Introduction to Data Visualization with R using ggplot2

Richard Johansen

Zhiyuan Yao

Jennifer Latessa

University of CINCINNATI

Workshop Agenda

Workshop Expectations
Understanding Data
Visualizations
Ggplot2
The Scenario

Workshop Agenda

Workshop Expectations

Understanding Data

Visualizations

Ggplot2

The Scenario

Workshop Expectations

- Prerequisites
  - R and R studio Installed
  - Install the ggplot2 package
- Goal
  - Conduct basic data exploration and data visualization
  - Allow you to (re)produce print-quality graphics in seconds

Workshop Agenda

Workshop Expectations

Understanding Data

Visualizations

Ggplot2

The Scenario

What is Data?

- Data is a collection of **objects** defined by **attributes**
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Synonyms: variables, fields, characteristics, features, columns, etc.
- A collection of attributes describe an object
  - Synonyms: records, points, cases, samples, instances, rows, etc.

Attribute

Object

	A	B	C	D	E
1	College Enrollment 2018-2019				
2					
3	Student ID	Last Name	Initial	Age	Program
4	ST348-245	Walton	L.	21	Drafting
5	ST348-246	Wilson	R.	19	Science
6	ST348-247	Thompson	G.	18	Business
7	ST348-248	James	L.	23	Nursing
8	ST348-249	Peterson	M.	37	Science
9	ST348-250	Graham	J.	20	Arts
10	ST348-251	Smith	F.	26	Business
11	ST348-252	Nash	S.	22	Arts
12	ST348-253	Russell	W.	19	Nursing
13	ST348-254	Robitaille	L.	20	Drafting

Attribute Values

- Each attribute has a potential set of values objects draw from.
- The same attribute can be mapped to different attribute values
  - Example: height can be measured in meters or feet
- Different attributes can be mapped to the same set of values
  - Example: Attribute values for ID and age are both integers

	A	B	C	D	E
1	College Enrollment 2018-2019				
2					
3	Student ID	Last Name	Initial	Age	Progr
4	ST348-245	Walton	L.	21	Drafting
5	ST348-246	Wilson	R.	19	Science
6	ST348-247	Thompson	G.	18	Business
7	ST348-248	James	L.	23	Nursing
8	ST348-249	Peterson	M.	37	Science
9	ST348-250	Graham	J.	20	Arts
10	ST348-251	Smith	F.	26	Business
11	ST348-252	Nash	S.	22	Arts
12	ST348-253	Russell	W.	19	Nursing
13	ST348-254	Robitaille	L.	20	Drafting

Attribute Classification

- Discrete Attribute – has an infinite or countably infinite set of values
  - Nominal - Data that can be counted, but not aggregated or ordered
    - Examples: Eye Color, Zip Code, Music Genre
  - Ordinal - Data that can be counted and ordered, but not aggregated.
    - Examples: Grades, Clothing Size, Positions (in a race)

- Continuous Attribute - has real numbers as attribute values
  - Interval (metrics) - The difference in values are constant and meaningful
    - Examples: The difference between a temperature of 100°F and 90°F is the same difference as between 90°F and 80°F.
  - Ratio - An interval scale with an absolute zero
    - Examples: Income, Height, Weight

Data Set Classification

- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data

Record Data

- Data that consist of a collection of records, each which consists of a fixed set of attributes (Tables)
  - Data Matrix – Entirely continuous numerical data
    - Can be plotted in multi-dimensional space (each dimension is an attribute)
  - Document Data – Each object is a “term” vector (count)
  - Transaction Data – Each record is a set of items (transactions)

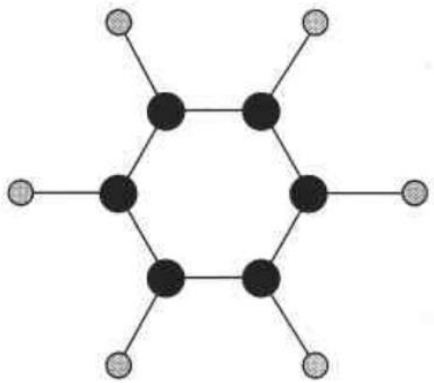
Projection of x Load	Projection of y load	Distance	Load
10.23	5.27	15.22	2.7
12.65	6.25	16.22	2.2

