Exploratory Data Analysis & Visualization

Business Analytics

Lesson Objectives

1. Exploratory Data Analysis & Visualization
    a. Exploratory Data Analysis
    b. Principles of Data Visualization

Understanding Your Data

Exploratory analysis of data is useful for:

- understanding data properties
- detecting errors, ensuring data quality
- finding patterns in data
- determining relationships among variables
- checking assumptions
- mapping business problems into data mining tasks and suggesting modeling strategies

Types of Exploratory Data Analysis

Exploratory data analysis is generally classified in two ways:

- non-graphical or graphical
- univariate or multivariate (usually just bivariate)

Types of Exploratory Data Analysis

- Non-graphical methods involve calculation of summary statistics.
- Graphical methods use charts and visual displays to summarize the data.
- Univariate methods look at one variable at a time.
- Multivariate methods look at two or more variables at a time to explore relationships.
- It is almost always a good idea to perform univariate EDA on each component of a multivariate EDA before performing the multivariate EDA.

Univariate Non-Graphical EDA

- A particular measurement on all of the subjects in a sample for a single characteristic such as age, gender, etc.
- Think of these measurements as representing a “sample distribution” of the variable.
- Categorical data -- the range of values and the frequency (or relative frequency) of occurrence for each value
- Quantitative data -- measures of central tendencies, spread, modality, outliers.

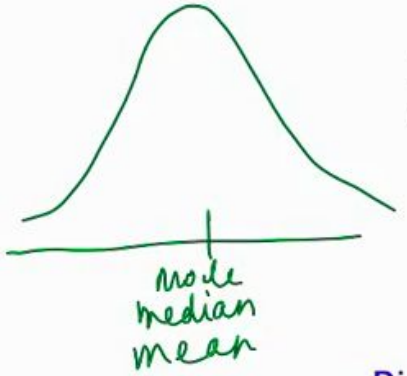
Univariate Graphical EDA

- Exploring the distribution of the sample graphically

Shape of Distribution

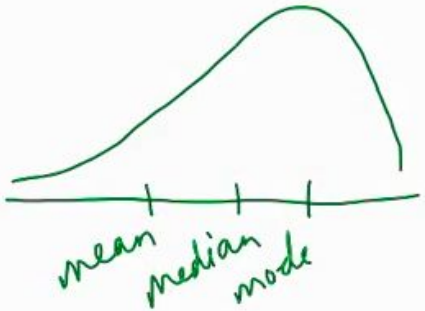
The relative location of the mode, median, and mean in a unimodal distribution:

Symmetric



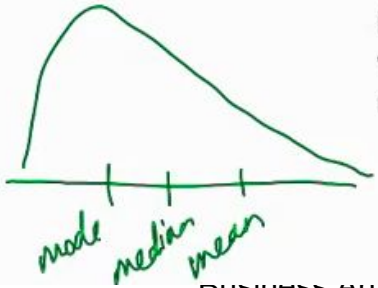
For a symmetric distribution, the mean, median, and mode are all approximately the same.

Left-skewed



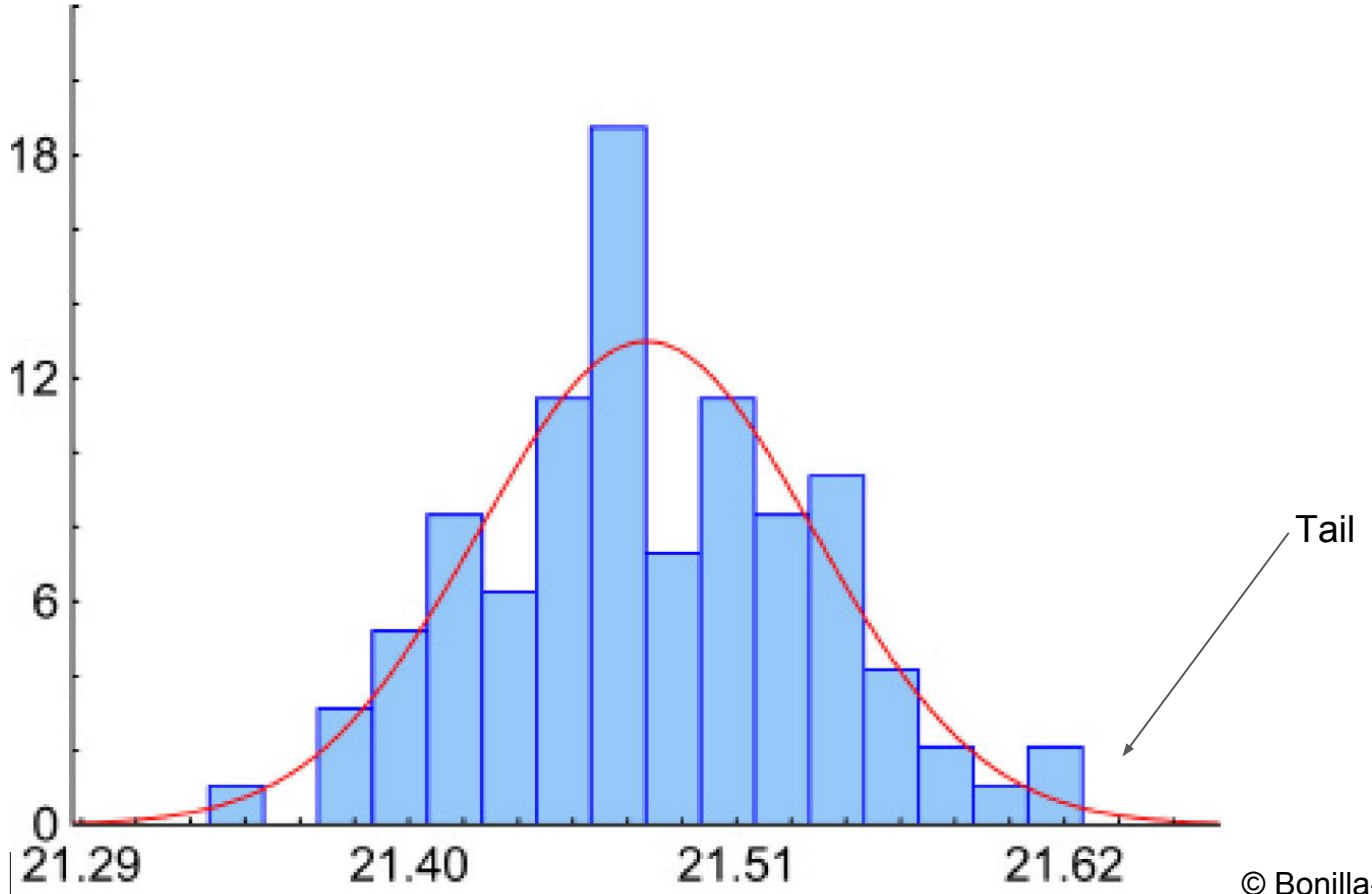
For a left-skewed distribution, the mode is larger than the median which is larger than the mean.

Right-skewed



For a right-skewed distribution, the mode is less than the median, which is less than the mean.

Histograms & Distributions



Case Study

Direct marketing campaigns of a Portuguese banking institution

- selling bank term deposit (CD)
- phone calls -- one or more contacts to the same client is often required

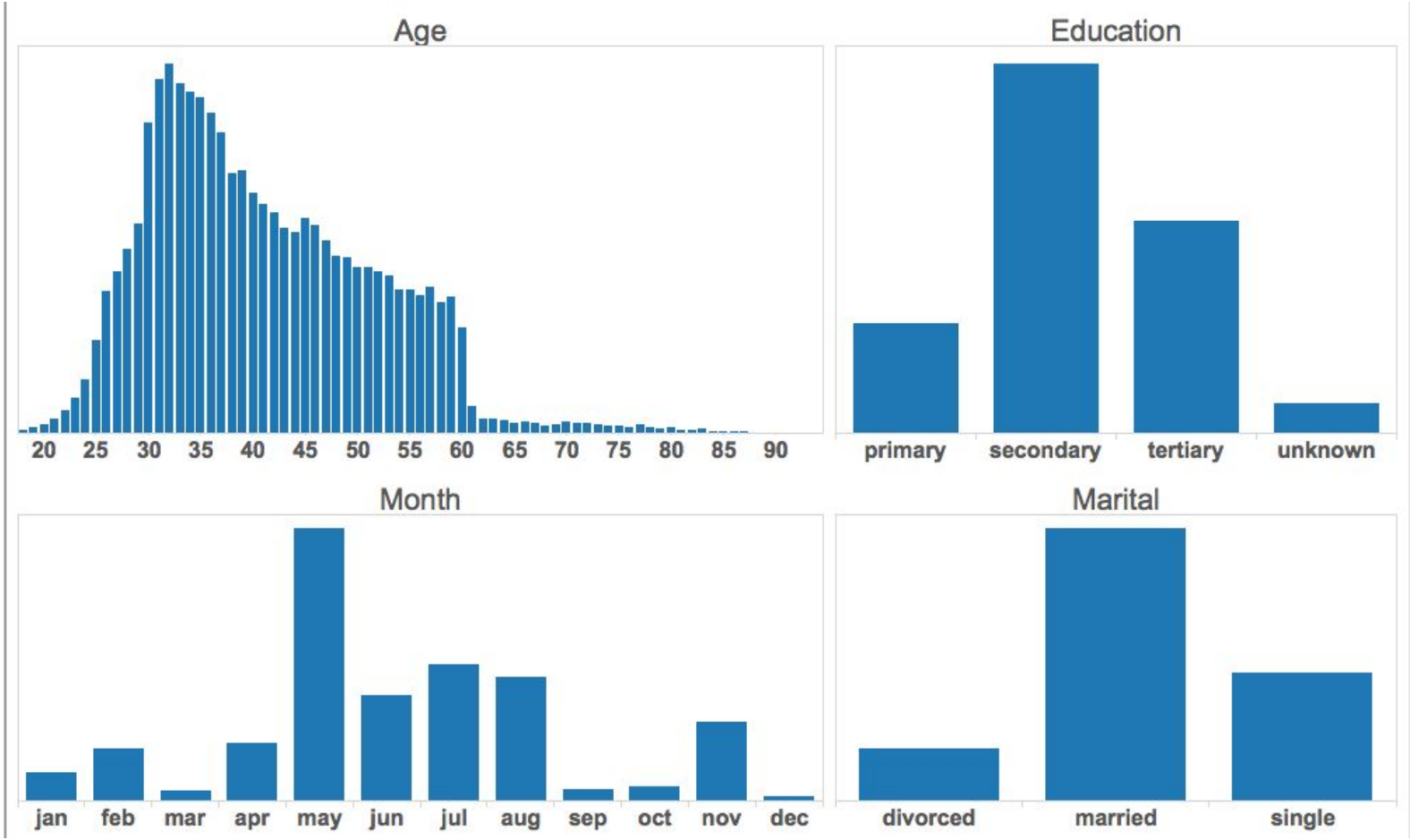
age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	subscribe
58	management	married	tertiary	no	$2,143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
44	technician	single	secondary	no	$29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
33	entrepreneur	married	secondary	no	$2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
47	blue-collar	married	unknown	no	$1,506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
33	unknown	single	unknown	no	$1	no	no	unknown	5	may	198	1	-1	0	unknown	no
35	management	married	tertiary	no	$231	yes	no	unknown	5	may	139	1	-1	0	unknown	no
28	management	single	tertiary	no	$447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
42	entrepreneur	divorced	tertiary	yes	$2	yes	no	unknown	5	may	380	1	-1	0	unknown	no
58	retired	married	primary	no	$121	yes	no	unknown	5	may	50	1	-1	0	unknown	no
43	technician	single	secondary	no	$593	yes	no	unknown	5	may	55	1	-1	0	unknown	no
41	admin.	divorced	secondary	no	$270	yes	no	unknown	5	may	222	1	-1	0	unknown	no
29	admin.	single	secondary	no	$390	yes	no	unknown	5	may	137	1	-1	0	unknown	no
53	technician	married	secondary	no	$6	yes	no	unknown	5	may	517	1	-1	0	unknown	no
58	technician	married	unknown	no	$71	yes	no	unknown	5	may	71	1	-1	0	unknown	no
57	services	married	secondary	no	$162	yes	no	unknown	5	may	174	1	-1	0	unknown	no
51	retired	married	primary	no	$229	yes	no	unknown	5	may	353	1	-1	0	unknown	no
45	admin.	single	unknown	no	$13	yes	no	unknown	5	may	98	1	-1	0	unknown	no

