KMMI 2021
Eksplorasi dan Visualisasi Data

Pertemuan 1:
Pengantar Eksplorasi dan Visualisasi Data

Highlights

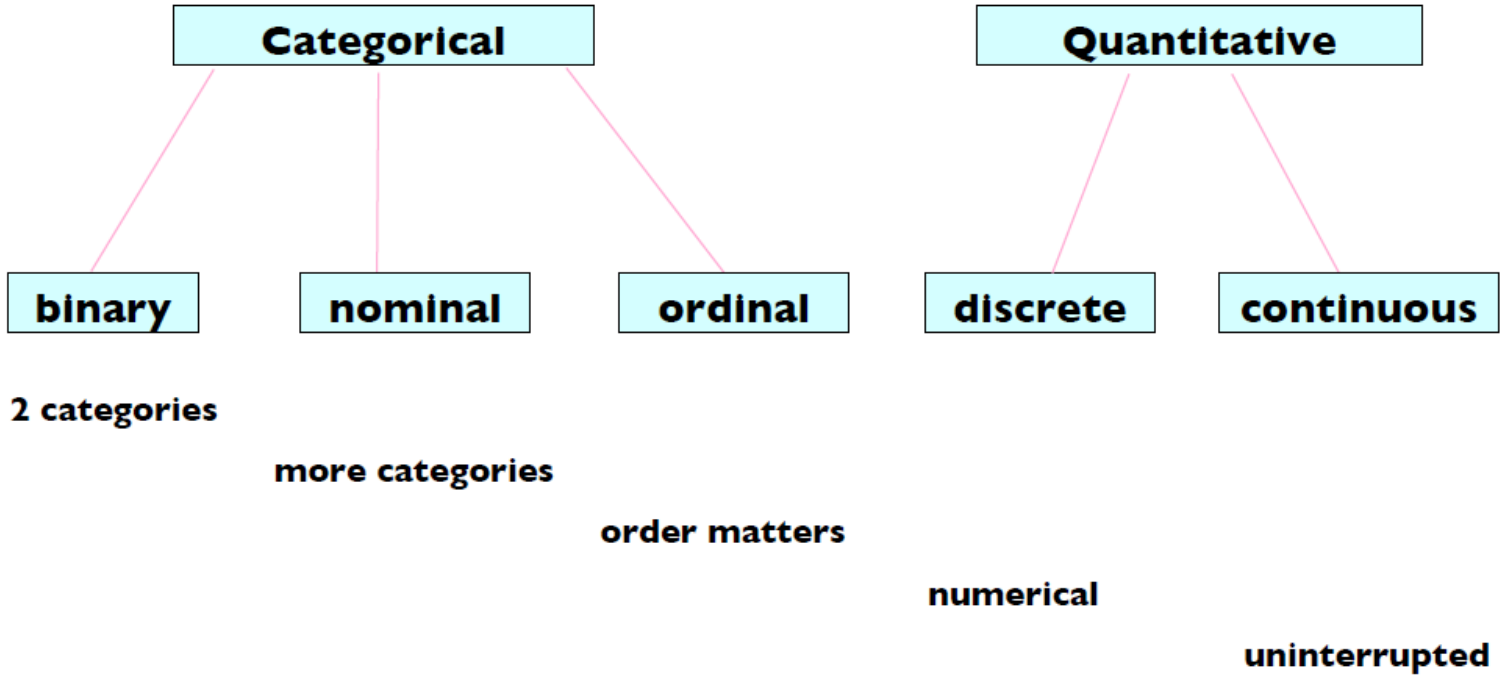
1. Data & Dataset
2. Introduction to Data Exploration
3. Introduction to Data Visualization
4. Good vs Bad Visualization

Background

- 74 zettabytes (1 zettabyte = 10^{21}) of data globally in 2021 (Statista)
- Data is very important, data is new money
- Data scientist is “sexiest job of the 21st century”
- **Data exploration and visualization** = one step closer to your data

