Data Science and Business Intelligence

BU.330.780

Session 2

Instructor: Changmi Jung, Ph.D.

Announcement

- Due next week
  - Assignment #1: Canvas > Week 3
    - For Q5: Render your R markdown file (.rmd file) to MS Word or HTML Document
  - Project Proposal: Modules > Project > Deliverables
    - Only the team leader submits the proposal
    - Max one-page Word document (no restriction on the structure)
    - Optional but highly recommended

Revisit: Data Mining

- Objectives:
  - find **patterns** and **relationships**
  - **classify** and **predict**

- There is a large number of data mining algorithms available, but only a limited number of data mining tasks
  - Classification/ Regression/ Clustering
  - Co-occurrence grouping
  - Similarity matching / Profiling
  - Link prediction

Lets talk about some of the tasks to improve "Business Understanding" step

- Decompose a data analytics problems into pieces such that you can solve a known task with a tool

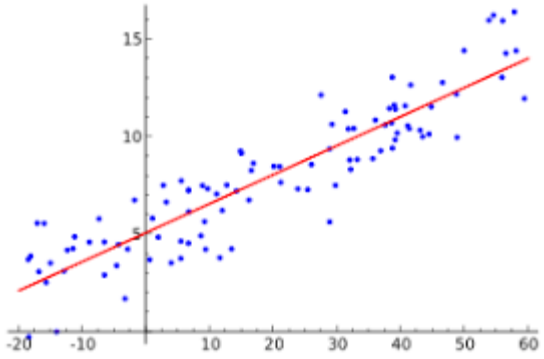
- **Analysts' creativity** plays an important role

Regression

- Regression (value estimation) attempts to estimate or predict, for each individual, the numerical value for that individual

> "How much will a given customer use the service?"
> → Predicted variable: service usage
>
> "How much profit margin can we expect a new customer to generate?"
> → Predicted variable: amount of profit margin

- Generate regression model by looking at other, similar individuals in the population



Classification

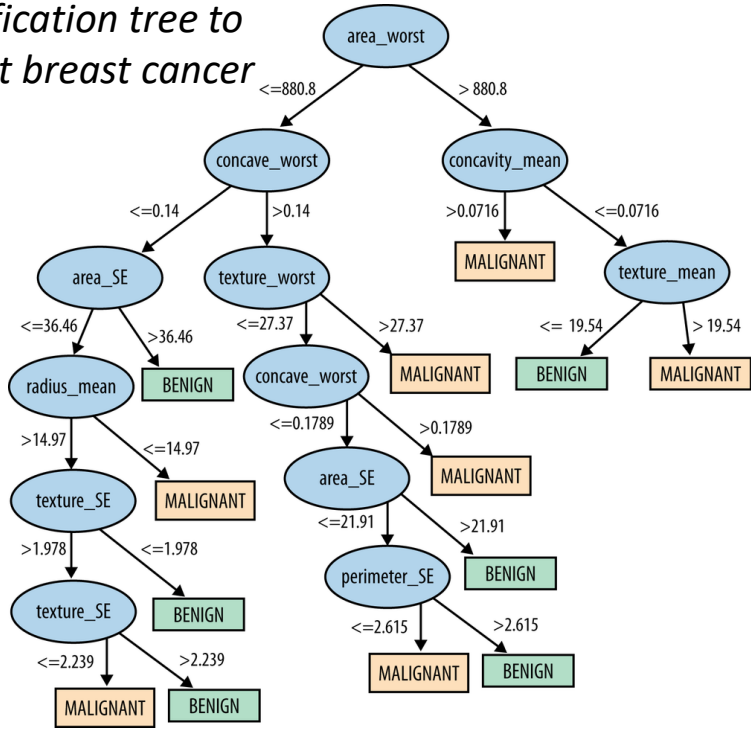
- Classification attempts to predict, for each individual in a population, to which of a (small) set of classes that individual belongs.

"Among all the customers of a cell phone carrier, which are likely to correspond to a given offer?"

"Which patients are likely to develop breast cancer?"

- Classification is related to scoring.

Classification tree to predict breast cancer



Q: Classification vs. Regression?

- The difference is the type of predicted (target) variable:
  - Classification → categorical target
  - Regression → numeric target

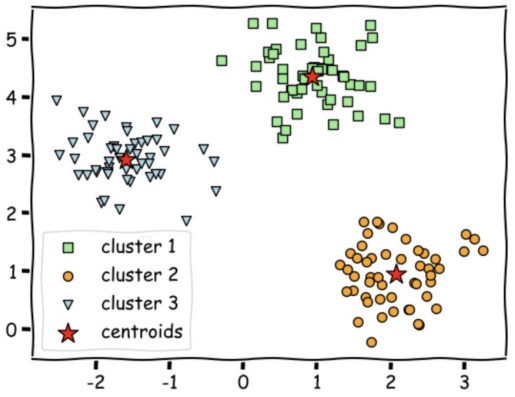
- Which one is classification, or which is regression?
  - Will this customer default on his/her loan? (e.g., yes/no)
  - Which service package (S1, S2, or none) will a customer purchase if offered the proposed incentive?
  - How frequently will this customer use the parcel service?
  - What is the probability that this prospect will default on his/her loan? (e.g., 78%)

Huh? It's confusing!

Clustering

- Clustering attempts to group individuals in a population together by their similarity, but without regard to any specific purpose

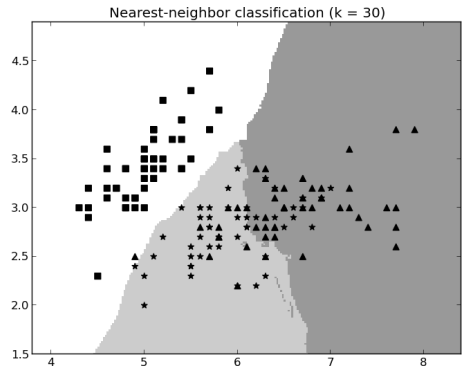
Do customer form natural groups or segments?



- Result: groupings of the individuals of a population

- Useful in preliminary domain exploration

vs



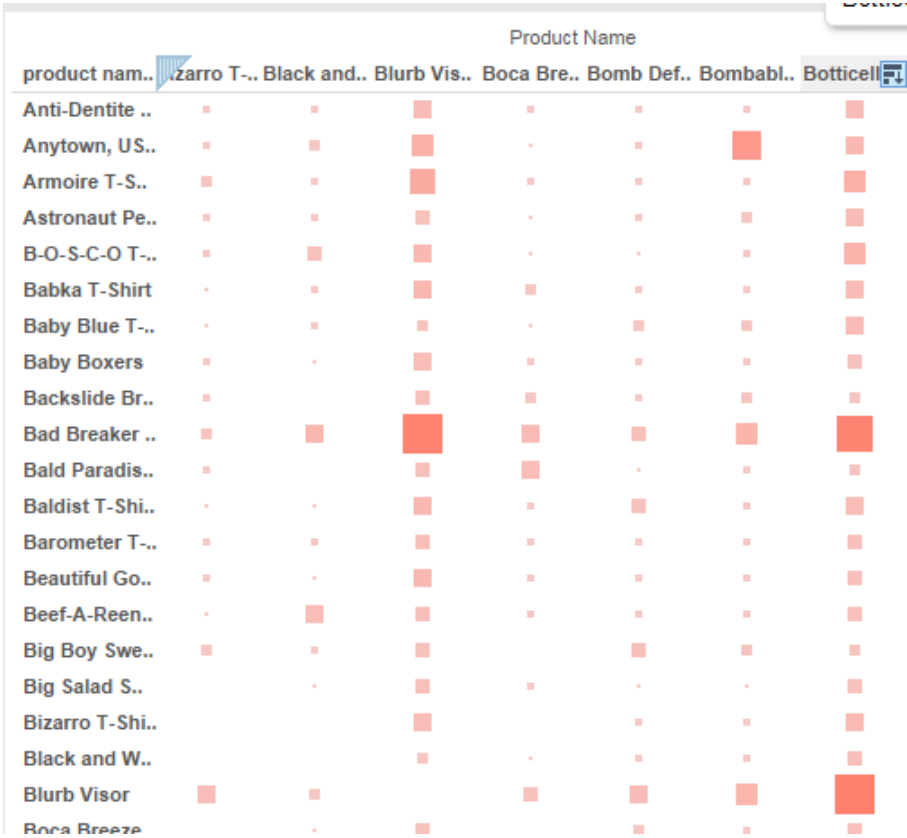
https://www.naukri.com/code360/library/knn-vs-k-means

Co-occurrence Grouping (Association Rule)

- Attempts to find associations between entities based on transactions involving them (a.k.a. association rules or market-basket analysis)

What items are commonly purchase together?

- Considers similarity of objects based on their appearing together in transactions
- Included in many recommendation systems
- Result: a description of items that occur together



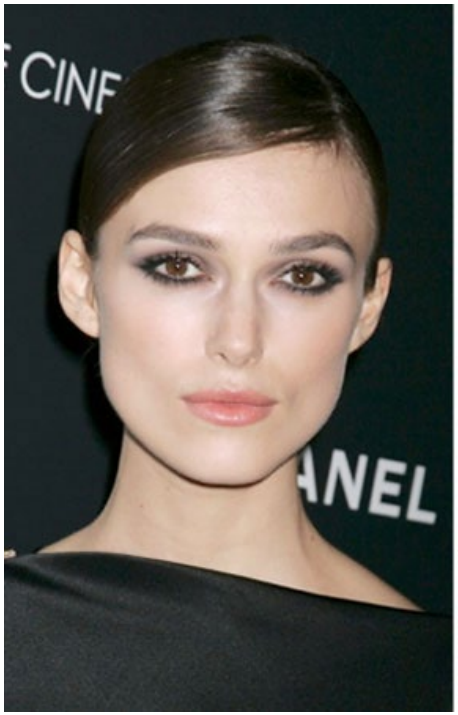
Similarity Matching

- Similarity matching attempts to identify similar individuals based on the data known about the individuals

- Can be used as a basis for making product recommendations

Find people who are similar to you in terms of the products they have liked or purchased

Find items that are similar to this product A based on the dimension, material, price, color, the number of compartments, etc.



Profiling

- Attempts to characterize the typical behavior of an individual, group or population

- A.k.a behavior description

  What is the typical cellphone usage of this customer segment?

  Where and how do you access your account normally?

- Often used to establish behavioral norms for anomaly detection (fraud detection)

Q: Business Question Examples

- What items are commonly purchased together at Barnes & Noble?

→ Co-occurrence grouping

- You are working for AMEX. What other companies are like our best small business customers?

→ Similarity Matching

- What does “normal behavior” in credit card spending look like among cardholders in group A?

→ Profiling

- Do my customers form natural groups?

→ Clustering

Recap: Supervised vs. Unsupervised

- Is there a specific, quantifiable target that we are interested in or trying to predict?
  - If yes, supervised. Otherwise, unsupervised.

- Examples
  - "Do our customers naturally fall into different groups?"

    : no specific target → unsupervised

  - "Can we find groups of customers who have particularly high likelihoods of cancelling their service soon after their contracts expire?"

    : specific target → supervised

We talked about it last week.

