DATA ANALYSIS AND VISUALIZATION

INSTRUCTOR: UMME AMMARAH

INTRODUCTION

DATA

Data are raw facts, that have not been processed to explain their meaning.

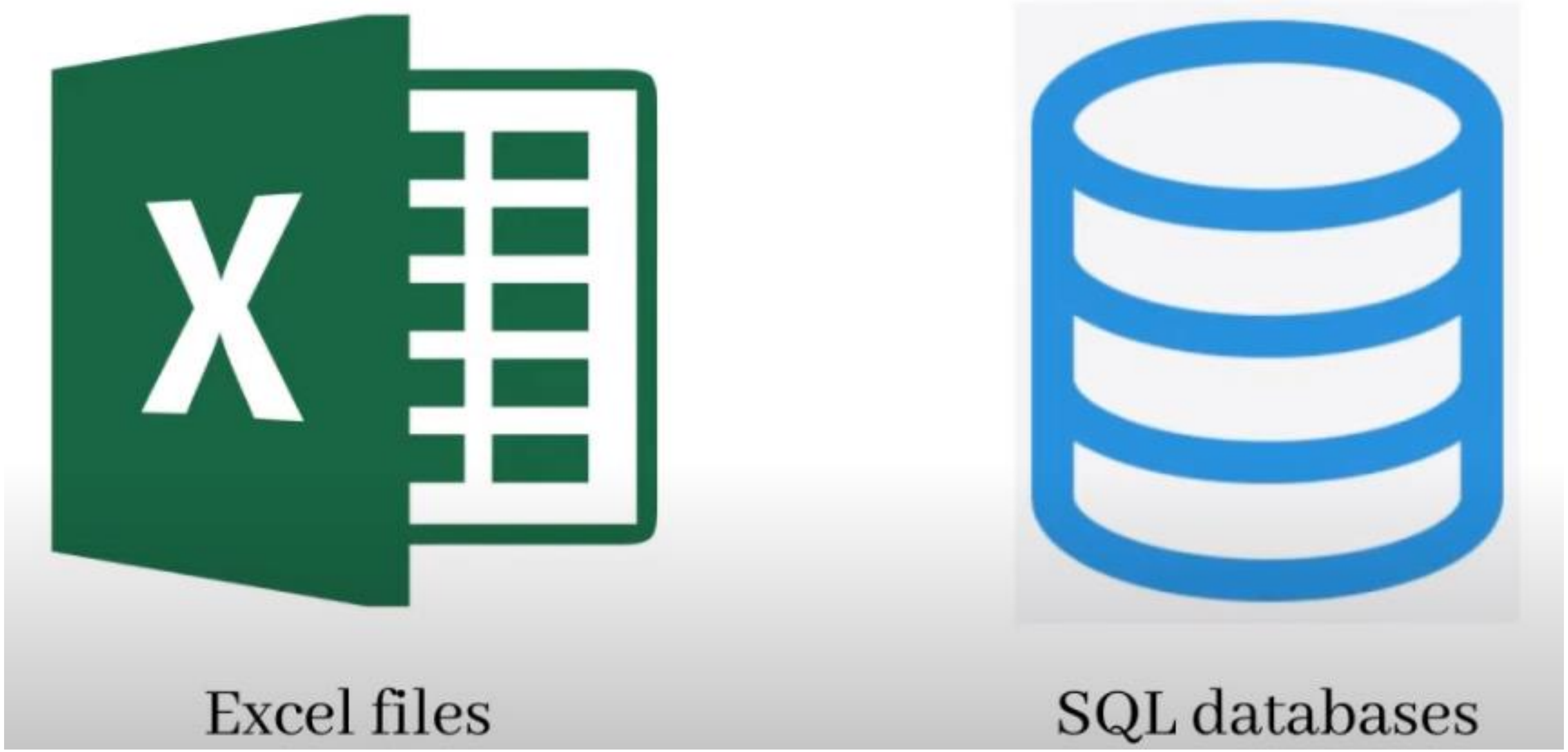
There are 3 different types of data:

- **Structured Data**
- **Unstructured Data**
- **Semi-structured Data**

STRUCTURED DATA

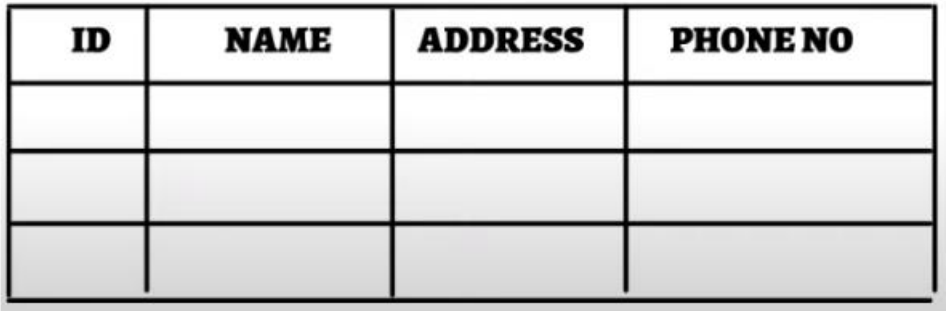
- Stored in a tabular format
- Clearly defined
- Stored in a predefined data model

EXAMPLE



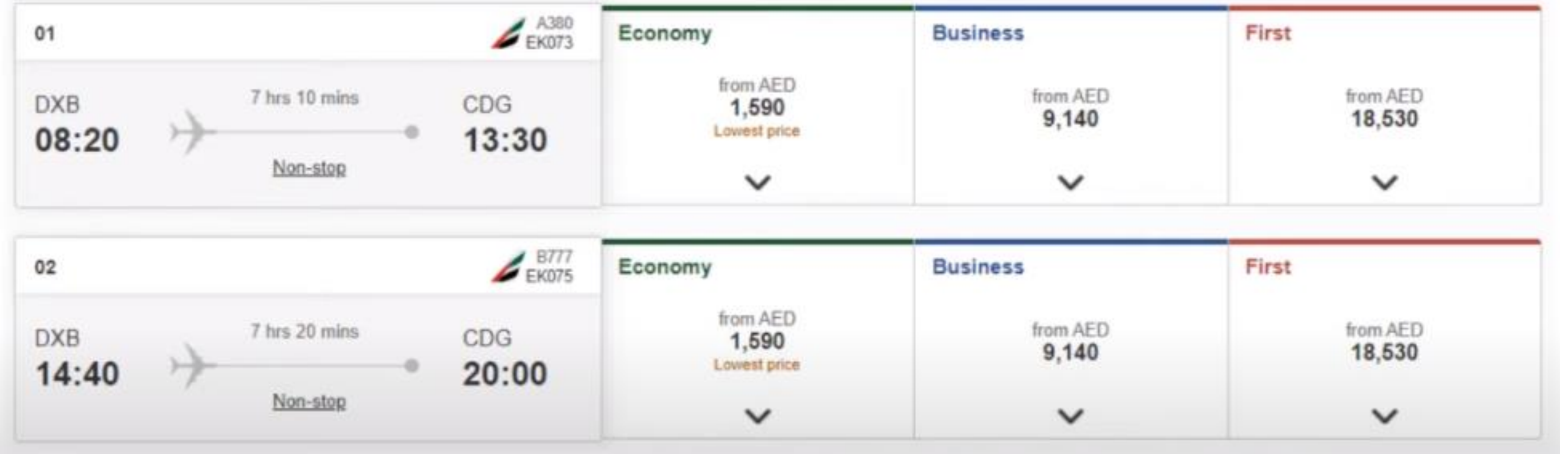
STRUCTURED DATA

- Rows and columns are related to each other
- Proper view and understanding of data



