GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 2: EXPLORATORY DATA ANALYSIS

NOVEMBER 24, 2014

DAT11-SF

AGENDA

I. EXPLORATORY ANALYSIS

II. VALUE OF VISUALIZATION

LAB:

III. DATA ANALYSIS WITH PANDAS

I. EXPLORATORY DATA ANALYSIS

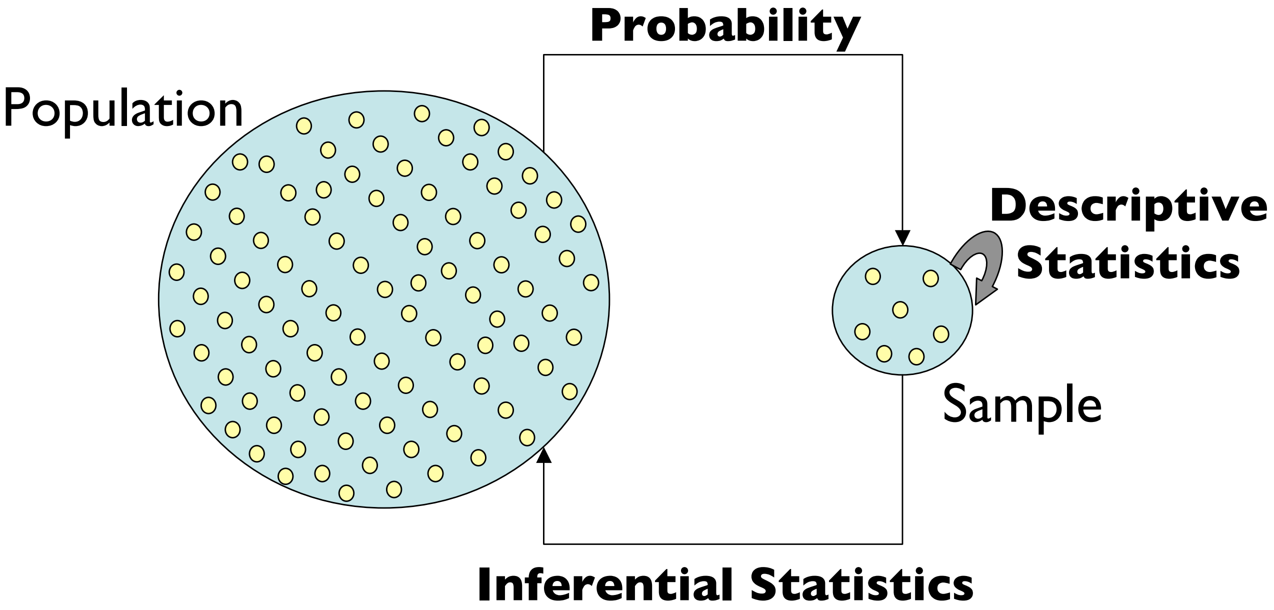
EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) is an approach for data analysis without statistical model or formulated prior hypothesis :

- Maximize insight into a data set
- Uncover underlying structure
- Detect outliers and anomalies
- Detect missing data
- Rank important factors
- Perform “sanity check”

John Tukey, “Exploratory Data Analysis”, 1977

"CENTRAL DOGMA" OF STATISTICS



Probability

Population

Descriptive Statistics

Sample

Inferential Statistics

From http://cs109.github.io/2014/

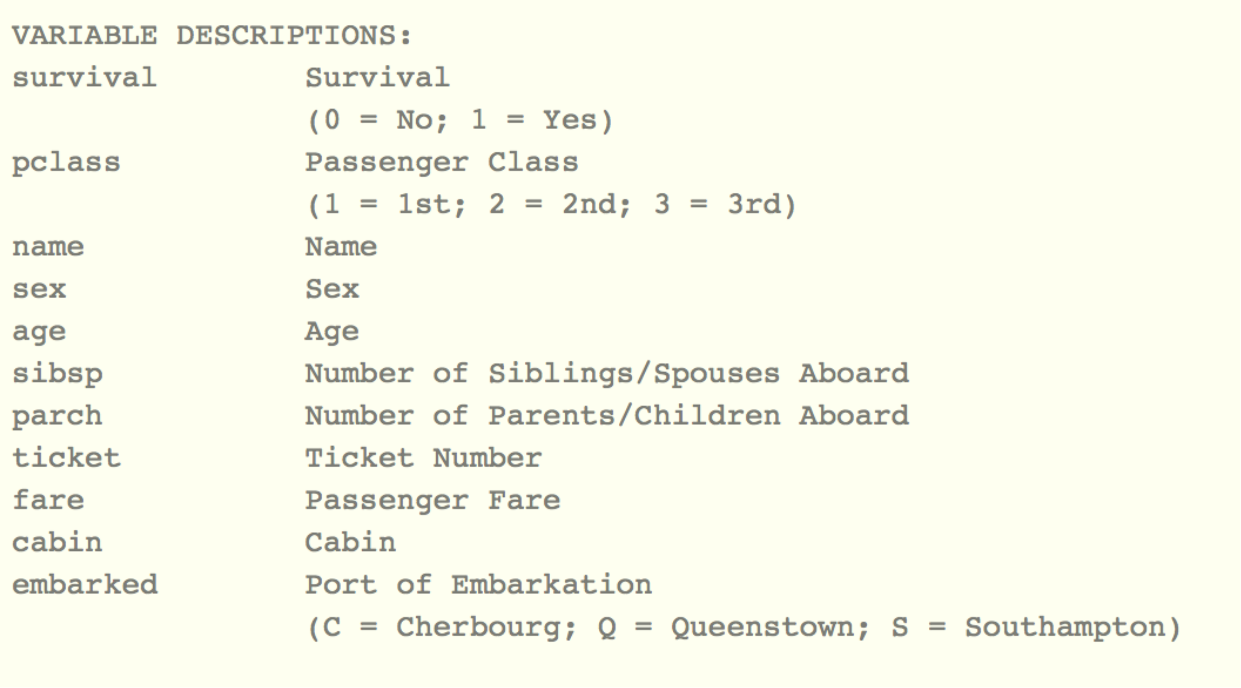
DATA TYPES

- Categorical data (nominal)
- Quantitative data (numerical, real values)
- Ordinal (ordered)

