WEEK 3: BASICS OF DATA ANALYTICS

Dr. Kai Li
School of Information Sciences
University of Tennessee, Knoxville
Spring 2025

Review of Week 2

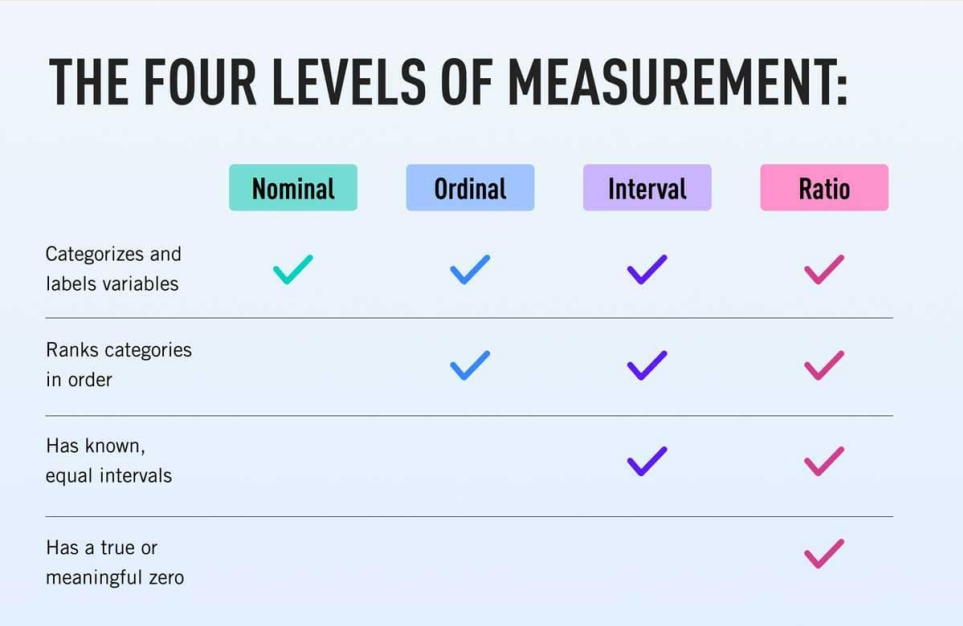
- Data science
- Data scientists

- Reading?
- Assignment?
- R exercise?

Overview of this week

- In this week, we will discuss the categories of data, questions, and statistical methods and their alignment.
  - *Data value types*
  - *Research questions*
  - *Types of statistical analysis*
    - We will come back to this topic from Week 6
  - *Demonstration:*
    - Project organization

Data value types



Data type	Mathematical operations	Measures of central tendency	Measures of variability
Nominal	• Equality (=, ≠)	• Mode	• None
Ordinal	• Equality (=, ≠) • Comparison (>, <)	• Mode • Median	• Range • Interquartile range
Interval	• Equality (=, ≠) • Comparison (>, <) • Addition, subtraction (+,−)	• Mode • Median • Arithmetic mean	• Range • Interquartile range • Standard deviation • Variance
Ratio	• Equality (=, ≠) • Comparison (>, <) • Addition, subtraction (+,−) • Multiplication, division (×, ÷)	• Mode • Median • Arithmetic mean • *Geometric mean	• Range • Interquartile range • Standard deviation • Variance • **Relative standard deviation

Data value types

- Nominal:
  - *They are names and cannot be calculated as themselves.*
  - *It is always a categorical / discrete variable.*
- Ordinal:
  - *Ranking data, such as 1st, 2nd, et al.*
- Interval/Ratio:
  - *They are continuous variables and can be calculated.*

Data value types

- How would you classify these variables into nominal, ordinal, interval, and ratio data?
  - *Color (red, yellow, blue…)*
  - *Ranking (1^{st}, 7^{th}…)*
  - *Temperature (75 degree Celsius…)*
  - *Time (1975/1/1)*
  - *Count of people*

Data value types

- How would you classify these variables into nominal, ordinal, interval, and ratio data?
  - *Color (red, yellow, blue…)* *Nominal*
  - *Ranking (1st, 7th…)* *Ordinal*
  - *Temperature (75 degree Celsius…)* *Interval*
  - *Time (1975/1/1)* *Interval (but not the difference of time, like 2 hours!)*
  - *Count of people* *Ratio*

Likert scales

- Likert scale is believed to be ordinal data, but not interval.
- However, it can be treated as interval data in statistical analyses (by calculating the mean value).

How satisfied are you with our services?

