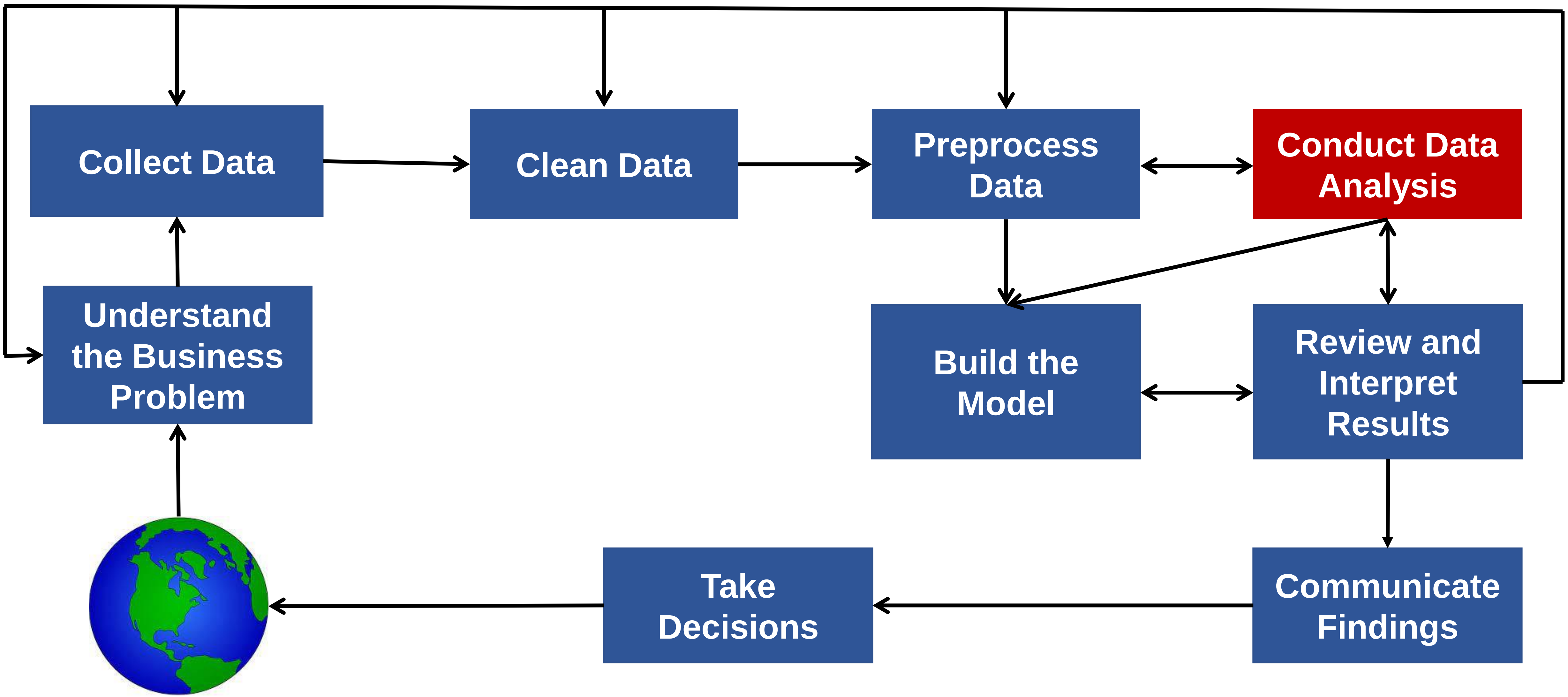
Exploratory Data Analysis

Dr. Sandareka Wickramanayake

sandarekaw@cse.mrt.ac.lk

Data Science Process



Recommended Reading

Chapter 3 - Art of Data Science

Chapter 15 - Head First Statistics

Data Analysis

- Ask good questions.
- Seek answers to those questions.

Descriptive	Seek to summarize characteristics of a set of data.
Exploratory	Seek for patterns, trends, or relationships between variables.
Predictive	Make predictions about future or otherwise unknown events based on data.

Data Analysis

- Ask good questions.
- Seek answers to those questions.

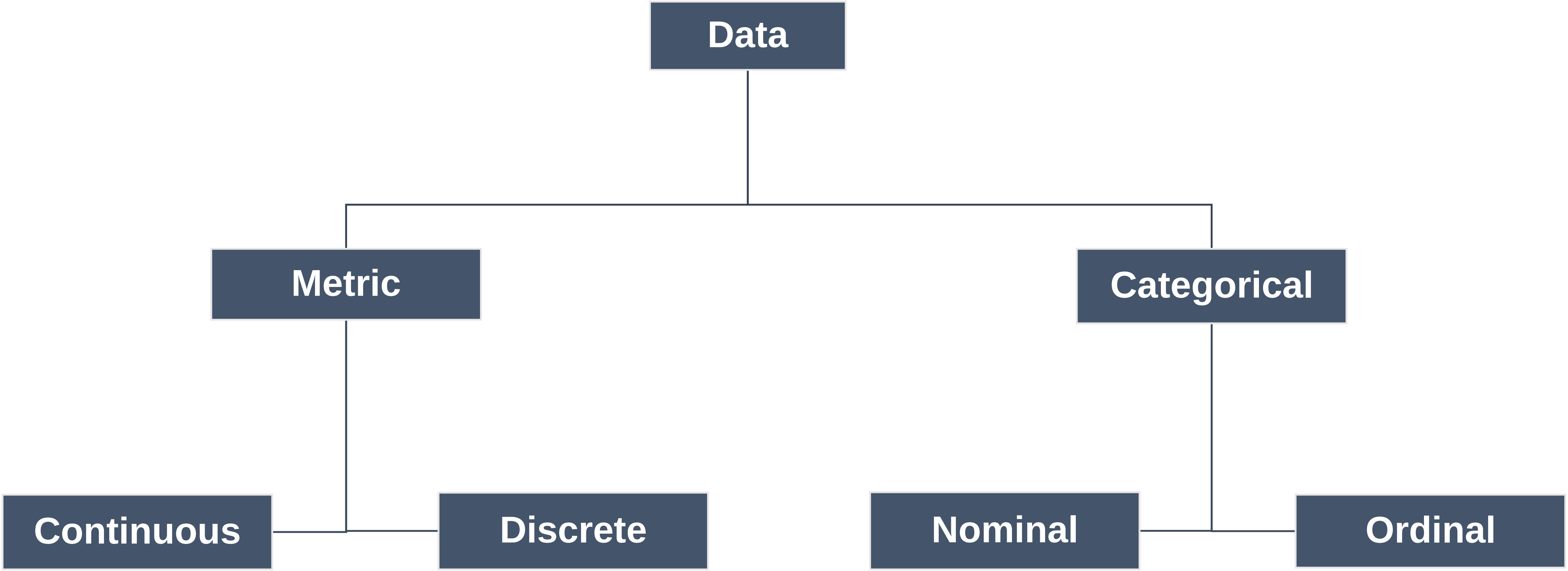
Descriptive Seek to summarize a characteristic of a set of data.

Exploratory Seek for patterns, trends, or relationships between variables.

Predictive Make predictions about future or otherwise unknown events based on data.

Data Types (Recap)

- Four fundamental data types.



Data

Metric

Categorical

Continuous

Discrete

Nominal

Ordinal

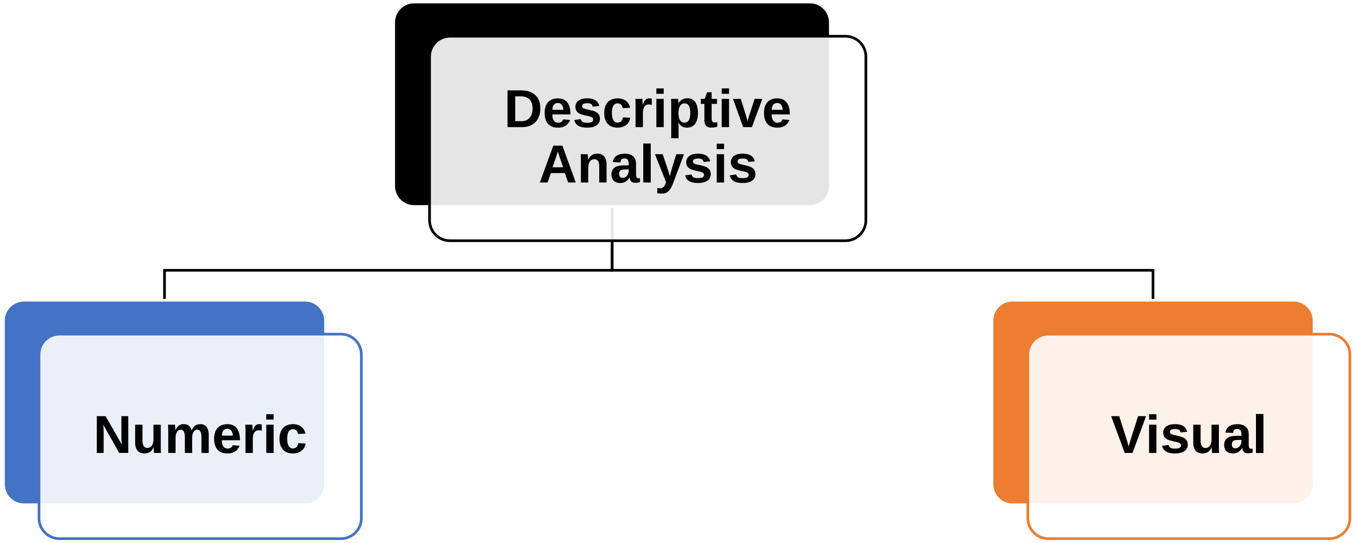
Descriptive Analysis

- Summarizes the properties of the dataset.
- Questions can be answered descriptive analysis:
  - *Determine the proportion of males.*
  - *What is the mean age of the participants?*
  - *What percentage of participants “rarely” uses the seatbelt?*
  - Etc …

- *An important consideration is the type of data!*
  - Step 1: Use data type classification tree.
  - Step 2: Apply appropriate descriptive methods.

