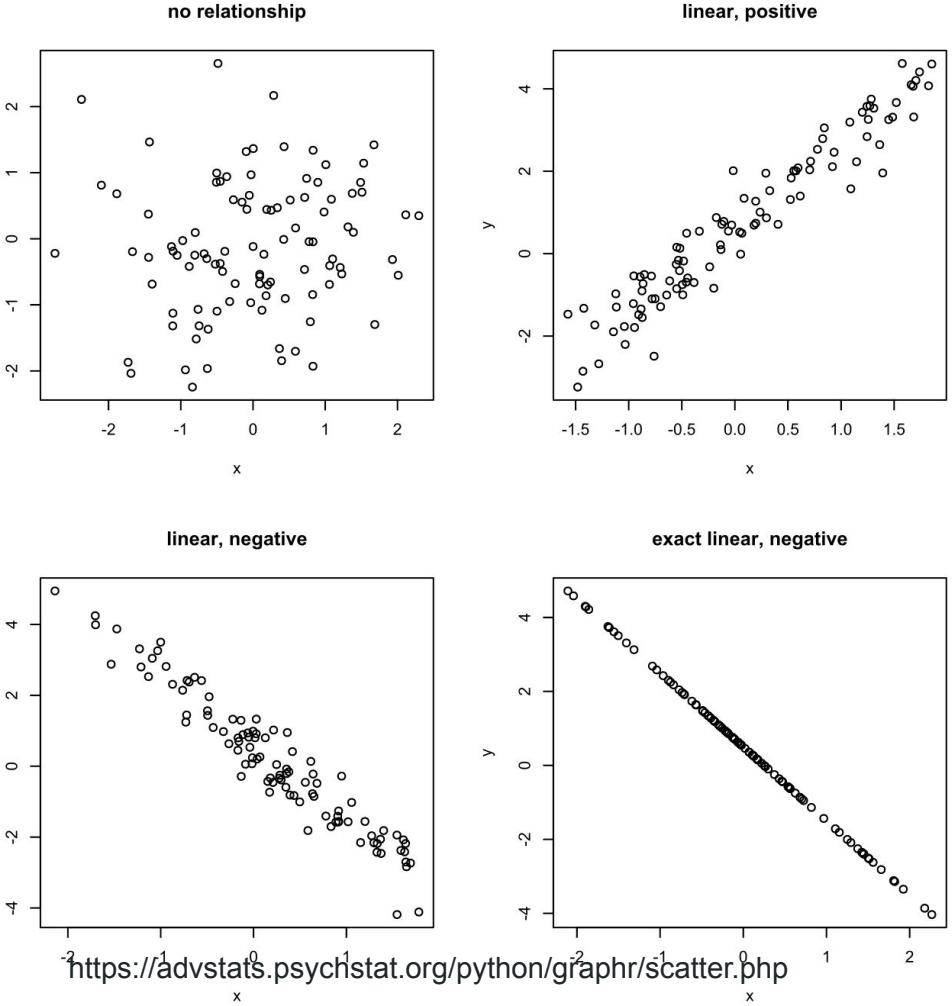




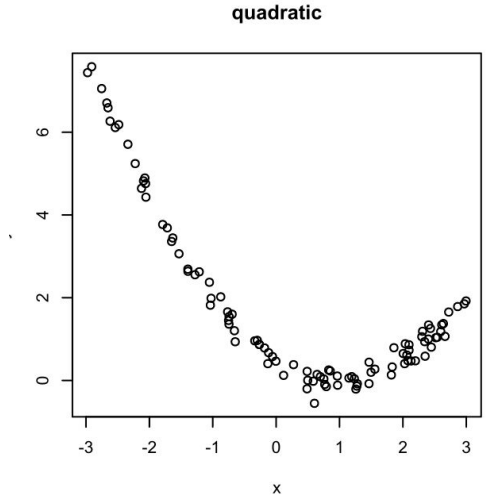
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