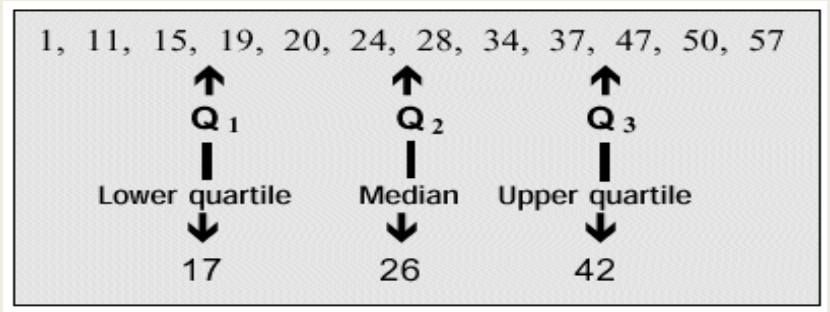
- Very Unsatisfied – 1
- Unsatisfied – 2
- Neutral – 3
- Satisfied – 4
- Very Satisfied – 5

[redacted]

No	Item	Mean	Std. Deviation	Rank	Level of influence
1	I feel valued in my role	2.36	1.47	1	Moderate
2	I'm clear about what is expected of me at work	2.35	1.34	2	Moderate
3	My skills are well-used	2.35	1.40	3	Moderate
Overall		**2.35**	**0.07**		**Moderate**

[redacted]

Transformation of Data value types



- We can transform one type of data value into another:
  - *Grouping interval/ratio data and then get the ranking of values*
    - Instead of getting 75 points, I am in the first quarter of the class in an exam.
  - *In many statistical models, we may want to consider time as a categorical data point (nominal).*

Statistical methods

- A set of methods to describe, interpret, and analyze data.
- There are two major categories of statistical methods:
  - *Descriptive methods: focusing on describing a variable*
  - *Inferential methods: focusing on making inferences and predictions*

Types of Questions

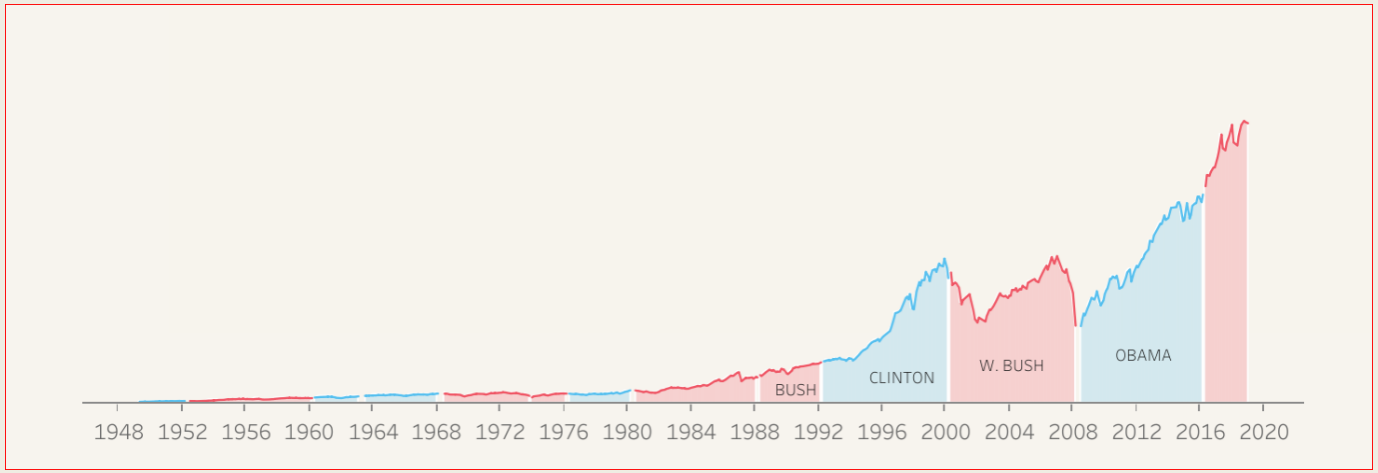
- Descriptive questions:
  - *How old are the students?*
  - *Where are the students from?*
- Inferential questions:
  - *Is the student's gender related to one's academic performance?*
  - *How much the performance will be improved if the university is only using small classes?*