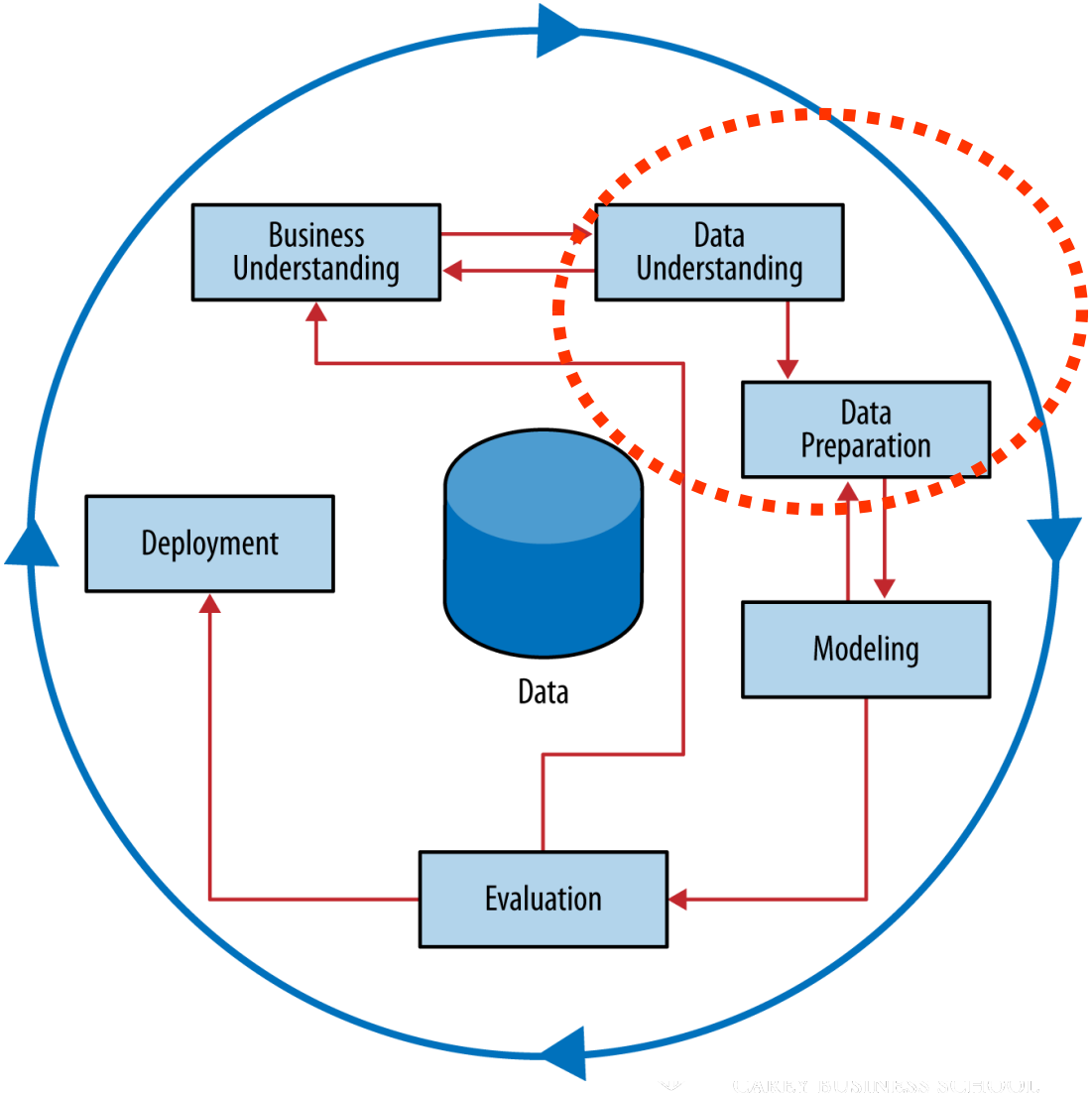
Supervised vs. unsupervised

- Supervised and unsupervised tasks require different techniques
- Classification and regression are generally solved with supervised techniques
- Clustering, co-occurrence grouping, and profiling are generally unsupervised
- Similarity matching can be either supervised or unsupervised based on the purpose and algorithms used (Sup: face recognition, signature verification)
- There is no guarantee that unsupervised tasks provide meaningful results

Data Understanding & Preparation Exercise

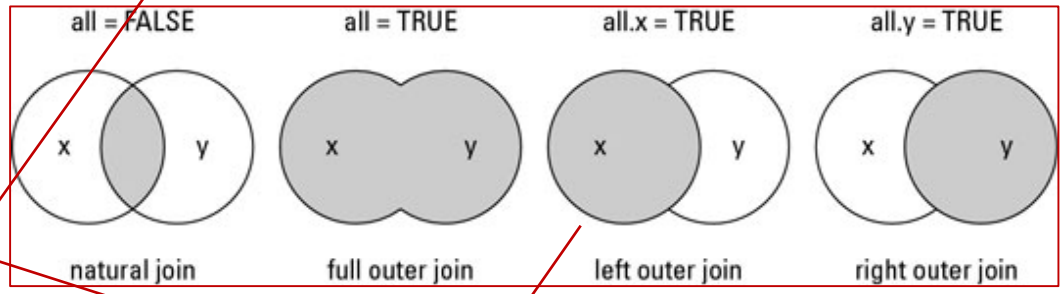
Data Understanding & Preparation

- Goals: Gain **general insights** about the data (independent of the project goal)

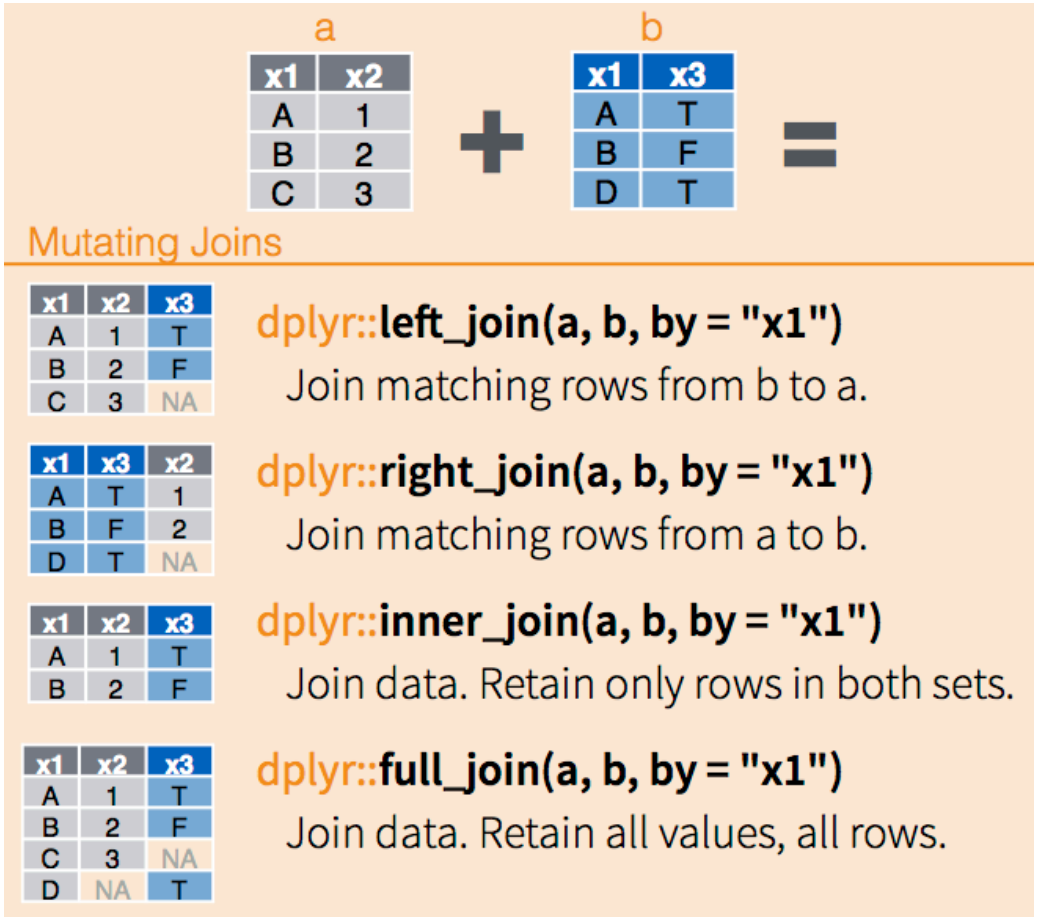


Merging dataset (Basic R)

```
> head(order)
  customer_id order_id order_short_date order_total zip_code
1           1        1       2009-01-01      404.72    32435
2           1    33658       2013-02-15       71.10    32435
3           2        2       2009-01-01      288.62     1099
4           2     1653       2010-01-03      182.86     1099
5           2     4193       2010-09-12      108.43     1099
6           3        3       2009-01-01       27.15    66063
> head(zipcode)
  zip state
1 501    NY
2 544    NY
3 601    PR
4 602    PR
5 603    PR
6 604    PR
> order<-merge(order,zipcode,by.x="zip_code",by.y="zip",all.x=TRUE)
> head(order)
  zip_code customer_id order_id order_short_date order_total state
1    10013        3881     7382       2011-03-26       93.68    NY
2    10013        3881    11784       2011-09-30      360.39    NY
3    10013        3881    50215       2013-09-21       23.12    NY
4    10017         568      712       2009-08-06      237.23    NY
5    10017         568    24273       2012-09-04       30.08    NY
6    10017         568    21458       2012-07-05       34.35    NY
```



Merging dataset (in a tidy way)



a

x1	x2
A	1
B	2
C	3

+

b

x1	x3
A	T
B	F
D	T

=

Mutating Joins

x1	x2	x3
A	1	T
B	2	F
C	3	NA

dplyr::**left_join(a, b, by = "x1")**

Join matching rows from b to a.

x1	x3	x2
A	T	1
B	F	2
D	T	NA

dplyr::**right_join(a, b, by = "x1")**

Join matching rows from a to b.

x1	x2	x3
A	1	T
B	2	F

dplyr::**inner_join(a, b, by = "x1")**

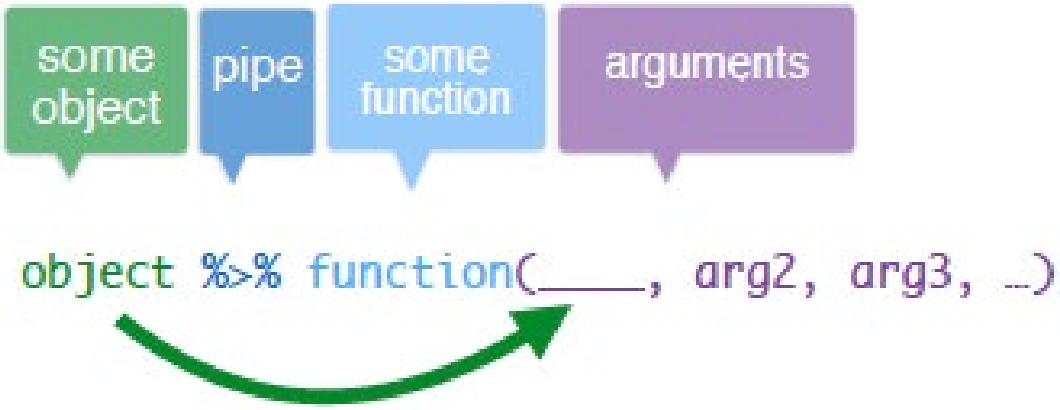
Join data. Retain only rows in both sets.

x1	x2	x3
A	1	T
B	2	F
C	3	NA
D	NA	T

dplyr::**full_join(a, b, by = "x1")**

Join data. Retain all values, all rows.