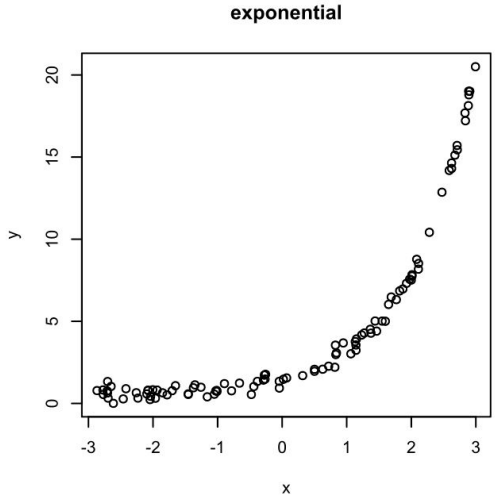
Check for outliers

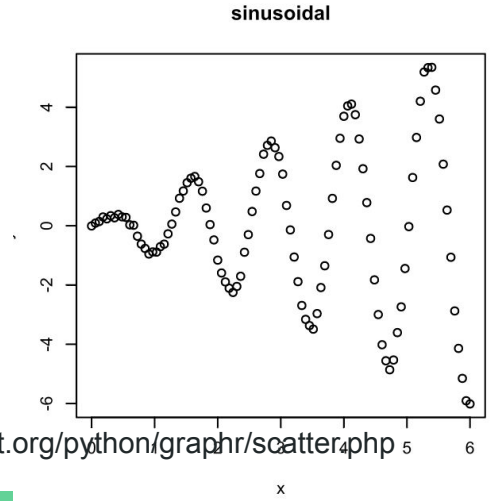


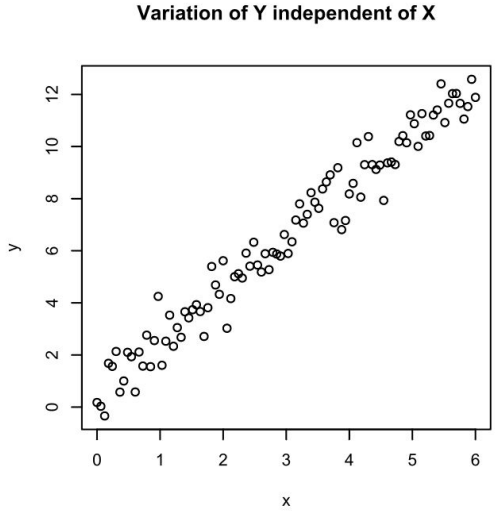


https://advstats.psychstat.org/python/graphr/scatter.php

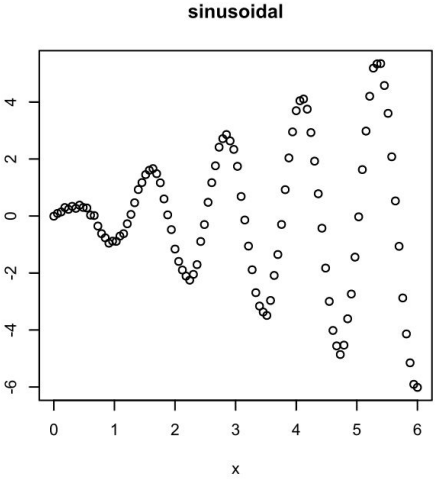


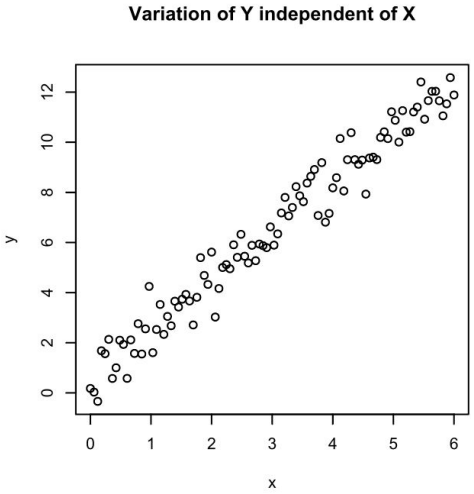


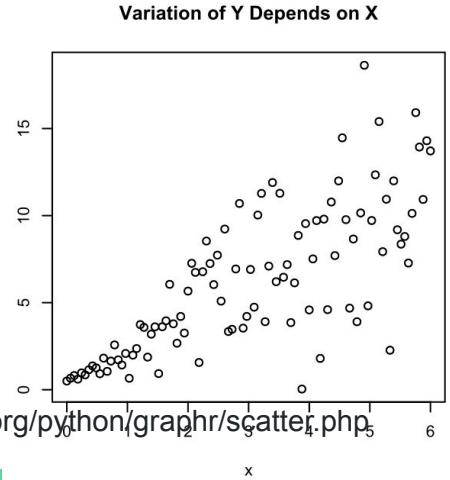


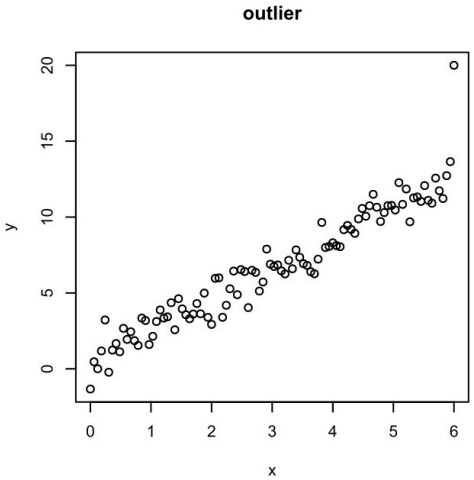


https://advstats.psychstat.org/python/graphr/scatter.php









https://advstats.psychstat.org/python/graphr/scatter.php

Correlation does not equal causation

Correlation ≠ causation