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0

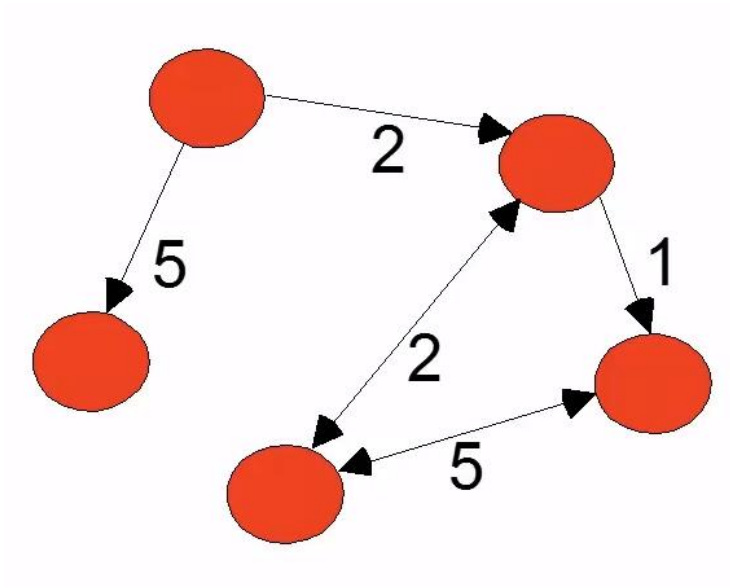
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Graph Data

- World Wide Web
  - Nodes, Edges, Direction, Weight
- Molecular Structures



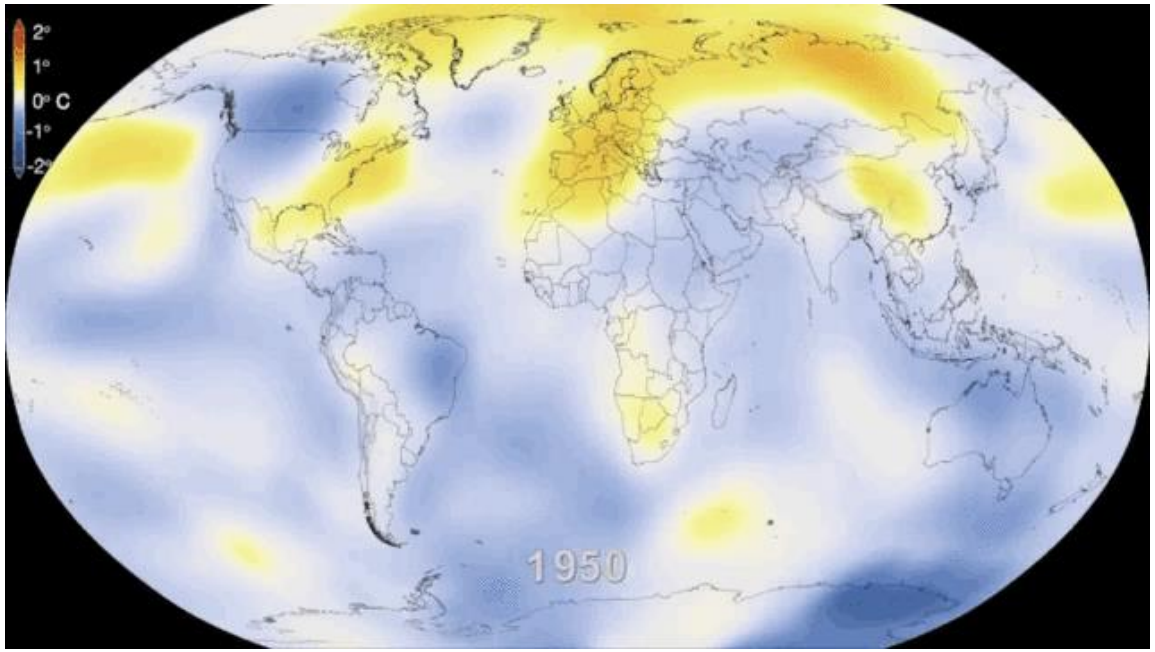
Benzene molecule



https://towardsdatascience.com/types-of-data-sets-in-data-science-data-mining-machine-learning-eb47c80af7a

Ordered Data

- Spatial and Temporal Data
  - Global Temperature
- Sequential Data
  - Genetic Sequence
    - GAGAAGGCCTTCCC



https://www.albert.io/learn/ngss-earth-space-sciences/hsess36-anthropogenic-impact-on-systems/analyzing-human-contributions-to-carbon-dioxide-levels-and-ocean-acidification/global-temperature-and-carbon-dioxide-levels?page=1

Data Quality

- Main Issues with Data Quality:
  - Noise and Outliers
  - Missing Values
  - Duplicate Data

A mistake or a millionaire?

Missing values

Inconsistent duplicate entries

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
9	No	Single	90K	No

Data Preprocessing Exploratory Analysis Post-processing Published by[Milo Cunliffe]

