Introduction to Exploratory Data Analysis

What is Data Analysis?

- Data has no value in itself!
- Collecting, organizing and applying analytical tools to gain information from data which can be used to make decisions.
- A blend of statistics, computer programming, computer science, and domain knowledge.
- 4 types of data analytics: descriptive, diagnostic, predictive, and prescriptive

Exploratory Data Analysis

- “We look at numbers or graphs and try to find patterns. We pursue leads suggested by background information, imagination, patterns perceived, and experience with other data analyses.”

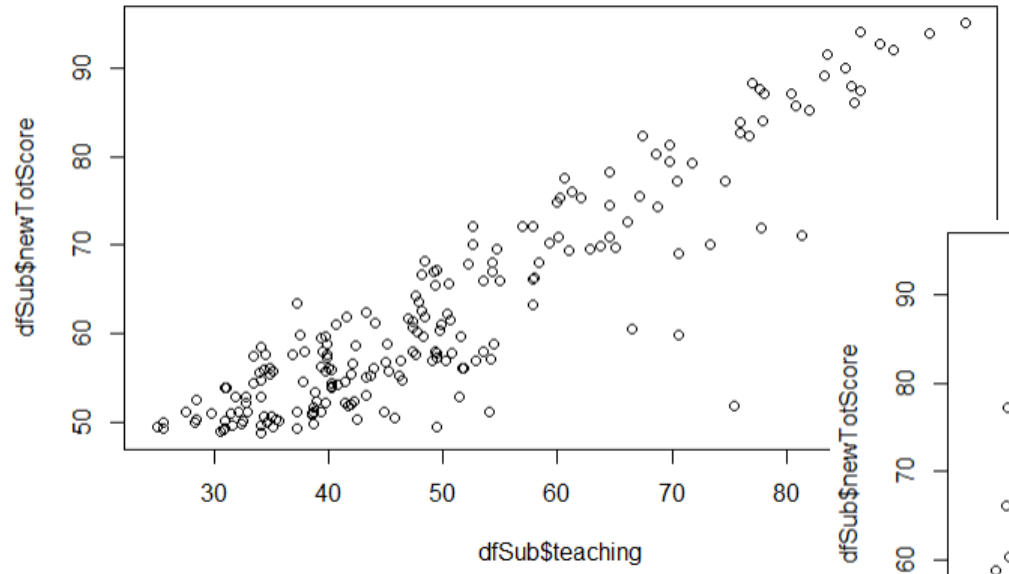
(P. Diaconis. Theories of data analysis: From magical thinking through classical statistics. In D.C. Hoaglin, F. Mosteller, and J.W. Tukey, editors, Exploring Data Tables, Trends, and Shapes, chapter 1. Wiley, 1985.)

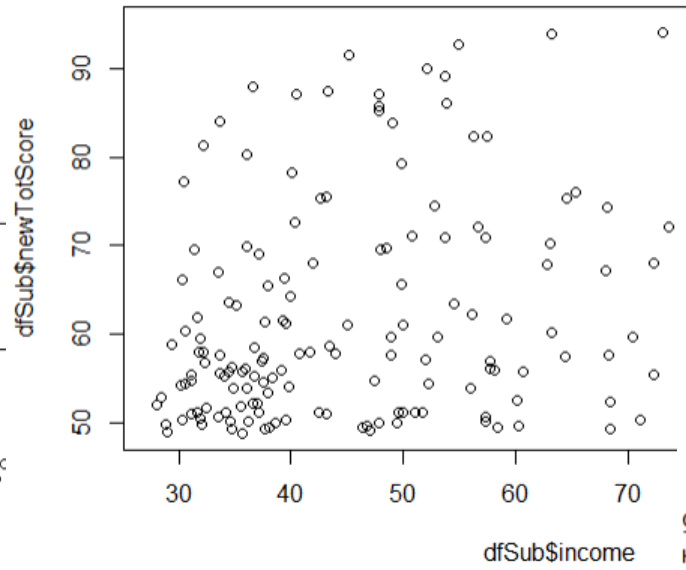
- “It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it.”