Pipe Operation (%>%)

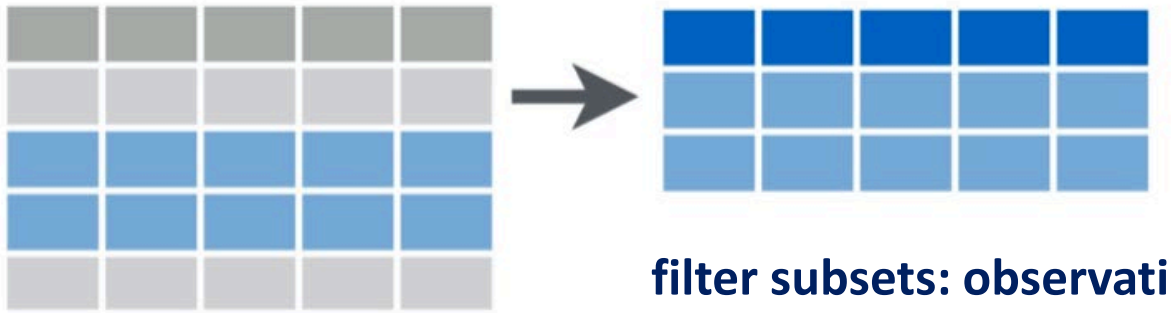


For example,

filter(data, state == “CA”) → data%>%filter(state == “CA”)

Filter(), Select() – Creating Subset

filter( )



filter subsets: observations

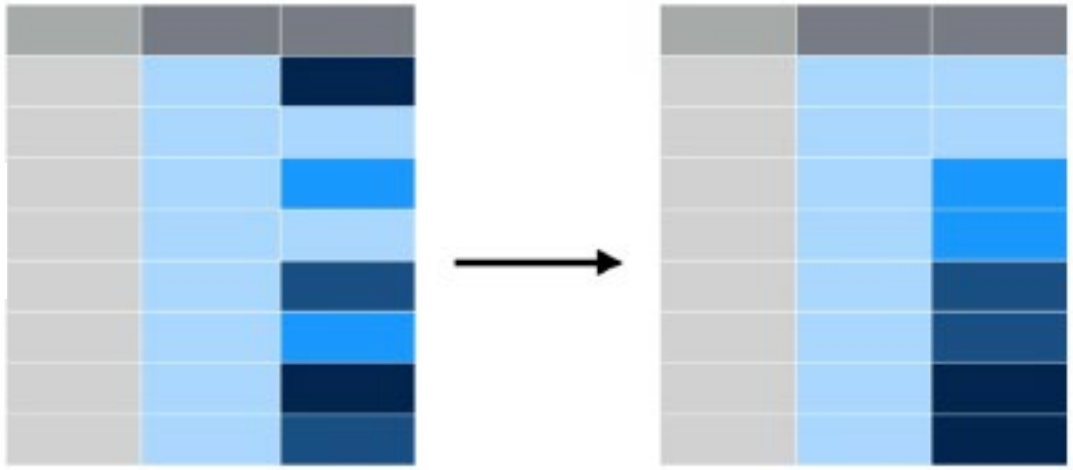
select ( )

filter subsets: variables

Arrange() and Mutate()

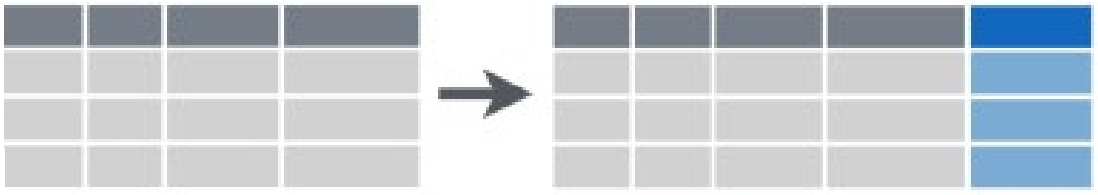
arrange( )

:sorts a table based on a variable

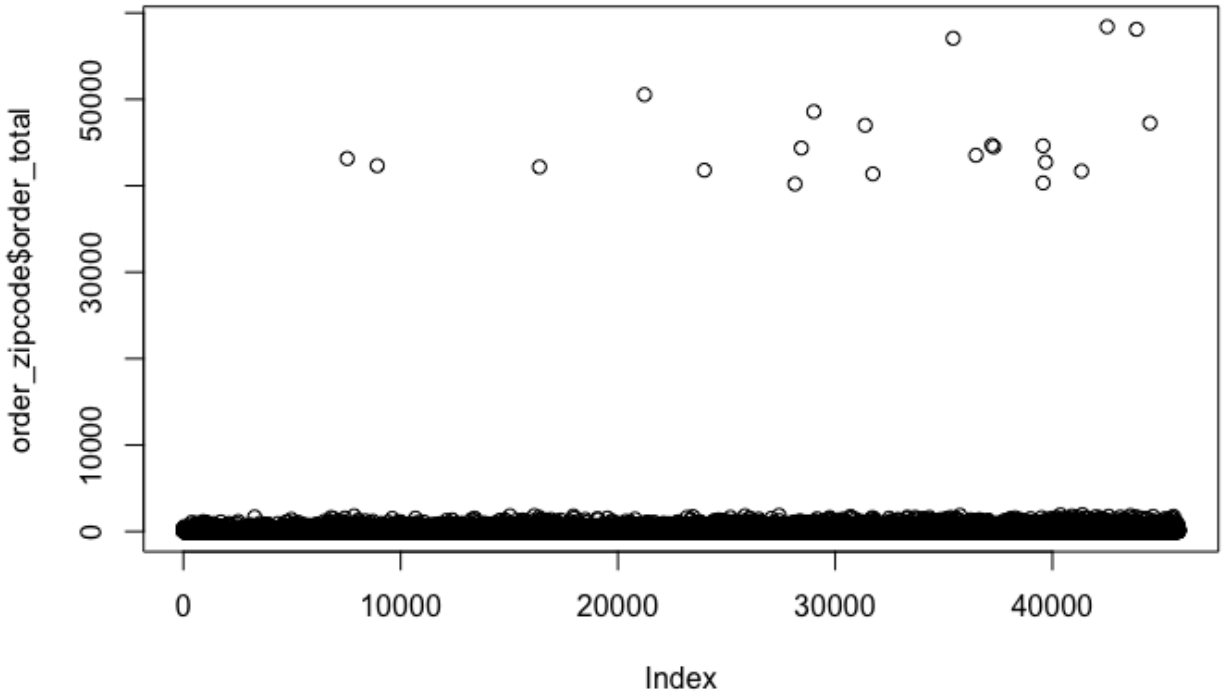


mutate( )

:changes or adds variables



Be Careful in Cleansing Data

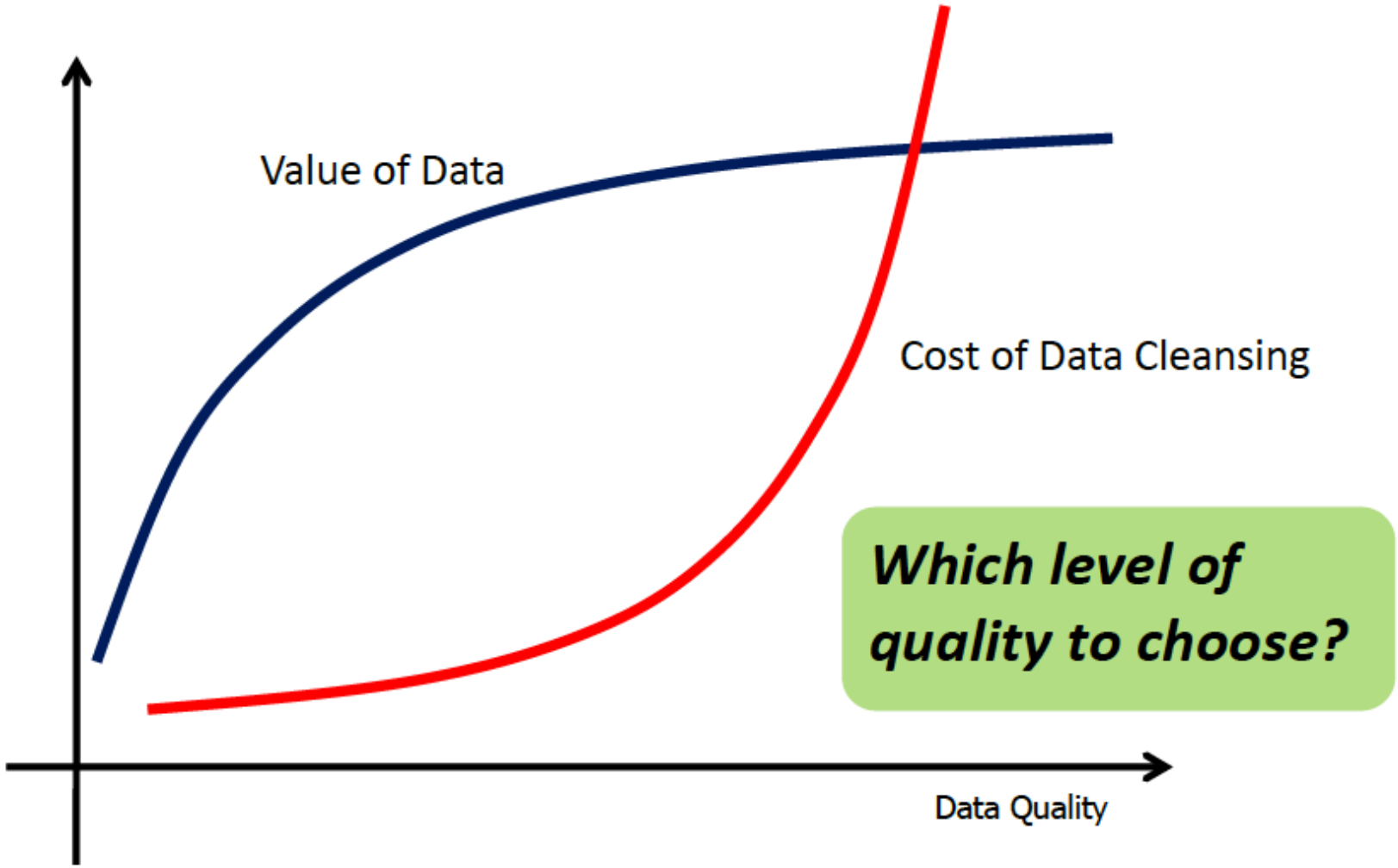


- Are they outliers?
- Do we need to "fix" it?
- It depends on cause.

What you will do with them depends on the business contexts, understanding of the data collection, and the purpose of the analytics.

- If yes, how can we tell if it is a real error or an unusual but correct data?

Trade-Off in Data Cleansing

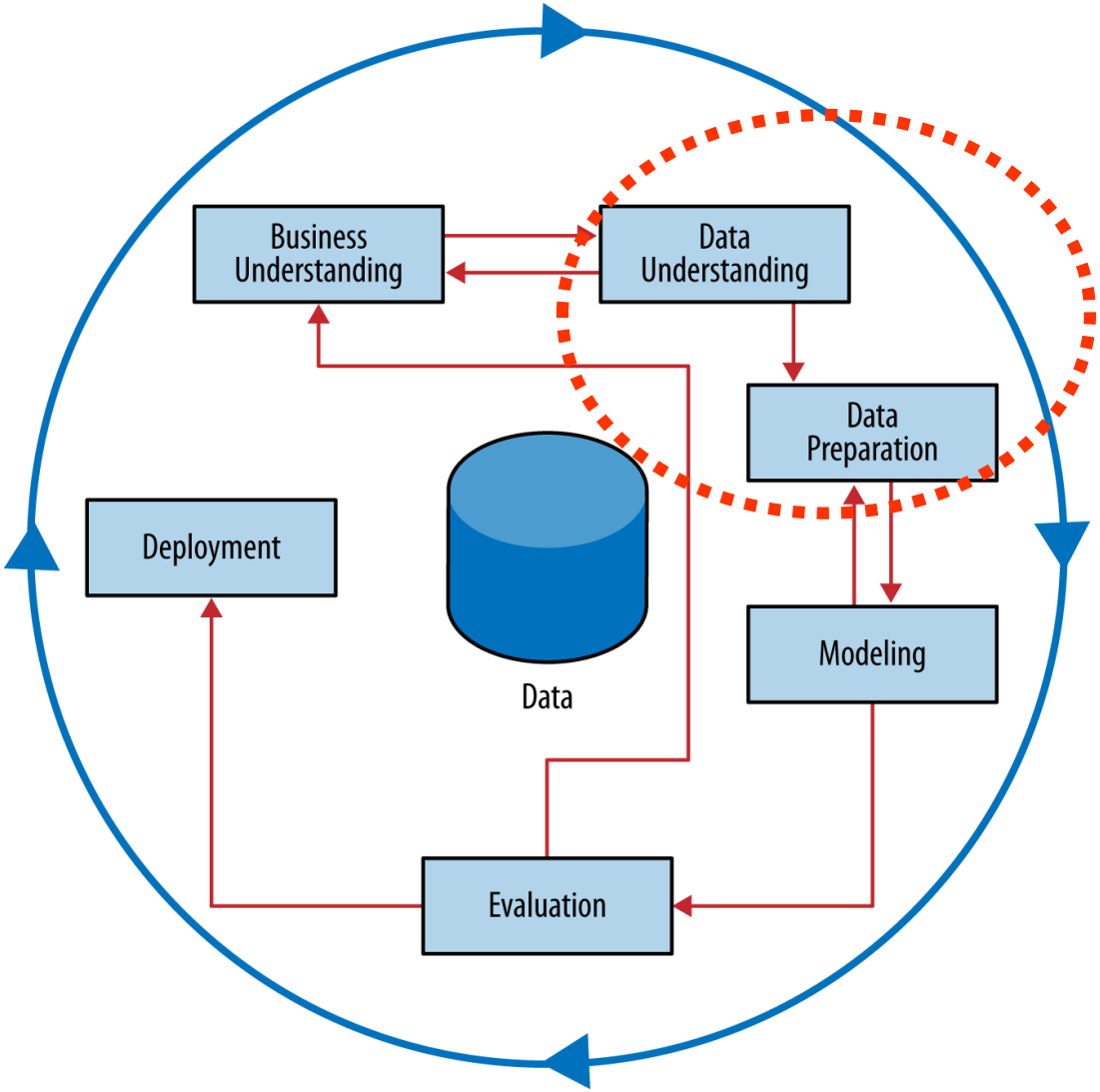


Data Visualization

Data Visualization

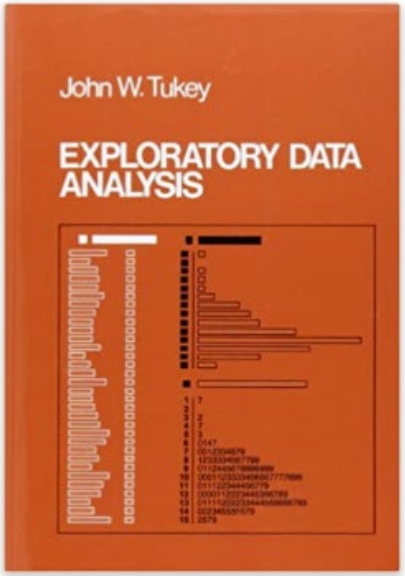
- Goals: Gain **general insights** about the data (independent of the project goal)

It may reveal many things about the data that may not be captured by a simple descriptive statistics summary!