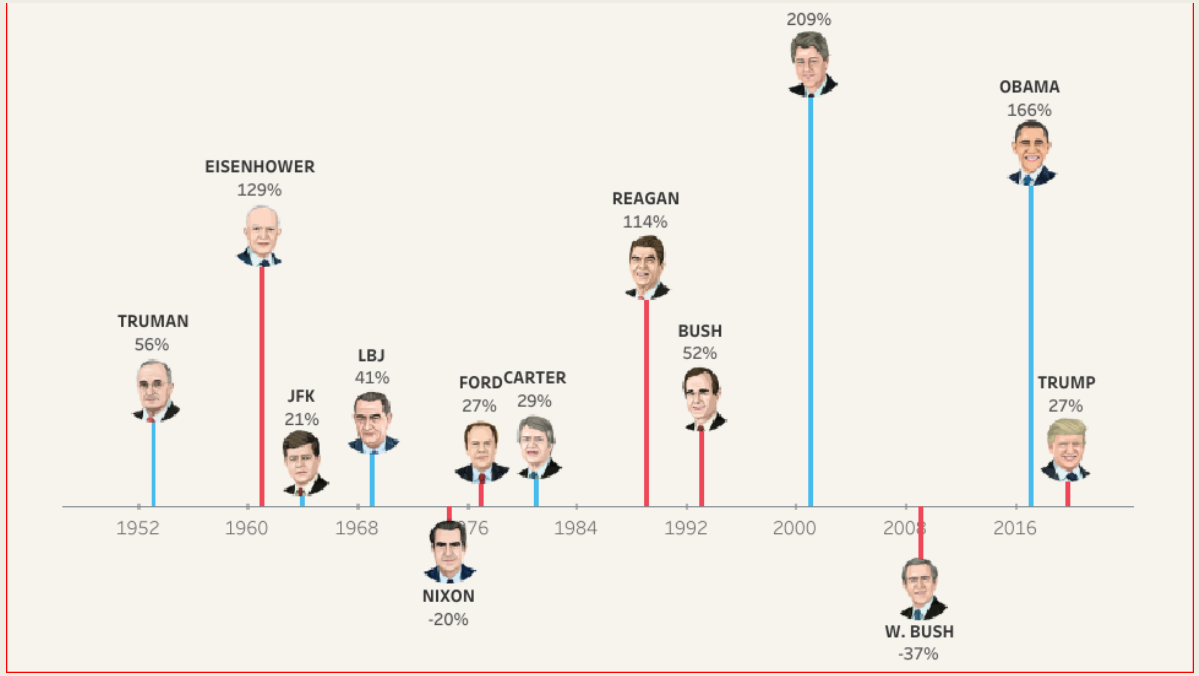
One more comment

- No one type of questions can “rule them all.”
  - *Each type of question is just one perspective we can take on a topic.*
- In a research paper and report, we may want to use multiple questions to tell a more comprehensive story.
- We need to make sure that our analysis is corresponding to the type of questions we ask.

Stock Market as an Indicator of Economy

- S&P 500 overall growth (1948-2019) – by Spencer Baucke
- https://public.tableau.com/profile/spencer.baucke#!/vizhome/IstheTrumpStockMarkettheBestinUSHistory_15710191293610/Analysis