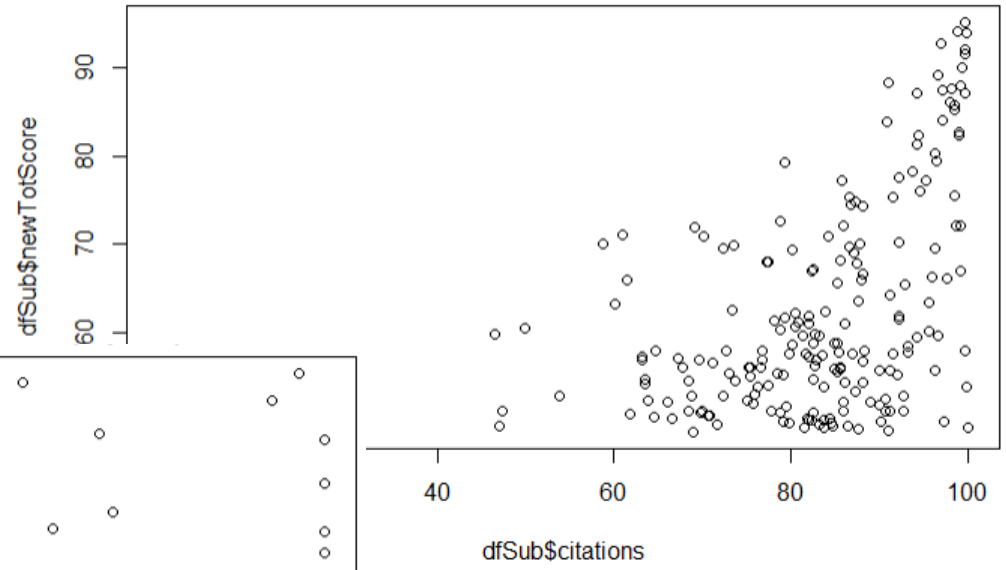
(J. Tukey, Exploratory data analysis, Addison-Wesley Publishing Company, 1977)

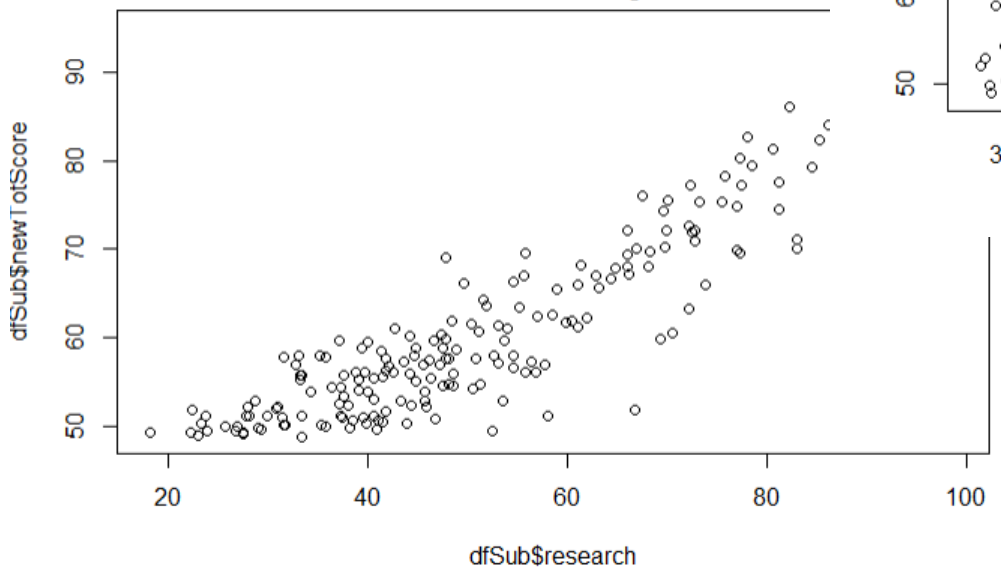
- “Humans are much better at seeing patterns in graphs than in large collections of numbers”
- “The greatest value of a picture is when it forces us to notice what we never expected to see”