Descriptive Analysis

- Two main approaches: Numeric and Visual descriptions.



Descriptive Analysis: Numeric

Attribute Type	Categorical		Metric	
	Nominal	Ordinal	Discrete	Continuous
Frequency	Yes	Yes	Yes	Under grouped representation
Normalized Frequency	Yes	Yes	Yes	Under grouped representation
Cumulative Frequency	No	Yes	Yes	Under grouped representation
Normalized Cumulative Frequency	No	Yes	Yes	Under grouped representation
Mode	Yes	Yes	Yes	No
Mean	No	No	Yes	Yes
Median	No	Yes	Yes	Yes
Range	No	No	Yes	Yes
Spread	No	No	Yes	Yes
Five Number Summary	No	No	Yes	Yes

Descriptive Analysis: Visual

Attribute Type	Categorical		Metric	
	Nominal	Ordinal	Discrete	Continuous
Pie Chart	Yes	Yes	No	No
Tag Cloud	Yes	Yes	Possible	No
Bar Chart	Yes	Yes	Yes	No
Clustered/Stacked Bar Chart	Yes	Yes	Yes	No
Step Chart	No	Yes	Yes	No
Box Plot	No	No	Yes	Yes
Histogram	No	No	Yes	Yes
Cumulative Histogram	No	No	Yes	Yes

Data Analysis

- Ask good questions.
- Seek answers to those questions.

Descriptive Seek to summarize a characteristic of a set of data.

Exploratory Seek for patterns, trends, or relationships between variables.

Predictive Make predictions about future or otherwise unknown events based on data.

Exploratory Data Analysis

- **Association**
- **Correlation**
- **Agreement**

Exploratory Data Analysis - **Association**

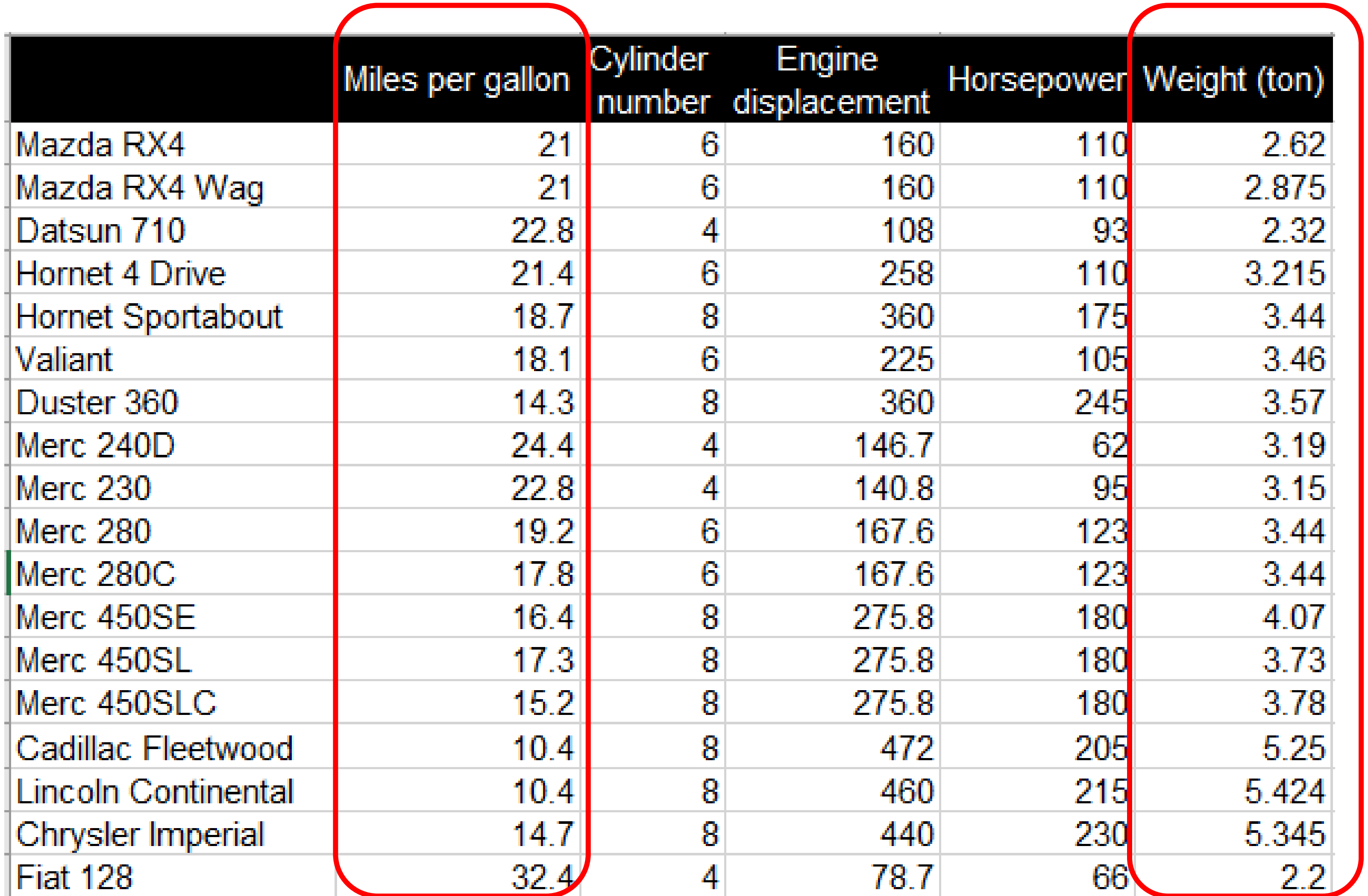
- Do changes in X (seem to) coincide with changes in Y?
- This does not mean changes in X cause changes in Y!

- Metric – Scatter plots
- Categorical – Contingency Table

Exploratory Data Analysis - **Association**

- Do changes in X (seem to) coincide with changes in Y?
- Example question – *Lighter car seems to run longer?*

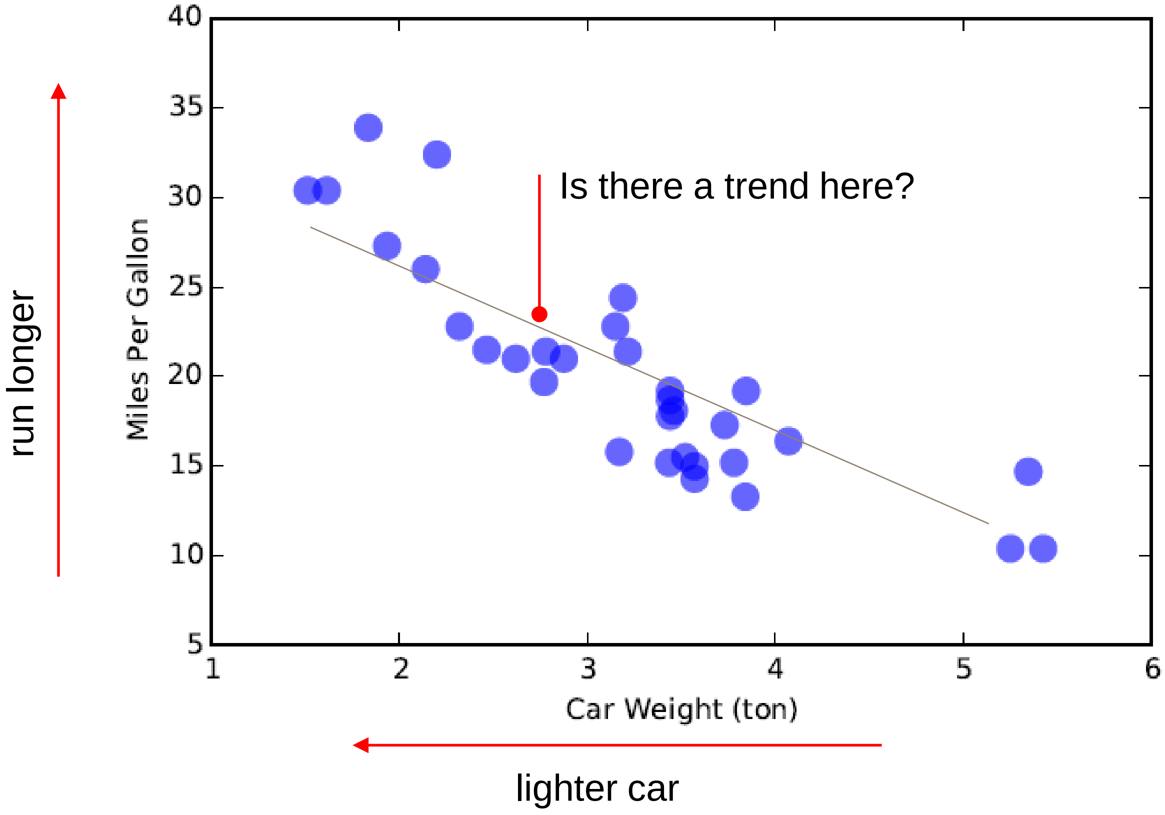
1974 Motor Trend Data



	Miles per gallon	Cylinder number	Engine displacement	Horsepower	Weight (ton)
Mazda RX4	21	6	160	110	2.62
Mazda RX4 Wag	21	6	160	110	2.875
Datsun 710	22.8	4	108	93	2.32
Hornet 4 Drive	21.4	6	258	110	3.215
Hornet Sportabout	18.7	8	360	175	3.44
Valiant	18.1	6	225	105	3.46
Duster 360	14.3	8	360	245	3.57
Merc 240D	24.4	4	146.7	62	3.19
Merc 230	22.8	4	140.8	95	3.15
Merc 280	19.2	6	167.6	123	3.44
Merc 280C	17.8	6	167.6	123	3.44
Merc 450SE	16.4	8	275.8	180	4.07
Merc 450SL	17.3	8	275.8	180	3.73
Merc 450SLC	15.2	8	275.8	180	3.78
Cadillac Fleetwood	10.4	8	472	205	5.25
Lincoln Continental	10.4	8	460	215	5.424
Chrysler Imperial	14.7	8	440	230	5.345
Fiat 128	32.4	4	78.7	66	2.2

