DATA SCIENCE

¿Cómo te engañan con datos?

En DS pasamos de tener **datos** a obtener **información**.

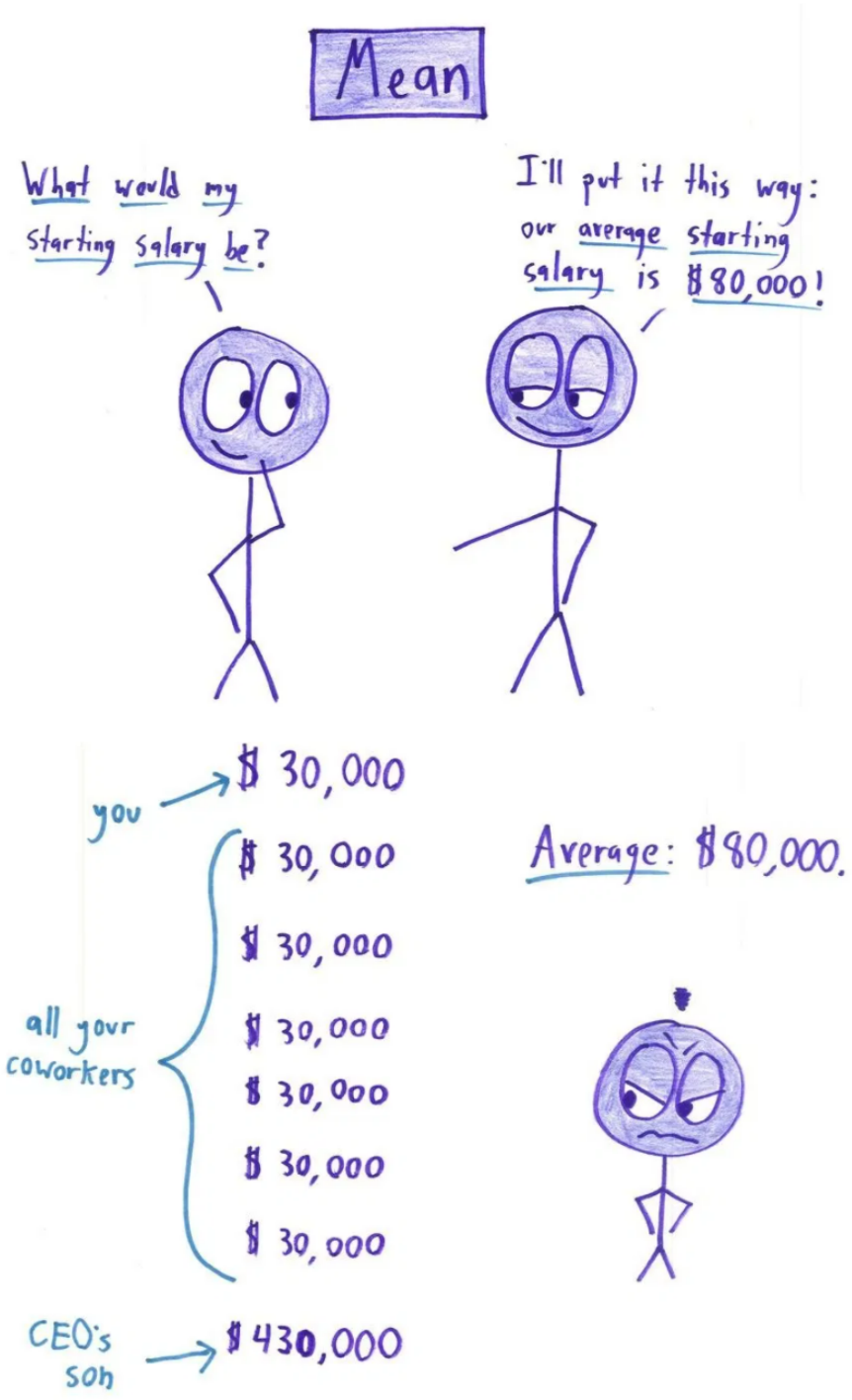
En este taller vamos a ver:

- Buenos y malos estadísticos para resumir los datos
- Buenas y malas prácticas de visualización
- Paradojas matemáticas de datos
- Equidad aplicada a modelos

A veces juegan con nuestros sesgos en la percepción para presentarnos los datos de una forma que nos va a inducir a interpretarlos erróneamente o de manera poco parcial.

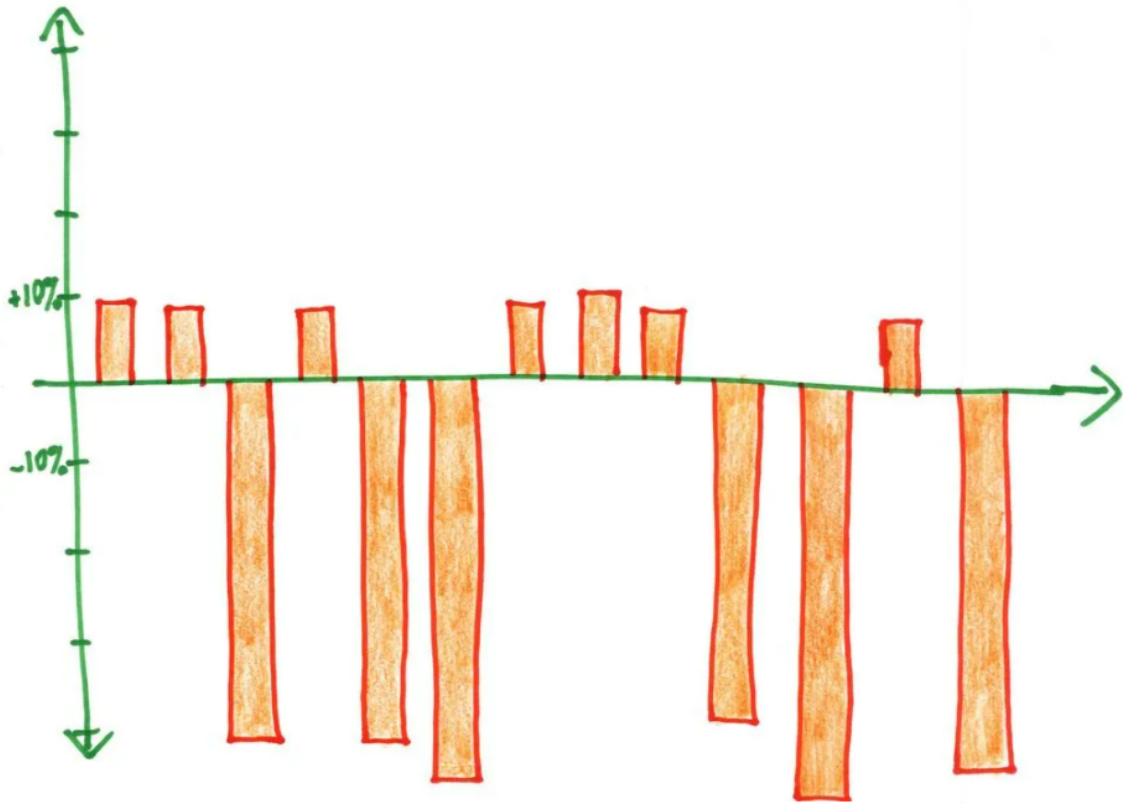
Estadísticos descriptivos

Veamos cómo nos pueden engañar con la estadística descriptiva, ¡empezando por la famosa media!

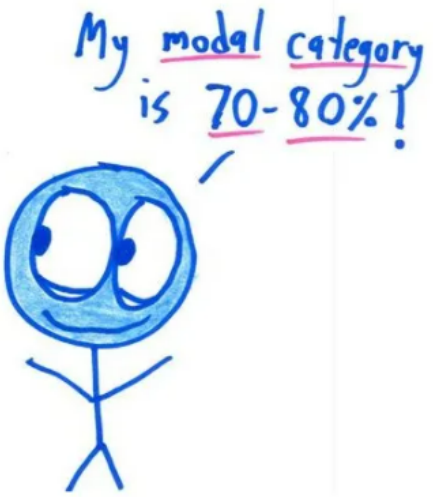


Median


So, why should I invest with you?
Well, not to brag, but my fund has a median gain of 8% per year!