TABLE ROWS = instances, examples, data points, observations

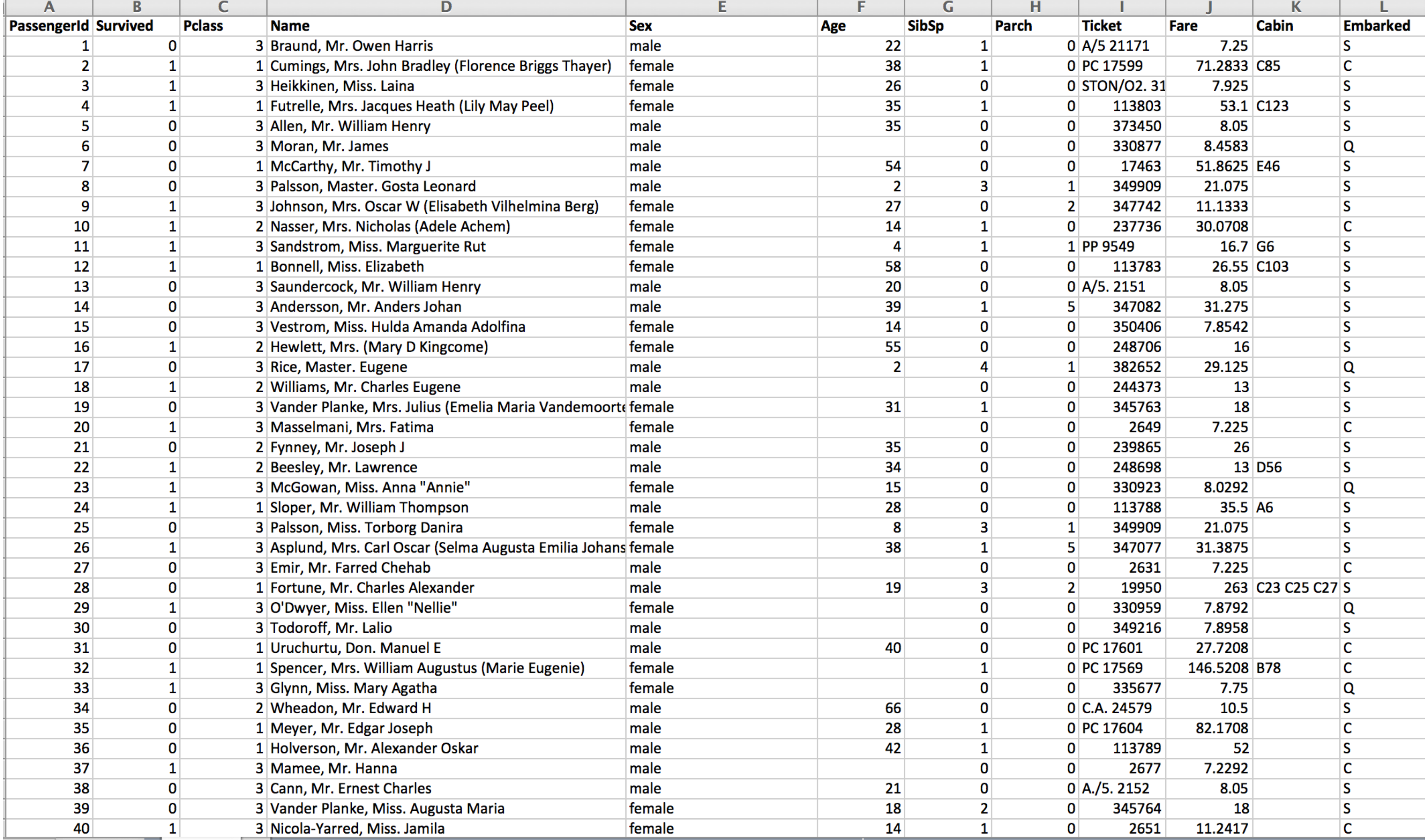
TABLE COLUMNS = attributes, features, variables

TITANIC DATASET



```
VARIABLE DESCRIPTIONS:
survival        Survival
                (0 = No; 1 = Yes)
pclass          Passenger Class
                (1 = 1st; 2 = 2nd; 3 = 3rd)
name            Name
sex             Sex
age             Age
sibsp           Number of Siblings/Spouses Aboard
parch           Number of Parents/Children Aboard
ticket          Ticket Number
fare            Passenger Fare
cabin           Cabin
embarked        Port of Embarkation
                (C = Cherbourg; Q = Queenstown; S = Southampton)
```