Primitive Data Types

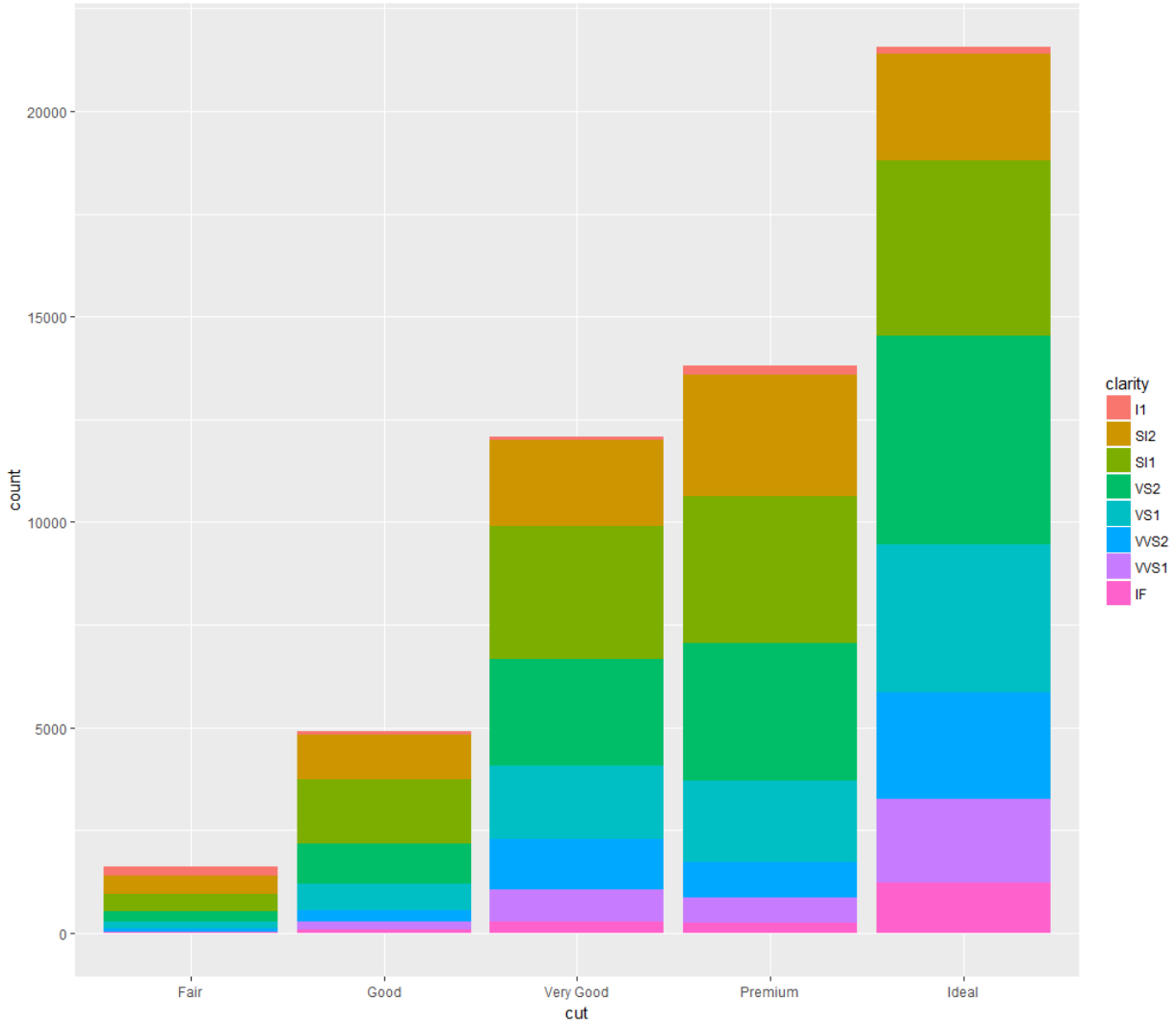
Primitive data types are predefined **types** of **data**, which are supported by the programming language.

- Boolean:
  - True (T) or False (F)
- Char:
  - Characters and Strings – “A”, “Beta”, “There are different data types!”
- Factors:
  - Ordinal Data – 1st, 2nd, 3rd Or High, Medium, Low
- Int:
  - Integers – (1, 2, 100)
- Float/Double:
  - Decimal – (0.1, 0.2, 0.1352)

Workshop Agenda

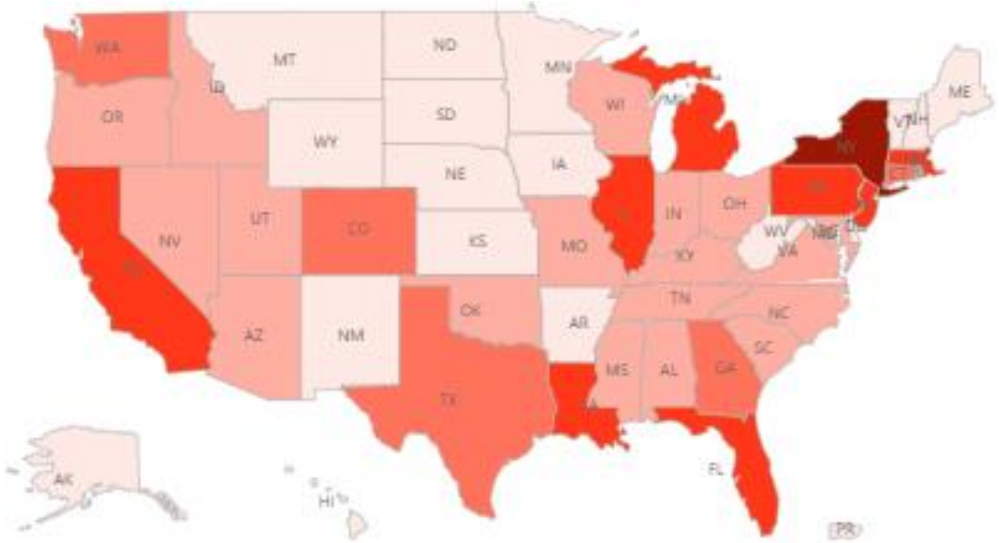
Workshop Expectations
Understanding Data
Visualizations
The Scenario
Ggplot2

