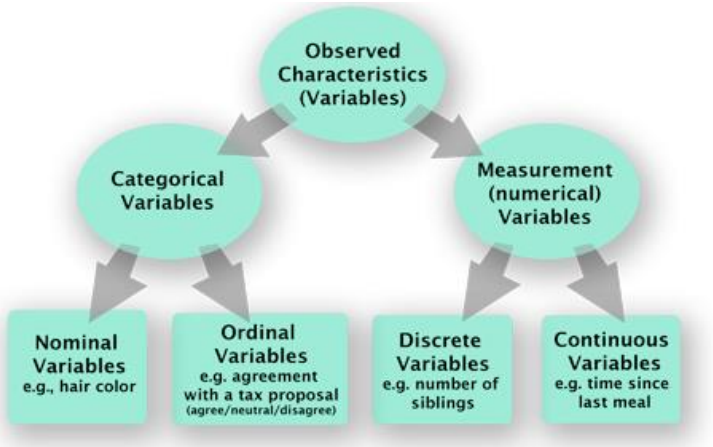
BRSM
Data Visualization & Summarization

Vinoo Alluri & Bapi Raju

ORGANIZATION

DATA VISUALIZATION

Data Organization



- identify variables (IV, DV) and respective types
- identify different levels of measurement

- missing data?
  - replace with mean
  - remove

20-25 years = 1
26-30 years = 2
31-35 years = 3
36-40 years = 4
41-45 years = 5
46 years and older= 6

Continuous → Categorical

Table format: XY		X	A		
		minutes	Test group A		
		X	A:Y1	A:Y2	A:Y3
1	Title	0	0.0	0.0	0.0
2	Title	2			
3	Title	4			
4	Title	6			
5	Title	8			
6	Title	10			
7	Title	12			
8	Title	14			
9	Title	16			
10	Title	18			
11	Title	20			

3	1	5.611248	4.1174493	5.4884806
4	2	5.5560017	3.9532921	5.2730727
5	3	4.5405	4.5603814	4.4634323
6	4	5.236287	3.8760467	5.198486
7	5	5.9417286	3.398312	5.965598
8	6	5.4199543	4.0421543	5.181767
9	7	4.4019384	3.394504	4.5349746
10	8	5.1843286	4.168893	4.9799395
11	9	5.3209386	3.9951186	5.300459
12	10	5.1961555	3.8243186	5.080454
13	11	5.5065527	3.938081	5.2821956
14	12	5.118871	3.8536696	5.1487226
15	13	5.4678555	3.9871855	5.308297
16	14	5.261652	3.4055495	5.6112976
17	15	5.9904175	4.116685	5.461459
18	16	3.838822	4.4964914	4.35598
19	17	5.68176	3.9998796	5.340737
20	18	4.433616	4.4853745	4.5518494
21	19	5.4475813	3.1434624	6.075916
22	20	5.3806806	3.8687606	5.3088202
23	21	5.417145	4.1244016	5.288509
24	22	5.8884277	4.254202	5.5335727

Summarize

to tell, in your own words, what has happened in the story

Summarize

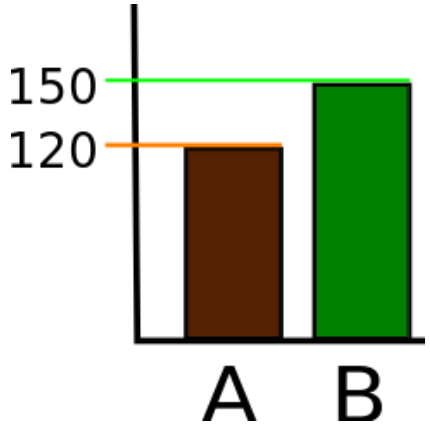
How?