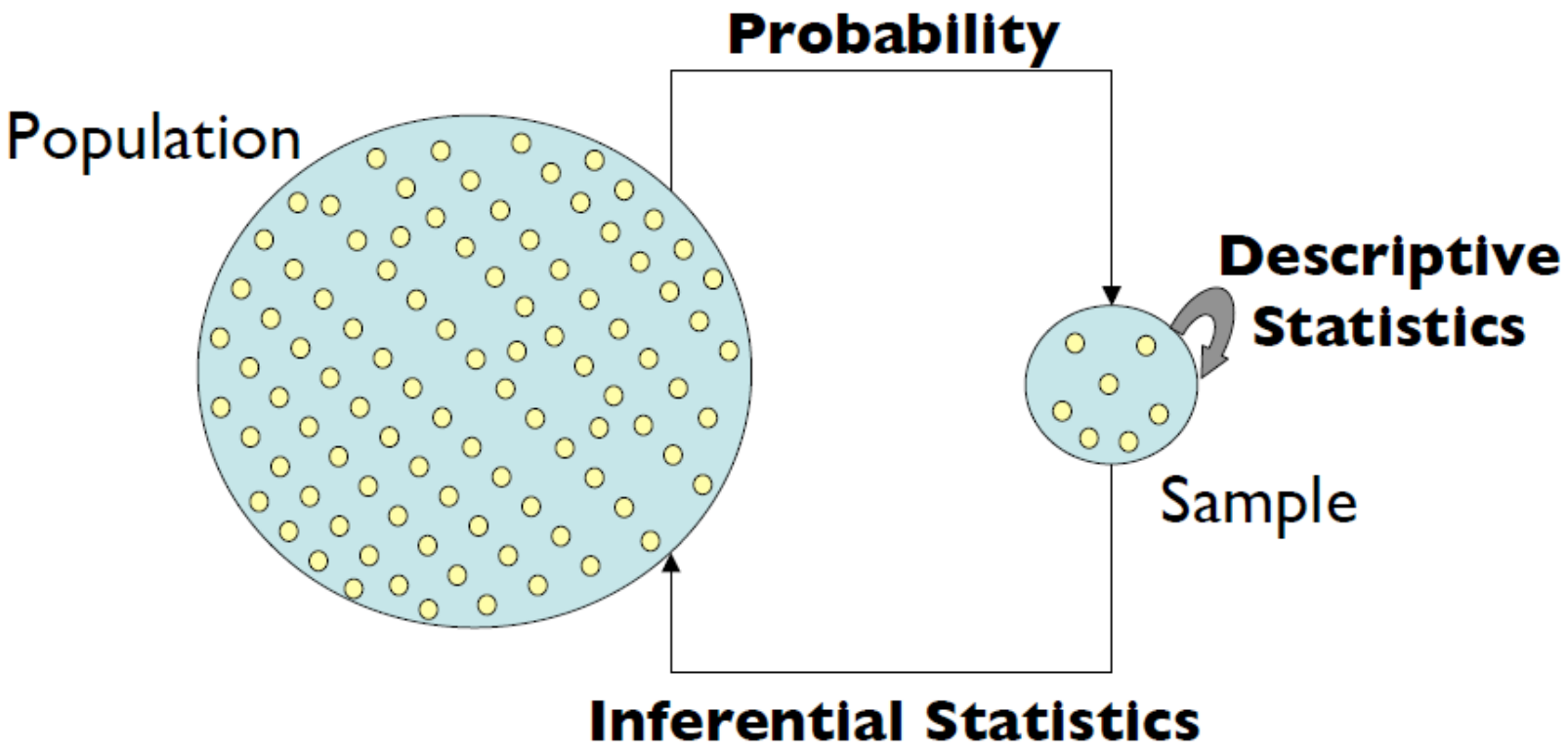
Data

- **A collection of facts** (numbers, words, measurements, observations, etc)
  1. Quantitative/Qualitative
  2. Categorical/Numerical
  3. Univariat/Bivariat/Multivariat