British Board of Trade (1990), *Report on the Loss of the 'Titanic' (S.S.)*

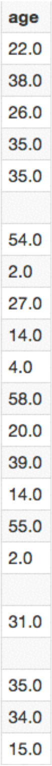


A	B	C	D	E	F	G	H	I	J	K	L
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 31	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16		S
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
18	1	2	Williams, Mr. Charles Eugene	male		0	0	244373	13		S
19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoorte	female	31	1	0	345763	18		S
20	1	3	Masselmani, Mrs. Fatima	female		0	0	2649	7.225		C
21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26		S
22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56	S
23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0292		Q
24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5	A6	S
25	0	3	Palsson, Miss. Torborg Danira	female	8	3	1	349909	21.075		S
26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johans	female	38	1	5	347077	31.3875		S
27	0	3	Emir, Mr. Farred Chehab	male		0	0	2631	7.225		C
28	0	1	Fortune, Mr. Charles Alexander	male	19	3	2	19950	263	C23 C25 C27	S
29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female		0	0	330959	7.8792		Q
30	0	3	Todoroff, Mr. Lalio	male		0	0	349216	7.8958		S
31	0	1	Uruchurtu, Don. Manuel E	male	40	0	0	PC 17601	27.7208		C
32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female		1	0	PC 17569	146.5208	B78	C
33	1	3	Glynn, Miss. Mary Agatha	female		0	0	335677	7.75		Q
34	0	2	Wheadon, Mr. Edward H	male	66	0	0	C.A. 24579	10.5		S
35	0	1	Meyer, Mr. Edgar Joseph	male	28	1	0	PC 17604	82.1708		C
36	0	1	Holverson, Mr. Alexander Oskar	male	42	1	0	113789	52		S
37	1	3	Mamee, Mr. Hanna	male		0	0	2677	7.2292		C
38	0	3	Cann, Mr. Ernest Charles	male	21	0	0	A./5. 2152	8.05		S
39	0	3	Vander Planke, Miss. Augusta Maria	female	18	2	0	345764	18		S
40	1	3	Nicola-Yarred, Miss. Jamila	female	14	1	0	2651	11.2417		C

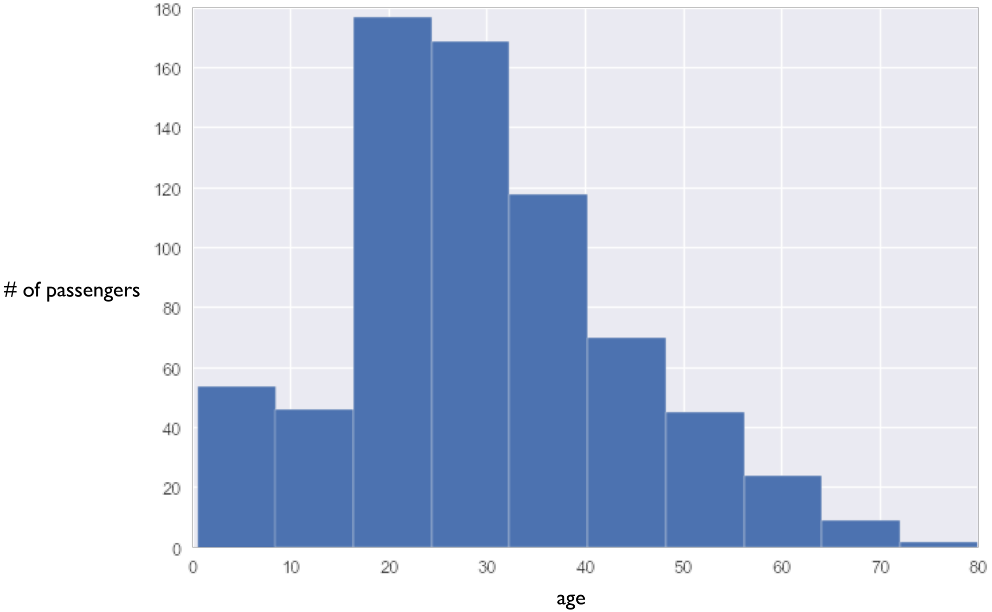
Some techniques for EDA:

- Summary statistics:
    - min, max, mean, median, standard deviation, quartiles
- Histograms
- Scatter plots
- Simple pairwise relationships between variables, correlation analysis

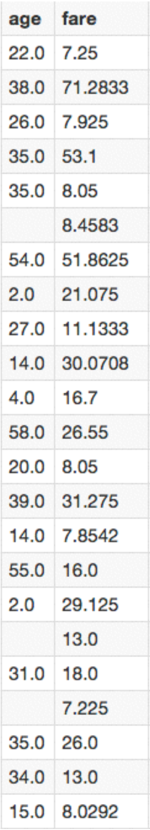
UNIVARIATE DATA



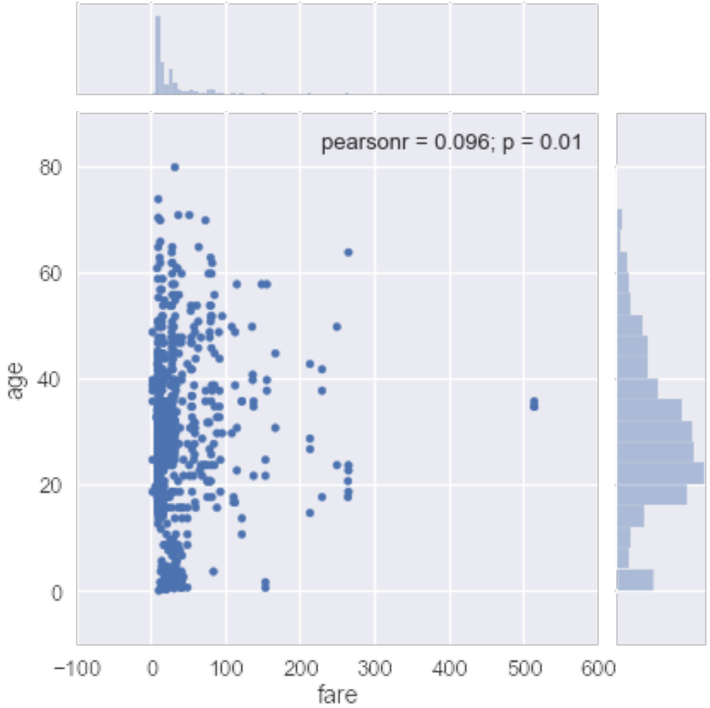
age
22.0
38.0
26.0
35.0
35.0
54.0
2.0
27.0
14.0
4.0
58.0
20.0
39.0
14.0
55.0
2.0
31.0
35.0
34.0
15.0