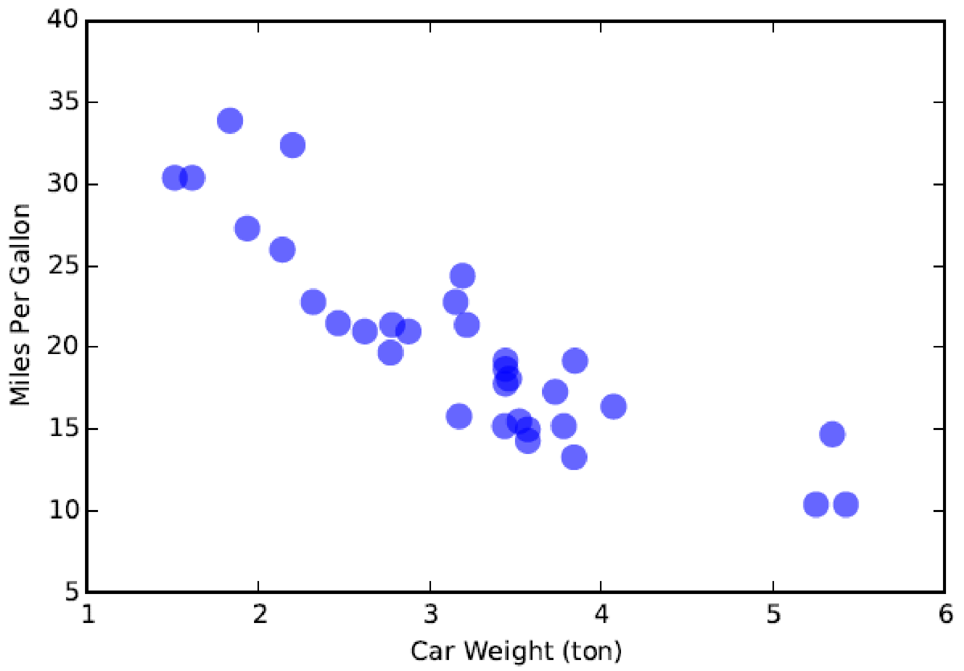
Exploratory Data Analysis - **Association**

- Do changes in X (seem to) coincide with changes in Y?
- Example question – *Lighter car seems to run longer?*
- Scatter plots



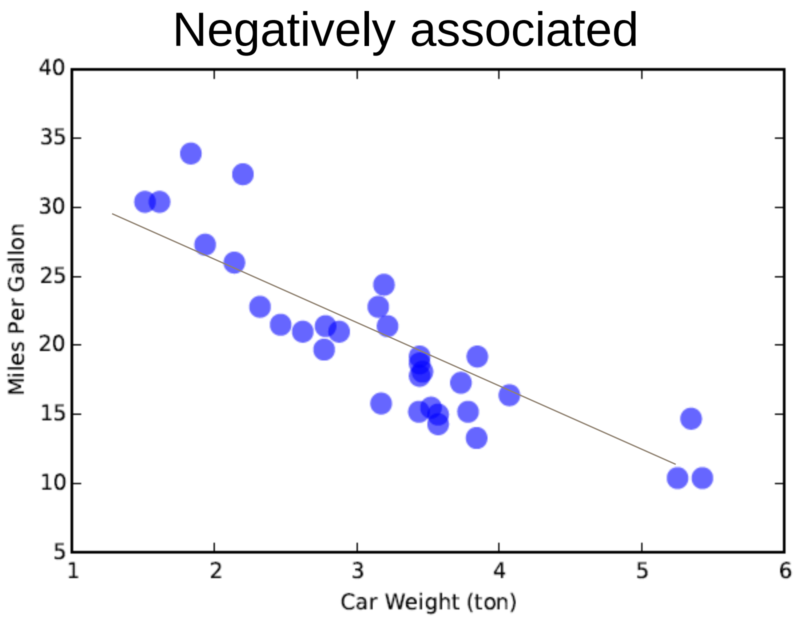
Exploratory Data Analysis - Association

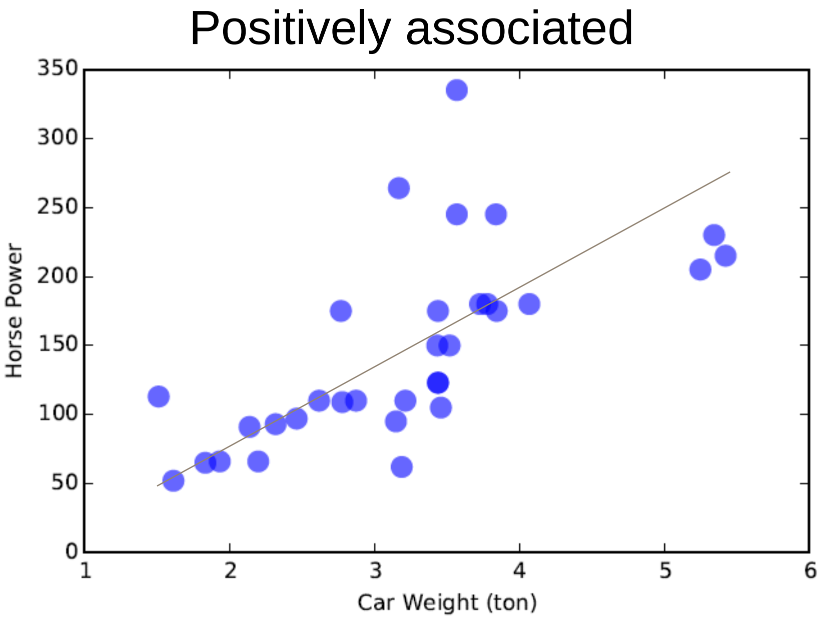
- Do changes in X (seem to) coincide with changes in Y?
- Example question – *Lighter car seems to run longer?*
- Scatter plots
    - Enables the visual inspection of association between variables.
    - Attribute values determine the position.
    - Two dimensional scatter plots are useful to understand the relationship between two (or more) **continuous variables**.
    - We can create three-dimensional scatter plots.
    - *First step* in examining relationships among continuous variables.



Exploratory Data Analysis - **Association**

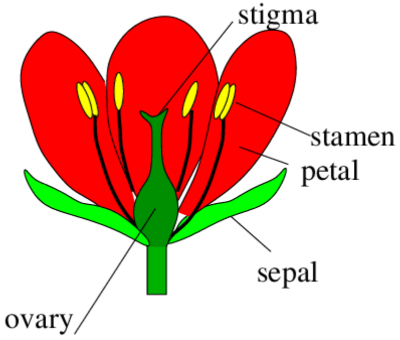
- Do changes in X (seem to) coincide with changes in Y?
- Example question – *Lighter car seems to run longer?*
- Scatter plots
  - Roughly three types of associations.
    - No association, Positively associated, Negatively associated


Negatively associated

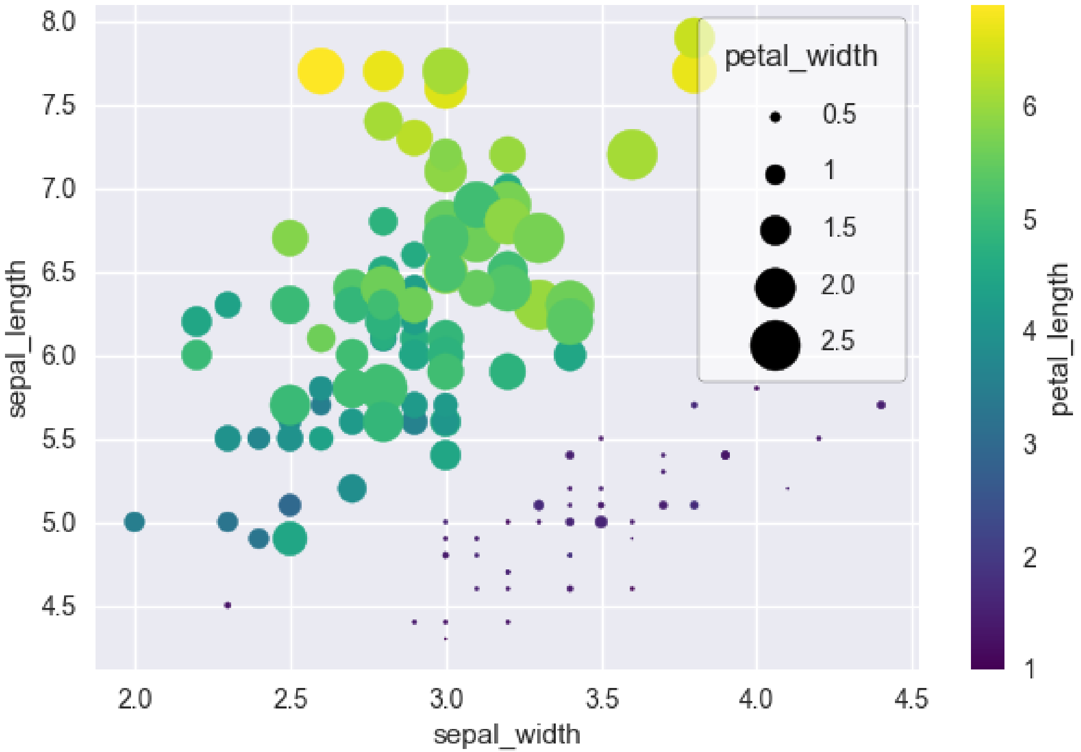

Positively associated

Exploratory Data Analysis - **Association**

- Do changes in X (seem to) coincide with changes in Y?
- Scatter plots
    - Additional attributes can be displayed by using the size, shape, and colour of the markers that represent objects.



