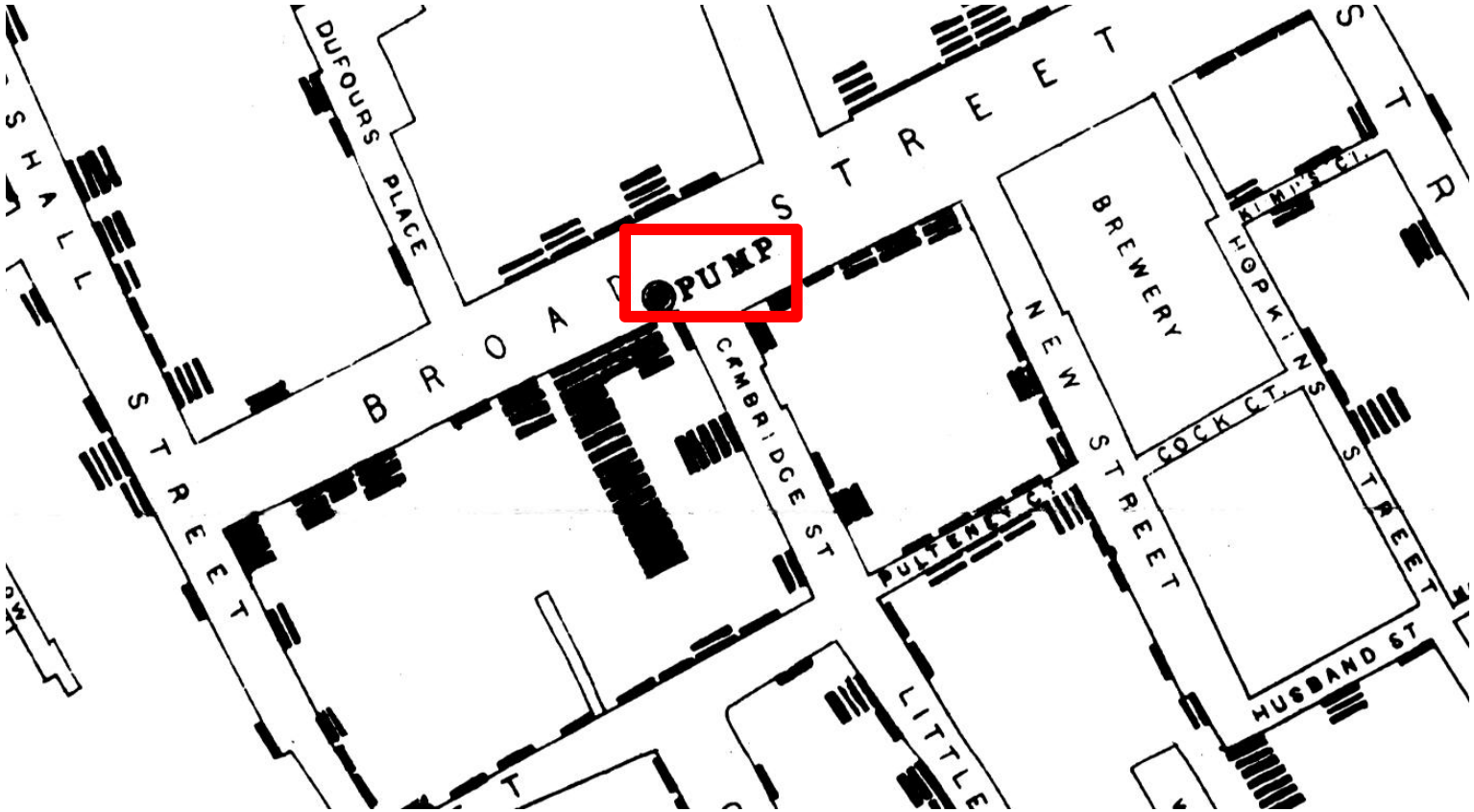
POST-PROCESSING

Post-processing

- Visualization
  - The human eye is a powerful analytical tool
  - If we visualize the data properly, we can discover patterns and demonstrate trends
  - Visualization is the way to present the data so that patterns can be seen
    - E.g., histograms and plots are a form of visualization
    - There are multiple techniques (a field on its own)

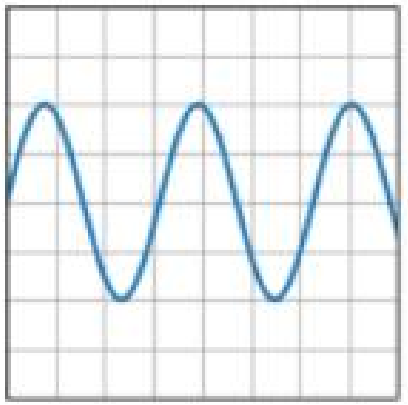
Visualization on a map

- John Snow, London 1854



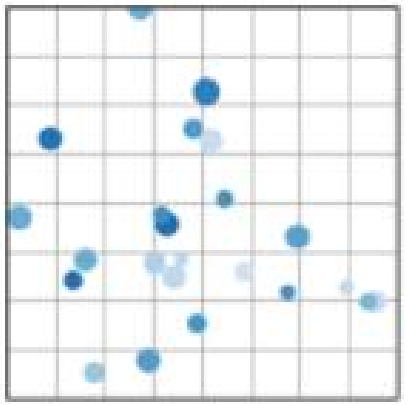
Basic

Basic plot types, usually y versus x.



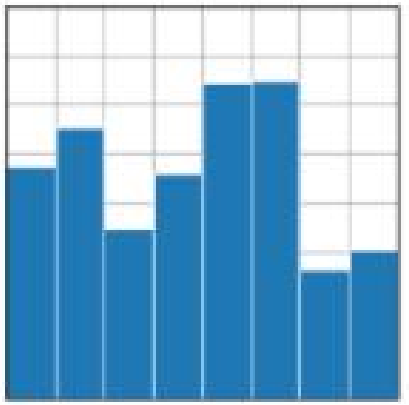
plot(x, y)

折线图



scatter(x, y)

散点图



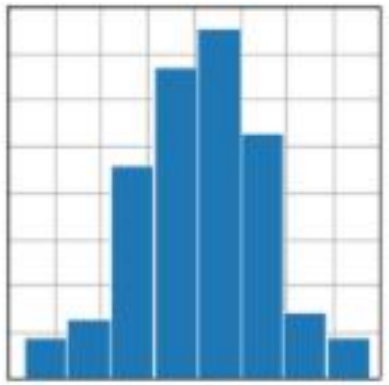
bar(x, height) / barh(y, width)

条形图

https://matplotlib.org/

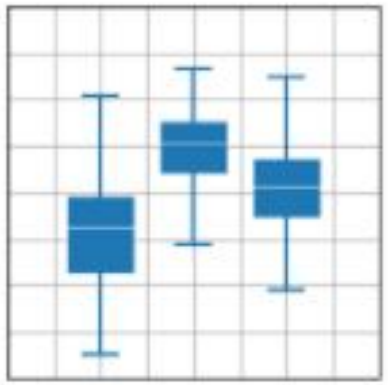
Statistics plots

Plots for statistical analysis. https://matplotlib.org/



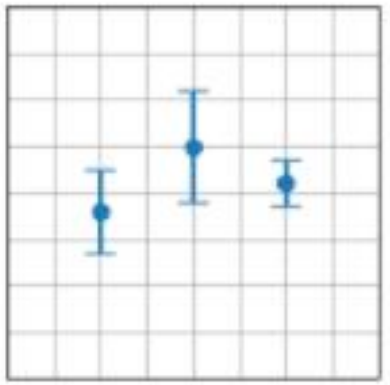
hist(x)

Histogram 直方图
分布展示



boxplot(X)