Count of subscribe	
Row Labels	Total
18	12
19	35
20	50
21	79
22	129
23	202
24	302
25	527
26	805
27	909
28	1038
29	1185
30	1757
31	1996
32	2085
33	1972
34	1930
35	1894
36	1806
37	1696
38	1466
39	1487
40	1355
41	1291
42	1242
43	1161
44	1136
45	1216

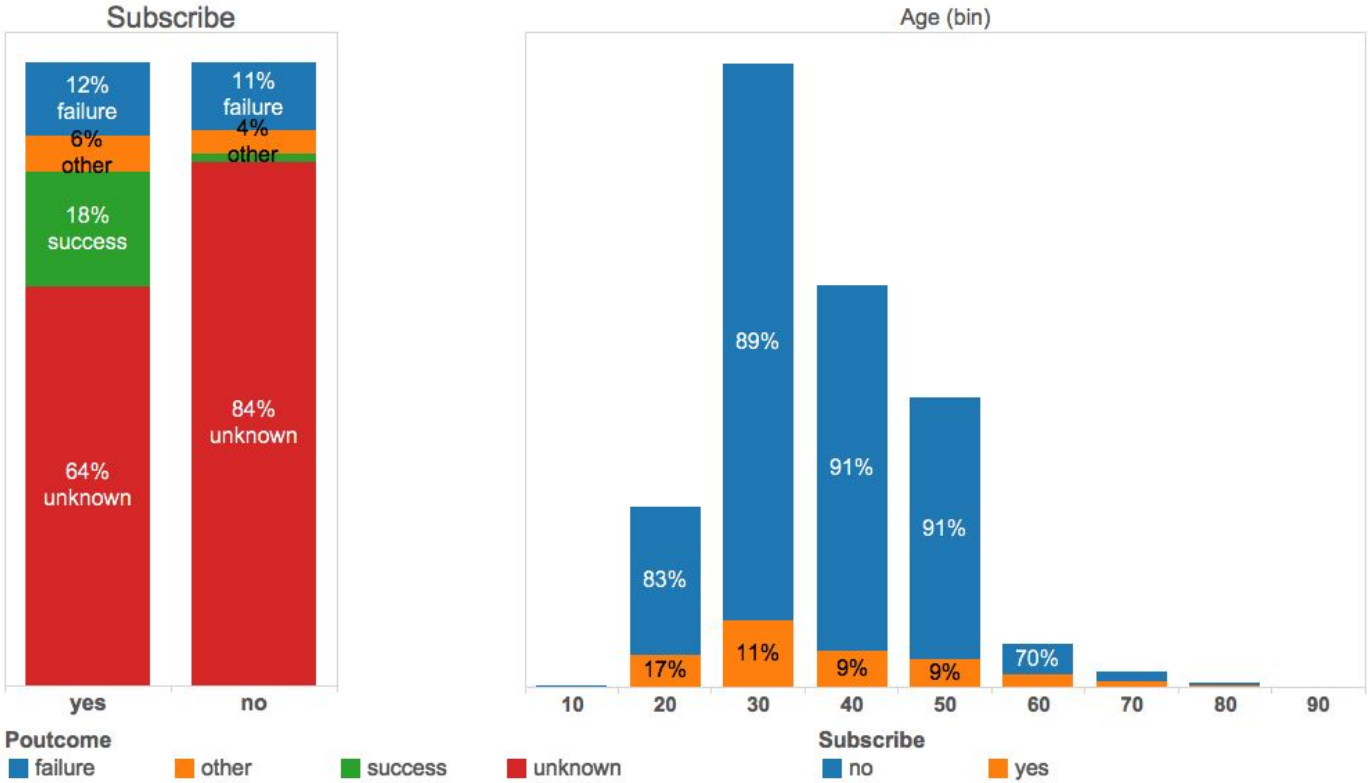
Count of subscribe	
Row Labels	Total
divorced	5207
married	27214
single	12790
(blank)	
Grand Total	**45211**

Count of subscribe	
Row Labels	Total
jan	1403
feb	2649
mar	477
apr	2932
may	13766
jun	5341
jul	6895
aug	6247
sep	579
oct	738
nov	3970
dec	214
(blank)	
Grand Total	**45211**

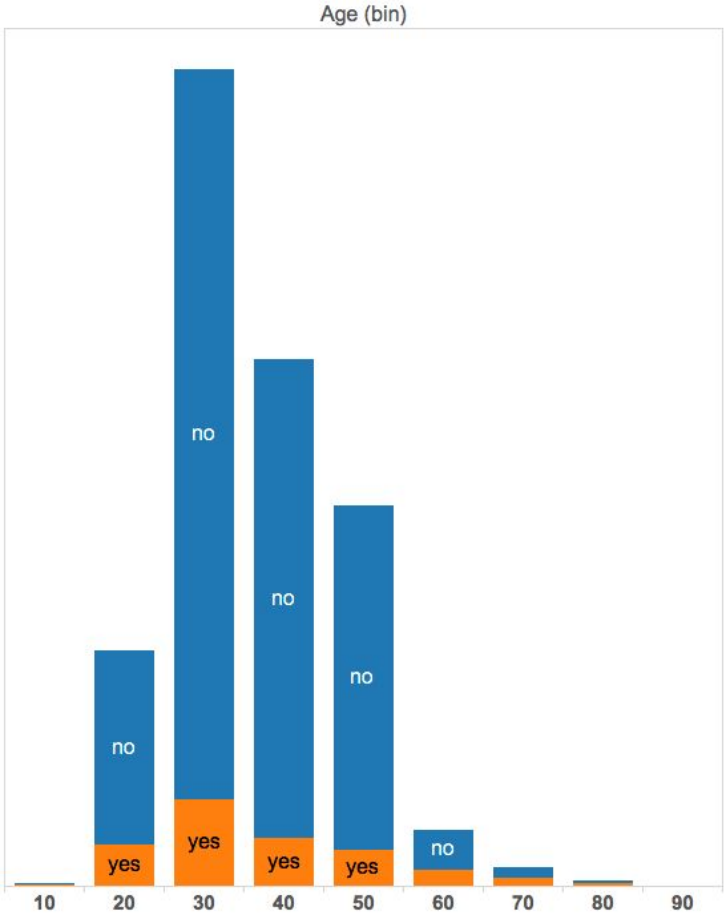
Count of subscribe	
Row Labels	Total
primary	6851
secondary	23202
tertiary	13301
unknown	1857
(blank)	
Grand Total	**45211**


Age
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
Education
primary
secondary
tertiary
unknown
Month
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec
Marital
divorced
married
single