+10%
-10%

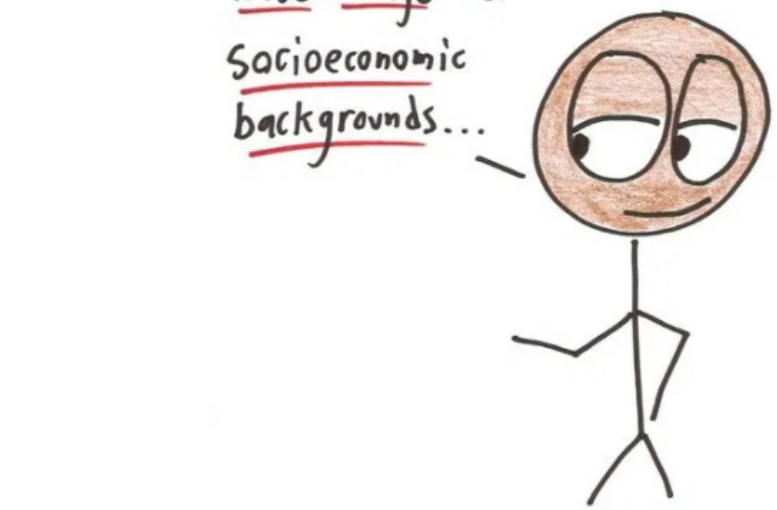
Mode

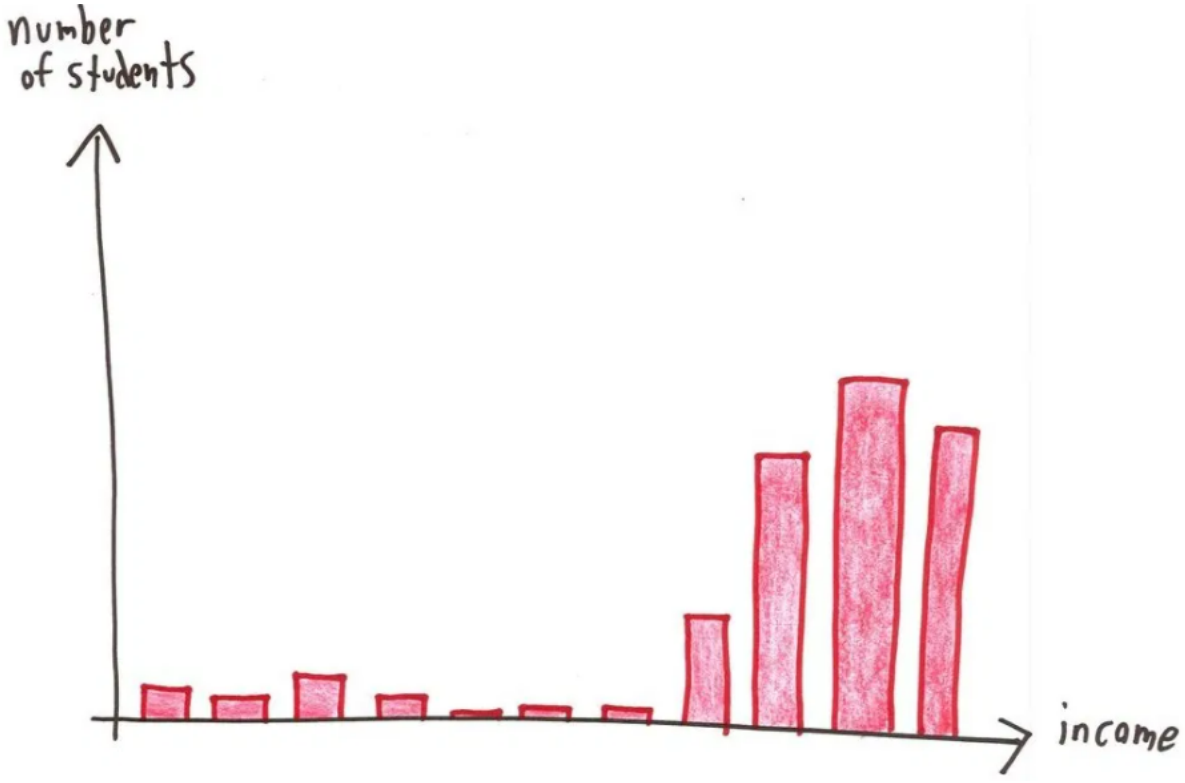


Score Category	Number of Tests
90s	0
80s	0
70s	2
60s	1
50s	1
40s	1
30s	1
20s	1

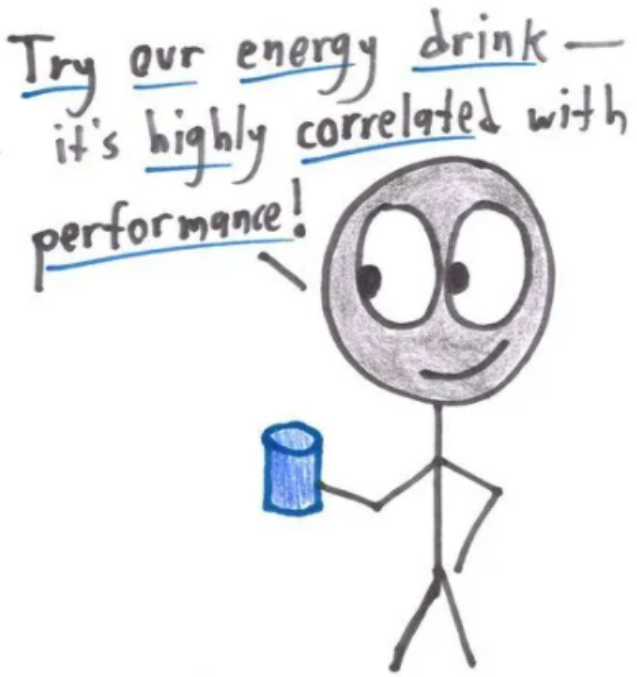
Range

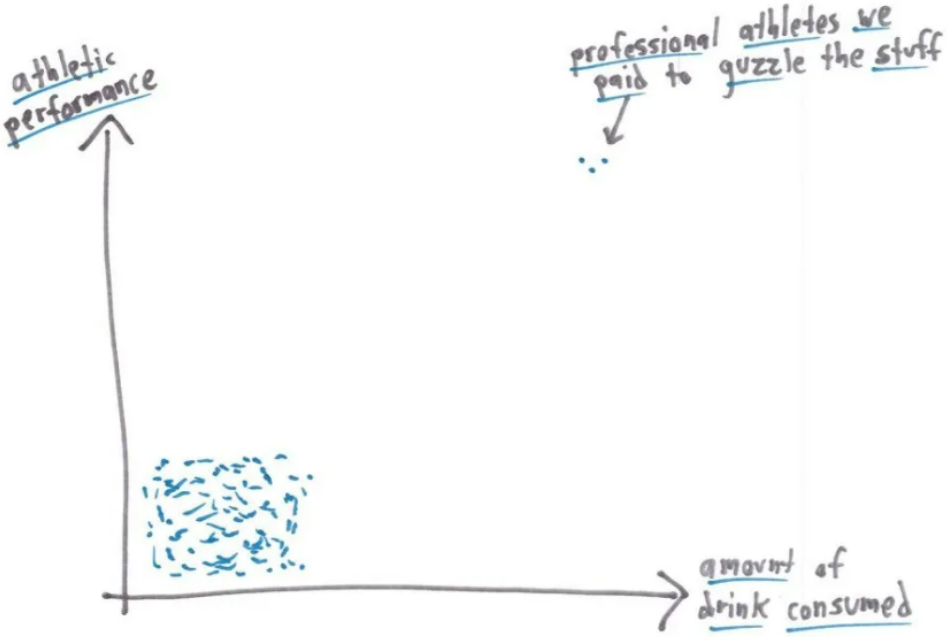
Our students come from a wide range of socioeconomic backgrounds...

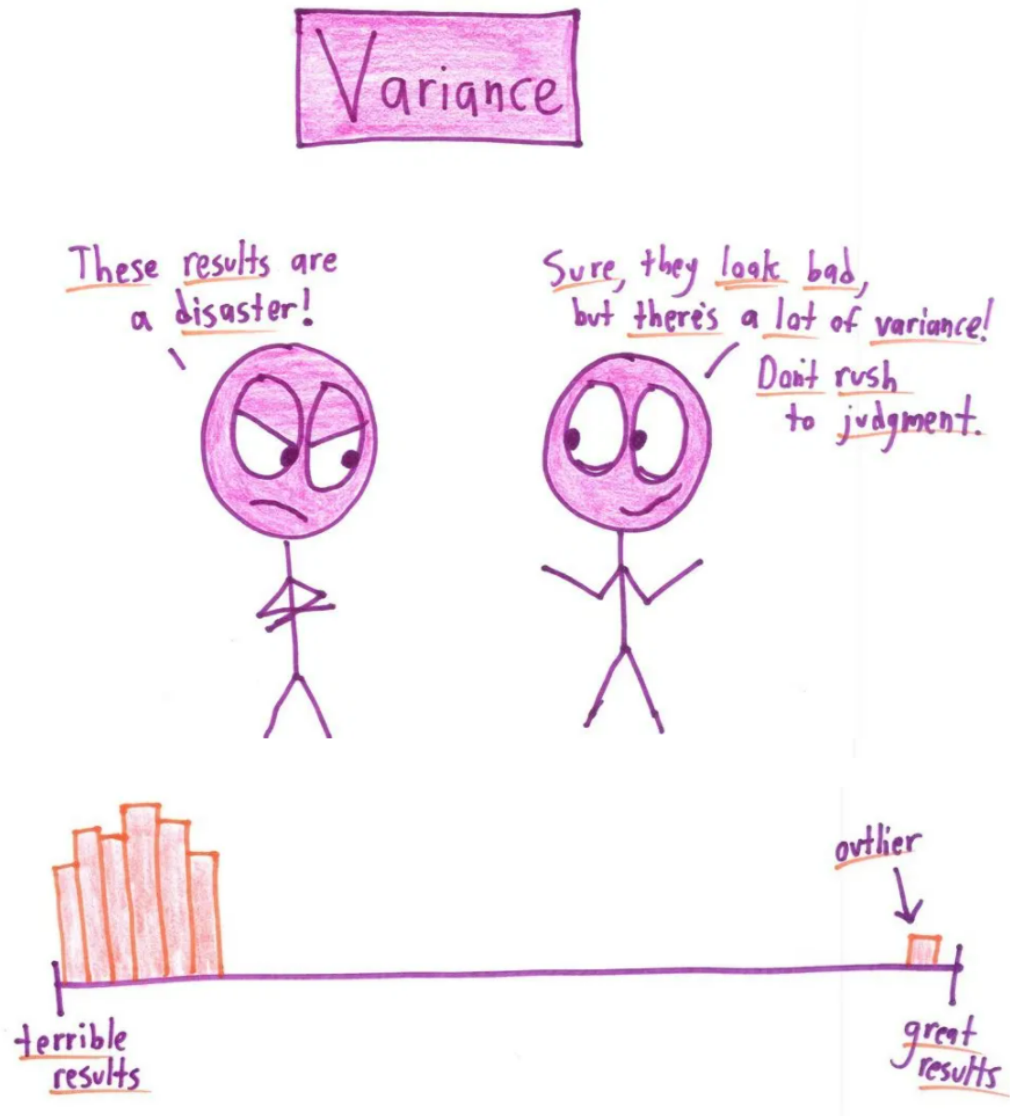




Correlation Coefficient

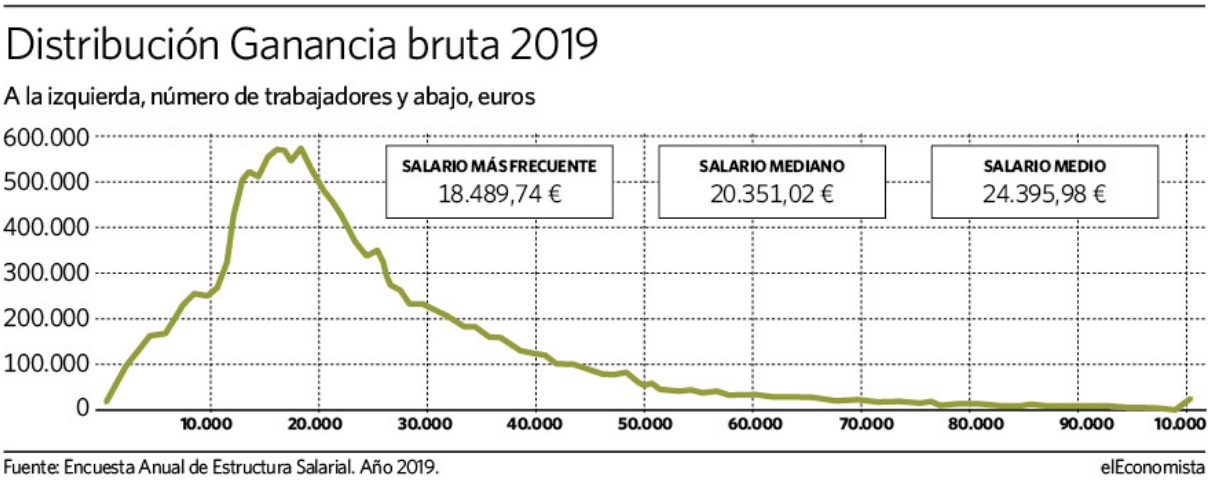






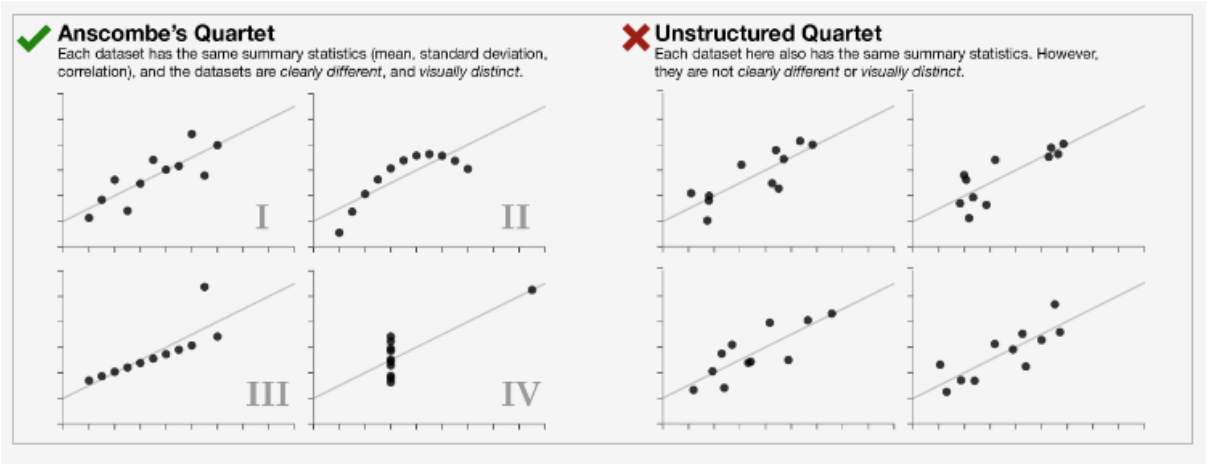
¿Y por qué **SIEMPRE** nos dan el dato **MEDIO**?

