HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

IT4142E

Introduction to Data Science

Chapter 4: Introduction to Exploratory Data Analysis

Lecturer:

Muriel VISANI: murielv@soict.hust.edu.vn

Acknowledgements:

Khoat Than
Viet-Trung Tran

Department of Information Systems
School of Information and Communication Technology - HUST

Contents of the course

- Chapter 1: Overview
- Chapter 2: Data scraping
- Chapter 3: Data cleaning, pre-processing and integration
- Chapter 4: Introduction to Exploratory Data Analysis
- Chapter 5: Introduction to Data visualization
- Chapter 6: Introduction to Machine Learning
  - Performance evaluation
- Chapter 7: Introduction to Big Data Analysis
- Chapter 8: Applications to Image and Video Analysis

Learning outcomes

- Understand key elements in exploratory data analysis (EDA)
- Explain and use common summary statistics for EDA
- Plot and explain common graphs and charts for EDA

Goals of this chapter

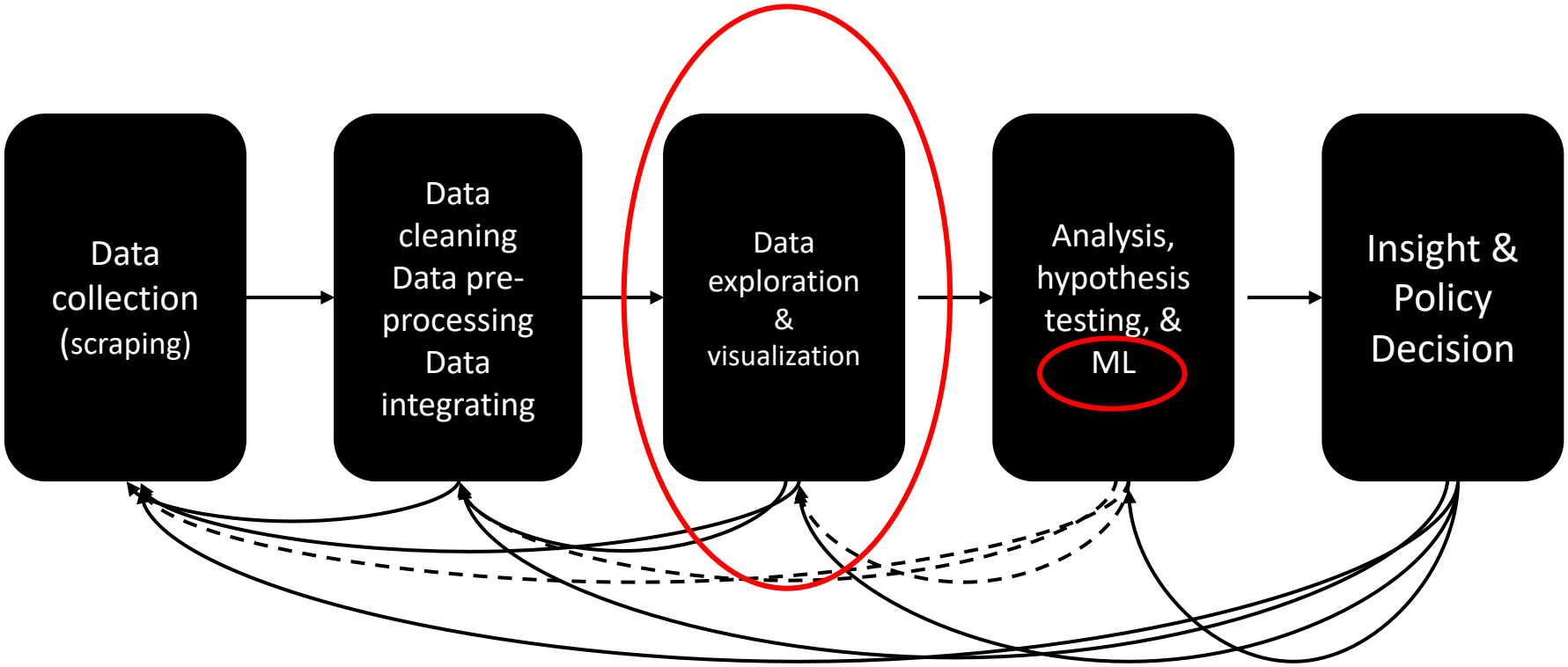
Goal	Description of the goal
M1	**Understand and be able to design and manage the systems which are based on Data Science (DS)**
M1.2	Identify, compare, and categorize the data types and systems in practice
M1.3	Be able to design systems based on DS in their future organizations
M2	**Identify and manage the opportunities from DS to boost the existing organizations, or develop new organizations**
M2.1	Understand and promote the use of DS to support their organizations
M2.2	Identify the (possible) impacts of Data Science on their organizations

Contents of this chapter

- **Chapter 4: Exploratory Data Analysis (EDA)**
  - Introduction and definitions
  - Univariate EDA measures
    - Central tendency (location)
    - Scale (data spread)
    - Shape of the distribution
  - Univariate EDA graphics
  - Multivariate EDA measures
  - Multivariate EDA graphics
  - More advanced EDA techniques
  - Summary
  - Homework

Introduction and definition

Recall: DS methodology



Motivation for EDA

- Before making inferences from data, it is essential to examine all your variables.
  - To understand your data
- Why?
  - To "listen" to the data:
    - to catch mistakes
    - to see patterns in the data
    - to find violations of statistical assumptions
    - to generate hypotheses
  - …and because if you don't, you will have trouble later

What is EDA?

- EDA **summarizes** all of the available data, so that it is **understandable by humans**
  - Summary statistics
  - Visualization (basic graphics shown in this chapter or more refined visualization in Chapter 5)
  - Dimensionality reduction
    - Principal Component Analysis (for instance)
  - Clustering and outliers - anomaly detection
    - Clustering will be seen in Chapter 6 (Machine Learning)
    - But, you can still use off-the-shelf clustering methods for EDA!

- Thanks to EDA, data scientists might find good models for the data
  - For instance, the data is / is not Gaussian
    - -> impact of the data models that can be used for data mining

EDA common questions

- What is a typical value for attribute X?
- What is the uncertainty for a typical value?
- What is a good distributional fit for attribute X / attributes X&Y?
- Does the data have outliers?
- Can we separate signal from noise in time dependent data?
- What are the most « important » attributes / combination of attributes?
- Is the data consistent or inconsistent among different data sources?

What is EDA?

- EDA is an iterative process: repeat...
  - Identify and prioritize relevant questions in decreasing order of importance
  - Ask these questions (might require data processing)
  - Construct graphics to address these questions
  - Inspect the answers and derive new questions

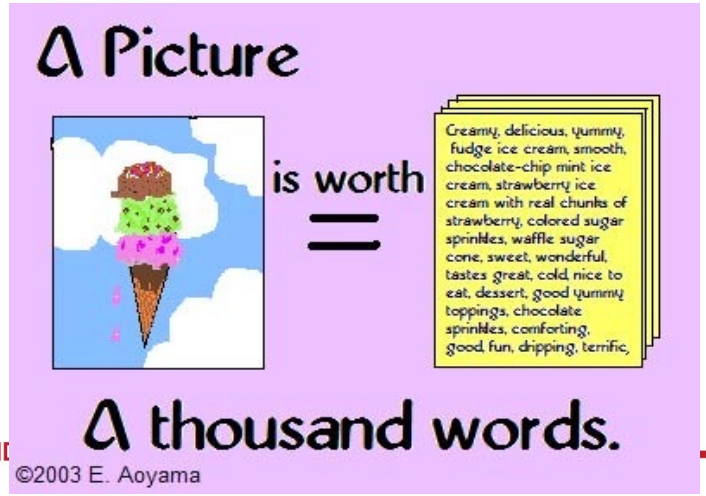
EDA strategy

- Examine variables one by one (univariate analysis)…
- … then, explore the relationships among the different variables (multivariate analysis)…
- … by applying techniques that are **adapted** to the types of the attributes (variables)
  - Categorical vs. Numeric

EDA techniques

- Quantitative techniques
  - Mostly statistics / statistical models – some ML
    - Most quantitative measures are implemented in the **pandas** library (Python)

- Graphical techniques (visualization)
  - Box plots, scatter plots, bar plots, histograms, probability plots, residual plots…
    - Most basic graphical representations, seen in this chapter, can be implemented using the **pandas / matplotlib** libraries (Python)

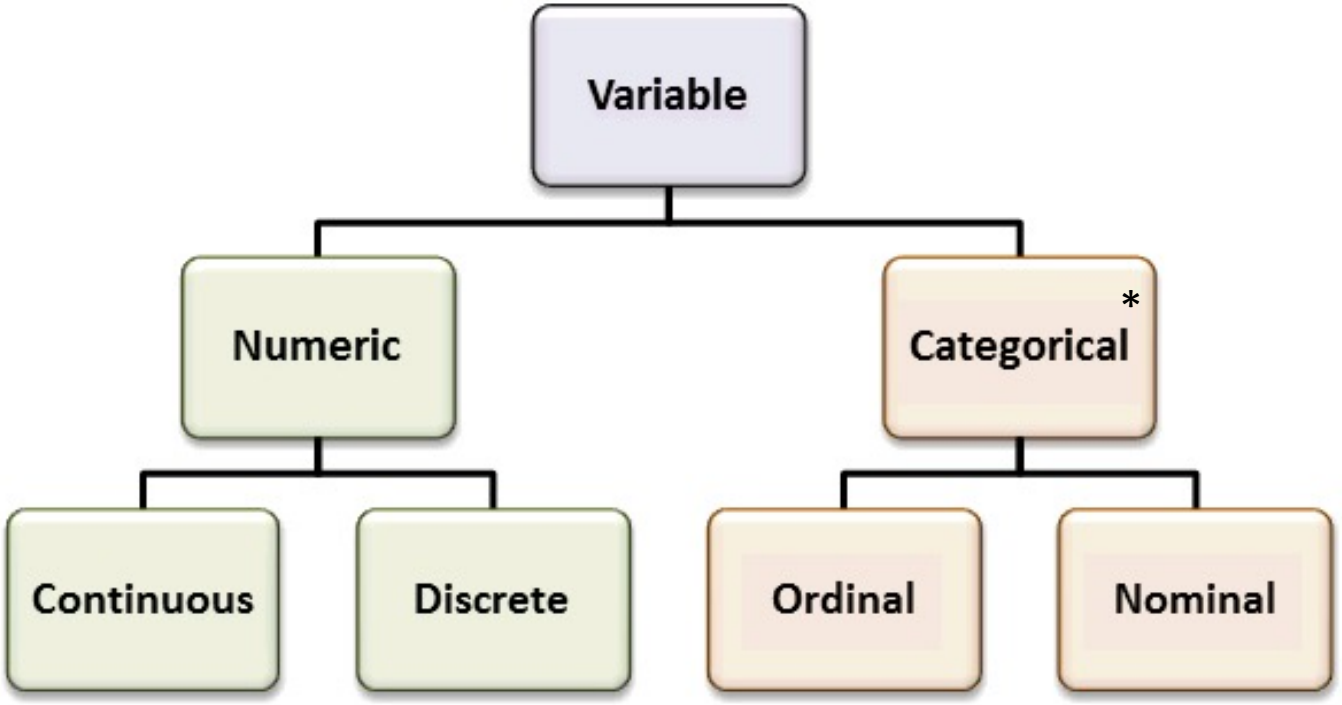


Defs: observations and variables

- Data is an collection of observations (records, subjects)
  - Rows in the data table
- An attribute is a set of values describing some aspect across all observations, it is called a variable
  - Columns in the data table

HR Information		Contact	
Position	**Salary**	**Office**	**Extn.**
Accountant	$162,700	Tokyo	5407
Chief Executive Officer (CEO)	$1,200,000	London	5797
Junior Technical Author	$86,000	San Francisco	1562
Software Engineer	$132,000	London	2558

Types of variables



*The set of possible values of categorical variables are called modalities

* Some authors call categorical variables “discrete variables”

- I do **not**, because I find this confusing, but you might find it on the web…

Dimensionality of EDA

- Univariate: analysis of **one variable**
- Bivariate: analysis of **two variable** simultaneously
- Multivariate: analysis of **several variables** simultaneously
  - Bivariate is a special case of multivariate

Univariate EDA measures

Univariate EDA measures

- Measures of central tendency (location):
  - typical or central value that best describes the data
- Measures of scale (spread):
  - estimate of the attribute's variability
- Measures of shape
  - How flat is the distribution?
  - How symmetric / asymmetric is the distribution?

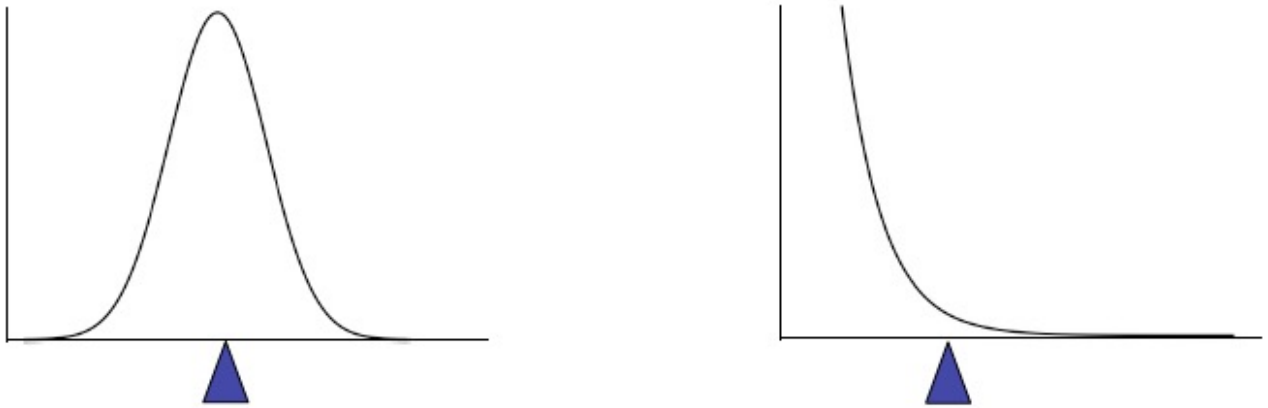
Univariate EDA measures

Central tendency (location)

Mean

- To calculate the average value of a set of observations, sum of their values divided by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



Median

- The median is the value having half the data smaller, and half the data larger than this value
- Calculation
  - If there are an odd number of observations, median is the middle value
  - If there are an even number of observations, find the middle two values and average them -> median
- Example
  - Age of participants: 17 19 21 22 23 23 23 38
  - **Median = (22+23)/2 = 22.5**