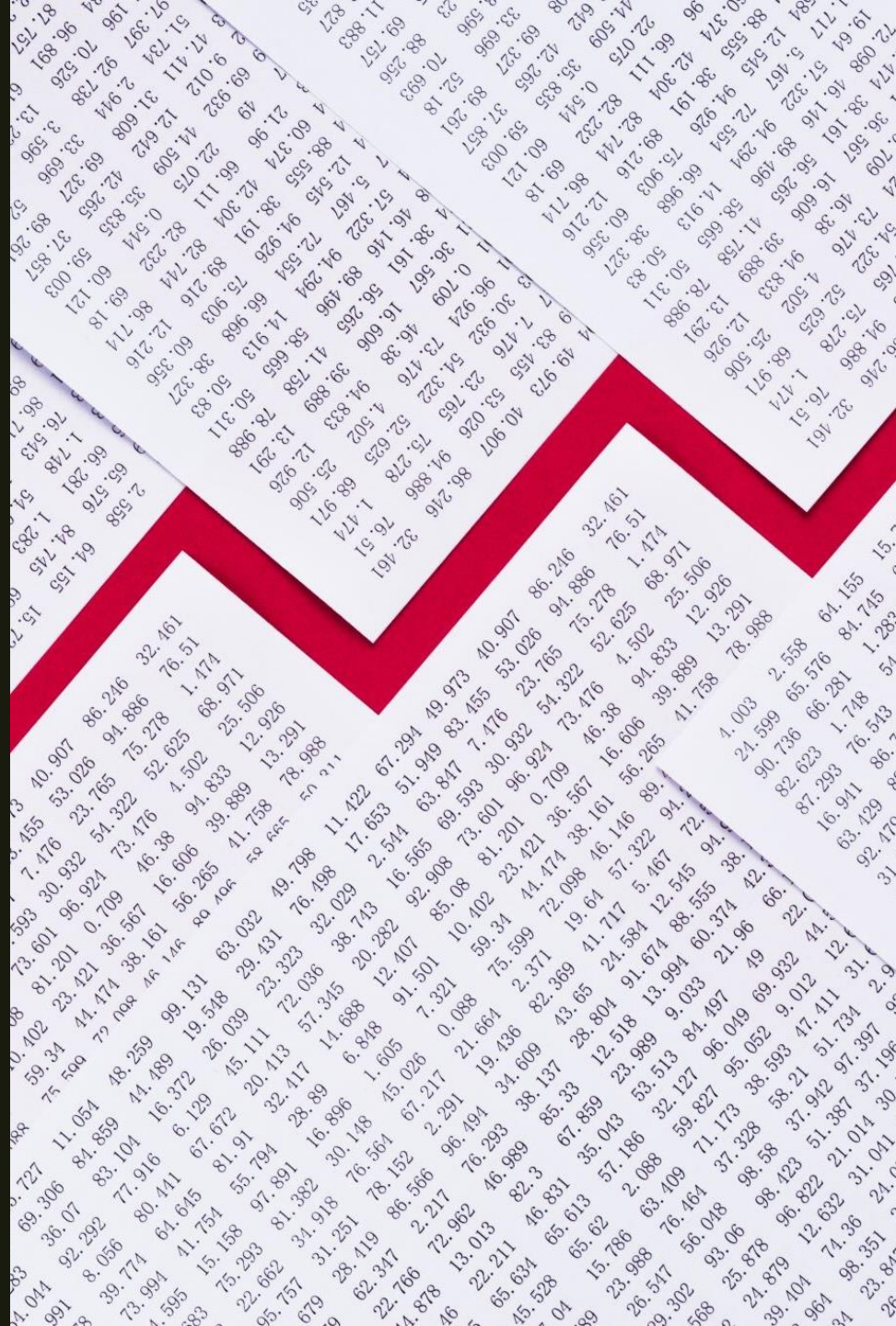


FROM RAW DATA TO PERCENTAGE GROWTH

What different patterns do you think are displayed in these two figures?

A COMPARISON

- The raw number view:
  - *A historical trend of growth*
  - *Larger numbers tend to be exaggerated and vice versa*
- The percentage view:
  - *Changes under each presidency*
- These are two different perspectives of the same data from which we can get different stories.



What is a research question?

- A research question is a clear and concise question serving as the focal point of a research project.
  - *It is generally phrased as a question.*
- It is different from a research problem statement.
  - *Research problem statement is focused on the bigger problem/issue/gap, whereas research questions are focused on more specific angles in the research problem.*

For example

- https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/asi.24986?saml_referrer

What is a good research question?

- The FINER criteria (Cummings et al., 2013):
  - *Feasible: a question should be researchable*
  - *Interesting: a question should have significance to something else (practice, policy, methods…)*
  - *Novel: a question should lead to new information and a gap in existing literature*
  - *Ethical:*
  - *Relevant: a question should be relevant to some broader communities*

Cummings, S.R., Browner, W.S., & Hulley, S.B. (2013). Conceiving the research question and developing the study plan. In: Designing clinical research (Hulley, S. R. Cummings, W. S. Browner, D. Grady, & T. B. Newman, Eds.; Fourth edition.). Wolters Kluwer/Lippincott Williams & Wilkins. Pp. 14-22.

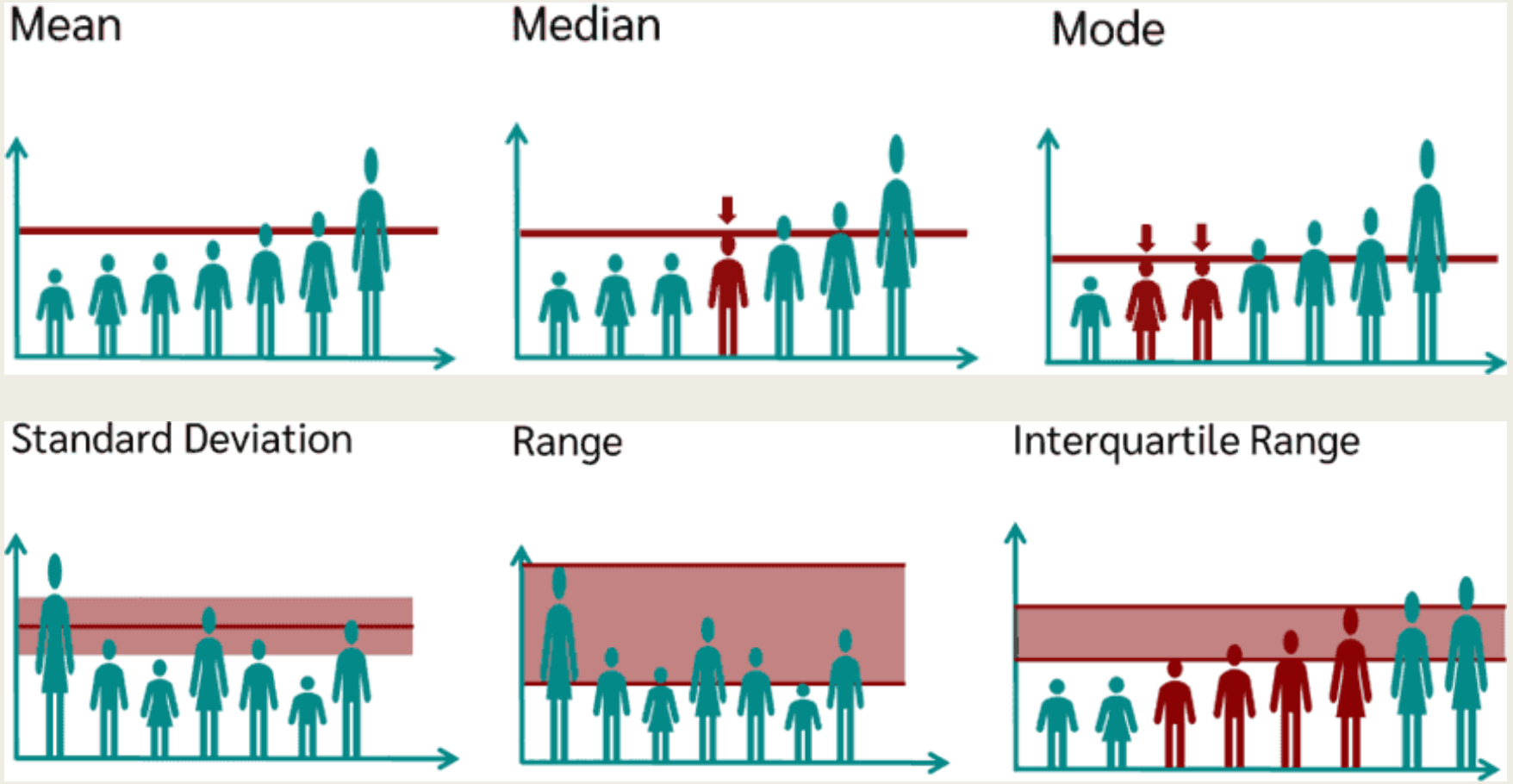
How to?