La media no te dice **NADA** de la dispersión de los datos.



¿Y noticias con el sueldo mediano, para cuándo?

Entonces, si conozco la media y la varianza, ya sé mucho sobre mis datos, ¿no?

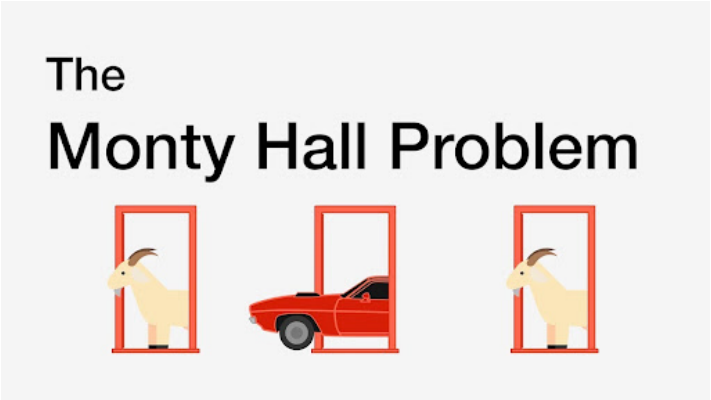


También juegan con nuestros **sesgos mentales**:

¿Cuánta gente distinta crees que tendrías que juntar aleatoriamente para que dos de esas personas cumplan años el mismo día?

Solo con 23 personas ya tienes un 50% de probabilidad de tener 2 personas con el mismo cumpleaños. Solemos pensar que necesitamos muchas más personas, porque SOLO pensamos en nosotros mismos y no en la combinación de cada persona con los demás.

¿Y si vas a un concurso de televisión?



¿Te quedas con tu puerta o la cambias?

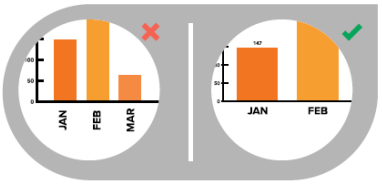
Te están engañando si no conoces la probabilidad CONDICIONADA.
La probabilidad de acertar a la primera es ⅓ y NUNCA CAMBIA.
Tras abrir una puerta, tu intento inicial SIGUE SIENDO ⅓ , luego la puerta cerrada tendrá una probabilidad de ⅔

¿Quieres verlo más claro? Piensa en 1000 puertas iniciales con un coche. Tras elegir una puerta, te abren todas las restantes menos una. Tienes una probabilidad de acertar a la primera de 1/1000 = 0.001y una probabilidad de que esté en la otra puerta cerrada de 0.999. Ahora lo vemos más claro.

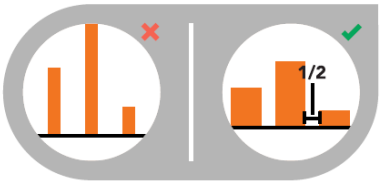
Buenas y malas prácticas en visualización

BAR CHART

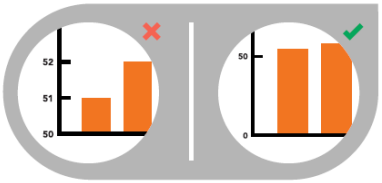
DESIGN BEST PRACTICES



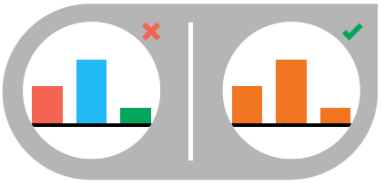
USE HORIZONTAL LABELS
Avoid steep diagonal or vertical type, as it can be difficult to read.



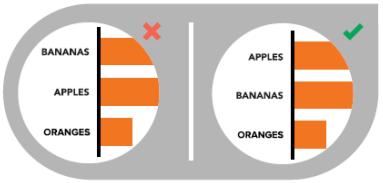
SPACE BARS APPROPRIATELY
Space between bars should be ½ bar width.



START THE Y-AXIS VALUE AT 0
Starting at a value above zero truncates the bars and doesn't accurately reflect the full value.



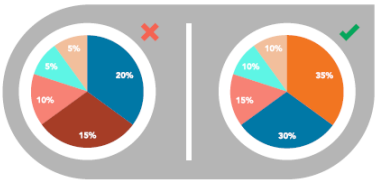
USE CONSISTENT COLORS
Use one color for bar charts. You may use an accent color to highlight a significant data point.



ORDER DATA APPROPRIATELY
Order categories alphabetically, sequentially, or by value.

VISUALIZE NO MORE THAN 5 CATEGORIES PER CHART
It is difficult to differentiate between small values; depicting too many slices decreases the impact of the visualization. If needed, you can group smaller values into an "other" or "miscellaneous" category, but make sure it does not hide interesting or significant information.

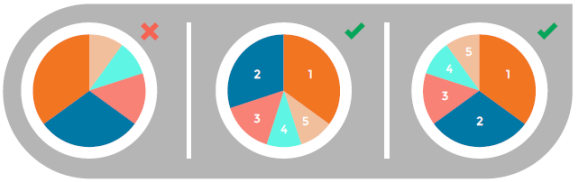
DON'T USE MULTIPLE PIE CHARTS FOR COMPARISON
Slice sizes are very difficult to compare side-by-side. Use a stacked bar chart instead.



MAKE SURE ALL DATA ADDS UP TO 100%
Verify that values total 100% and that pie slices are sized proportionate to their corresponding value.

PIE CHART

DESIGN BEST PRACTICES



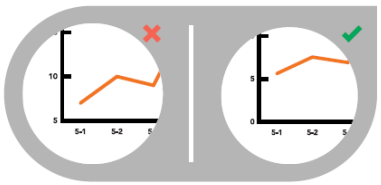
ORDER SLICES CORRECTLY
There are two ways to order sections, both of which are meant to aid comprehension:

OPTION 1
Place the largest section at 12 o'clock, going clockwise. Place the second largest section at 12 o'clock, going counterclockwise. The remaining sections can be placed below, continuing counterclockwise.

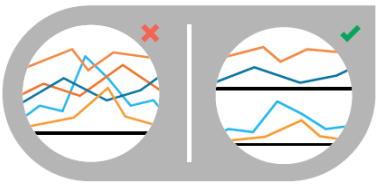
OPTION 2
Start the largest section at 12 o'clock, going clockwise. Place remaining sections in descending order, going clockwise.

LINE CHART

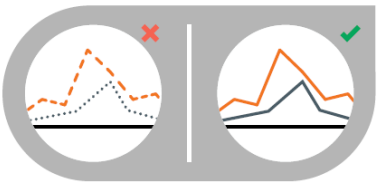
DESIGN BEST PRACTICES



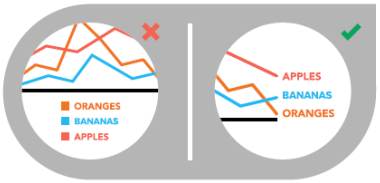
INCLUDE A ZERO BASELINE IF POSSIBLE
Although a line chart does not have to start at a zero baseline, it should be included if possible. If relatively small fluctuations in data are meaningful (e.g., in stock market data), you may truncate the scale to showcase these variances.



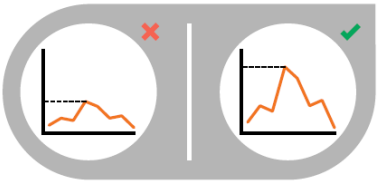
DON'T PLOT MORE THAN 4 LINES
If you need to display more, break them out into separate charts for better comparison.



USE SOLID LINES ONLY
Dashed and dotted lines can be distracting.



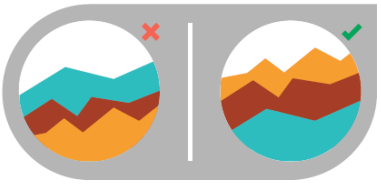
LABEL THE LINES DIRECTLY
This lets readers quickly identify lines and corresponding labels instead of referencing a legend.



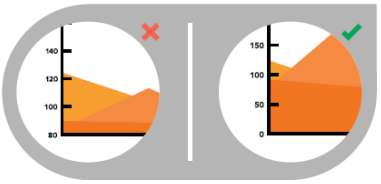
USE THE RIGHT HEIGHT
Plot all data points so that the line chart takes up approximately two-thirds of the y-axis' total scale.

AREA CHART

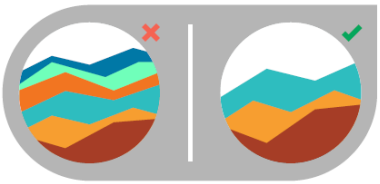
DESIGN BEST PRACTICES



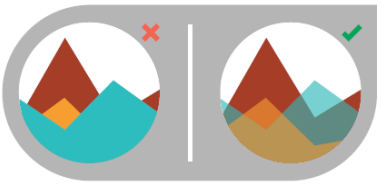
MAKE IT EASY TO READ
In stacked area charts, arrange data to position categories with highly variable data on the top of the chart and low variability on the bottom.



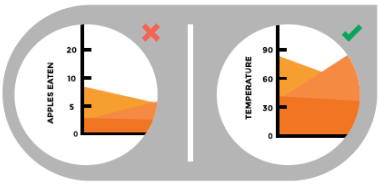
START Y-AXIS VALUE AT 0
Starting the axis above zero truncates the visualization of values.



DON'T DISPLAY MORE THAN 4 DATA CATEGORIES
Too many will result in a cluttered visual that is difficult to decipher.