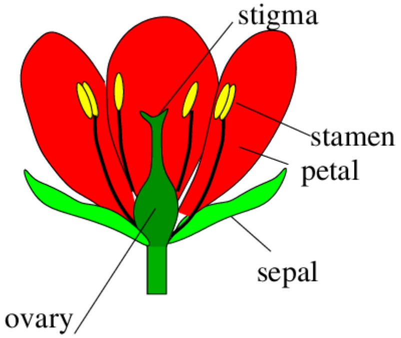
https://www.researchgate.net/publication/265877256_How_plants_grow_and_move



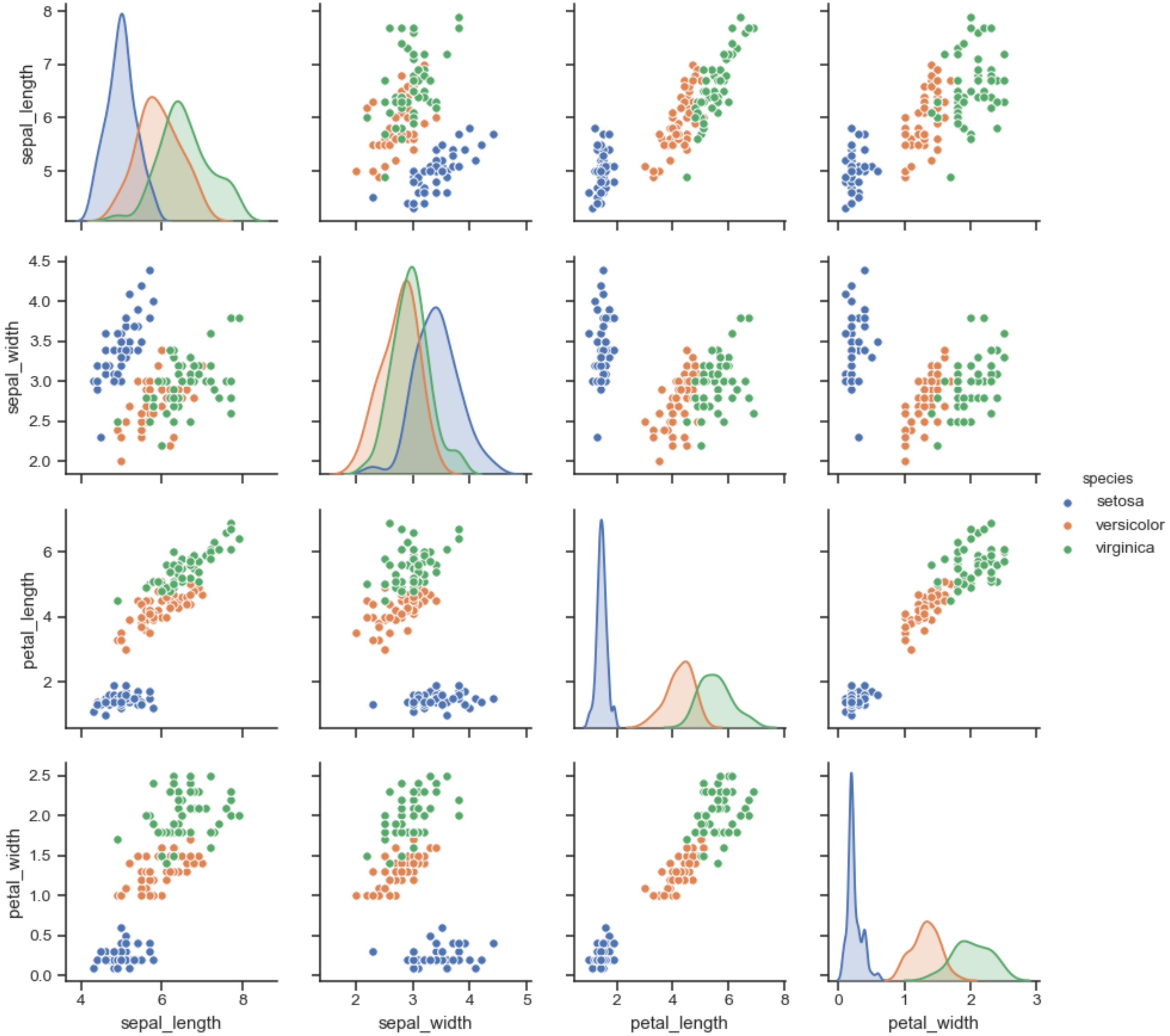
https://stackoverflow.com/questions/42754458/scatter-plots-in-seaborn-matplotlib-with-point-size-and-color-given-by-continuou

Exploratory Data Analysis - **Association**

- Scatter plots
  - Arrays of scatter plots are useful when we want to compactly summarize the relationships of several pairs of attributes.



https://www.researchgate.net/publication/265877256_How_plants_grow_and_move

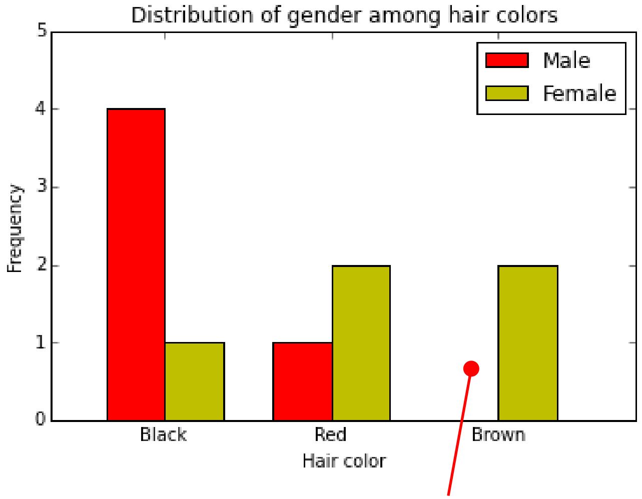


https://stats.stackexchange.com/questions/429920/how-can-i-leverage-a-pairwise-scatter-plot-to-help-choose-a-machine-learning-mod

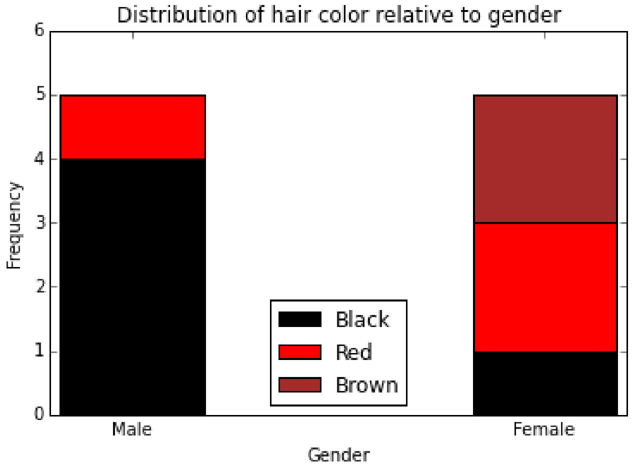
Exploratory Data Analysis - **Association**

- Contingency Table

		Gender	
		Male	Female
Hair Color	Black	4	1
	Red	1	2
	Brown	0	2



No association between Male with Brown hair



Exploratory Data Analysis

- **Association**
- **Correlation**
- **Agreement**

Exploratory Data Analysis - **Correlation**

- How to quantity the association between X and Y?
  - Measure the degree to which X and Y co-behave.

- There are many metrics to measure correlation
  - Check the types of the attributes you want to consider
  - Check the distribution of each attribute
  - Check the association of the attributes
  - Check the assumptions of each correlation metric

Exploratory Data Analysis - **Correlation**

Comparison	Test
Relationship between 2 continuous variables	Pearson correlation (When the relationship is linear) Spearman's Correlation Coefficient
Relationship between 2 discrete variables	Pearson correlation (When the relationship is linear) *Spearman's Correlation Coefficient*
Influence of one or more categorical variables on a continuous variable	ANOVA test
Relationship between a continuous variable and binary categorical variable	Point-biserial correlation (A special case of Pearson correlation)
Relationship between 2 ordinal variables	Spearman's Correlation Coefficient Kendall's rank-order correlation coefficient
Relationship between 2 categorical variables	Chi-squared test

https://stats.oarc.ucla.edu/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/

Exploratory Data Analysis - **Correlation**

- Pearson correlation
  - Quantify the association with a numeric measure of strength.

  - Given data $(X_i, Y_i), i = 1,2, \dots, n$
  - Step 1: compute the means for X and Y

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \,, \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

  - Step 2: compute the standard deviation for X and Y

$$\sigma_X = \sqrt{\frac{1}{n}\sum_{i}^{n}(X_i - \bar{X})^2} \quad \sigma_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

  - Step 3: compute the covariance of X and Y

$$\mathrm{cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

  - Step 4: compute the Pearson correlation:

$$\rho(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Exploratory Data Analysis - **Correlation**

- Pearson correlation
  - Quantify the association with a numeric measure of strength.

Pearson Correlation Coefficient

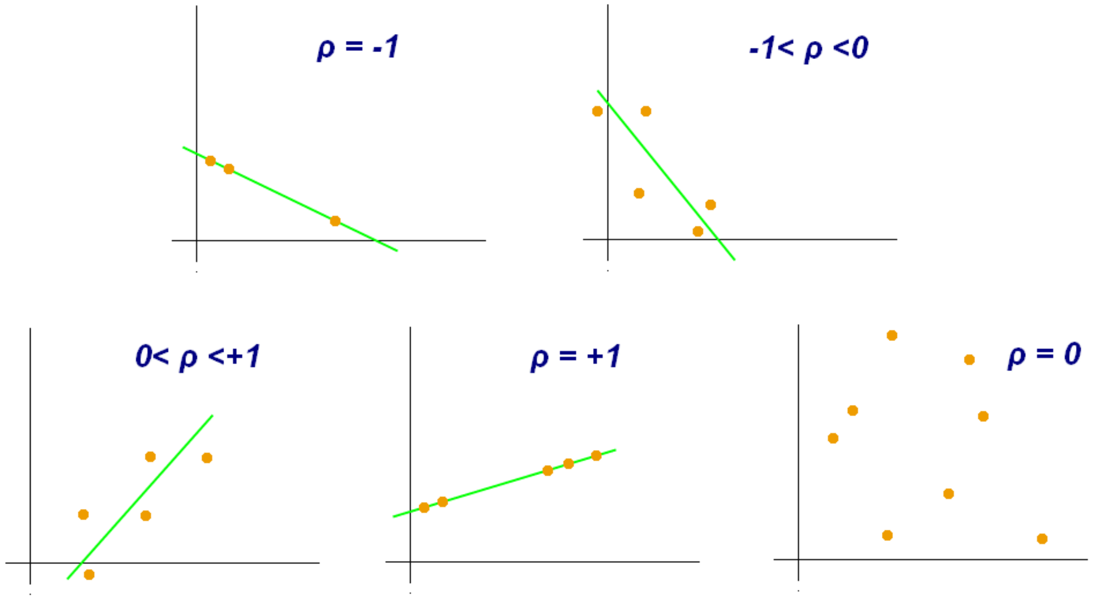
$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \frac{\mathbb{E}(X-\mu_X)(Y-\mu_Y)}{\sigma_X \sigma_Y}$$