Why is Data Visualization Important?



Why Do Visualization

- Reasons for doing data visualization
- Exploration
  - Use visualizations as a means of data exploration
- Analysis
  - Verify (or Falsify) a hypothesis
- Presentation
  - Visualization is used to communicate results or findings



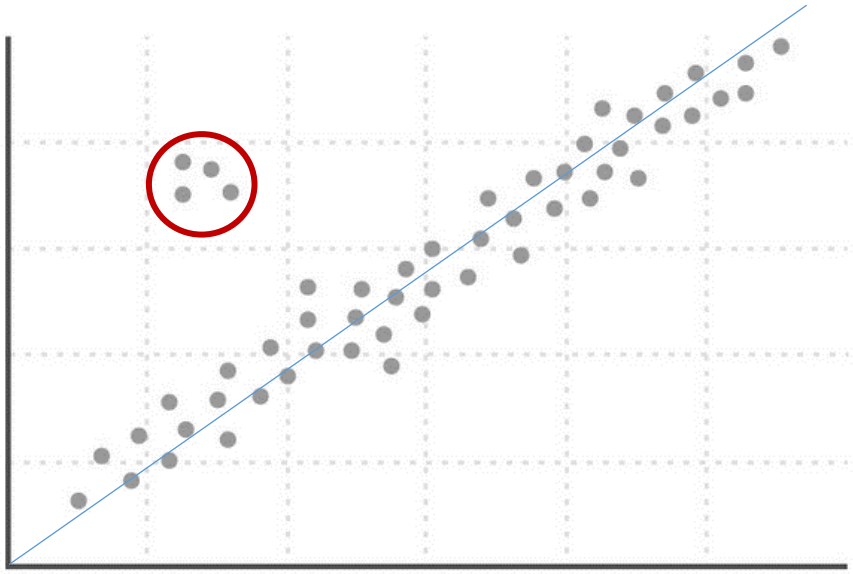
Why Do Visualization

- Reasons for doing data visualization

- Exploration
  - Use visualizations as a means of data exploration

- Analysis
  - Verify (or Falsify) a hypothesis

- Presentation
  - Visualization is used to communicate results or findings



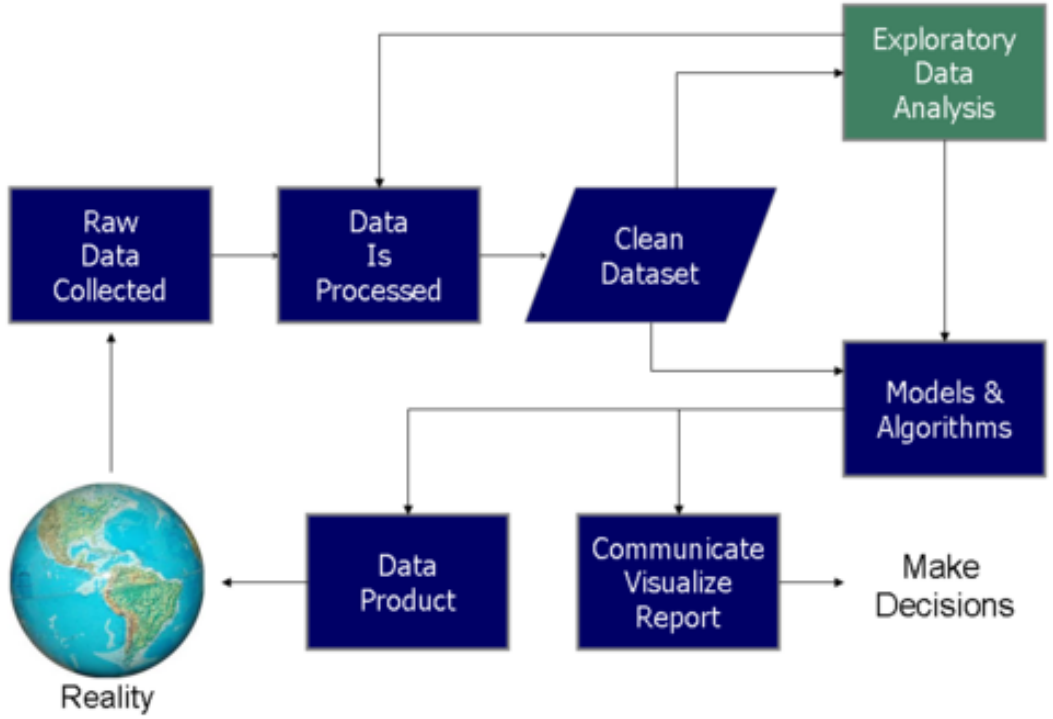
Why Do Visualization

- Reasons for doing data visualization

- Exploration
  - Use visualizations as a means of data exploration
- Analysis
  - Verify (or Falsify) a hypothesis
- Presentation
  - Visualization is used to communicate results or findings