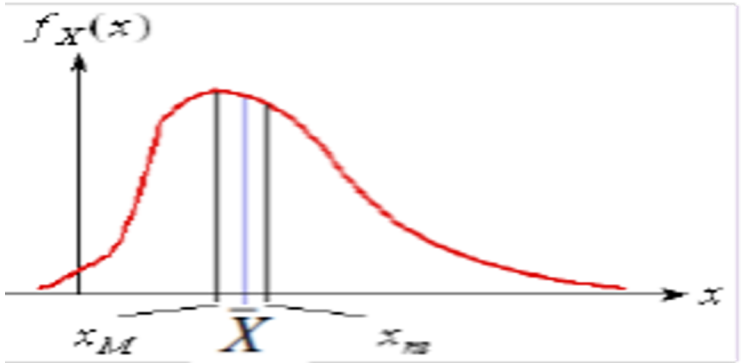
Mode

- mode is the most commonly reported value for a particular variable
  - Eg. 3, 4, 5, 6, 7, 7, 7, 8, 8, 9. Mode = 7
  - Eg. 3, 4, 5, 6, 7, 7, 7, 8, 8, 8, 9. Mode = {7, 8} = 7.5

Summary of location measures

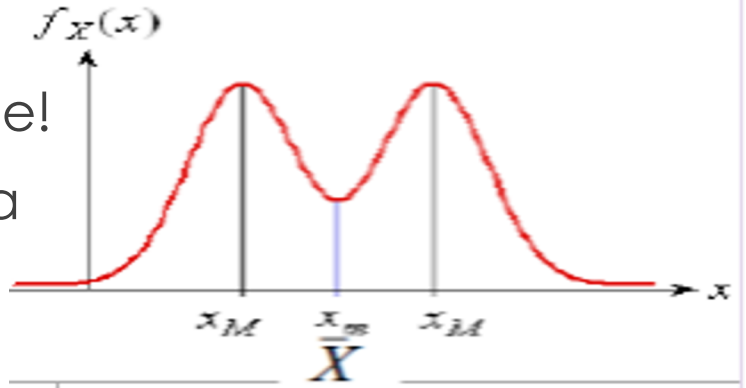
- Notations:
  - $\bar{x}$=mean
  - x_m =median
  - x_M =mode

In the general case,
$\bar{x} \neq x_m \neq x_M$

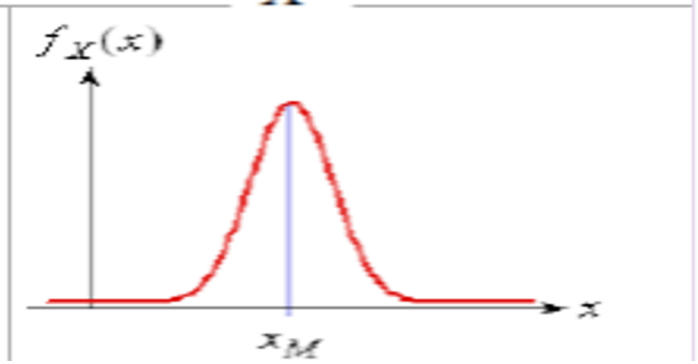


The mode is not necessarily unique!

$\bar{x} = x_m$only for symmetric data

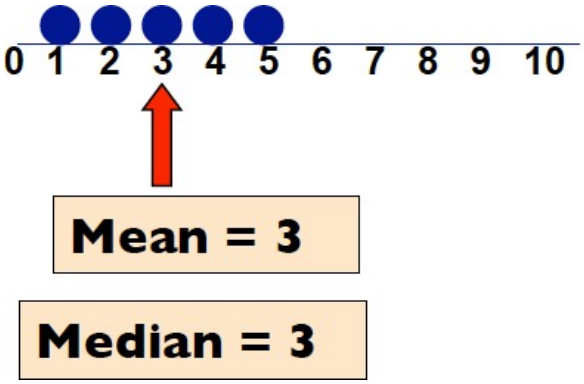


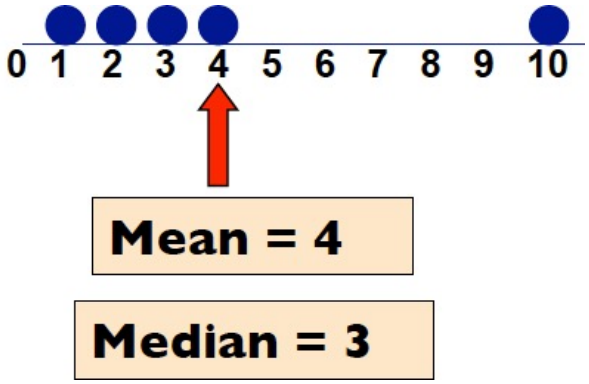
$\bar{x} = x_m = x_M$ only for symmetric, unimodal data



Which location measure is best?

- Well, it all depends on your data distribution and what you want to do with the measure!
  - Mean is best for symmetric distributions without outliers
  - But, median is more robust than mean towards outliers





  - Median is also (in general) more useful for skewed distributions

Beware – when using tools/libraries

- Sometimes, the categorical data is encoded by numbers
  - For instance, gender (male/female) can be encoded by Boolean (0/1)
  - Either because the source data was already encoded…
  - … or because, during data pre-processing (see Chapter 3-part2), we applied some **encoding** techniques
- By default, your tool (or Python library) does not "understand" your variables!!!
  - Beware: do **not** consider the mean of a categorical variable encoded as numerical!!!
  - Example of pandas .describe()

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
count	364	364	364.000000	364	364	364	364.000000	364	3.640000e+02
unique	364	30	NaN	5	22	17	NaN	115	NaN
top	Cleanthony Early	New Orleans Pelicans	NaN	SG	24.0	6-9	NaN	Kentucky	NaN
freq	1	16	NaN	87	41	49	NaN	22	NaN
mean	NaN	NaN	16.829670	NaN	NaN	NaN	219.785714	NaN	4.620311e+06
std	NaN	NaN	14.994162	NaN	NaN	NaN	24.793099	NaN	5.119716e+06
min	NaN	NaN	0.000000	NaN	NaN	NaN	161.000000	NaN	5.572200e+04
20%	NaN	NaN	4.000000	NaN	NaN	NaN	195.000000	NaN	9.472760e+05
40%	NaN	NaN	9.000000	NaN	NaN	NaN	212.000000	NaN	1.638754e+06
50%	NaN	NaN	12.000000	NaN	NaN	NaN	220.000000	NaN	2.515440e+06
60%	NaN	NaN	17.000000	NaN	NaN	NaN	228.000000	NaN	3.429934e+06
80%	NaN	NaN	30.000000	NaN	NaN	NaN	242.400000	NaN	7.838202e+06
max	NaN	NaN	99.000000	NaN	NaN	NaN	279.000000	NaN	2.287500e+07

Univariate EDA measures

Scale (spread)

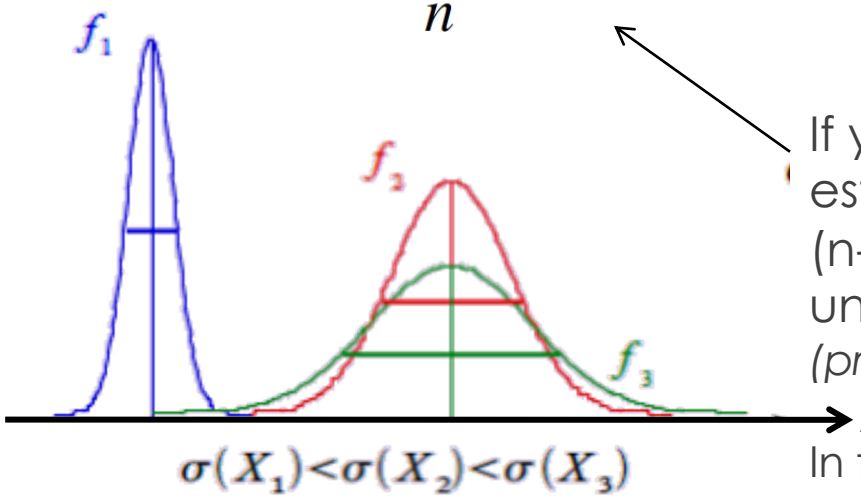
Variance and standard deviation

- Variance: the average of squared differences between the observation values and the mean

$$\hat{\sigma}^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n}$$

- Standard Deviation: just the square root of the variance

$$\hat{\sigma} = \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{n}}$$



If you want unbiased, convergent estimations, you would have to divide by (n-1), because the actual mean is unknown (estimated using $\bar{x}$)
(proof outside of the scope of this course)

In this course, we'll use only the formulas above

Quizz

- Can variance, standard deviation be used for
  - Numeric, continuous variables?
    - Answer:
  - Numeric, discrete variables?
    - Answer :
  - Categorical, nominal variables?
    - Answer :
  - Categorical, ordinal variables?
    - Answer :
- For categorical variables: gini index, entropy…

Scale measures for categorical variables

- Gini index and entropy can measure the scale or categorical variables
  - *I.e.* how much they're spread among their *K* modalities

- Gini index:

$$Gini = 1 - \sum_{k=1}^{K} P[k]^2$$

- If *Gini* is close to 0, then the values are concentrated on one modality
- If *Gini* tends to $1 - \frac{1}{K}$, then the attribute is uniformly spread amongst modalities

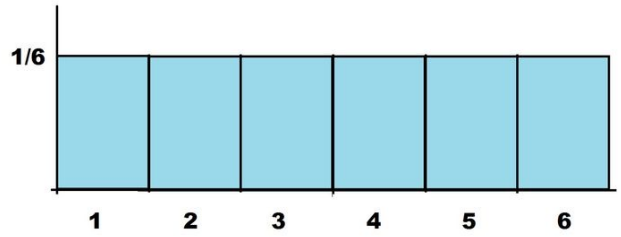
- Entropy:

$$Entropy = -\sum_{k=1}^{K} P[k] \ln(P[k])$$

- If *Entropy* is close to 0, then the values are concentrated on one modality
- If *Entropy* tends to *ln(K)*, then the attribute is uniformly spread amongst modalities

Exercises

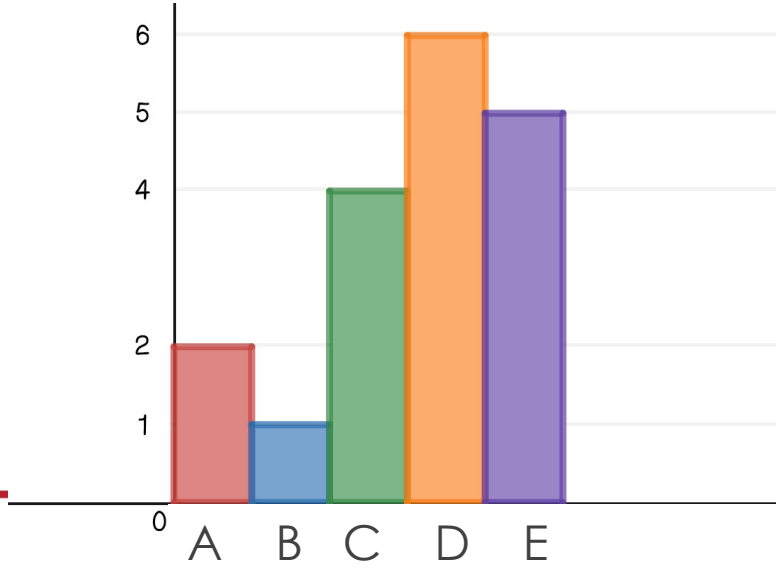
- How much are the Gini index and the entropy of a uniform distribution with 6 modalities? (*e.g.* throwing a dice)



  - Answer:

- How much are the Gini index and the entropy of this distribution?

  - Answer:



Univariate EDA measures

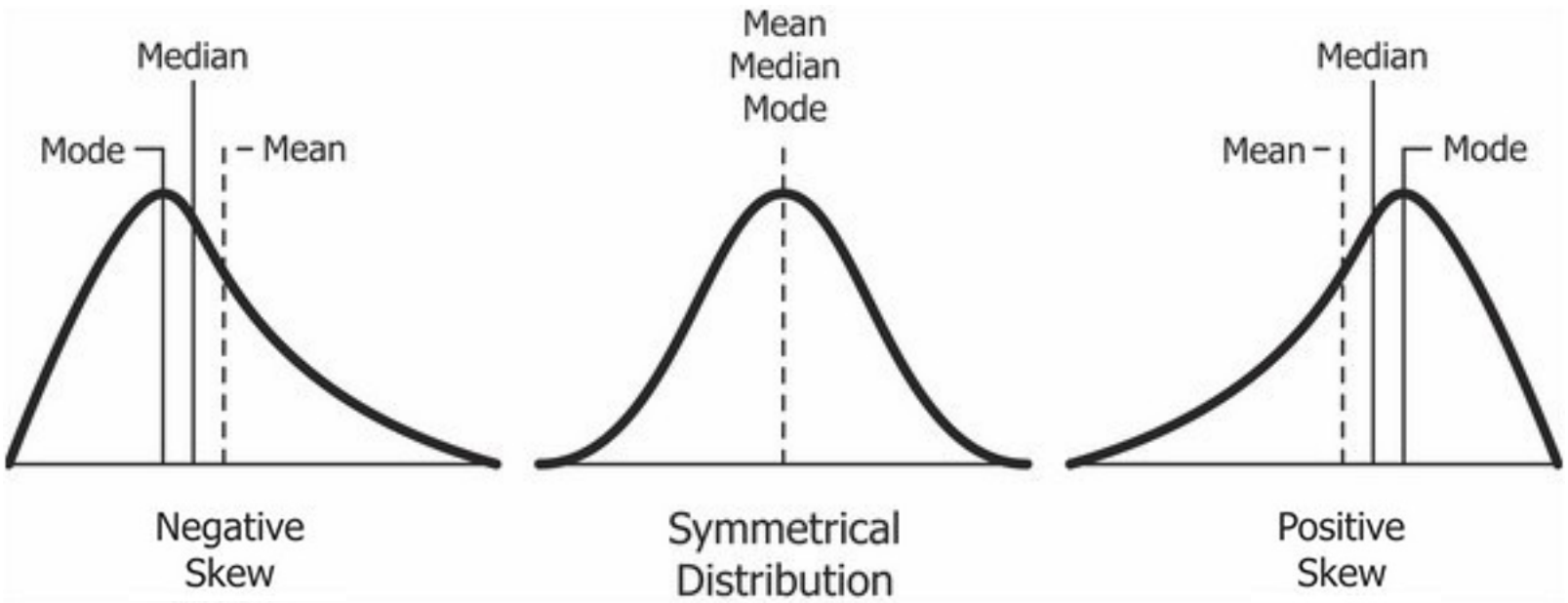
Shape of the distribution

Shape measures (for numeric variables)

- Fisher parameters characterize the shape of a distribution
  - Moment of order 1: mean $\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Moment of order 2: variance $\hat{\sigma}^2 = \frac{\sum_{i}^{n} (x_i - \bar{x})^2}{n}$
  - Fisher parameters=moments of order >2
    - Skewness: moment of order 3
      - Measures the **symetry / asymetry** of a distribution
    - Kurtosis: moment of order 4
      - Measures how **flat / peaky** a distribution is