- The *correlation* statistic only measures the relationship between variables
- It does *not* assume any functional relationship

Correlation ≠ causation

- Consider confounding variables and spurious relationships

Regression

Regression

Regression models the relationship
between a numeric dependent variable (Y)
and one (or more) independent variables (X)

Terminology

Y

= dependent variable

= outcome variable

= response variable

= endogenous variable

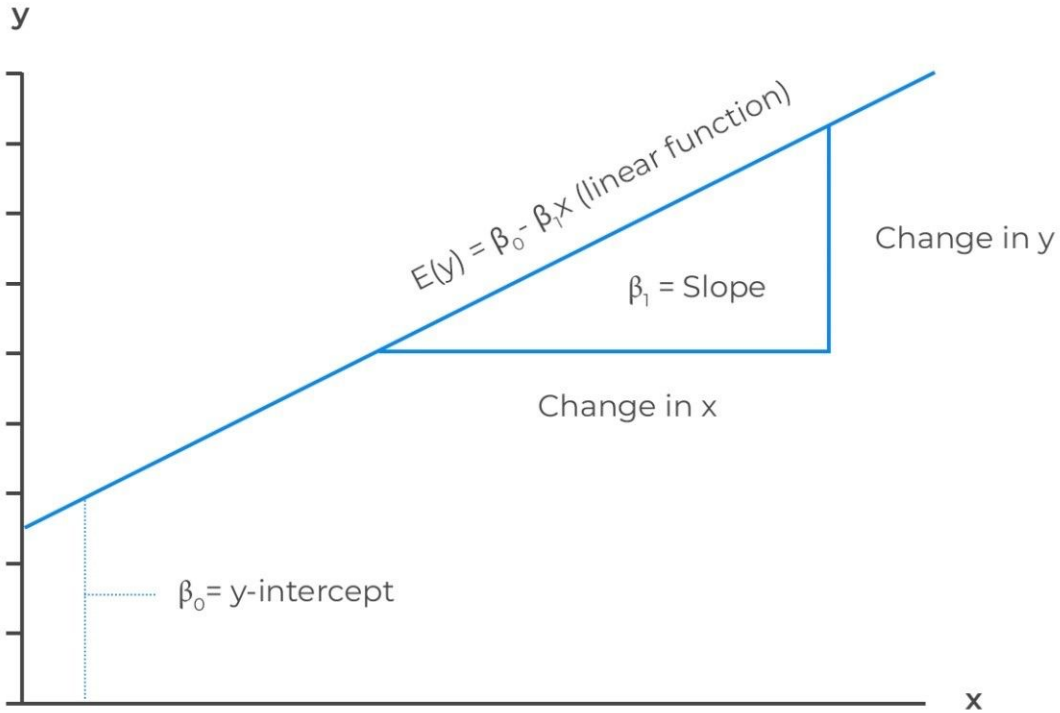
Terminology

X
= independent variable
= predictor
= regressor
= explanatory variable
= exogenous variable
= feature