Data Science Process

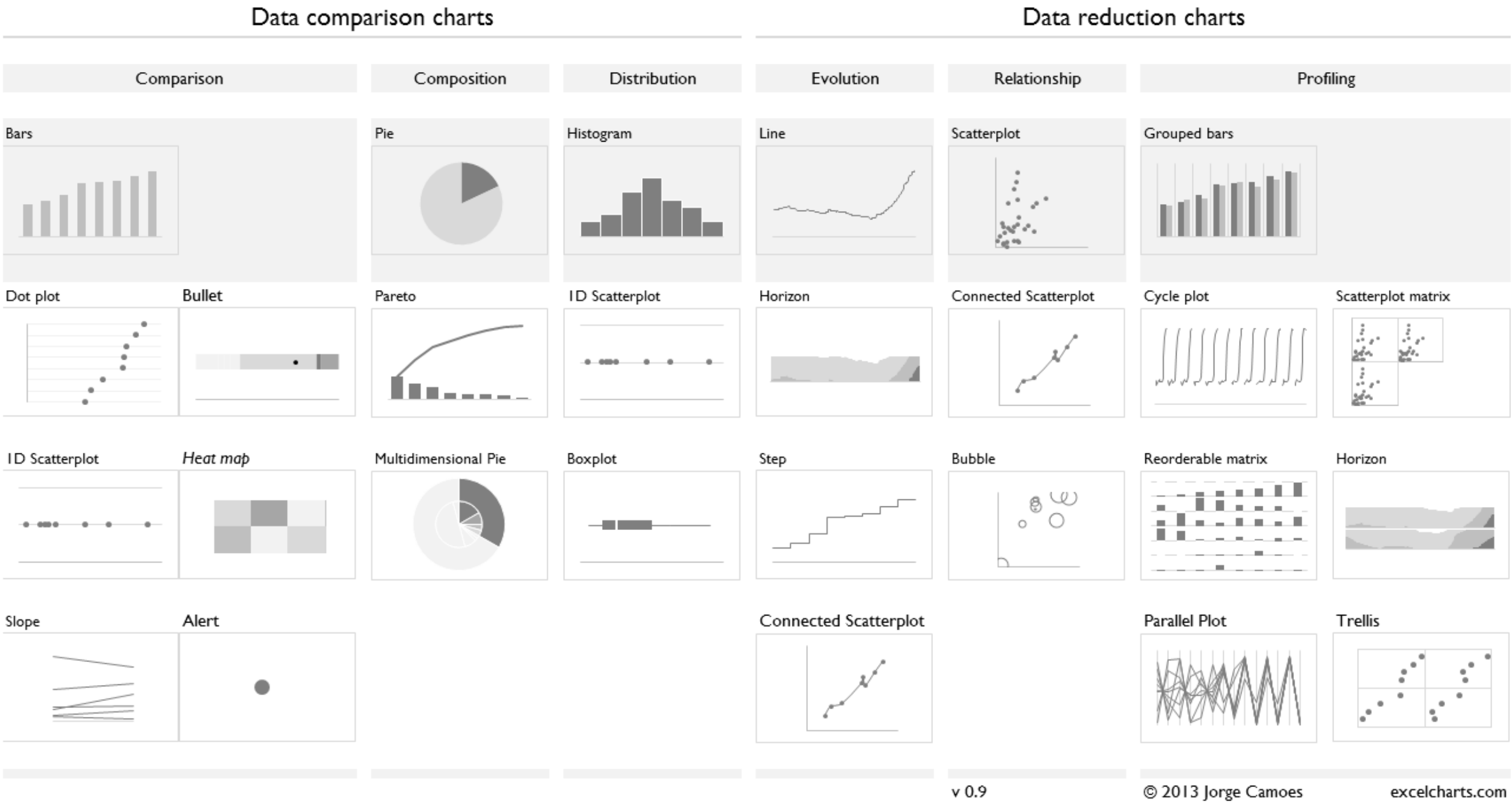
- Acquire Data
  - "Know your data"
- Clean and Pre-Process
- Visualize (explore)
- Model/Analyze
- Communicate Findings/ Data Production