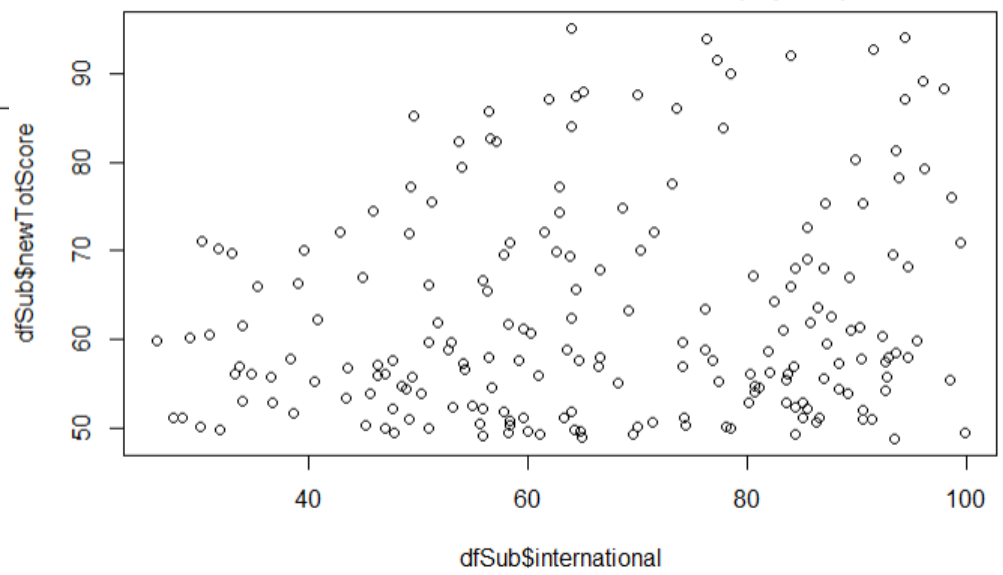
Exploratory Data Analysis











Exploratory Data Analysis

EDA helps build intuition about the data and validate understanding of the data – be it primary or secondary.

The following basic sequence can be helpful in exploring an unfamiliar dataset.

- **Step 1**: Assess the general characteristics of the dataset
  - How many records does this dataset contain?
  - How many fields (variables) are there?
  - What kind of variables are they?
  - How many unique values does each variable have?
  - Which values happen more frequently? Or what is the distribution for continuous variables?
  - Is there missing data? If so, how are missing values represented? How often does this happen?
  - Are the values of these variables consistent with what we expect?

Exploratory Data Analysis

- **Step 2**: Applying descriptive statistics to variables

- **Step 3**: Examine exploratory visualisations
  - Univariate, e.g. histograms, boxplots, bar plots
  - Bivariate, e.g. scatterplot

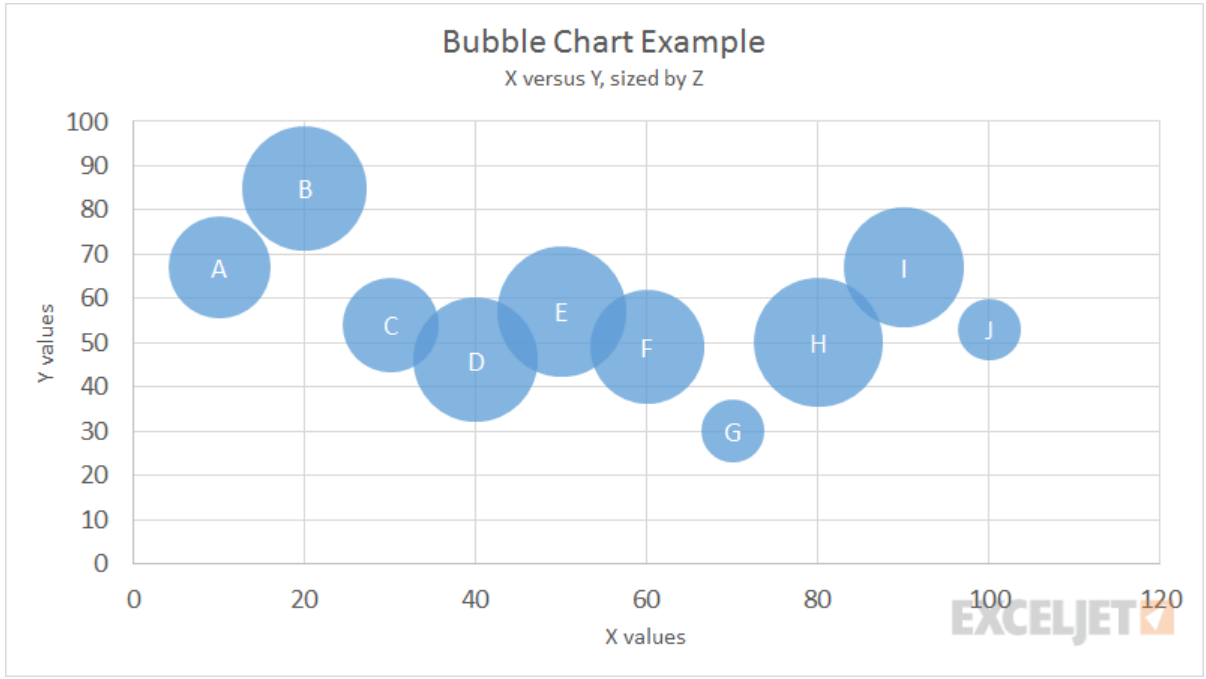
- **Step 4**: Applying techniques to look for data anomalies