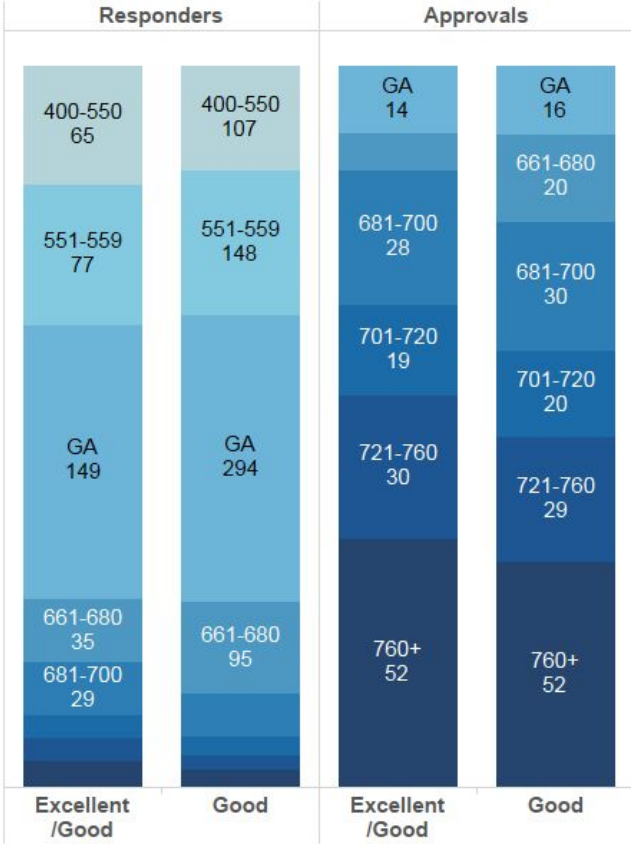
Multivariate Graphical EDA

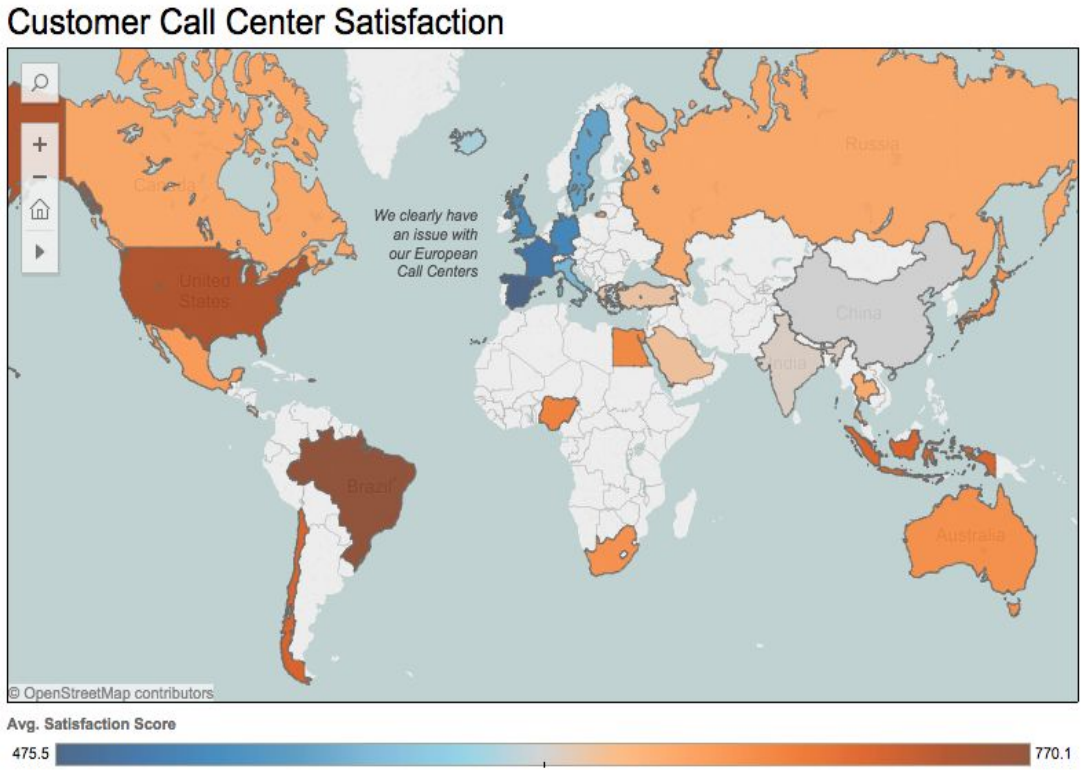


Multivariate Graphical EDA

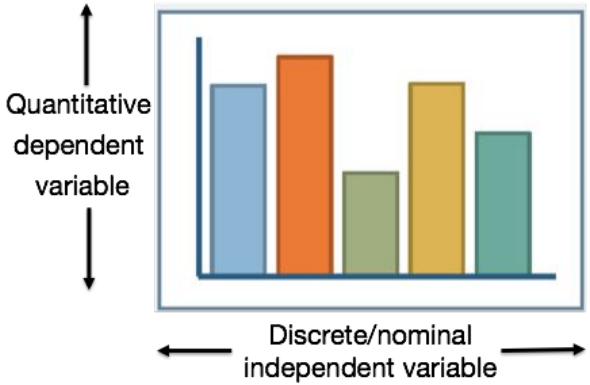


Multivariate Graphical EDA

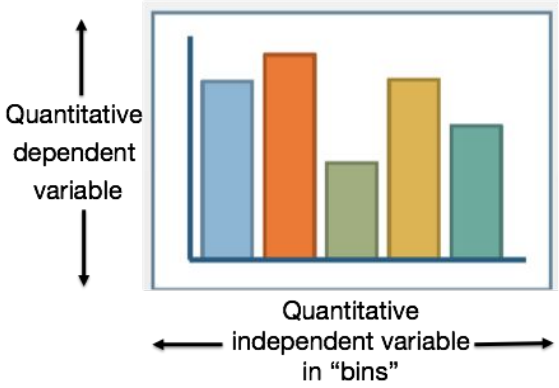




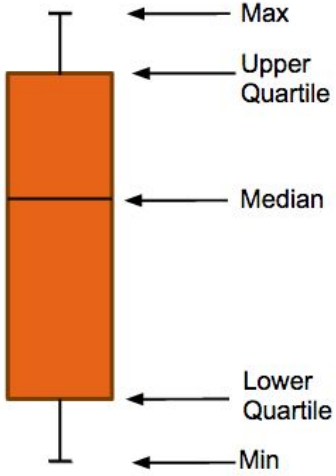
Bar Chart



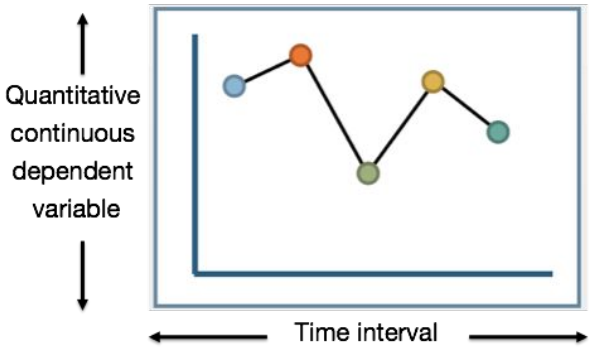
Histogram



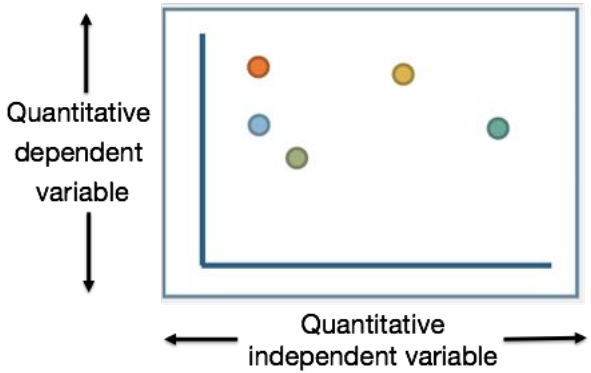
Box Plot



Time Series



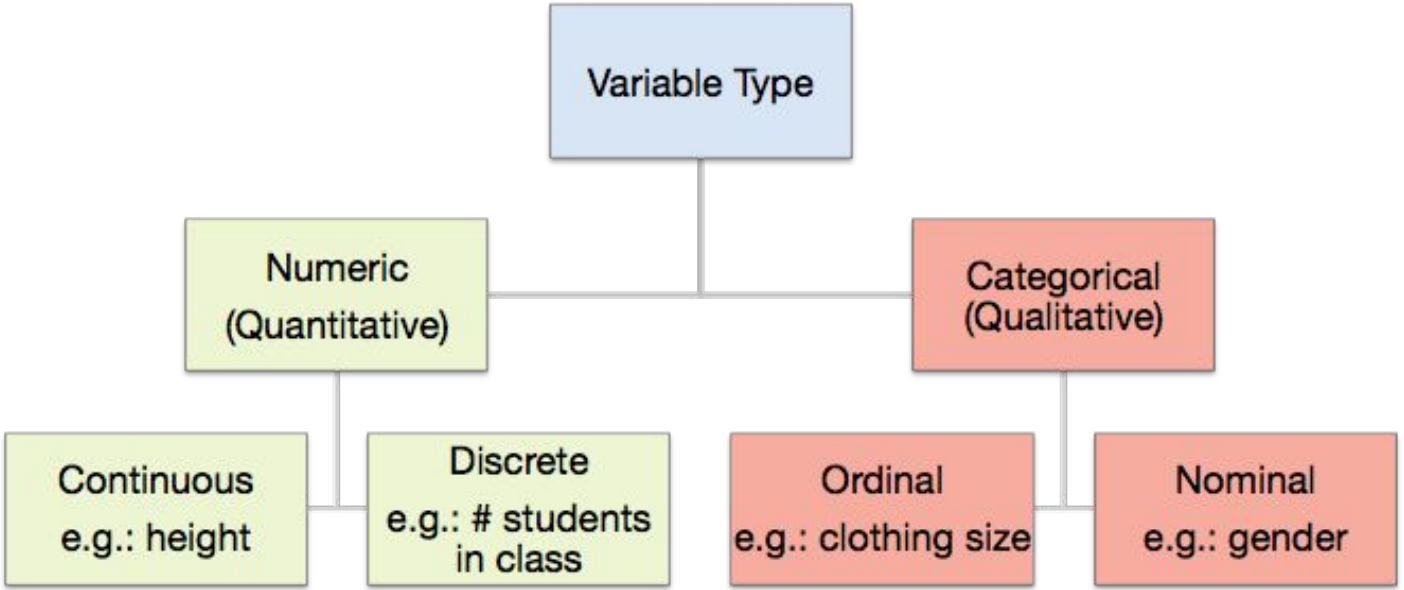
Scatter Plot



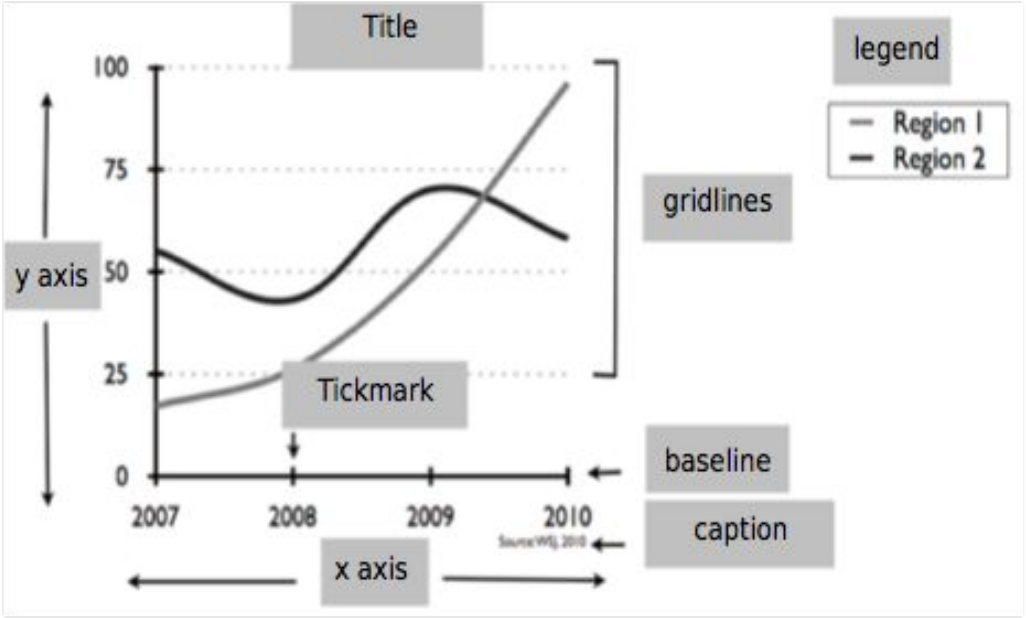
Visualization for Descriptive Analytics

- Data types
- Data transformation - percentages, proportions...
- Chart types - visualizing patterns, relationships, comparisons, or distributions?

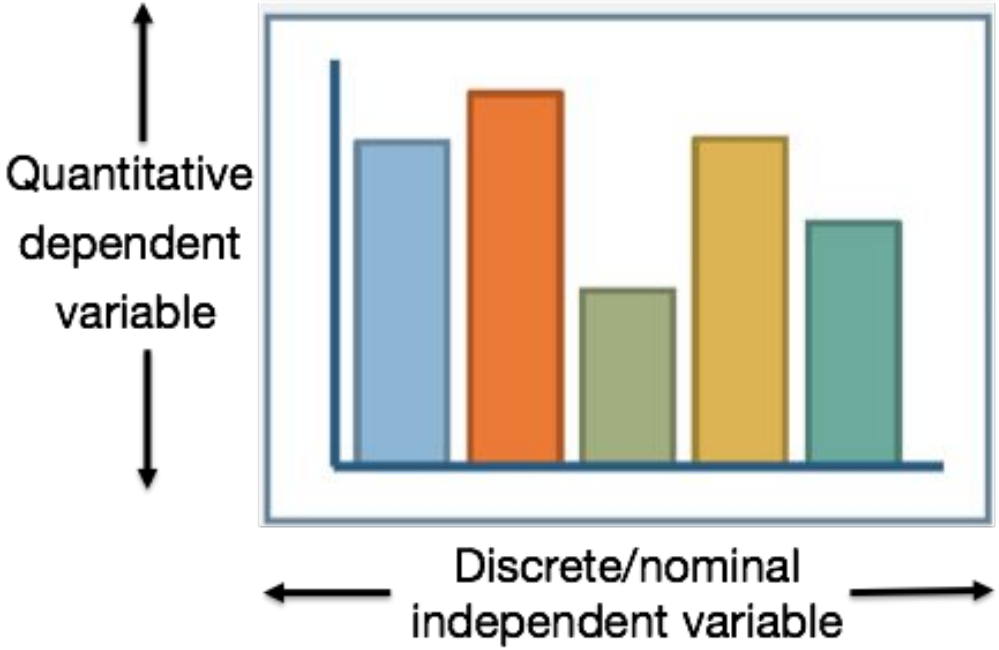
Data Types

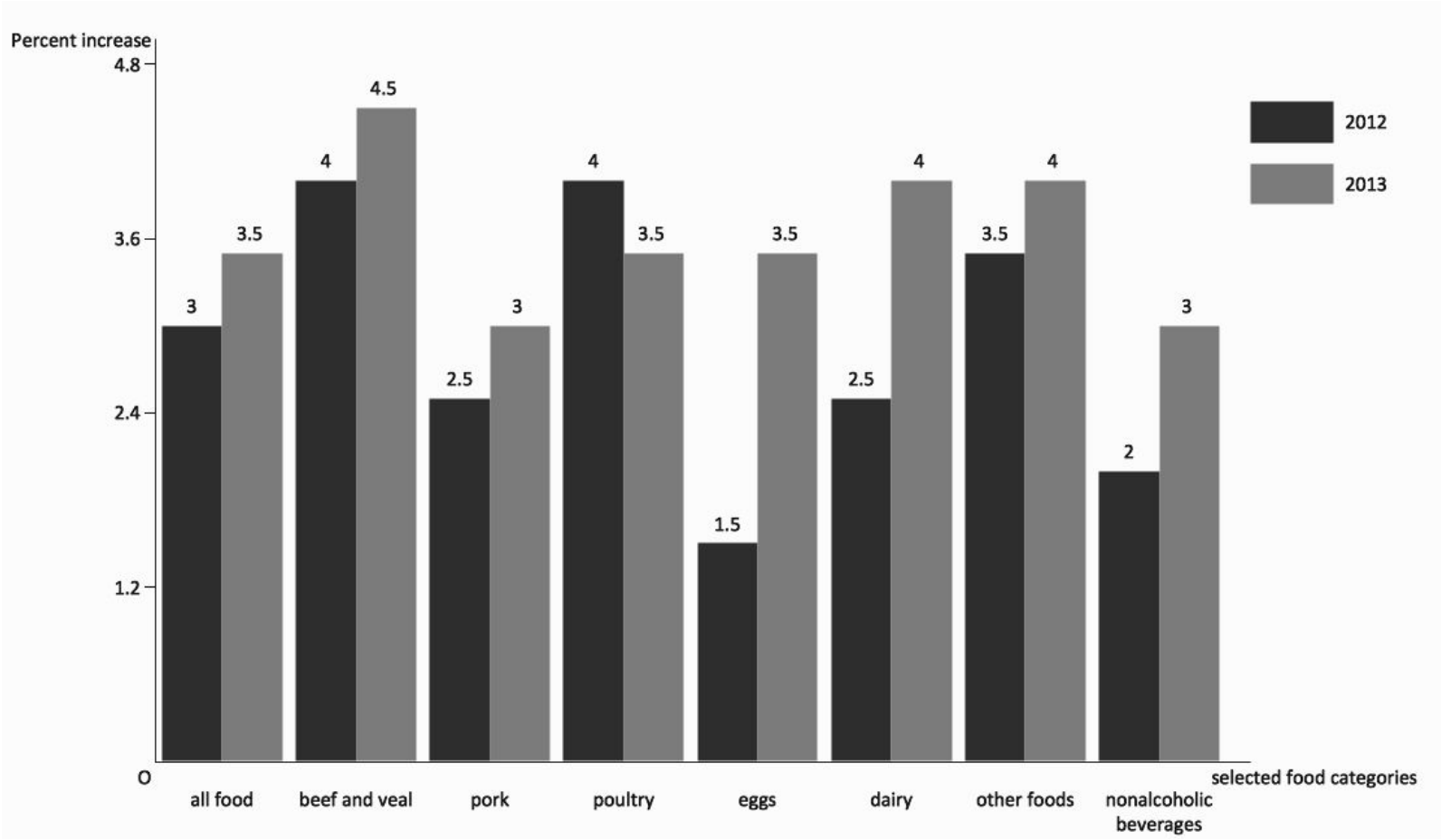

Variable Type
Numeric
(Quantitative)
Categorical
(Qualitative)
Continuous
e.g.: height
Discrete
e.g.: # students
in class
Ordinal
e.g.: clothing size
Nominal
e.g.: gender