ID	NAME	ADDRESS	PHONE NO

<u>Example of Emirates airlines from Dubai to Paris</u>



01 A380 EK073

DXB 08:20 — 7 hrs 10 mins — Non-stop — CDG 13:30

Economy: from AED 1,590 Lowest price

Business: from AED 9,140

First: from AED 18,530

02 B777 EK075

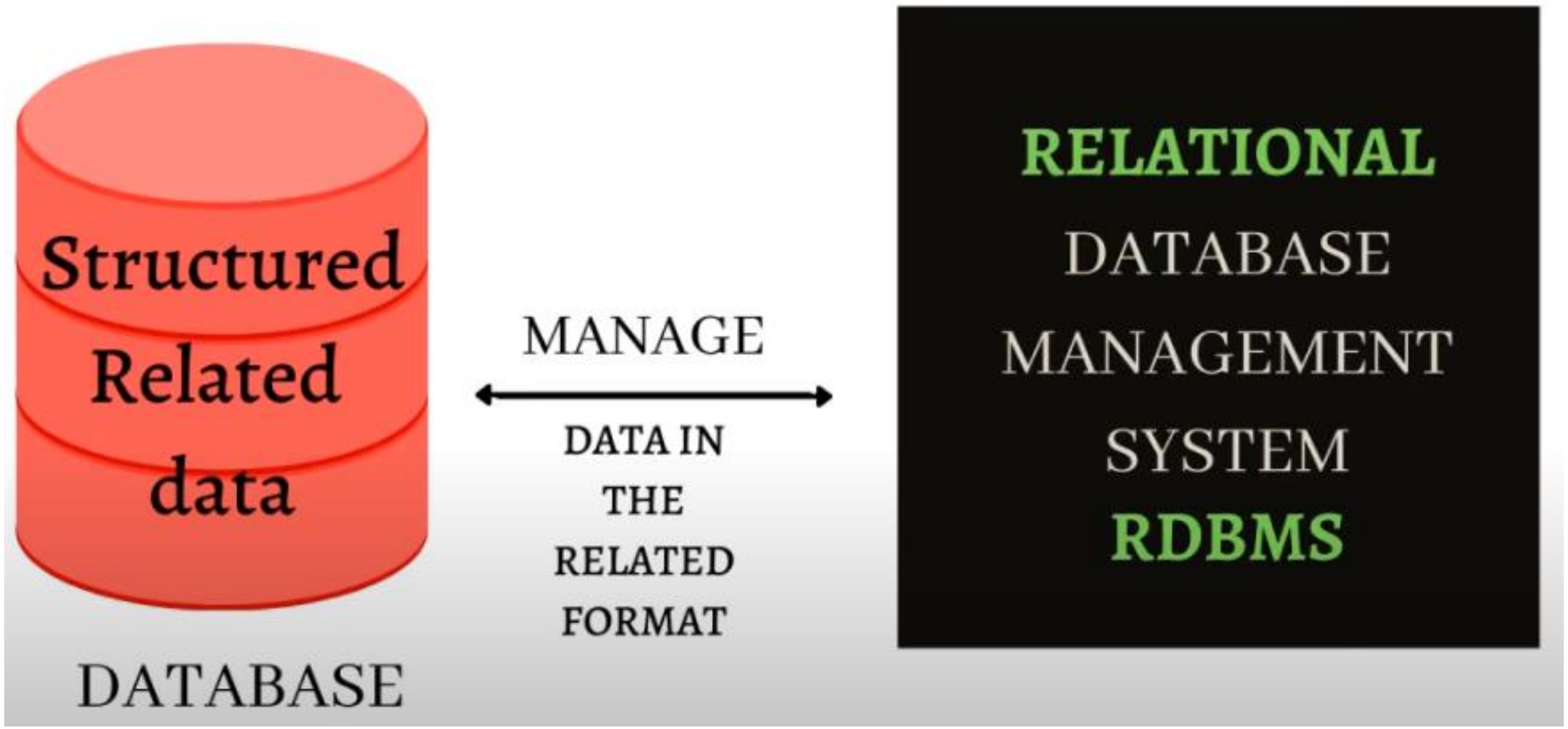
DXB 14:40 — 7 hrs 20 mins — Non-stop — CDG 20:00

Economy: from AED 1,590 Lowest price

Business: from AED 9,140

First: from AED 18,530

STRUCTURED DATA IS STORED IN RELATIONAL DATABASES



UNSTRUCTURED DATA

- No predefined structure
- No data model
- Irregular and ambiguous
- Easiest to extract data
- 80 to 90% data available is unstructured
- combination of text, images, videos, surveys, messages, numbers
- complex to analyze

EXAMPLE

SEMI STRUCTURED DATA

- Falls between structured and unstructured data type
- Combination of both
- Example: Emails, WWW, XML

DATA ANALYSIS

The Process of Analyzing the data

Huge amount of data

- Social media posts
- Products on e-commerce sites etc

But this data is not

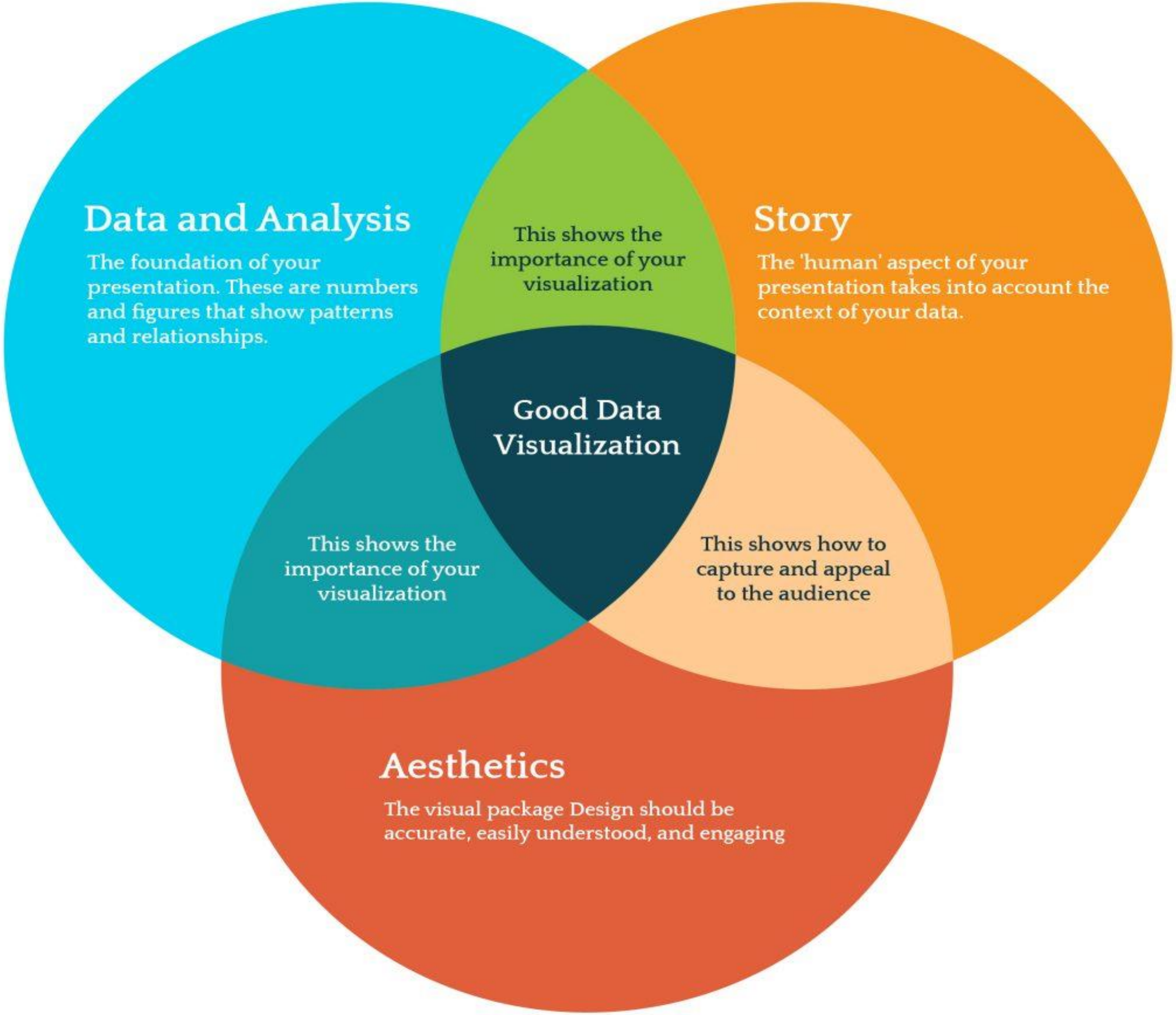
- Not accurate
- Not in one place
- Not directly useful

PHASES OF DATA ANALYSIS

- Data requirements
- Data collection
- Data processing
- Data cleaning
- Exploratory Data Analysis
- Modelling and algorithms
- Data product

DATA VISUALIZATION

- **The graphical representation of information and data**



Data and Analysis

The foundation of your presentation. These are numbers and figures that show patterns and relationships.

This shows the importance of your visualization

Story

The 'human' aspect of your presentation takes into account the context of your data.

