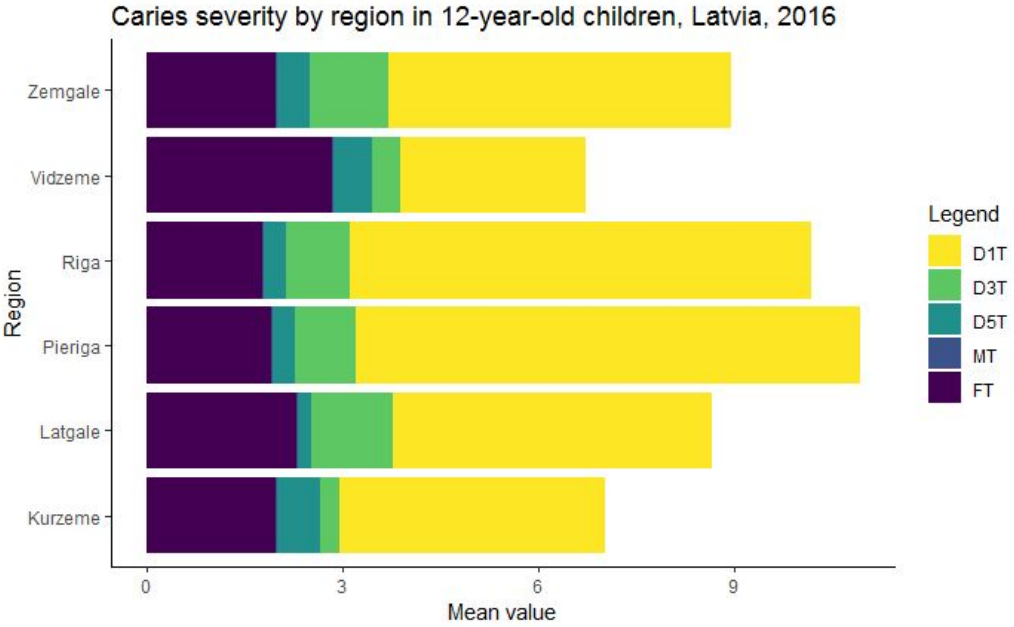
Introduction to Data Visualization for Research



Sergio Uribe, DDS, MSc, PhD
Assoc Professor
Universidad Austral de Chile

Objectives

Define data visualization

Identify the basic principles of data visualization with examples (good and bad)

Uses of data visualization

How to prepare the dataset for visualization

Use R and ggplot package for data visualization

Audience

Anyone who wants to learn principles of good data visualization

What is Data Visualization?

Visual Representation of Data

Useful for exploration

discovery

insight

What is Data Visualization?

Visual Representation of Data

For exploration, discovery, insight

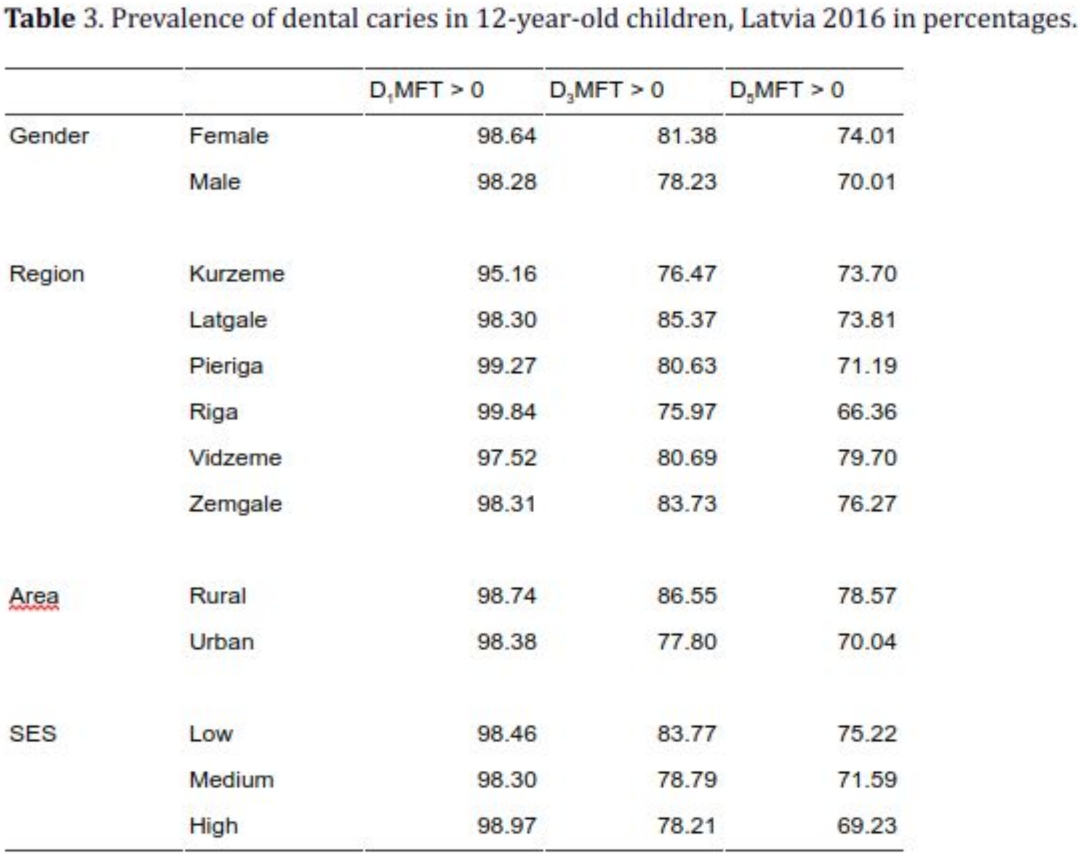


Table 3. Prevalence of dental caries in 12-year-old children, Latvia 2016 in percentages.

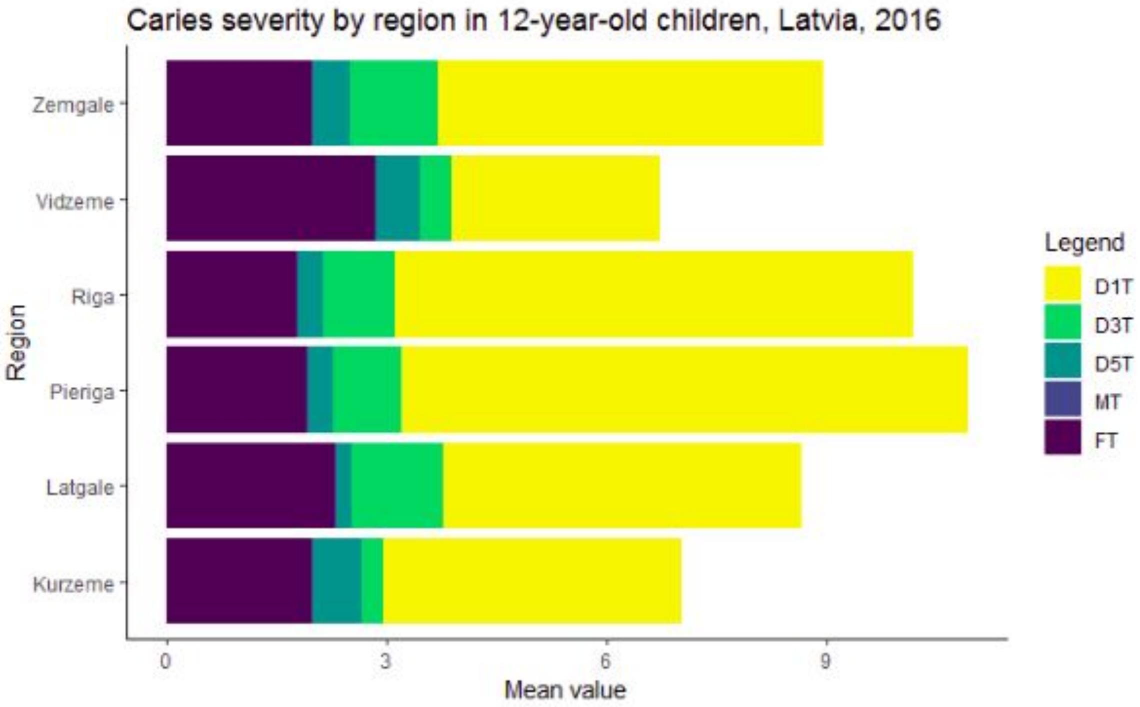
		$D_1MFT > 0$	$D_3MFT > 0$	$D_5MFT > 0$
Gender	Female	98.64	81.38	74.01
	Male	98.28	78.23	70.01
Region	Kurzeme	95.16	76.47	73.70
	Latgale	98.30	85.37	73.81
	Pieriga	99.27	80.63	71.19
	Riga	99.84	75.97	66.36
	Vidzeme	97.52	80.69	79.70
	Zemgale	98.31	83.73	76.27
Area	Rural	98.74	86.55	78.57
	Urban	98.38	77.80	70.04
SES	Low	98.46	83.77	75.22
	Medium	98.30	78.79	71.59
	High	98.97	78.21	69.23