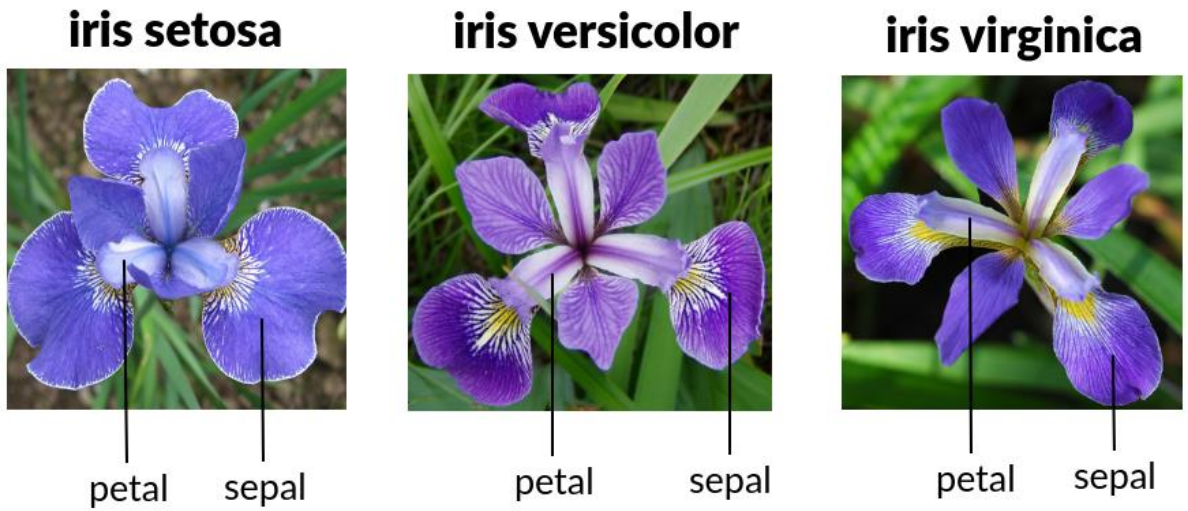
Data Collection

1. Census
2. Sampling



Dataset

- A data set (or dataset) is a **collection of data**.
- Data set refers to a file that contains **one or more records**.
- Usually presented in tabular form (**row and column**)

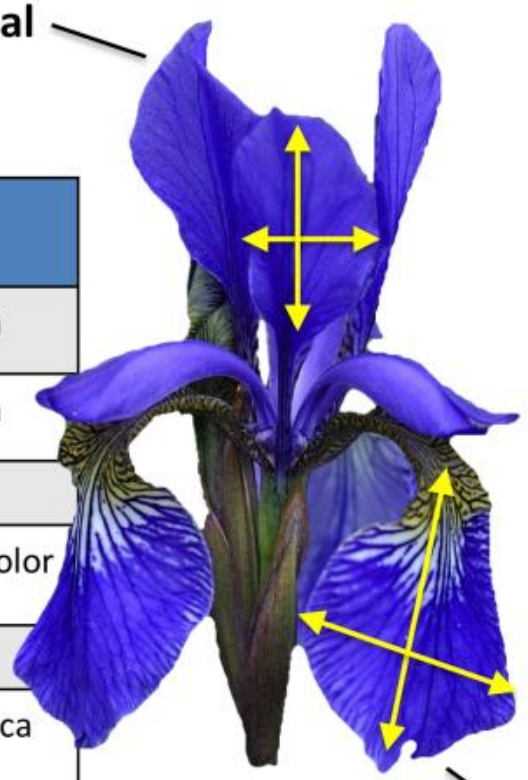


Dataset

- Contains:
  - Data object/
  - Samples/
  - Data points/
- Example:
  - Iris flower

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica



Class labels
(targets)

Features
(attributes, measurements, dimensions)

Data Exploration

- A preliminary exploration of the data to **better understand its characteristics**.
- Related to the area of Exploratory Data Analysis (EDA)
- Focus on:
  - Summary statistics
  - Visualization
- Why?
  - Helping to **select the right tool** for preprocessing or analysis
  - Making use of humans' abilities to **recognize patterns**

Data Exploration Tasks

1. Data understanding
2. Preprocessing
   - Join, cleaning, noise, outliers, duplicate, missing value, incomplete data
3. Basic/Summary Statistics
4. Data Visualization
5. Hypotheses
6. Assumption Checking
7. Story Telling (Reporting)

4 EDA Techniques

1. Univariate non-graphical
2. Univariate graphical
3. Multivariate non-graphical
4. Multivariate graphical

Summary Statistic

- Summary statistics are **numbers that summarize properties of the data**
- Summarized properties include **frequency, location, and spread**
- Example:
  - Location - mean
  - Spread - standard deviation
  - Frequency - mode