箱形图：max,min,median, 25th percentile, 75th percentile



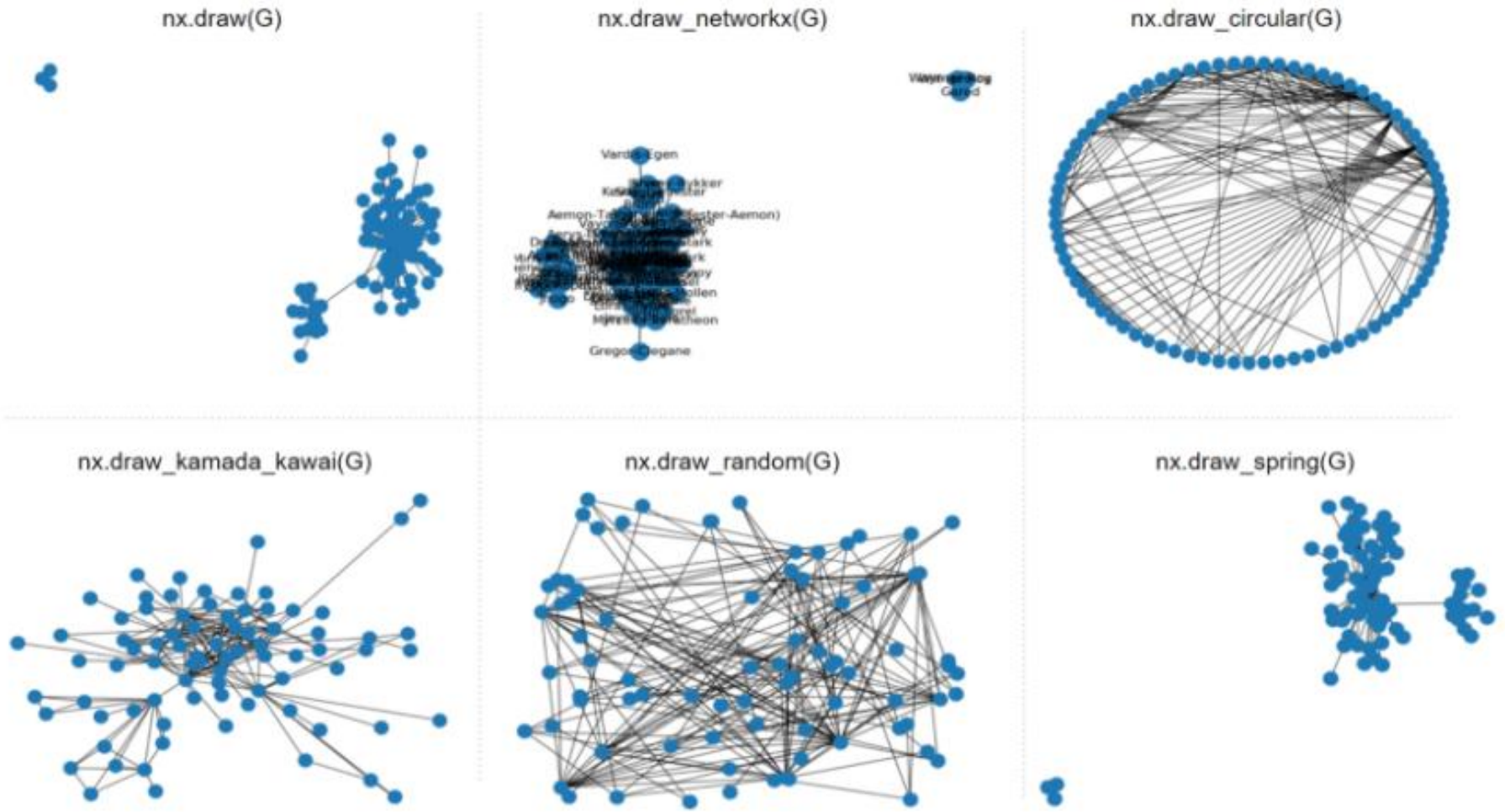
errorbar(x, y, yerr, xerr)

误差线：表示标准差或标准误

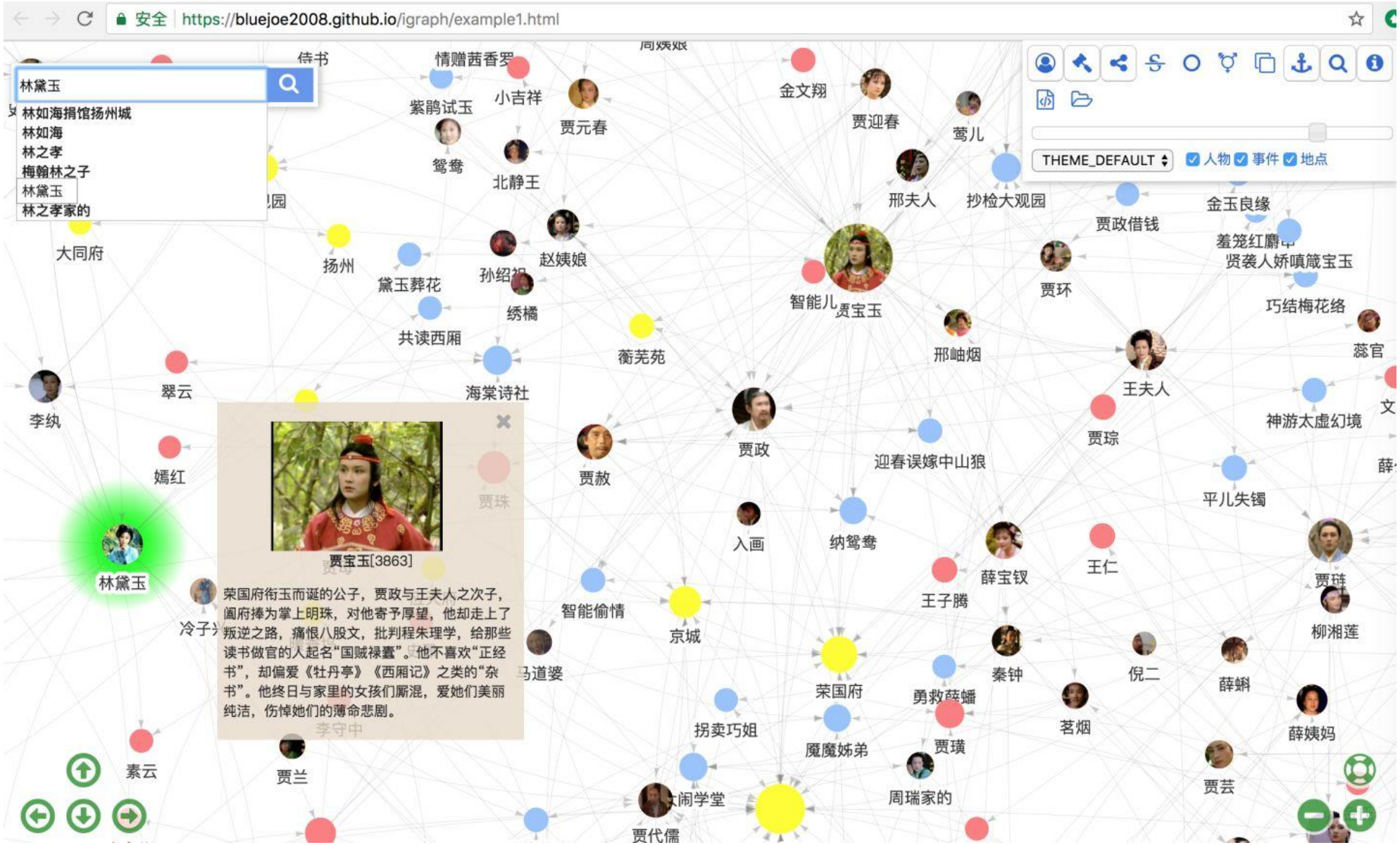
pie(x)

饼图

标准误Standard Error：衡量样本均值与总体均值之间的差异，样本均值的标准差除以样本容量的平方根



https://networkx.org/



https://grapheco.org/InteractiveGraph/dist/examples/example1.html

Dimensionality Reduction

- The human eye is limited to processing visualizations in two (at most three) dimensions
- One of the great challenges in visualization is to visualize high-dimensional data into a two-dimensional space
  - Dimensionality reduction
  - Distance preserving embeddings
- Dimensionality reduction is also a preprocessing technique:
  - Reduce the amount of data
  - Extract the useful information.