What is Data Visualization?

Visual Representation of Data

For exploration, discovery, insight

Figure 1. Caries severity by region per tooth in 12-year-old children, Latvia 2016.

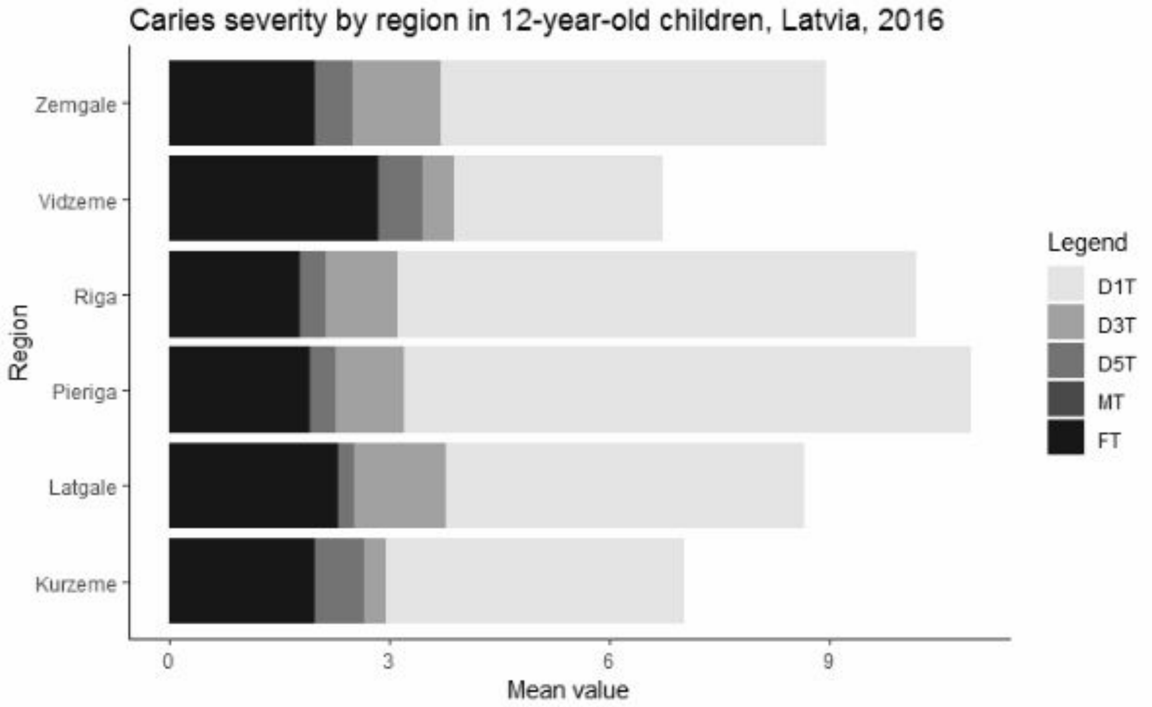


What is Data Visualization?

Visual Representation of Data

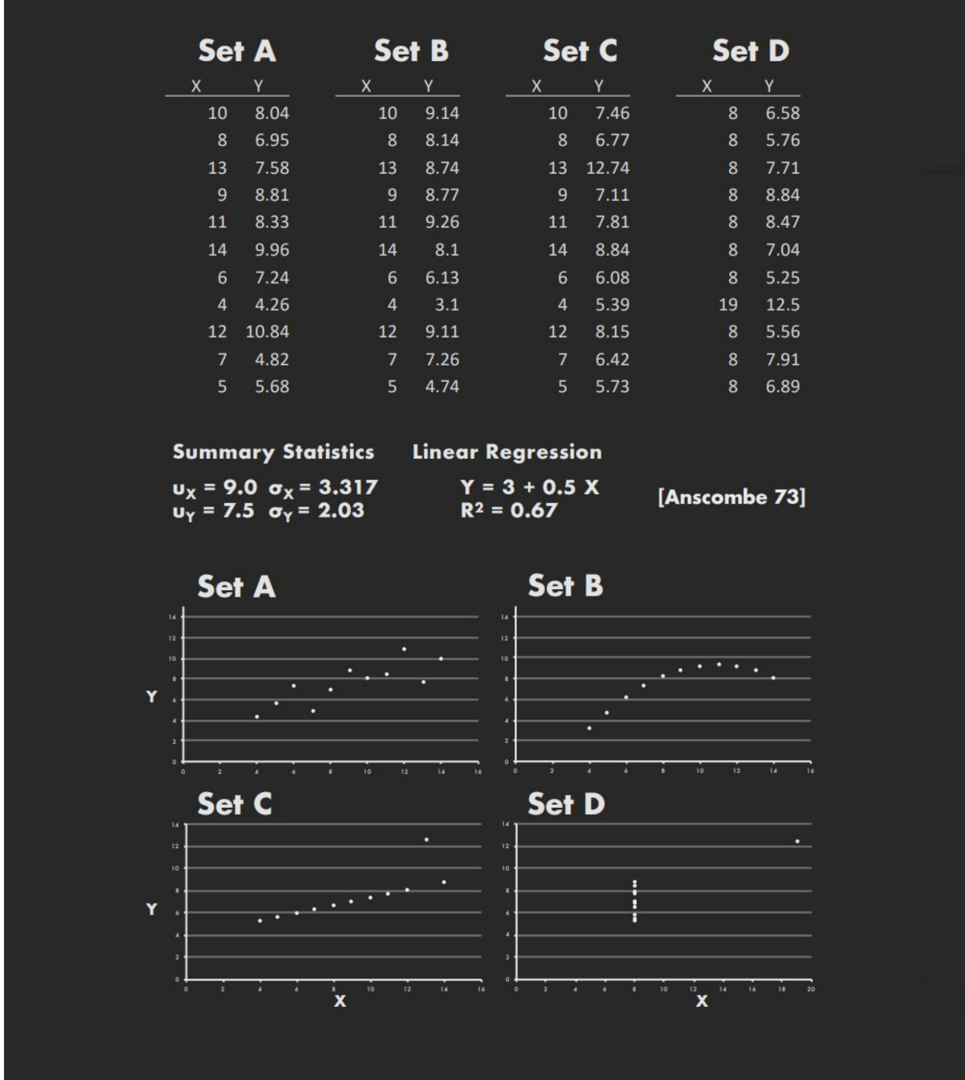
For exploration, discovery, insight

Figure 1. Caries severity by region per tooth in 12-year-old children, Latvia 2016.