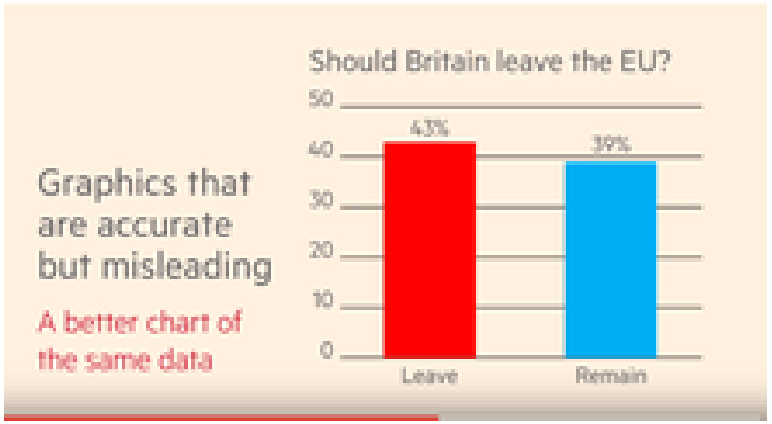
What information does it give ???

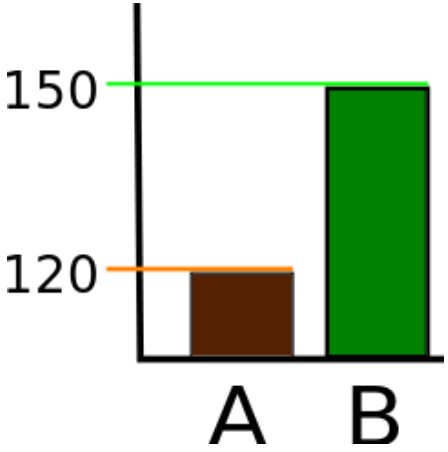
Outline

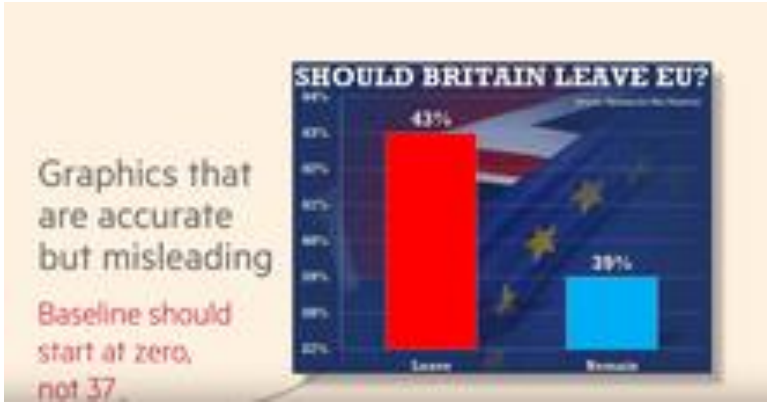
- **Visualization**
  - **why we visualise**
  - how to pick a plot
  - initial data vs final results visualization (some examples)
  - bad designs and misleading graphs
- **Summarization**
  - measures of central tendency & dispersion
  - which measure to pick