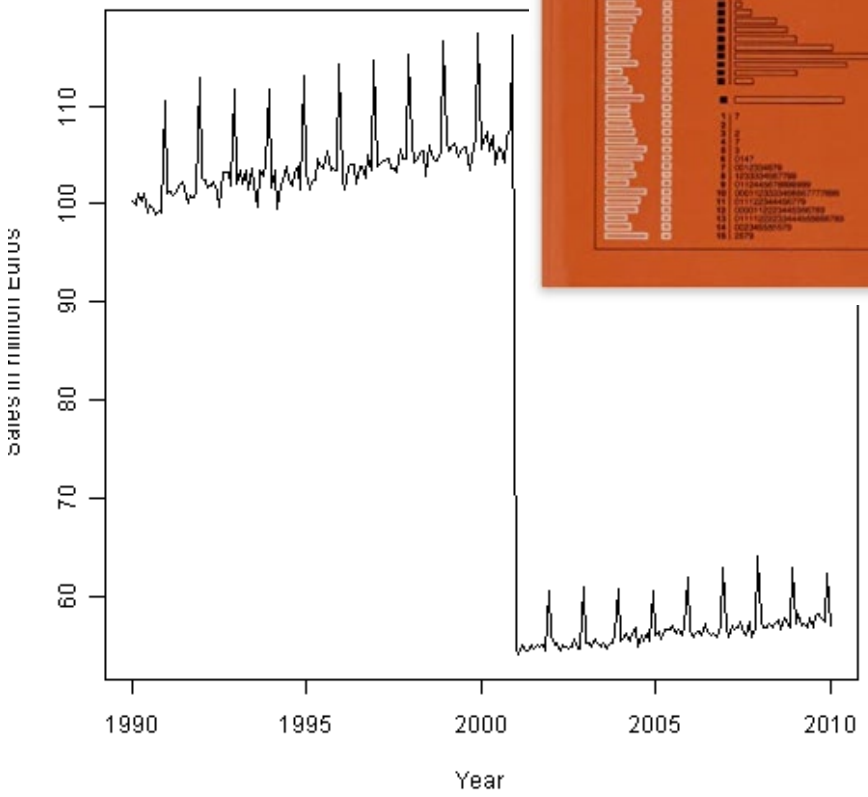


Simple Sanity Check!

"There is no excuse for failing to plot and look"
–Tukey (1977)

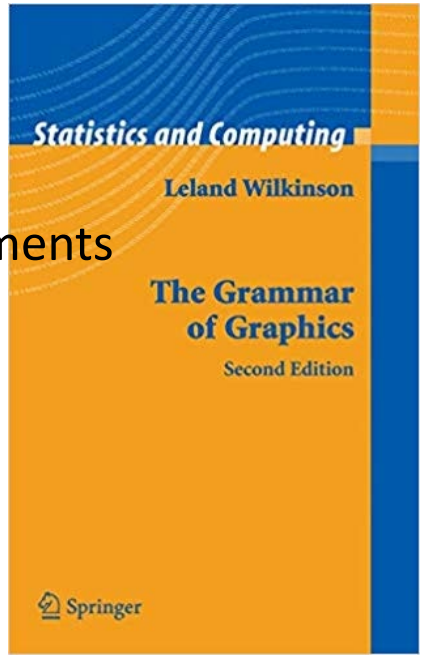


- Checking the assumptions made during the business understanding phase.
- **Never trust any data** as long as you have not carried out some simple plausibility checks!



ggplot: Grammar of Graphics

- Plotting Framework
- Two principles
  - Graphics = combination of distinct layers of grammatical elements
  - Meaningful plots through aesthetic mapping



Leland Wilkinson, Grammar of Graphics, 1999

Element	Description
Data	The dataset being plotted.
Aesthetics	The scales onto which we *map* our data.
Geometries	The visual elements used for our data.
Facets	Plotting small multiples.
Statistics	Representations of our data to aid understanding.
Coordinates	The space on which the data will be plotted.
Themes	All non-data ink.

data

Mapping rule

Chart Type

Layers

ggplot2 Layers

Element	Description
Data	The dataset being plotted.
Aesthetics	The scales onto which we *map* our data.
Geometries	The visual elements used for our data.
Facets	Plotting small multiples.
Statistics	Representations of our data to aid understanding.
Coordinates	The space on which the data will be plotted.
Themes	All non-data ink.

```
ggplot(data, aes(x=var1, y=var2))+
  geom_point(aes(col=var3))+
  geom_smooth() +
  theme_bw()
```

```
crimes%>%
  ggplot(aes(x=violent_crime_rate,y=property_crime_rate))+
  geom_point(shape=1)+
  geom_smooth(method=lm,fullrange=TRUE)+
  facet_wrap(~region)
```

Data
Aesthetics
Geometries
Statistics
Facets

Typical Aesthetics

Aesthetic	Description
x	X axis position
y	Y axis position
colour	Colour of dots, outlines of other shapes
fill	Fill colour
size	Diameter of points, thickness of lines
alpha	Transparency
linetype	Line dash pattern
labels	Text on a plot or axes
shape	Shape

No need to memorize! Just play with it.

ggplot2 compared to base R graphic package

- Creates plotting objects, which can be manipulated (by modifying and adding elements and layers).

- It takes care of a lot of the legwork for you, such as choosing nice color palettes and making legends.

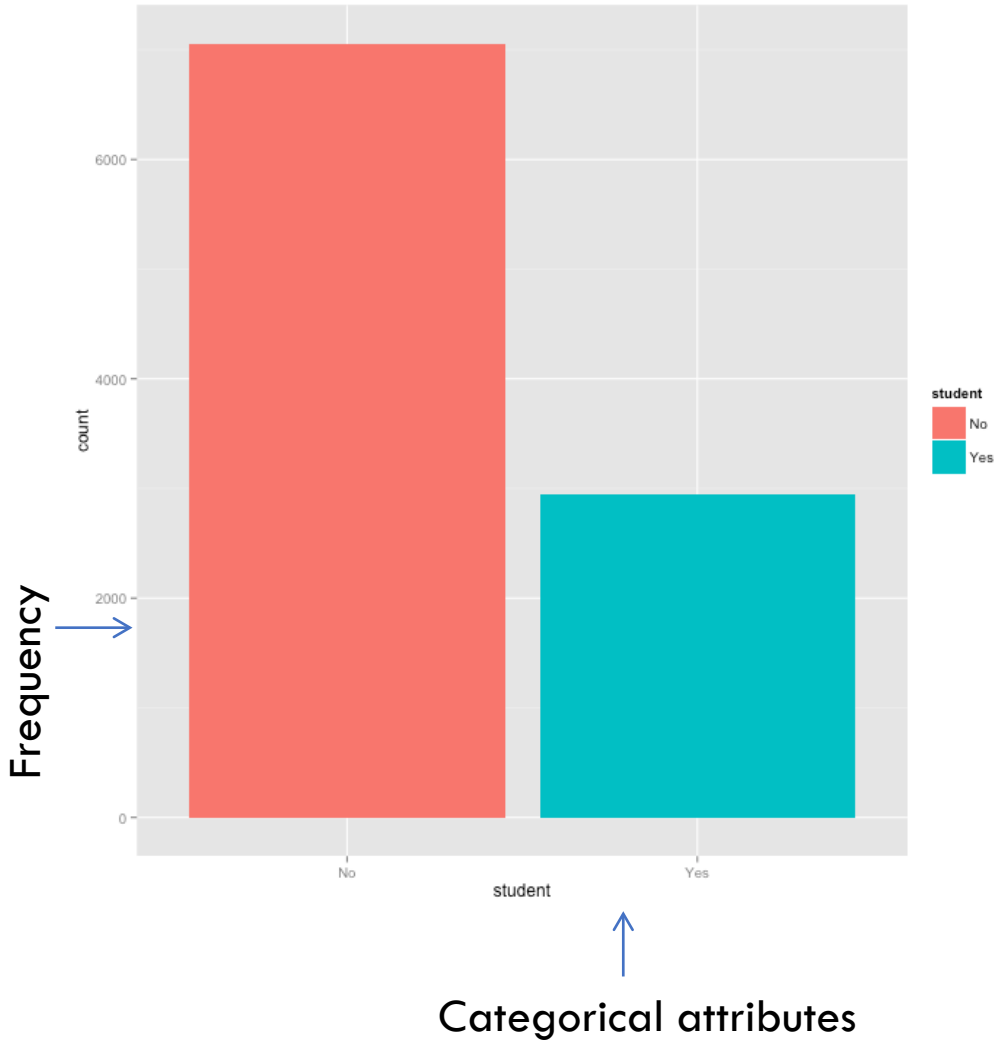
- Is built upon the grammar of graphics plotting philosophy, making it more flexible and intuitive for understanding the relationship between your visuals and your data.

Bar charts

- A bar chart is a simple way to depict the **frequencies/volumes** of the values of a **categorical** attribute.

Among the customers who defaulted, how many are students and non-students?

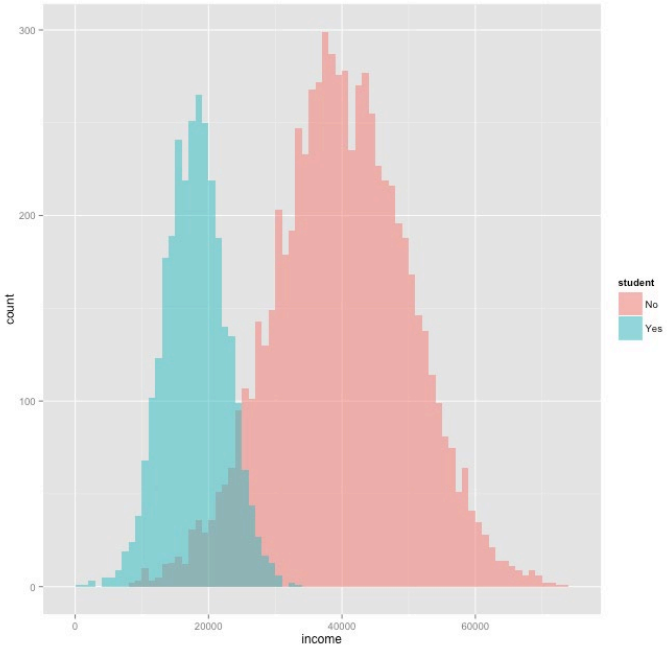
- What about numerical variables?