Data Visualization

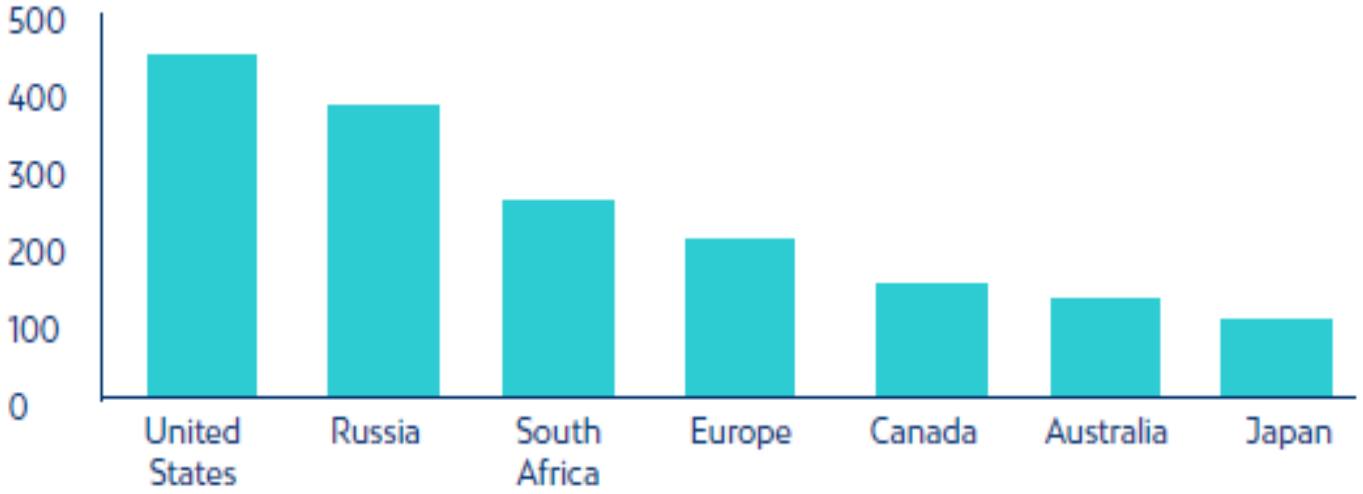
A (Good) Picture Is
Worth A 1,000 Words

Data Visualization

- Visualization is **the conversion of data into a visual or tabular format** so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.
- Visualization of data is one of **the most powerful techniques** for data exploration.
  - Humans have a well developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