Why visualize?

to understand data



Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Summary Statistics

$u_X = 9.0$ $\sigma_X = 3.317$
$u_Y = 7.5$ $\sigma_Y = 2.03$

Linear Regression

$Y = 3 + 0.5\ X$
$R^2 = 0.67$

[Anscombe 73]

Three principles for visualization

1. **be true to your research** – design your display to illustrate a particular point
2. **maximize information, minimize ink** –use the simplest possible representation for the bits you want to convey
3. **organize hierarchically** – what should a viewer see first? what if they look deeper?

Exploratory Data Analysis: Visualization

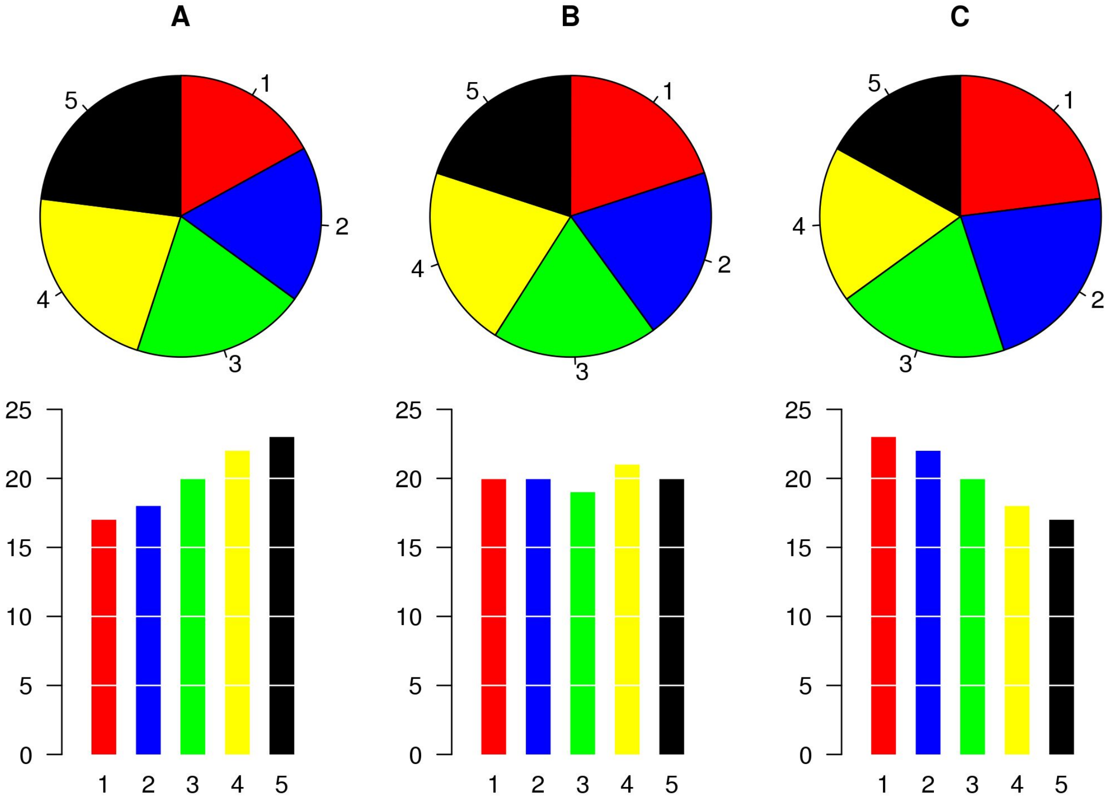
Distribution

Compare

Change

Association

Don’t use pie charts



Get the right tools

Introduction to R

Install R

Install R Rstudio

Components of RStudio

Install libraries

Load libraries

Load data

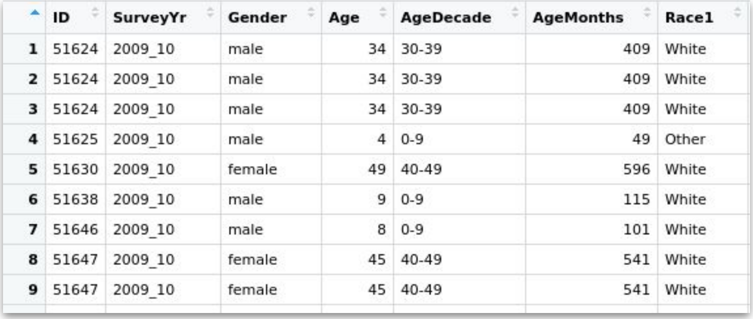
Exploratory Data Analysis: basic commands

View the dataset	`View(dataset)`
Summaries of data	`summary(dataset)`
Create simple tables	`table(dataset$columnA, dataset$columnB)`
See column names	`names(dataset)`
Take a look	`glimpse(dataset)`

ggplot2

Grammar of **G**raphics plot

Data

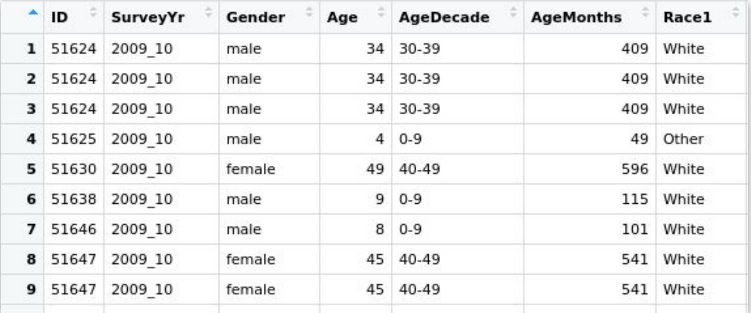


	ID	SurveyYr	Gender	Age	AgeDecade	AgeMonths	Race1
1	51624	2009_10	male	34	30-39	409	White
2	51624	2009_10	male	34	30-39	409	White
3	51624	2009_10	male	34	30-39	409	White
4	51625	2009_10	male	4	0-9	49	Other
5	51630	2009_10	female	49	40-49	596	White
6	51638	2009_10	male	9	0-9	115	White
7	51646	2009_10	male	8	0-9	101	White
8	51647	2009_10	female	45	40-49	541	White
9	51647	2009_10	female	45	40-49	541	White

Aesthetic
(x, y)