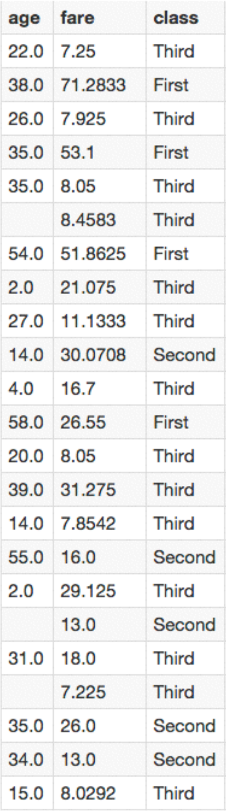
BIVARIATE DATA



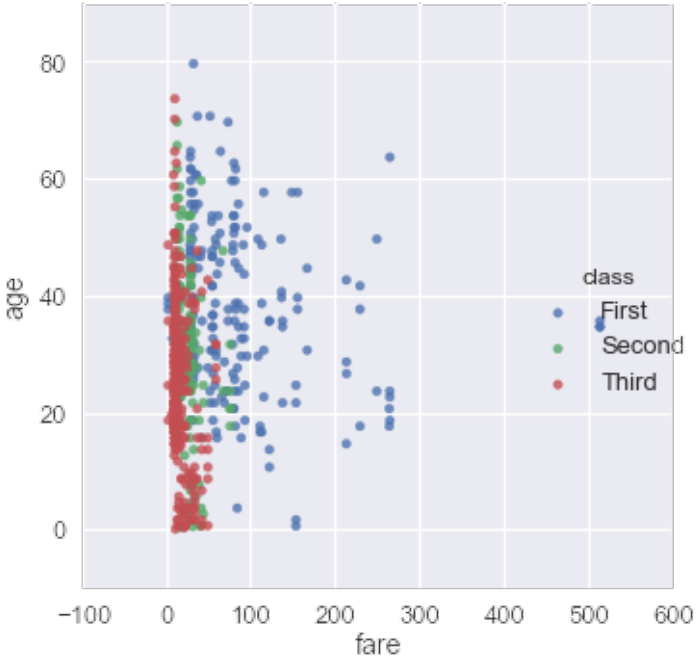
age	fare
22.0	7.25
38.0	71.2833
26.0	7.925
35.0	53.1
35.0	8.05
	8.4583
54.0	51.8625
2.0	21.075
27.0	11.1333
14.0	30.0708
4.0	16.7
58.0	26.55
20.0	8.05
39.0	31.275
14.0	7.8542
55.0	16.0
2.0	29.125
	13.0
31.0	18.0
	7.225
35.0	26.0
34.0	13.0
15.0	8.0292



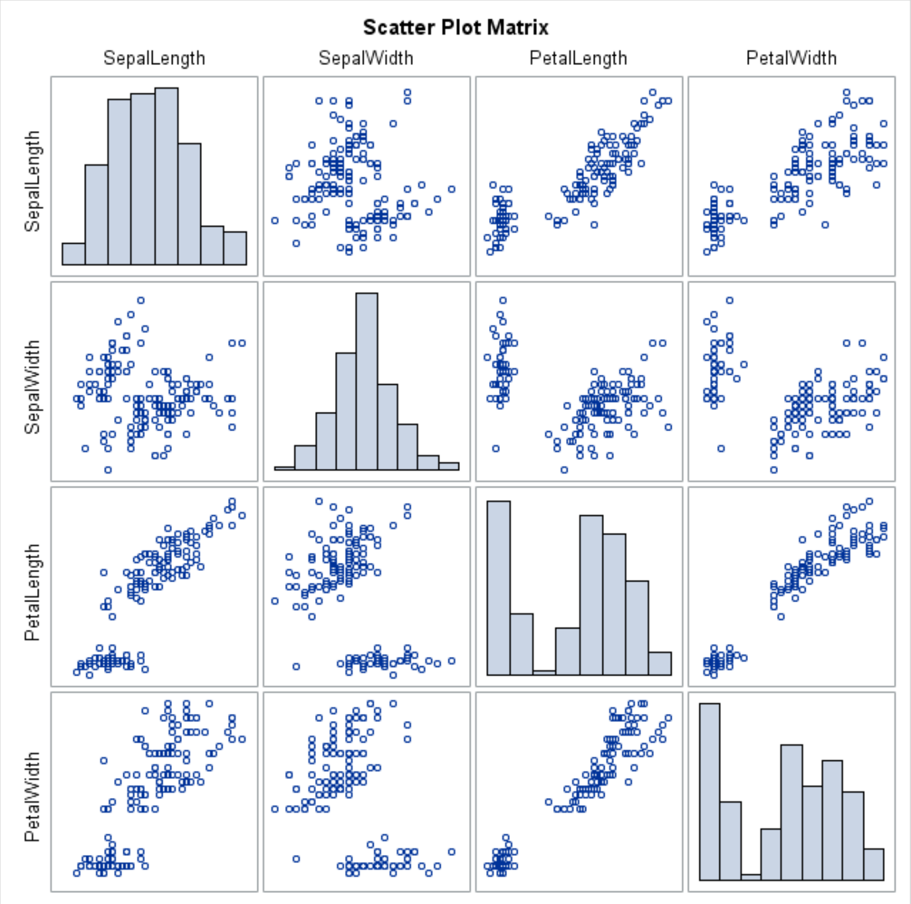
TRIVARIATE DATA (2 REAL + 1 CATEGORICAL)



age	fare	class
22.0	7.25	Third
38.0	71.2833	First
26.0	7.925	Third
35.0	53.1	First
35.0	8.05	Third
	8.4583	Third
54.0	51.8625	First
2.0	21.075	Third
27.0	11.1333	Third
14.0	30.0708	Second
4.0	16.7	Third
58.0	26.55	First
20.0	8.05	Third
39.0	31.275	Third
14.0	7.8542	Third
55.0	16.0	Second
2.0	29.125	Third
	13.0	Second
31.0	18.0	Third
	7.225	Third
35.0	26.0	Second
34.0	13.0	Second
15.0	8.0292	Third



SCATTERPLOT MATRIX



SUMMARY STATISTICS

Sample mean, average:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad \bar{x} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} \qquad \bar{x} = n \cdot \left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1}$$

Sample standard deviation – amount of variation from the average

$$s_N = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2}, \qquad s_N^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2,$$

std

variance

Quartile of a ranked data set are the three points that divide the data set into four equal groups, each group comprising a quarter of the data.