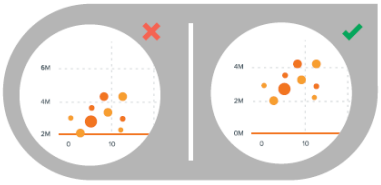
USE TRANSPARENT COLORS
In standard area charts, ensure data isn't obscured in the background by ordering thoughtfully and using transparency.



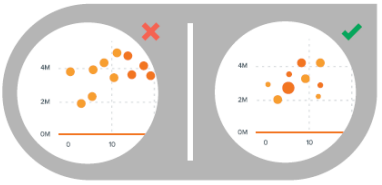
DON'T USE AREA CHARTS TO DISPLAY DISCRETE DATA
The connected lines imply intermediate values, which only exist with continuous data.

SCATTER PLOT

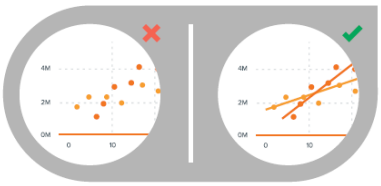
DESIGN BEST PRACTICES



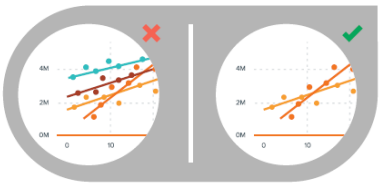
START Y-AXIS VALUE AT 0
Starting the axis above zero truncates the visualization of values.



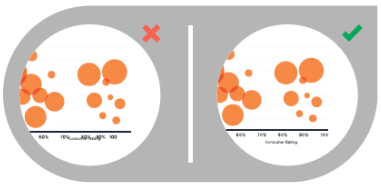
INCLUDE MORE VARIABLES
Use size and dot color to encode additional data variables.



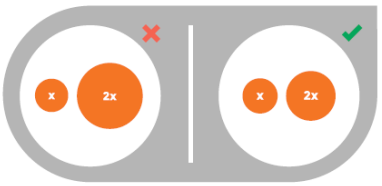
USE TREND LINES
These help draw correlation between the variables to show trends.



DON'T COMPARE MORE THAN 2 TREND LINES
Too many lines make data difficult to interpret.



MAKE SURE LABELS ARE VISIBLE
All labels should be unobstructed and easily identified with the corresponding bubble.



SIZE BUBBLES APPROPRIATELY
Bubbles should be scaled according to area, not diameter.

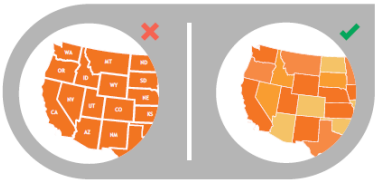
DON'T USE ODD SHAPES
Avoid adding too much detail or using shapes that are not entirely circular; this can lead to inaccuracies.

BUBBLE CHART

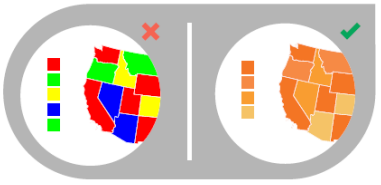
DESIGN BEST PRACTICES

HEAT MAP

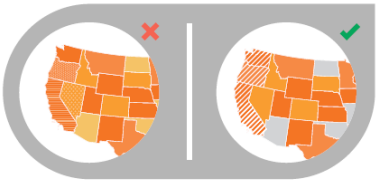
DESIGN BEST PRACTICES



USE A SIMPLE MAP OUTLINE
These lines are meant to frame the data, not distract.



SELECT COLORS APPROPRIATELY
Some colors stand out more than others, giving unnecessary weight to that data. Instead, use a single color with varying shade or a spectrum between two analogous colors to show intensity. Also remember to intuitively code color intensity according to values.



USE PATTERNS SPARINGLY
A pattern overlay that indicates a second variable is acceptable, but using multiple is overwhelming and distracting.

CHOOSE APPROPRIATE DATA RANGES
Select 3-5 numerical ranges that enable fairly even distribution of data between them. Use +/- signs to extend high and low ranges.

10 DATA DESIGN DOS AND DON'TS

Designing your data doesn't have to be overwhelming. With a basic understanding of how different data sets should be visualized, along with a few fundamental design tips and best practices, you can create more accurate, more effective data visualizations. Follow these 10 tips to ensure your design does your data justice.

1 | DO USE ONE COLOR TO REPRESENT EACH CATEGORY.

2 | DO ORDER DATA SETS USING LOGICAL HEIRARCHY.

3 | DO USE CALLOUTS TO HIGHLIGHT IMPORTANT OR INTERESTING INFORMATION.

4 | DO VISUALIZE DATA IN A WAY THAT IS EASY FOR READERS TO COMPARE VALUES.

5 | DO USE ICONS TO ENHANCE COMPREHENSION AND REDUCE UNNECESSARY LABELING.

6 | DON'T USE HIGH CONTRAST COLOR COMBINATIONS SUCH AS RED/GREEN OR BLUE/YELLOW.

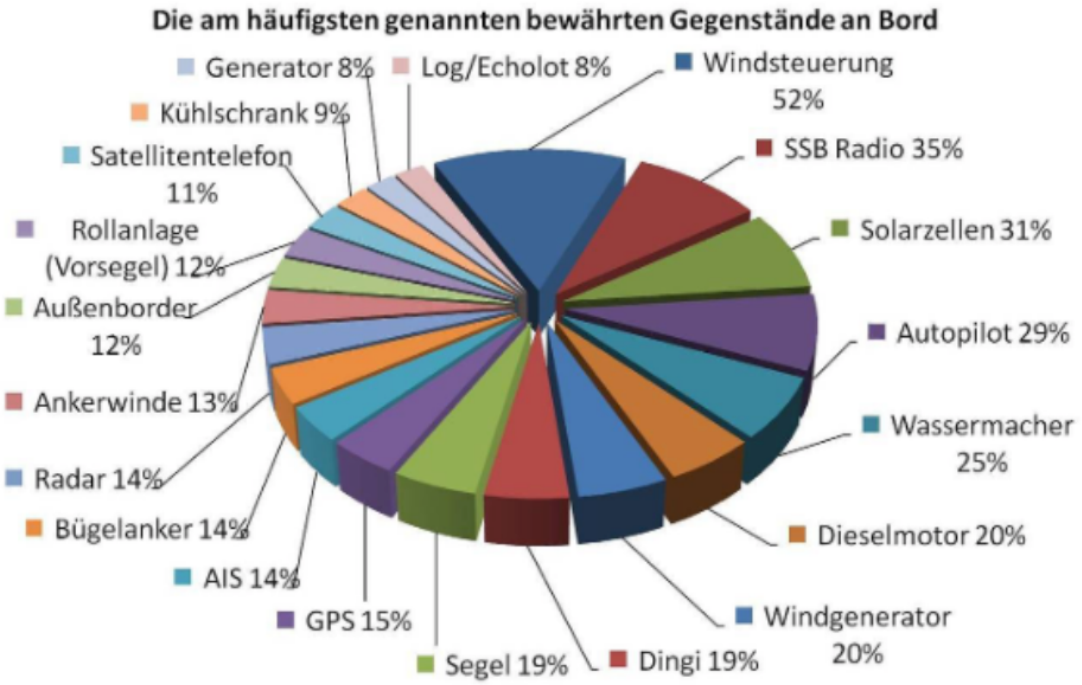
7 | DON'T USE 3D CHARTS. THEY CAN SKEW PERCEPTION OF THE VISUALIZATION.

8 | DON'T ADD CHART JUNK. UNNECESSARY ILLUSTRATIONS, DROP SHADOWS, OR ORNAMENTATIONS DISTRACT FROM THE DATA.

9 | DON'T USE MORE THAN 6 COLORS IN A SINGLE LAYOUT.

10 | DON'T USE DISTRACTING FONTS OR ELEMENTS (SUCH AS BOLD, ITALIC, OR UNDERLINED TEXT).

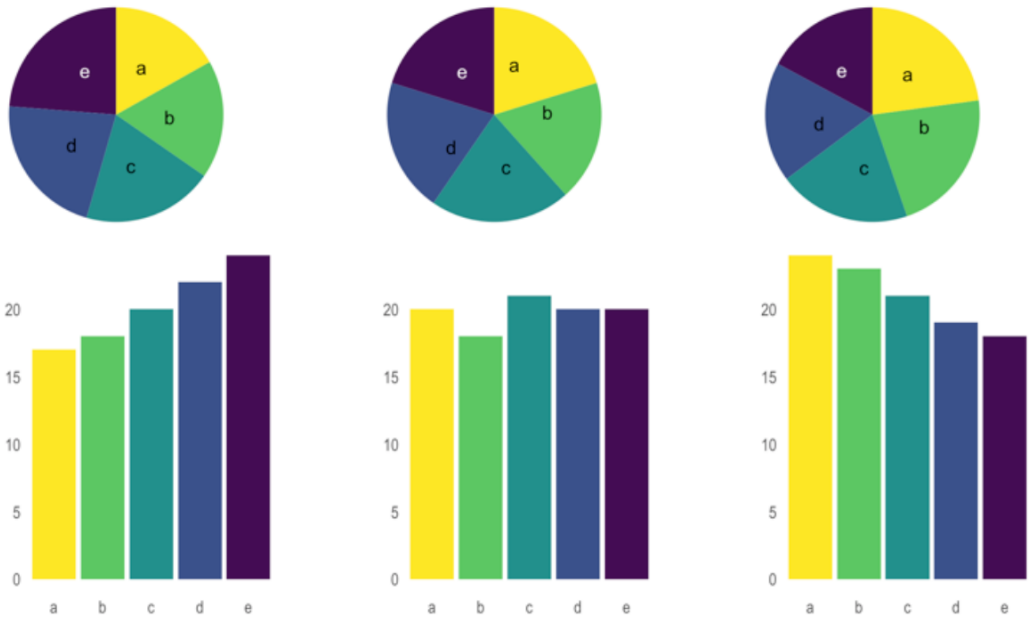
Quiz time! :)

