EDA

The average x value is 9 for each dataset I, II, III, IV.

The average y value is 7.50 for each dataset, I, II, III, IV.

The variance for x is 11 and the variance for y is 4.12.

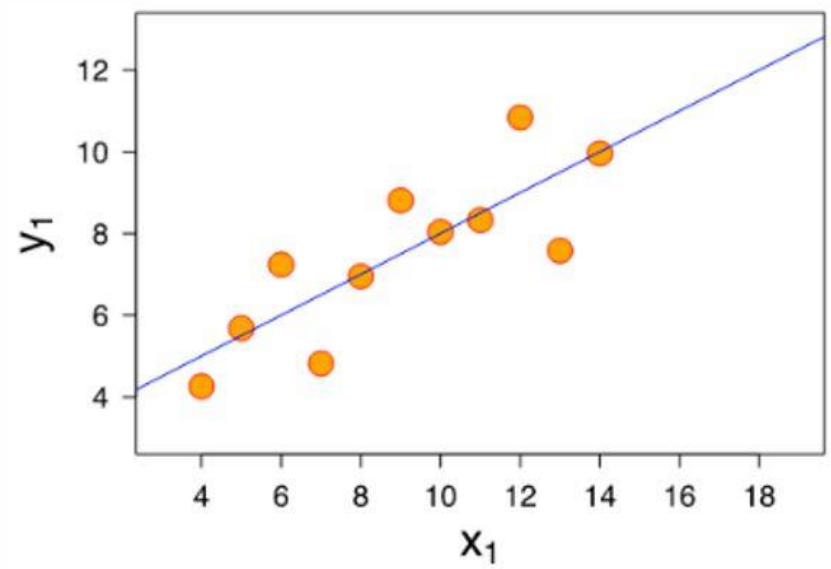
Are these 4 datasets similar?

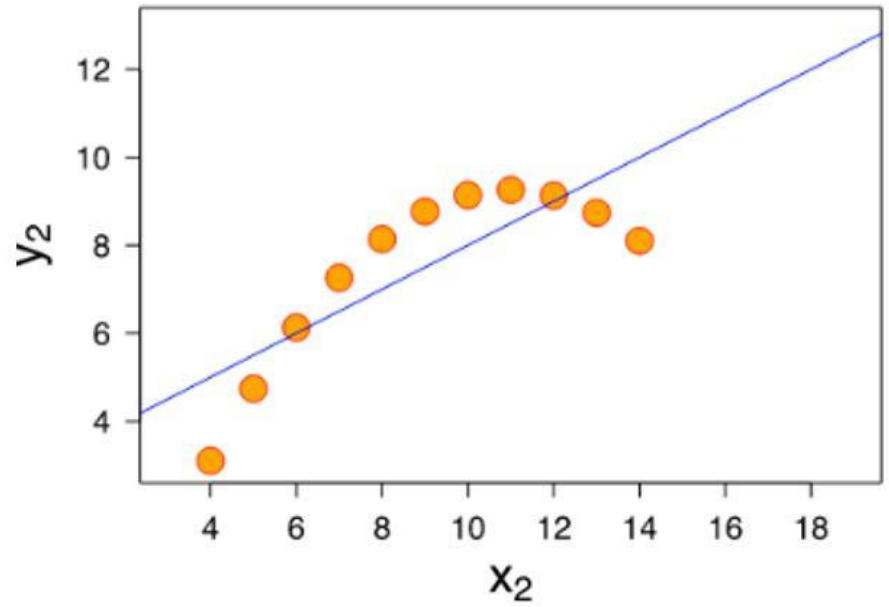
What does it look like if we plot this data?