Q1 = 25% (splits off lowest 25% of the data)

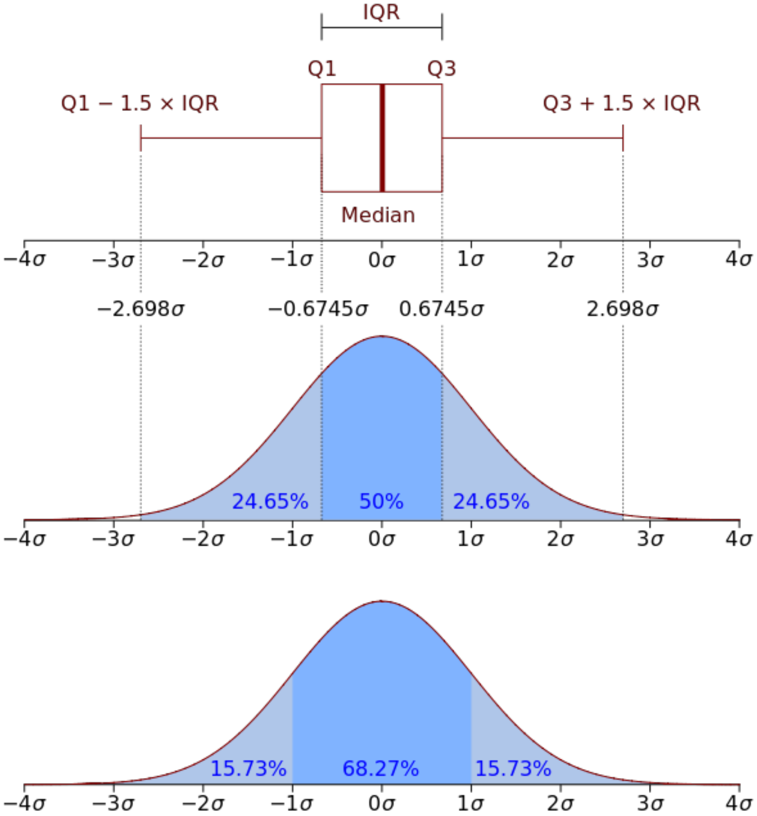
Q2 = median (half of the data)

Q3 = 75% (splits off highest 25% of the data)

Lower fence =Q1 – 1.5 IQR

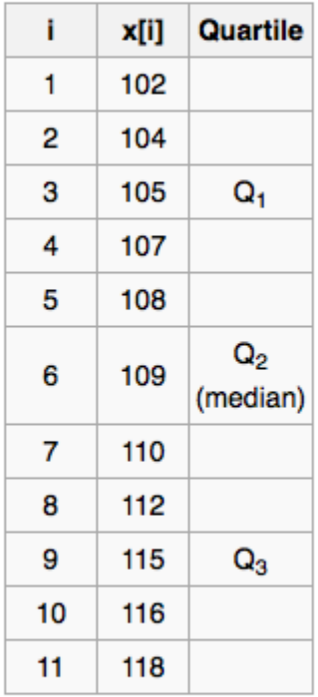
Upper fence = Q3 + 1.5 IQR

IQR = interquartile range



From http://en.wikipedia.org/wiki/Quartile

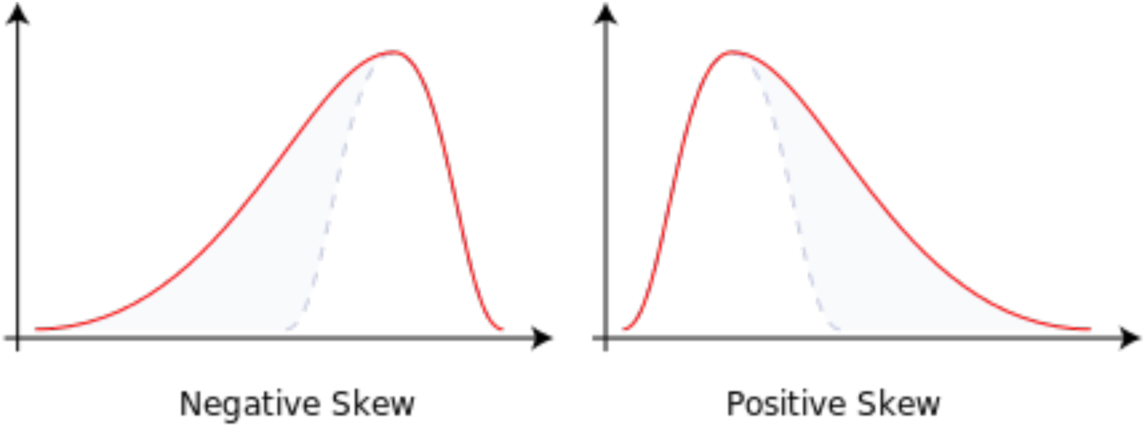
SUMMARY STATISTICS



i	x[i]	Quartile
1	102	
2	104	
3	105	Q_1
4	107	
5	108	
6	109	Q_2 (median)
7	110	
8	112	
9	115	Q_3
10	116	
11	118	