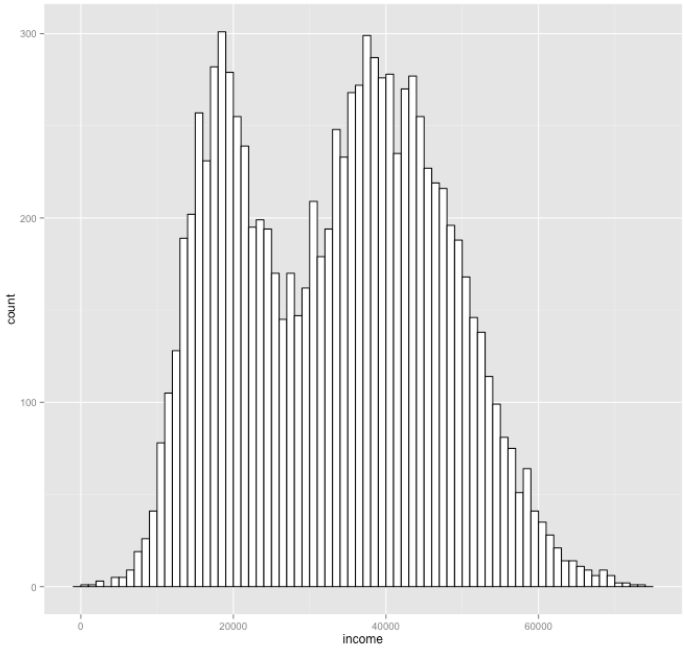


Histograms

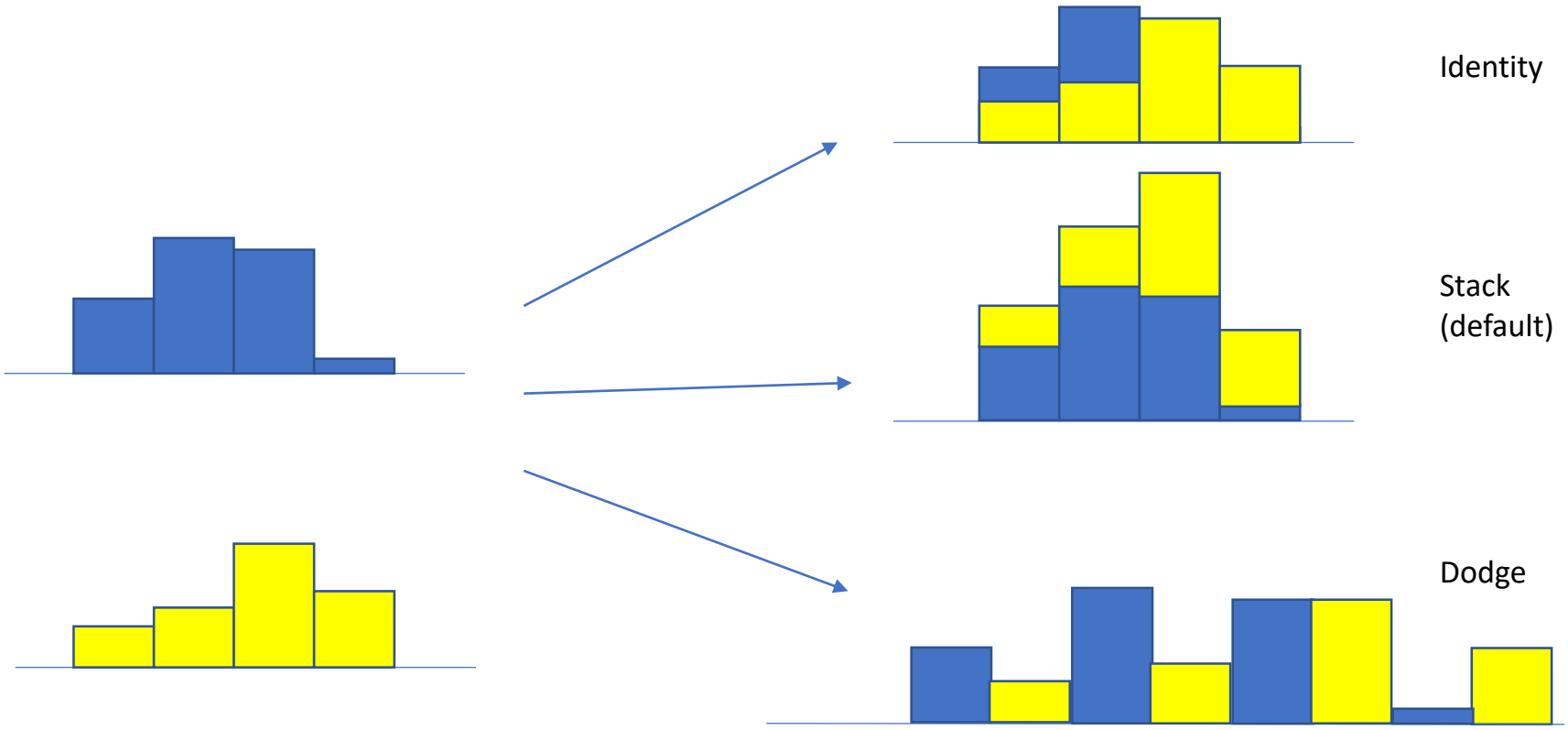
- A histogram shows the **frequency** distribution for a **numerical** attribute.
- The range of numerical attribute is discretized into a fixed number of intervals ("bins"), usually of equal length. For each interval, the (absolute) frequency of values falling into it is indicated by the height of a bar.

How many data observations have the exact value of 2.731?

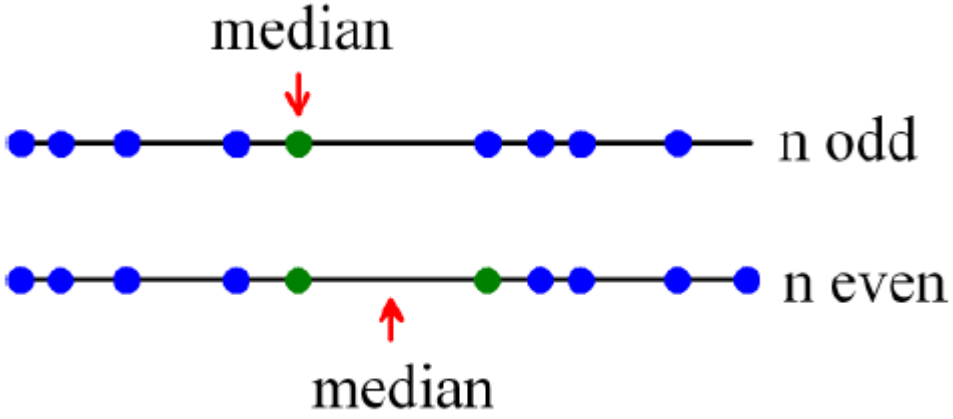




geom_histogram (position= _____)



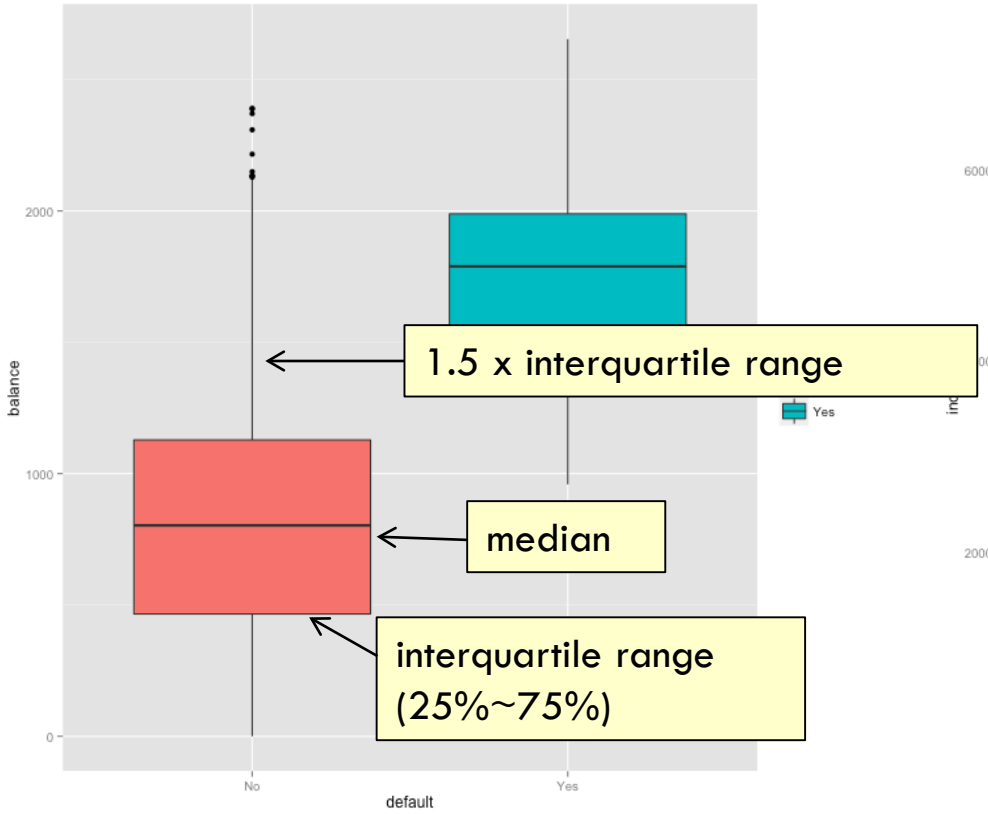
Review: median, quantiles, quartiles, IQR

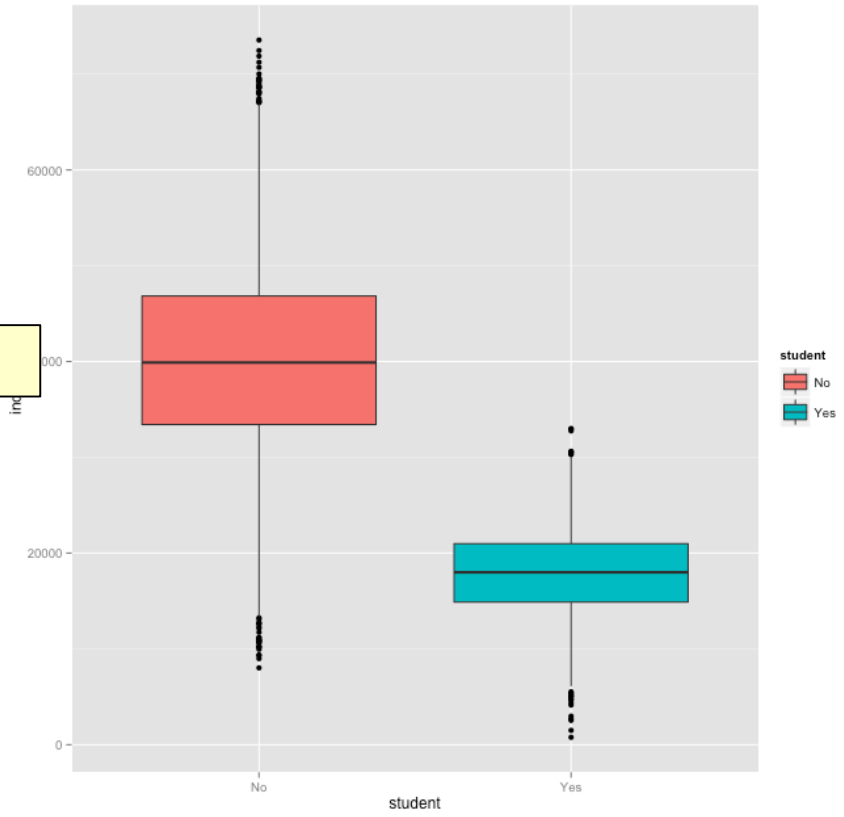


- Median: the value in the middle (for the values given in increasing order)
- q%- quantile (0<q<100): The value for which q% of the values are smaller and 100-q% are larger. The median is the 50%-quantile
- Quartiles: 1st (25% quantile), 2nd (median), 3rd (75% quantile)
- Interquartile range: 3rd quartile – 1st quartile