Basic Chart Terminology

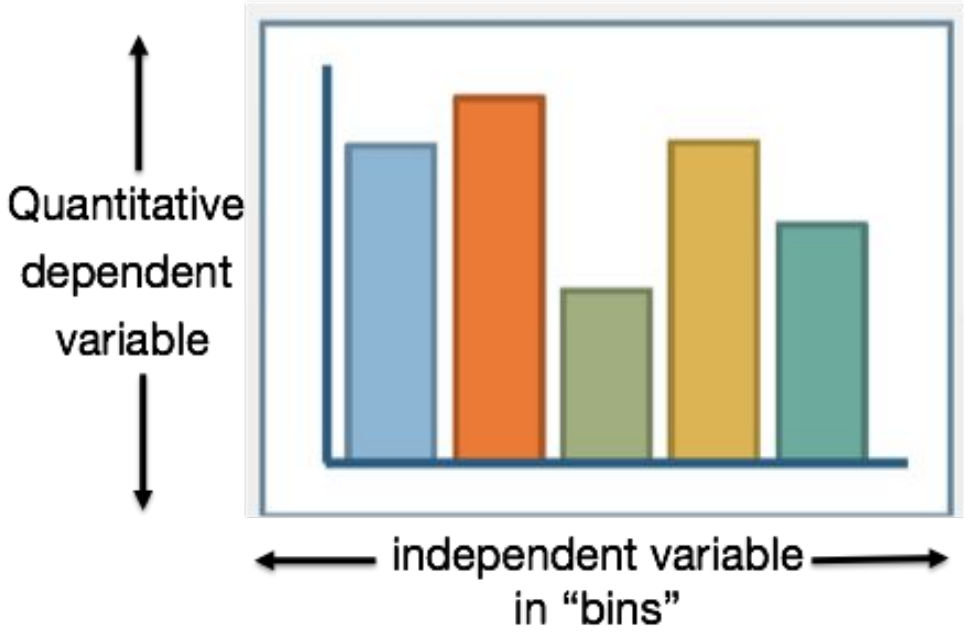


Bar Chart


Quantitative
dependent
variable
Discrete/nominal
independent variable


Percent increase
4.8
3.6
2.4
1.2
O
2012
2013
3
3.5
4
4.5
2.5
3
4
3.5
1.5
3.5
2.5
4
3.5
4
2
3
all food
beef and veal
pork
poultry
eggs
dairy
other foods
nonalcoholic beverages
selected food categories

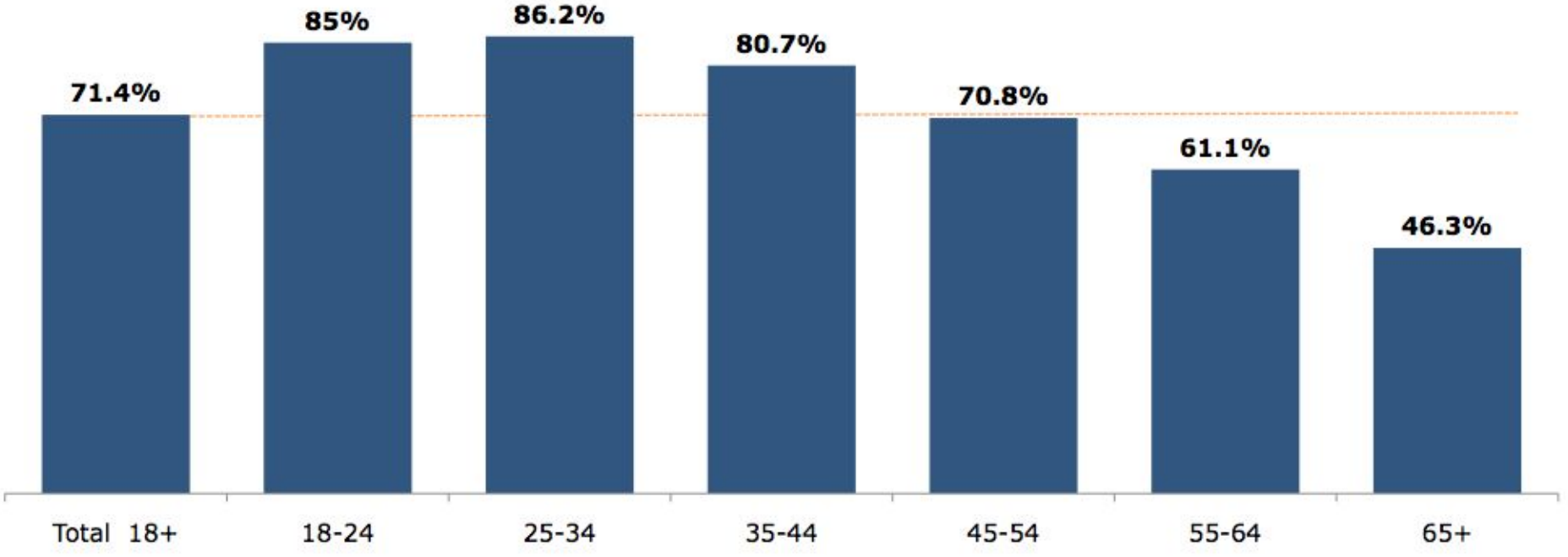
Histogram


Quantitative
dependent
variable
independent variable
in “bins”

US Smartphone Penetration Rate, by Age Group

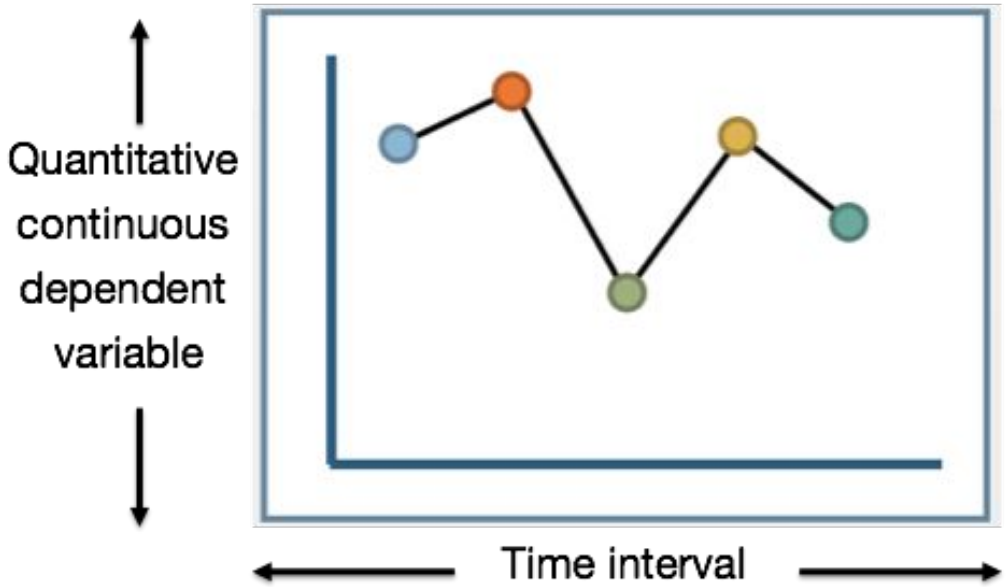
among mobile subscribers in the US

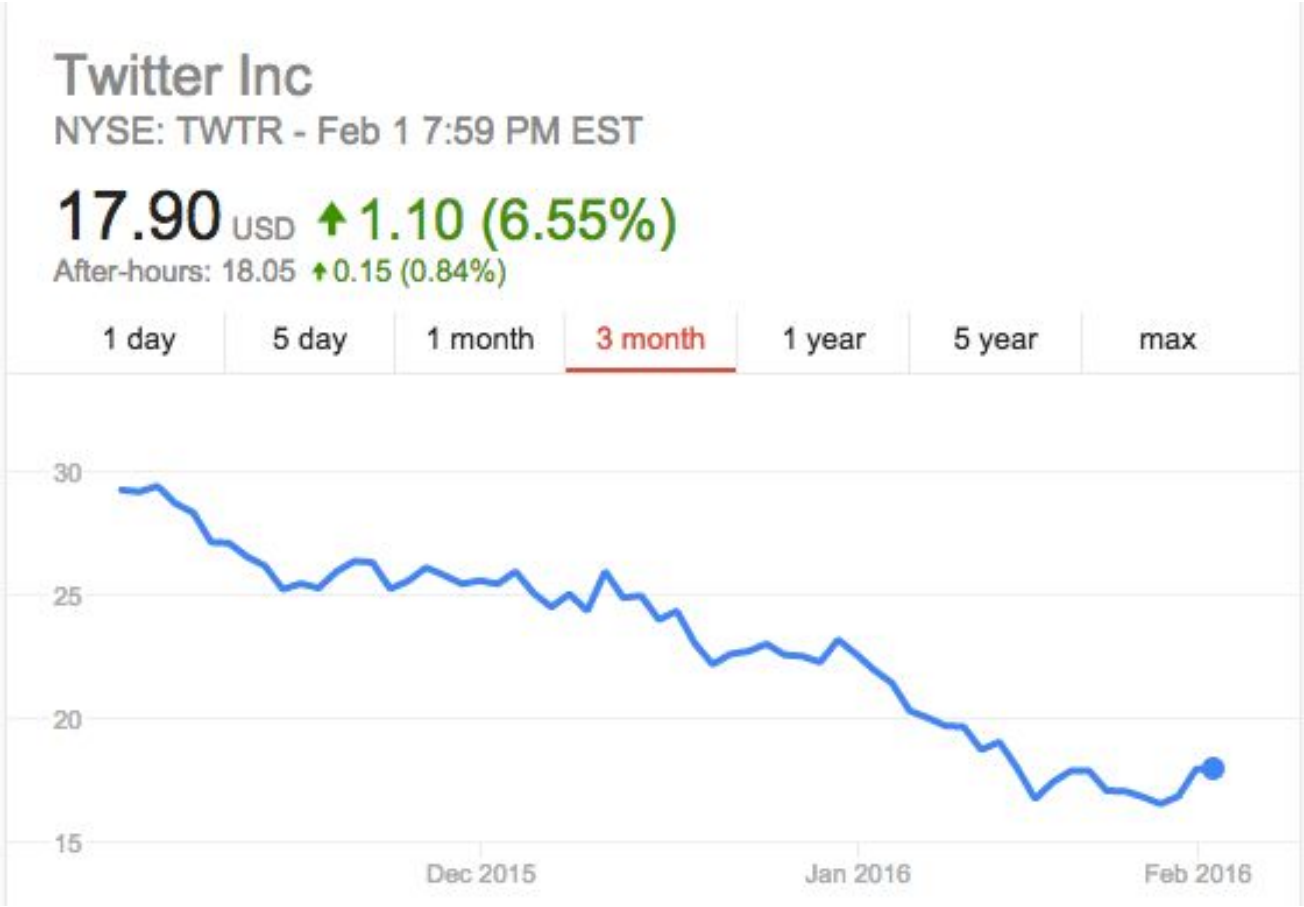
During Q2 2014



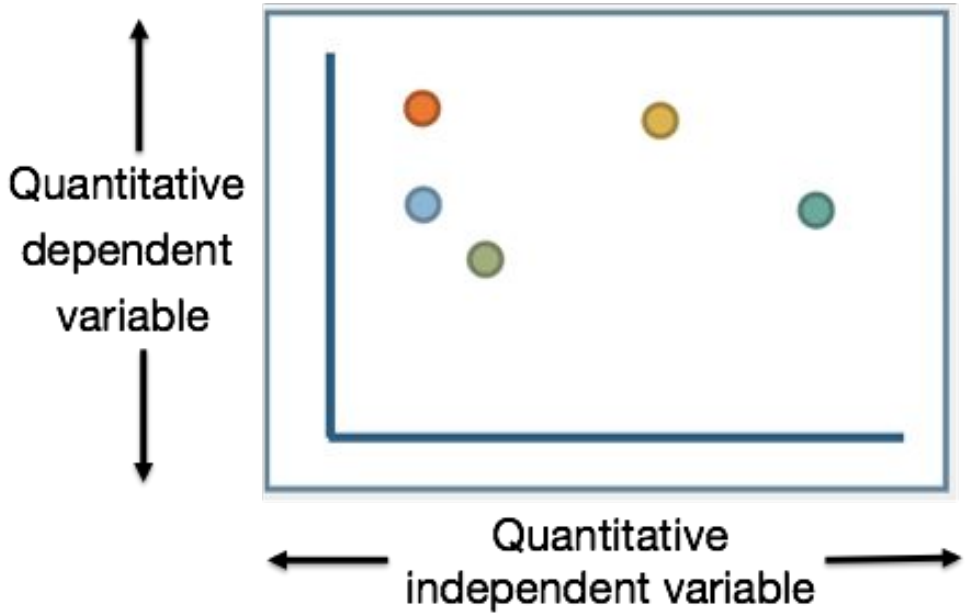
MarketingCharts.com | Data Source: Nielsen