Data Visualization

- Main Goal of Data Visualization:
  - Explaining
  - Exploring
  - Analyzing

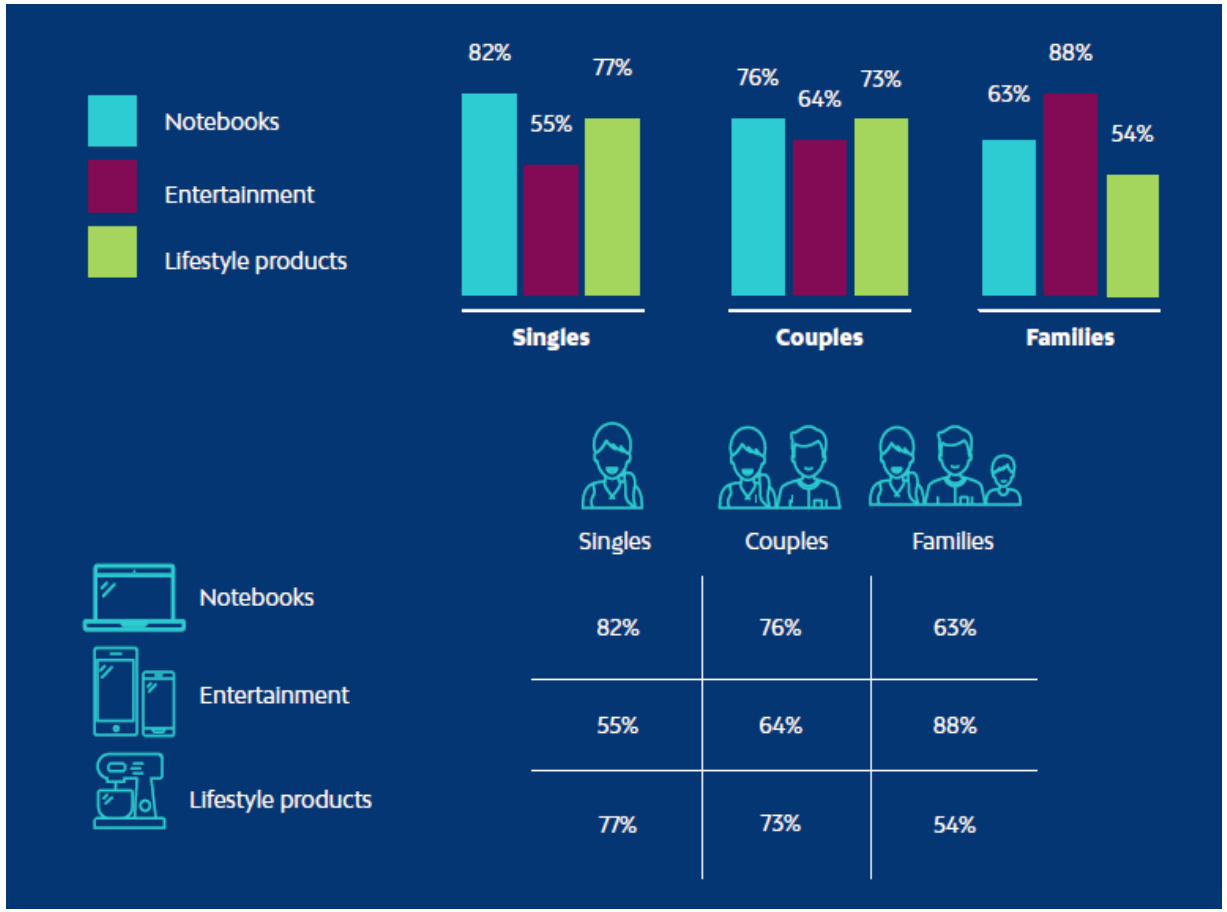


Good Visualization

- Display data accurately and clearly
  - Layout and design
  - Visual variables dan semantics
  - Consistent - colors
  - Simple icon and symbols