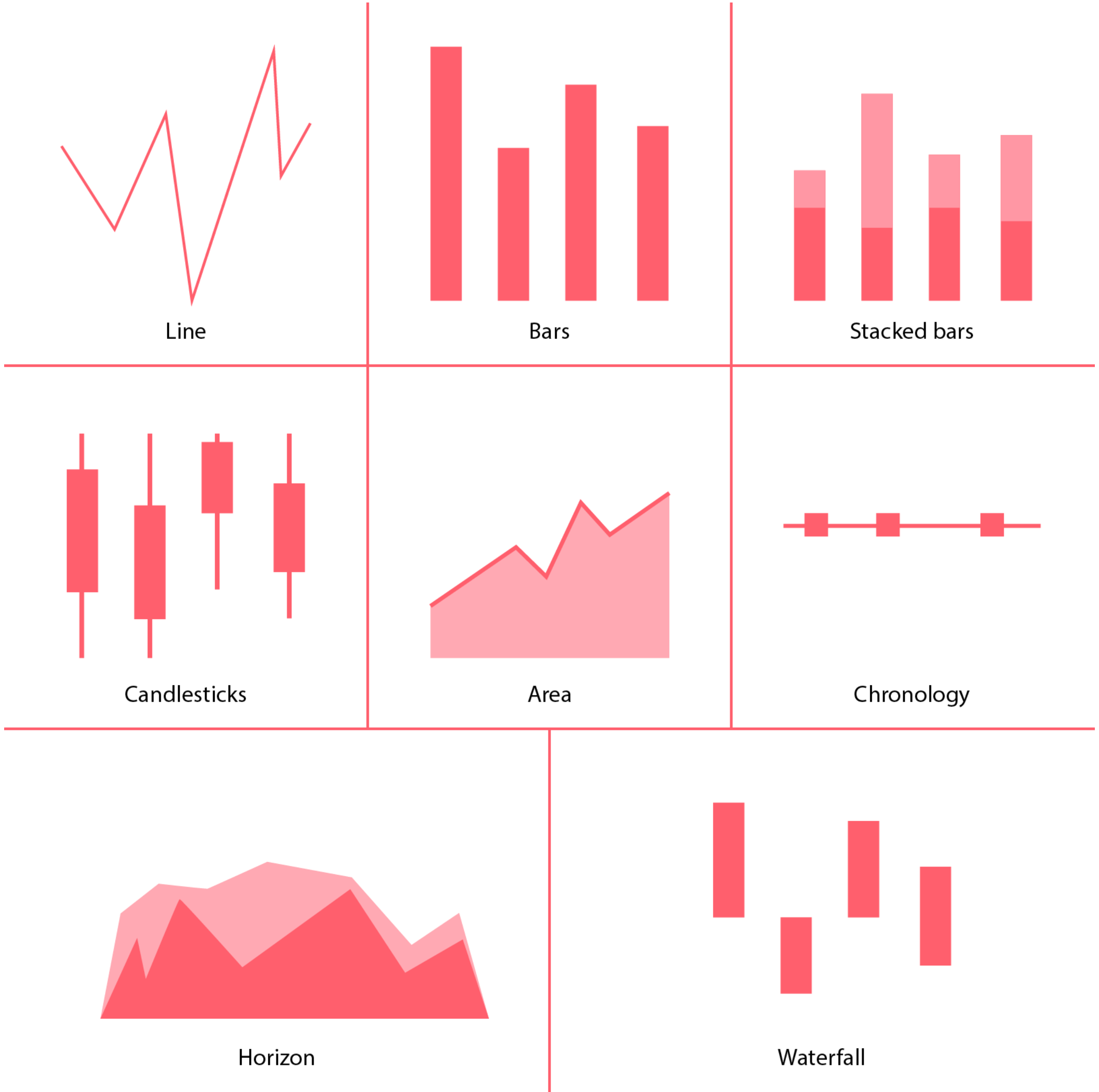
Good Data Visualization

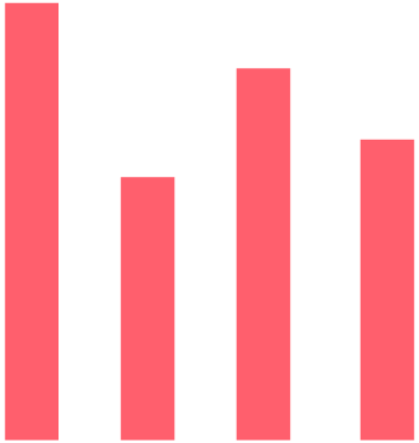
This shows the importance of your visualization