Skewness

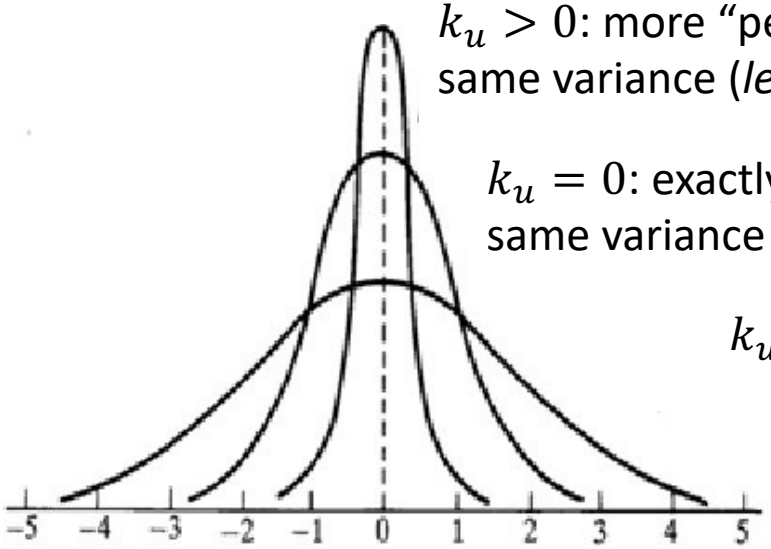
- Skewness: $s_k = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n\sigma^3}$



$s_k < 0$ $s_k = 0$ $s_k > 0$

Kurtosis

- Sample excess kurtosis: $k_u = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n\sigma^4} - 3$
  - Kurtosis measures peakedness / flatness of a distribution…
  - … relative to a Gaussian (normal) distribution



$k_u > 0$: more "peaky" than a Gaussian (normal) distribution with same variance (*leptokurtic* distribution)

$k_u = 0$: exactly as "peaky" as a Gaussian (normal) distribution with same variance (*mesokurtic* distribution)

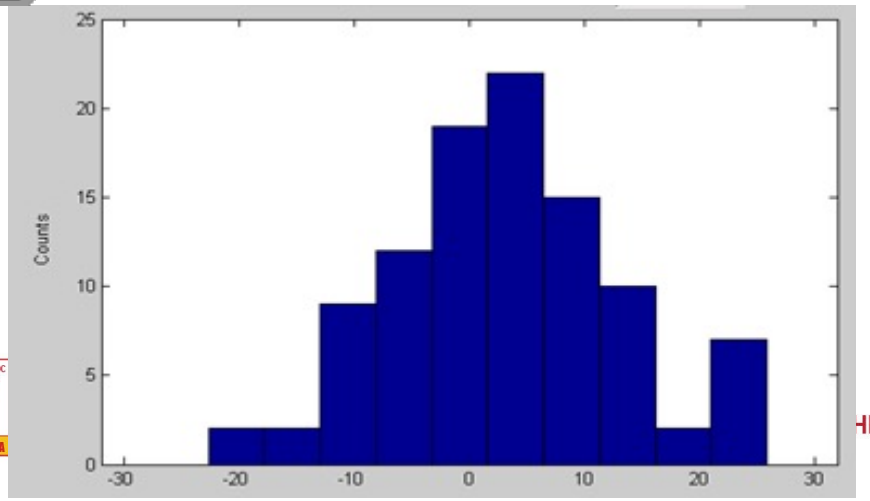
$k_u < 0$: more "flat" than a Gaussian (normal) distribution with same variance (*platykurtic* distribution)

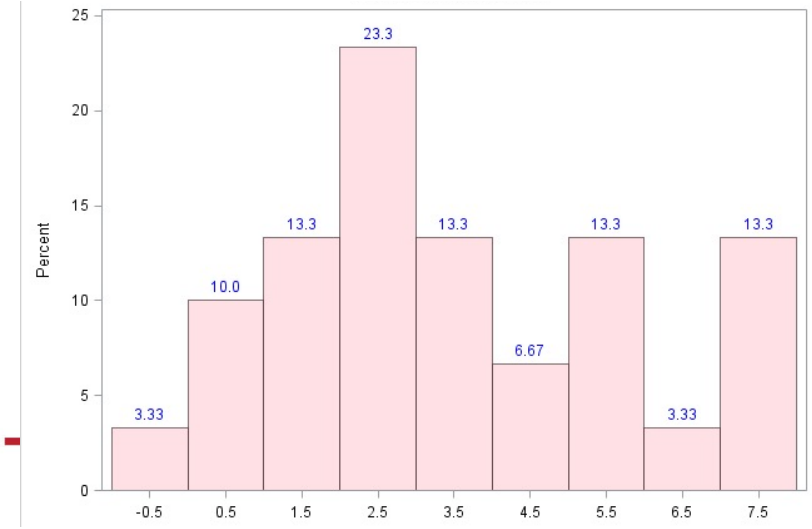
Univariate EDA graphics

Histogram plot

- An **histogram** graphically summarizes one **numeric** variable's distribution
  - The ranges used in the x-axis are called **bins**
- Histograms can be used to answer the following questions:
  - What kind of probability distribution do the data come from?
  - Where is the data located?
  - How spread out is the data?
  - Is the data symmetric or skewed?
  - Are there outliers in the data?
- Histograms can diplay either counts, or percentages

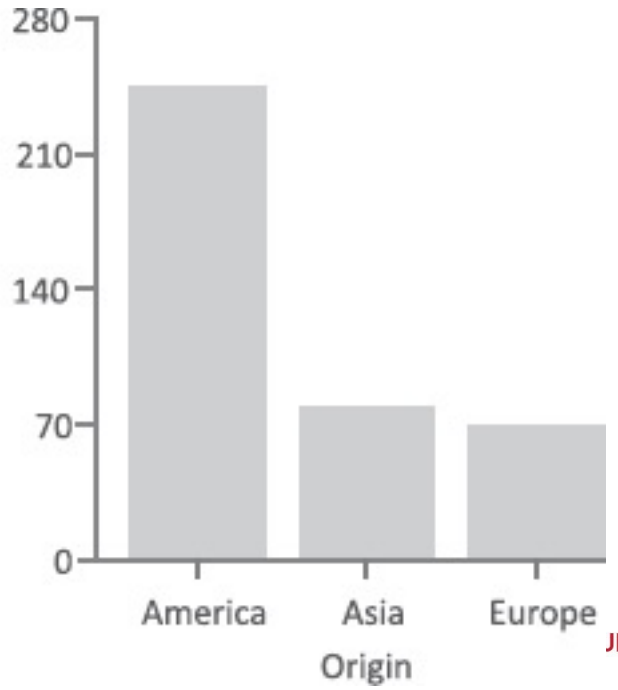
You **HAVE** to mention if it's count or percentage in the caption or y-axis

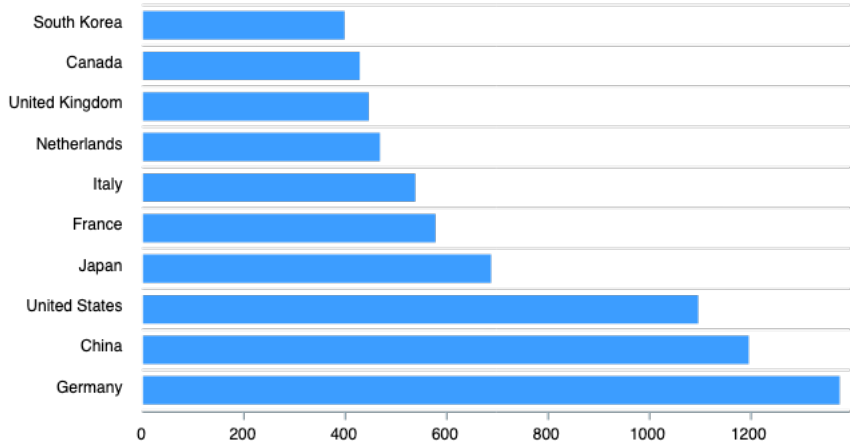




Bar charts

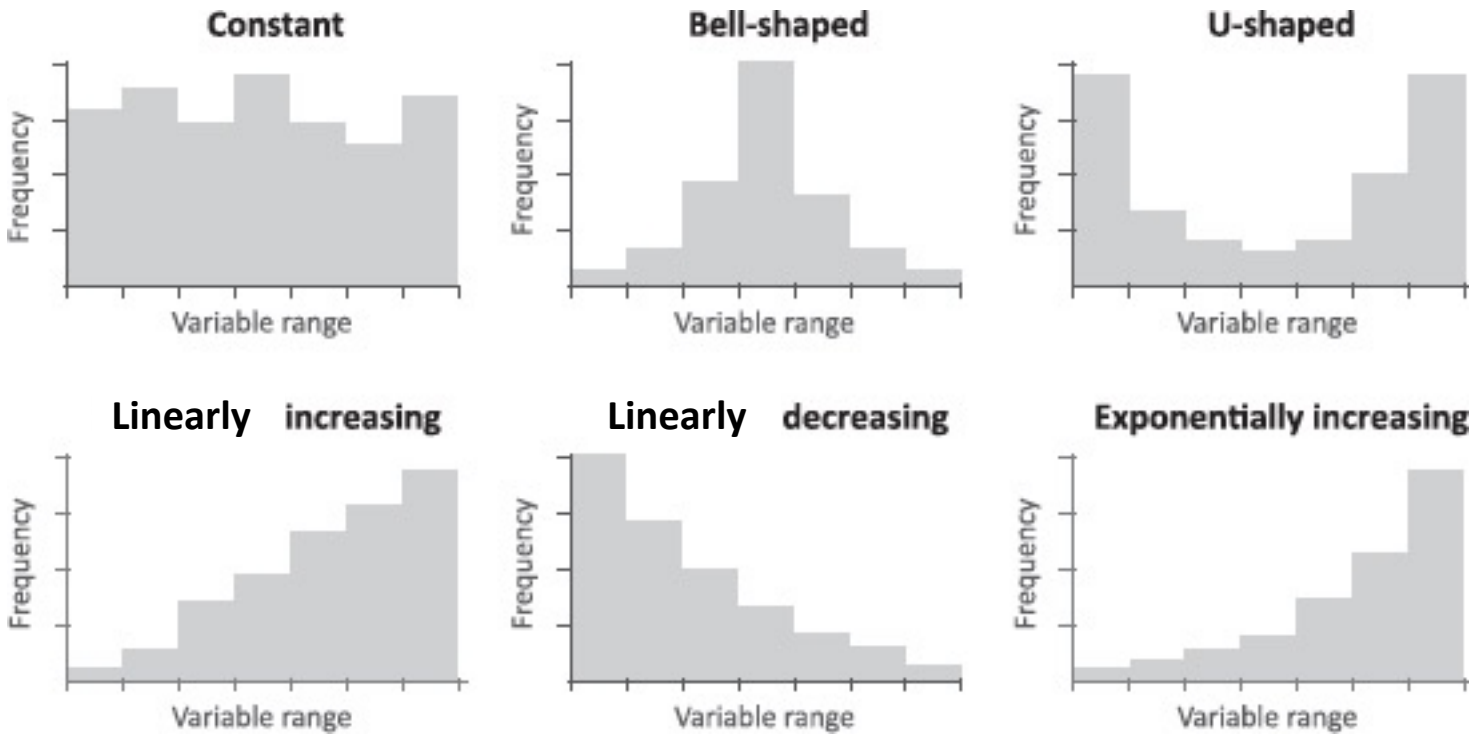
- An bar chart graphically summarizes one categorical variable's distribution
  - Similar to histograms, but for **categorical** variables
  - Instead of **bins**, bar charts use **modalities**





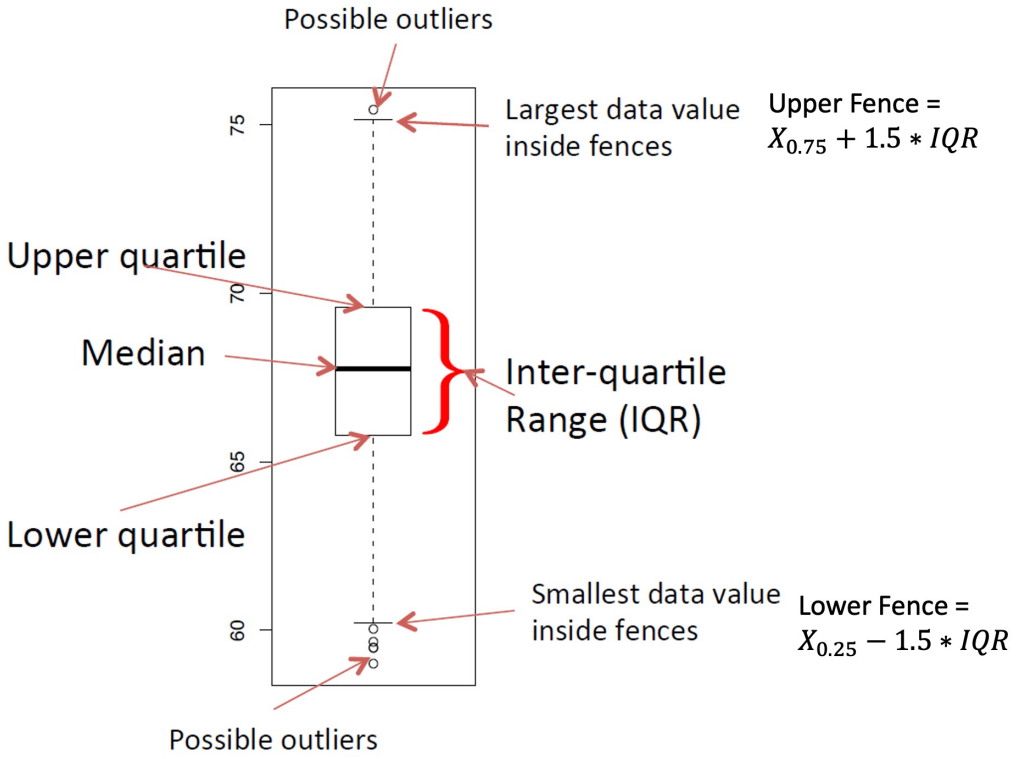
Example of distribution shapes

monotonous



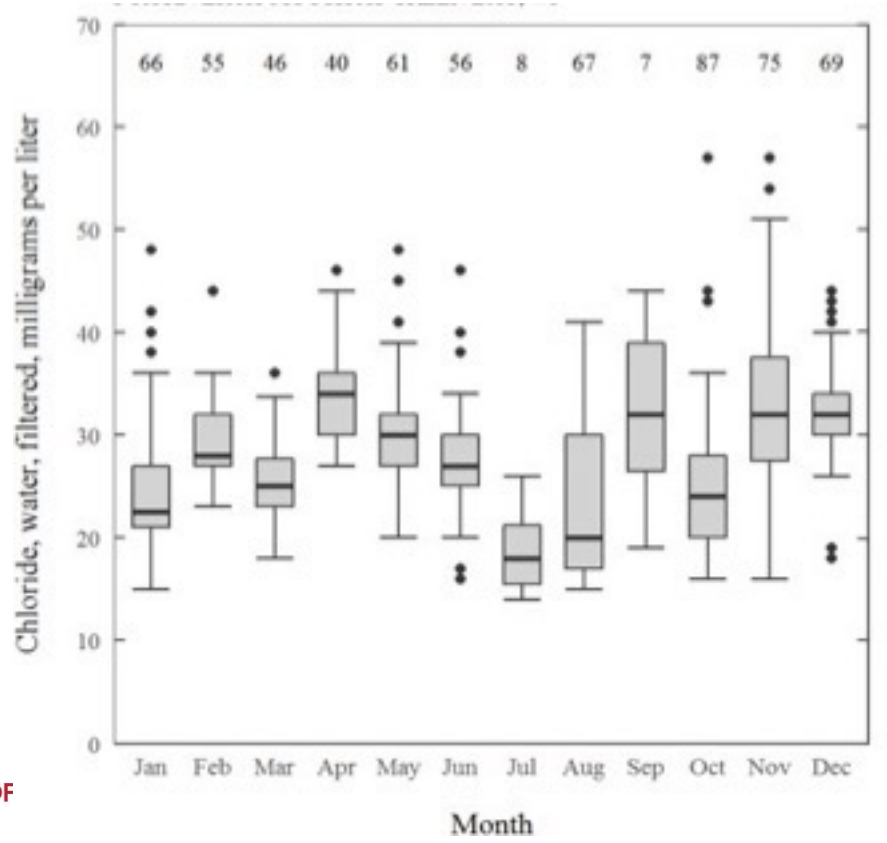
Box plot (*a.k.a.* whisker plot)