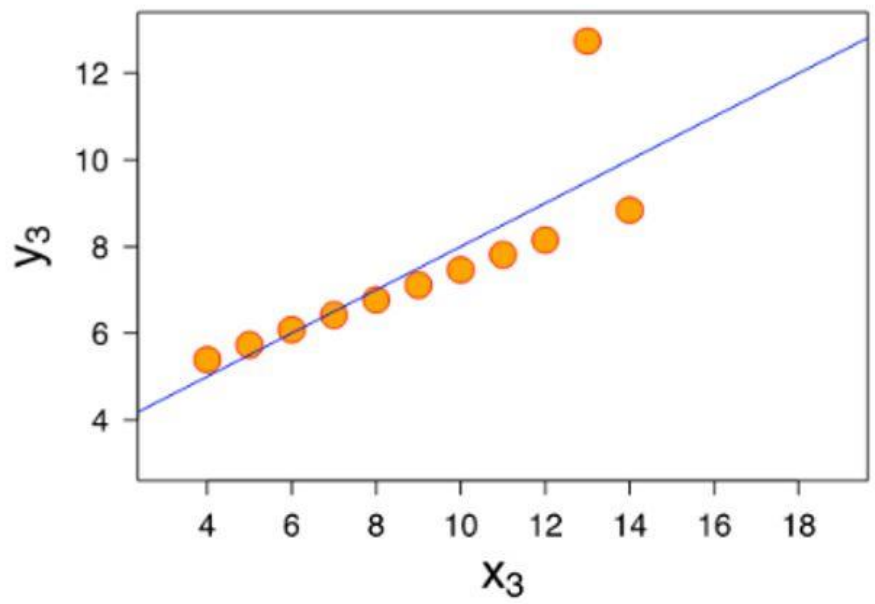
Anscombe's quartet

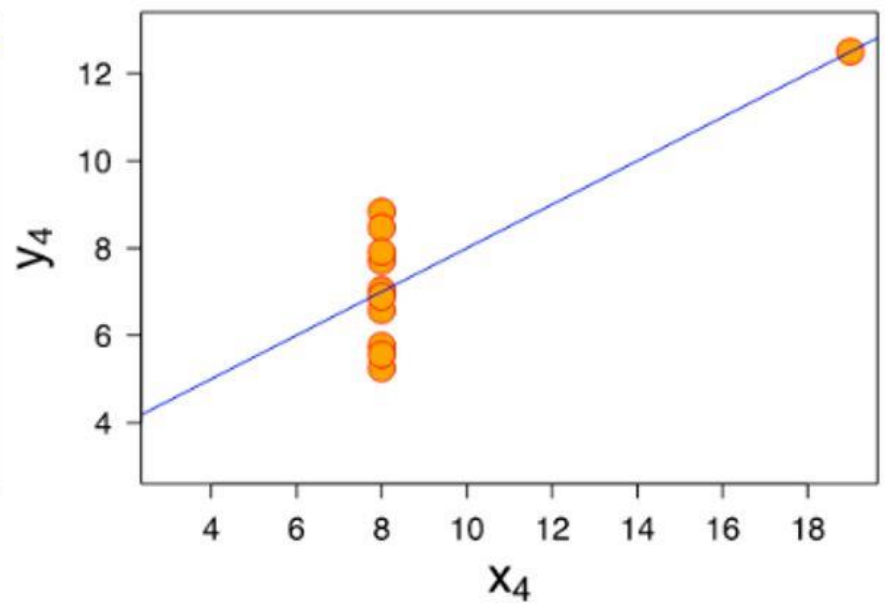
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

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


y1
x1
4
6
8
10
12
14
16
18


y2
x2
4
6
8
10
12
14
16
18


y3
x3
4
6
8
10
12
14
16
18


y4
x4
4
6
8
10
12
14
16
18

Common Terms in EDA

- numeric variables, categorical variables (including binary)
- nominal and ordinal variables
- mean, median, standard deviation, range, percentiles and quantiles
- univariate and bivariate summaries and visualizations
- histograms, boxplots, scatter plots

Short Quiz

- True or false: a feature is a column in the data, a variable is a row in the data.
- In Hong Kong, the average monthly salary is HK$30k, but the median monthly salary is $14k. Why is there around a double of difference?