This shows how to capture and appeal to the audience

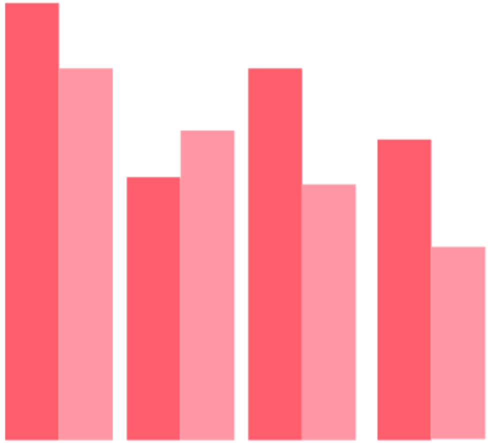
Aesthetics

The visual package Design should be accurate, easily understood, and engaging

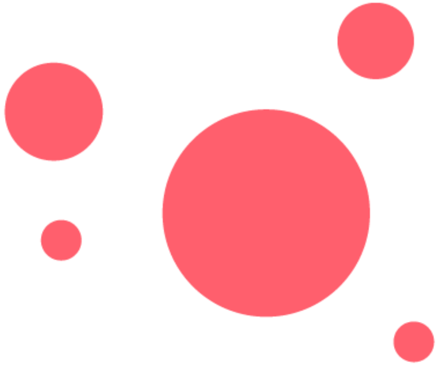

Line
Bars
Stacked bars
Candlesticks
Area
Chronology
Horizon
Waterfall



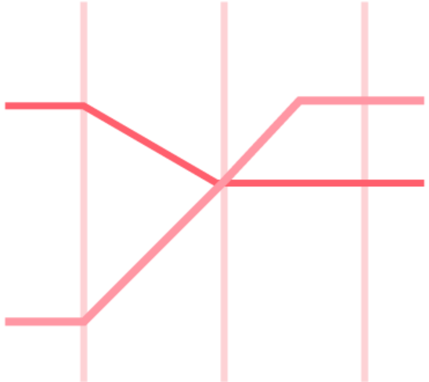
Bars



Grouped bars



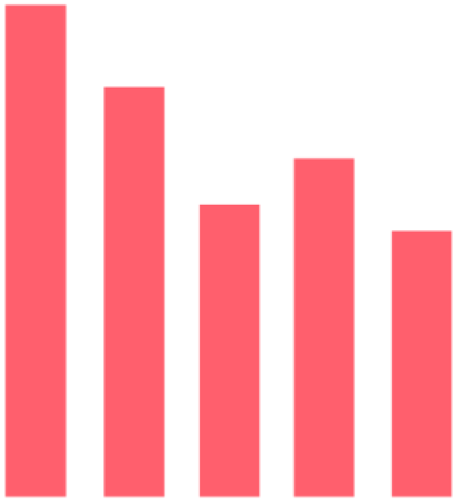
Bubles



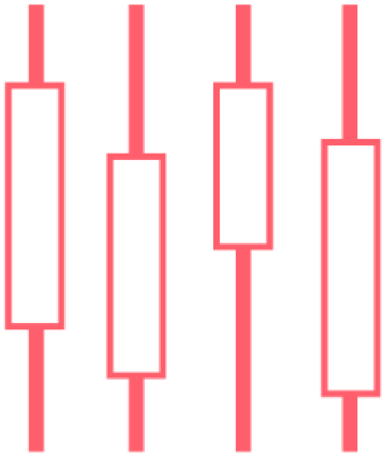
Multi-lines

Parallel coordinates

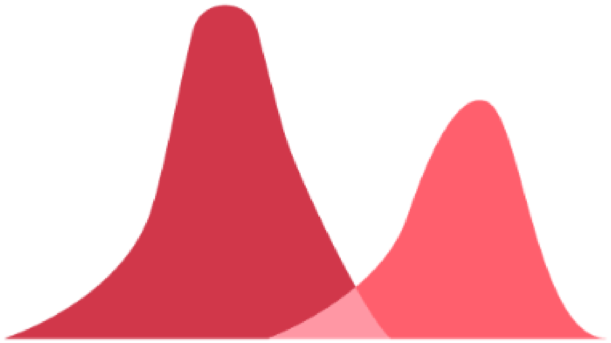
Bullet chart



Histograms

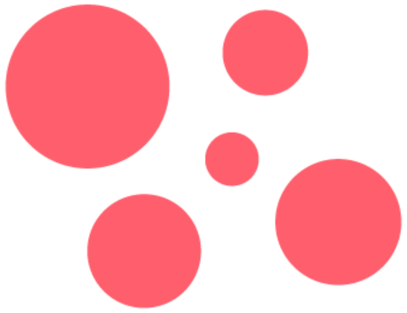


Boxes

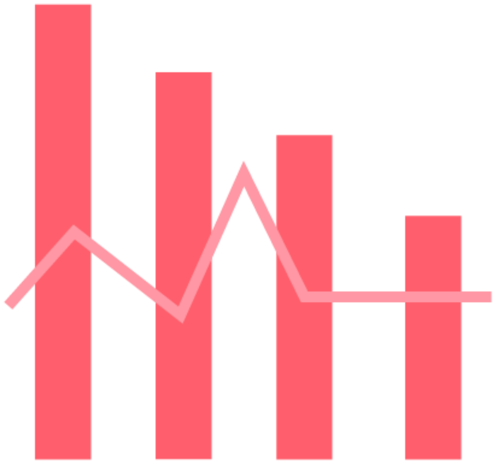


Density

Points clouds



Bubles



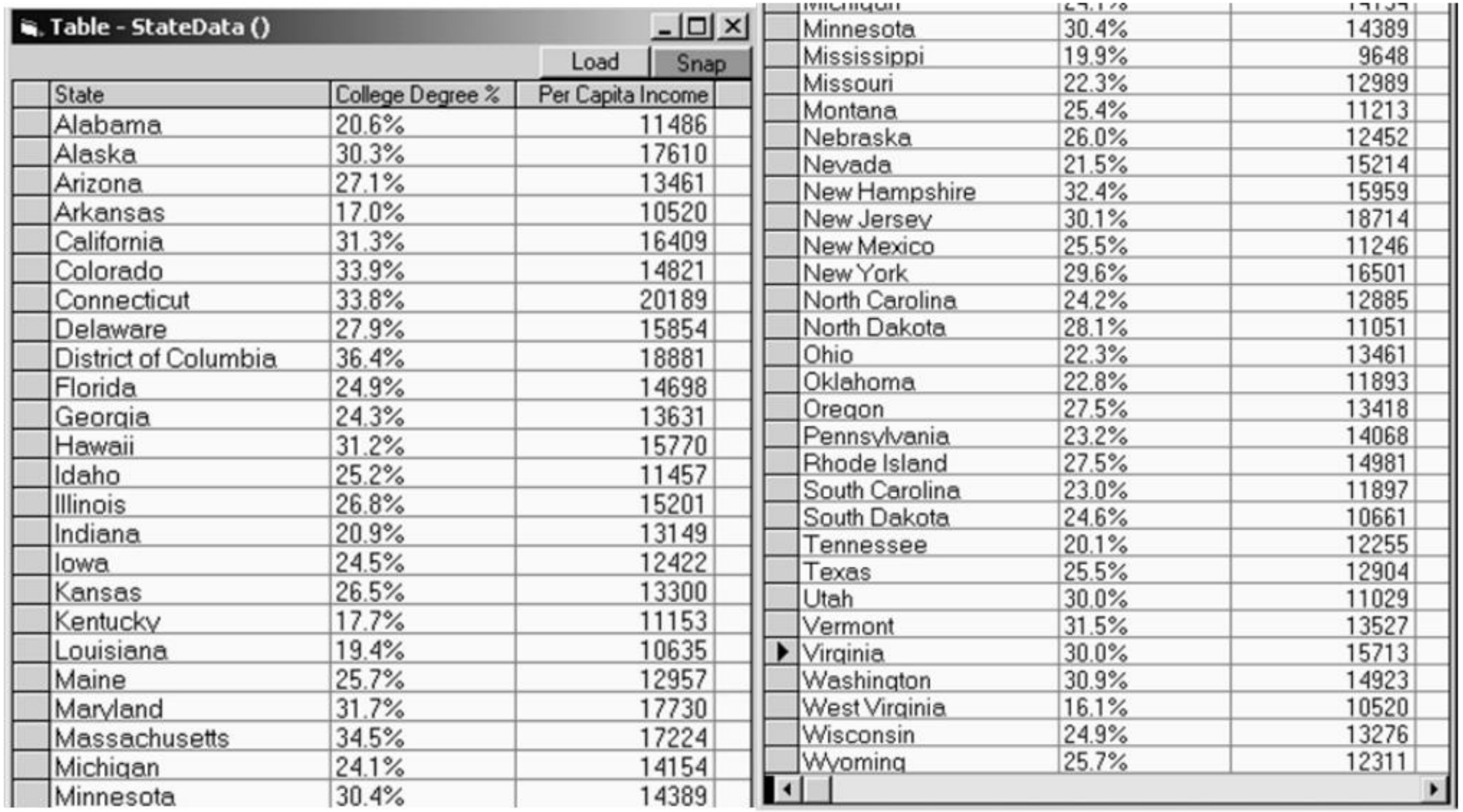
Columns and lines

Heat map

WHY?

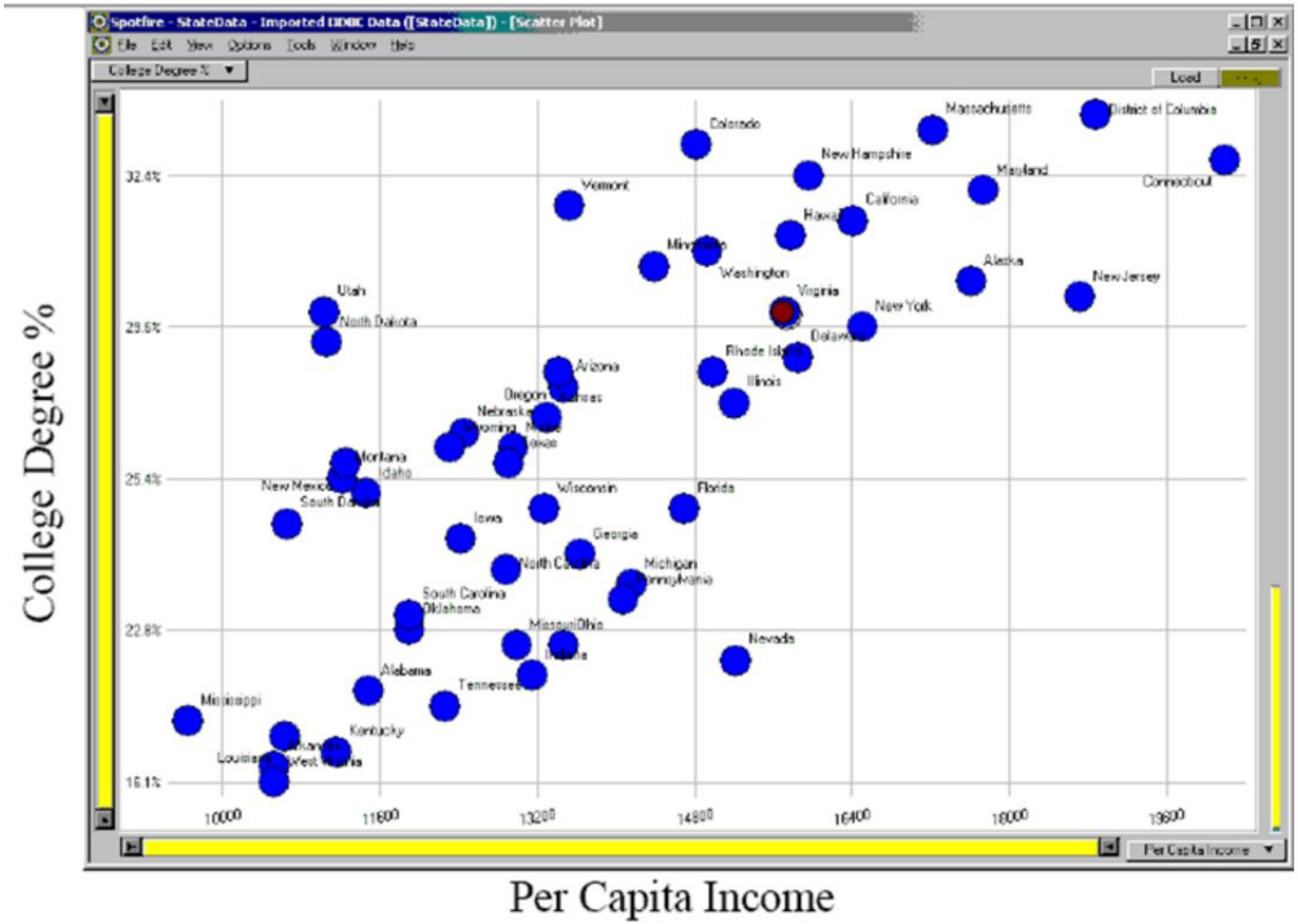
- Reveals invisible parts in data
- Analyze things that are otherwise difficult
- Magnifies ability to understand things
- Help us tell a story
- Efficient way to understand Big Data

- Which state has the largest and the smallest?
- Which states are outliers if any?
- How is income related to college degree?