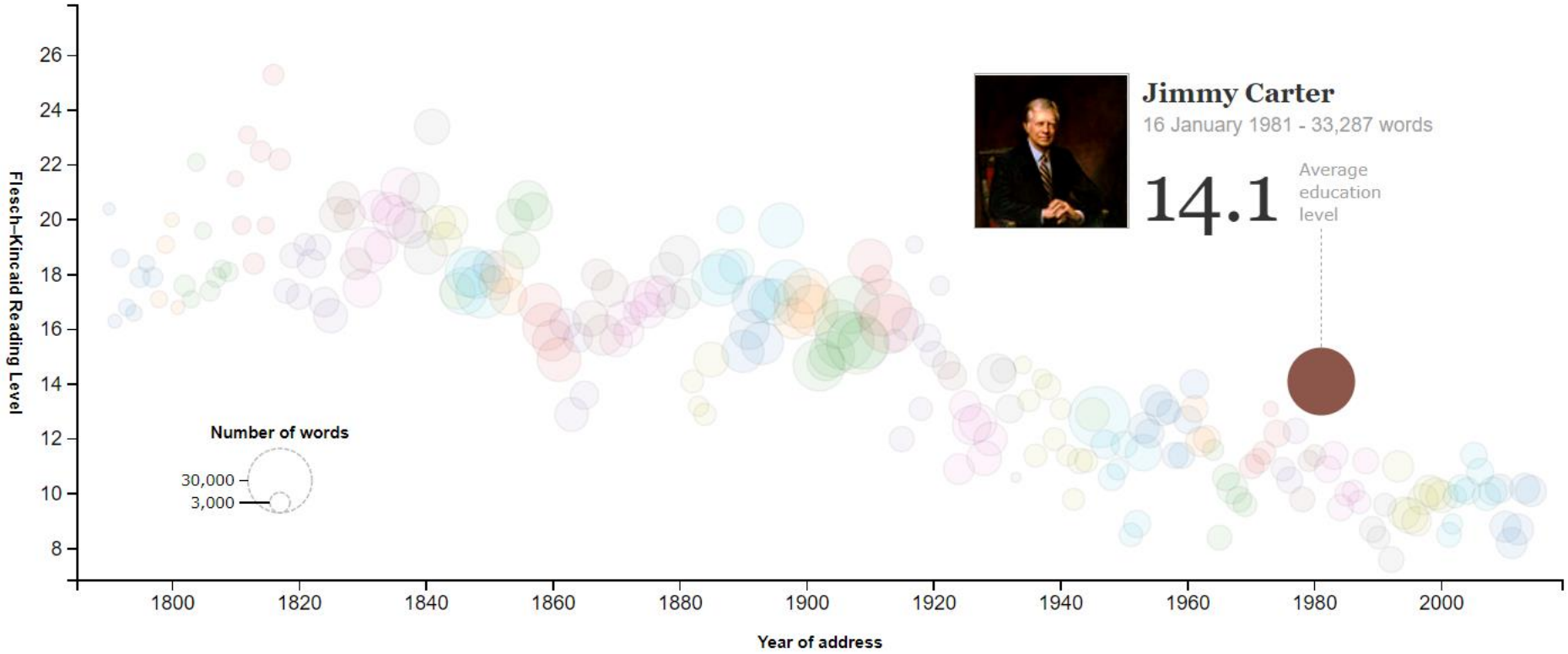
From Netquest.com

Good Visualization



From Netquest.com

Good Visualization



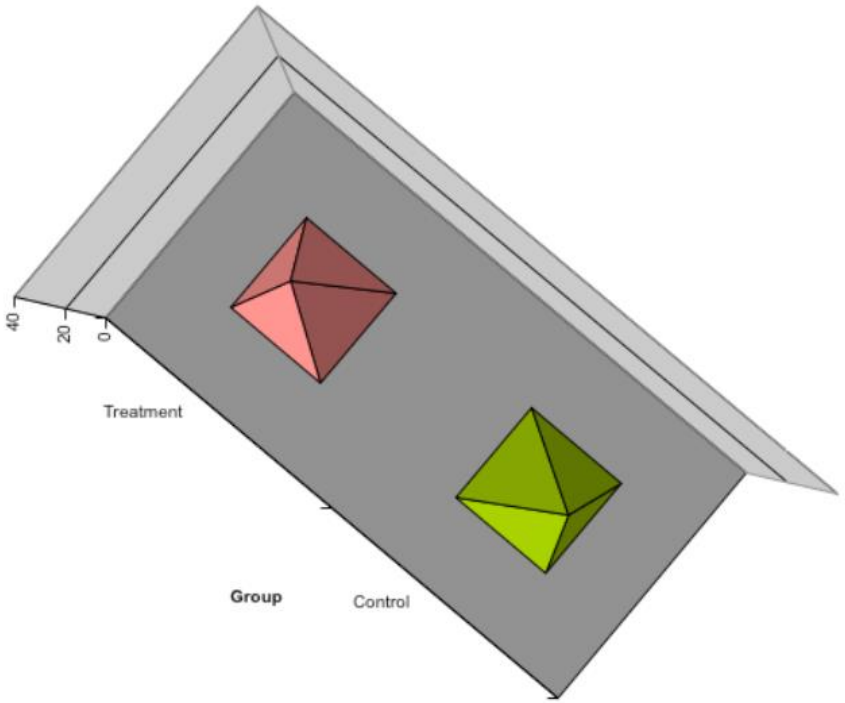
https://www.theguardian.com/world/interactive/2013/feb/12/state-of-the-union-reading-level