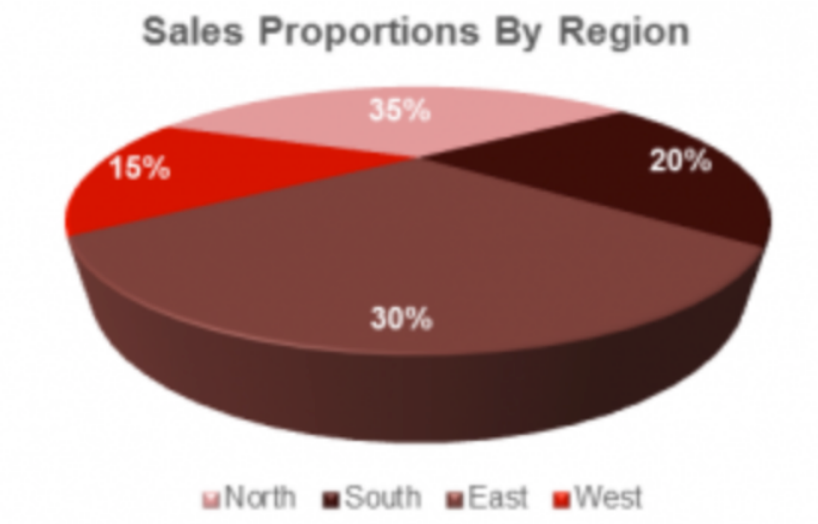


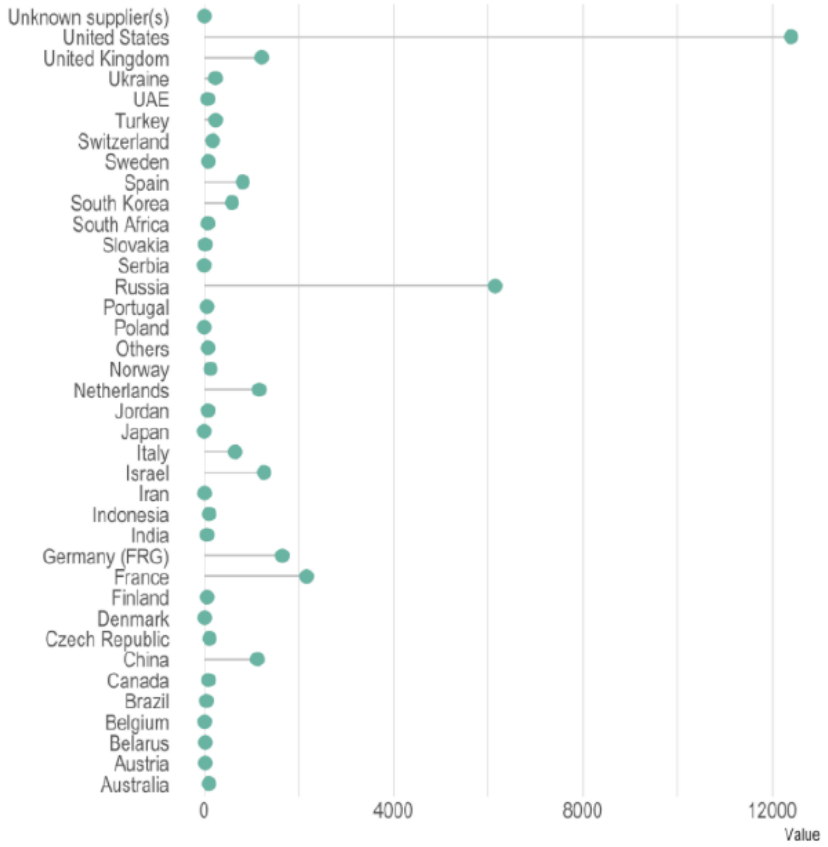
Cálculo + 3D + pie chart

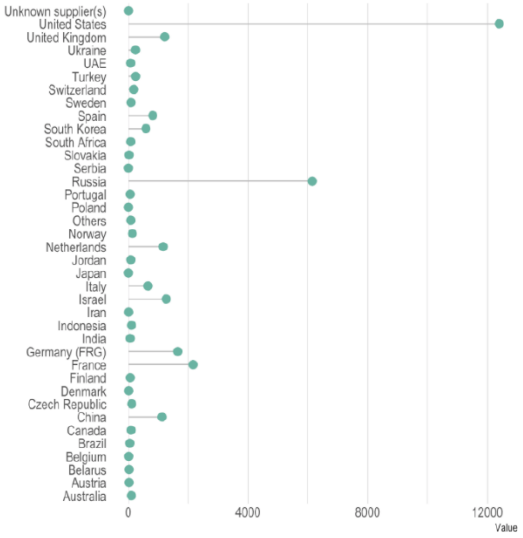
What's wrong with pie chart?

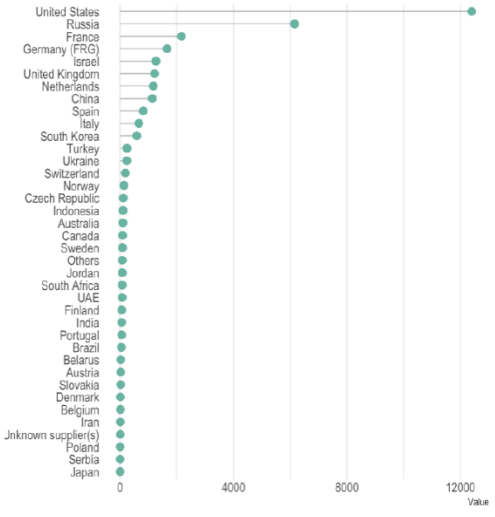
It is hard to distinguish angles

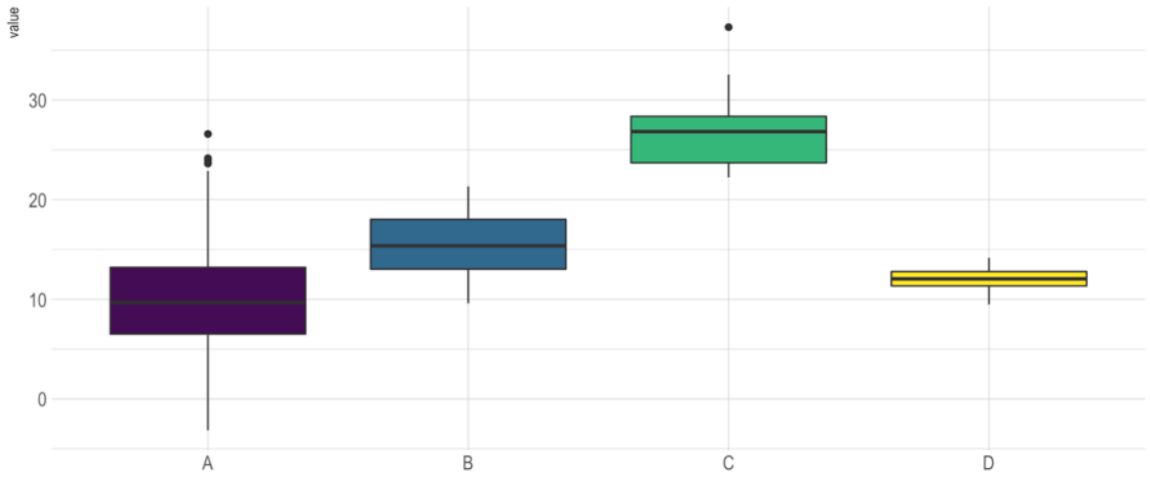


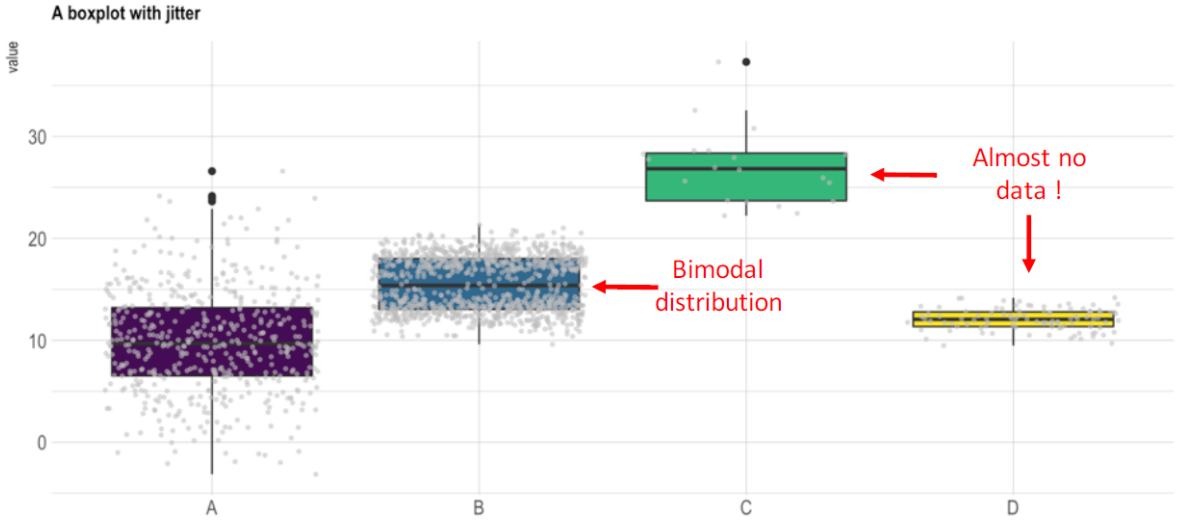


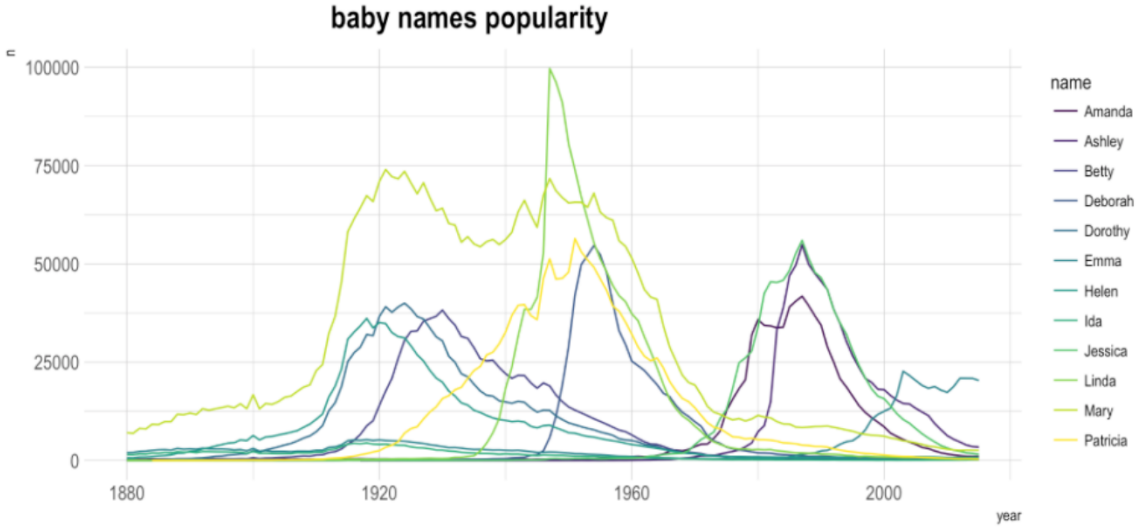

Unknown supplier(s)
United States
United Kingdom
Ukraine
UAE
Turkey
Switzerland
Sweden
Spain
South Korea
South Africa
Slovakia
Serbia
Russia
Portugal
Poland
Others
Norway
Netherlands
Jordan
Japan
Italy
Israel
Iran
Indonesia
India
Germany (FRG)
France
Finland
Denmark
Czech Republic
China
Canada
Brazil
Belgium
Belarus
Austria
Australia
0
4000
8000
12000
Value