Data Visualisation



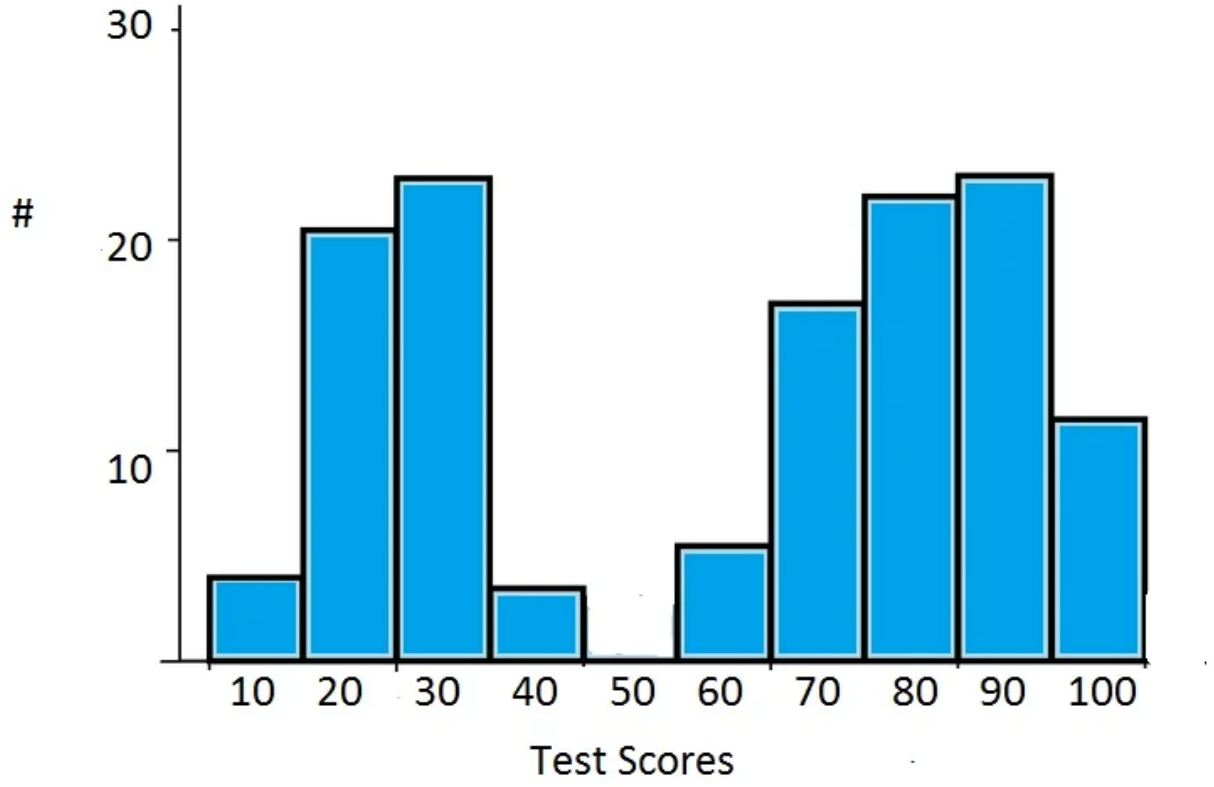






EXAMPLE

Mean Mid-Sem Test Score = 65.5



How can i summarise this data?

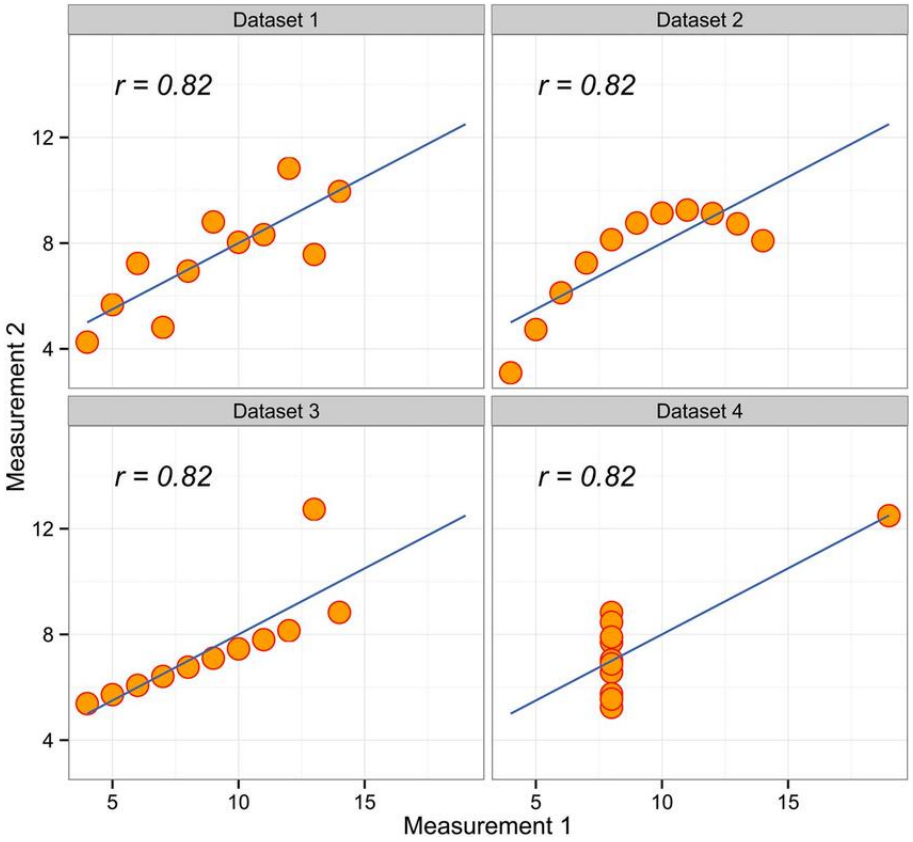
Anscombe's Quartet

- same mean, std, correlation, regression line

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	5.76
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	8.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	7.26	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