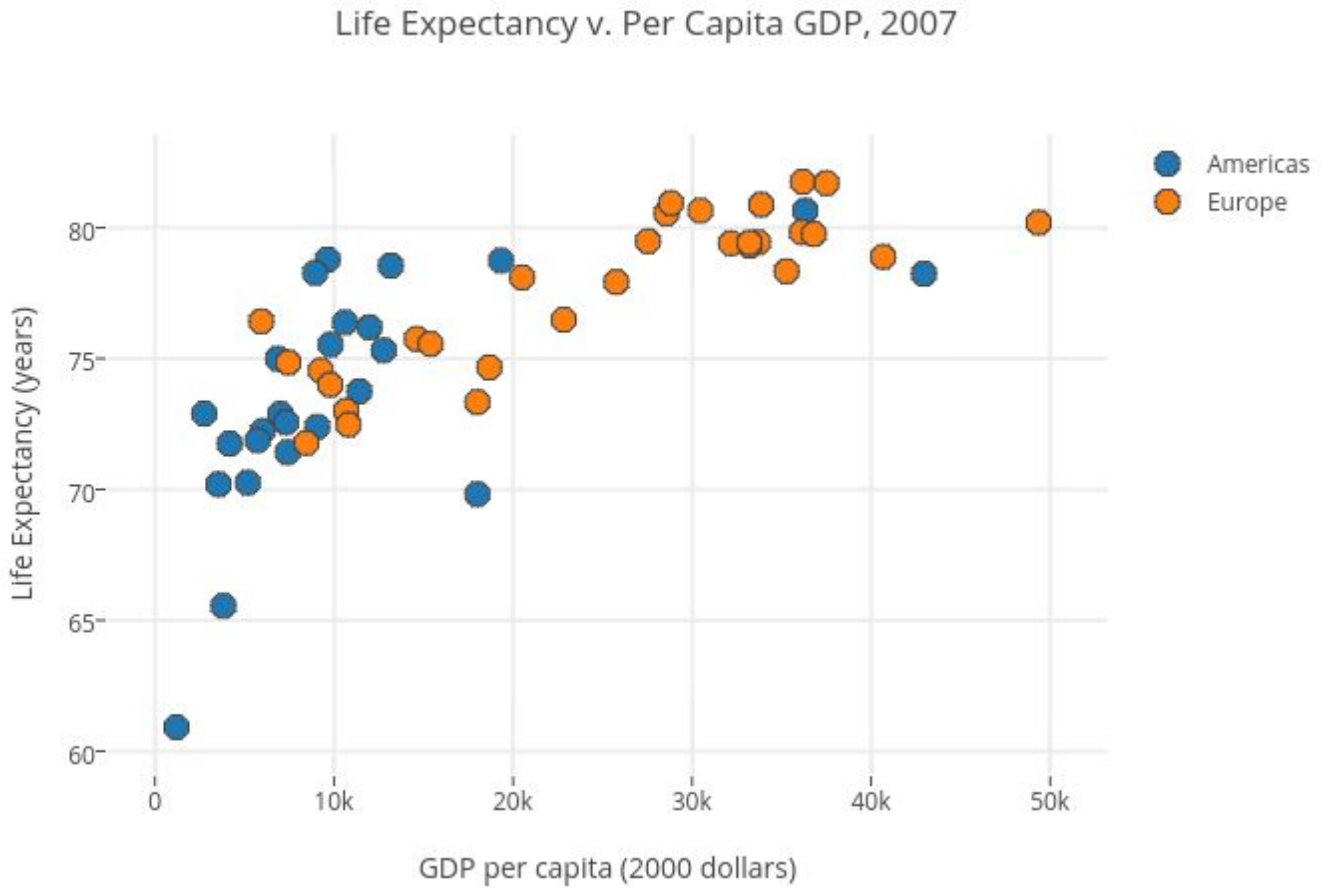
Time Series


Quantitative
continuous
dependent
variable
Time interval

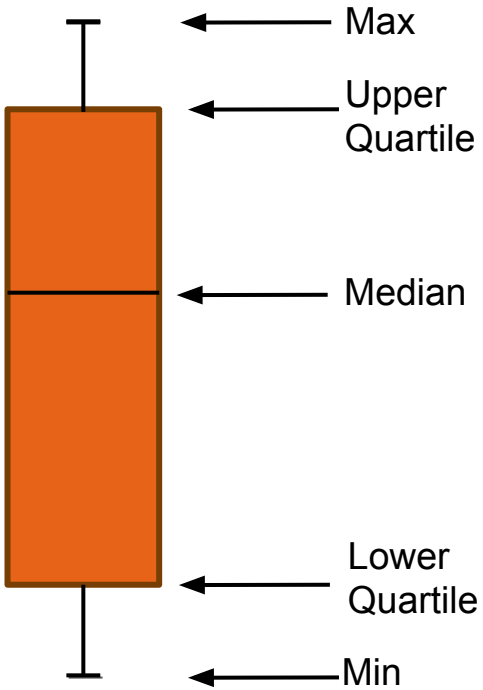

Twitter Inc
NYSE: TWTR - Feb 1 7:59 PM EST
17.90 USD ↑1.10 (6.55%)
After-hours: 18.05 ↑0.15 (0.84%)
1 day
5 day
1 month
3 month
1 year
5 year
max
30
25
20
15
Dec 2015
Jan 2016
Feb 2016

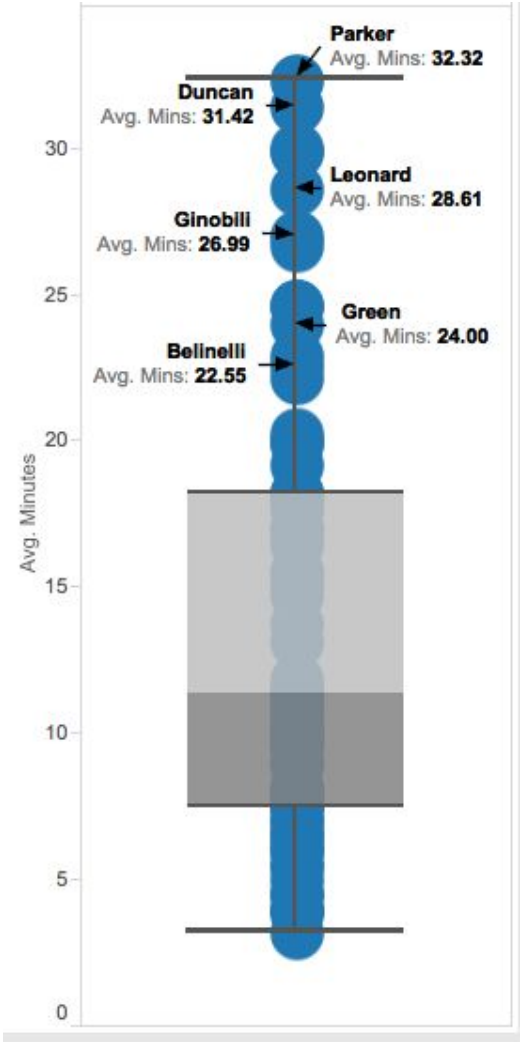
Scatter Plot




Life Expectancy v. Per Capita GDP, 2007
Americas
Europe
Life Expectancy (years)
80
75
70
65
60
0
10k
20k
30k
40k
50k
GDP per capita (2000 dollars)

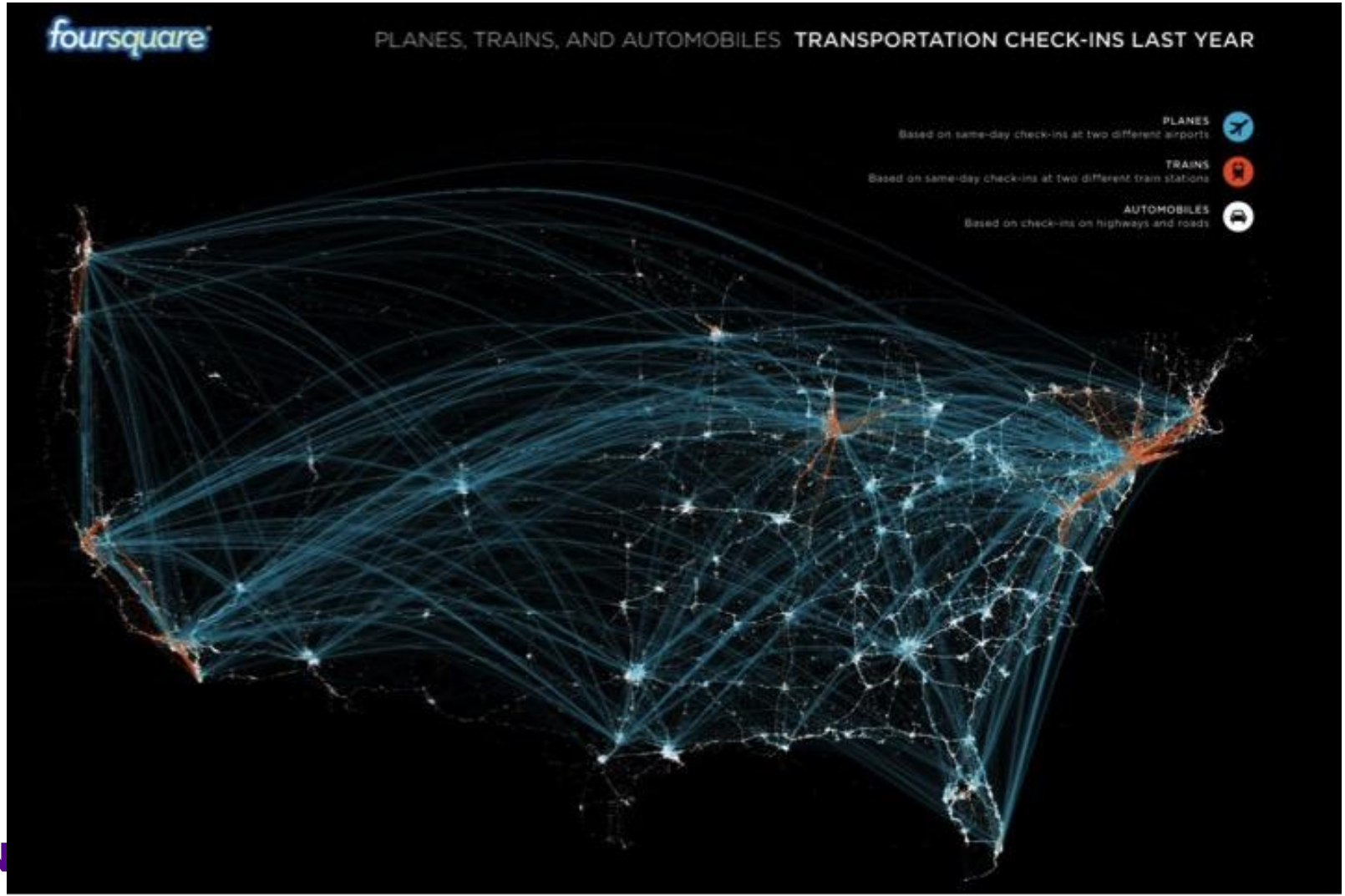
Box Plot (Box & Whisker diagram)