Example

- Consider the following 6-dimensional dataset

$$D = \begin{bmatrix} 1 & 2 & 3 & 0 & 0 & 0 \\ 2 & 4 & 6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 2 & 4 & 6 \\ 1 & 2 & 3 & 1 & 2 & 3 \\ 2 & 4 & 6 & 2 & 4 & 6 \end{bmatrix}$$

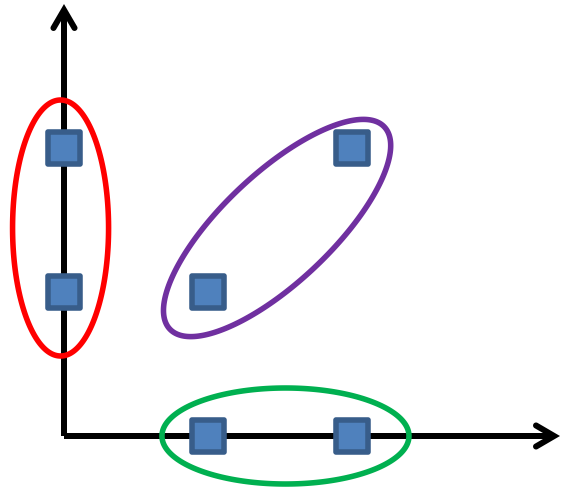
- What do you observe? Can we reduce the dimension of the data?

Example

$$D = \begin{bmatrix} 1 & 2 & 3 & 0 & 0 & 0 \\ 2 & 4 & 6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 2 & 4 & 6 \\ 1 & 2 & 3 & 1 & 2 & 3 \\ 2 & 4 & 6 & 2 & 4 & 6 \end{bmatrix}$$

- Each row is a multiple of two vectors
  - $x = [1,\ 2,\ 3,\ 0,\ 0,\ 0]$
  - $y = [0,\ 0,\ 0,\ 1,\ 2,\ 3]$
- We can rewrite D as

$$D = \begin{matrix} x & y \\ \end{matrix} \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 0 & 1 \\ 0 & 2 \\ 1 & 1 \\ 2 & 2 \end{bmatrix}$$



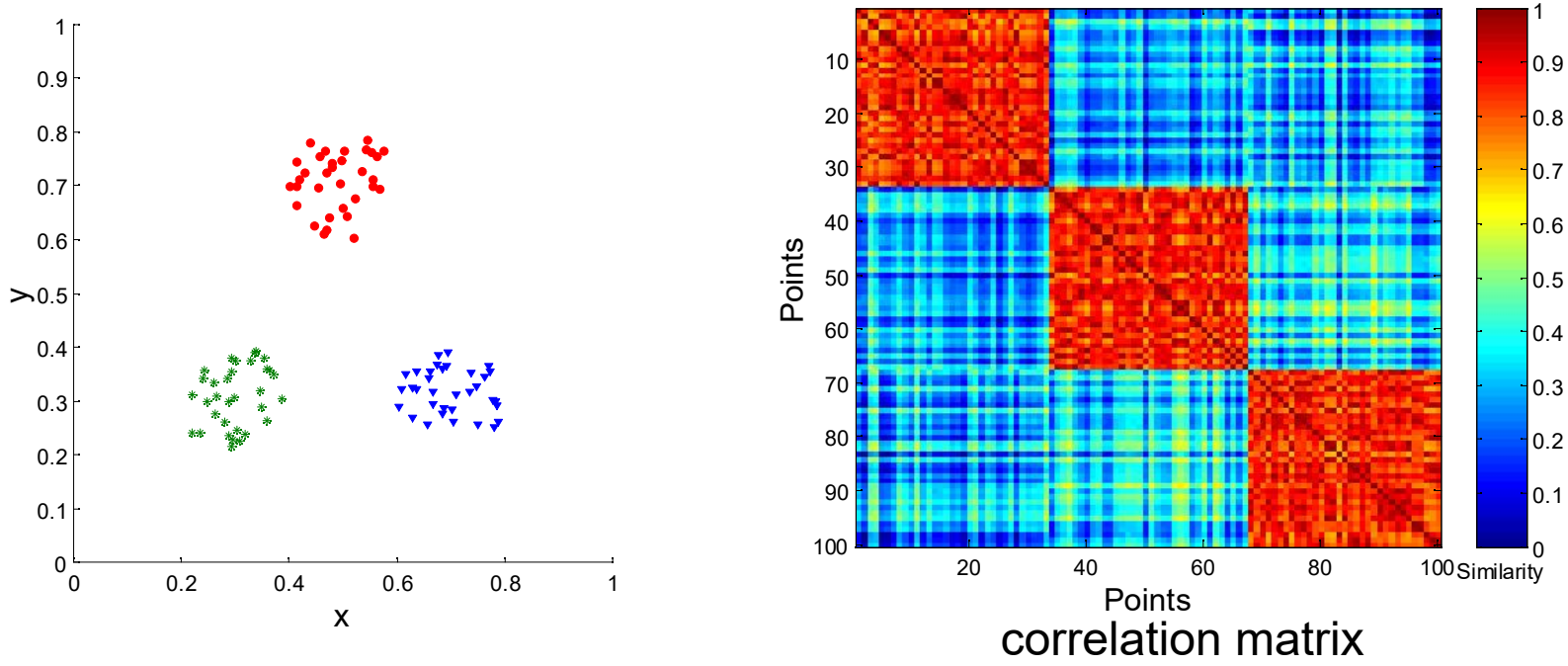
Three types of data points

Word Clouds

- A fancy way to visualize a document or collection of documents.