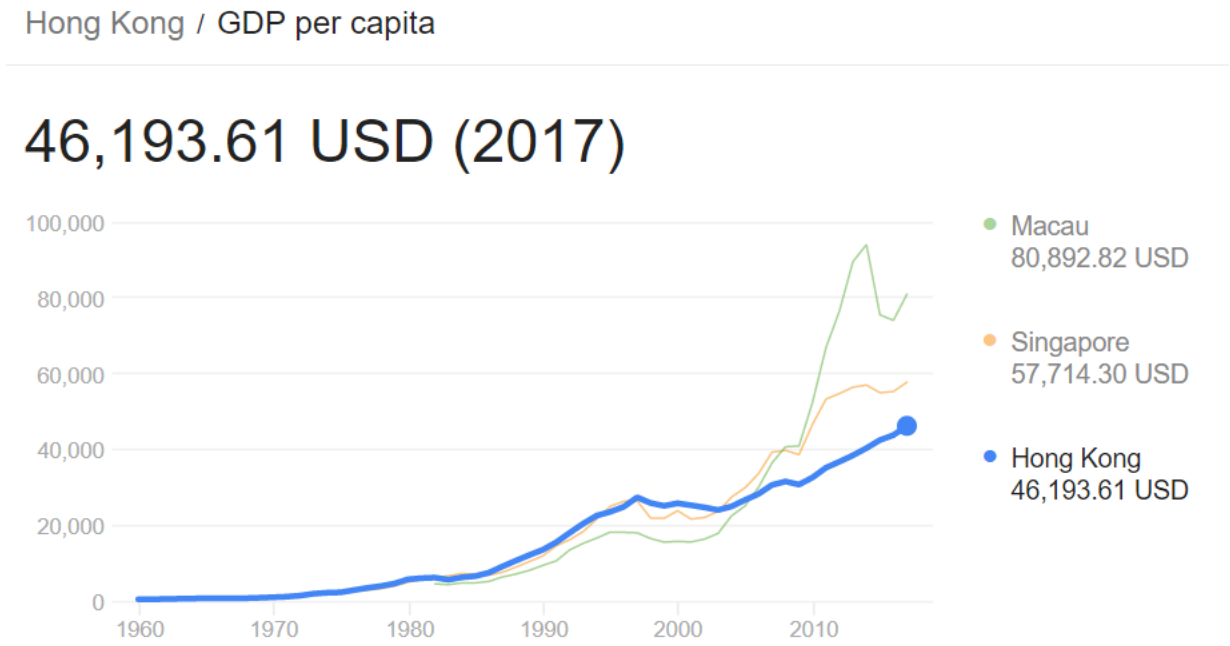


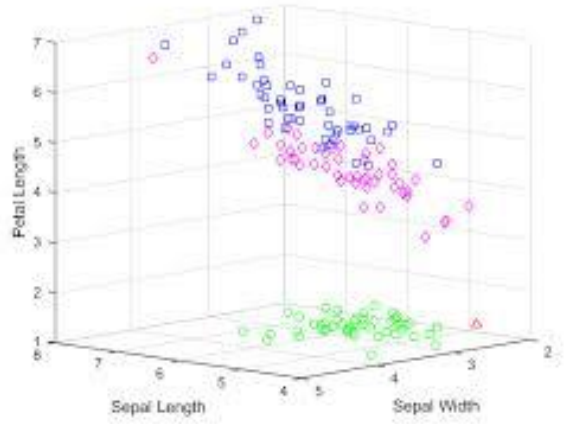
Table E012 Median Monthly Wage of Employees Analysed by Sex, Age Group, Educational Attainment, Occupational Group and Industry Section, May – June 2017			
		Median monthly wage (HK$)	
By sex			
	Male	19,100	(+3.9%)
	Female	14,700	(+4.4%)
By age group			
	15-24	12,400	(+4.6%)
	25-34	17,600	(+4.6%)
	35-44	19,700	(+3.7%)
	45-54	17,200	(+3.6%)
	≥55	14,000	(+4.7%)
By educational attainment			
	Primary and below	11,500	(+4.5%)
	Secondary 1 to 3	13,300	(+4.2%)
	Secondary 4 to 7	16,000	(+3.9%)
	Tertiary education	26,400	(+3.1%)

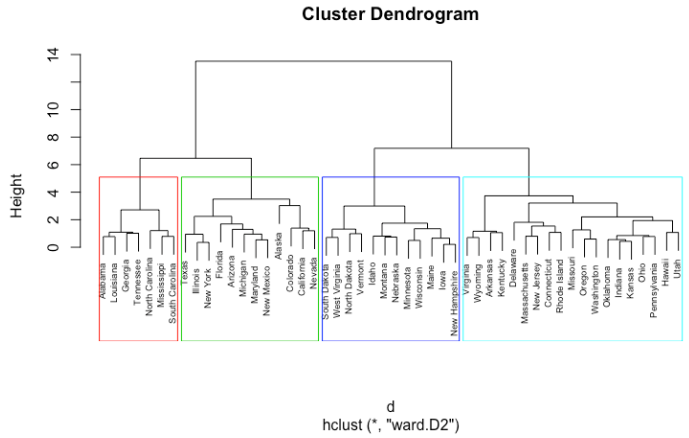
Unsupervised Machine Learning

- Clustering: k-means clustering vs hierarchical clustering

  The k-means clustering is an algorithm that attempts to find grouping in the **Records(rows)** of the data.

- It finds similar data points (observations) when we compare their features. k-means clustering finds redundancy in the data across rows.





Unsupervised Machine Learning

- Principle Component Analysis

  Principal component analysis attempts to find groupings of the **features(Columns).**

  It finds features that are similar (relay similar information because they are highly correlated) and combines them into one feature called a principal component.