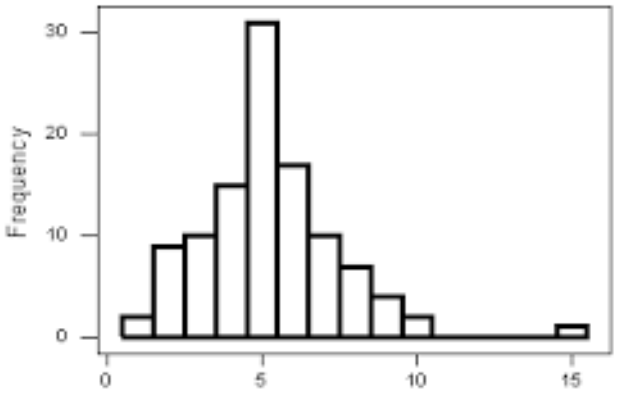
Anscombe's Quartet

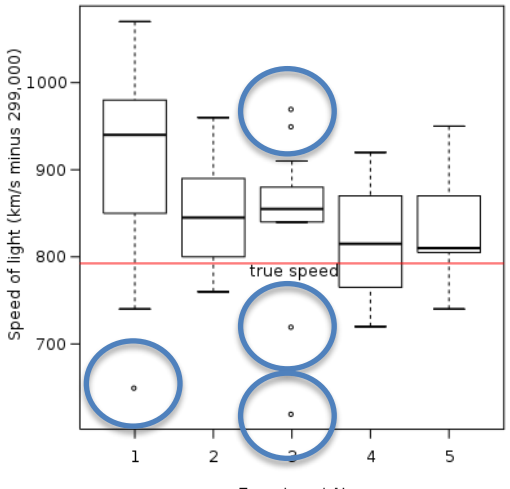
- same mean, std, correlation, regression line

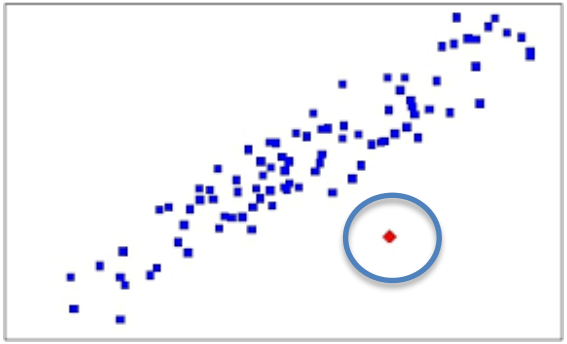


Why do we visualise?

- allows for initial guesses of data distribution
- direction of effect
- error detection (eg: missing, NaNs)
- outlier detection
- present results