Heatmaps

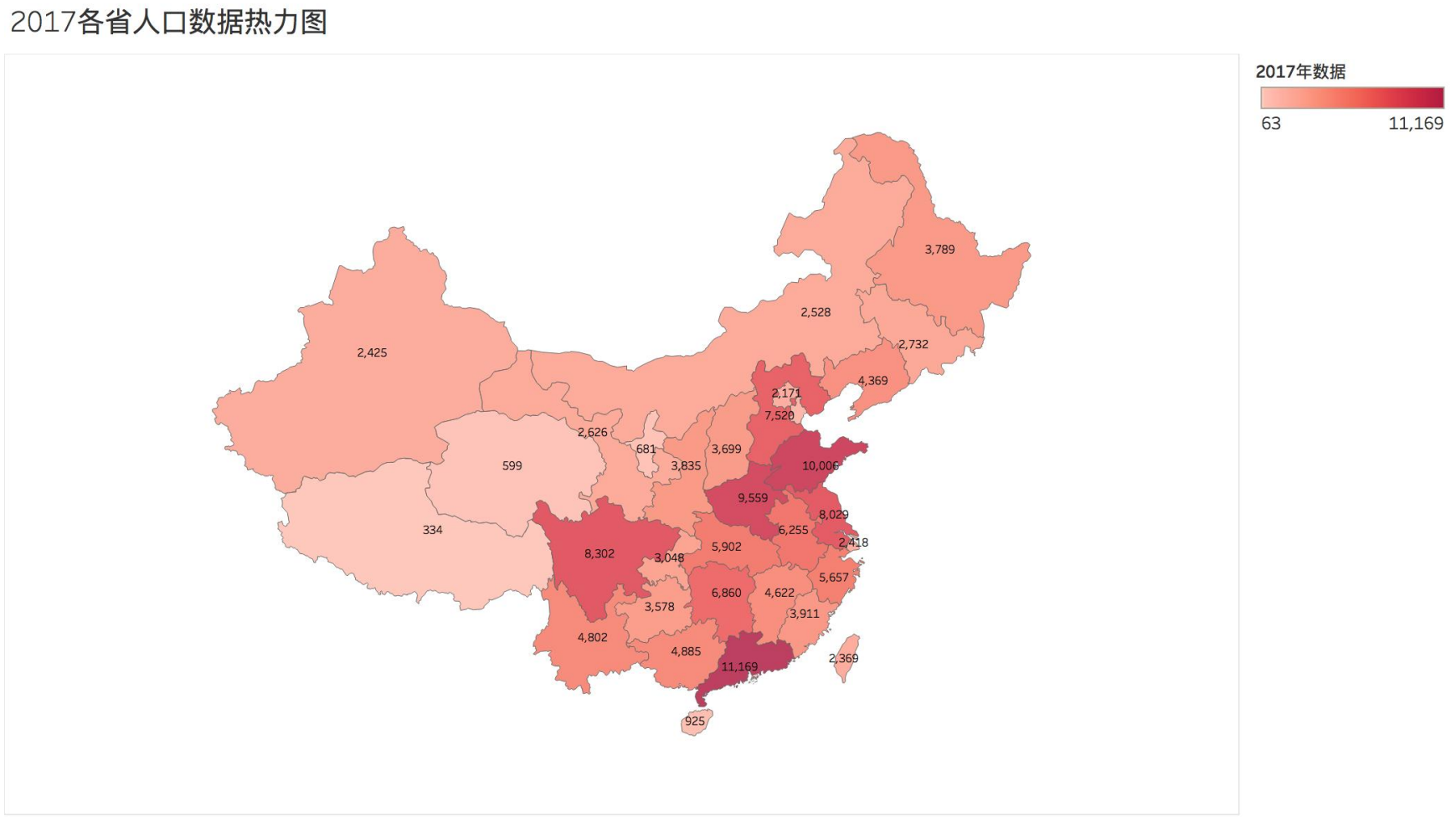
- Plot a point-to-point similarity matrix using a heatmap:
  - Deep red = high values (hot)
  - Dark blue = low values (cold)



The clustering structure becomes clear in the heatmap

Heatmaps

A very popular way to visualize data



Statistical Significance

- When we extract knowledge from a large dataset we need to make sure that what we found is not an artifact of randomness
  - E.g., we find that many people buy milk and beer together.
  - But many (more) people buy milk and beer independently
- Statistical tests compare the results of an experiment with those generated by a null hypothesis
  - E.g., a null hypothesis is that people select items independently.
- A result is interesting if it cannot be produced by randomness.
  - An important problem is to define the null hypothesis correctly: What is random?

P436 of《数据挖掘导论》避免错误发现

P-value

- A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true. The lower the p-value, the greater the statistical significance of the observed difference.
- 例子：连续抛一枚硬币5次，每次都正面朝上，判断硬币是否均匀。
  - null hypothesis：硬币是均匀的（正面朝上和反面朝上的概率一样，各50% ）
  - 如果原假设成立，结果（5次都是正面朝上）发生的概率是0.5^5=0.03125. 所以p-value = 0.03125，可以拒绝原假设。

EXPLORATORY DATA ANALYSIS

What does my data look like?

Exploratory analysis of data

- Summary statistics: numbers that summarize properties of the data
- Summarized properties include frequency（频率）, location（定位）and spread（离散程度）
  - Examples: location - mean
    spread - standard deviation
- Most summary statistics can be calculated in a single pass through the data
- Computing data statistics is one of the first steps in understanding our data

Frequency and Mode

- The frequency of an attribute value is the percentage of time the value occurs in the data set
  - For example, given the attribute ‘gender’ and a representative population of people, the gender ‘female’ occurs about 50% of the time.
- The mode（众数） of an attribute is the most frequent attribute value
- The notions of frequency and mode are typically used with categorical data