IQR = 115-105 = 10

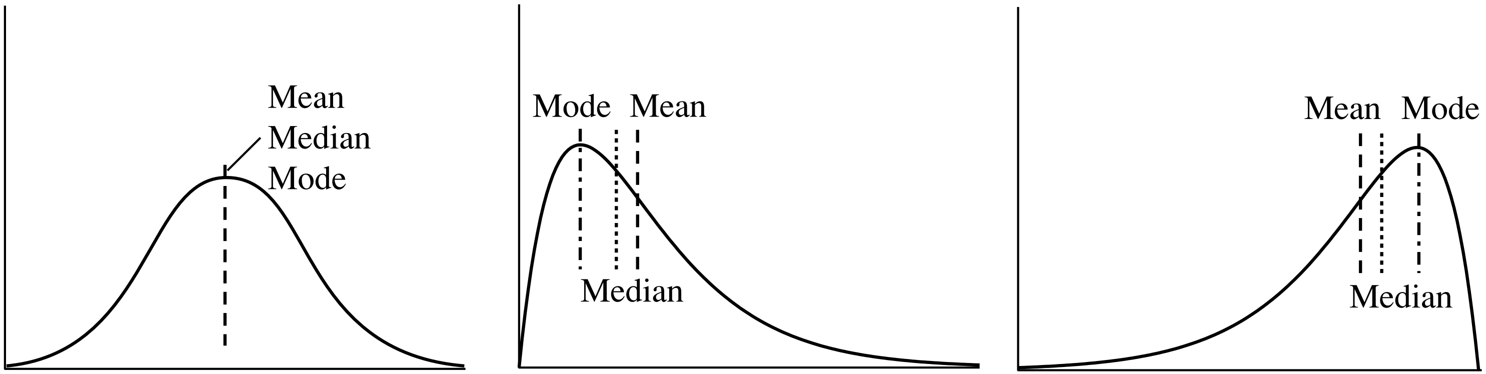
Sample **skewness** – measure of the asymmetry of distribution



Moment coefficient

$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{3/2}} ,$$

SUMMARY SATISTICS



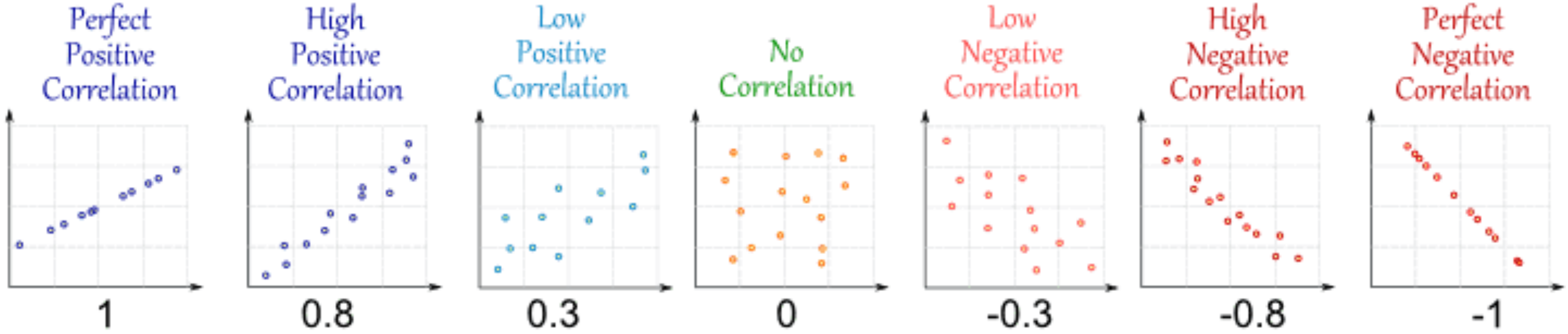
(a) Symmetric data

(b) Positively skewed data

(c) Negatively skewed data

Correlation coefficient (Pearson's correlation coefficient) – measure of dependency between two variables (how much they change together)

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$



From http://www.mathsisfun.com/data/correlation.html

MISSING DATA

Types of missing data points:

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

Treatment of missing data points:

- Deletion:
  - listwise – delete data point (table row)
  - pairwise – only for analysis when required
- Single imputation - mean, regression, random, last values
- Multiple imputation – average over multiple randomly imputed datasets
- Extra indictor variable

EDA COMMON QUESTIONS

1. What is a typical value?
2. What is the uncertainty for a typical value?
3. What is a good distributional fit for a set of numbers?
4. What is a percentile?
5. Does a factor have an effect?
6. What are the most important factors?
7. What is the best function for relating a response variable to a set of factor variables?
8. Can we separate signal from noise in time dependent data?
9. Can we extract any structure from multivariate data?
10. Does the data have outliers?

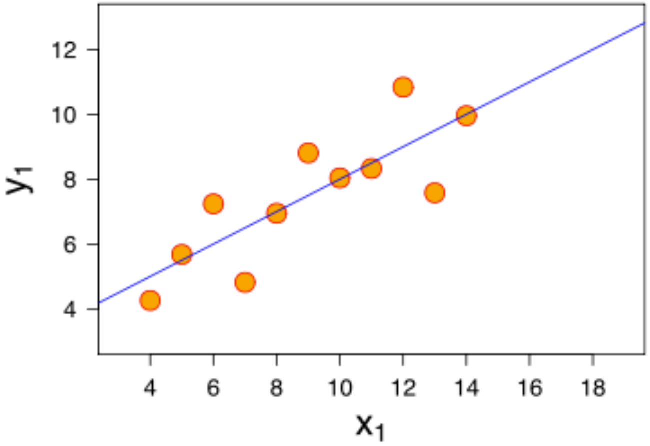
II. VALUE OF VISULAIZATION

“The greatest value of a picture is when it forces us to notice what we never expected to see.”

-John Tukey (1915 - 2000)

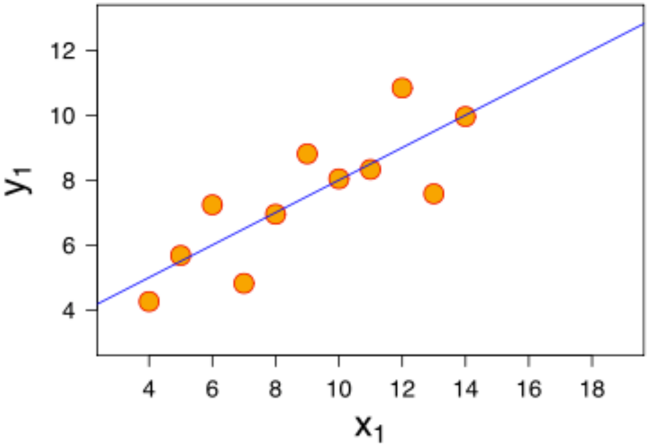
Consider the following dataset:

- eleven (x, y) points



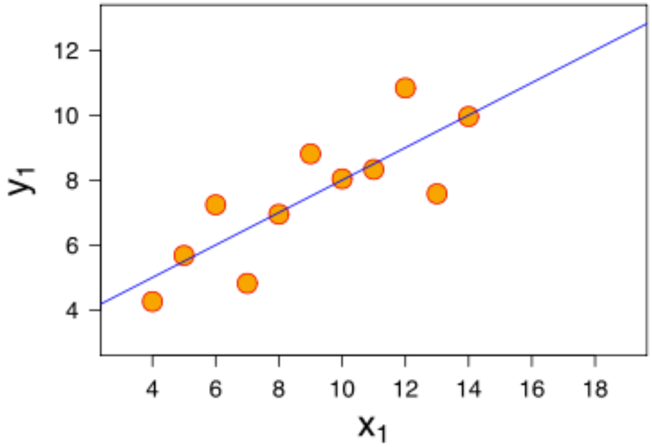
Consider the following dataset:

- eleven (x, y) points
- mean of x = 9, mean of y = 7.5



Consider the following dataset:

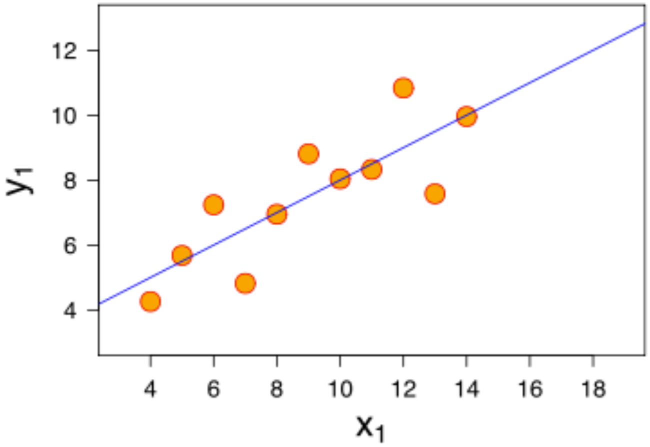
- eleven (x, y) points
- mean of x = 9, mean of y = 7.5
- variance of x = 11, variance of y = 4.1



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

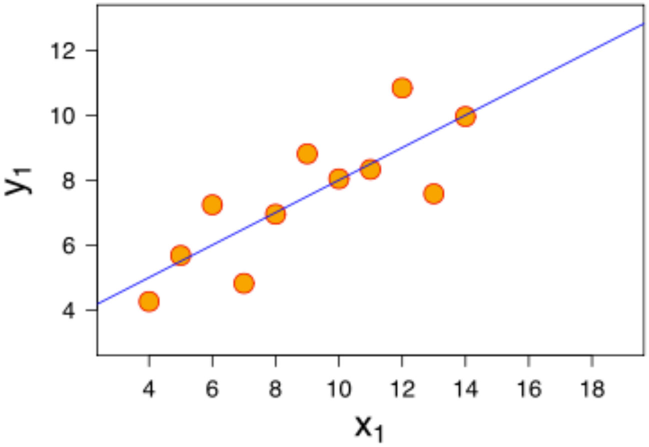
- eleven (x, y) points
- mean of x = 9, mean of y = 7.5
- variance of x = 11, variance of y = 4.1
- correlation of x and y = 0.8



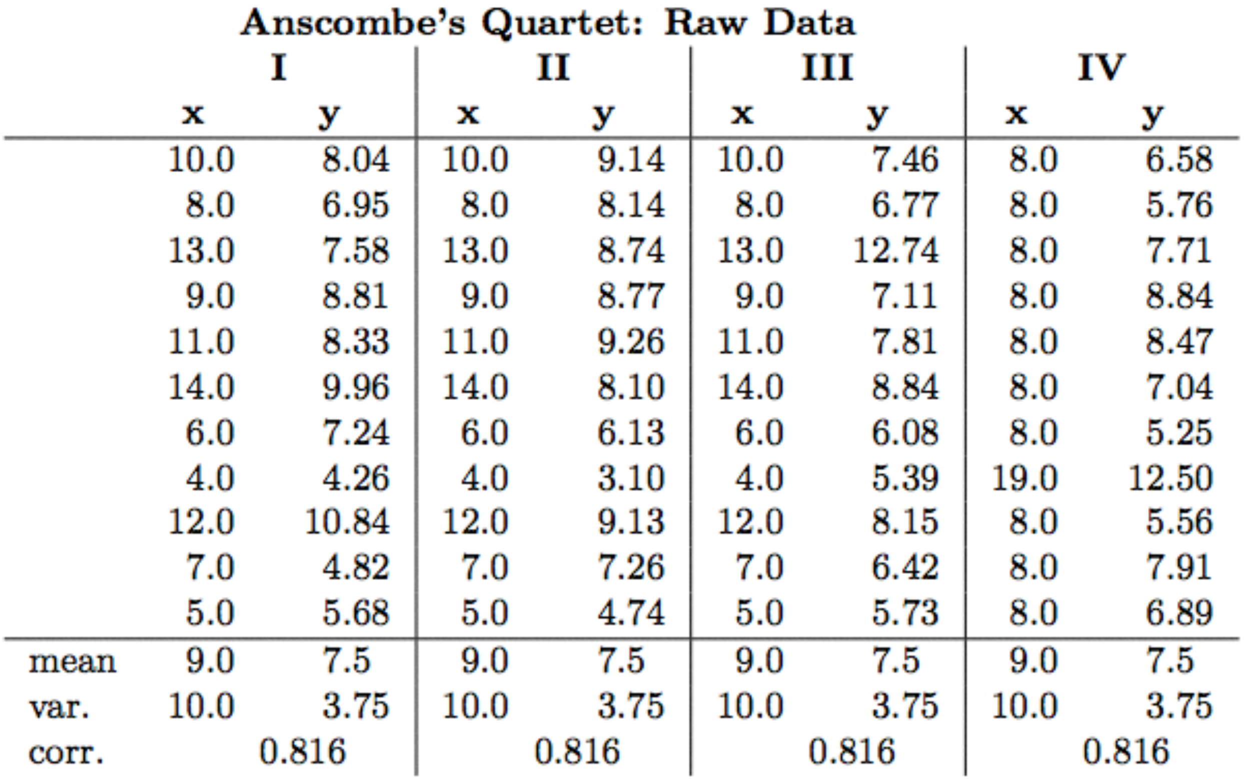
EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

- eleven (x, y) points
- mean of x = 9, mean of y = 7.5
- variance of x = 11, variance of y = 4.1
- correlation of x, y = 0.8
- line of best fit: y = 3.00 + 0.500x



EXERCISE – WHY VISUALIZE DATA?