$-1 \leq \rho \leq 1$

$\rho = 1$: perfect positive (linear) correlatio

$\rho =$-1 perfect negative correlation

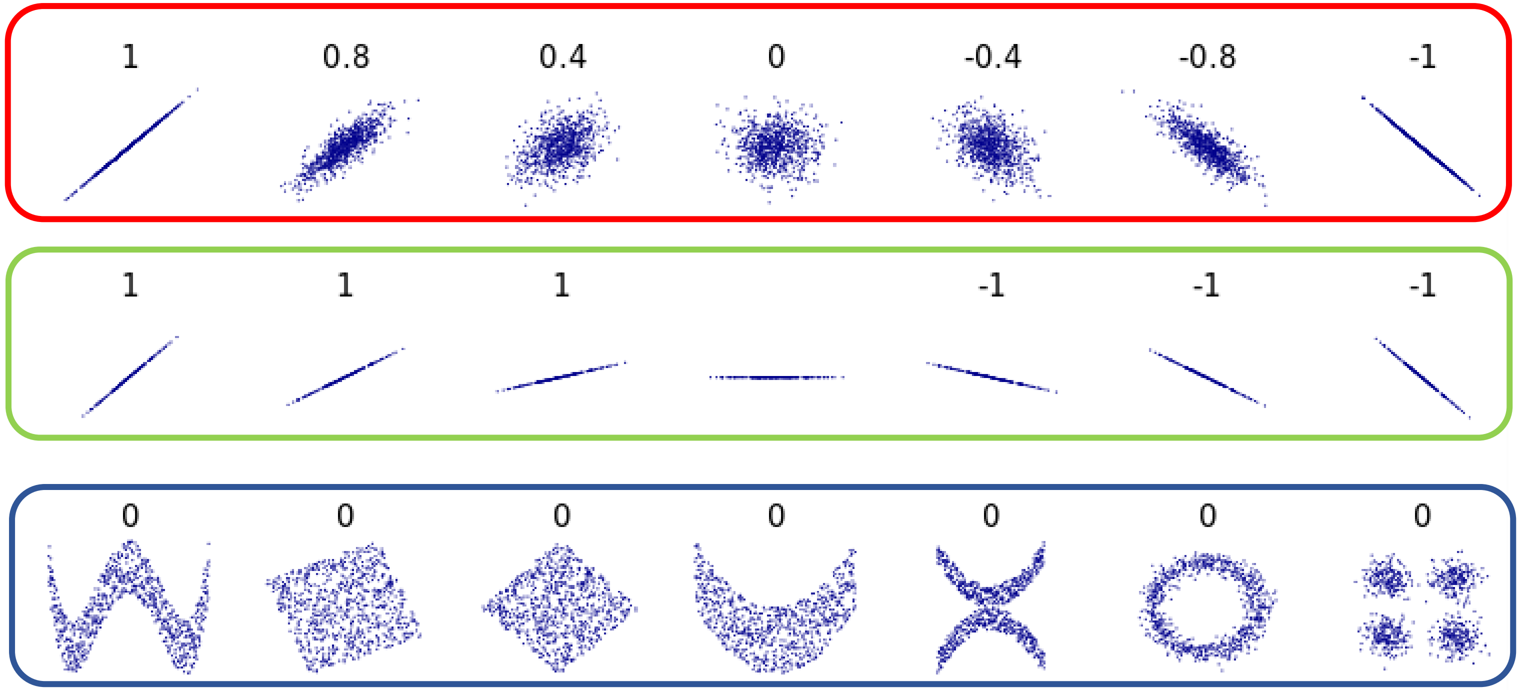
$\rho =$0 no correlation



https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Exploratory Data Analysis - **Correlation**

- Pearson correlation
  - Scatter plots for variable pairs of different Pearson correlations.
  - Correlation reflects the strength and direction of a linear relationship but not the slop of that relationship, nor many aspects of nonlinear relationships.

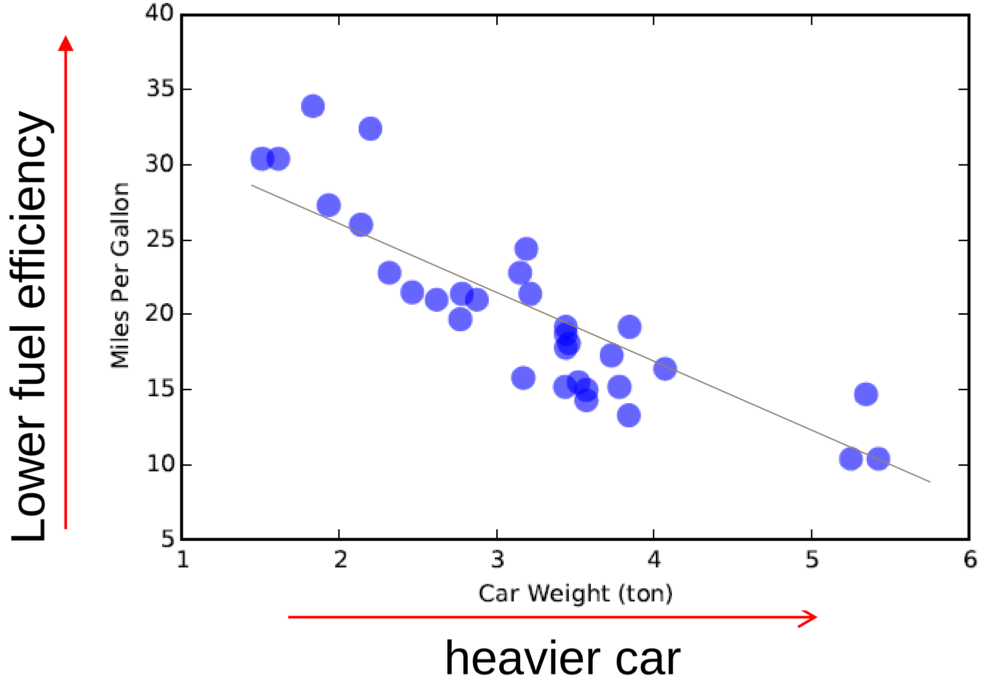


https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Exploratory Data Analysis - **Correlation**

- Pearson correlation

Negative Correlation



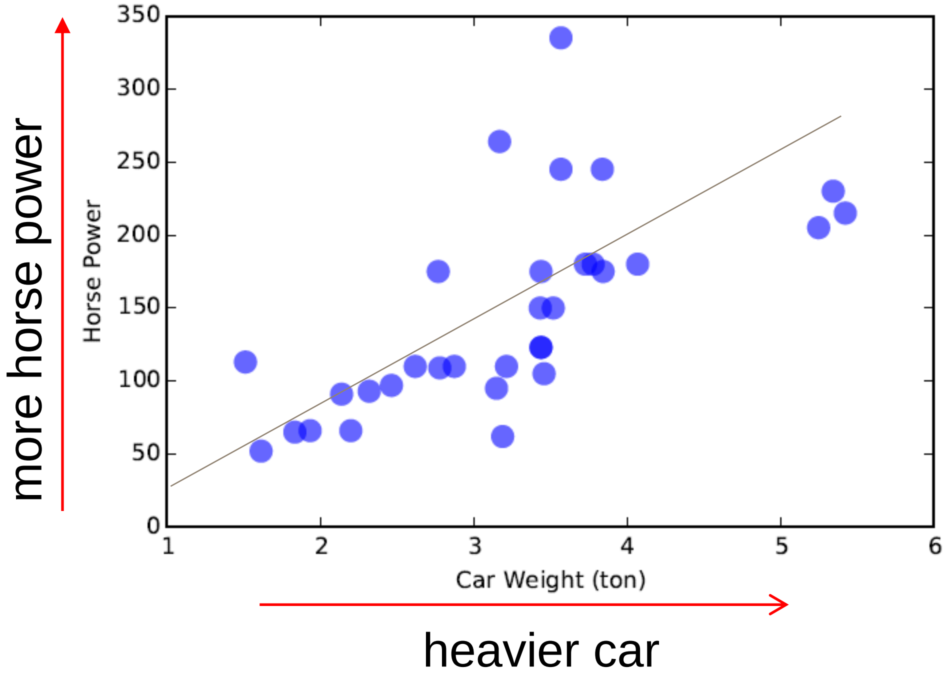
Lower fuel efficiency

heavier car

- The heavier the car is, the lower fuel efficiency.
- Negatively associated.
- How much they are associated?

Pearson correlation coefficient $\rho = -0.87$

Positive Correlation



more horse power

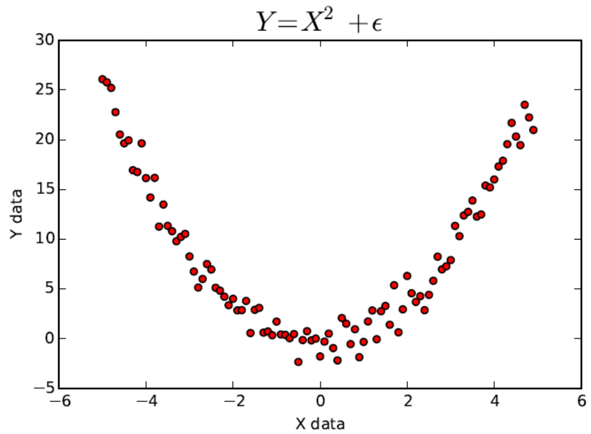
heavier car

- The heavier the car is, the more horsepower the car has.
- Positively associated.
- How much they are associated?