Linear regression: Functional form

$$Y = \beta_o + \beta_1 X + \epsilon$$

- Y :outcome; dependent variable
- X : predictor; independent variable
- β_o : intercept
- β_1 : slope (change in Y per unit change in X)
- ϵ : error (unexplained variation)



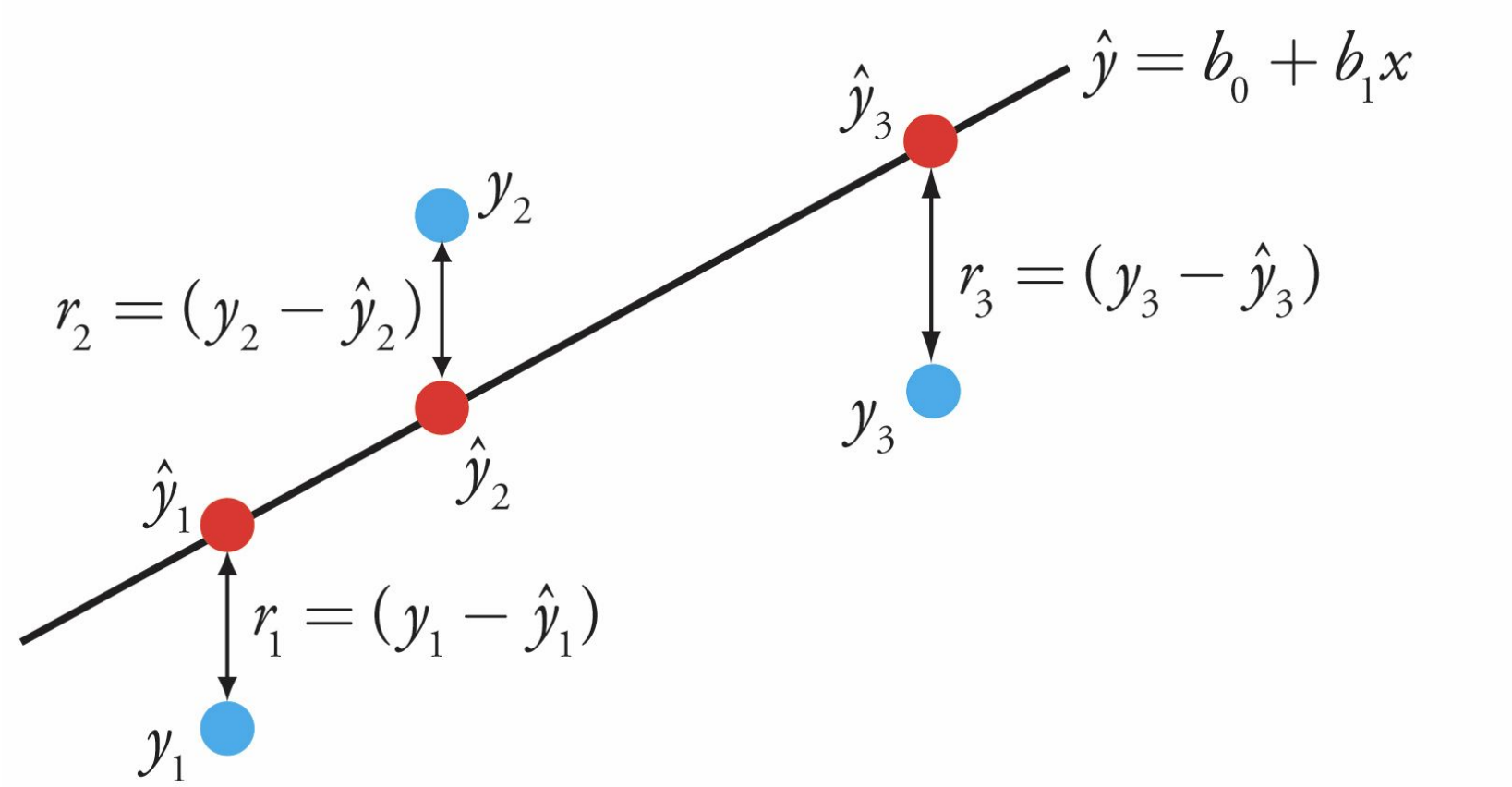
https://analystprep.com/cfa-level-1-exam/quantitative-methods/introduction-to-linear-regression/