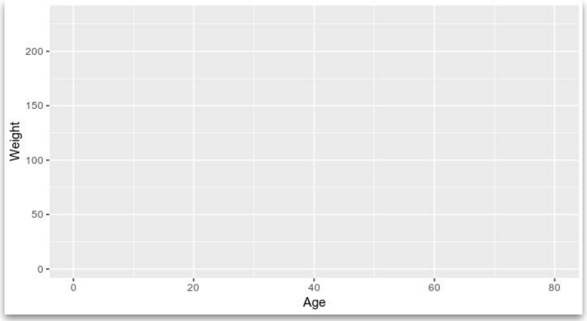
Geometry

Data



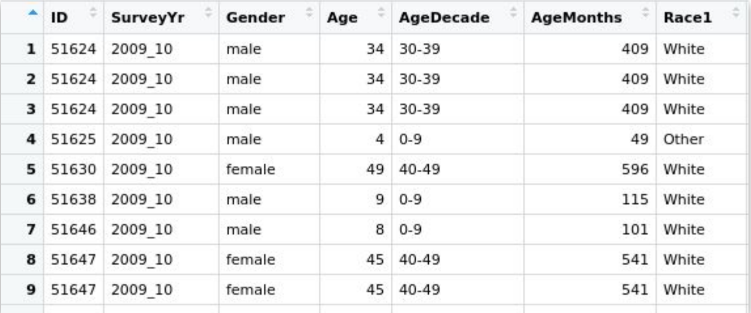
	ID	SurveyYr	Gender	Age	AgeDecade	AgeMonths	Race1
1	51624	2009_10	male	34	30-39	409	White
2	51624	2009_10	male	34	30-39	409	White
3	51624	2009_10	male	34	30-39	409	White
4	51625	2009_10	male	4	0-9	49	Other
5	51630	2009_10	female	49	40-49	596	White
6	51638	2009_10	male	9	0-9	115	White
7	51646	2009_10	male	8	0-9	101	White
8	51647	2009_10	female	45	40-49	541	White
9	51647	2009_10	female	45	40-49	541	White

Aesthetic
(x, y)



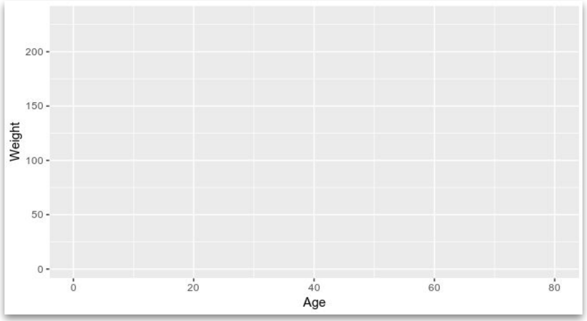
Geometry

Data

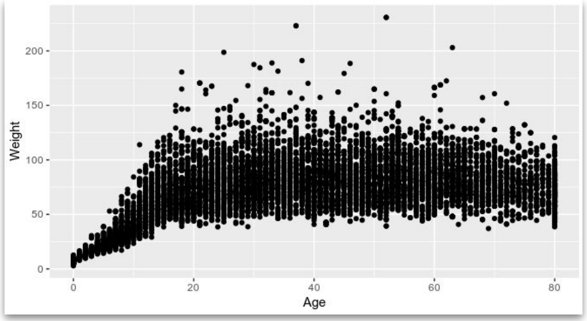


	ID	SurveyYr	Gender	Age	AgeDecade	AgeMonths	Race1
1	51624	2009_10	male	34	30-39	409	White
2	51624	2009_10	male	34	30-39	409	White
3	51624	2009_10	male	34	30-39	409	White
4	51625	2009_10	male	4	0-9	49	Other
5	51630	2009_10	female	49	40-49	596	White
6	51638	2009_10	male	9	0-9	115	White
7	51646	2009_10	male	8	0-9	101	White
8	51647	2009_10	female	45	40-49	541	White
9	51647	2009_10	female	45	40-49	541	White

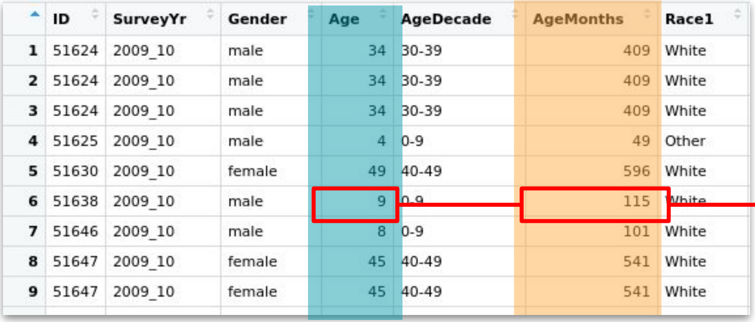
Aesthetic
(x, y)



Geometry

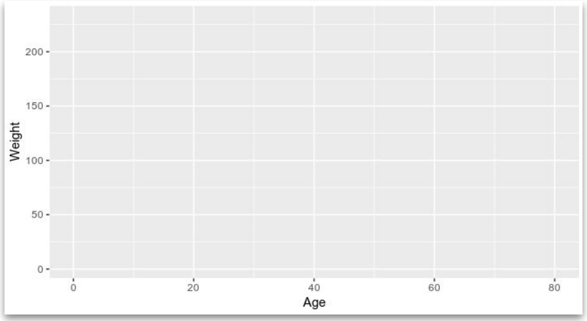


Data

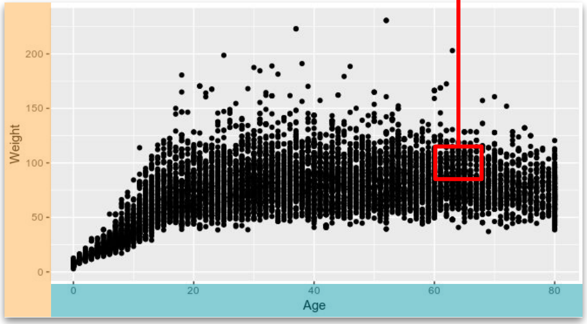


	ID	SurveyYr	Gender	Age	AgeDecade	AgeMonths	Race1
1	51624	2009_10	male	34	30-39	409	White
2	51624	2009_10	male	34	30-39	409	White
3	51624	2009_10	male	34	30-39	409	White
4	51625	2009_10	male	4	0-9	49	Other
5	51630	2009_10	female	49	40-49	596	White
6	51638	2009_10	male	9	0-9	115	White
7	51646	2009_10	male	8	0-9	101	White
8	51647	2009_10	female	45	40-49	541	White
9	51647	2009_10	female	45	40-49	541	White

Aesthetic
(x, y)