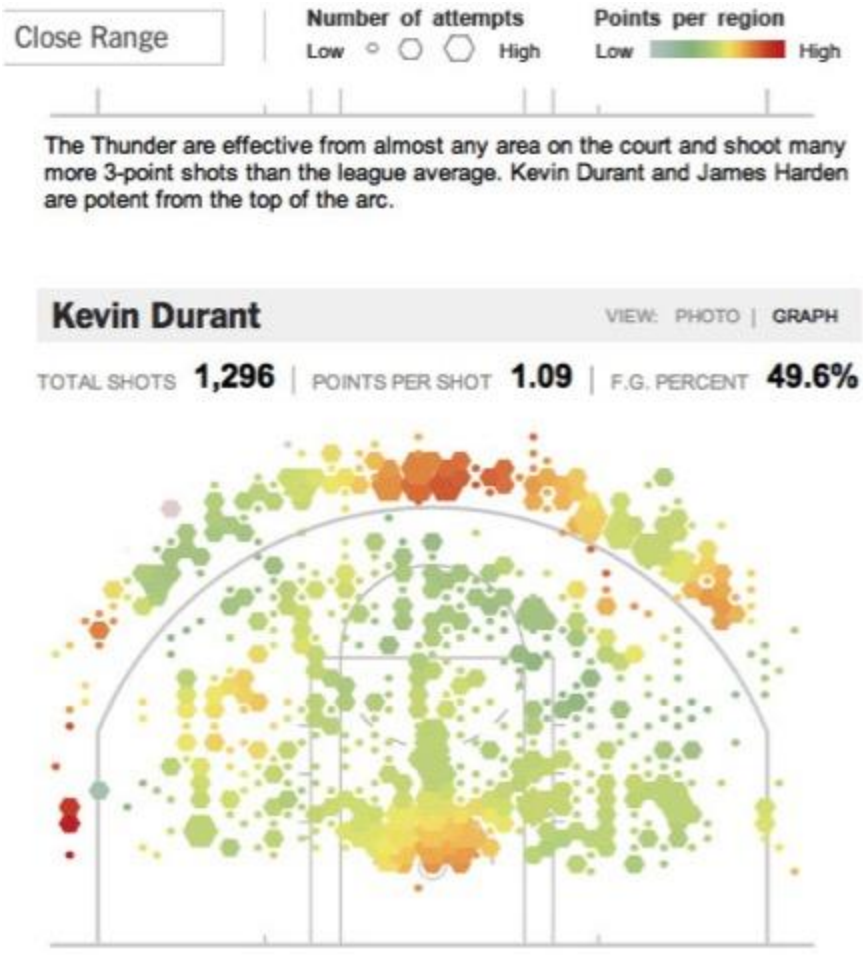



Parker
Avg. Mins: 32.32
Duncan
Avg. Mins: 31.42
Leonard
Avg. Mins: 28.61
Ginobili
Avg. Mins: 26.99
Green
Avg. Mins: 24.00
Belinelli
Avg. Mins: 22.55
Avg. Minutes
30
25
20
15
10
5
0

What is data visualization?

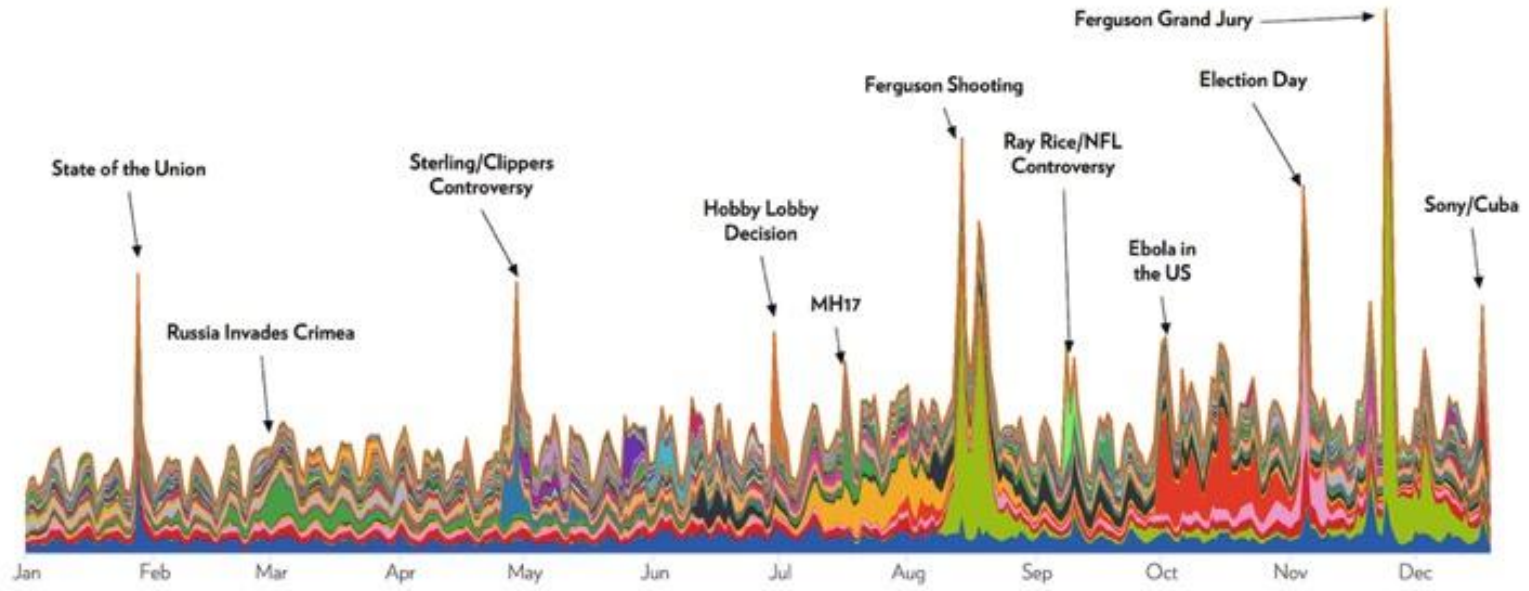
- Representation of data in a pictorial or graphical format.
- A general way of talking about anything that converts data sources into a visual representation:
  - charts, graphs, maps, sometimes even just tables
- Combination of many disciplines
  - statistics, perception, graphic design, cognitive psychology, information design, communications, and data mining


foursquare
PLANES, TRAINS, AND AUTOMOBILES TRANSPORTATION CHECK-INS LAST YEAR
PLANES
Based on same-day check-ins at two different airports
TRAINS
Based on same-day check-ins at two different train stations
AUTOMOBILES
Based on check-ins on highways and roads


Close Range
Number of attempts
Low
High
Points per region
Low
High
The Thunder are effective from almost any area on the court and shoot many more 3-point shots than the league average. Kevin Durant and James Harden are potent from the top of the arc.
Kevin Durant
VIEW: PHOTO | GRAPH
TOTAL SHOTS 1,296 | POINTS PER SHOT 1.09 | F.G. PERCENT 49.6%

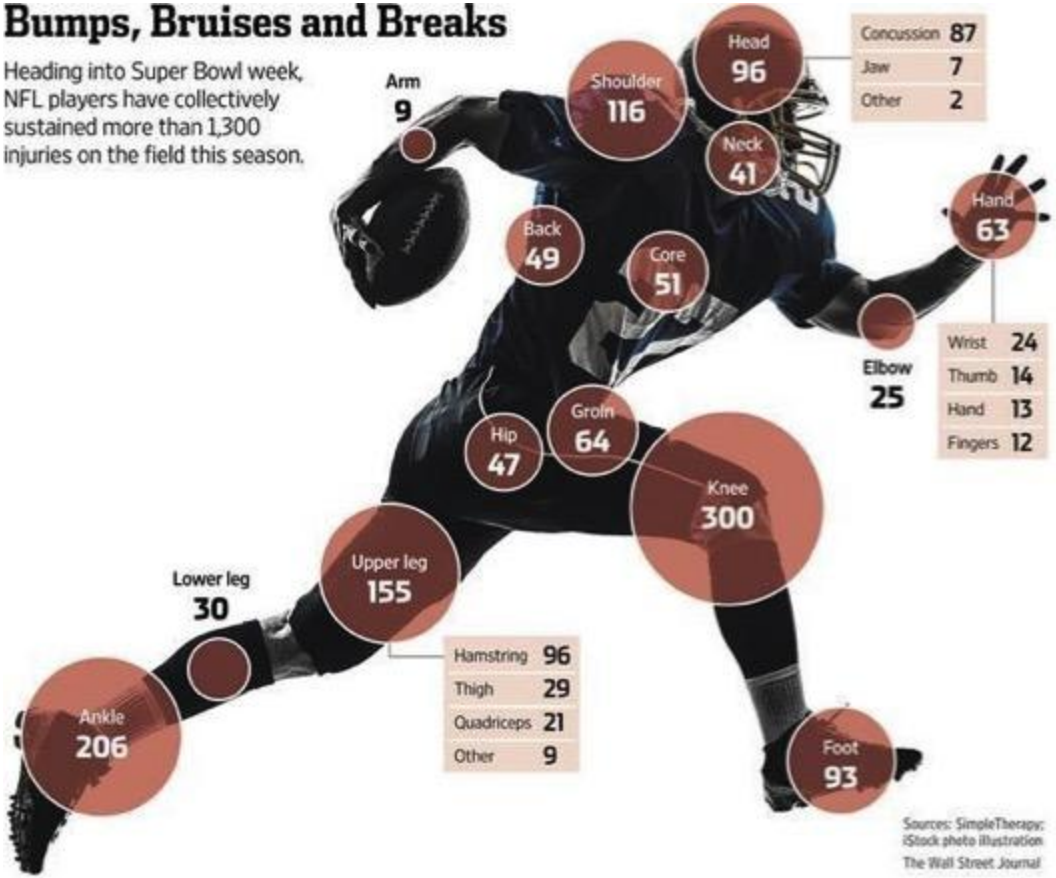
THE YEAR IN NEWS from ECHELON INSIGHTS

What America talked about in 2014, as viewed through 184.5 million Twitter mentions.



Bumps, Bruises and Breaks

Heading into Super Bowl week, NFL players have collectively sustained more than 1,300 injuries on the field this season.

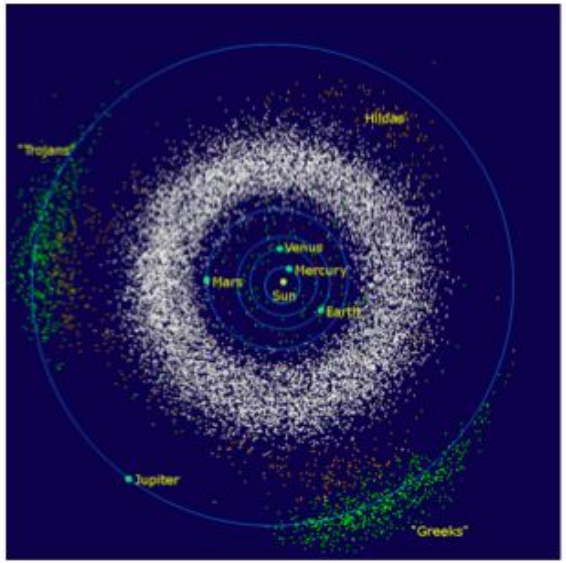


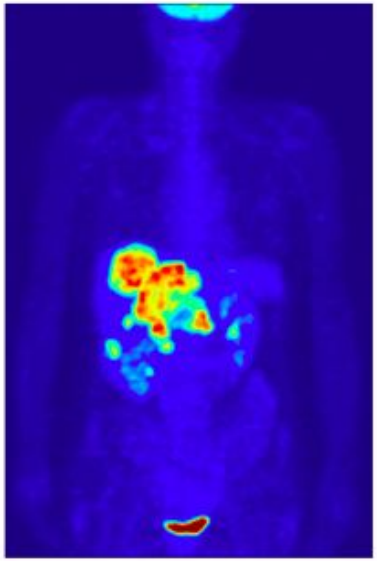
Sources: SimpleTherapy; iStock photo illustration