- By default in most tools/libraries, a box plot displays:
  - The lower fence value, the 25% quartile ($X_{0.25}$), the median ($X_{0.5}$), the 75% quartile ($X_{0.75}$), the upper fence value
  - Possible outliers (outside of the range between fences)



Box plot (cont'd)

- **Multiple boxplots** can provide answers to the following questions:
  - Does the central tendency differ between subgroups?
  - Does the scale differ between subgroups?
  - Are there any outliers?



Multivariate EDA **measures**

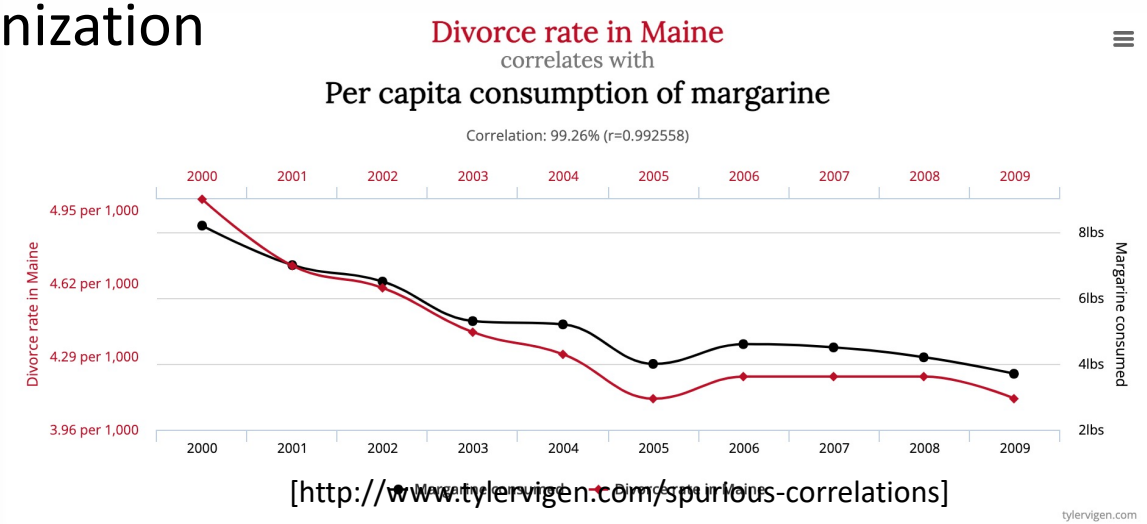
Identifying “links” (“relationships”) between multiple variables

Goal of multivariate EDA measures

- Mostly, multivariate EDA measures aim at discovering links / « relationahips » between multiple variables
  - Covariance / correlation (numeric / ordinal variables)
  - χ^2coefficient (nominal variables)

Link is **different** from causality (cause-effect)

- Link between 2 variables can be induced by other factors
  - For instance, two variables which happen to vary similarly with time/modernization



[http://www.tylervigen.com/spurious-correlations]

Covariance

- Covariance measures the linear link between numeric variables

- Covariance between 2 numeric variables *X* and *Y*

$$\sigma(X,Y)=\mathrm{cov}(X,Y)=\mathrm{cov}(Y,X) \qquad = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

- Covariance **matrix** between *p*>2 numeric variables

$$S = \begin{pmatrix} \sigma^2(X) & \sigma(X,Y) & \sigma(X,Z) \\ \sigma(X,Y) & \sigma^2(Y) & \sigma(Y,Z) \\ \sigma(X,Z) & \sigma(Y,Z) & \sigma^2(Z) \end{pmatrix}$$

Pearson's correlation

- Problem with covariance: it is not bounded
  - Difficult to say if it's high or low: depends on the range of the variables' values
- Hence, Pearson's correlation (linear correlation) coefficient
  - Bounded in [-1, 1] -> more often used than covariance
- Pearson's correlation between 2 numeric variables *X* and *Y*

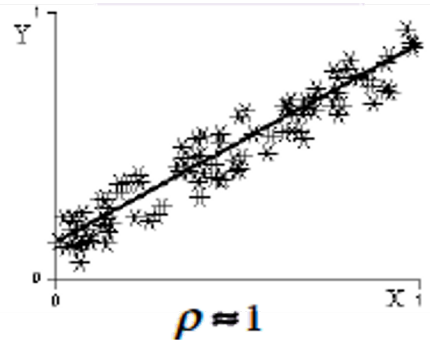
$$\rho(X,Y) = corr(X,Y) = corr(Y,X) = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i-\mu(X))(y_i-\mu(Y))}{\sigma(X)\sigma(Y)} = \frac{\sigma(X,Y)}{\sigma(X)\sigma(Y)}$$

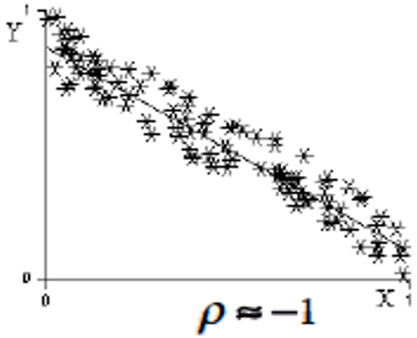
- Pearson's correlation **matrix** between *p*>2 numeric variables

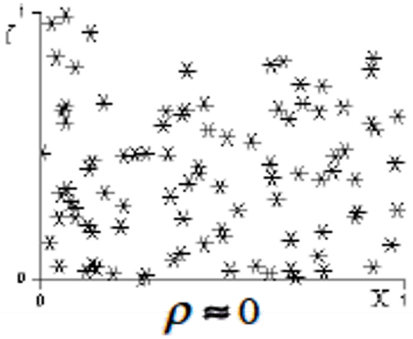
$$R = \begin{pmatrix} 1 & \rho(X,Y) & \rho(X,Z) \\ \rho(X,Y) & 1 & \rho(Y,Z) \\ \rho(X,Z) & \rho(Y,Z) & 1 \end{pmatrix}$$

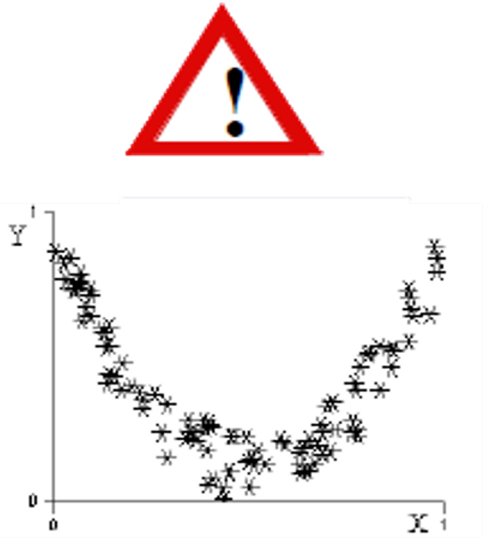
Pearson's correlation - limitations

- Pearson's correlation can **only** be applied to numeric variables
- Pearson's correlation can **only** detect linear links between variables
- Therefore, $\rho=0$ **does not imply** that the variables are statistically independent
  - Except for normal (Gaussian) variables
  - But statistically independent => $\rho=0$

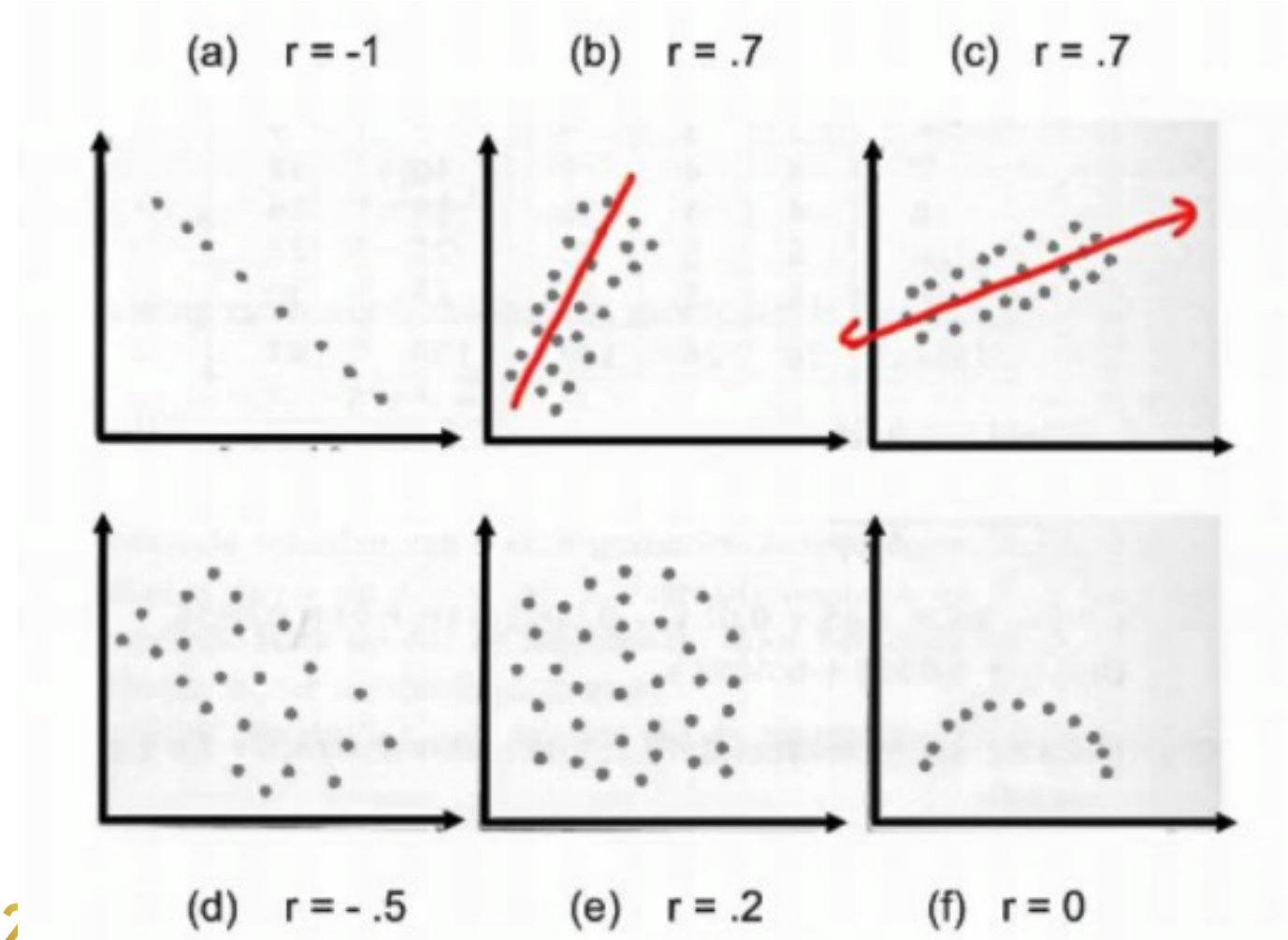






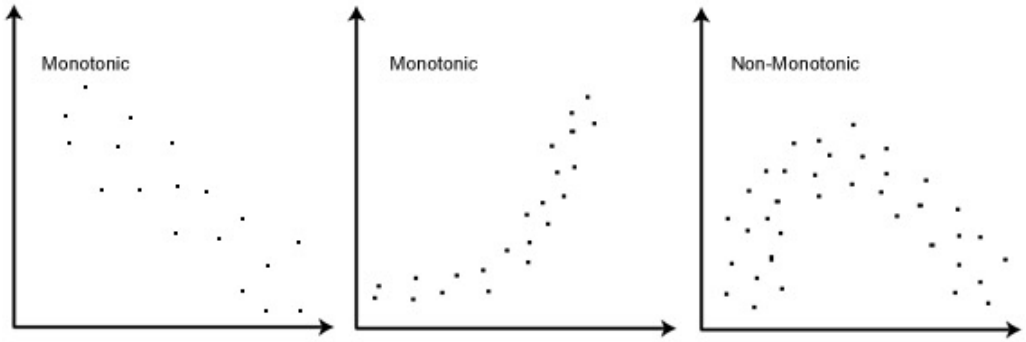


Some real examples



Spearman's correlation

- Spearman's correlation can be applied to numeric variables and to ordinal categorical variables
- Spearman's correlation
  - can detect monotonic links, whatever their shape
    - linear, exponential, ...