https://commons.wikimedia.org/wiki/File:Data_visualization_process_v1.png

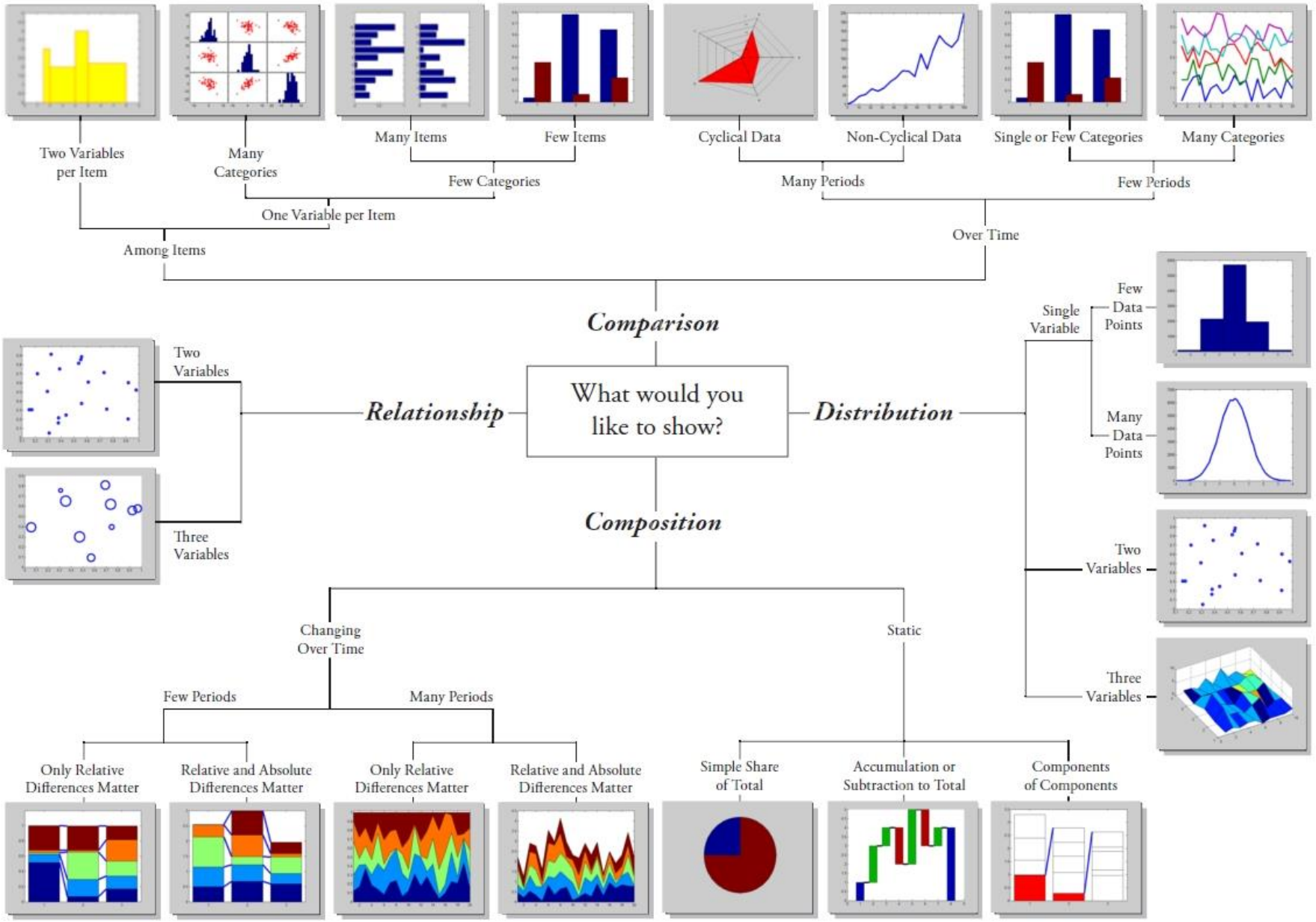
Chart Types

- Common Chart Types
  - Comparison - comparing and sorting data points;
  - Composition - part-to-whole comparisons;
  - Distribution - comparison of data points along an axis;
  - Relationship - relationship patterns between two or more variables;