Pearson correlation coefficient $\rho = +0.66$

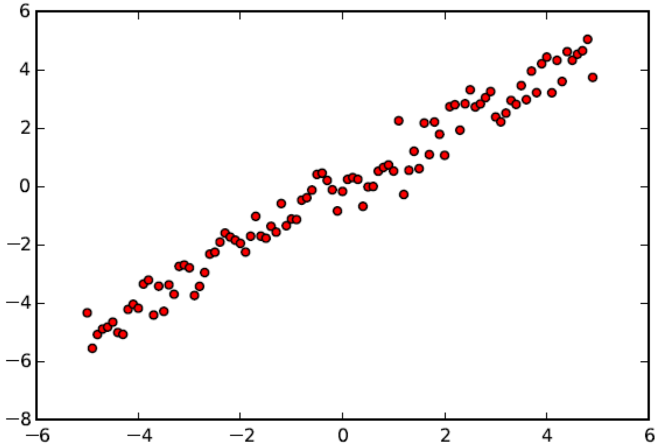
Exploratory Data Analysis - **Correlation**

- Pearson correlation
  - When does it fail?
    - Non-linear relationships
      - Y can perfectly explained by X.
      - But, $\rho = -0.09$

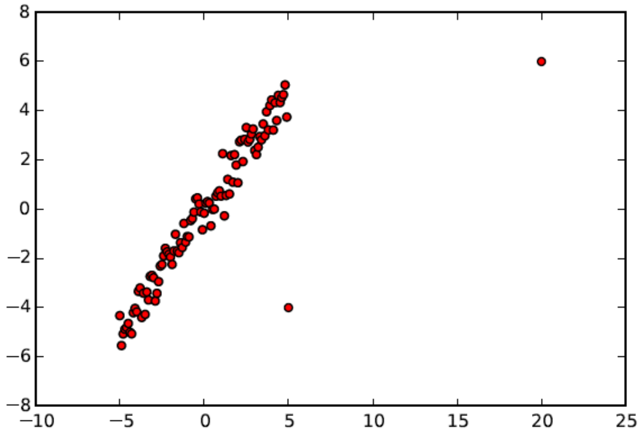


$\rho = -0.09$

- Presence of outliers
  - Y can almost perfectly explained by X.



$\rho = 0.98$



$\rho = 0.87$

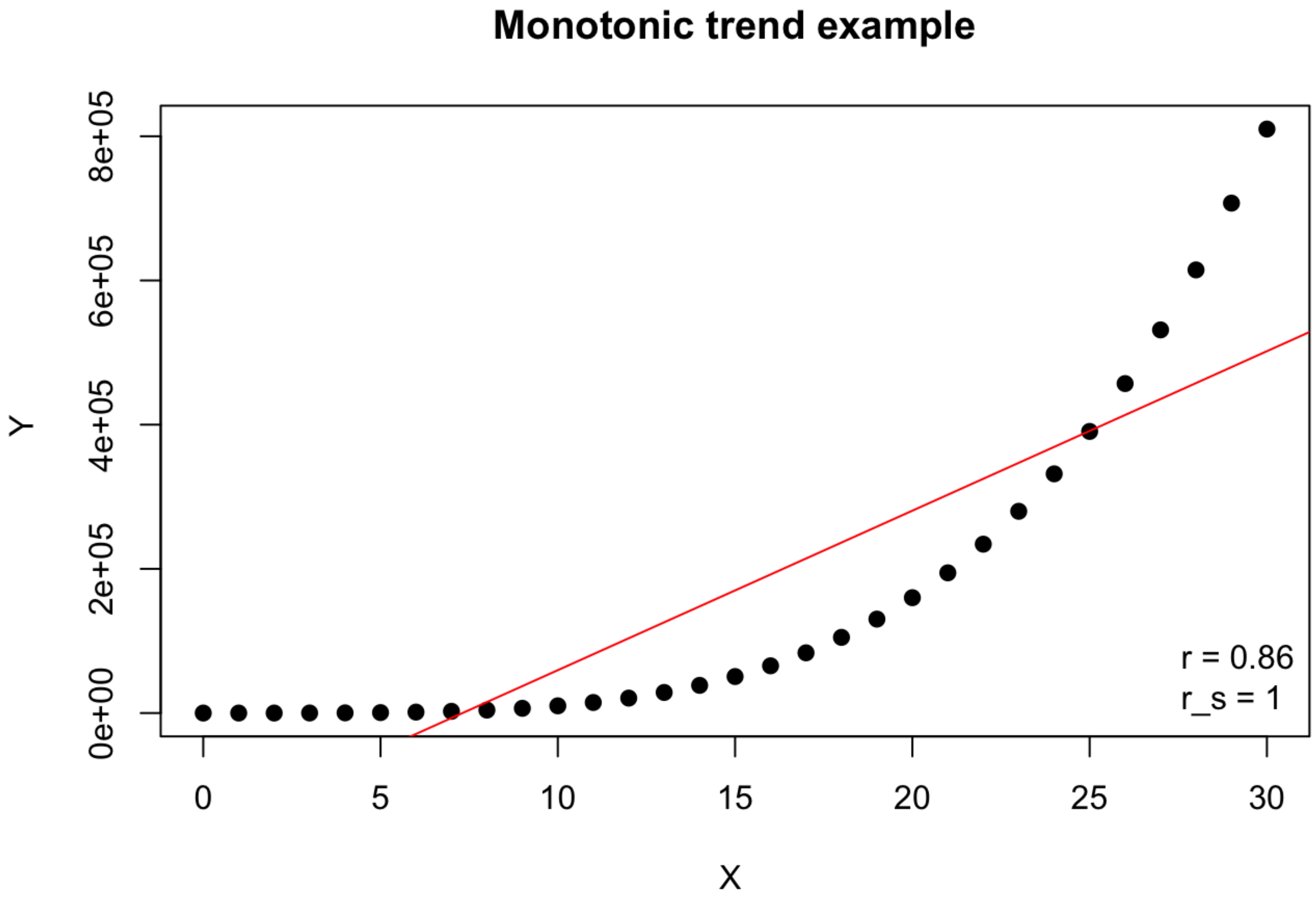
Exploratory Data Analysis - **Correlation**

- Pearson correlation
  - Limitations
    - Only capture linear correlation!
    - Sensitive to outliers.
    - Assume data is normally distributed

- What to use if we have a monotonic trend?

Exploratory Data Analysis - **Correlation**

- Monotonic trend



Y never decreases as X increases but the trend may or may not be linear.

Positively correlated

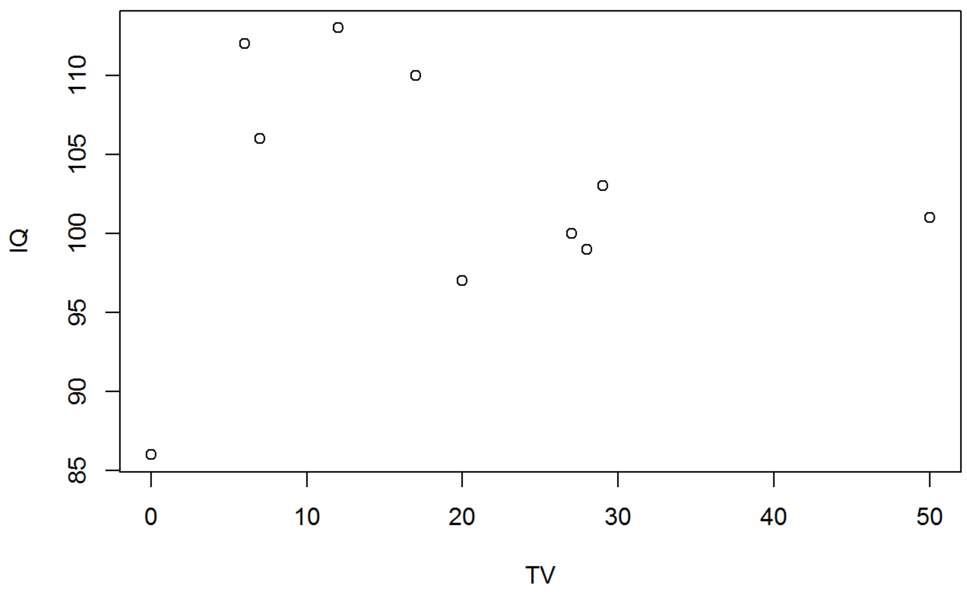
Or, Y never increases as X increases the trend may or may not be linear.

Negatively correlated

Exploratory Data Analysis - **Correlation**

- Spearman’s rank correlation - ρ_s
  - To analyse two ordinal or discrete variables.
  - Detect monotonic trends.
  - $-1 \leq \rho_s \leq 1$
  - How to calculate Spearman’s rank correlation?

National IQ estimate	**Average TV viewing (hrs/week)**
106	7
86	0
100	27
101	50
99	28
103	29
97	20
113	12
112	6
110	17



Pearson correlation $\rho(X,Y) = -0.038$

Exploratory Data Analysis - **Correlation**

- Spearman's rank correlation - ρ_s
  - How to calculate Spearman's rank correlation?
    - Step 1: Calculate the rank for IQ.
    - Step 2: Calculate the rank for TV.
    - Step 3: Calculate the rank differences.
    - Step 4: Calculate the square of the rank differences.