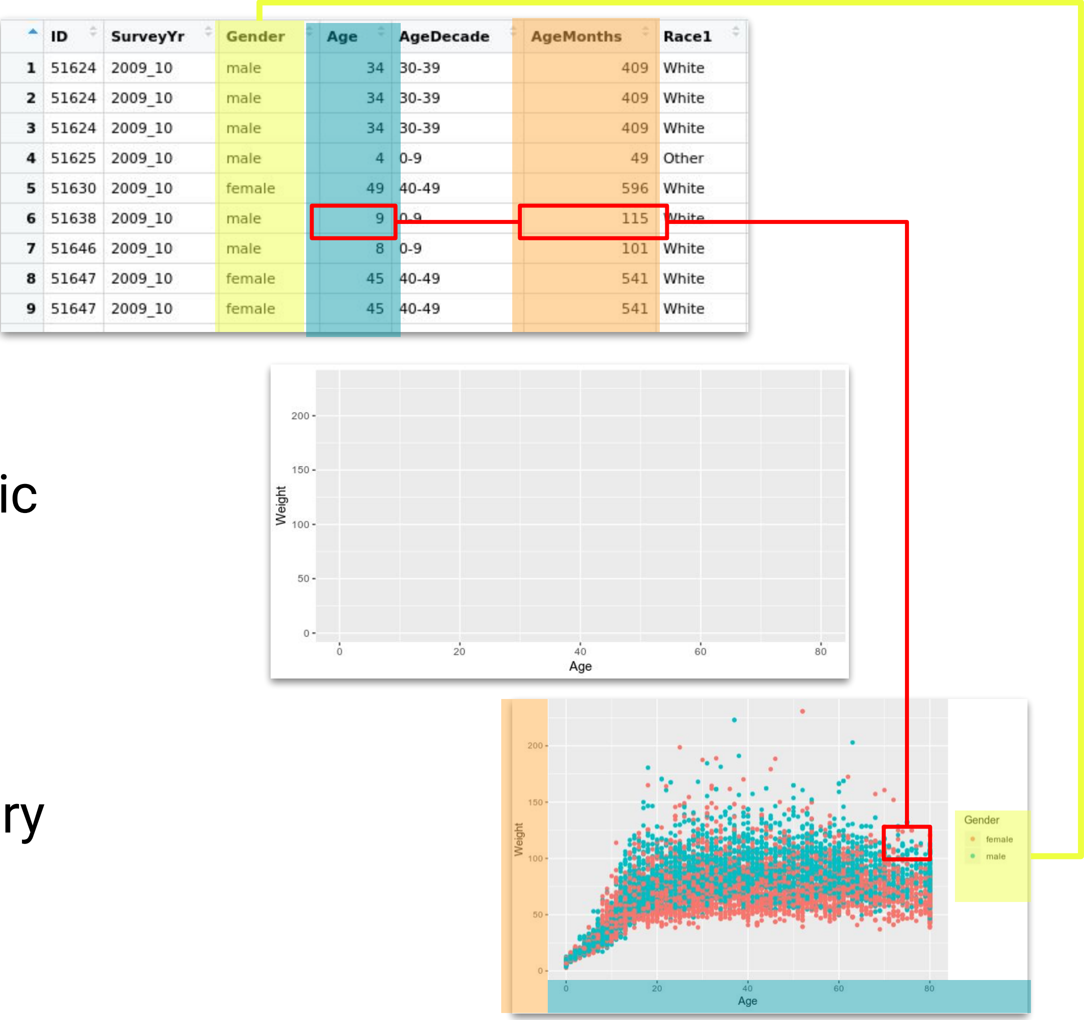
Geometry



Data

Aesthetic
(x, y)

Geometry



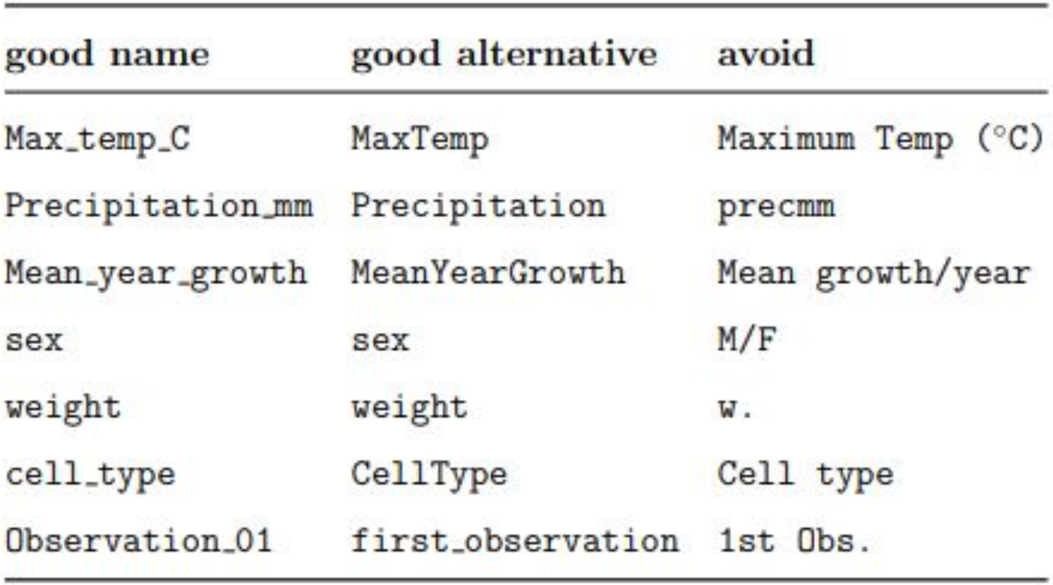
	ID	SurveyYr	Gender	Age	AgeDecade	AgeMonths	Race1
1	51624	2009_10	male	34	30-39	409	White
2	51624	2009_10	male	34	30-39	409	White
3	51624	2009_10	male	34	30-39	409	White
4	51625	2009_10	male	4	0-9	49	Other
5	51630	2009_10	female	49	40-49	596	White
6	51638	2009_10	male	9	0-9	115	White
7	51646	2009_10	male	8	0-9	101	White
8	51647	2009_10	female	45	40-49	541	White
9	51647	2009_10	female	45	40-49	541	White

Workshop

How to prepare the dataset for visualization

Choose good variable names

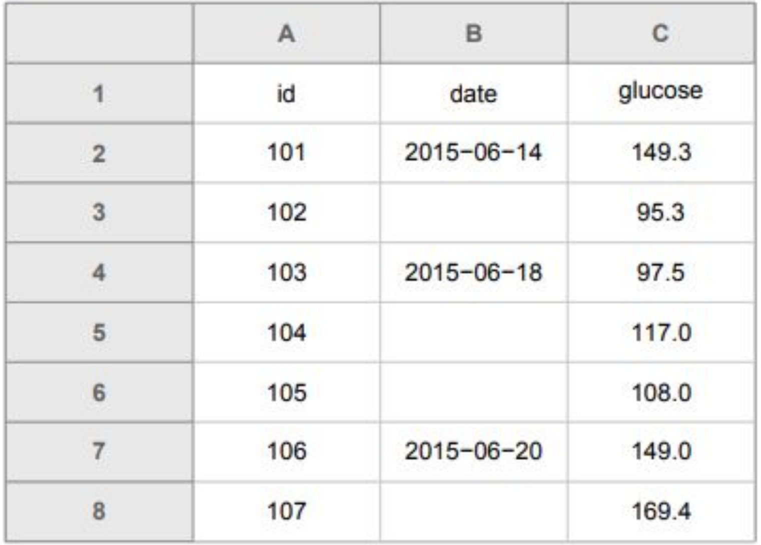
Table 1: Examples of good and bad variable names.



good name	good alternative	avoid
Max_temp_C	MaxTemp	Maximum Temp (°C)
Precipitation_mm	Precipitation	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell type
Observation_01	first_observation	1st Obs.