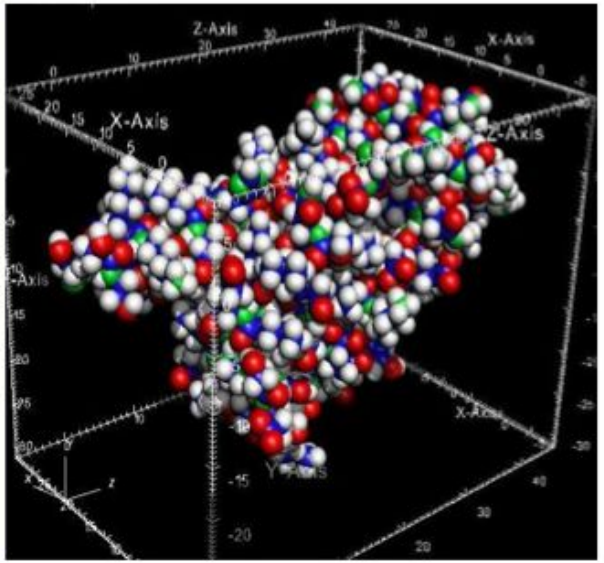
The Wall Street Journal

Types of Data Visualization

- Scientific visualization







Types of Data Visualization

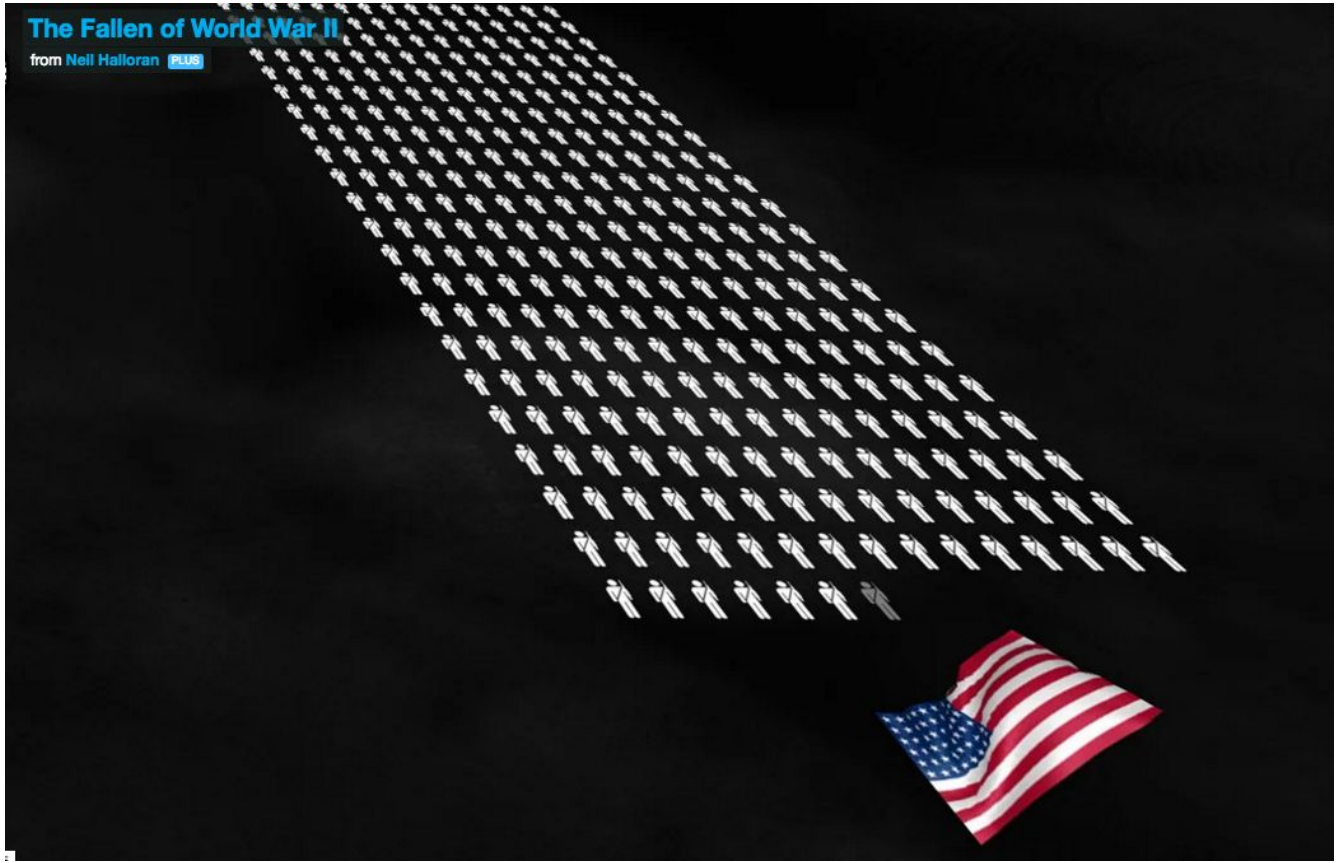
- Information visualization
  - Covers statistical charts and graphs as well as other visual/spatial metaphors that can be used to represent data sets that don't have inherent spatial components.
  - Relies more heavily on processing abstract data into a more concrete form that can be more effectively perceived by an observer

Why data visualization?

- Exploring and analyzing
- Presenting and communicating


BBC FOUR
lifespan
75 years
50 years
25 years
1927
income
$400
$4 000
$40 000

https://vimeo.com/128373915



Good Visualizations

- Present a visual interpretation of data and do so by improving comprehension, communication, and decision making
- Consider whom the visualization is targeting
- Set up a clear framework
- Tell a story

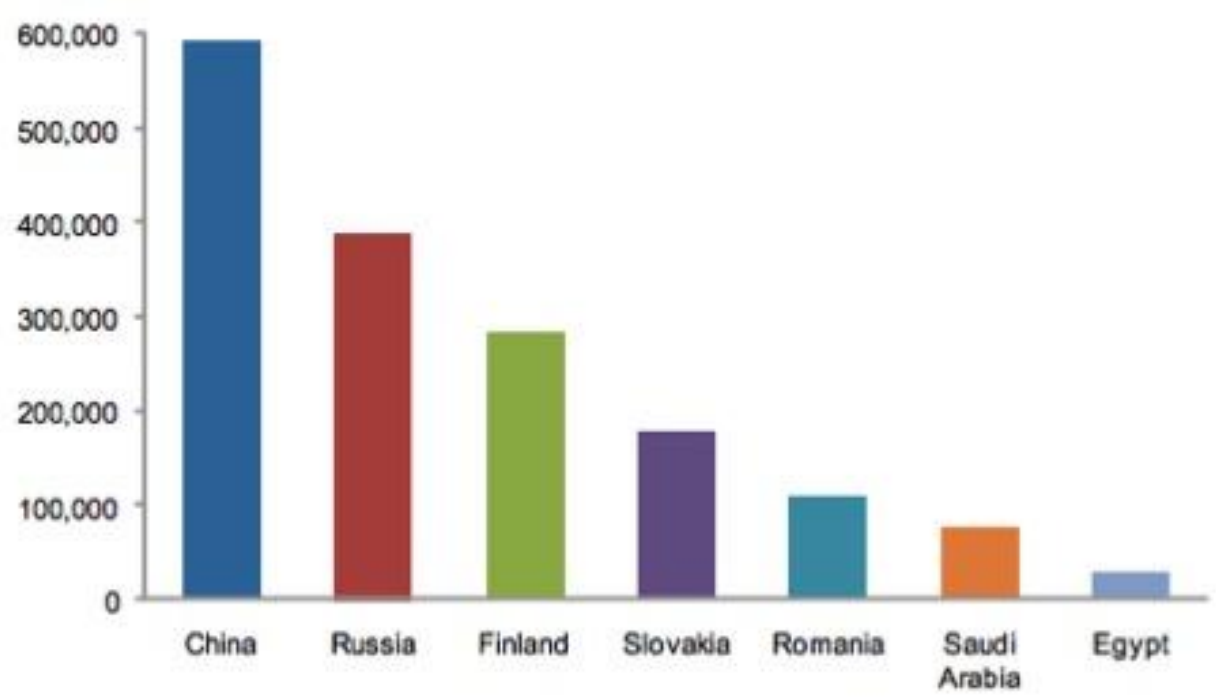
Principles of Good Visualizations

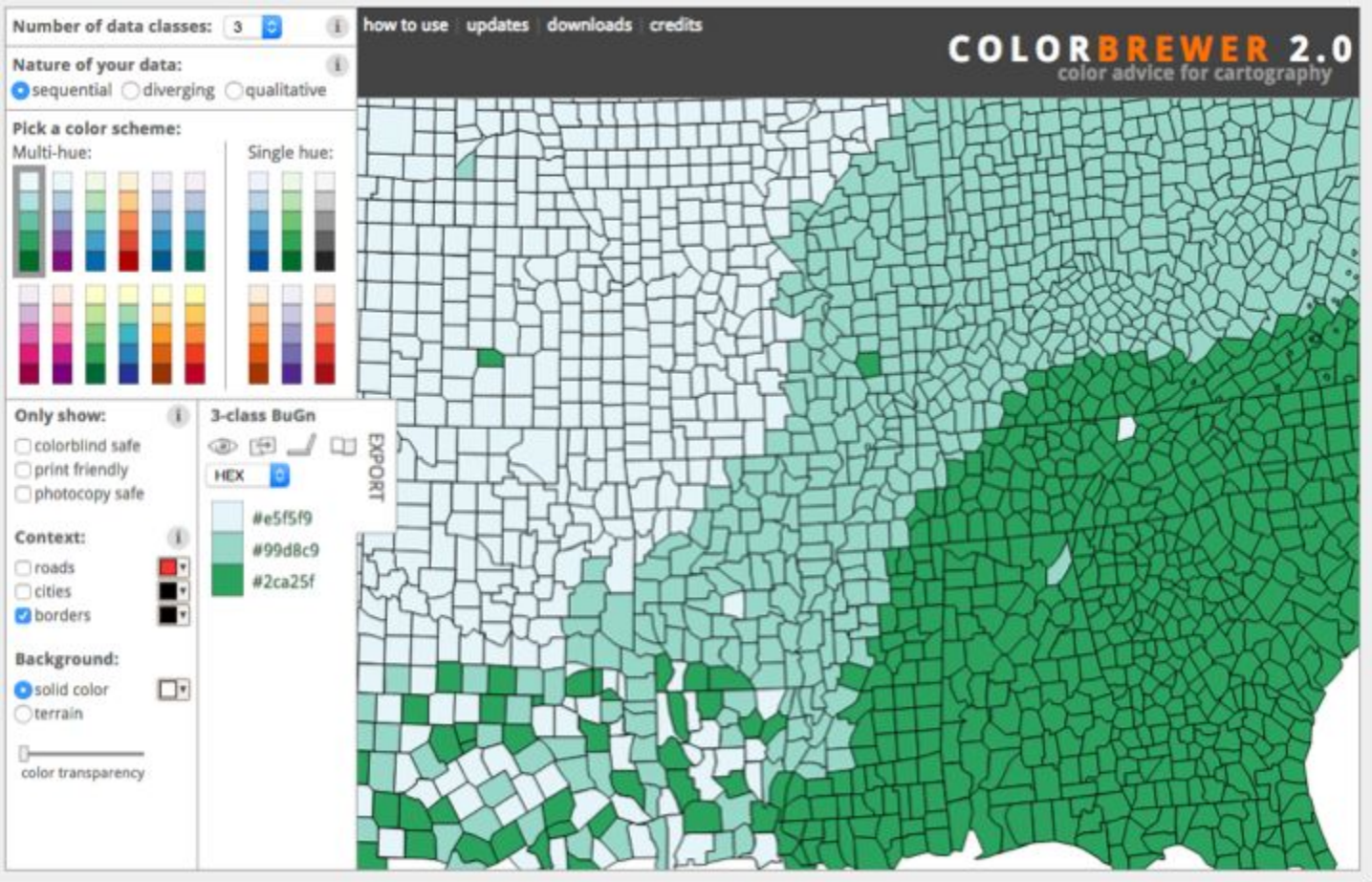
- “Above all else, show the data.”
- Help the audience think about the substance rather than about methodology (graphic design, the technology of graphic production, etc.), or something else.
- Avoid distorting what the data have to say.
- Present many numbers in a small space - but also emphasize the important values.
- Make large data sets coherent, and encourage the audience to compare different pieces of data.
- Reveal the data at several levels of detail, from a broad overview to the fine structure.

Design Principles

- Chart type
  - Select the appropriate chart type for your data and audience. Emphasize the data.
- Color
  - Use color sparingly. Use to highlight a data point.
  - Avoid decorative usage of color.
  - Only add color to an information display to communicate something in particular.
  - Use bright and/or dark colors to highlight information that requires greater attention.
  - Use lighter, soft, natural contrasting colors for the rest.
  - Consider using gray scale shading over color.
  - When encoding a sequential range of quantitative values:
    - Stick with a single hue (or a small set of closely related hues).
    - Vary intensity from pale colors for low values to increasingly darker and brighter colors for high values.