Boxplots

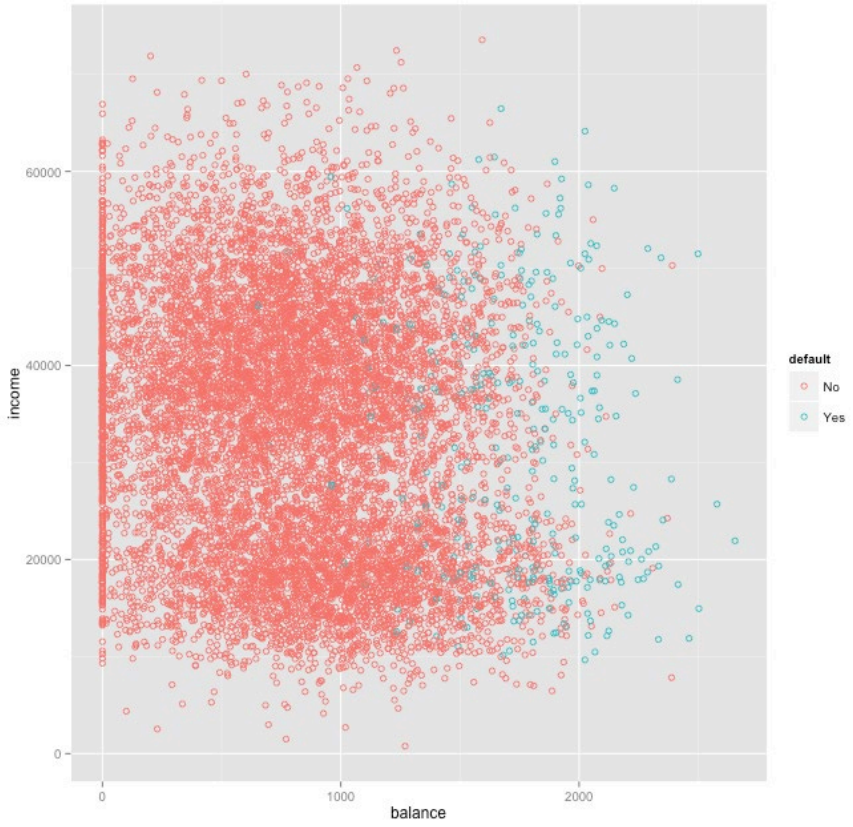
- Boxplots are a very compact way to visualize and **summarize** main characteristics of a sample from a **numeric** attribute





Scatter Plots

- Scatter plots visualize **two variables** in a two dimensional plot
- Each axes corresponds to one variable
- Can “visually” reveal correlations or dependencies between two variables



Let’s Check!

More on R Graphics with ggplot2

www.cookbook-r.com/Graphs/

Apps | WSJ-Tech | NYT-Tech | NYT Tech | Databases @ JH: Bu | Carey library | JH Carey Business S | WRDS | Instructions for Auth | HB Case | Listonic

Cookbook for R » Graphs

Graphs

My book about data visualization in R is available! The book covers many of the same topics as the Graphs and Data Manipulation sections of this website, but it goes into more depth and covers a broader range of techniques. You can preview it at Google Books.

Purchase it from Amazon, or direct from O'Reilly.

There are many ways of making graphs in R, each with its advantages and disadvantages. The focus here is on the ggplot2 package, which is based on the Grammar of Graphics (by Leland Wilkinson) to describe data graphics.

Graphs with ggplot2

1. Bar and line graphs (ggplot2)
2. Plotting means and error bars (ggplot2)
3. Plotting distributions (ggplot2) - Histograms, density curves, boxplots
4. Scatterplots (ggplot2)
5. Titles (ggplot2)
6. Axes (ggplot2) - Control axis text, labels, and grid lines.
7. Legends (ggplot2)
8. Lines (ggplot2) - Add lines to a graph.
9. Facets (ggplot2) - Slice up data and graph the subsets together in a grid.
10. Multiple graphs on one page (ggplot2)
11. Colors (ggplot2)

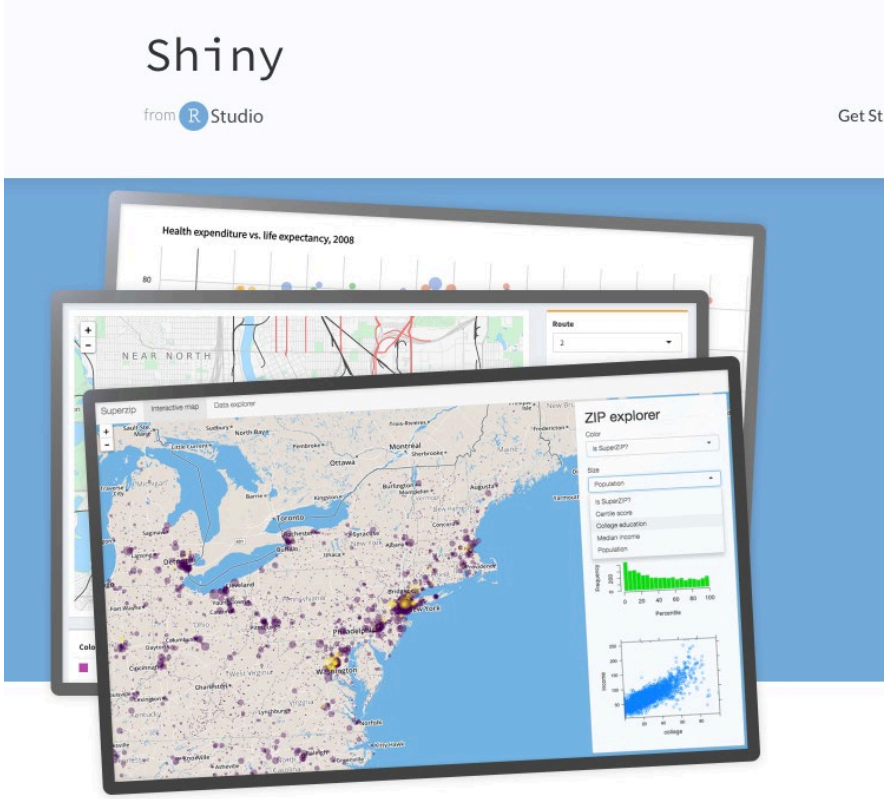
Miscellaneous

1. Output to a file - PDF, PNG, TIFF, SVG
2. Shapes and line types - Set the shape of points and patterns used in lines.
3. Fonts - Use different fonts in your graphs
4. Antialiased bitmap output - If your plots look jagged or pixelated. (Not yet finished)

More Tools for Advanced R Visualization

- **ggmap**
- R package that makes it easy to retrieve raster map tiles from popular online mapping services like Google Maps and Stamen Maps and plot them using the ggplot2 framework
- https://builtin.com/data-science/ggmap
- https://www.rdocumentation.org/packages/ggmap/versions/4.0.0

- **R Shiny** - Interactive visualization + dashboard
- https://shiny.posit.co/



Data Visualization with Tableau

Tableau mapping issue

- If you cannot see any points on the map, look for the message that will appear on the bottom of the screen.
- The message will say something like “96 missing matches”. Click the message.
- In the pop-up window, click “Edit Location.”
- You will see that the current location is set to a country other than United States. Change it to U.S.

Telling Stories Without Distortion

Which is the largest U.S. state by land area?

In Reality…

worldatlas
explore your world…

United States

An **interactive** list of U.S. States, including their *(2012)* populations, as well as land sizes and densities.

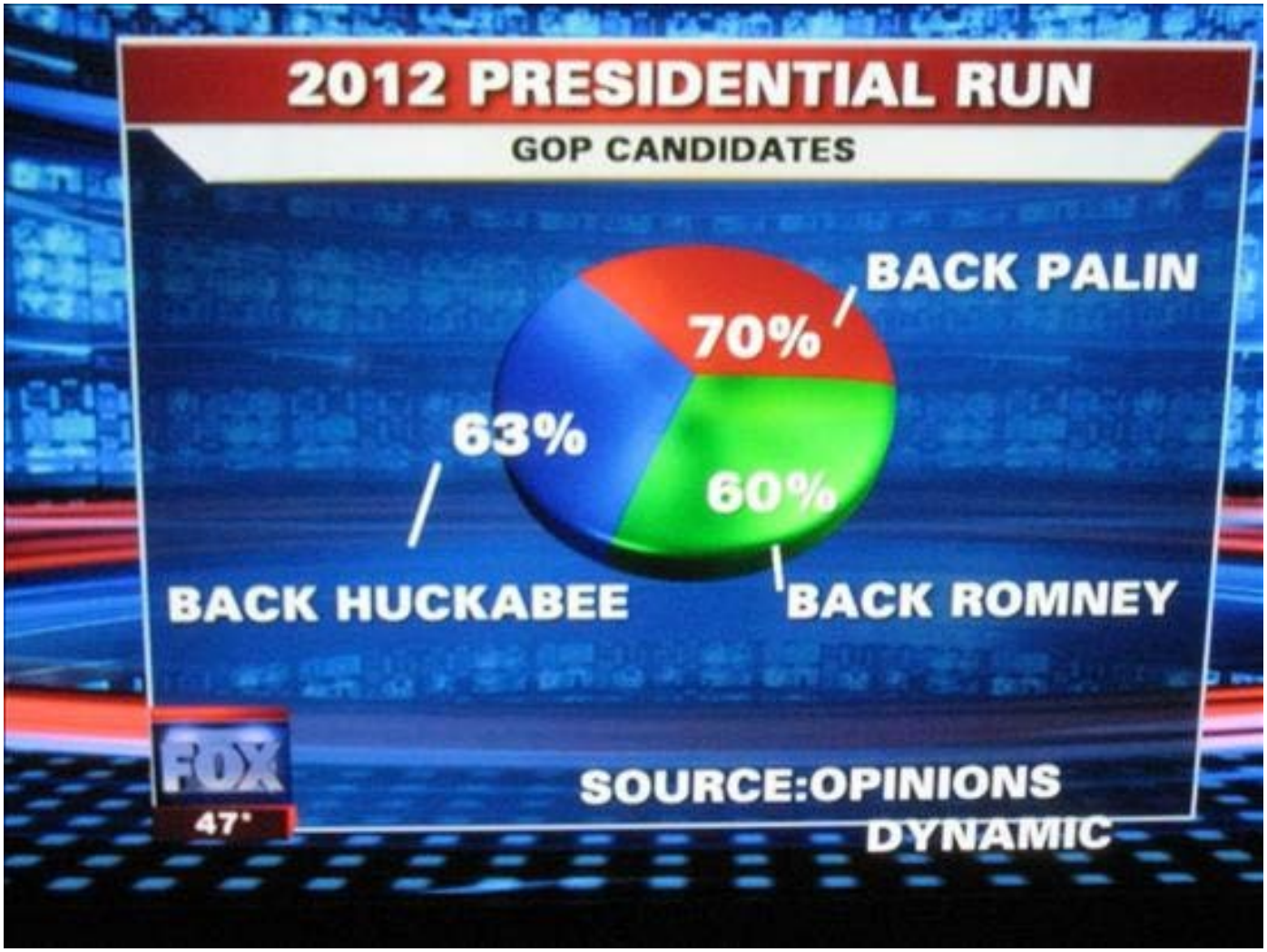
Click on any of the bolded headings to sort this state list.
Type the name of a state you're looking for in the search bar to find its individual stats.