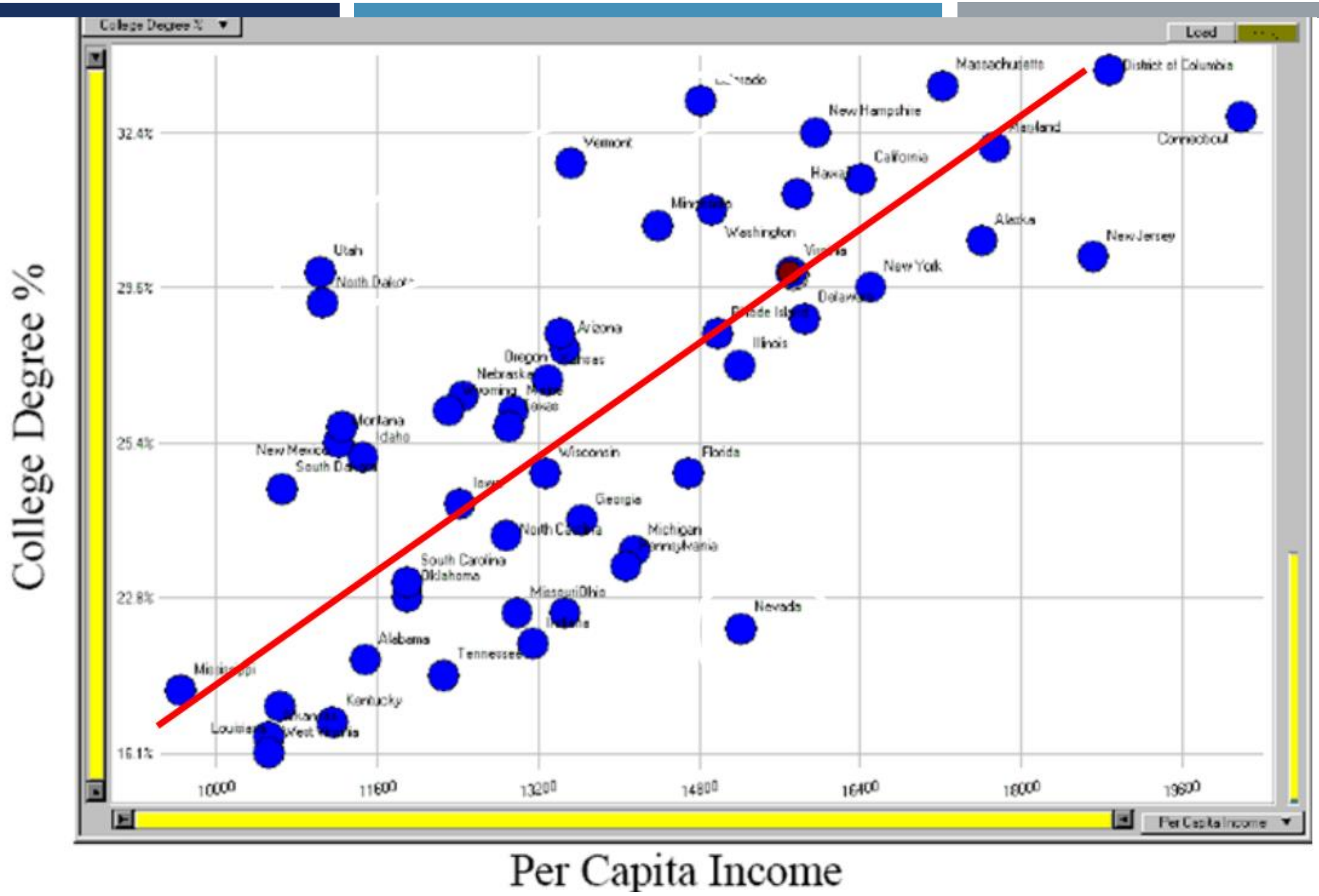


State	College Degree %	Per Capita Income
Alabama	20.6%	11486
Alaska	30.3%	17610
Arizona	27.1%	13461
Arkansas	17.0%	10520
California	31.3%	16409
Colorado	33.9%	14821
Connecticut	33.8%	20189
Delaware	27.9%	15854
District of Columbia	36.4%	18881
Florida	24.9%	14698
Georgia	24.3%	13631
Hawaii	31.2%	15770
Idaho	25.2%	11457
Illinois	26.8%	15201
Indiana	20.9%	13149
Iowa	24.5%	12422
Kansas	26.5%	13300
Kentucky	17.7%	11153
Louisiana	19.4%	10635
Maine	25.7%	12957
Maryland	31.7%	17730
Massachusetts	34.5%	17224
Michigan	24.1%	14154
Minnesota	30.4%	14389
Mississippi	19.9%	9648
Missouri	22.3%	12989
Montana	25.4%	11213
Nebraska	26.0%	12452
Nevada	21.5%	15214
New Hampshire	32.4%	15959
New Jersey	30.1%	18714
New Mexico	25.5%	11246
New York	29.6%	16501
North Carolina	24.2%	12885
North Dakota	28.1%	11051
Ohio	22.3%	13461
Oklahoma	22.8%	11893
Oregon	27.5%	13418
Pennsylvania	23.2%	14068
Rhode Island	27.5%	14981
South Carolina	23.0%	11897
South Dakota	24.6%	10661
Tennessee	20.1%	12255
Texas	25.5%	12904
Utah	30.0%	11029
Vermont	31.5%	13527
Virginia	30.0%	15713
Washington	30.9%	14923
West Virginia	16.1%	10520
Wisconsin	24.9%	13276
Wyoming	25.7%	12311

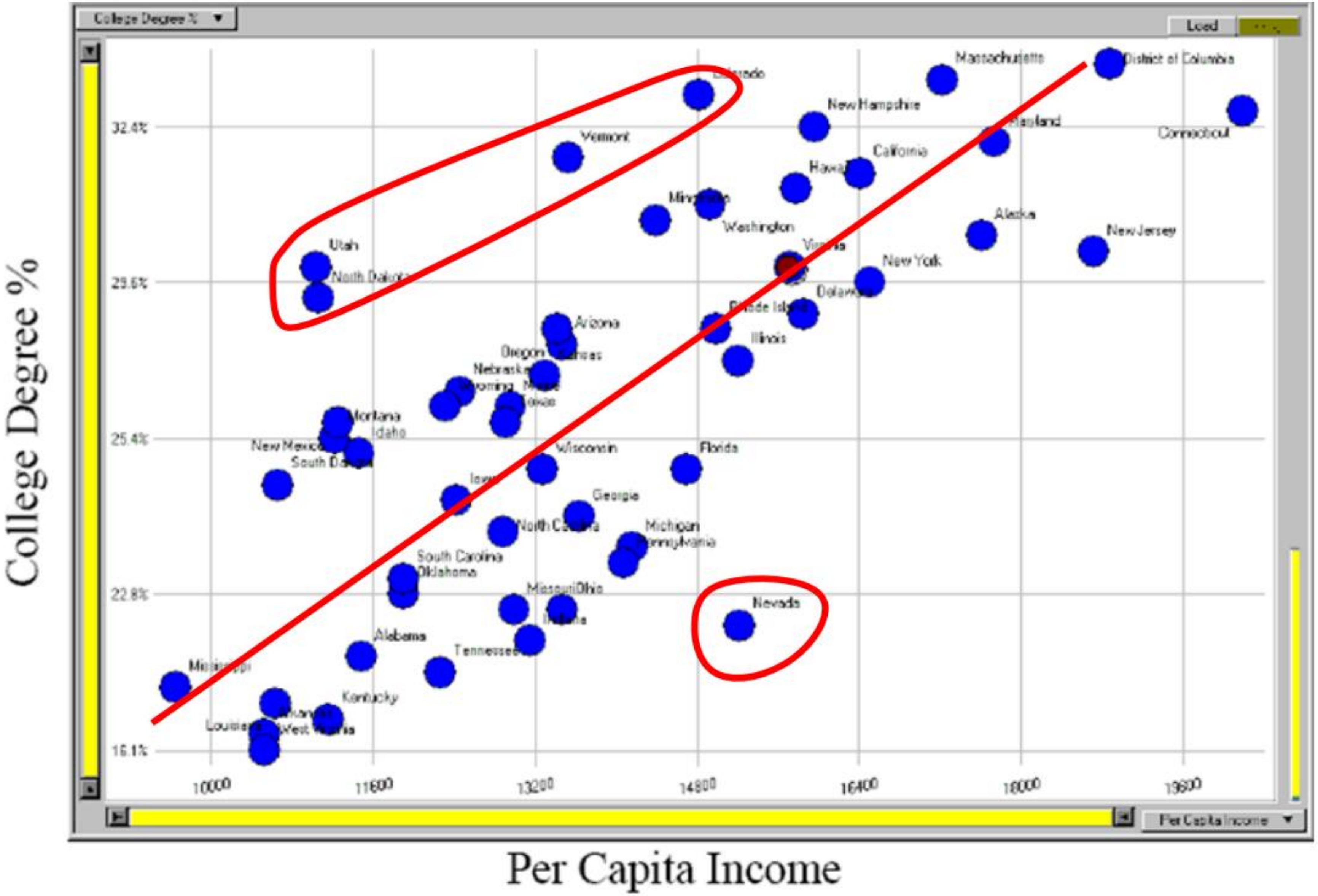
source: Bradley Hemminger, Uni. of North Carolina