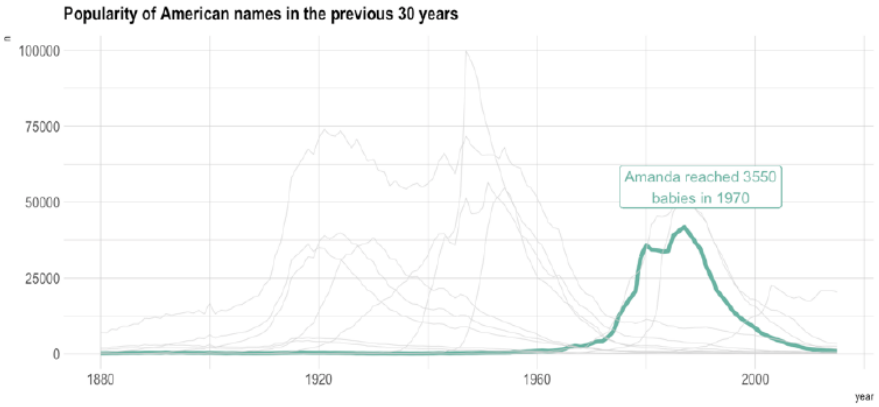

Unknown supplier(s)
United States
United Kingdom
Ukraine
UAE
Turkey
Switzerland
Sweden
Spain
South Korea
South Africa
Slovakia
Serbia
Russia
Portugal
Poland
Others
Norway
Netherlands
Jordan
Japan
Italy
Israel
Iran
Indonesia
India
Germany (FRG)
France
Finland
Denmark
Czech Republic
China
Canada
Brazil
Belgium
Belarus
Austria
Australia
0
4000
8000
12000
Value

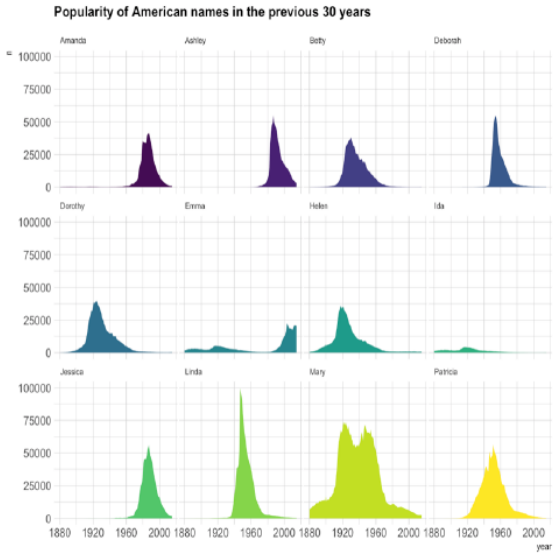

United States
Russia
France
Germany (FRG)
Israel
United Kingdom
Netherlands
China
Spain
Italy
South Korea
Turkey
Ukraine
Switzerland
Norway
Czech Republic
Indonesia
Australia
Canada
Sweden
Others
Jordan
South Africa
UAE
Finland
India
Portugal
Brazil
Belarus
Austria
Slovakia
Denmark
Belgium
Iran
Jnknown supplier(s)
Poland
Serbia
Japan
0
4000
8000
12000
Value

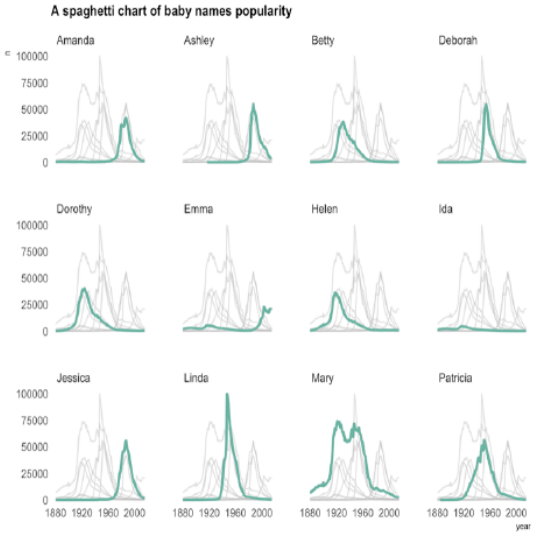

value
30
20
10
0
A
B
C
D

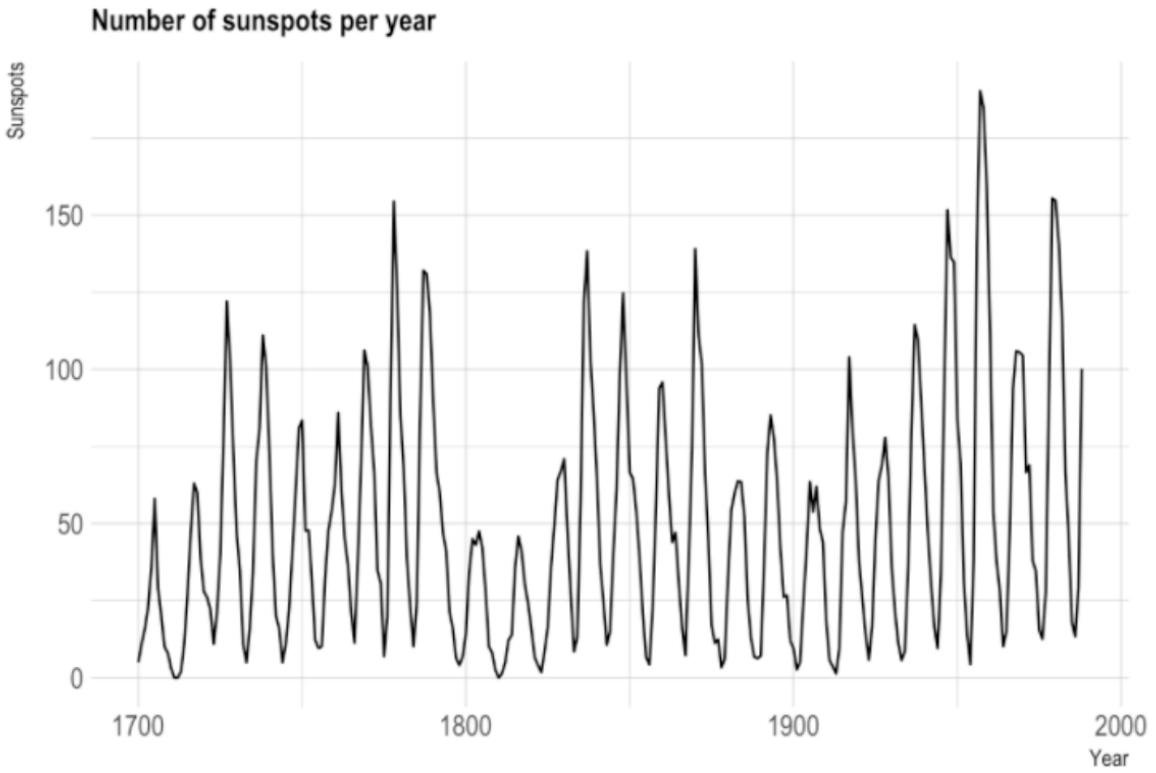

A boxplot with jitter
value
30
20
10
0
A
B
C
D
Almost no data !
Bimodal distribution


baby names popularity
n
100000
75000
50000
25000
0
1880
1920
1960
2000
year
name
Amanda
Ashley
Betty
Deborah
Dorothy
Emma
Helen
Ida
Jessica
Linda
Mary
Patricia


Popularity of American names in the previous 30 years
n
100000
75000
50000
25000
0
Amanda reached 3550
babies in 1970
1880
1920
1960
2000
year


Popularity of American names in the previous 30 years
Amanda
Ashley
Betty
Deborah
Dorothy
Emma
Helen
Ida
Jessica
Linda
Mary
Patricia
n
100000
75000
50000
25000
0
1880
1920
1960
2000
year

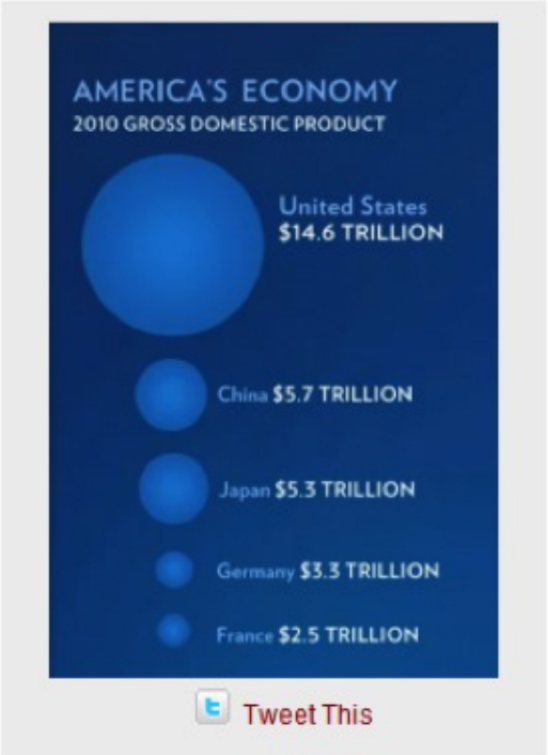

A spaghetti chart of baby names popularity
Amanda
Ashley
Betty
Deborah
Dorothy
Emma
Helen
Ida
Jessica
Linda
Mary
Patricia
n
100000
75000
50000
25000
0
1880
1920
1960
2000
year