Show 50 entries Search:

	STATE	Abbr.	Population	Land Area (sq. km.)	Pop. Density (per sq. km.)	Land Area (sq. mi.)	Pop. Density (per sq. mi.)
1	Alaska	AK	731,449	1,481,346.00	0.46	571,951	1.20
2	Texas	TX	26,059,203	678,051.12	35.88	261,797	92.92
3	California	CA	38,041,430	403,931.96	91.00	155,959	235.68
4	Montana	MT	1,005,141	376,977.95	2.57	145,552	6.65
5	New Mexico	NM	2,085,538	314,310.60	6.31	121,356	16.35

Don’t mess with Alaska

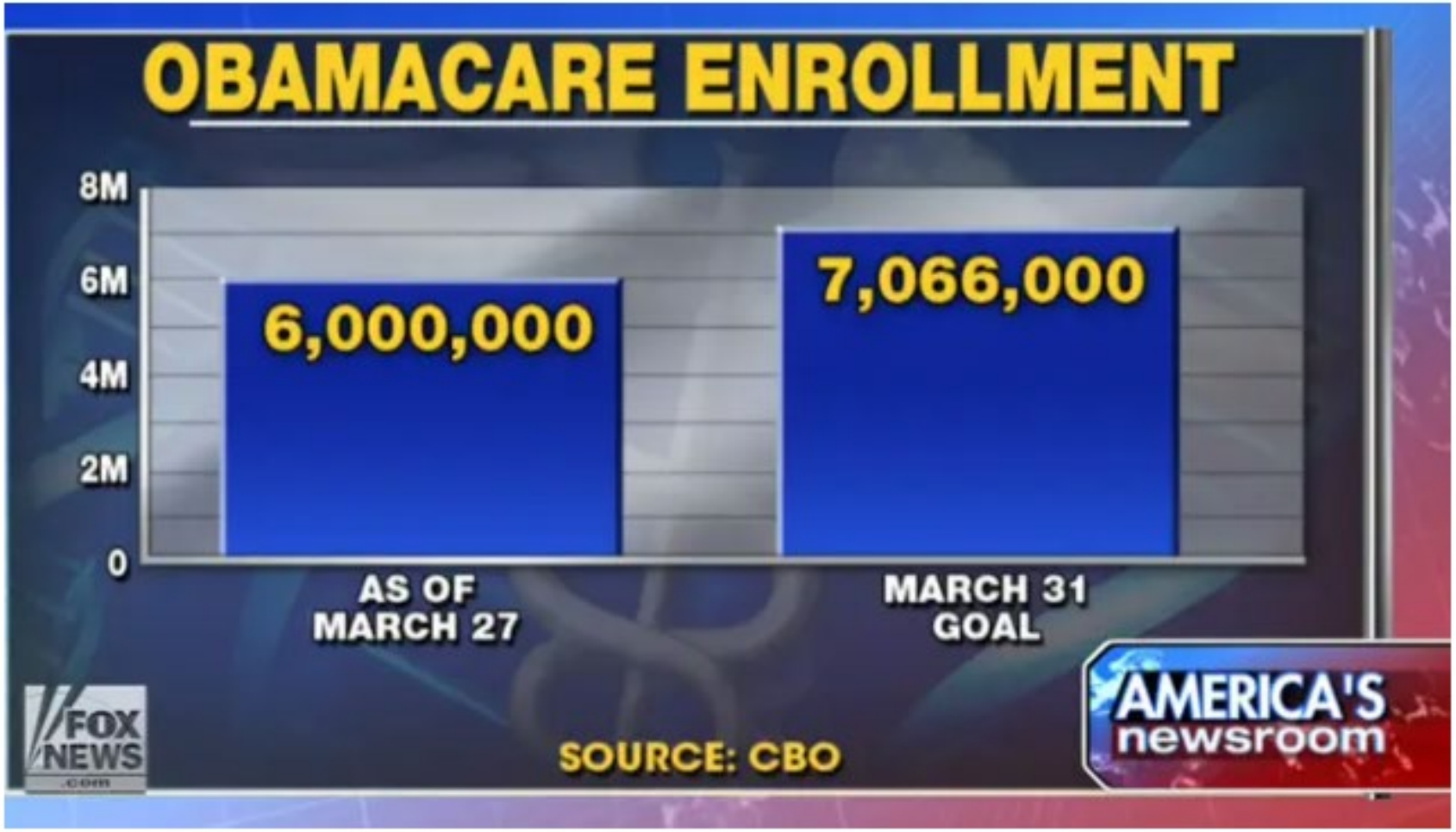
The New Principle of Pie Chart



It happened again!

What's wrong with this graph?

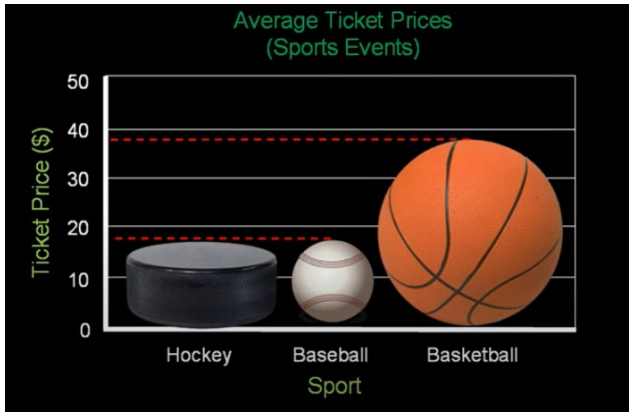
Correction Made



https://www.politico.com/blogs/media/2014/04/fox-news-corrects-obamacare-chart-186120

What are misleading graphs?

- Any graph that distracts the reader from fully understanding what the reader is viewing can be considered misleading.
  - Graphs with no title
  - Graphs with no labels on the axes
  - Graphs with a vertical axis that doesn't start at 0
  - Graphs with differently sized graphing icons or bars
  - Graphs without equal intervals on any axis
  - Graphs with a broken or squished axis
  - And more! (Cherry picking, inconsistent units, etc.)



Source: https://www.youtube.com/watch?v=ETbc8GIhfHo

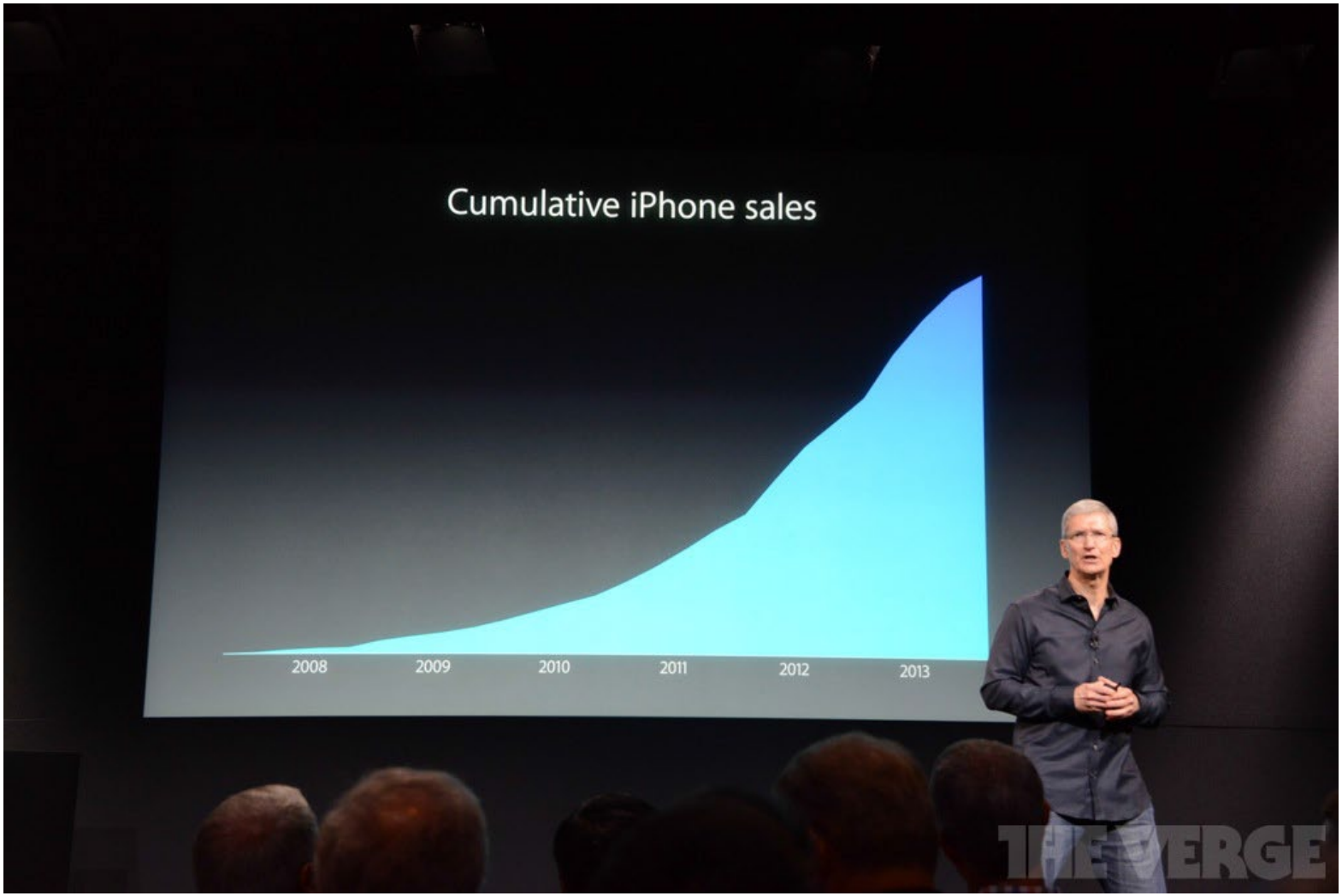
Misleading graphs?

Broken (squished) y-axis

Corrected

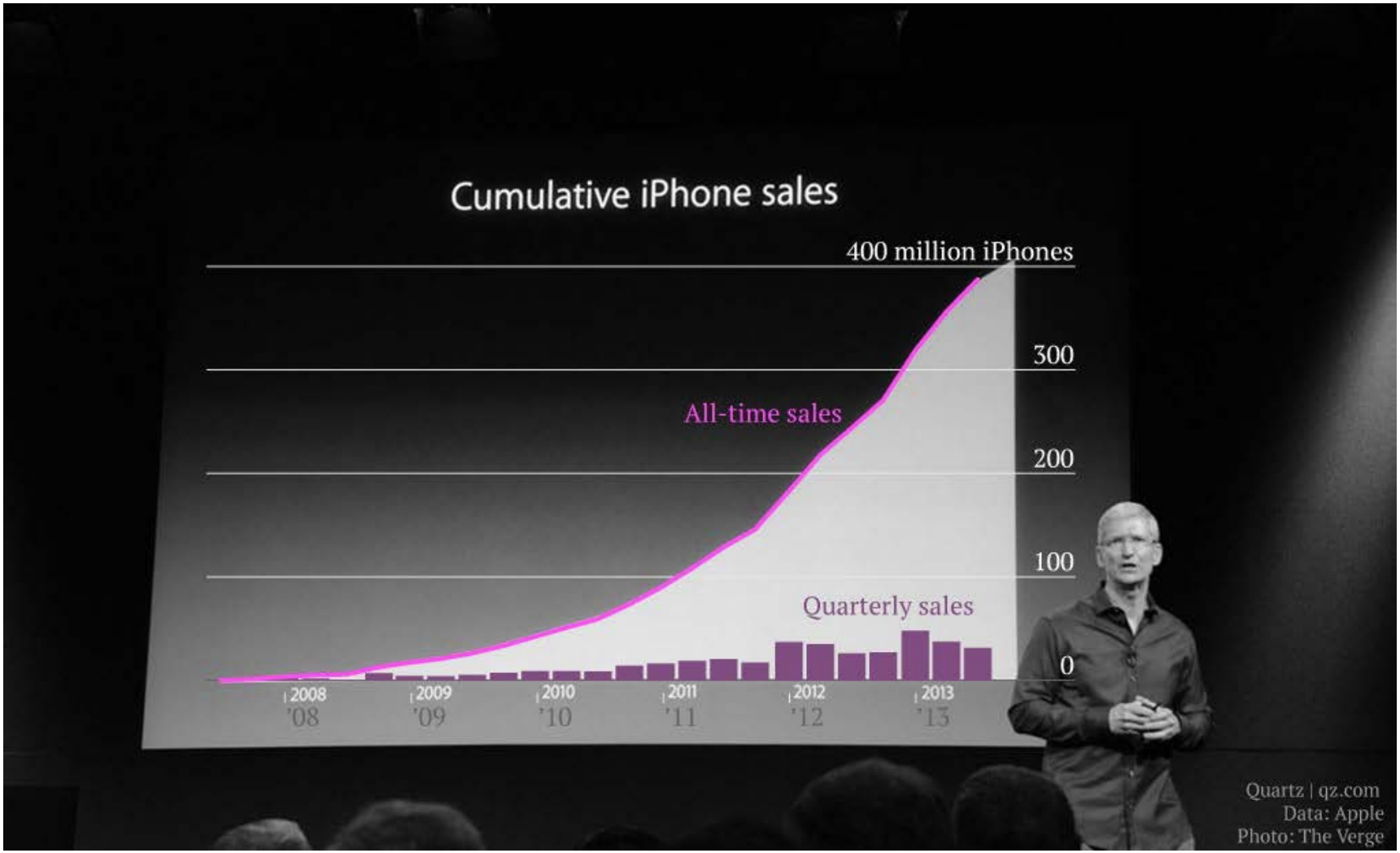
Recommend to watch: https://youtu.be/E91bGT9BjYk?si=2ClvaSgATdJWCRub

Tim Cook is Cooking His Graph!