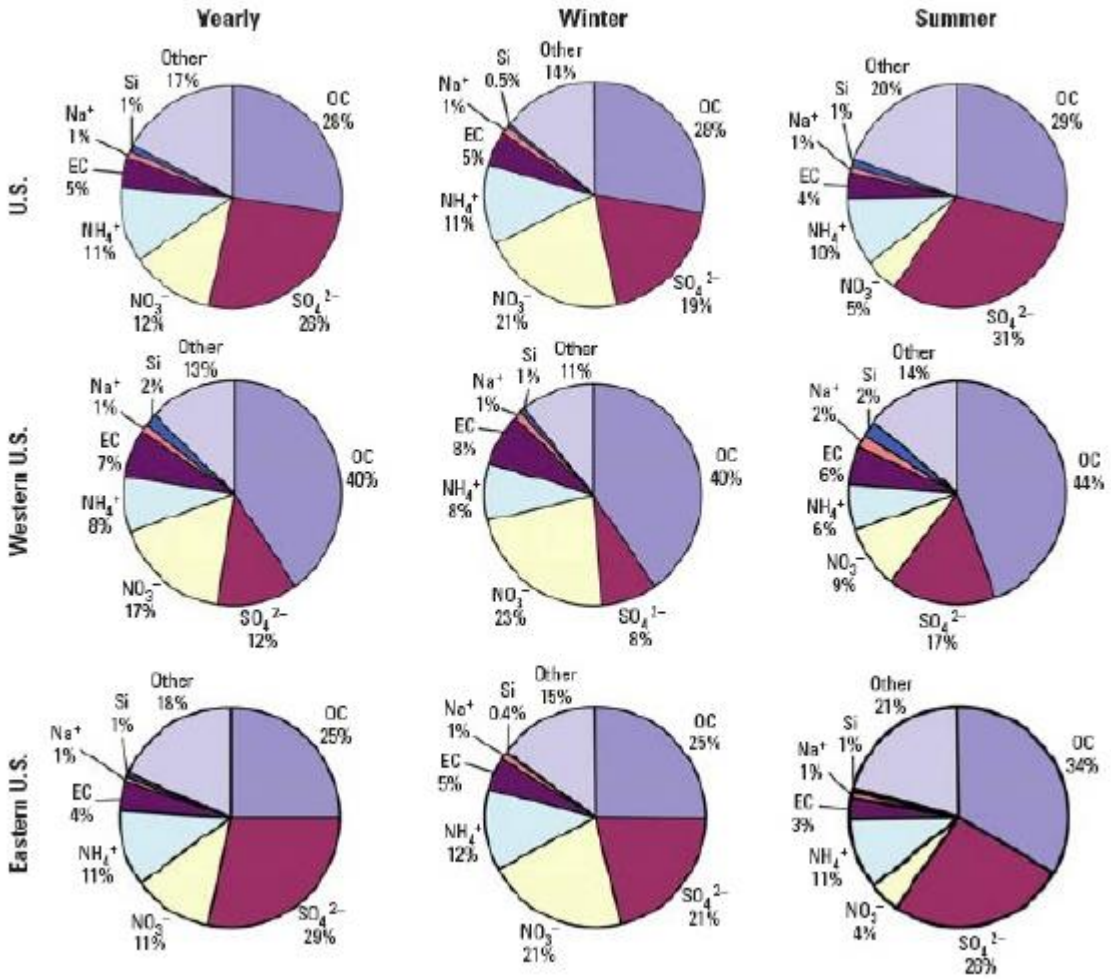
Bad Visualization

- Display as little/much information as possible
- Obscure what you do show (with chart junk)
- Use pseudo-3d and color gratuitously
- Make a pie chart (preferably in color and 3d)
- Use a poorly chosen scale