- **Step 5**: Look at the relations between *key* variable pairs.

- **Step 6**: Summarise results in the form of a data dictionary

Notes in visualisations in Excel

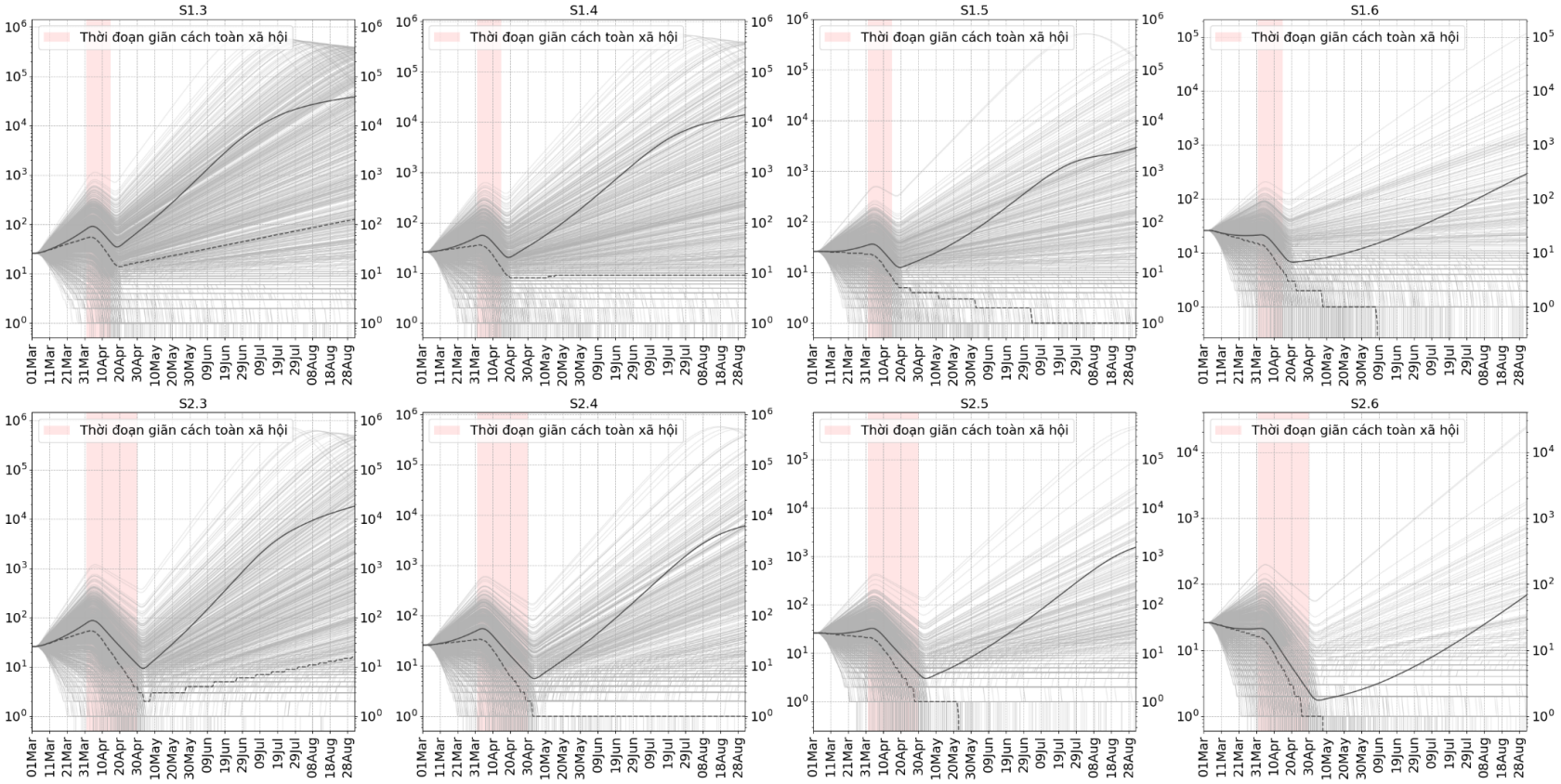
- Bin width of a histogram – Freedman Diaconis rules $bin\ width = 2 * IQR * n^{-1/3}$
- Plotting more than 2 variables
  - Bubble plots