ϵ: error

- Variation in the outcome variable unexplained by the model
- Random variation + excluded factors

Residuals

: the differences between the observed values and the values predicted by a model



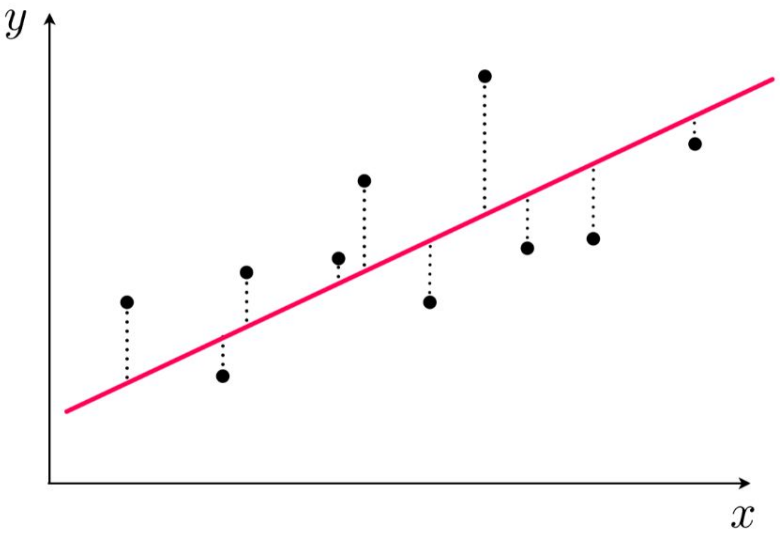
https://chem.libretexts.org/Courses/Duke_University/CHEM_310L%3A_Physical_Chemistry_I_Laboratory/CHEM310L_-_Physical_Chemistry_I_Lab_Manual/03%3A_Estimating_and_Reporting_Experimental_Error/3.04%3A_Least_Squares_Linear_Regression

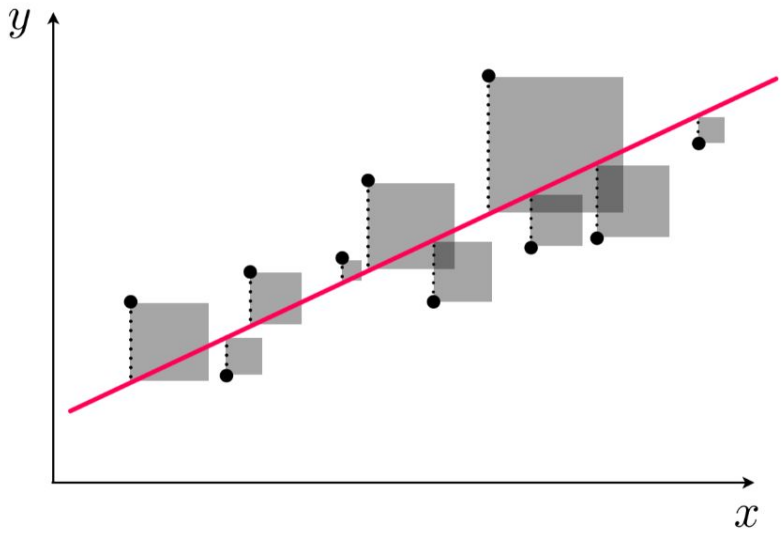
R^2

Amount of variance explained by the model

- Inversely related to sum of square residuals
- Better fit, higher R^2

Ordinary least squares





https://kenndanielso.github.io/mlrefined/blog_posts/8_Linear_regression/8_1_Least_squares_regression.html

Ordinary least squares

```
formula = 'outcome_column ~ predictor_column'

model = smf.ols(data=data, formula=formula)

result = model.fit()
```