  - is based on the links between the **ranks** of observations

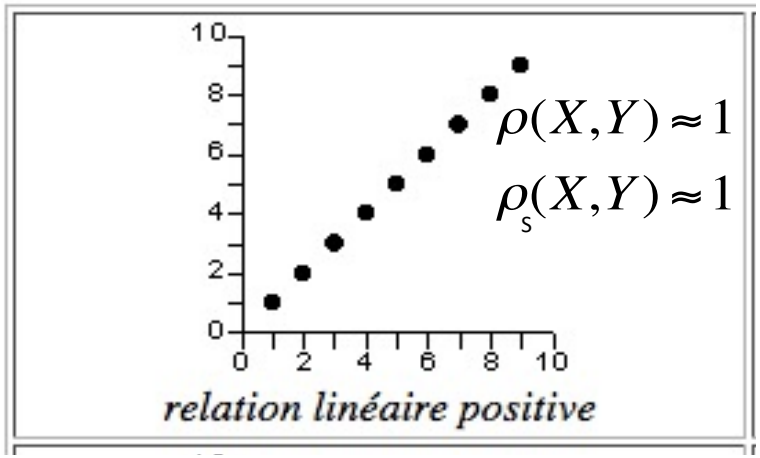
$$\rho_S(X,Y) = 1 - 6\frac{\sum_{i=1}^{N}(r(X_i) - r(Y_i))^2}{N^3 - N}$$

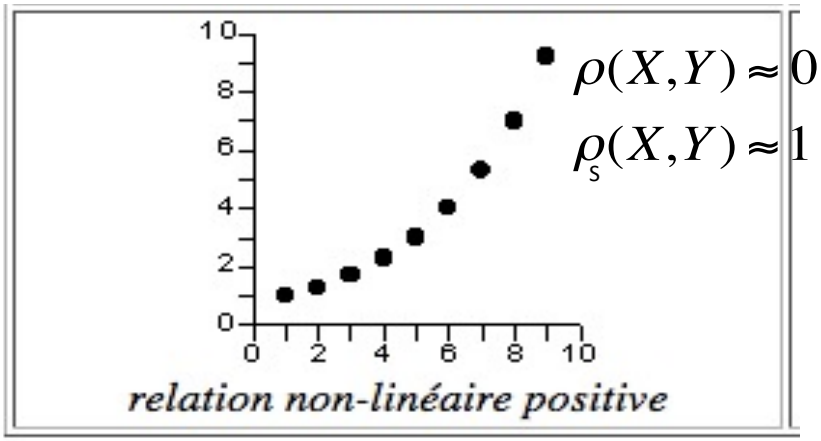
Where $r(X_i)$ is the **rank** of X_i in the distribution of $X_1, \dots X_n$ and $r(Y_i)$ is the **rank** of Y_i in the distribution of $Y_1, \dots Y_n$

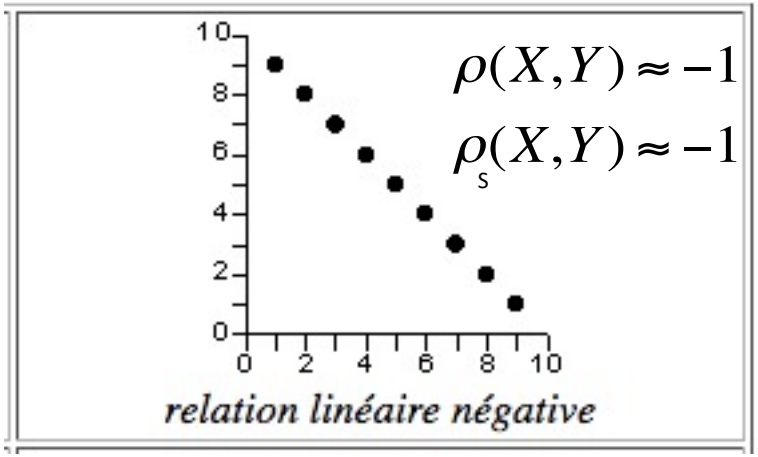
Spearman's correlation

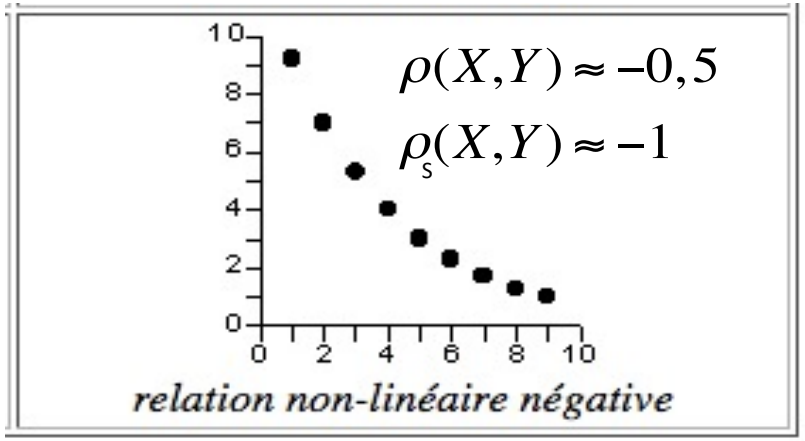
- Spearman's correlation is preferrable to Pearson's correlation if:
  - The variables are ordinal
  - The variables are numeric, but visibly non-Gaussian (*e.g.* asymmetric)
  - The link between the two variables is visibly non-linear
  - There are outliers in the distributions

Pearson *vs.* Spearman: example









Exercise 1

- Yesterday, I went to the market and bought 8 oranges
- I weighted them, wrote down their initial weight, then squizzed them for juice and weighted the waste
- So, I could determine, for each orange, the percentage of juice that each orange had
- I got the following observations:

Orange number	1	2	3	4	5	6	7	8
Orange weight	50	30	48	45	42	53	32	38
% of juice	80%	53%	70%	65%	58%	100%	53%	54.5%

Exercise 1 (cont'd)

- Draw roughly (on a piece of paper) the percentage of juice as a function of the (initial) orange weight
- By hand, calculate Pearson's correlation coefficient between the initial weight and the percentage of juice
- By hand, calculate Spearman's correlation coefficient between the initial weight and the percentage of juice
- What conclusions can you draw?

Orange number	1	2	3	4	5	6	7	8
Orange weight	50	30	48	45	42	53	32	38
% of juice	80%	53%	70%	65%	58%	100%	53%	54.5%

Exercise 1 solution

Exercise 2

- I did the same today, and I got the following observations
- Draw roughly (on a piece of paper) the percentage of juice as a function of the (initial) orange weight
- By hand, calculate Pearson's correlation coefficient between the initial weight and the percentage of juice
- By hand, calculate Spearman's correlation coefficient between the initial weight and the percentage of juice
- What conclusions can you draw?

orange number	1	2	3	4	5	6	7	8
orange weight	48	30	40	45	42	53	32	38
% of juice	73,00%	58,50%	55,00%	65,00%	58,00%	85,00%	53,00%	54,50%

Exercise 2 solution

Measuring the link between categorical variables

- When both variables are categorical and at least one is nominal, we can use the χ^2 coefficient
- Idea: measure the difference with the ***expected*** values (theoretical values, if the variables were independent)

	Like Scary Movies		
	Yes	No	Total
Girls	32	38	70
Men	30	12	42
Total	62	50	112
Percentage	55.4%	44.6%	

Question

- What are the expected values (if girls and men liked scary movies as much)?
  - N.B. The following table is called a contingency table

	Like Scary Movies		
	Yes	No	Total
Girls	32	38	70
Men	30	12	42
Total	62	50	112
Percentage	55.4%	44.6%	

Answer

- What are the expected values (if girls and men liked scary movies as much)?
  - N.B. The following table is called a contingency table

	Like Scary Movies		
	Yes	No	Total
Girls	32 *38.75*	38 *31.25*	70
Men	30 *23.25*	12 *18.75*	42
Total	62	50	112
Percentage	55.4%	44.6%	

Definition of the χ^2 coefficient

- Notations

Y / X	x_1	x_l	x_L	Σ
y_1				
y_k		n_{kl}		$n_{k.}$
y_K				
Σ		$n_{.l}$		n

- χ^2 coefficient

$$\chi^2(X,Y) = \sum_{k=1}^{K}\sum_{l=1}^{L} \frac{\left(n_{kl} - \frac{n_{k.}n_{.l}}{n}\right)^2}{\frac{n_{k.}n_{.l}}{n}}$$

Expected values

Quiz

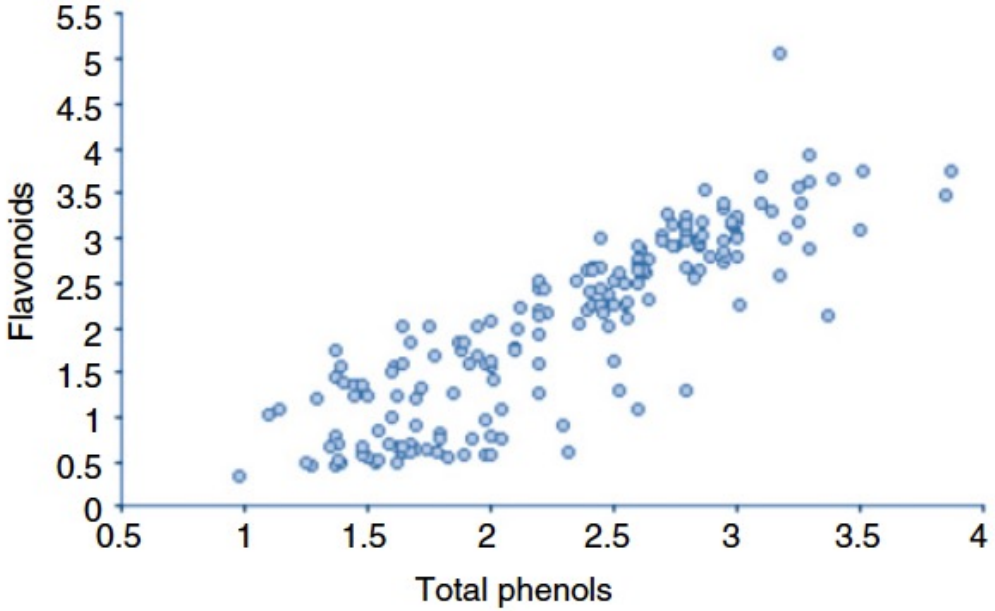
- If the χ^2 coefficient is almost 0, is there a link between the two variables, or are they not linked?
  - Answer:
- In the example of scary movies, how much is the χ^2 coefficient value?
  - Answer:
  - Homework: verify that you find the same number at home

- But, is this value **high enough** to say that liking scary movies is linked to gender, or not???
  - This is a tricky question actually, especially with few observations…
  - To answer this question, we need **hypothesis testing** (here a χ^2 test): out of the scope of this course…
    - https://en.wikipedia.org/wiki/Chi-squared_test
  - FYI: in this case, yes, liking scary movies is **significantly** linked to the gender…
  - Hypothesis testing is also required to determine if Pearson's / Spearman correlation is significant, or not
    - https://online.stat.psu.edu/stat501/lesson/1/1.9

Multivariate EDA **graphics**

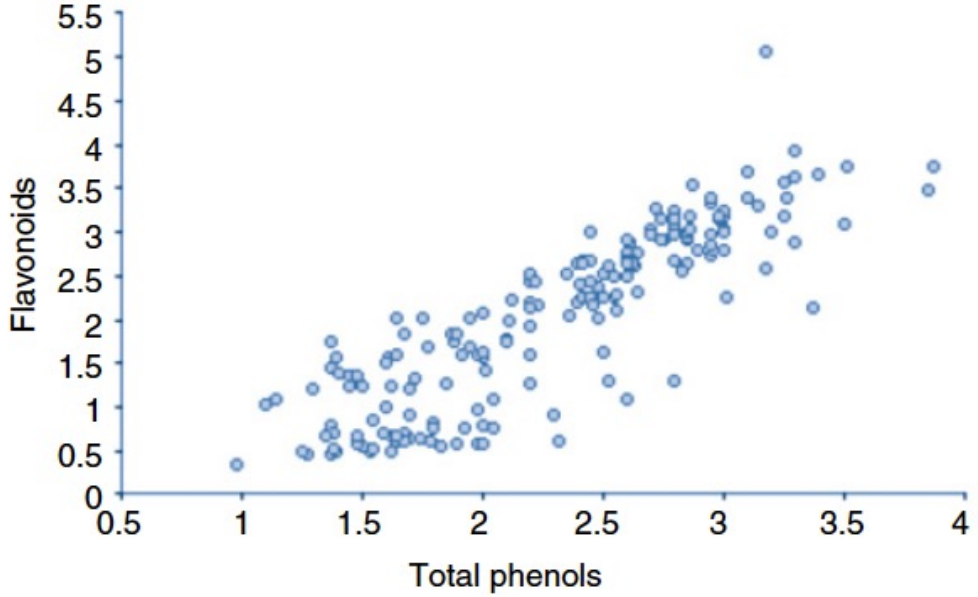
Scatter plot

- Objective: detect possible links between two **numeric, continuous** variables
  - two variables are plotted on the x-and y-axis
  - each point is a single observation.



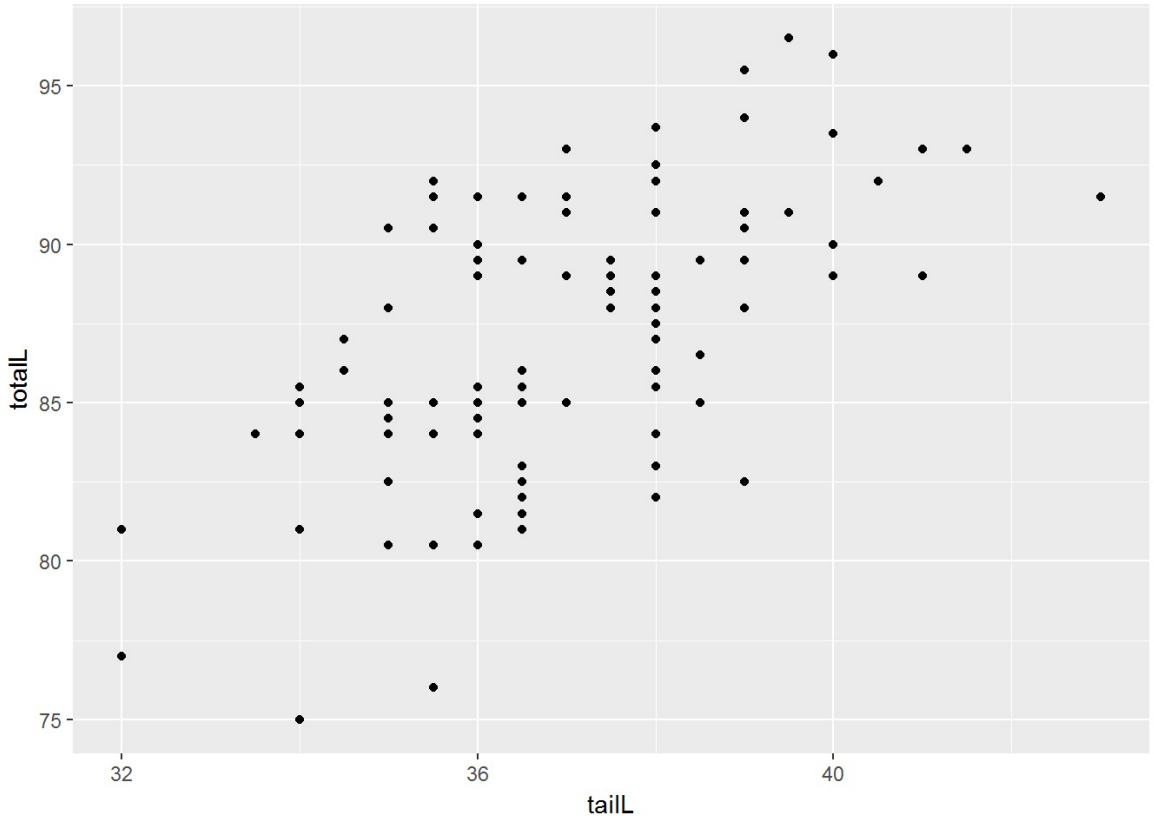
Scatter plot

- Scatter plots can provide answers to the following questions:
  - Are variables X and Y related?
  - Are variables X and Y linearly related?
  - Are variables X and Y non-linearly related?
  - Does the variation in Y change depending on X?
  - Are there outliers?