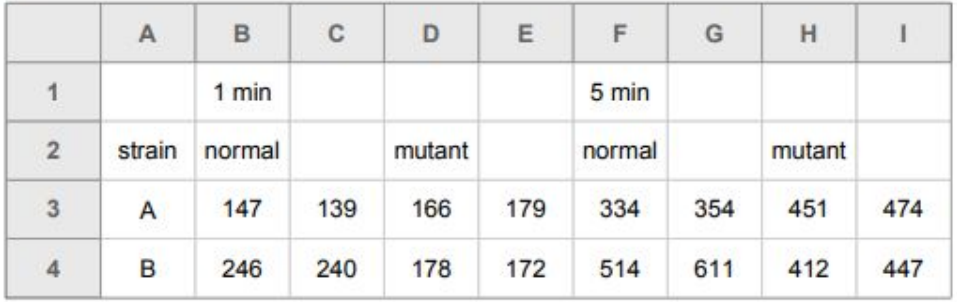
Tidy your data

A

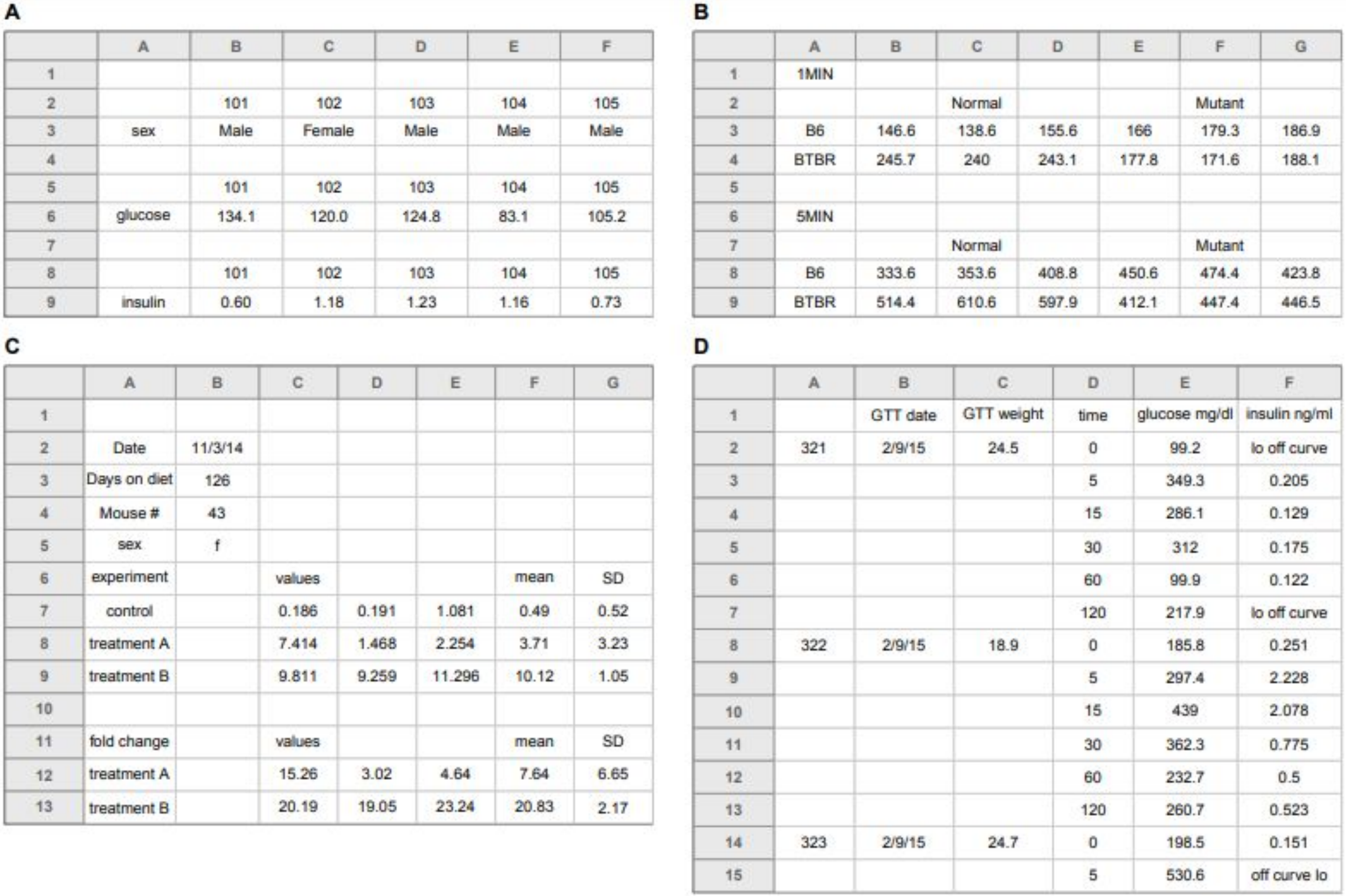


	A	B	C
1	id	date	glucose
2	101	2015-06-14	149.3
3	102		95.3
4	103	2015-06-18	97.5
5	104		117.0
6	105		108.0
7	106	2015-06-20	149.0
8	107		169.4

B



	A	B	C	D	E	F	G	H	I
1		1 min				5 min			
2	strain	normal		mutant		normal		mutant	
3	A	147	139	166	179	334	354	451	474
4	B	246	240	178	172	514	611	412	447



A

	A	B	C	D	E	F
1						
2		101	102	103	104	105
3	sex	Male	Female	Male	Male	Male
4						
5		101	102	103	104	105
6	glucose	134.1	120.0	124.8	83.1	105.2
7						
8		101	102	103	104	105
9	insulin	0.60	1.18	1.23	1.16	0.73

B

	A	B	C	D	E	F	G
1	1MIN						
2			Normal			Mutant	
3	B6	146.6	138.6	155.6	166	179.3	186.9
4	BTBR	245.7	240	243.1	177.8	171.6	188.1
5							
6	5MIN						
7			Normal			Mutant	
8	B6	333.6	353.6	408.8	450.6	474.4	423.8
9	BTBR	514.4	610.6	597.9	412.1	447.4	446.5

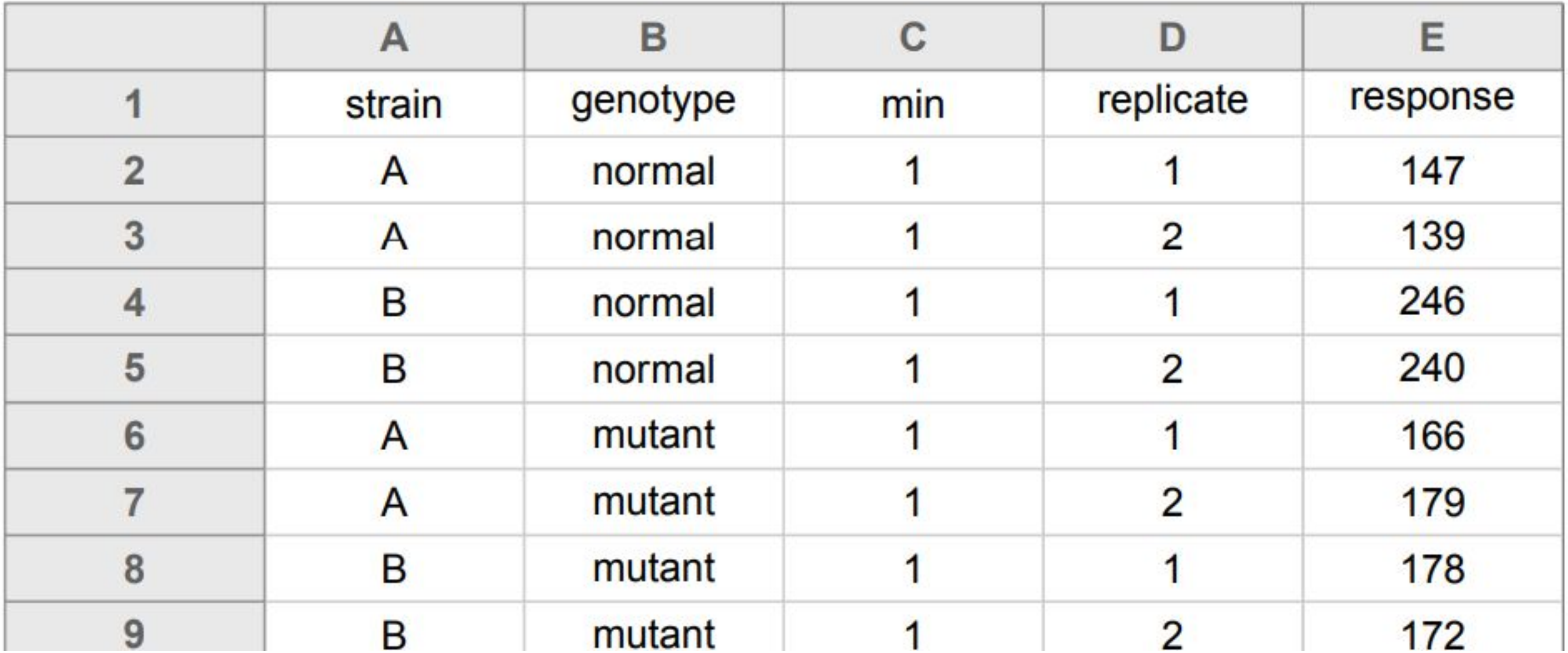
C

	A	B	C	D	E	F	G
1							
2	Date	11/3/14					
3	Days on diet	126					
4	Mouse #	43					
5	sex	f					
6	experiment		values			mean	SD
7	control		0.186	0.191	1.081	0.49	0.52
8	treatment A		7.414	1.468	2.254	3.71	3.23
9	treatment B		9.811	9.259	11.296	10.12	1.05
10							
11	fold change		values			mean	SD
12	treatment A		15.26	3.02	4.64	7.64	6.65
13	treatment B		20.19	19.05	23.24	20.83	2.17