- There could be many recipes of getting good research questions.
  - *A question before doing the research, based on which we collect data and design and apply the methods*
  - *A question that emerges during the research design*
- But in any case, we want our questions to be concise and clear, so that readers can understand the RQs without any difficulty.
  - *All concepts used in RQs should be explained before the questions.*
  - *Use plain language.*

Descriptive methods

- Central tendency:
  - *Mean: average of all numbers*
  - *Median: the middle number for all sorted values*
- Dispersion:
  - *Variation: the difference between minimum and maximum values*
  - *Range (Min, Max, Quartiles)*
  - *Standard deviation: how disperse the sample is*

Descriptive methods



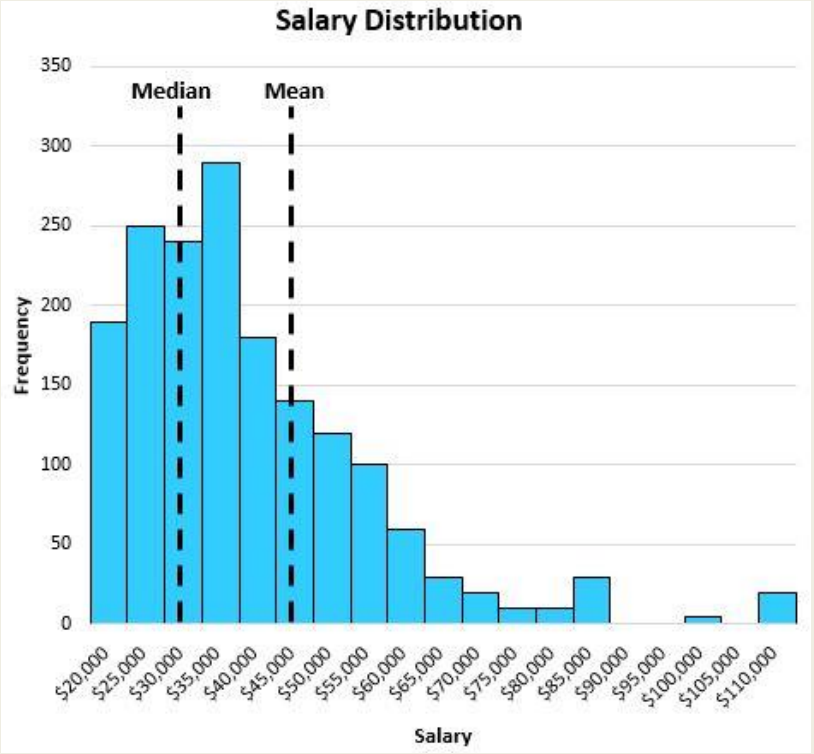
https://datatab.net/tutorial/descriptive-inferential-statistics

Reporting descriptive statistics

- It probably depends on the nature of research design.
- But a general guideline is here: https://about.illinoisstate.edu/mshesso-test2/reporting-statistics-in-apa-style/.
  - *Mean and standard deviation are the two central measurements.*
- But we should also try to understand:
  - *If there is any outlier (i.e., very different max or min values and very large range)*
  - *If mean and median are very different from each other.*

Mean vs. Median

- We should consider using median if mean is not able to represent the whole sample:
  - *For example, when there are extreme values in the dataset*
  - *We can also calculate median for ordinal data.*