Number of sunspots per year
Sunspots
150
100
50
0
1700
1800
1900
2000
Year

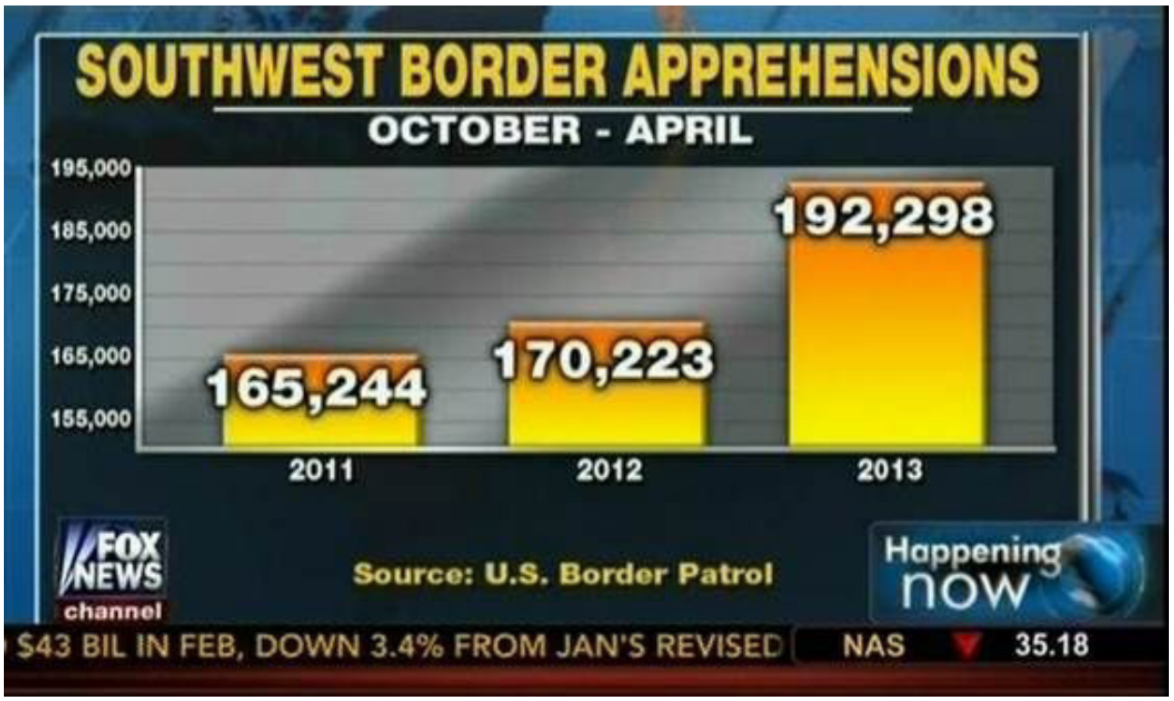

Number of sunspots per year
Sunspots
100
50
0
1800
1825
1850
1875
1900
Year

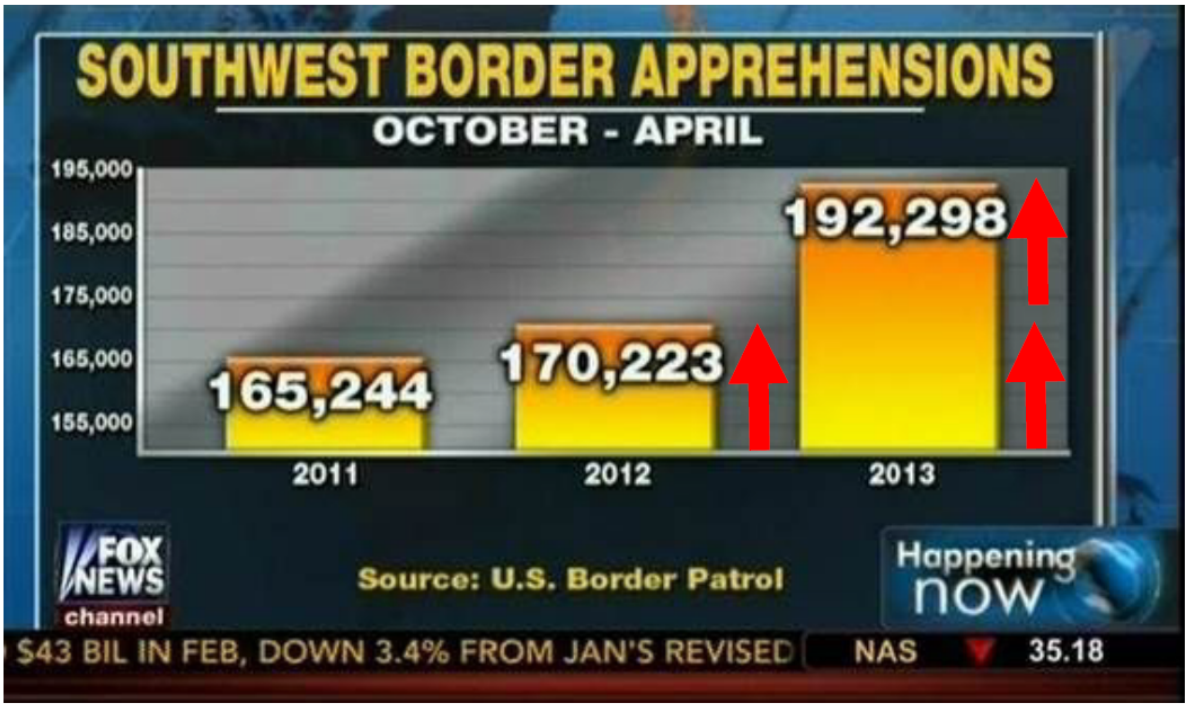

AMERICA'S ECONOMY
2010 GROSS DOMESTIC PRODUCT
United States
$14.6 TRILLION
China $5.7 TRILLION
Japan $5.3 TRILLION
Germany $3.3 TRILLION
France $2.5 TRILLION

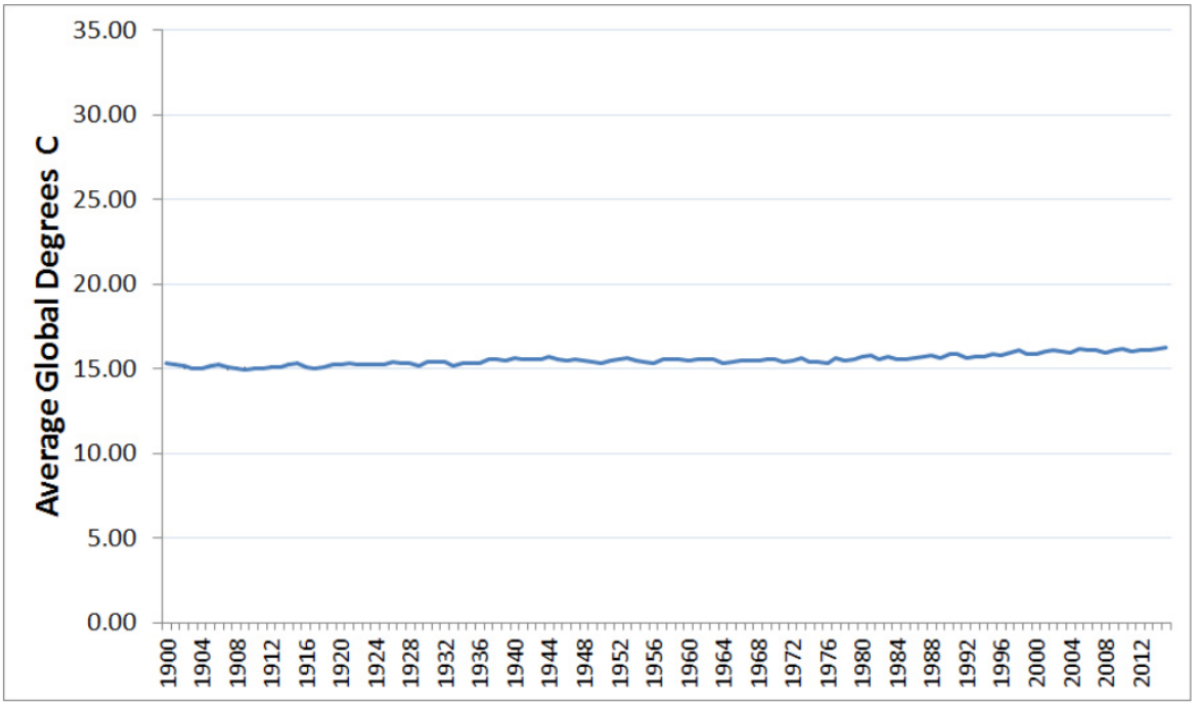
Size = radius

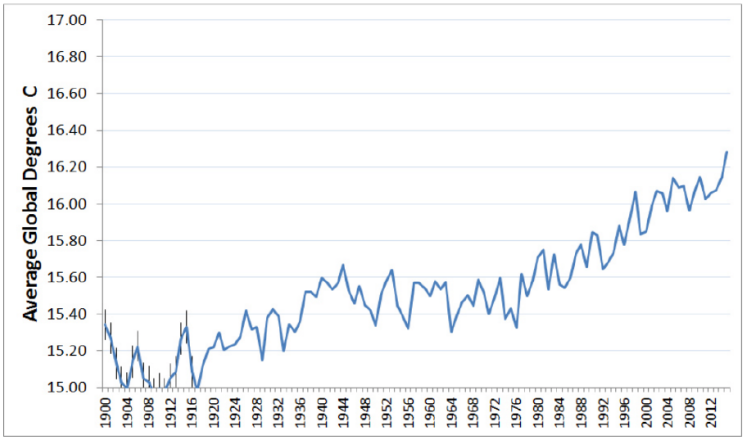


Size = area

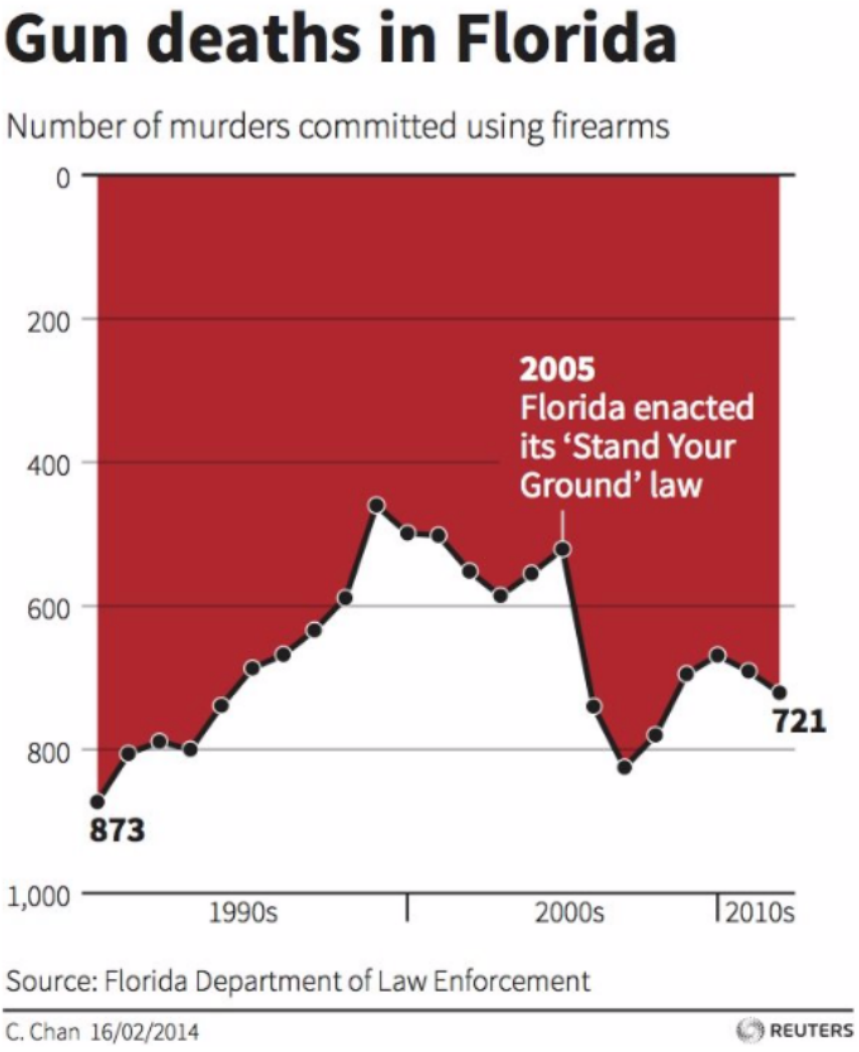


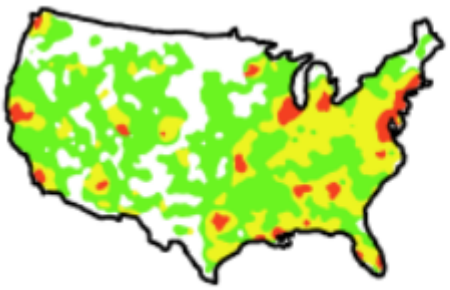

SOUTHWEST BORDER APPREHENSIONS
OCTOBER - APRIL
195,000
185,000
175,000
165,000
155,000
165,244
170,223
192,298
2011
2012
2013
FOX NEWS channel
Source: U.S. Border Patrol
Happening now
$43 BIL IN FEB, DOWN 3.4% FROM JAN'S REVISED
NAS
35.18