OLS Regression Results

Dep. Variable:	Q("Total GHG Emissions (Metric Tons CO2e)")	R-squared:	0.607
Model:	OLS	Adj. R-squared:	0.607
Method:	Least Squares	F-statistic:	3.936e+04
Date:	Sat, 05 Apr 2025	Prob (F-statistic):	0.00
Time:	13:50:47	Log-Likelihood:	-1.7106e+05
No. Observations:	25455	AIC:	3.421e+05
Df Residuals:	25453	BIC:	3.421e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	59.0695	2.033	29.059	0.000	55.085	63.054
Q("Property GFA - Calculated (Buildings) (ft²)")	0.0040	2e-05	198.393	0.000	0.004	0.004

Omnibus:	8266.733	Durbin-Watson:	1.709
Prob(Omnibus):	0.000	Jarque-Bera (JB):	100921.120
Skew:	1.212	Prob(JB):	0.00
Kurtosis:	12.449	Cond. No.	1.64e+05

OLS Regression Results

Dep. Variable:	Q("Total GHG Emissions (Metric Tons CO2e)")	R-squared:	0.607
Model:	OLS	Adj. R-squared:	0.607
Method:	Least Squares	F-statistic:	3.936e+04
Date:	Sat, 05 Apr 2025	Prob (F-statistic):	0.00
Time:	13:50:47	Log-Likelihood:	-1.7106e+05
No. Observations:	25455	AIC:	3.421e+05
Df Residuals:	25453	BIC:	3.421e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	59.0695	2.033	29.059	0.000	55.085	63.054
Q("Property GFA - Calculated (Buildings) (ft²)")	0.0040	2e-05	198.393	0.000	0.004	0.004

Omnibus:	8266.733	Durbin-Watson:	1.709
Prob(Omnibus):	0.000	Jarque-Bera (JB):	100921.120
Skew:	1.212	Prob(JB):	0.00
Kurtosis:	12.449	Cond. No.	1.64e+05

OLS Regression Results

Dep. Variable:	Q("Total GHG Emissions (Metric Tons CO2e)")	R-squared:	0.607
Model:	OLS	Adj. R-squared:	0.607
Method:	Least Squares	F-statistic:	3.936e+04
Date:	Sat, 05 Apr 2025	Prob (F-statistic):	0.00
Time:	13:50:47	Log-Likelihood:	-1.7106e+05
No. Observations:	25455	AIC:	3.421e+05
Df Residuals:	25453	BIC:	3.421e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	59.0695	2.033	29.059	0.000	55.085	63.054
Q("Property GFA - Calculated (Buildings) (ft²)")	0.0040	2e-05	198.393	0.000	0.004	0.004

Omnibus:	8266.733	Durbin-Watson:	1.709
Prob(Omnibus):	0.000	Jarque-Bera (JB):	100921.120
Skew:	1.212	Prob(JB):	0.00
Kurtosis:	12.449	Cond. No.	1.64e+05

OLS Regression Results

Dep. Variable:	Q("Total GHG Emissions (Metric Tons CO2e)")	R-squared:	0.607
Model:	OLS	Adj. R-squared:	0.607
Method:	Least Squares	F-statistic:	3.936e+04
Date:	Sat, 05 Apr 2025	Prob (F-statistic):	0.00
Time:	13:50:47	Log-Likelihood:	-1.7106e+05
No. Observations:	25455	AIC:	3.421e+05
Df Residuals:	25453	BIC:	3.421e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	59.0695	2.033	29.059	0.000	55.085	63.054
Q("Property GFA - Calculated (Buildings) (ft²)")	0.0040	2e-05	198.393	0.000	0.004	0.004

Omnibus:	8266.733	Durbin-Watson:	1.709
Prob(Omnibus):	0.000	Jarque-Bera (JB):	100921.120
Skew:	1.212	Prob(JB):	0.00
Kurtosis:	12.449	Cond. No.	1.64e+05

Assumptions for linear regression

- Linearity – the relationship is linear
- Independence – observations are independent
- Homoscedasticity – constant variance of errors
- Normality – residuals/errors are normally distributed

Regression analysis

How to use regression:

1. Prediction
    a. e.g. weather forecasts, Spotify playlists
2. Inference
    a. e.g. does residential density increase subway ridership?