From Karl Broman: http://www.biostat.wisc.edu/~kbroman/

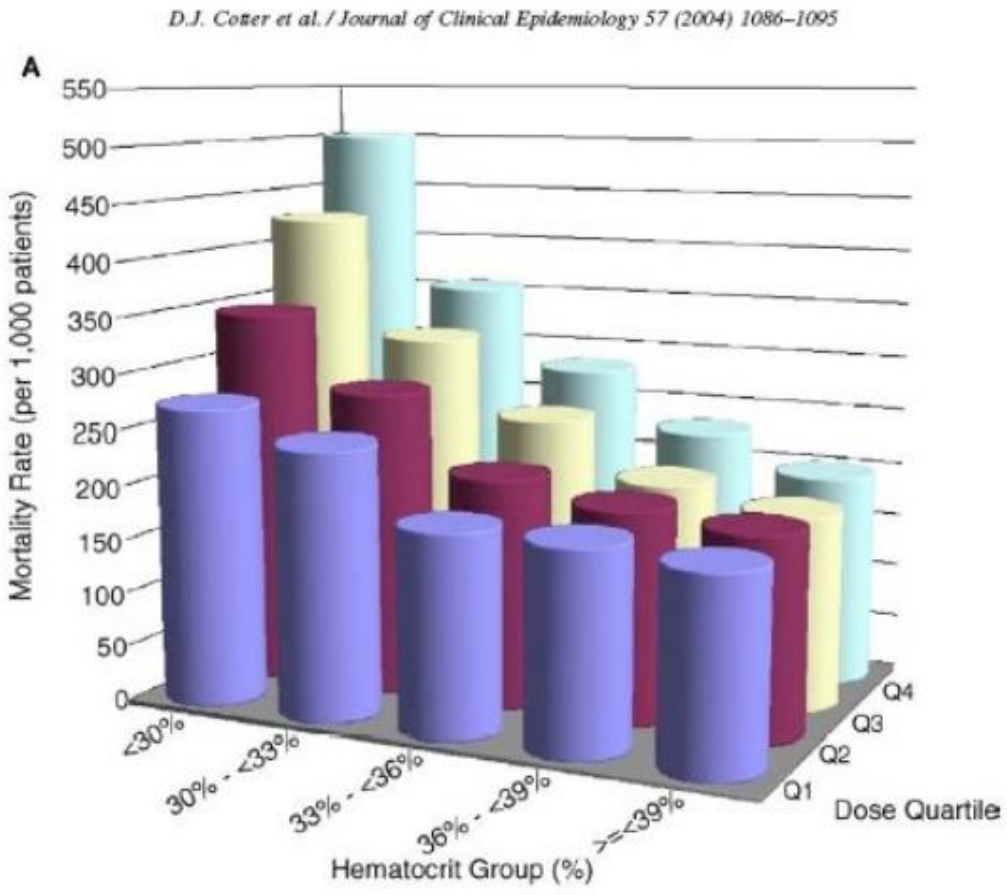
Bad Visualization

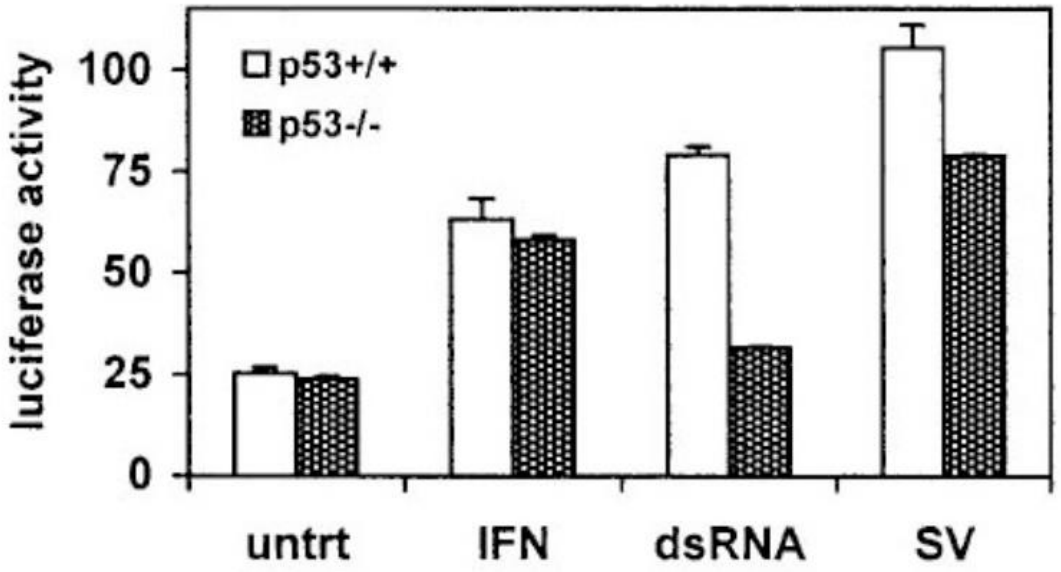




From Karl Broman: http://www.biostat.wisc.edu/~kbroman/

Bad Visualization





From Karl Broman: http://www.biostat.wisc.edu/~kbroman/

Data Exploration Tools

- Python
- R
- Etc..

Thank You