Average Global Degrees C
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
1900
1904
1908
1912
1916
1920
1924
1928
1932
1936
1940
1944
1948
1952
1956
1960
1964
1968
1972
1976
1980
1984
1988
1992
1996
2000
2004
2008
2012

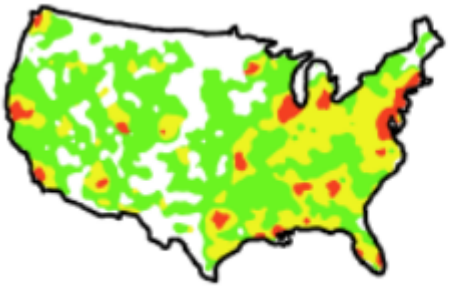


"In general, in a time-series, use a baseline that shows the data not the zero point" - Edward Tufte

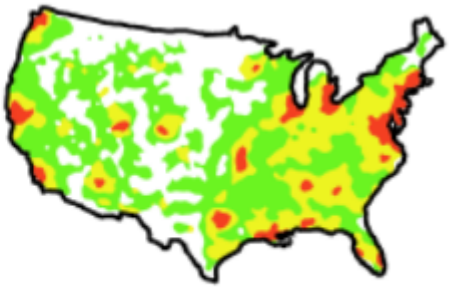




OUR SITE'S USERS

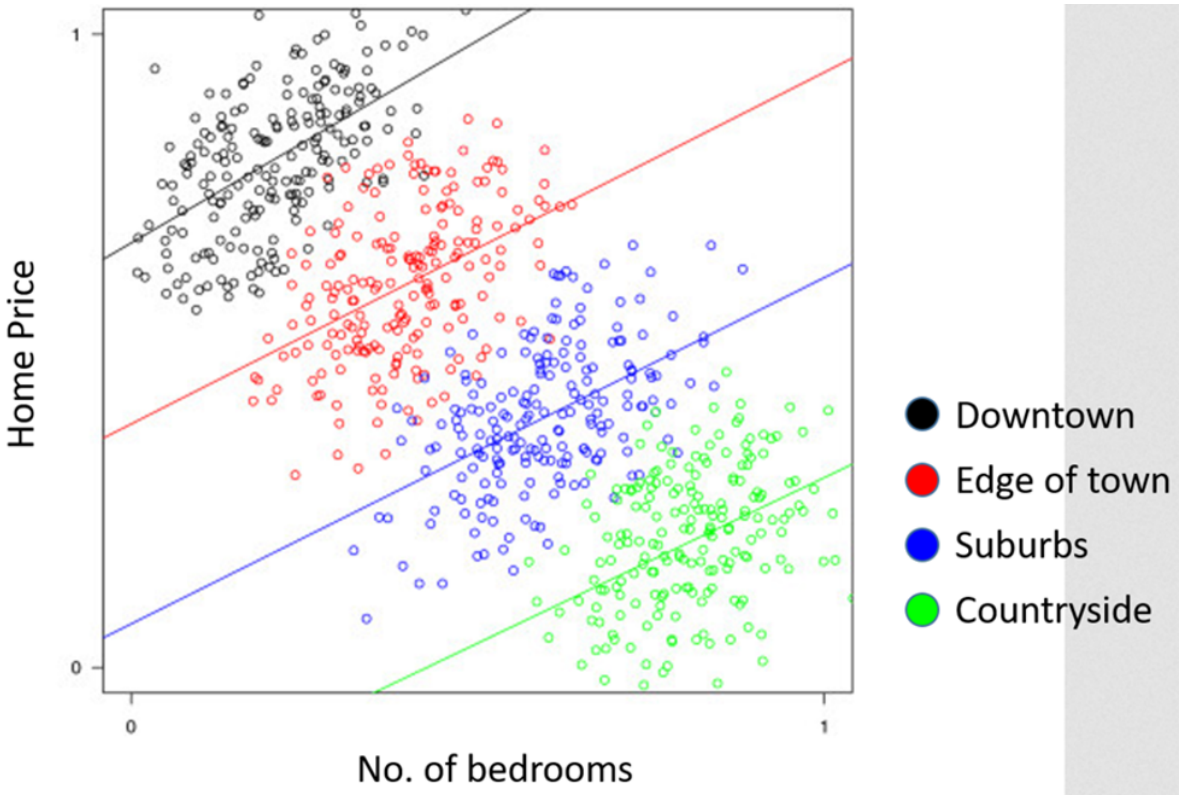


OUR SITE'S USERS



SUBSCRIBERS TO
MARTHA STEWART LIVING

Paradojas matemáticas: Simpson



Fairness

Equidad en los modelos de Machine Learning

Current Charges

☐ Homicide	☑ Weapons	☑ Assault	☐ Arson
☐ Robbery	☐ Burglary	☐ Property/Larceny	☐ Fraud
☐ Drug Trafficking/Sales	☐ Drug Possession/Use	☐ DUI/OUIL	☑ Other
☐ Sex Offense with Force	☐ Sex Offense w/o Force		

1. Do any current offenses involve family violence?
   ☑ No ☐ Yes
2. Which offense category represents the most serious current offense?
   ☐ Misdemeanor ☐ Non-violent Felony ☑ Violent Felony
3. Was this person on probation or parole at the time of the current offense?
   ☑ Probation ☐ Parole ☐ Both ☐ Neither
4. Based on the screener's observations, is this person a suspected or admitted gang member?
   ☐ No ☑ Yes
5. Number of pending charges or holds?
   ☑ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4+
6. Is the current top charge felony property or fraud?
   ☑ No ☐ Yes

Criminal History

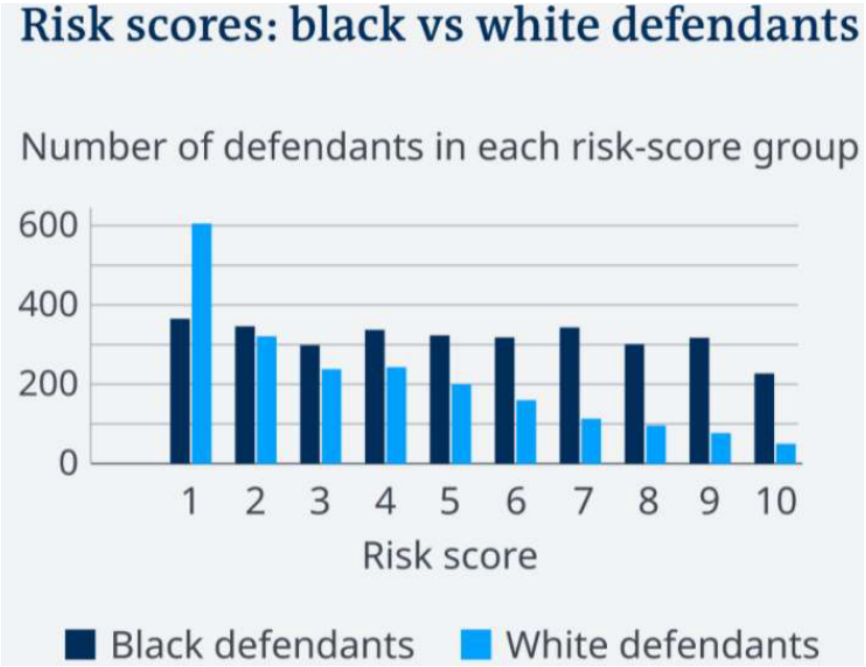
Exclude the current case for these questions.

7. How many times has this person been arrested before as an adult or juvenile (criminal arrests only)?
   5
8. How many prior juvenile felony offense arrests?
   ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☑ 4 ☐ 5+
9. How many prior juvenile violent felony offense arrests?
   ☐ 0 ☐ 1 ☑ 2+
10. How many prior commitments to a juvenile institution?
    ☐ 0 ☑ 1 ☐ 2+

All Defendants	Low	High
Survived	2681	1282
Recidivated	1216	2035
FP rate: 32.35		
FN rate: 37.40		
PPV: 0.61		
NPV: 0.69		
LR+: 1.94		
LR-: 0.55		

Black Defendants	Low	High
Survived	990	805
Recidivated	532	1369
FP rate: 44.85		
FN rate: 27.99		
PPV: 0.63		
NPV: 0.65		
LR+: 1.61		
LR-: 0.51		

White Defendants	Low	High
Survived	1139	349
Recidivated	461	505
FP rate: 23.45		
FN rate: 47.72		
PPV: 0.59		
NPV: 0.71		
LR+: 2.23		
LR-: 0.62		



All Defendants	Low	High
Survived	2681	1282
Recidivated	1216	2035
FP rate: 32.35		
FN rate: 37.40		
PPV: 0.61		
NPV: 0.69		
LR+: 1.94		
LR-: 0.55		

Black Defendants	Low	High
Survived	990	805
Recidivated	532	1369
FP rate: 44.85		
FN rate: 27.99		
PPV: 0.63		
NPV: 0.65		
LR+: 1.61		
LR-: 0.51		

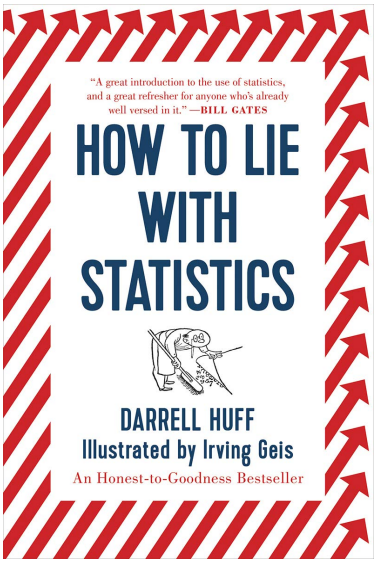
White Defendants	Low	High
Survived	1139	349
Recidivated	461	505
FP rate: 23.45		
FN rate: 47.72		
PPV: 0.59		
NPV: 0.71		
LR+: 2.23		
LR-: 0.62		

¿Es el modelo justo?

En el bootcamp de Data Science aprenderás programación para explorar la data, procesarla, aplicar modelos matemáticos, deep learning, nube… Lo más importante: aprenderás a pensar

en abstracto, a plantear soluciones que un ordenador puede ejecutar

Referencias



https://mathwithbaddrawings.com/

¡GRACIAS! :)