CAN EASILY TELL WHAT IS LARGEST/SMALLEST IN EVERY DIMENSION



Visualization helps identify relationship easily as compared to raw data

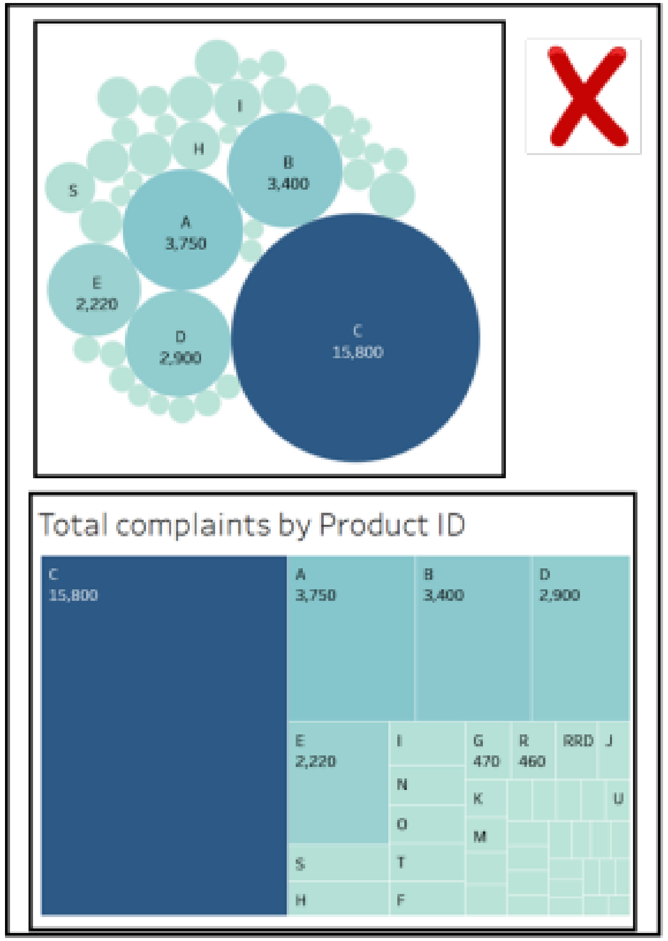


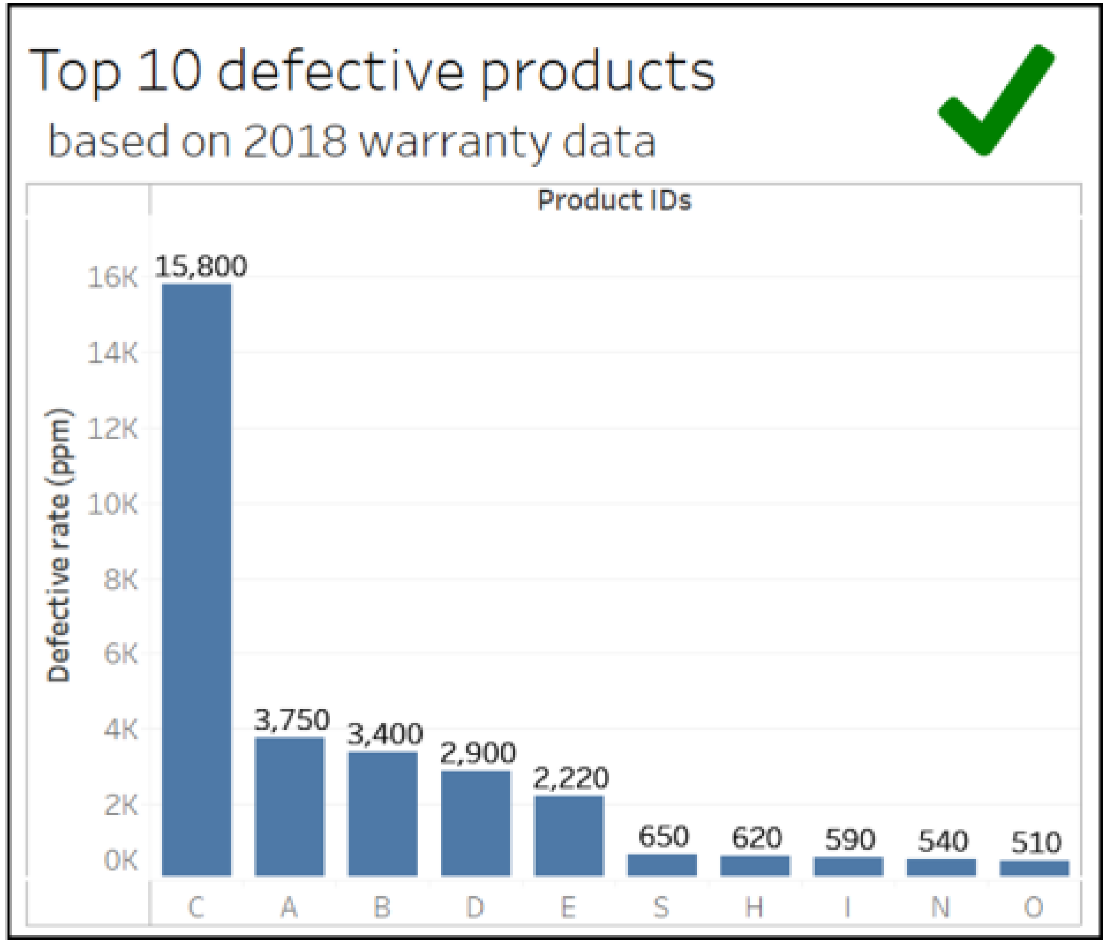
Outliers stand out and get identified easily

USES OF DATA VISUALIZATION

- Decision Making
- Finding solution to problems
- For understanding data clearly
- To find relationship among data
- Comparative analysis