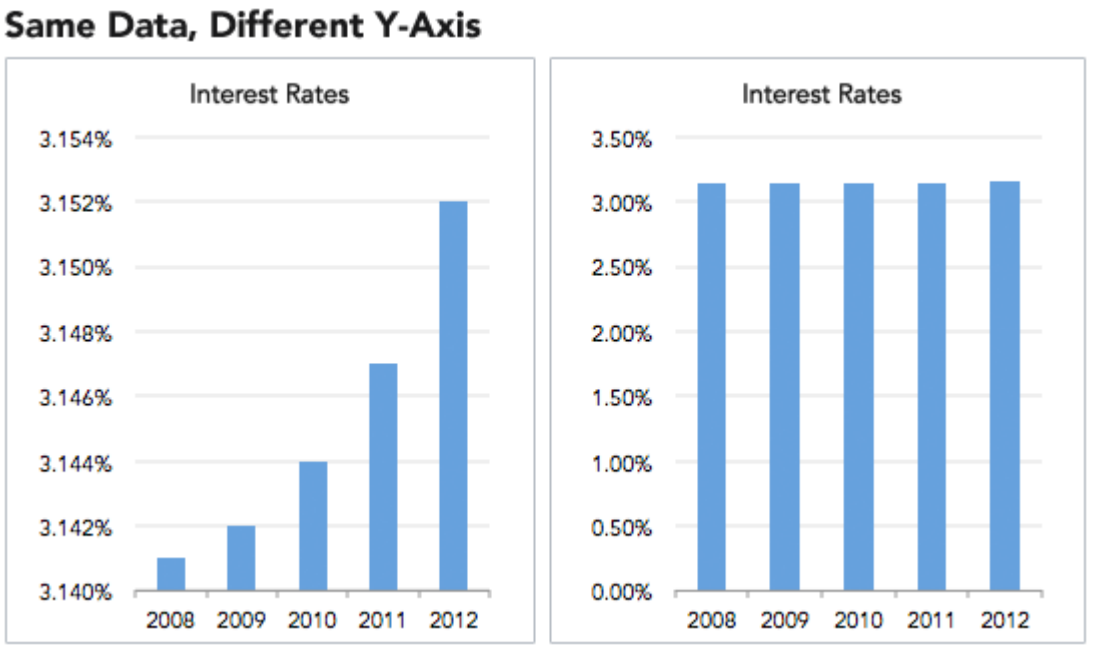
Source: https://exceljet.net/chart-type/bubble-chart

Notes in visualisations in Excel

- Plotting more than 2 variables with side-by-side plots
  - Consistency: chart type, axis scale, colour scheme
  - Arrangement: for easy comparison
  - Sequence: should follow some natural orders