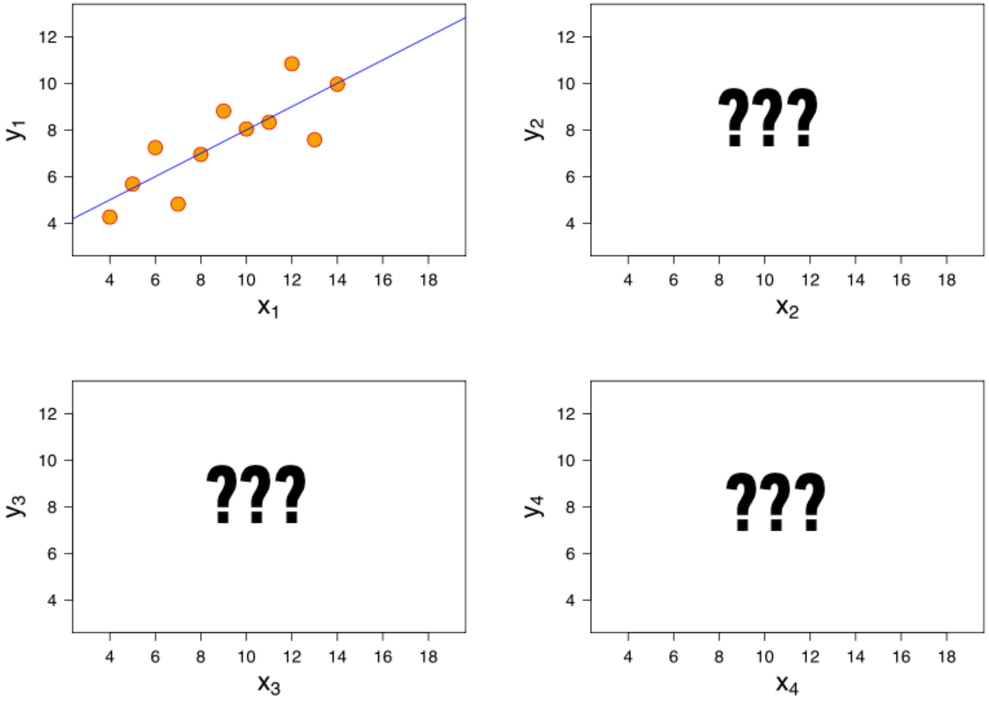
Anscombe's Quartet: Raw Data

	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
mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
corr.	0.816		0.816		0.816		0.816	

EXERCISE – WHY VISUALIZE DATA?

Now, suppose I give you three more datasets with exactly the same characteristics…

Q: how similar are these datasets?

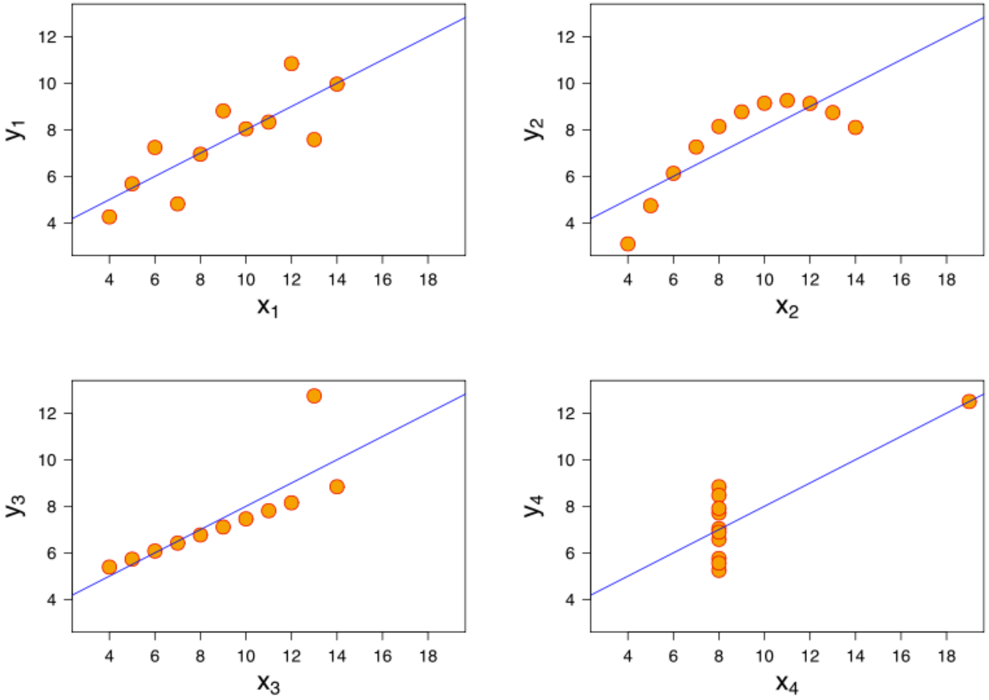


EXERCISE – WHY VISUALIZE DATA?

Now, suppose I give you three more datasets with exactly the same characteristics.

Q: how similar are these datasets?

A: not very!



http://en.wikipedia.org/wiki/Anscombe's_quartet

LAB