- We can visualize the data frequencies using a value histogram

Example

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
10	No	Single	90K	No

Marital Status

Single	Married	Divorced	NULL
4	3	2	1

Mode: Single

Example

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
10	No	Single	90K	No

Marital Status

Single	Married	Divorced	NULL
40%	30%	20%	10%

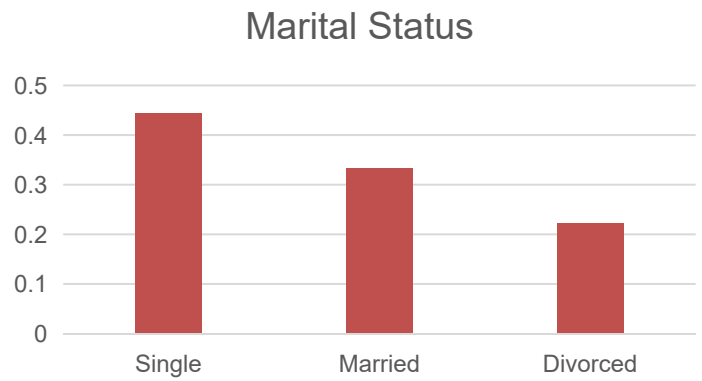
Example

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
10	No	Single	90K	No

We can choose to ignore NULL values

Marital Status

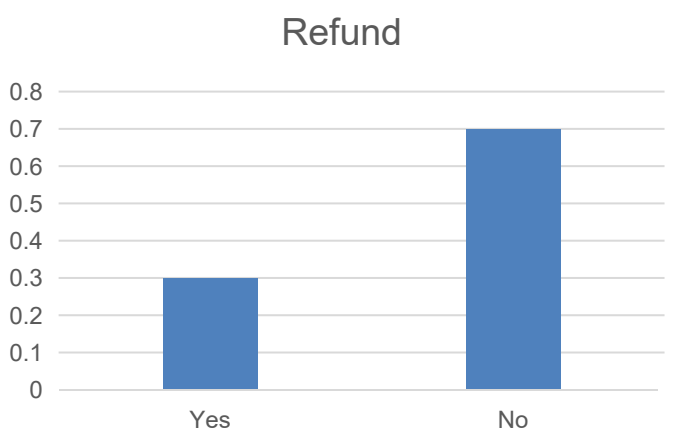
Single	Married	Divorced
44%	33%	22%

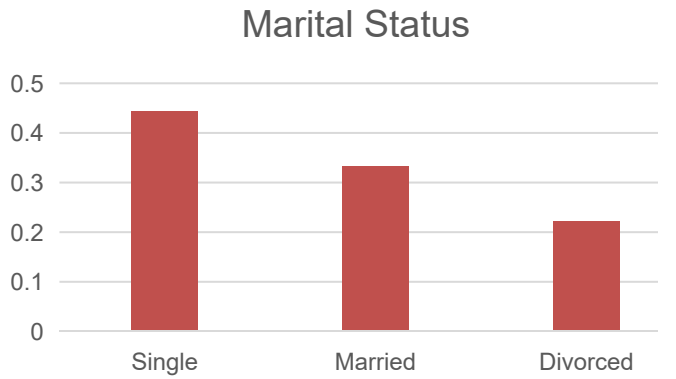


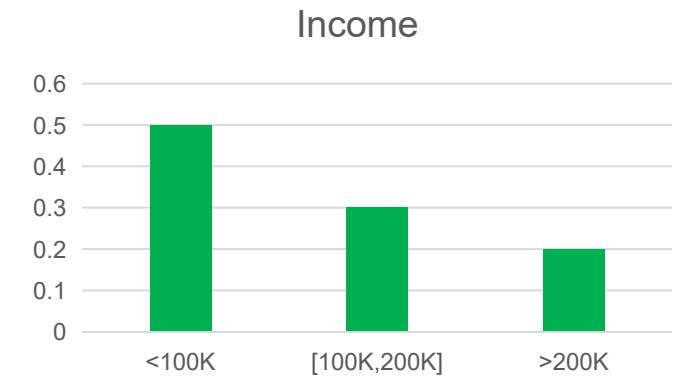
Data histograms

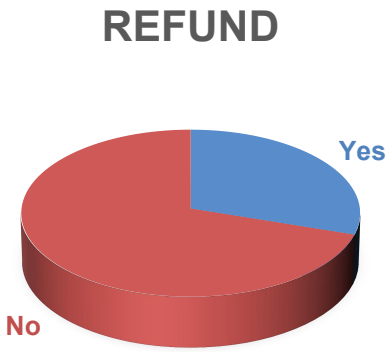
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
10	No	Single	90K	No

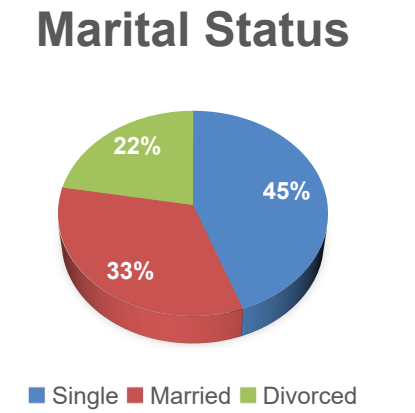
Use binning for numerical values

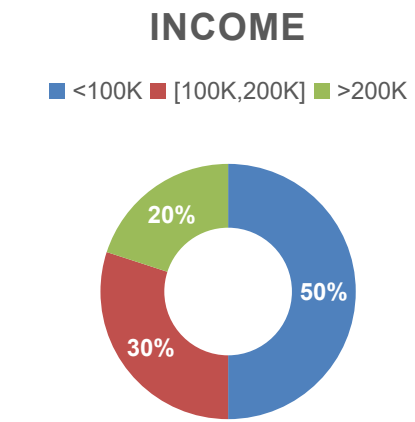












Percentiles

- For continuous data, the notion of a percentile（百分位数） is more useful.

Given an ordinal or continuous attribute x and a number p between 0 and 100, the p^{th} percentile（第p百分位数） is a value x_p of x such that $p\%$ of the observed values of x are less than or equal to x_p.

- For instance, the 80th percentile is the value $x_{80\%}$ that is greater or equal than 80% of all the values of x we have in our data.

Example

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
10	No	Single	90K	No

Taxable Income
10000K
220K
125K
120K
100K
90K
90K
85K
70K
60K

$x_{80\%}$ = 125K

Measures of Location: Mean and Median

- The mean is the most common measure of the location of a set of points.

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

- However, the mean is very sensitive to outliers.
- Thus, the median is also commonly used.

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

- Or the trimmed mean（裁剪平均值）: the mean after removing min and max values

Example

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
10	No	Single	90K	No

Mean: 1090K

Trimmed mean (remove min, max): 105K

Median: (90+100)/2 = 95K

Measures of Spread: Range and Variance

- Range（极差、全距） is the difference between the max and min

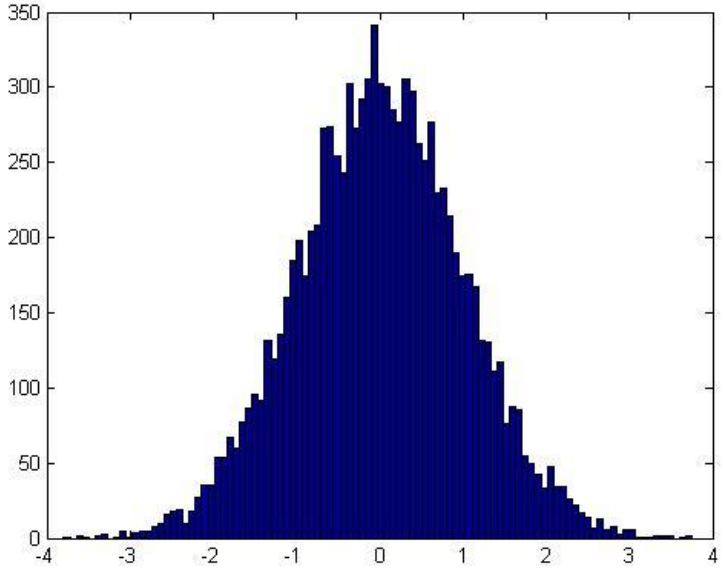
- The variance（方差） or standard deviation（标准差） is the most common measure of the spread of a set of points.

$$var(x) = \frac{1}{m}\sum_{i=1}^{m}(x - \bar{x})^2$$

$$\sigma(x) = \sqrt{var(x)}$$

Normal Distribution

- $\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$



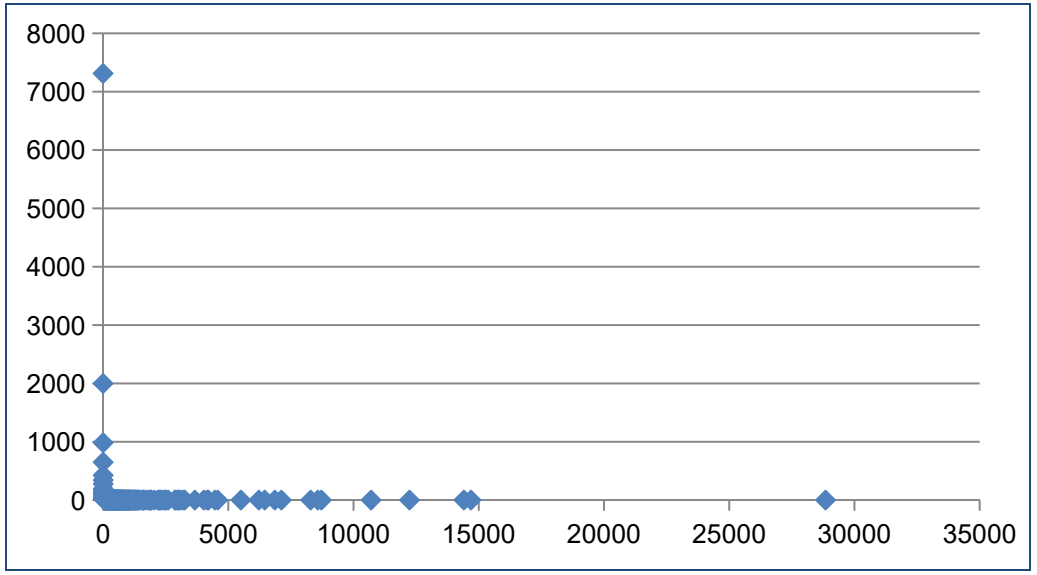
This is a value histogram

- An important distribution that characterizes many quantities and has a central role in probabilities and statistics.
- Fully characterized by the mean μ and standard deviation σ