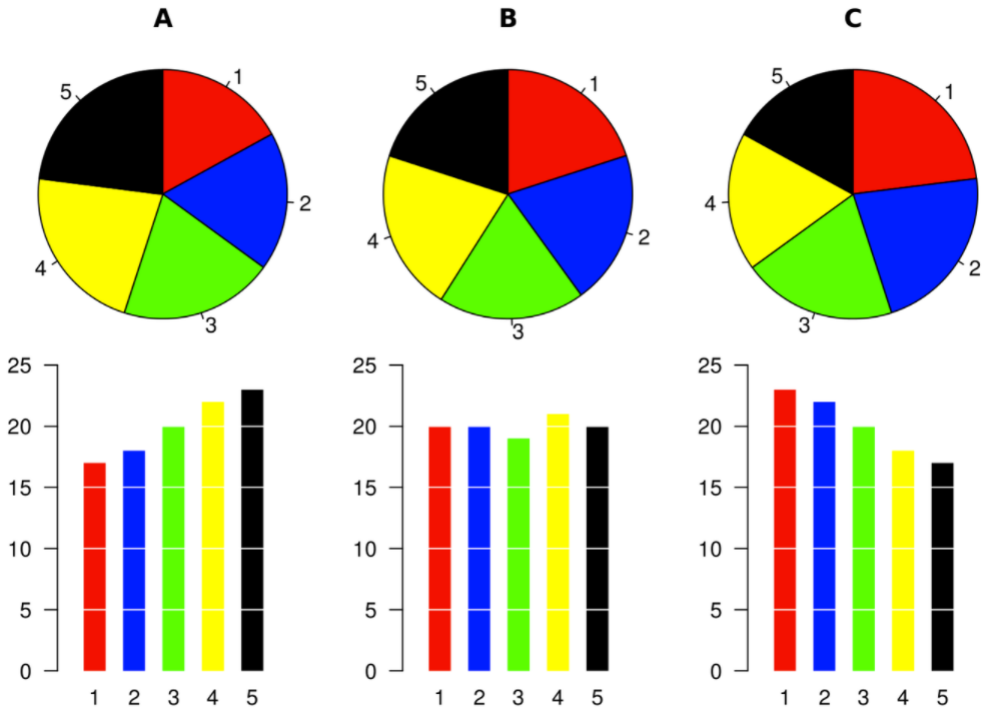


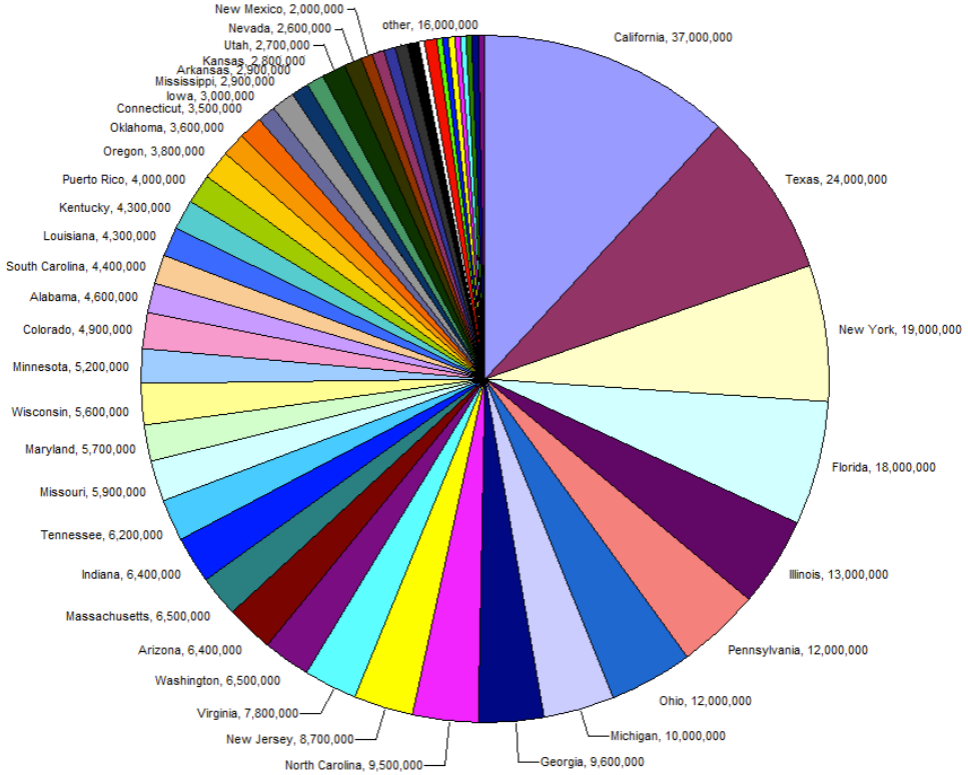
Plots to avoid



Source: https://www.datasciencejunction.com/2019/01/how-to-be-cautious-about-misleading.html