IQ194	**TV**	**Rank IQ**	**Rank TV**	d_i	d_i^2
86	0	1	1	0	0
97	20	2	6	−4	16
99	28	3	8	−5	25
100	27	4	7	−3	9
101	50	5	10	−5	25
103	29	6	9	−3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36
					194

$n = 10$

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

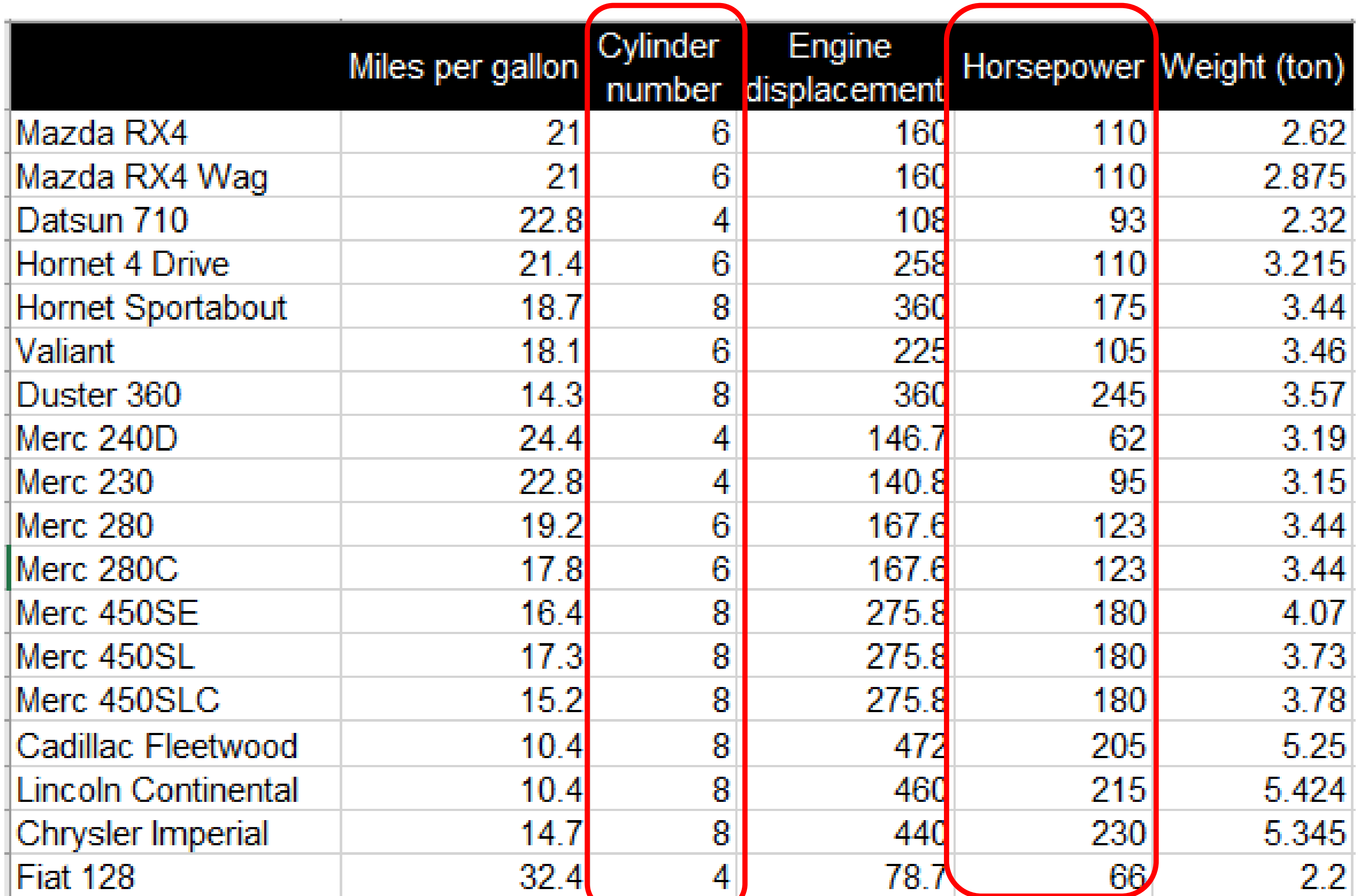
$$\rho_s = 1 - \frac{6 \times 194}{10(100-1)} = -0.176$$

$$\rho = -0.038$$

Exploratory Data Analysis - **Correlation**

- Spearman’s rank correlation - ρ_s
- *Higher number of cylinders seems to increase horsepower?*

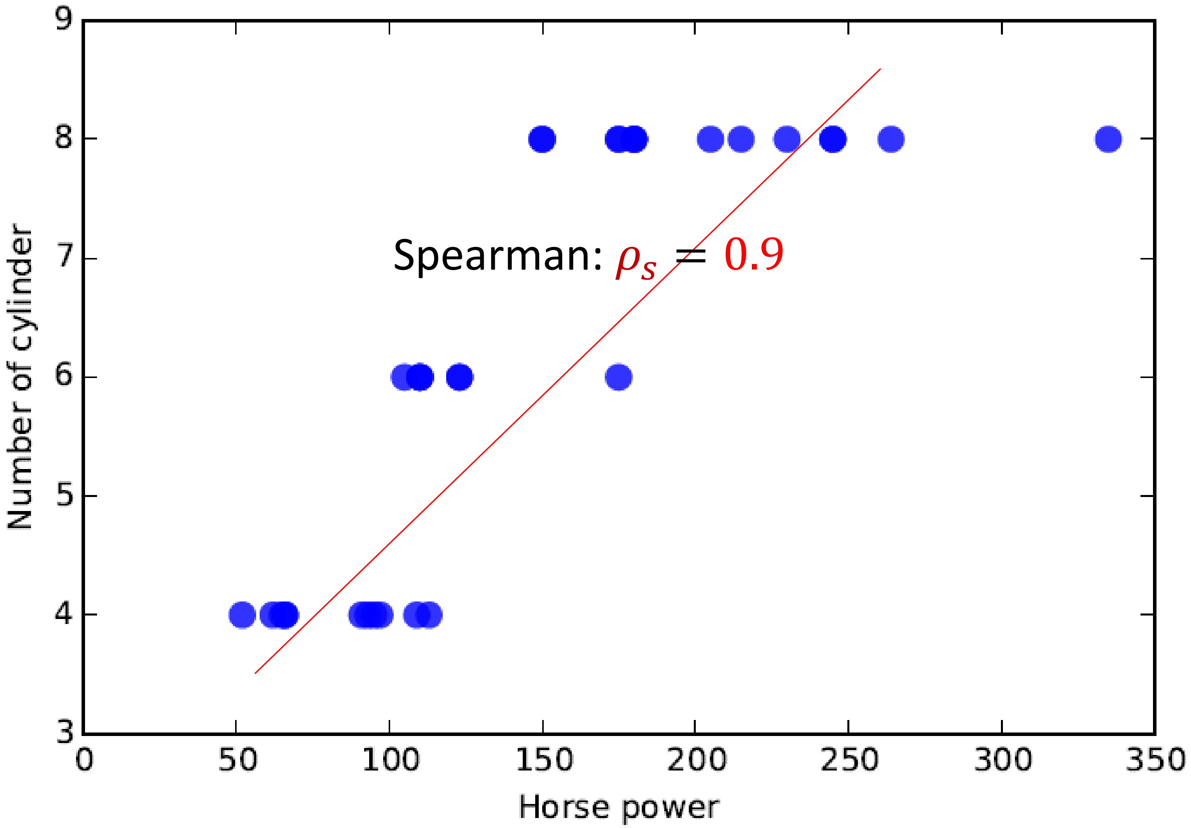
1974 Motor Trend Data



	Miles per gallon	Cylinder number	Engine displacement	Horsepower	Weight (ton)
Mazda RX4	21	6	160	110	2.62
Mazda RX4 Wag	21	6	160	110	2.875
Datsun 710	22.8	4	108	93	2.32
Hornet 4 Drive	21.4	6	258	110	3.215
Hornet Sportabout	18.7	8	360	175	3.44
Valiant	18.1	6	225	105	3.46
Duster 360	14.3	8	360	245	3.57
Merc 240D	24.4	4	146.7	62	3.19
Merc 230	22.8	4	140.8	95	3.15
Merc 280	19.2	6	167.6	123	3.44
Merc 280C	17.8	6	167.6	123	3.44
Merc 450SE	16.4	8	275.8	180	4.07
Merc 450SL	17.3	8	275.8	180	3.73
Merc 450SLC	15.2	8	275.8	180	3.78
Cadillac Fleetwood	10.4	8	472	205	5.25
Lincoln Continental	10.4	8	460	215	5.424
Chrysler Imperial	14.7	8	440	230	5.345
Fiat 128	32.4	4	78.7	66	2.2

Exploratory Data Analysis - **Correlation**

- Spearman's rank correlation - ρ_s
- *Higher number of cylinders seems to increase horsepower?*



Pearson: $\rho = 0.83$

Spearman: $\rho_s = 0.9$

Exploratory Data Analysis - **Correlation**

- Spearman’s rank correlation - ρ_s
- Summary
  - Pearson correlation is suitable for continuous data (with normality assumption on the distribution of the data)
  - Spearman correlation is suitable for ordinal/discrete data (but also continuous).
    - It is nonparametric and distribution-assumption free.
  - Pearson correlation detects linear trends.
  - Spearman correlation detects monotonic trends.

Exploratory Data Analysis

- **Association**
- **Correlation**
- **Agreement**

Exploratory Data Analysis - **Agreement**

- Measure if X and Y *agree* - Nominal data

The decision by a psychiatrist and a psychiatric social worker whether or not to section 10 individuals suffering mental ill-health. [source: MDSS, p. 182]

	Patient	1	2	3	4	5	6	7	8	9	10
X →	Psychiatrist	Y	Y	N	Y	N	N	N	Y	Y	Y
Y →	PSW	Y	N	N	Y	N	N	Y	Y	Y	N

- X and Y are *perfectly* agree if every pair of values are the same.
  - Rarely happens in real-world data.

Exploratory Data Analysis - Agreement

- Measure if X and Y *agree* - Nominal data

The decision by a psychiatrist and a psychiatric social worker whether or not to section 10 individuals suffering mental ill-health. [source: MDSS, p. 182]

	Patient	1	2	3	4	5	6	7	8	9	10
X	Psychiatrist	Y	Y	N	Y	N	N	N	Y	Y	Y
Y	PSW	Y	N	N	Y	N	N	Y	Y	Y	N

- $$\%\ Observed\ Agreement = \frac{\#\ agreement\ cases}{\#\ total\ cases} = \frac{7}{10} = 0.7$$

Exploratory Data Analysis - **Agreement**

- But, random chance alone gives an agreement of 50%.
  - To be precise, it is an expected agreement due to random chance.
- Cohen’s Kappa

$$\text{Cohen's Kappa }(\kappa) = \frac{\%(\text{observed agreement}) - \%(\text{expected agreement})}{1 - \%(\text{expected agreement})}$$

We have computed this before = 0.7

How do we compute this expected agreement?