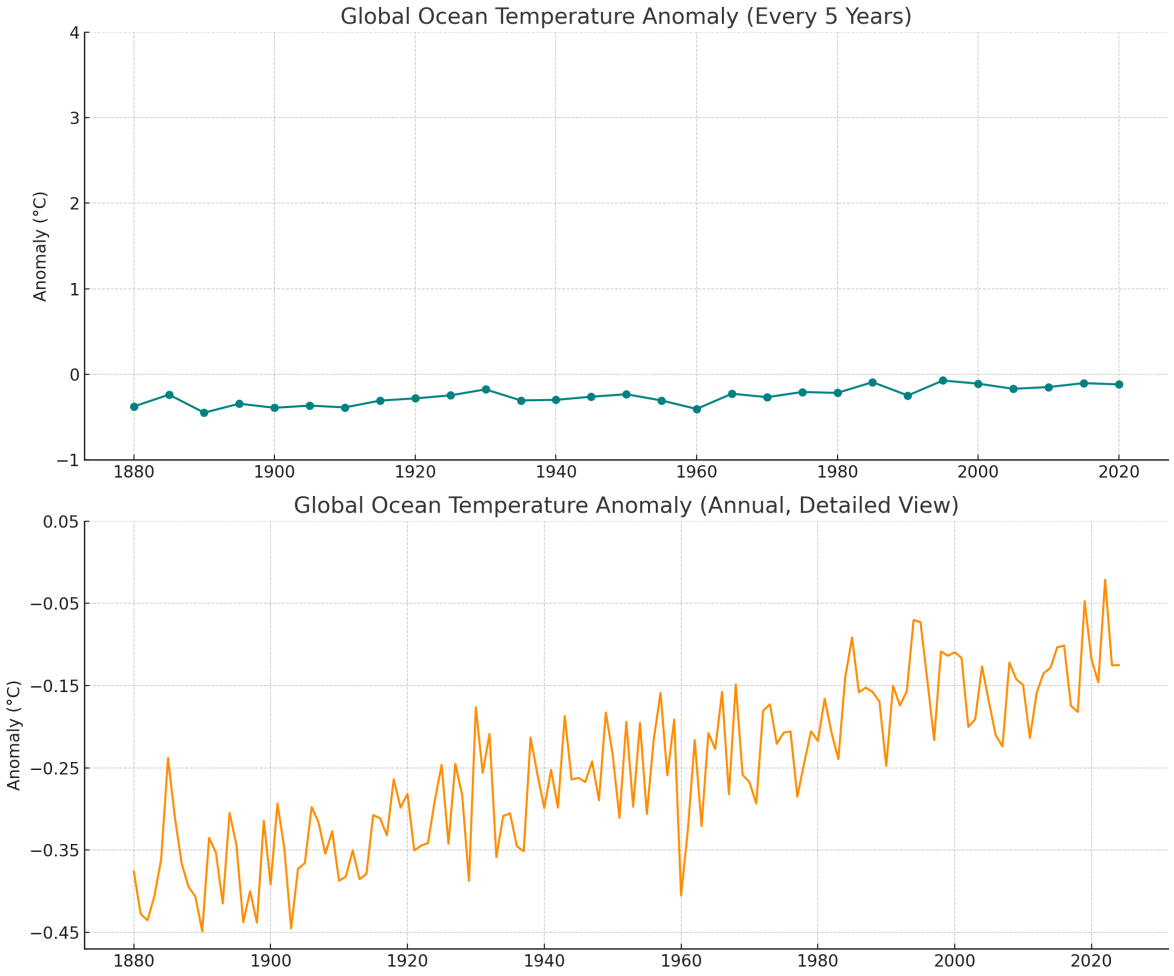
http://qz.com/122921/the-chart-tim-cook-doesnt-want-you-to-see/

This looks less cool, but less misleading.

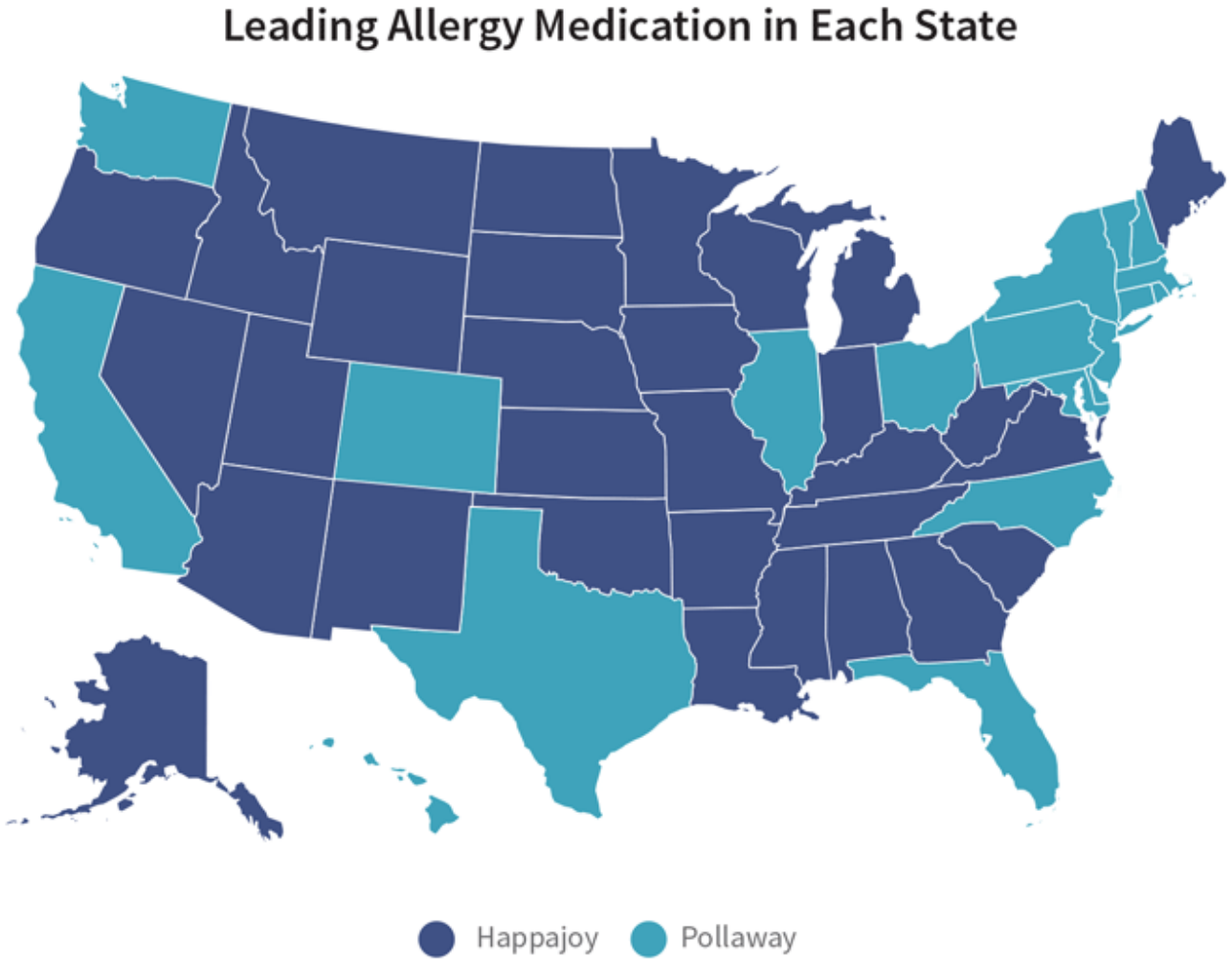


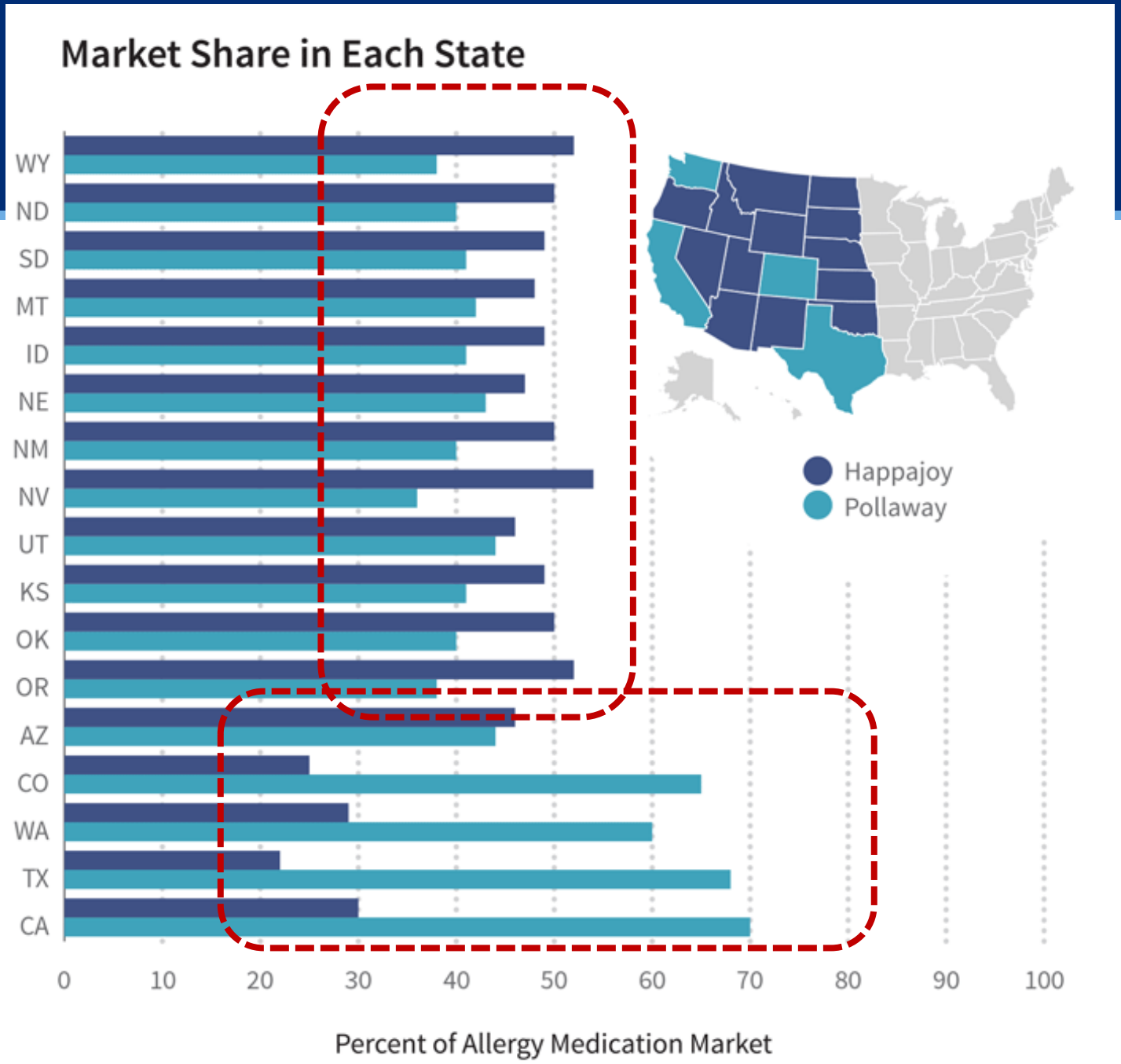
http://qz.com/122921/the-chart-tim-cook-doesnt-want-you-to-see/

Selecting the right scale and granularity



Which company is the market leader?




Market Share in Each State
WY
ND
SD
MT
ID
NE
NM
NV
UT
KS
OK
OR
AZ
CO
WA
TX
CA
0
10
20
30
40
50
60
70
80
90
100
Happajoy
Pollaway
Percent of Allergy Medication Market