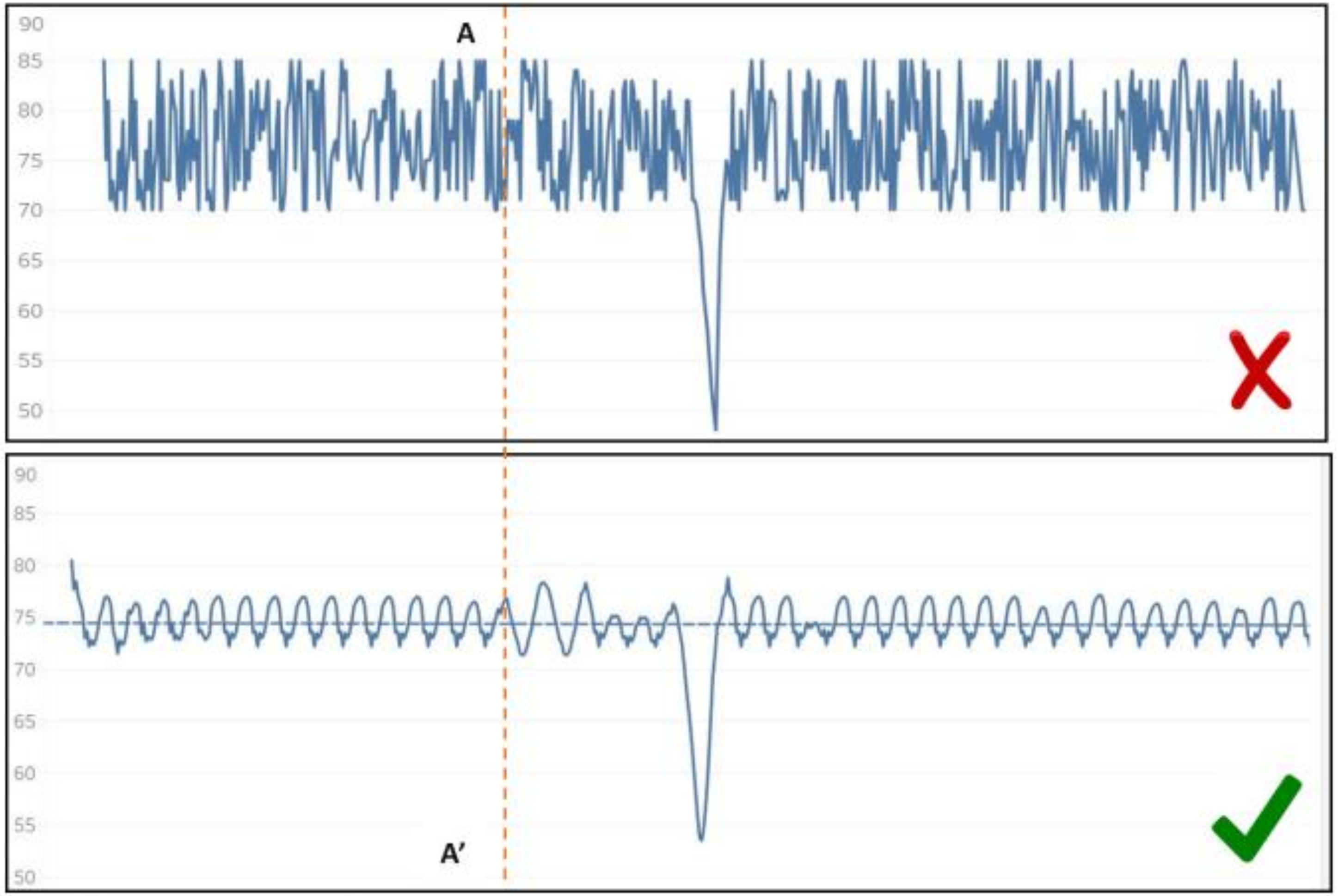
CHOSE RIGHT VISUALS, ACCORDING TO PURPOSE





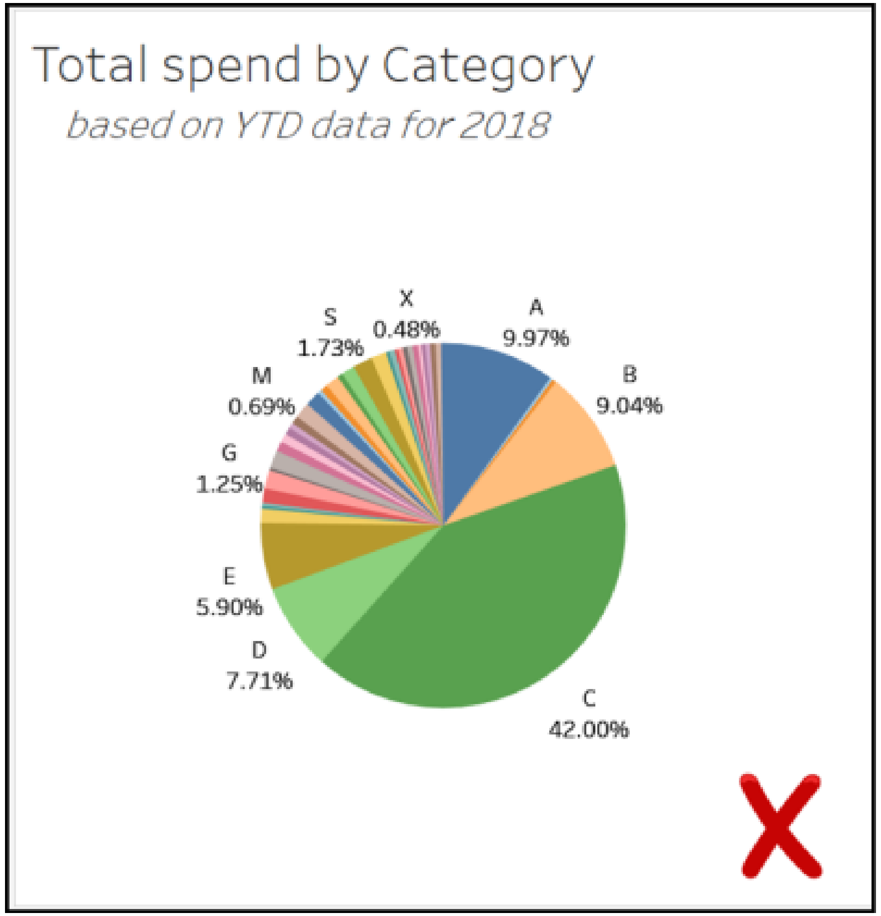
https://towardsdatascience.com/tips-for-effective-data-visualization-d4b2af91db37

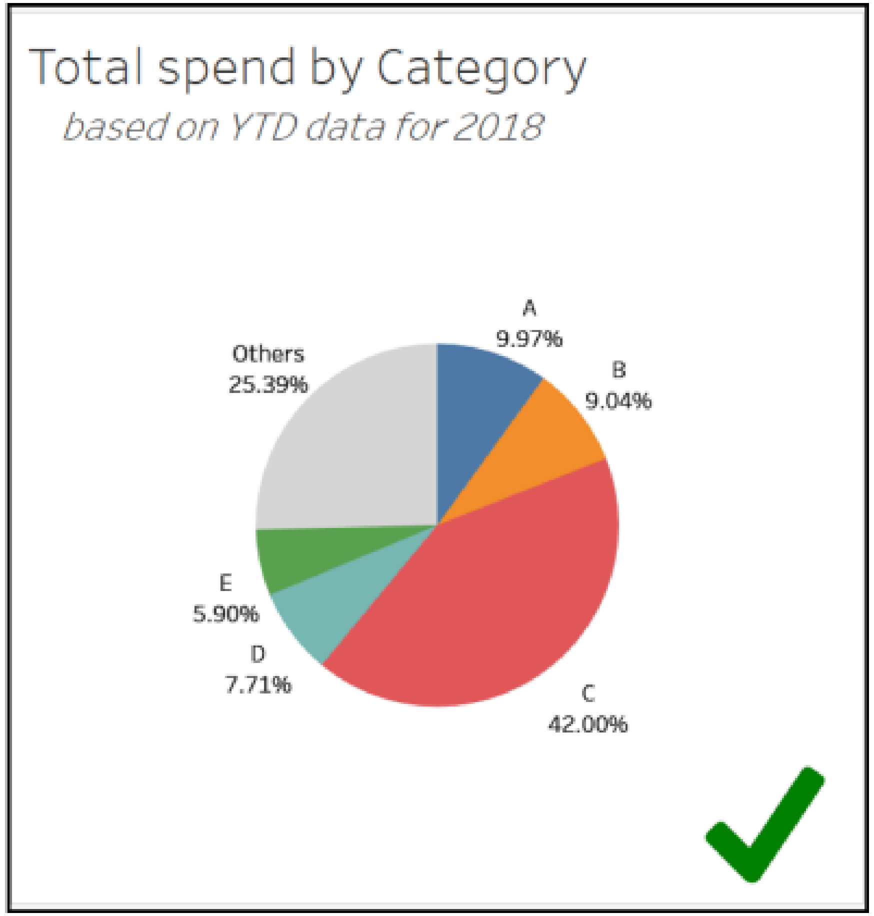
FOCUS ON VITAL DATA POINTS



https://towardsdatascience.com/tips-for-effective-data-visualization-d4b2af91db37