Exploratory Data Analysis - Agreement

- Cohen's Kappa

The decision by a psychiatrist and a psychiatric social worker whether or not to section 10 individuals suffering mental ill-health. [source: MDSS, p. 182]

Patient	1	2	3	4	5	6	7	8	9	10
Psychiatrist	Y	Y	N	Y	N	N	N	Y	Y	Y
PSW	Y	N	N	Y	N	N	Y	Y	Y	N

		Psychiatric Social Worker (PSW)		
		Yes	No	Total
Psychiatrist	Yes	4	2	6
	No	1	3	4
	Total	5	5	10

Column total

Overall total

Row total

$$\mathit{expected\ value} = \frac{\mathit{row\ total} \times \mathit{column\ total}}{\mathit{overal\ total}}$$

We need to compute the expected value for this

Exploratory Data Analysis - **Agreement**

- Cohen’s Kappa

The decision by a psychiatrist and a psychiatric social worker whether or not to section 10 individuals suffering mental ill-health. [source: MDSS, p. 182]

Patient	1	2	3	4	5	6	7	8	9	10
Psychiatrist	Y	Y	N	Y	N	N	N	Y	Y	Y
PSW	Y	N	N	Y	N	N	Y	Y	Y	N

Psychiatric Social Worker (PSW)

		Yes	No	Total
Psychiatrist	Yes	4 (3)	2	6
	No	1	3	4
	Total	5	5	10

← Column total

← Overall total

Row total

$$expected\ value = \frac{row\ total \times column\ total}{overal\ total}$$

We need to compute the expected value for this

$$\frac{\mathbf{5 \times 6}}{\mathbf{10}} = \mathbf{3}$$

Exploratory Data Analysis - Agreement

- Cohen’s Kappa

The decision by a psychiatrist and a psychiatric social worker whether or not to section 10 individuals suffering mental ill-health. [source: MDSS, p. 182]

Patient	1	2	3	4	5	6	7	8	9	10
Psychiatrist	Y	Y	N	Y	N	N	N	Y	Y	Y
PSW	Y	N	N	Y	N	N	Y	Y	Y	N

		Psychiatric Social Worker (PSW)		
		Yes	No	Total
Psychiatrist	Yes	4 (3)	2 (3)	6
	No	1 (2)	3 (2)	4
	Total	5	5	10

$$expected\ value = \frac{row\ total \times column\ total}{overal\ total}$$

Exploratory Data Analysis - Agreement

- Cohen’s Kappa

The decision by a psychiatrist and a psychiatric social worker whether or not to section 10 individuals suffering mental ill-health. [source: MDSS, p. 182]

Patient	1	2	3	4	5	6	7	8	9	10
Psychiatrist	Y	Y	N	Y	N	N	N	Y	Y	Y
PSW	Y	N	N	Y	N	N	Y	Y	Y	N

		Psychiatric Social Worker (PSW)		
		Yes	No	Total
Psychiatrist	Yes	4 (3)	2 (3)	6
	No	1 (2)	3 (2)	4
	Total	5	5	10

Since the number of expected agreements (both Yes, or both No) = 3 + 2 = 5

Hence, $\%(\text{expected agreement}) = \frac{5}{10} = 0.5$

Exploratory Data Analysis - **Agreement**

- Cohen’s Kappa (Chance-corrected proportional agreement statistic)

$$\text{Cohen}'\text{s Kappa } (\kappa) = \frac{\%(\text{observed agreement}) - \%(\text{expected agreement})}{1 - \%(\text{expected agreement})}$$

We have computed this before = 0.7

How do we compute this expected agreement?

$$\text{Cohen}'\text{s Kappa } (\kappa) = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

Exploratory Data Analysis - Agreement

- Cohen’s Kappa
  - After adjusting random chance, the agreement reduces from 70% to 40%.
  - How good is the Cohen’s kappa agreement?

Kappa	Strength of agreement
<.20	Poor
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Good
0.81 – 1.00	Very good

[Source: MDSS, p.183]

Exploratory Data Analysis - Agreement

- Cohen’s Kappa
  - When the agreement is completely random, Cohen’s Kappa value is zero.

The decision by a psychiatrist and a psychiatric social worker whether or not to section 10 individuals suffering mental ill-health. [source: MDSS, p. 182]

Patient	1	2	3	4	5	6	7	8	9	10	11	12
Psychiatrist	Y	Y	N	N	N	Y	N	Y	N	Y	Y	N
PSW	N	N	N	Y	N	Y	Y	Y	N	N	Y	Y

		Psychiatric Social Worker (PSW)		
		Yes	No	Total
Psychiatrist	Yes	3 (3)	3 (3)	6
	No	3 (3)	3 (3)	6
	Total	6	6	12

$$\frac{\#(\text{agreement cases})}{\#(\text{total})} = \frac{6}{12} = 0.5$$

$$\%(\text{expected agreement}) = \frac{6}{12} = 0.5$$

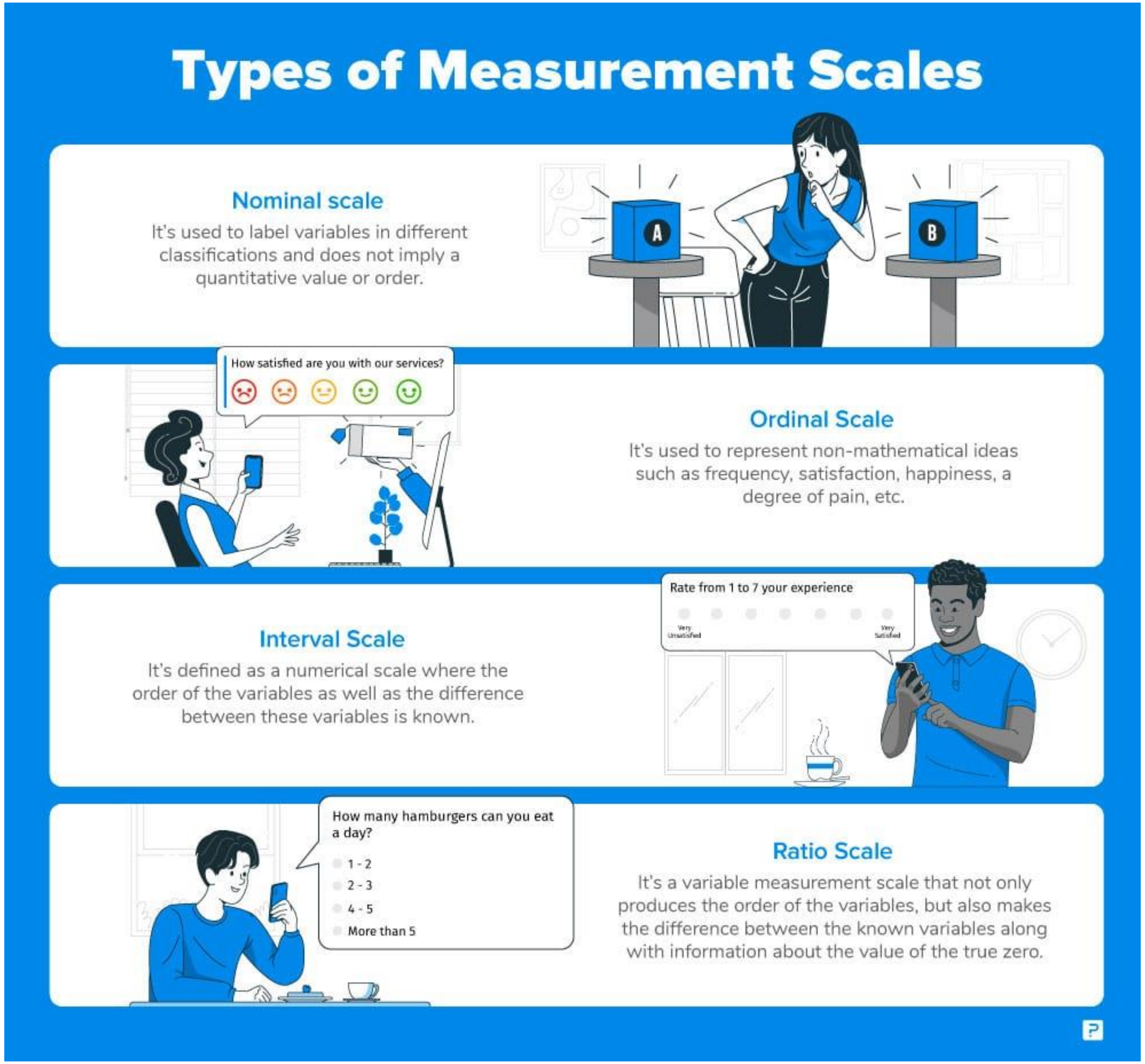
$$\kappa = \frac{0.5 - 0.5}{1 - 0.5} = 0.0$$

Exploratory Data Analysis - Agreement

	Level of Measurement		
	Nominal	**Ordinal**	**Interval and Ratio**
2 raters	Cohen’s Kappa	Cohen’s Weighted Kappa	Bland-Altman plots
	Inter Class Correlation (ICC)	ICC	ICC
> 2 raters	Fleiss’ Kappa	Kendall’s Coefficient of Concordance	
	ICC	ICC	ICC

Not an exhaustive list!

Exploratory Data Analysis - **Agreement**



Types of Measurement Scales

Nominal scale

It's used to label variables in different classifications and does not imply a quantitative value or order.

Ordinal Scale

It's used to represent non-mathematical ideas such as frequency, satisfaction, happiness, a degree of pain, etc.

Interval Scale

It's defined as a numerical scale where the order of the variables as well as the difference between these variables is known.

Ratio Scale

It's a variable measurement scale that not only produces the order of the variables, but also makes the difference between the known variables along with information about the value of the true zero.

Questions?