Design Principles - Color




Number of data classes: 3
Nature of your data:
sequential diverging qualitative
Pick a color scheme:
Multi-hue:
Single hue:
Only show:
colorblind safe
print friendly
photocopy safe
Context:
roads
cities
borders
Background:
solid color
terrain
color transparency
3-class BuGn
HEX
EXPORT
#e5f5f9
#99d8c9
#2ca25f
how to use | updates | downloads | credits
COLORBREWER 2.0
color advice for cartography

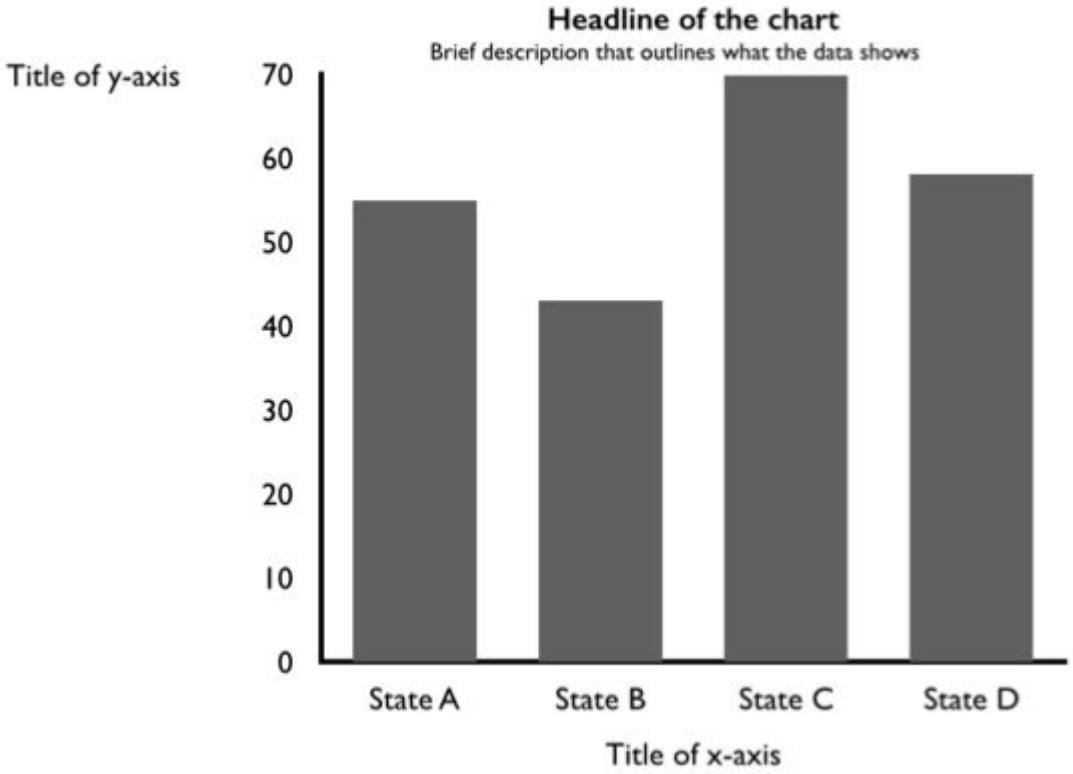
Design Principles - Labeling

- Use descriptive text and labels.
- Place label directly on the data.
- Use a legend when the chart encodings are too small to label and/or if they would impede readability.
- Add a description to guide readers in interpreting your visualization.
- Cite your data sources.

Design Principles - Text

- Readability - Font face, size, direction, and color affect the legibility.
  - Don’t set type too small or condensed.
  - Avoid all CAPS.
  - Avoid **bold** and *italic* at the same time.
  - Don’t use highly stylized fonts.
  - Do not set text at an angle or vertically.

Design Principles - Labeling & Text



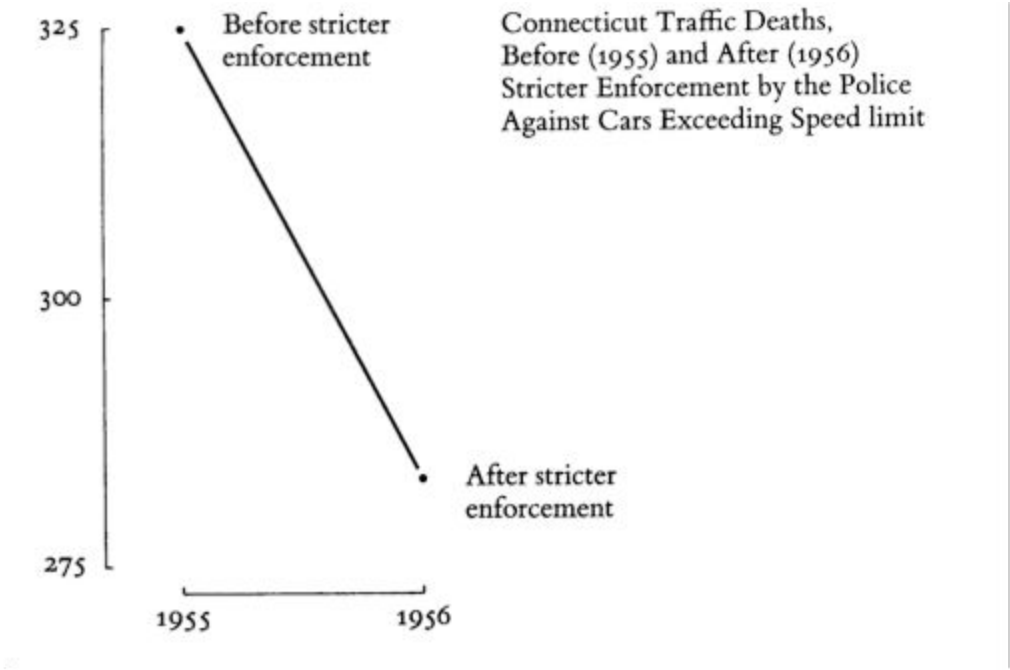
Design Principles - Scale

- Use natural increments for scales
  - 0,1, 2, 3, 4, 5
  - 0, 2, 4, 6, 8, 10
  - 0, 5, 10, 15, 20
  - 0, 10, 20, 30, 40, 50
  - 0, 25, 50, 75, 100
  - 0, 0.25, 0.50, 0.75, 1.00
- Avoid awkward scale increments
  - 0, 3, 6, 9, 12, 15
  - 0, 4, 8, 12, 16, 20
  - 0, 6, 12, 18, 24, 30
  - 0, 12, 24, 36, 28
  - 0, 0.4, 0.8, 1.2, 1.6

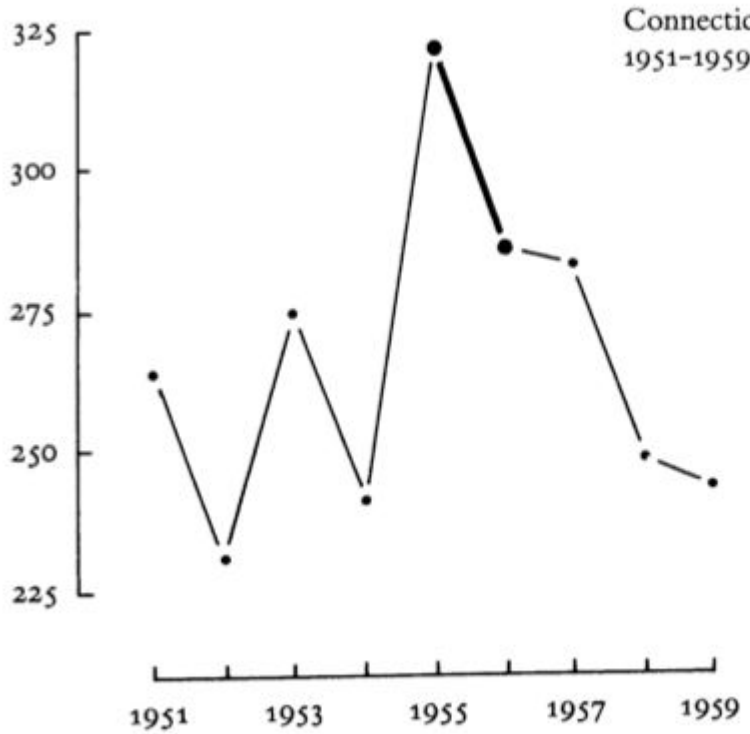
Design Principles - Data Integrity

- Show your data accurately and avoid distortions.
- Avoid fake perspectives, such as 3D.
- The graphics should bear the question “compared to what?” – presented within the right context.

Design Principles - Data Integrity



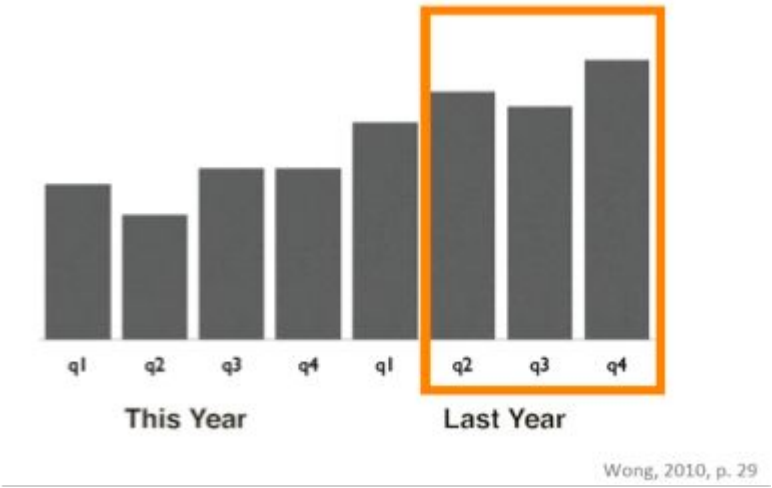
Design Principles - Data Integrity



Context is essential for data integrity.

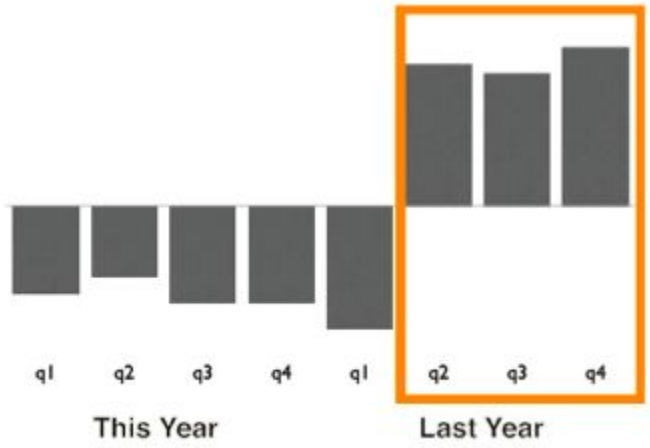
Design Principles - Data Integrity

It is acceptable to extract a few numbers out of a series if these data points tell a story without misleading the reader.



Design Principles - Data Integrity

It would be misleading to extract the last three quarters in the case below.



Wong, 2010, p. 29

Design Principles - Data Integrity

Provide context for your visualizations.

Design Principles - Chart Junk

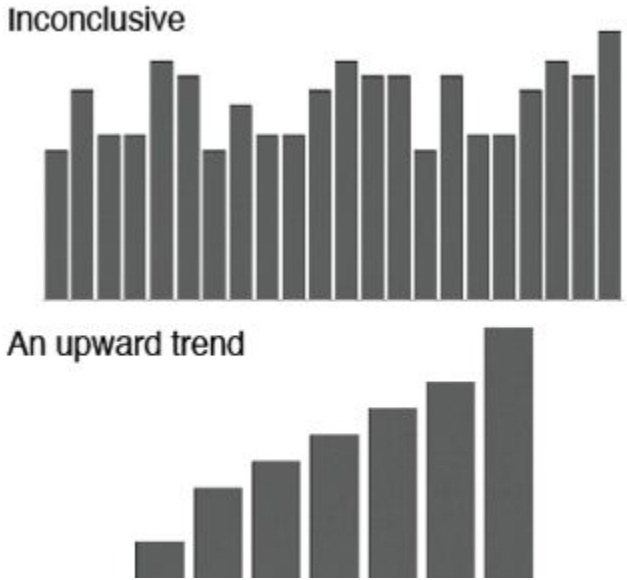
- Avoid chart junk.
- Useless, non-informative, or information-obscuring elements of quantitative information displays.

--Edward Tufte

- Reduce non-data graphic elements (e.g. reduce the thickness of the bars in a bar chart).
- Remove the grid (or use a light gray grid) and non-essential elements.
- Avoid using shadows.
- Stick to white or match the chart background.

Design Principles - Data Richness

Accurate data and effective filtering of your data based on audience.



Process of creating and selecting appropriate visual displays

Identify the following:

1. Audience: Who will be viewing and/or interacting with your visual displays?
2. Task: What is the message of your display? Is there something you want the reader to take away from your visual?
3. Data: Do you have the data to achieve the task? What are the tables/fields? Does the data need to be aggregated, transformed, etc?
4. Display: What is the best display type for my task, data, and audience? Do I want to show a pattern, relationship, proportions, comparisons, or distributions?

Tableau

https://www.tableau.com/academic/students