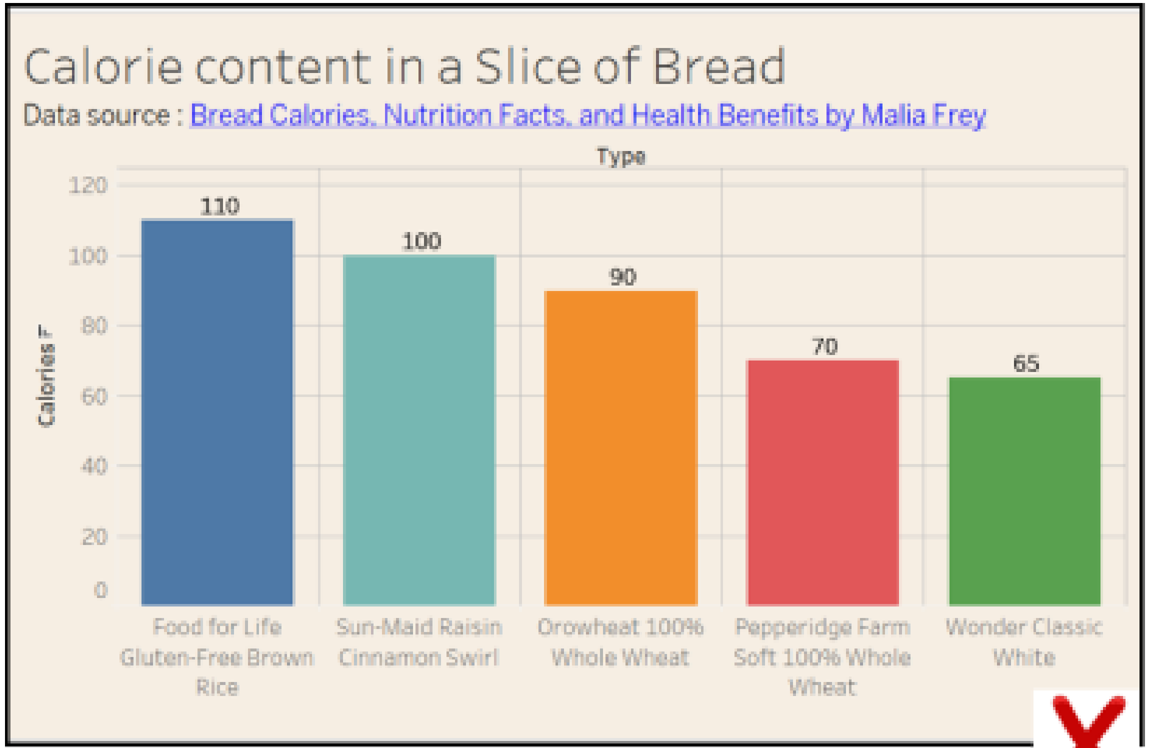
SUPPRESS THE NOISE





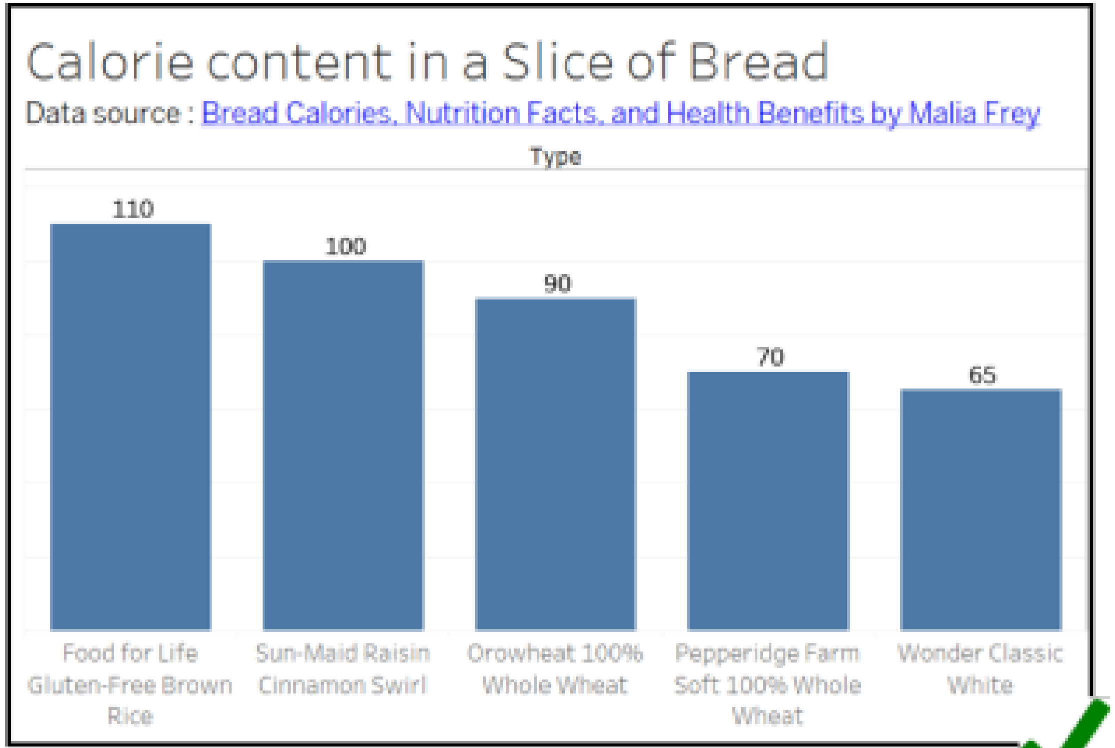
https://towardsdatascience.com/tips-for-effective-data-visualization-d4b2af91db37

USE COLORS WISELY



Calorie content in a Slice of Bread

Data source : Bread Calories, Nutrition Facts, and Health Benefits by Malia Frey

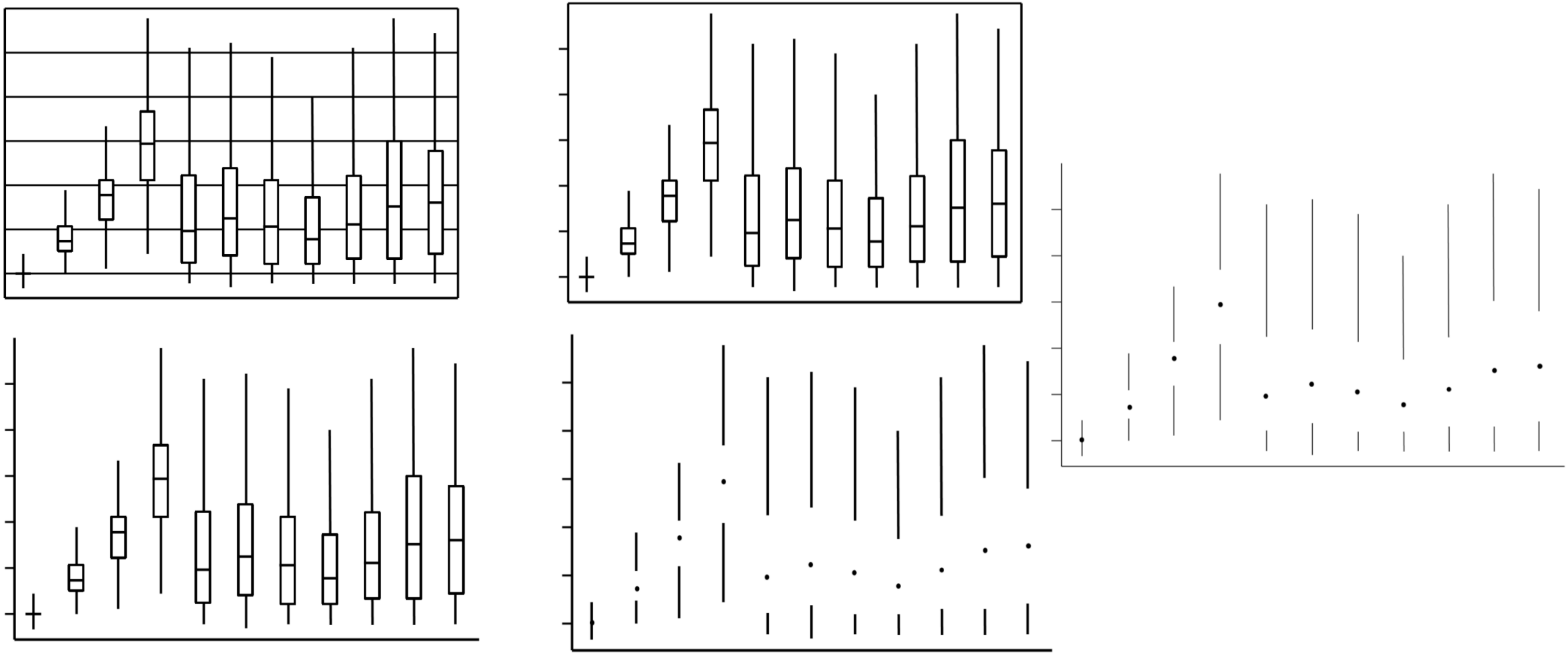


Calorie content in a Slice of Bread

Data source : Bread Calories, Nutrition Facts, and Health Benefits by Malia Frey

https://towardsdatascience.com/tips-for-effective-data-visualization-d4b2af91db37

AVOID UNNECESSARY AESTHETICS



PRINCIPLES OF VISUALIZATION

- Define what questions are you answering
- Use accurate data
- Experiment with ways to answer
- Go with cognitive research
- Faithfully represent your data
- Tailor it to your audience
- Make it as simple as possible
- Remove everything that you can