D

	A	B	C	D	E	F
1		GTT date	GTT weight	time	glucose mg/dl	insulin ng/ml
2	321	2/9/15	24.5	0	99.2	lo off curve
3				5	349.3	0.205
4				15	286.1	0.129
5				30	312	0.175
6				60	99.9	0.122
7				120	217.9	lo off curve
8	322	2/9/15	18.9	0	185.8	0.251
9				5	297.4	2.228
10				15	439	2.078
11				30	362.3	0.775
12				60	232.7	0.5
13				120	260.7	0.523
14	323	2/9/15	24.7	0	198.5	0.151
15				5	530.6	off curve lo

Tidy your data

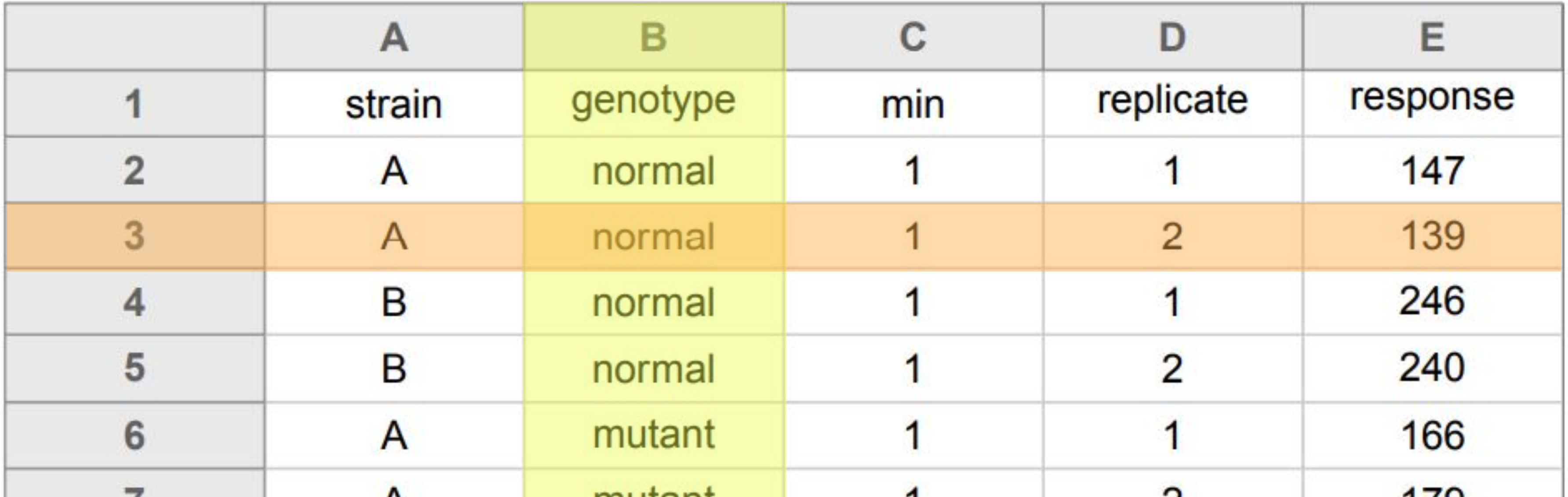


	A	B	C	D	E
1	strain	genotype	min	replicate	response
2	A	normal	1	1	147
3	A	normal	1	2	139
4	B	normal	1	1	246
5	B	normal	1	2	240
6	A	mutant	1	1	166
7	A	mutant	1	2	179
8	B	mutant	1	1	178
9	B	mutant	1	2	172

A tabular data set is tidy if:

1. Each variable is in its own column

2. Each observation is in its own row



	A	B	C	D	E
1	strain	genotype	min	replicate	response
2	A	normal	1	1	147
3	A	normal	1	2	139
4	B	normal	1	1	246
5	B	normal	1	2	240
6	A	mutant	1	1	166
7	A	mutant	1	2	170

R and Rstudio

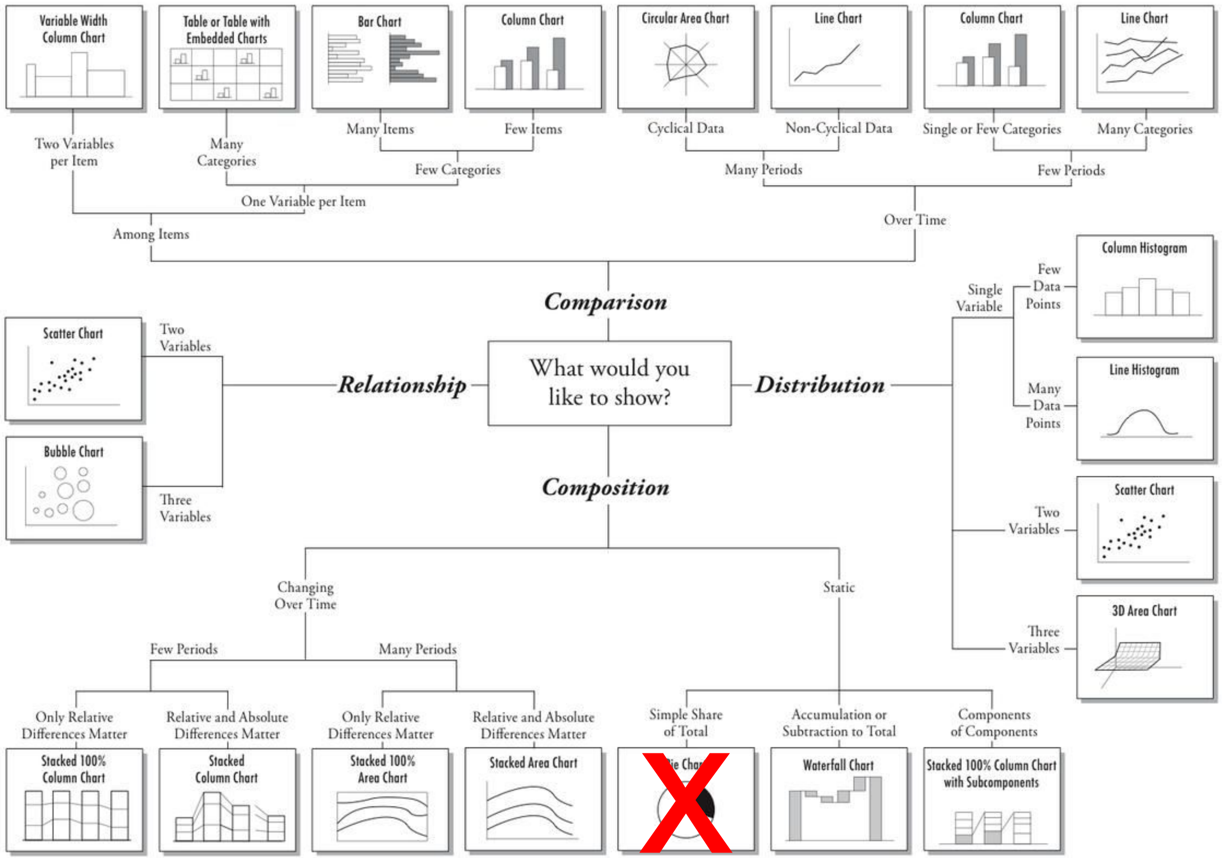
R and Rstudio

R:	Statistical languaje
RStudio:	Interface for R
Package:	Packages are collections of R functions, data, and compiled code

Useful shortcuts	⊞	🍎
Insert <-	Alt + -	Option + -
Insert %>%	Ctrl + Shift + M	Cmd + Shift + M

Which graph?