Descriptive methods

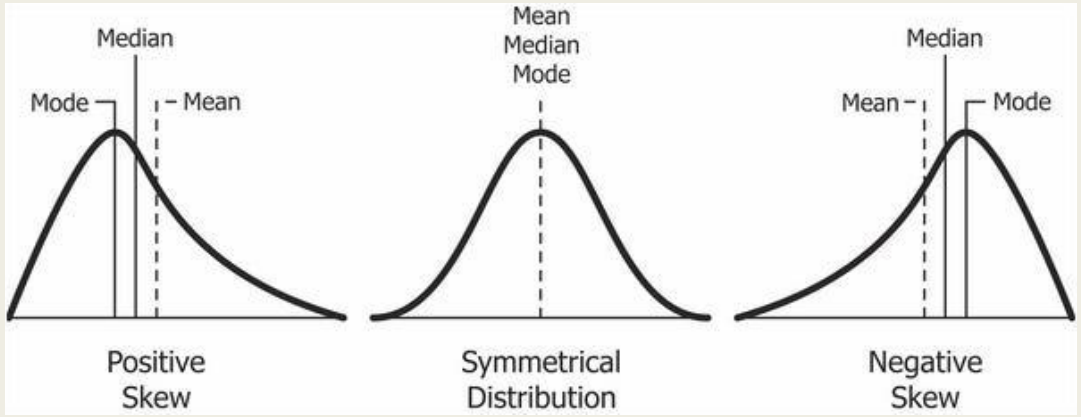
- All these descriptive statistical methods are very standardized and are implemented in nearly every statistical software.
- We cannot apply most of these techniques on categorical variables. So instead, we primarily reply on frequencies and percentage of values to offer description.

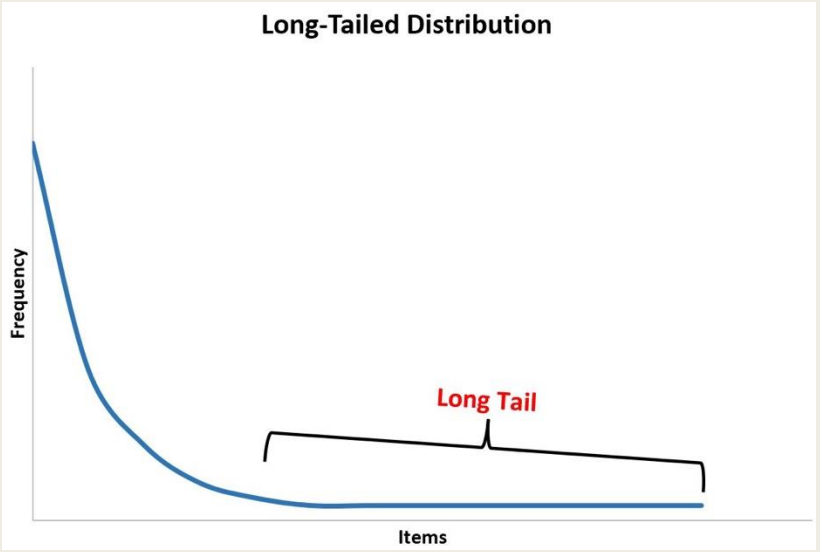
The use of descriptive analysis

- Doing descriptive analysis is generally the first and the most basic step of using the data.
  - *Establish basic understanding of the data*
  - *Identify outliers*
  - *Determine the usability of inferential methods*
    - Many inferential methods can only be applied to certain type of data, which we will discuss in the weeks of statistical methods.
- We can also include some descriptive results to the report as the first part of results.