https://excelcharts.com/classification-chart-types/

Chart Suggestions—A Thought-Starter



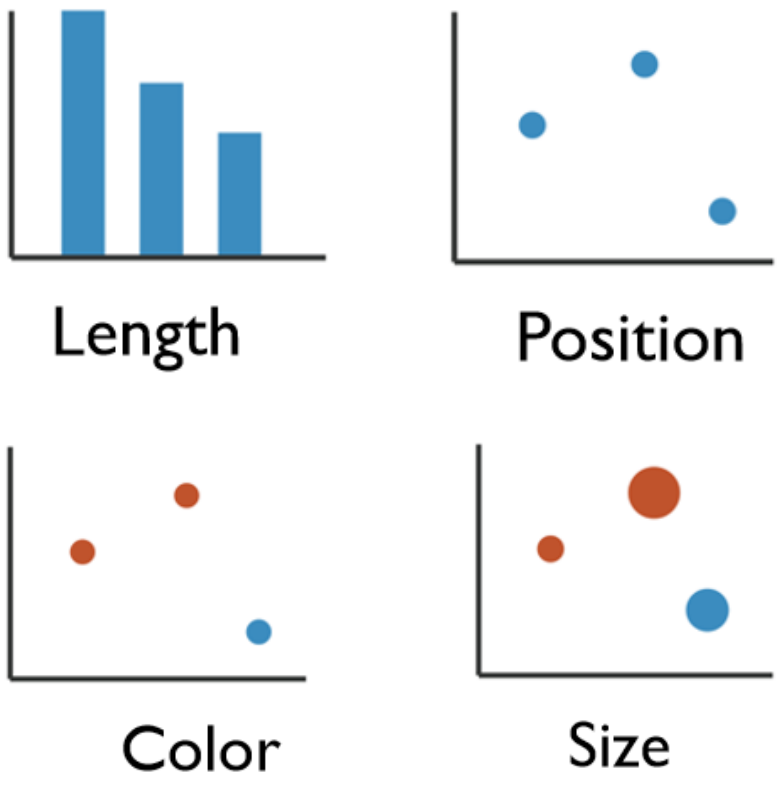
Modified with permision -Doug Hull
blogs.mathworks.com/videos
hull@mathworks.com 2009

www.ExtremePresentation.com
© 2009 A. Abela — a.v.abela@gmail.com

https://datavizblog.files.wordpress.com/2013/04/andrew-abela-chart-chooser-in-color.jpg

Enhancing Visualizations

- 1 Dimensional Data
  - Length
- 2 Dimensional Data
  - Position
- 2+ Dimensional Data
  - Position
  - Color Hue/Saturation
  - Size
  - Shape



Workshop Agenda

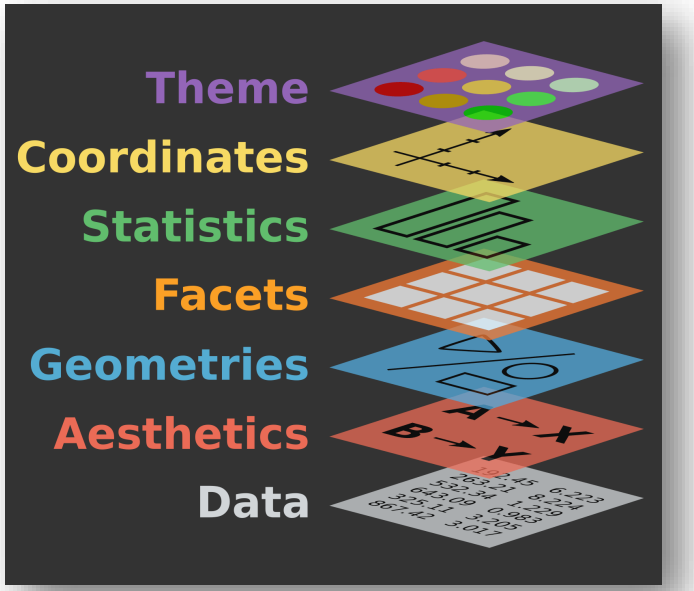
Workshop Expectations
Understanding Data
Visualizations
Ggplot2
The Scenario

Ggplot2 Full Template

ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics

Full template

- # ggplot(data = <DATA>) +
- # <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>),
- # stat = <STAT>,
- # position = <POSITION>) +
- # <COORDINATE_FUNCTION> +
- # <FACET_FUNCTION> +
- # <THEME_FUNCTION>



Grammar of Graphics

Originated by Leland Wilkinson, simplified by Hadley Wickham and others.