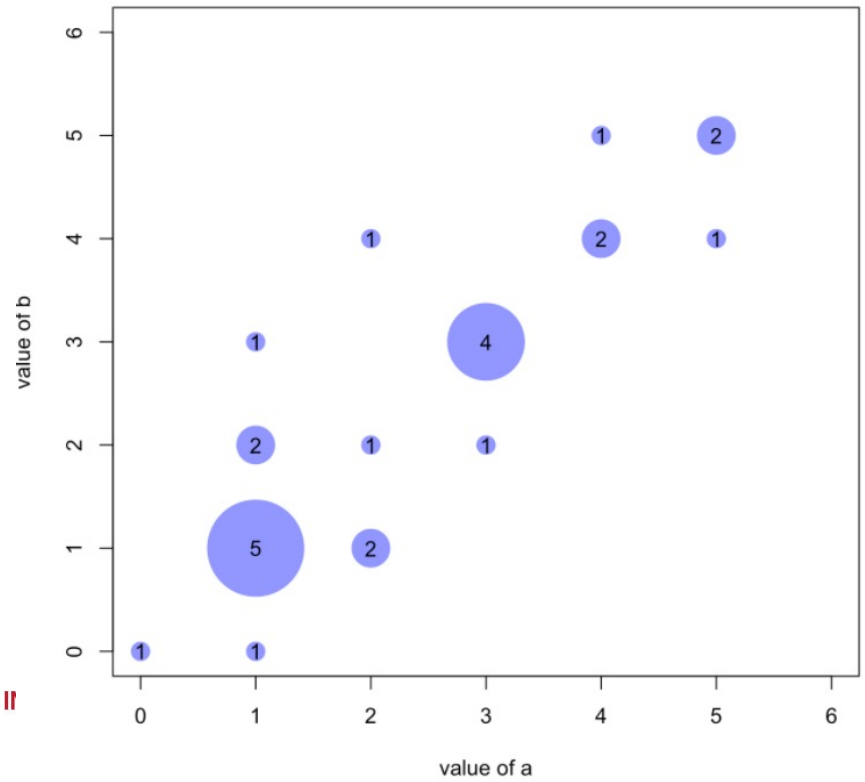
Scatter plot

- Scatter plots might also look like that
  - In particular, when one of the variables is **numeric**, but **discrete** (but not only)



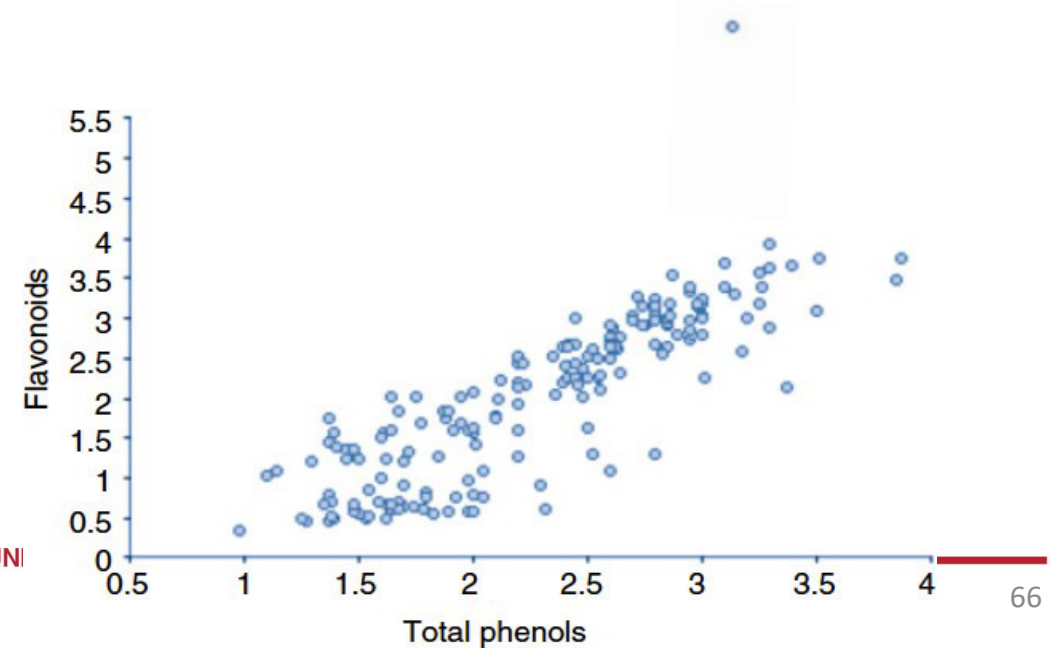
Scatter plot

- Scatter plot with **numeric, discrete** variables
  - Difficulty: if both variables are discrete, points might be cluttered
    - (overplotting)
  - Solution: Make bigger « circles » for the value sets with more data



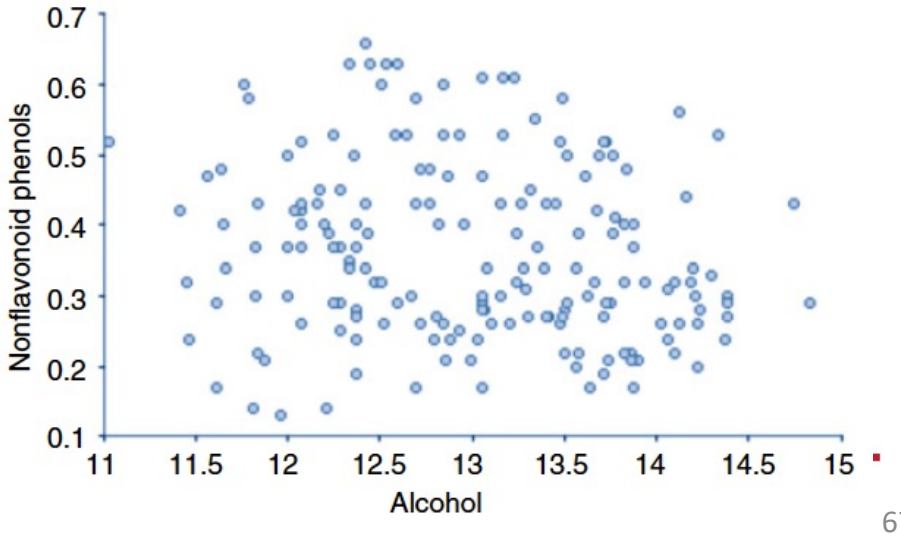
Quiz

- With this scatter plot, do you expect Pearson's coefficient to be: close to -1, close to 1, or close to 0?
  - Answer:
- With this scatter plot, do you expect Pearson's coefficient to be: close to -1, close to 1, or close to 0?
  - Answer:
- Are there any multivariate / univariate outlier(s)?
  - Answer:



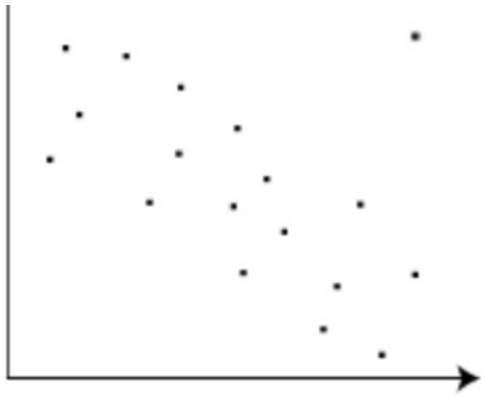
Quiz

- With this scatter plot, do you expect Pearson's coefficient to be: close to -1, close to 1, or close to 0?
  - Answer:
- With this scatter plot, do you expect Pearson's coefficient to be: close to -1, close to 1, or close to 0?
  - Answer:
- Are there any multivariate / univariate outlier(s)?
  - Answer:



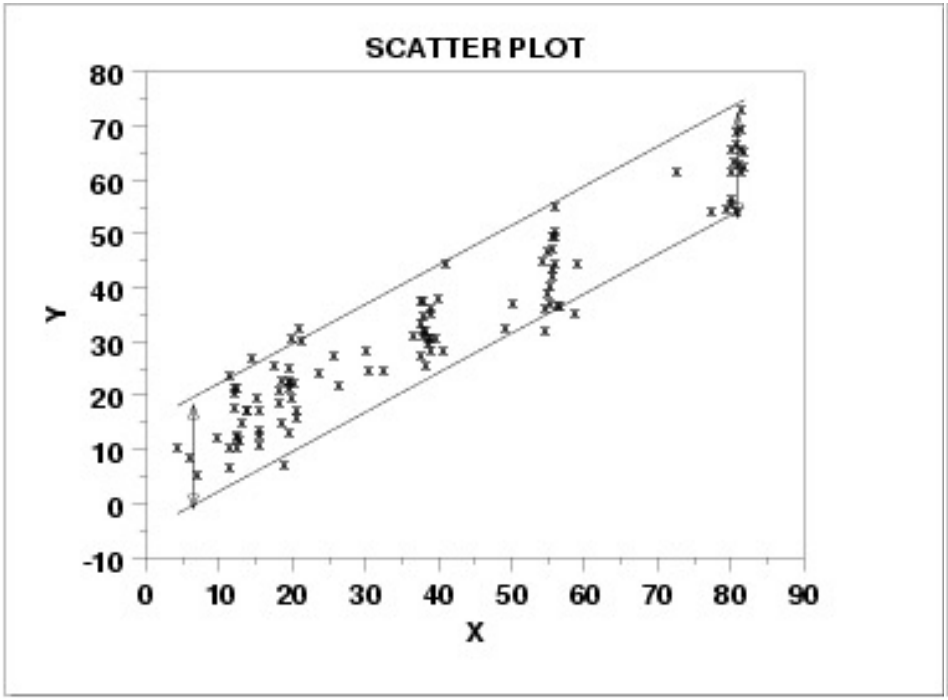
Quiz

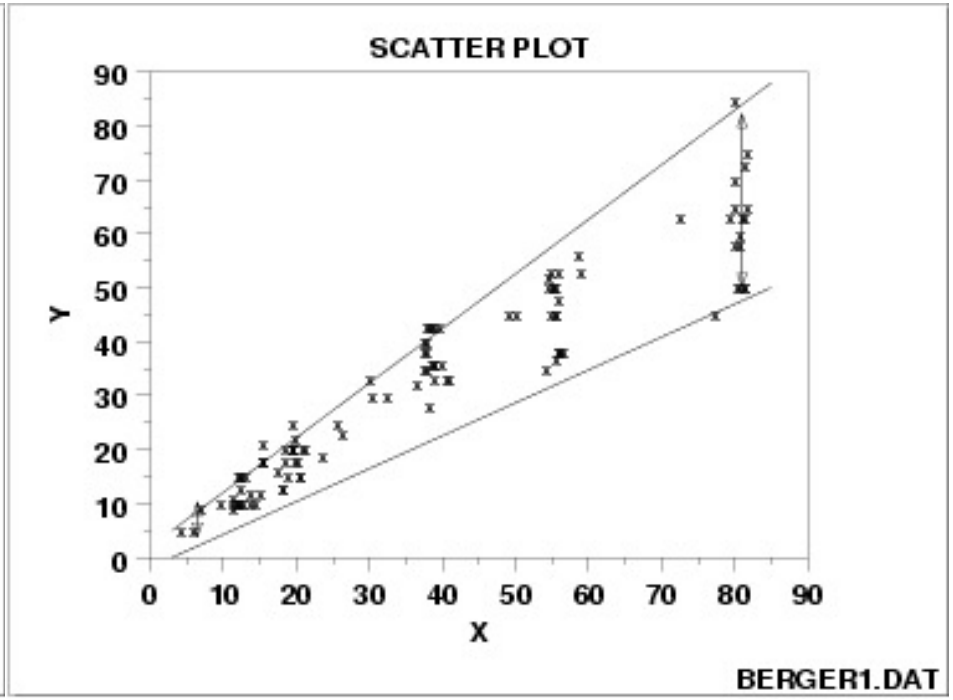
- With this scatter plot, do you expect Pearson's coefficient to be: close to -1, close to 1, or close to 0?
  - Answer:
- With this scatter plot, do you expect Pearson's coefficient to be: close to -1, close to 1, or close to 0?
  - Answer:
- Are there any multivariate / univariate outlier(s)?
  - Answer:



Scatter plot: variation of Y does not depend on X

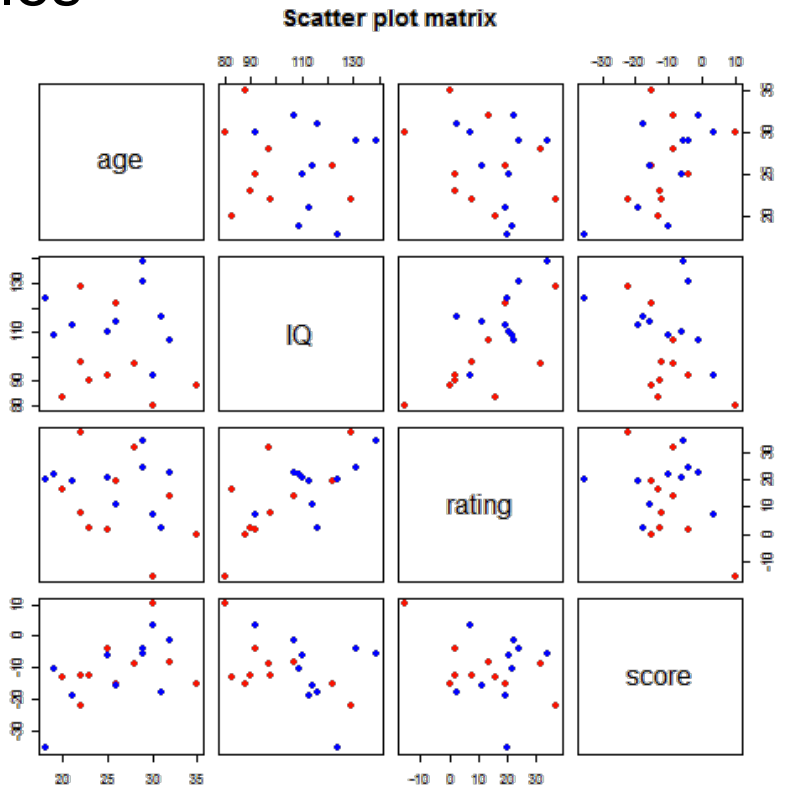
- Notion of homoscedasticity / heteroscedasticity
  - https://www.statisticshowto.com/homoscedasticity/