Chart Suggestions—A Thought-Starter



Variable Width Column Chart

Table or Table with Embedded Charts

Bar Chart

Column Chart

Circular Area Chart

Line Chart

Column Chart

Line Chart

Two Variables per Item

Many Categories

Many Items

Few Items

Cyclical Data

Non-Cyclical Data

Single or Few Categories

Many Categories

Few Categories

Many Periods

Few Periods

One Variable per Item

Over Time

Among Items

Comparison

Column Histogram

Single Variable

Few Data Points

Scatter Chart

Two Variables

Relationship

What would you like to show?

Distribution

Line Histogram

Many Data Points

Bubble Chart

Three Variables

Composition

Scatter Chart

Two Variables

Changing Over Time

Static

3D Area Chart

Three Variables

Few Periods

Many Periods

Only Relative Differences Matter

Relative and Absolute Differences Matter

Only Relative Differences Matter

Relative and Absolute Differences Matter

Simple Share of Total

Accumulation or Subtraction to Total

Components of Components

Stacked 100% Column Chart

Stacked Column Chart

Stacked 100% Area Chart

Stacked Area Chart

ie Char

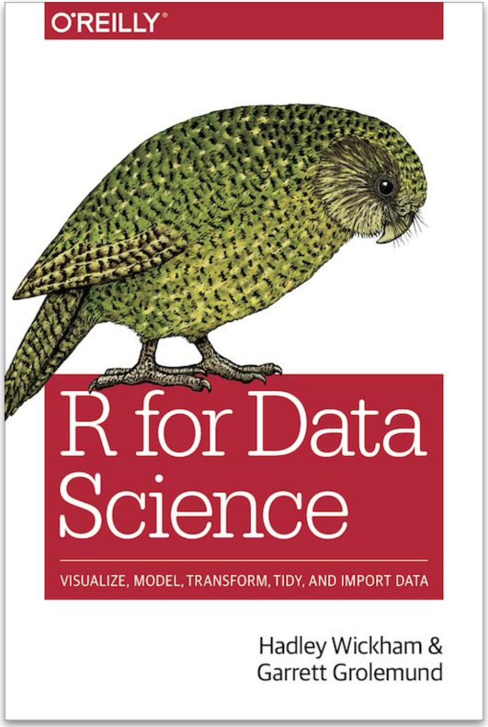
Waterfall Chart

Stacked 100% Column Chart with Subcomponents

© 2006 A. Abela — a.v.abela@gmail.com

Additional reading

Further reading

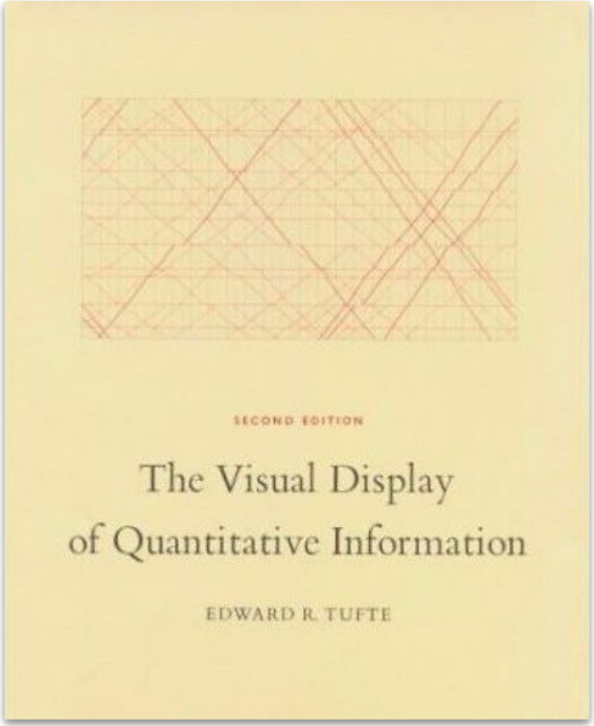


O'REILLY

R for Data Science

VISUALIZE, MODEL, TRANSFORM, TIDY, AND IMPORT DATA

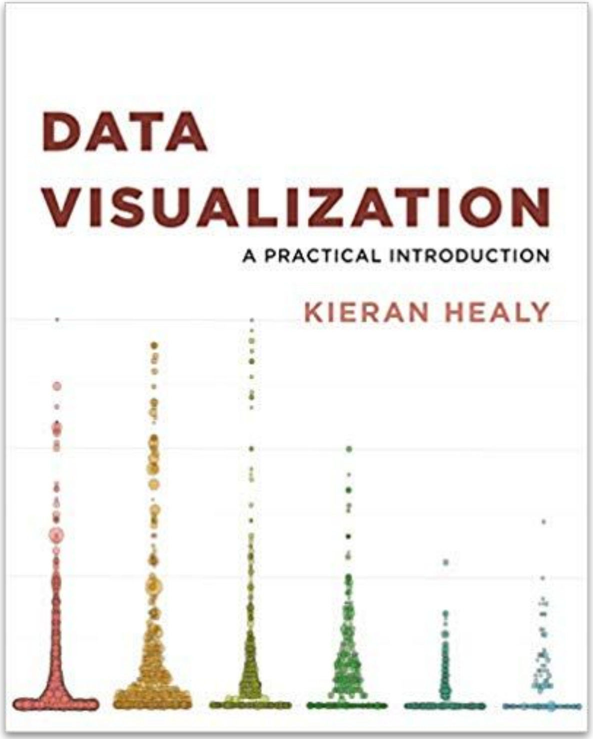
Hadley Wickham &
Garrett Grolemund



SECOND EDITION

The Visual Display
of Quantitative Information

EDWARD R. TUFTE



DATA
VISUALIZATION

A PRACTICAL INTRODUCTION

KIERAN HEALY

Recommended papers

Wagenmakers, E.-J., Gronau, Q.F., s. f. A Compendium of Clean Graphs in R [WWW Document]. URL http://shinyapps.org/apps/RGraphCompendium/index.php (accedido 12.18.18).

Puhan, M.A., ter Riet, G., Eichler, K., Steurer, J., Bachmann, L.M., 10/2006. More medical journals should inform their contributors about three key principles of graph construction. J. Clin. Epidemiol. 59, 1017.e1–1017.e8.

https://www.geckoboard.com/learn/data-literacy/data-visualization-tips/

http://shinyapps.stat.ubc.ca/r-graph-catalog/

Recommended papers

Broman, K.W., Woo, K.H., 2017. Data organization in spreadsheets (No. e3183v1). PeerJ Preprints. doi:10.7287/peerj.preprints.3183v1

Ellis, S.E., Leek, J.T., 2017. How to share data for collaboration (No. e3139v5). PeerJ Preprints.