Describe all the non-data ink

Plotting space for the data

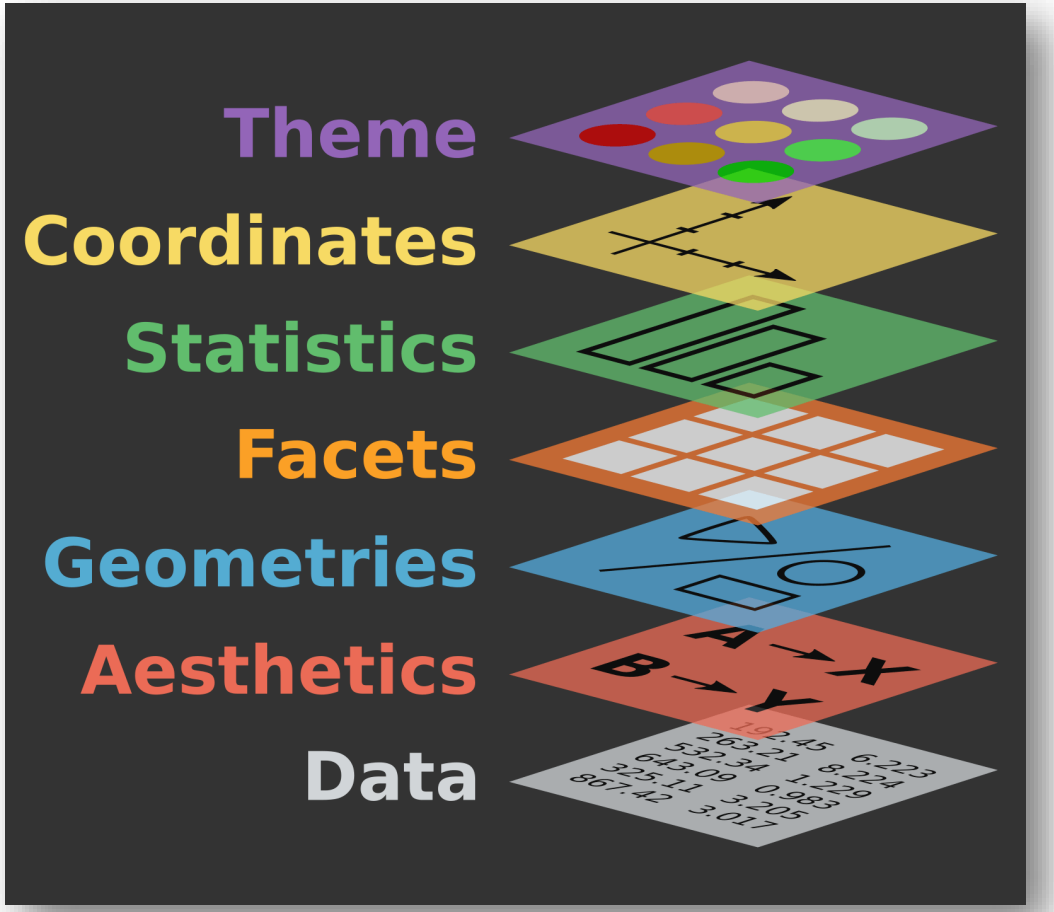
Statistical models & summaries

Rows and columns of sub-plots

Shapes used to represent the data

Scales onto which data is mapped

The actual variables to be plotted

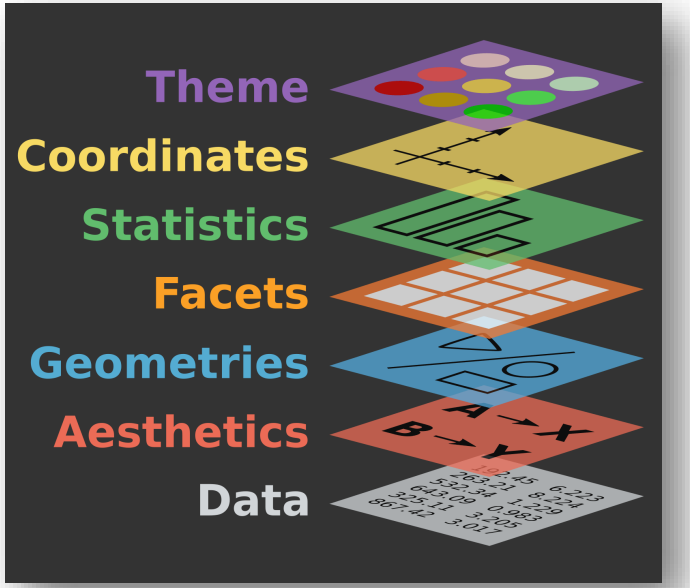


The Basics

- Data – The raw materials of your visualization

- Aesthetics – The mapping of your data to the visualization
  - X-axis is age
  - Y-axis is survival

- Geometrics – Any visualization requires at least one layer and in ggplot2 these are typically the geoms.
  - Example a barchart is geom_bar()



Workshop Agenda

Workshop Expectations
Understanding Data
Visualizations
Ggplot2
The Scenario

Titanic Data Set

- We will use the Kaggle Competition's Titanic Machine Learning from Disaster Dataset
  - Everyone is familiar with the Titanic
  - The data set is a good representation of real world data

- Following the teaching model from Dave Langer's presentation on Data Science Dojo

Titanic Data Dictionary

Variables:

- Survival – Survival (yes=1, no=0)
- Pclass – Ticket Class (1st class, 2nd class)
- Sex – Gender (Male or Female)
- Age – Passenger age
- Sibsp – # of Siblings/Spouse
- Parch – # of Parents/Children
- Ticket – Ticket Number
- Fare – Passenger Fare
- Cabin – Cabin Number
- Embarked – Port of Embarkation

Your Job

- You are hired as a consultant and have been tasked with analyzing the titanic data set.

- Your goal is to explore patterns and trends to explain what influenced the survival rate of the passengers on the Titanic.

Questions?

Contact the Visualization Laboratory

Email: AskData@uc.edu

Web: https://libraries.uc.edu/research-teaching-support/research-data-services.html

Visit: 240 Braunstein Hall (Geology-Math-Physics Library)

University of CINCINNATI