Not everything is normally distributed

- Plot of number of words with x number of occurrences

y: number of words with x number of occurrences



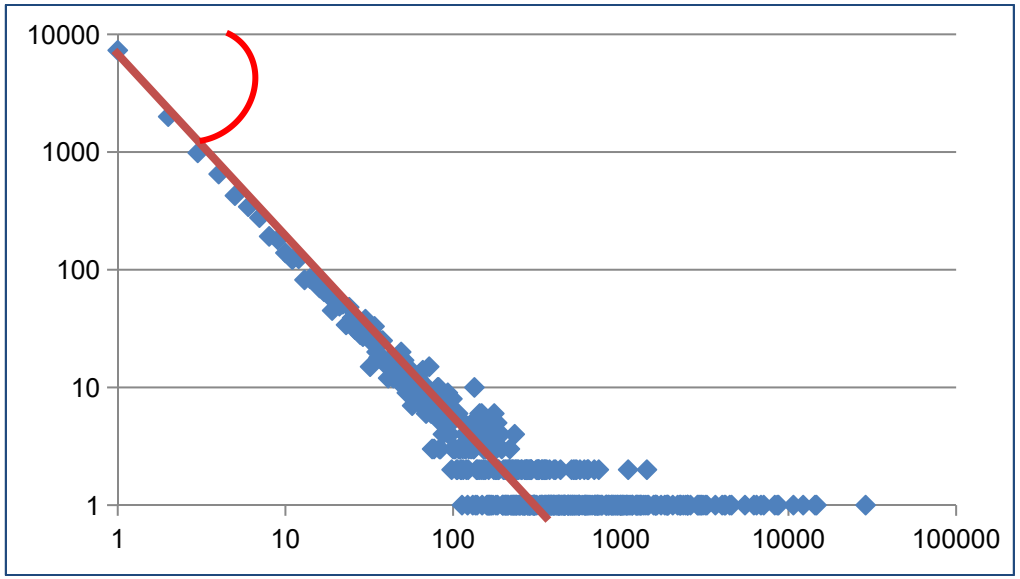
x: number of occurrences

- If this was a normal distribution we would not have number of occurrences as large as 28K

Power-law distribution

- We can understand the distribution of words if we take the log-log plot

y: logarithm of number of words with x number of occurrences



x: logarithm of number of occurrences

Power-law distribution:

$$p(k) = k^{-a}$$

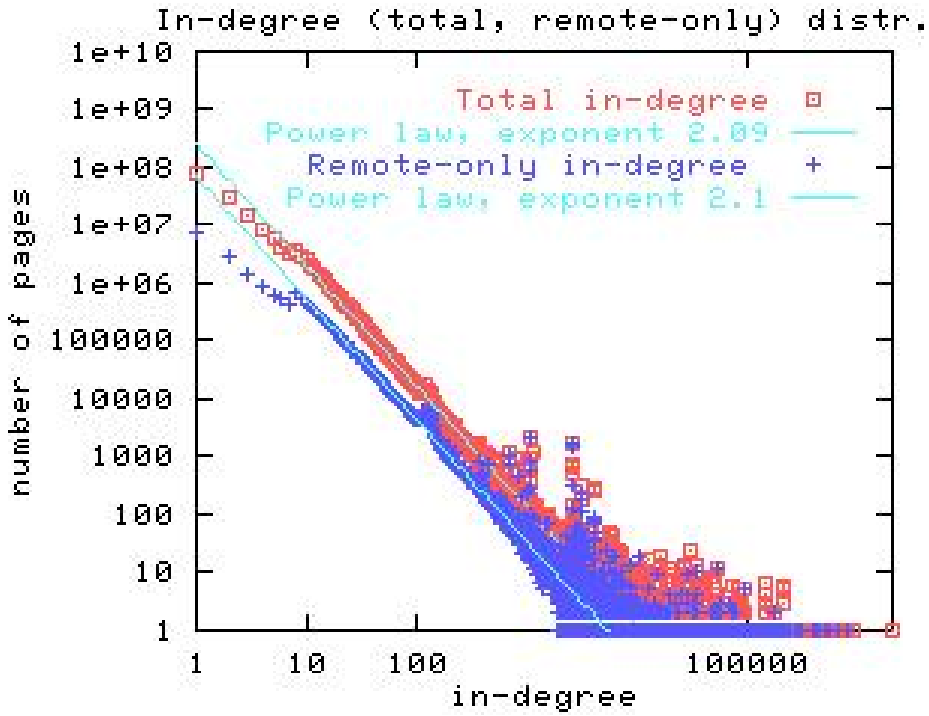
The slope of the line gives us the exponent α

Linear relationship in the log-log space

$$\log p(x = k) = -a \log k$$

Power-laws are everywhere

- number of friends in social networks, number of occurrences of words, city sizes, income distribution, popularity of products and movies
  - Signature of human activity?
  - A mechanism that explains everything?
  - Rich get richer process



Attribute relationships

- In many cases it is interesting to look at two attributes together to understand if they are correlated
  - Is there a relationship between years of study and income?
- How do we visualize these relationships?

Correlating numerical attributes

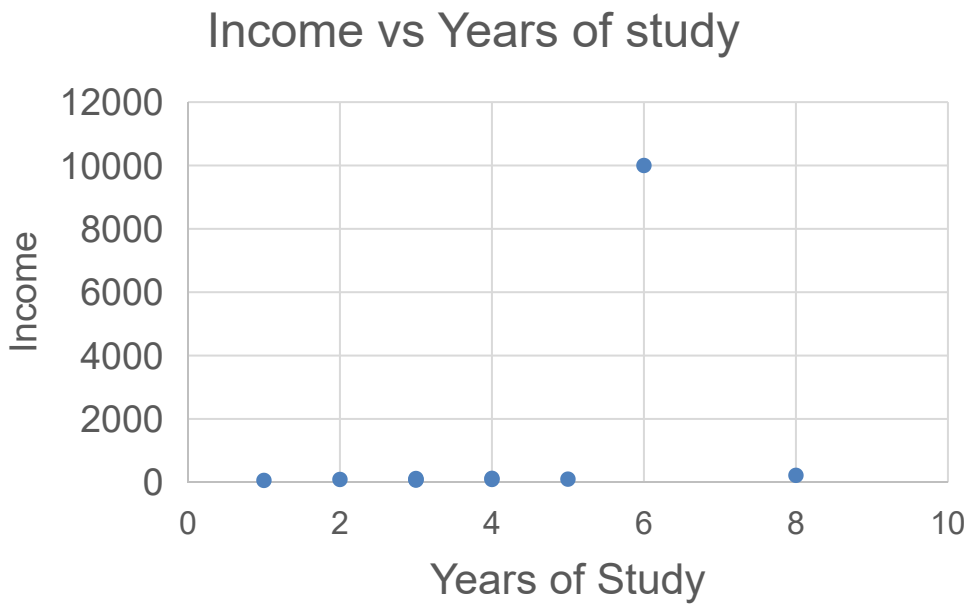
Tid	Refund	Marital Status	Taxable Income	Years of Study
1	Yes	Single	125K	4
2	No	Married	100K	5
3	No	Single	70K	3
4	Yes	Married	120K	3
5	No	Divorced	10000K	6
6	No	NULL	60K	1
7	Yes	Divorced	220K	8
8	No	Single	85K	3
9	No	Married	90K	2
10	No	Single	90K	4

Scatter plot:

X axis is one attribute, Y axis is the other

For each entry we have two values

Plot the entries as two-dimensional points



Correlating numerical attributes

Tid	Refund	Marital Status	Taxable Income	Years of Study
1	Yes	Single	125K	4
2	No	Married	100K	5
3	No	Single	70K	3
4	Yes	Married	120K	3
5	No	Divorced	10000K	6
6	No	NULL	60K	1
7	Yes	Divorced	220K	8
8	No	Single	85K	3
9	No	Married	90K	2
10	No	Single	90K	4