Visualization of distribution

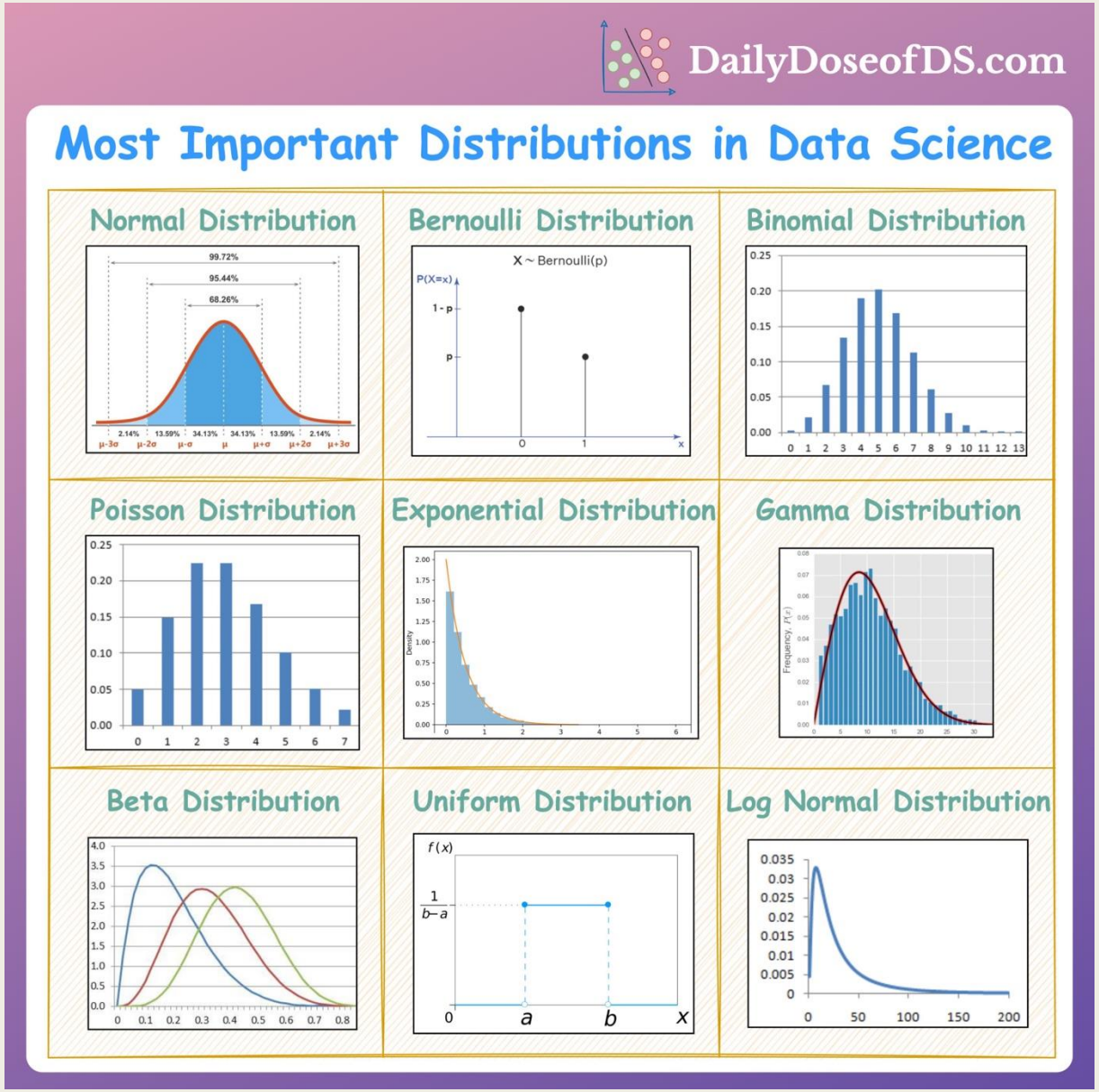
- Visualizing the distribution can also help us see many descriptive statistical measurements.
- Two methods of visualizing the distribution:
  - *Histogram or shape chart (it is technically NOT a line chart)*
  - *Box and violin charts*