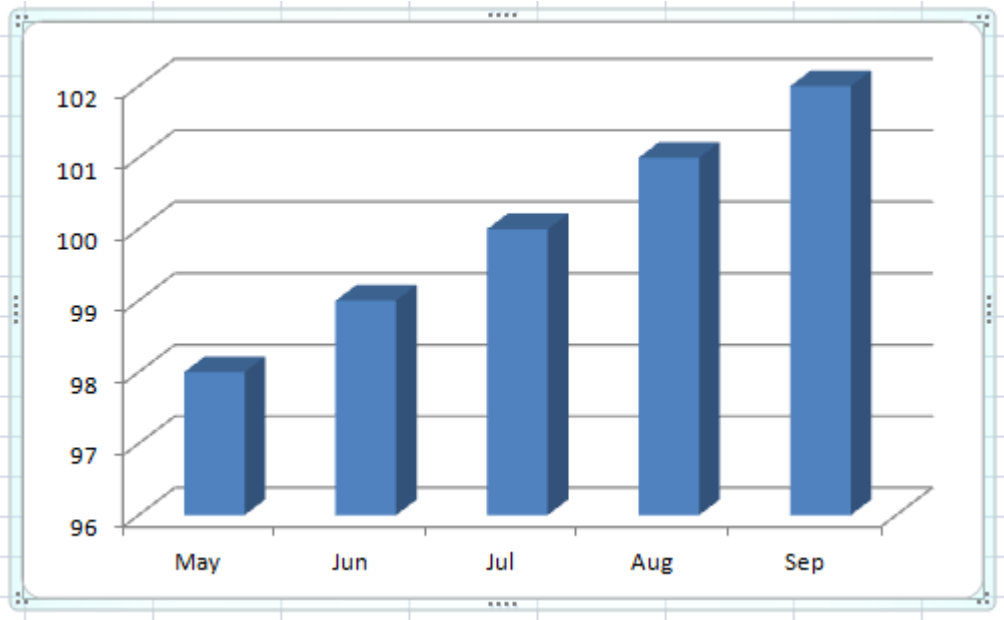
Plots to avoid



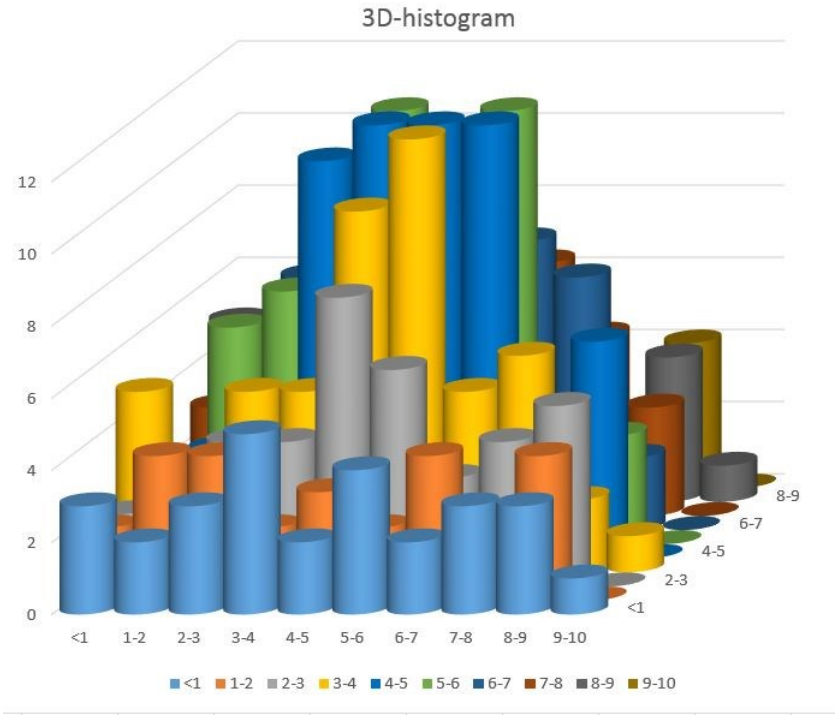


Source: https://eagereyes.org/pie-charts

Plots to avoid



Source: https://www.exceldashboardtemplates.com/wp-content/uploads/2012/11/image32.png



Source: http://excelgraphs.blogspot.com/2013/04/3d-histogram-in-excel.html

Other preprocessing considerations

- **Data transformation, e.g. centering and scaling**
- **Adding variables, e.g. one hot encoding**
- **Remove variables, e.g. those with zero or near zero variance**

Example

- **Practice example.** Download file bank.csv (downloaded from and extra information available here) and explore the dataset with Excel.