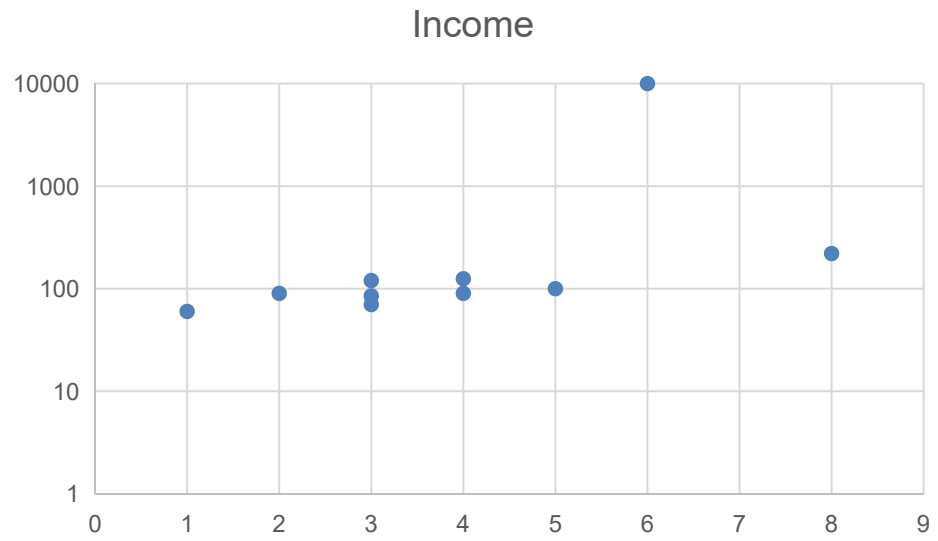
Scatter plot:

X axis is one attribute, Y axis is the other

For each entry we have two values

Plot the entries as two-dimensional points

Log-scale in y-axis makes the plot look a little better



Plotting attributes against each other

Tid	Refund	Marital Status	Taxable Income	Years of Study
1	Yes	Single	125K	4
2	No	Married	100K	5
3	No	Single	70K	3
4	Yes	Married	120K	3
5	No	Divorced	10000K	6
6	No	NULL	60K	1
7	Yes	Divorced	220K	8
8	No	Single	85K	3
9	No	Married	90K	2
10	No	Single	90K	4

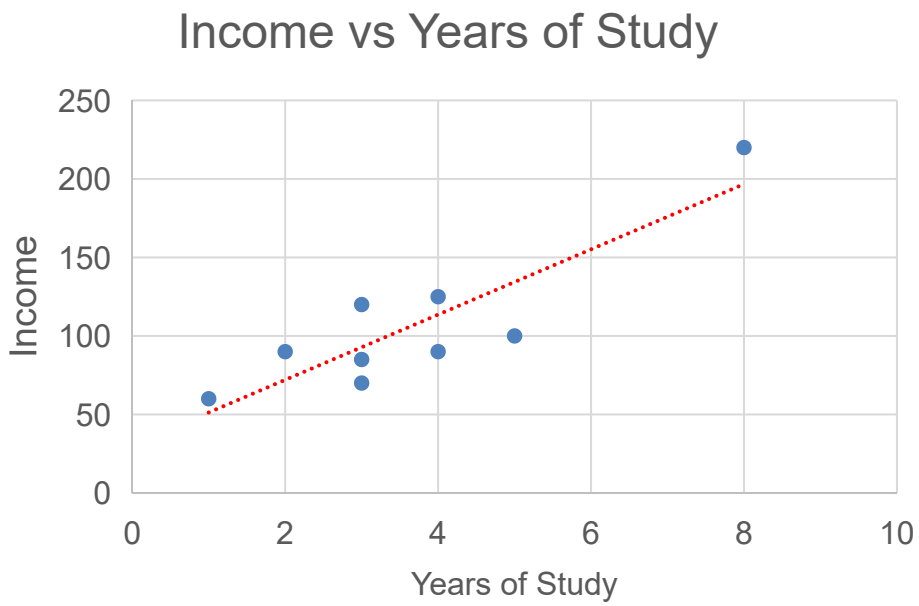
Scatter plot:

X axis is one attribute, Y axis is the other

For each entry we have two values

Plot the entries as two-dimensional points

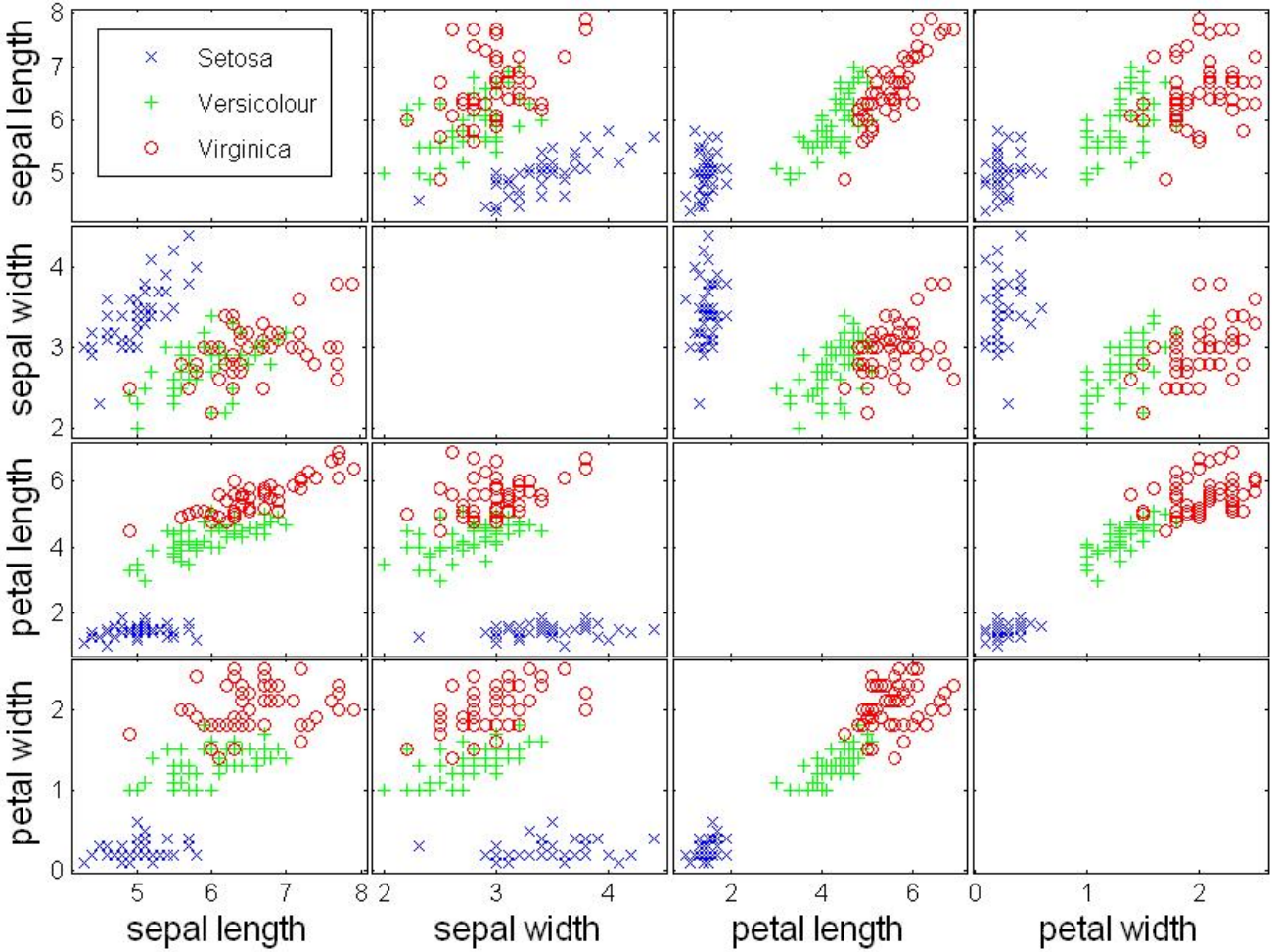
After removing the outlier value there is a clear correlation