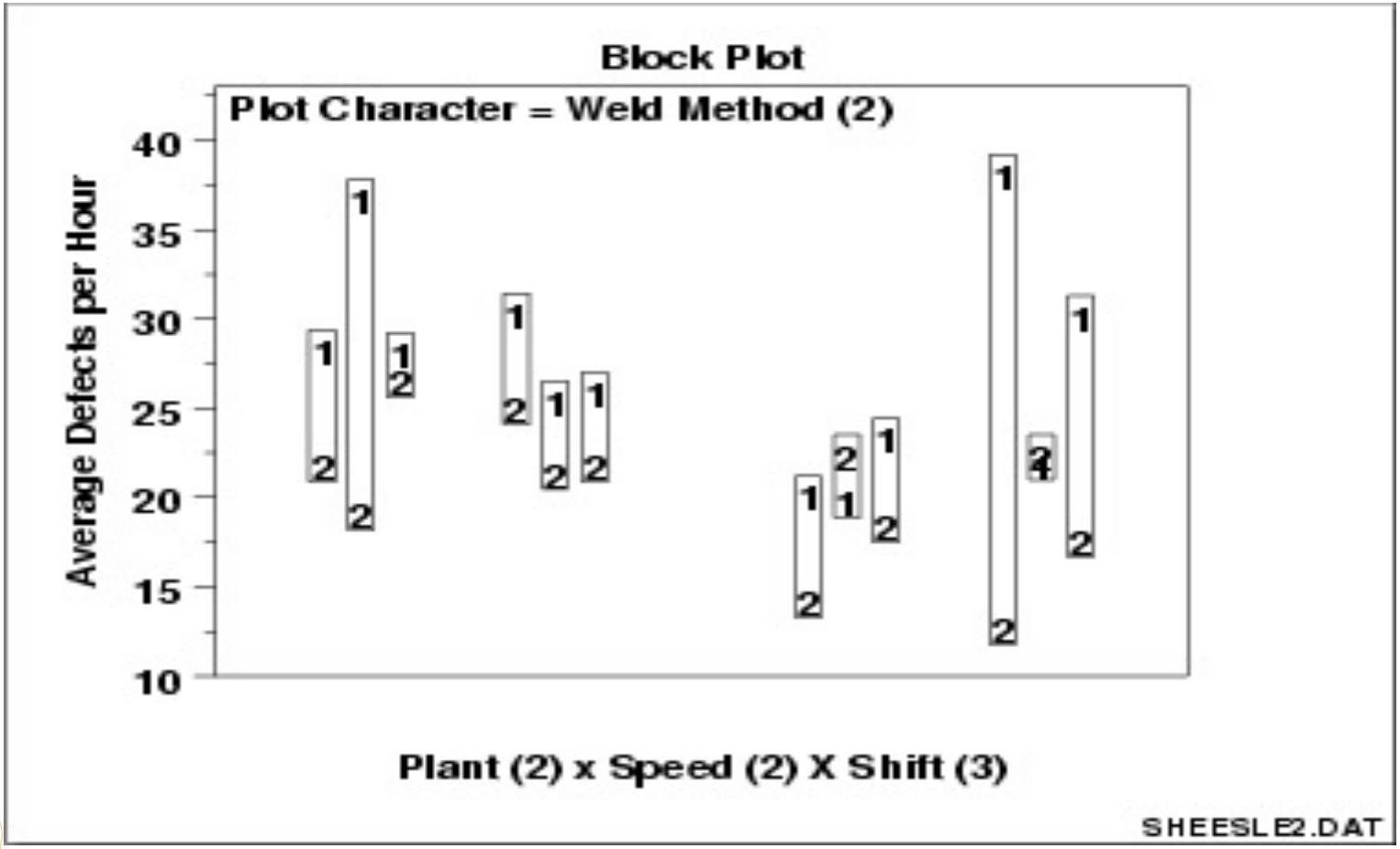
Scatterplot matrix

- Collection of **scatterplots** organized into a grid (or **matrix**)
- Each **scatterplot** shows the relationship between a pair of variables



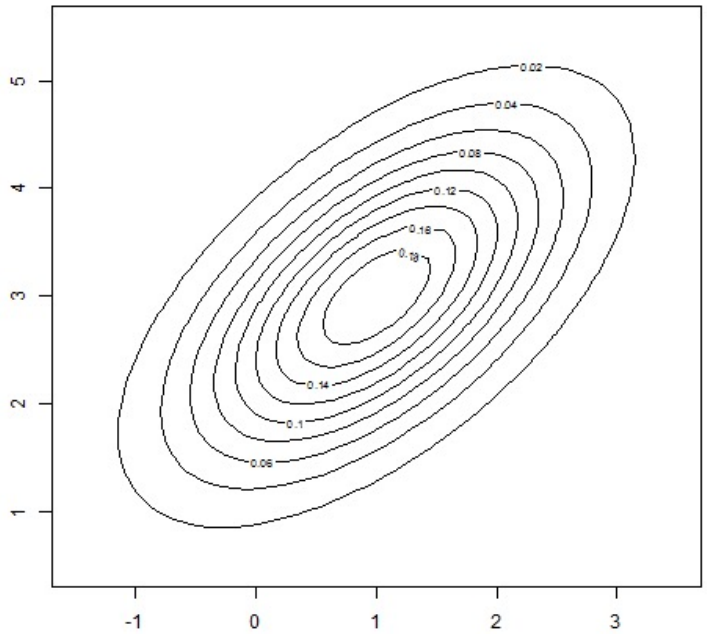
Block plot

- Useful to compare different groups of records using 2 categorical variables (with few modalities)



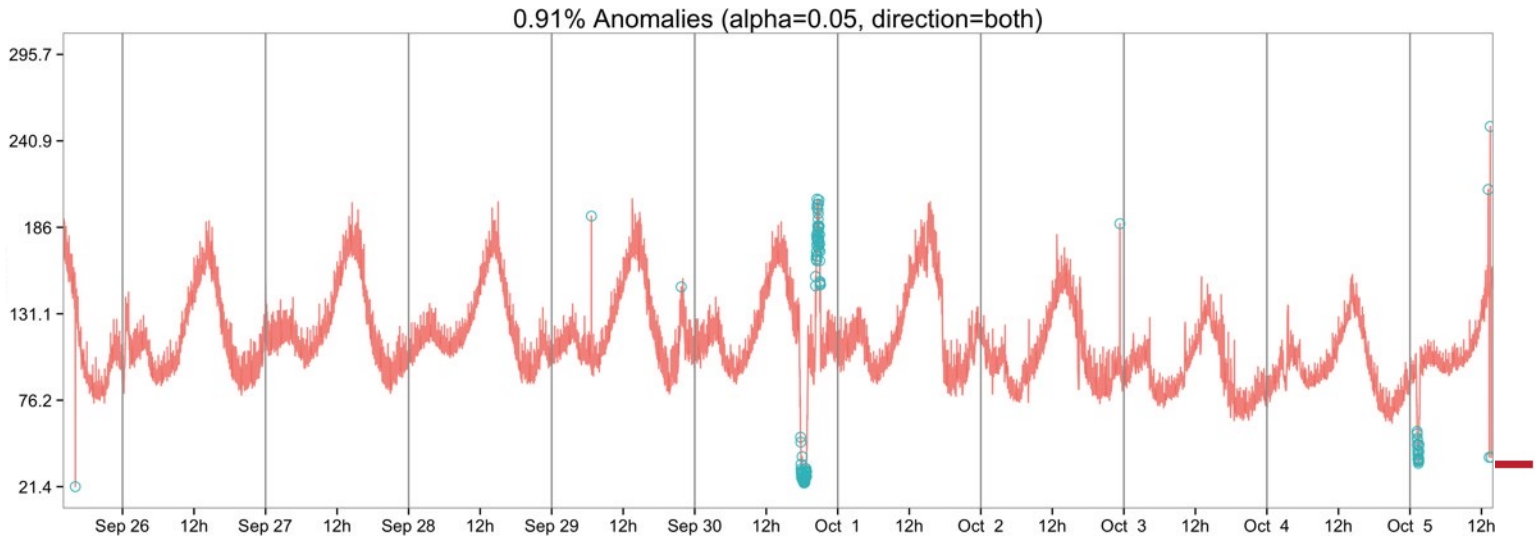
Contour plots

- Useful to visualize 3D data but on a two-dimensional plane
  - Contour lines indicate elevations that are the same
- The contour plot is used to answer the question
  - How does Z change as a function of X and Y?



Special case of time series

- Run sequence plot displays data in a time sequence
- Can be seen as a special kind of scatter plot
  - But, only one variable « of interest »: the variable on the y-axis
- The run sequence plot can be used to answer the following questions
  - Are there any shifts in central tendency (on the y-axis)?
  - Are there any shifts in variations(on the y-axis)?
  - Are there any outliers?



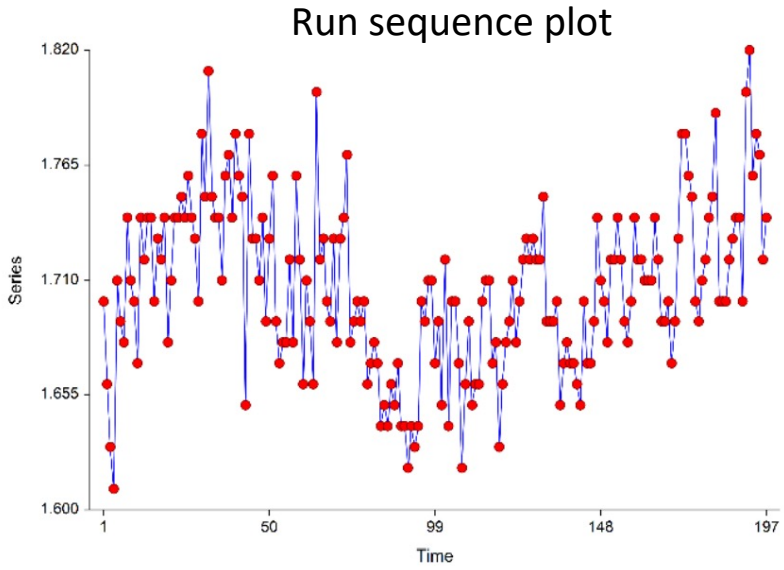
Special case of time series

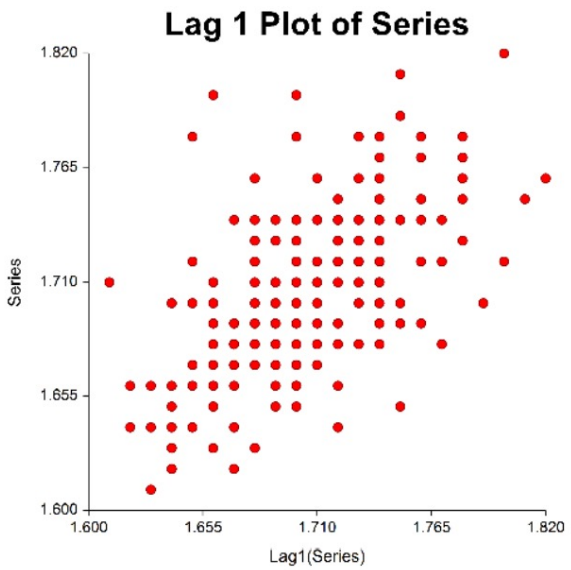
- **Autocorrelation** represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals.
  - For example, in places where the weather is quite stable:
    - If it's rainy today, then it's more likely to rain tomorrow also

- Autocorrelation is conceptually similar to the correlation between two different time series, but autocorrelation uses the same time series twice (1x original, 1x lagged)

Special case of time series

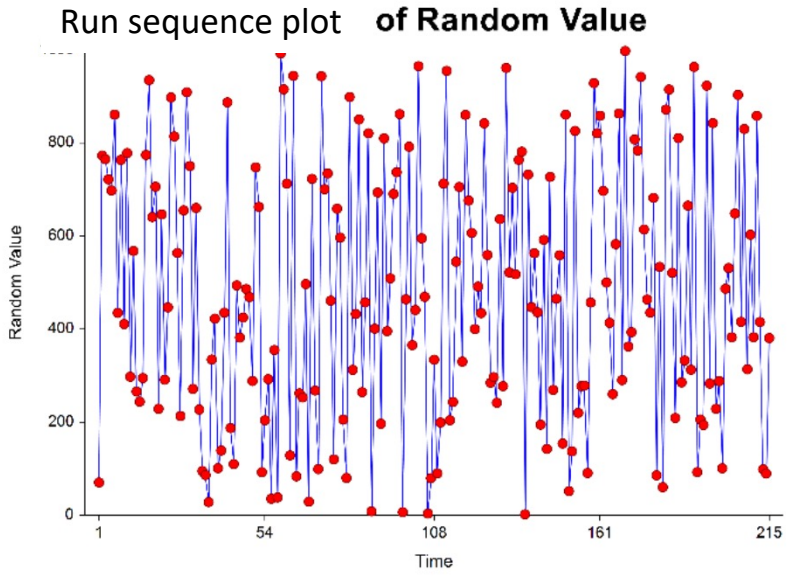
- To detect **auto-correlation** in a time series, one might use **lag plot**
  - The time interval (lag=k) is fixed for all data
    - If Y_i = value of the y-axis variable at time *I*, $\text{Lag}_k(Y_i) = Y_{i-k}$
    - For example, $\text{Lag}_1(Y_2) = Y_1$ and $\text{Lag}_3(Y_{10}) = Y_7$