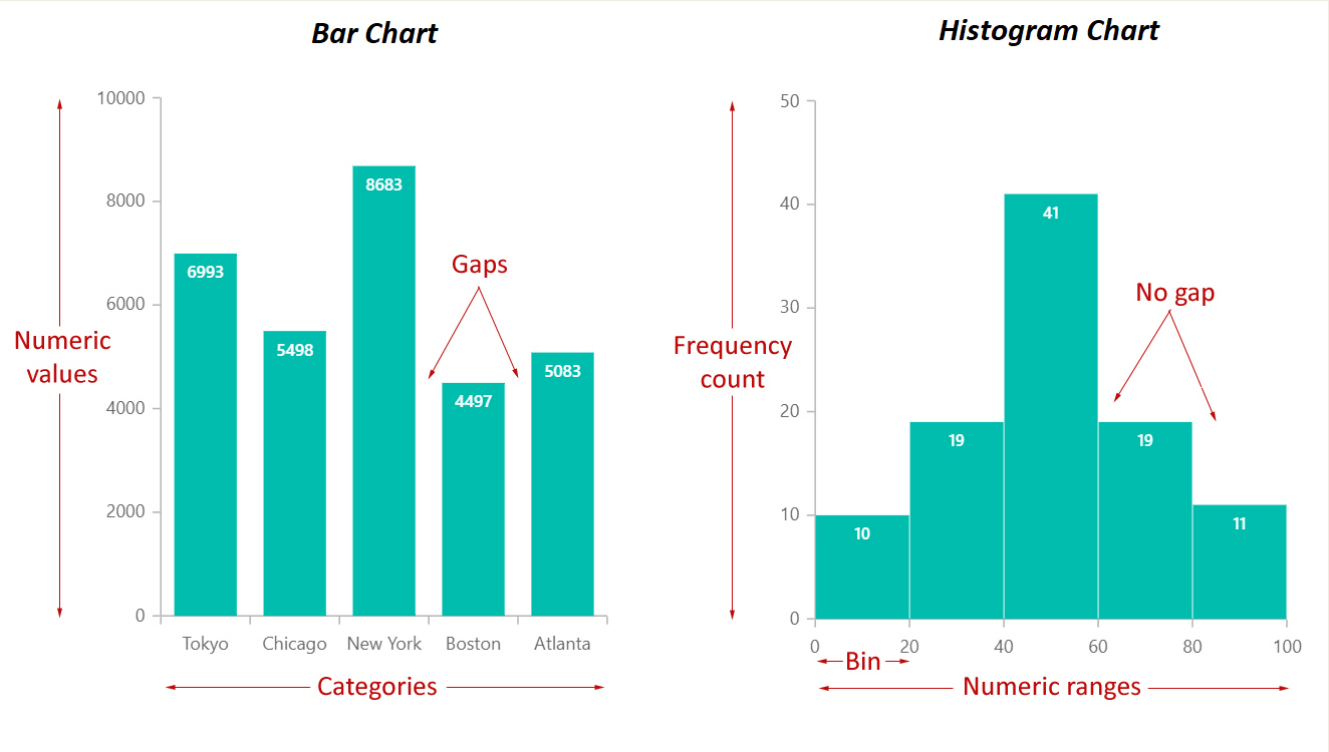


Types of distribution

- In histogram or shape chart, the y axis represents the no. of observations corresponding to the value.
- Many types of distribution
- Normal distribution
  - *Requirement for many statistical methods*

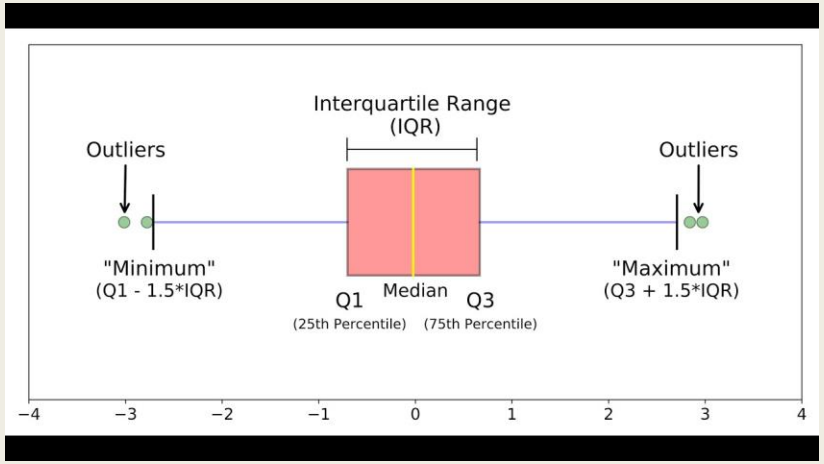


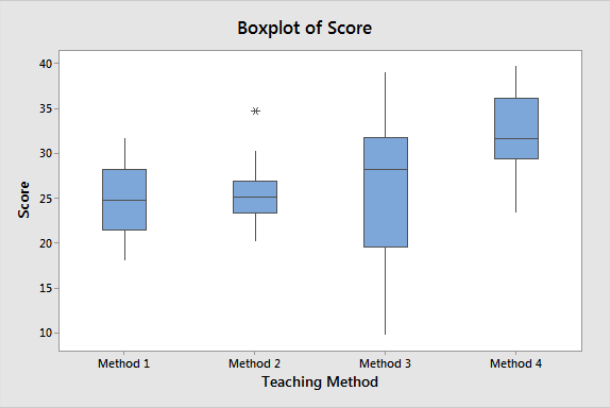
Histogram vs. Bar chart

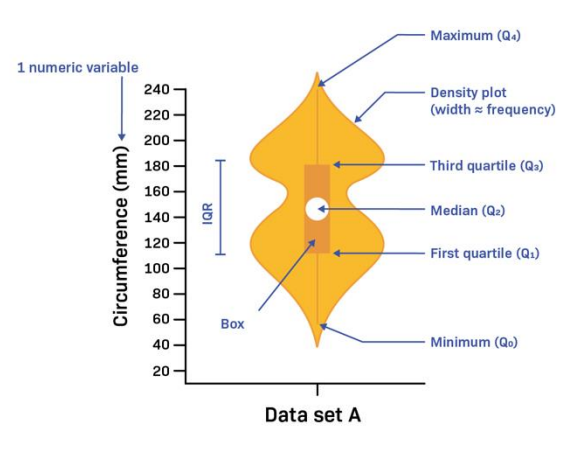


Box and violin charts

- Box plot can show key descriptive facts of one distribution.
  - *We can also use it to compare different distributions! (But the question is beyond descriptive analysis.)*
- Violin chart uses the shape to show the number of observations across data values.
  - *Box and violin charts can be used together.*







Demonstration

- 1. Data values in R
- 2. Core descriptive analysis functions in R
- 3. R files and GitHub
  - *There is a strong need for project organization (i.e., the management of your folders) in data analysis projects. → See the next slide*
- 4. GitHub:
  - *We may need to talk about it in the next week. But I will share a tutorial if you want to learn to use it.*
- 5. Class activity: descriptive analysis of the Boston dataset and RQs

Project organization

- Some general principles:
  - *One project has one main folder.*
  - *Separate code, data and outputs into separate sub-folders.*
    - We may want to further separate raw data and processed data, depending on the scope of the project.
  - *Choose a reasonable and consistent naming convention for your files and folders.*
    - Example: https://datamanagement.hms.harvard.edu/plan-design/file-naming-conventions
    - Also consider how to manage different versions of the code, data, outputs…
    - Write the ReadMe file.