Scatter Plot Array of Iris Attributes

For multiple attribute pairs

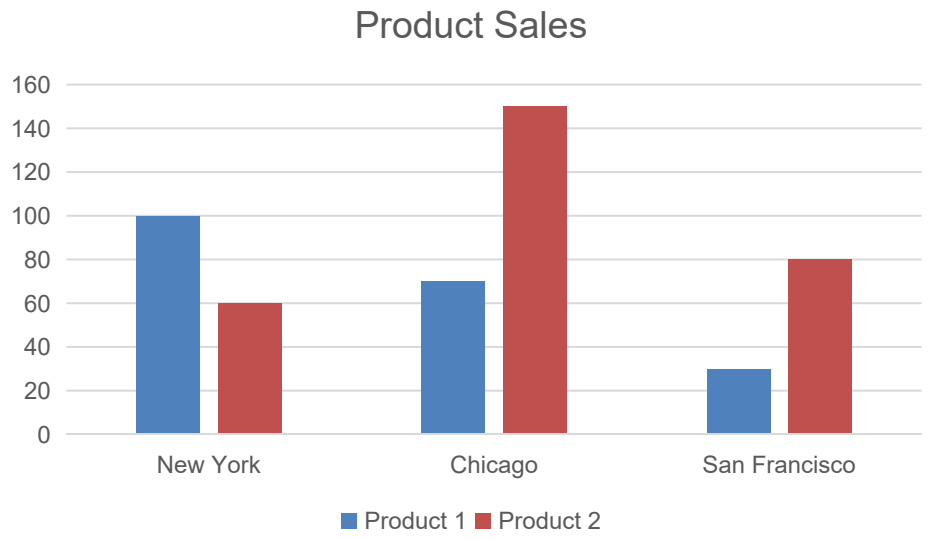
from sklearn import datasets
iris = datasets.load_iris()



https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html

Plotting attributes together

City	Product 1	Product 2
New York	100	60
Chicago	70	150
San Francisco	30	80

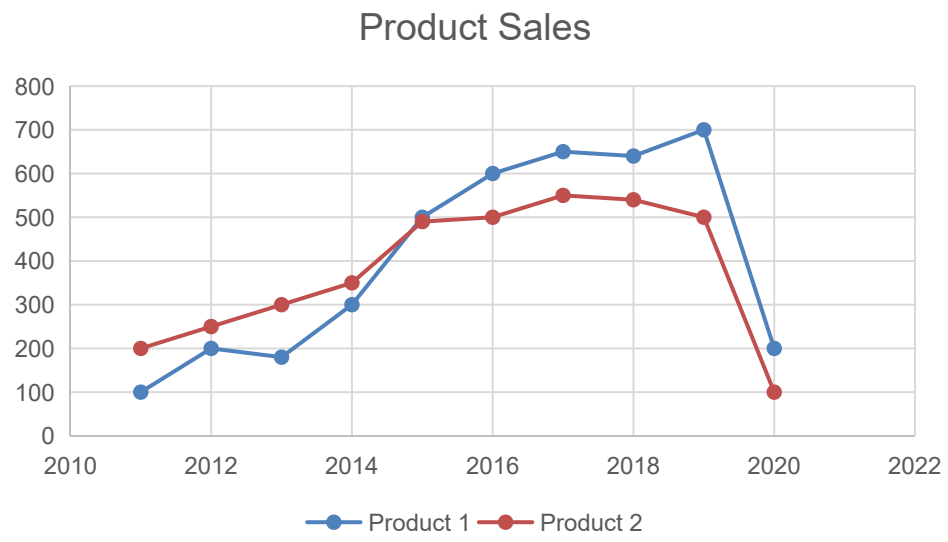


How would you visualize the differences between the product sales per city?

Plotting attributes together

Year	Product 1	Product 2
2011	100	200
2012	200	250
2013	180	300
2014	300	350
2015	500	490
2016	600	500
2017	650	550
2018	640	540
2019	700	500
2020	200	100

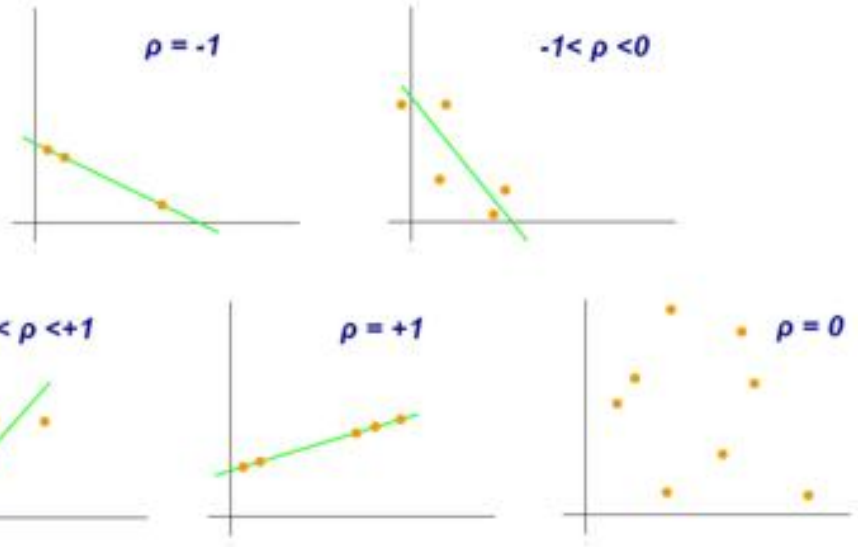
How would you visualize the differences between the product sales over time?



Measuring correlation

- Pearson correlation coefficient: measures the extent to which two variables are linearly correlated
  - $X = \{x_1, ..., x_n\}$
  - $Y = \{y_1, ..., y_n\}$

Must have pairs of observations



- $corr(X, Y) = \frac{\sum_i (x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_i (x_i - \mu_X)^2}\sqrt{\sum_i (y_i - \mu_Y)^2}}$
- It comes with a p-value
  - The p-value is the probability that the correlation was by chance.
- Assumes no outliers and that the variables are normally distributed

- Spearman rank correlation coefficient: tells us if two variable are rank-correlated
  - They place items in the same order – Pearson correlation of the rank vectors
  - For ranking without ties it looks at the differences between the ranks of the same items

The scores for nine students in physics and math are as follows:

Physics: 35, 23, 47, 17, 10, 43, 9, 6, 28

Mathematics: 30, 33, 45, 23, 8, 49, 12, 4, 31

Physics	Rank	Math	Rank
35	3	30	5
23	5	33	3
47	1	45	2
17	6	23	6
10	7	8	8
43	2	49	1
9	8	12	7
6	9	4	9
28	4	31	4

	名字	项目标题
1	黄河清 杨超凡	利用CBDB研究隋唐时期官员任用升迁与门阀科举之间的联系
2	张羽飞 刘晨	利用日常生活数据集探究日常习惯如何影响成绩
3	陈欣禾 刘文婷	利用GDT数据集预测恐怖袭击/新闻事件或人物分析
4	高胜寒 吴明正	
5	丁鹏程 林天卫	基于中国家谱总目的数据清洗及挖掘
6	李盛忻 董佳和	
7	刘鹏程	面向移动电子商务的淘宝用户购物行为预测