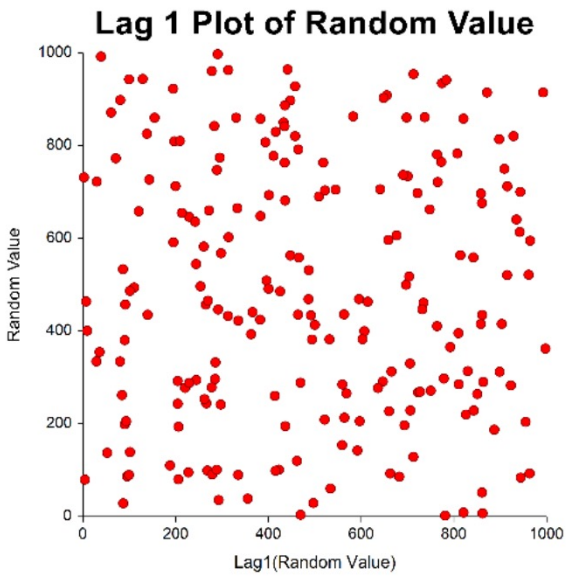




Example of lag plot

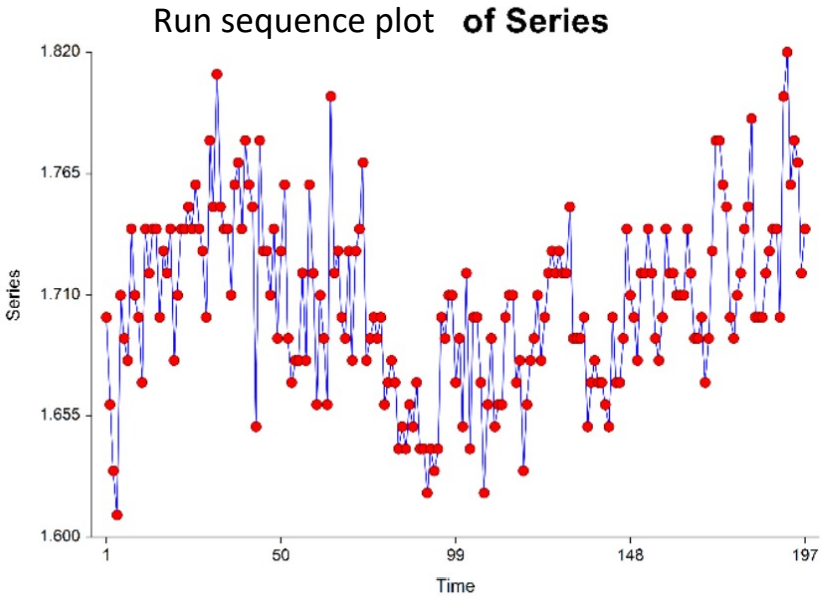
- Random Data (no autocorrelation)

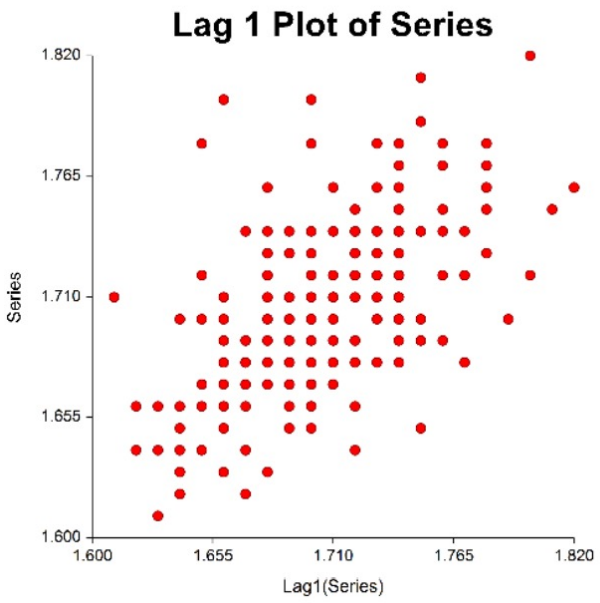




Example of lag plot

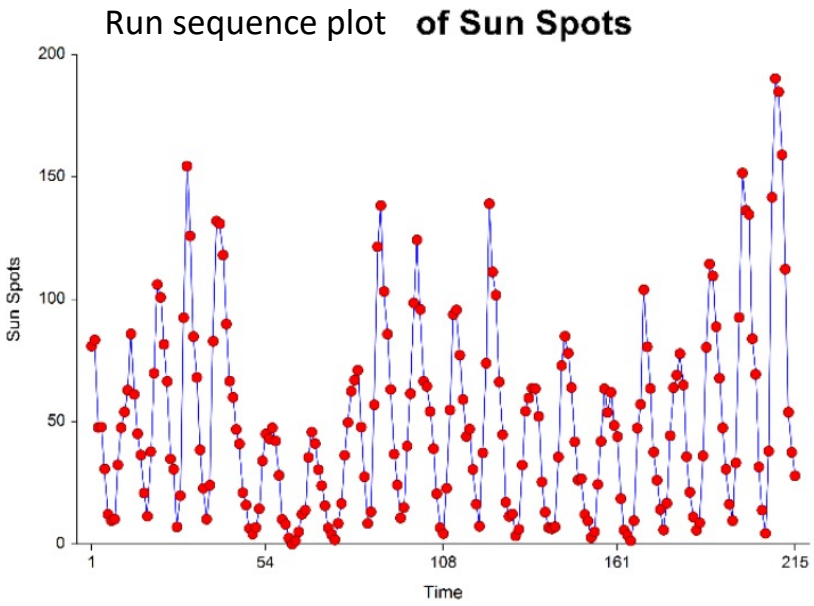
- Data with weak autocorrelation

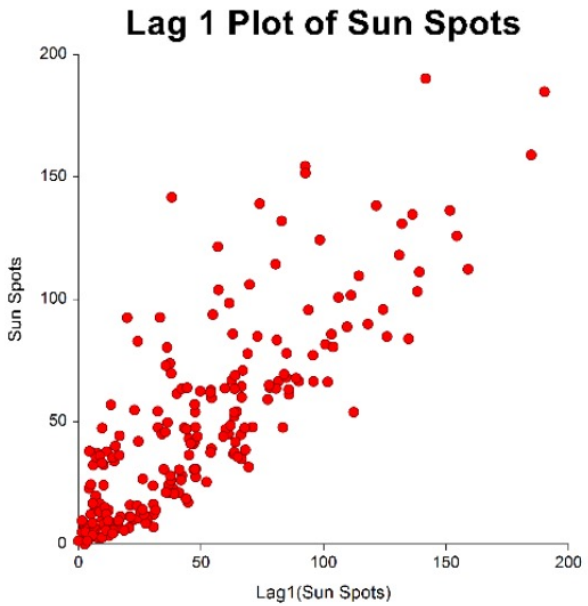




Example of lag plot

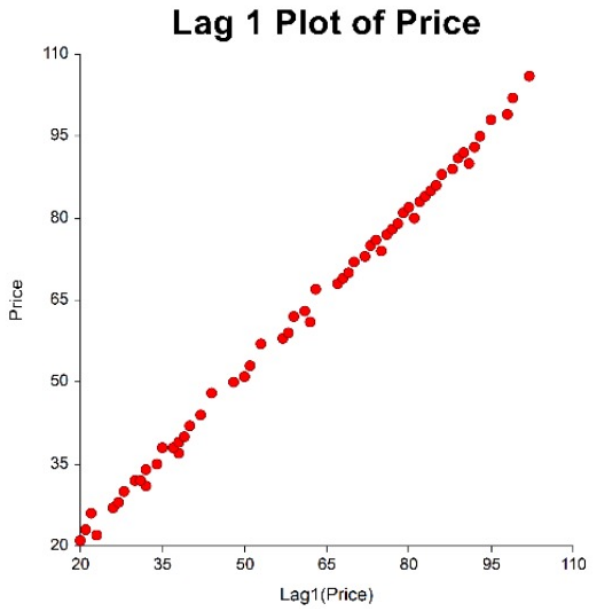
- Data with moderate autocorrelation





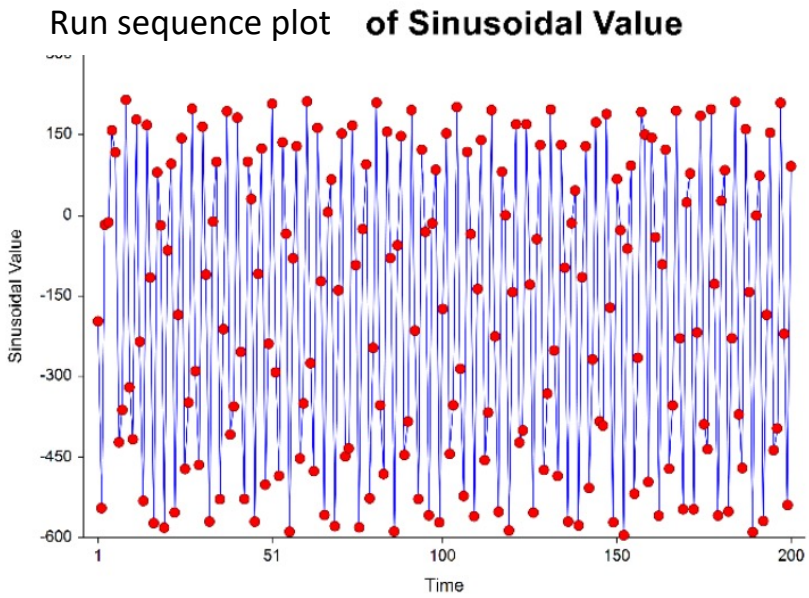
Example of lag plot

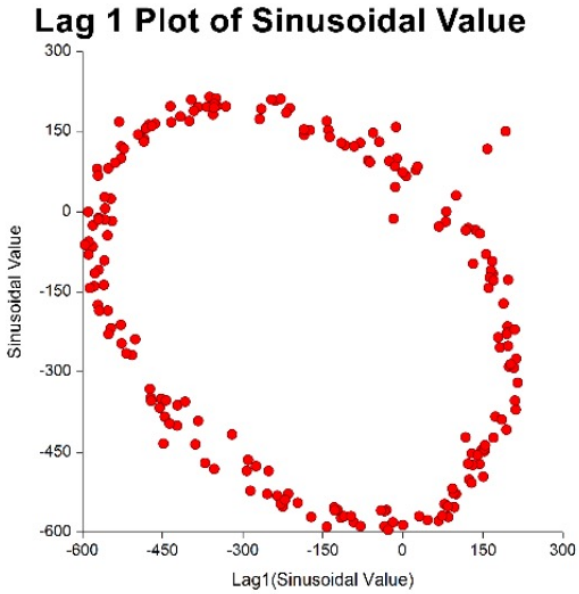
- Data with moderate autocorrelation



Example of lag plot

- Lag-plot of sinusoidal data



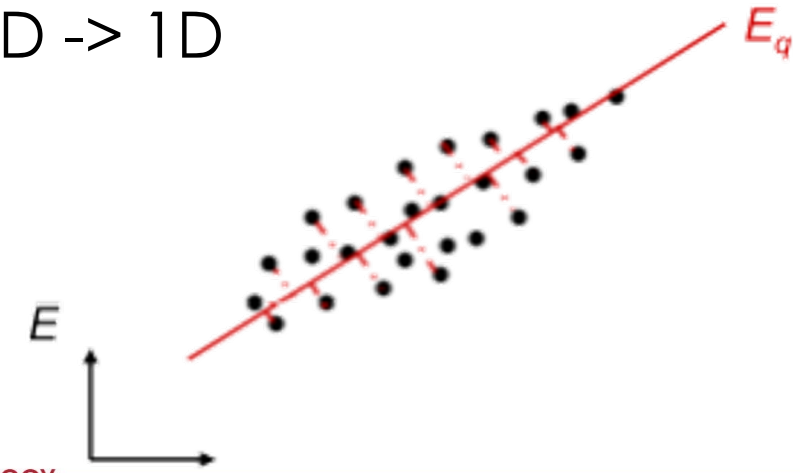


More advanced EDA techniques

Motivations

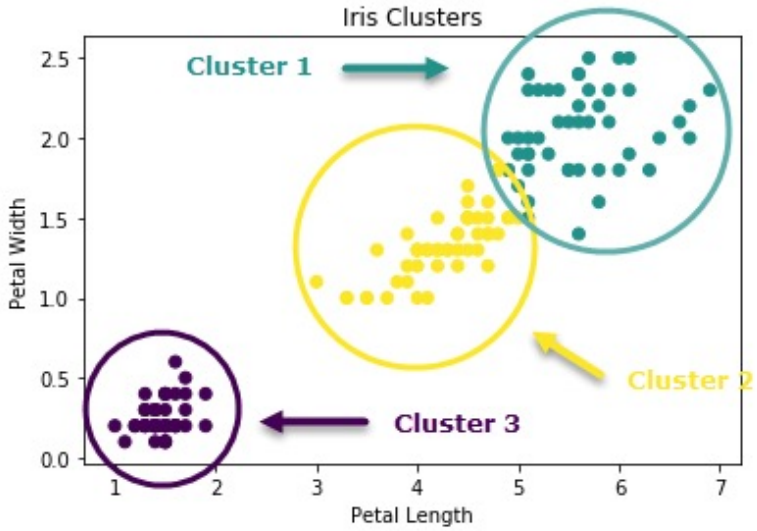
- Contour plots are fine with 3 variables, but what if we have 12 variables for instance?
  - Advanced **visualization** techniques (chapter 5)
  - **Dimensionality reduction** techniques (chapter 6)
    - *E.g.* Principal Component Analysis (**PCA**)
      - Finds the optimal hyperplane Eq of reduced dimensions that best represents the « shape » of the data (minimizes MSE)
      - Example: PCA for 2D -> 1D

MSE = $\sqrt{\frac{\sum_{i=1}^{n} ei^2}{n}}$; ei=distance between x_i in *E* and its projection on Eq



Motivations

- What if some scatter-plots show different *clusters* of data?
  - Might be interesting to observe in order to summarize the data (create customer groups for instance)
  - Clustering techniques are unsupervised Machine Learning techniques (Chapter 6)



Summary

Summary

- In this lecture, you've learned (1/2):
  - Univariate EDA **measures**
    - Central tendency (location)
      - Mean, median, mode
    - Scale (data spread)
      - Variance, standard deviation (for numeric variables)
      - Gini index, entropy (for categorical variables)
    - Shape of the distribution
      - Kurtosis, skewness (for numeric variables)
  - Univariate EDA **graphics**
    - Histogram, boxplot (for numeric variables)
    - Bar chart (for categorical variables)

Summary

- In this lecture, you've learned (2/2):
  - Multivariate EDA **measures**
    - Covariance and Pearson's coeffcient (for numeric variables)
    - Spearman coefficient (for numeric and categorical, ordinal variables)
    - Chi2 coefficient (for categorical variables)
  - Multivariate EDA **graphics**
    - Scatter plot and contour plots (for numeric variables)
    - Block plots (for categorical variables with few modalities)
    - Run sequence plot and lag plot (for time series)

Summary

- In this lecture, you've had an overview about:
  - More advanced EDA techniques
    - Visualization (see chapter 5)
    - Dimensionality reduction (see chapter 6)
    - Clustering (see chapter 6)

Homework

Homework (for next lesson)

- By hand:
  - If not finished yet, make the exercises on page 31 and 51-55
- Using a specific software (Tableau Public)
  - Register on Tableau Public
  - Do the following tutorial: https://help.tableau.com/current/guides/get-started-tutorial/en-us/get-started-tutorial-home.htm
  - Because this tutorial is based on Tableau Desktop (paying version) and not Tableau Public (free version), there are a few things that cannot be done, such as:
    - At the very end of "Step 3: Focus your results", File -> Save As
    - In "Step 5: Drill down into the details", some hierarchy-based actions are not possible to perform using the Public version
    - In "Step 8: Share your findings with others", do the section "Use Tableau Public" only

EDA: to be continued in the next chapters…

Questions

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attention!!!