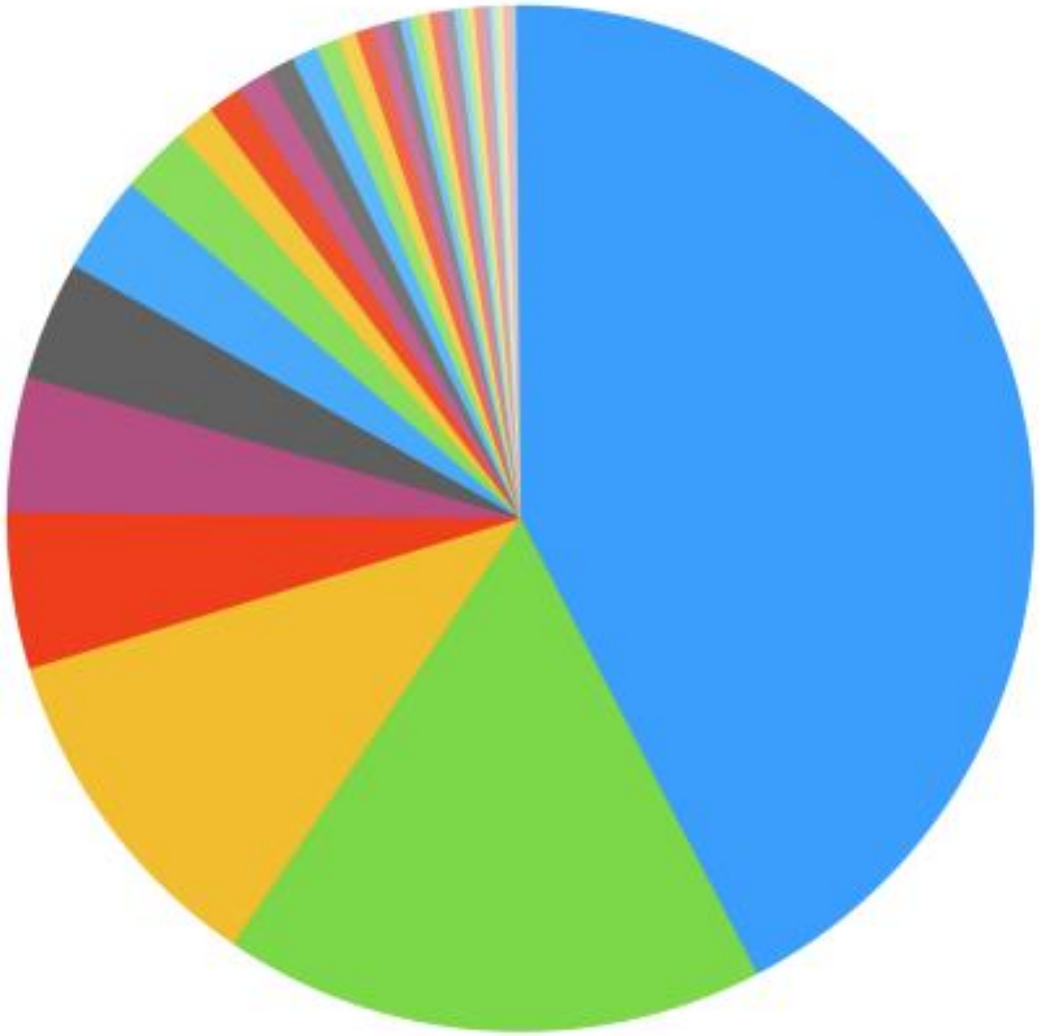


what makes them “good” or “bad”?

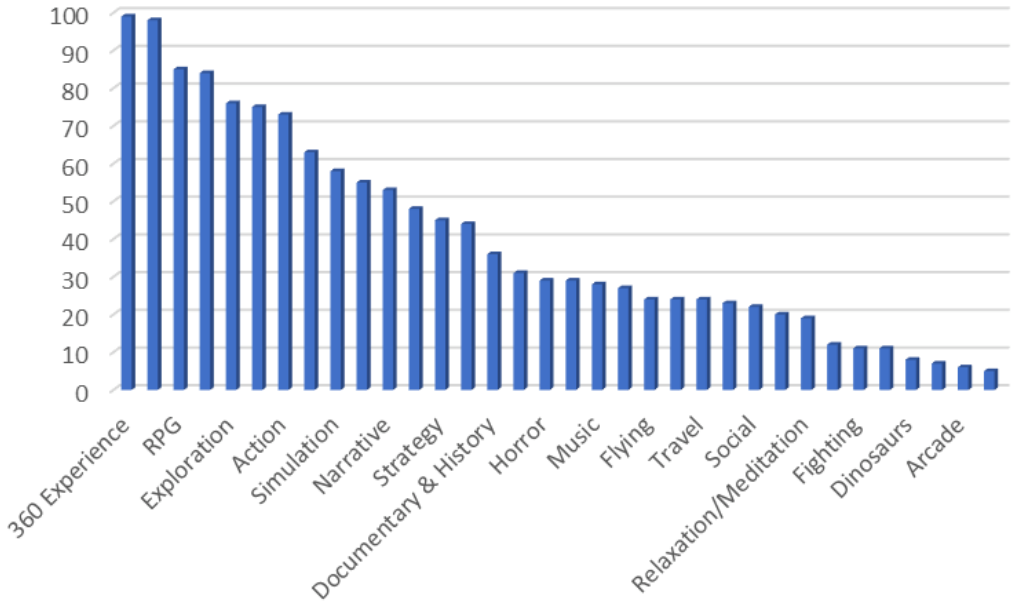
comment on these visualizations

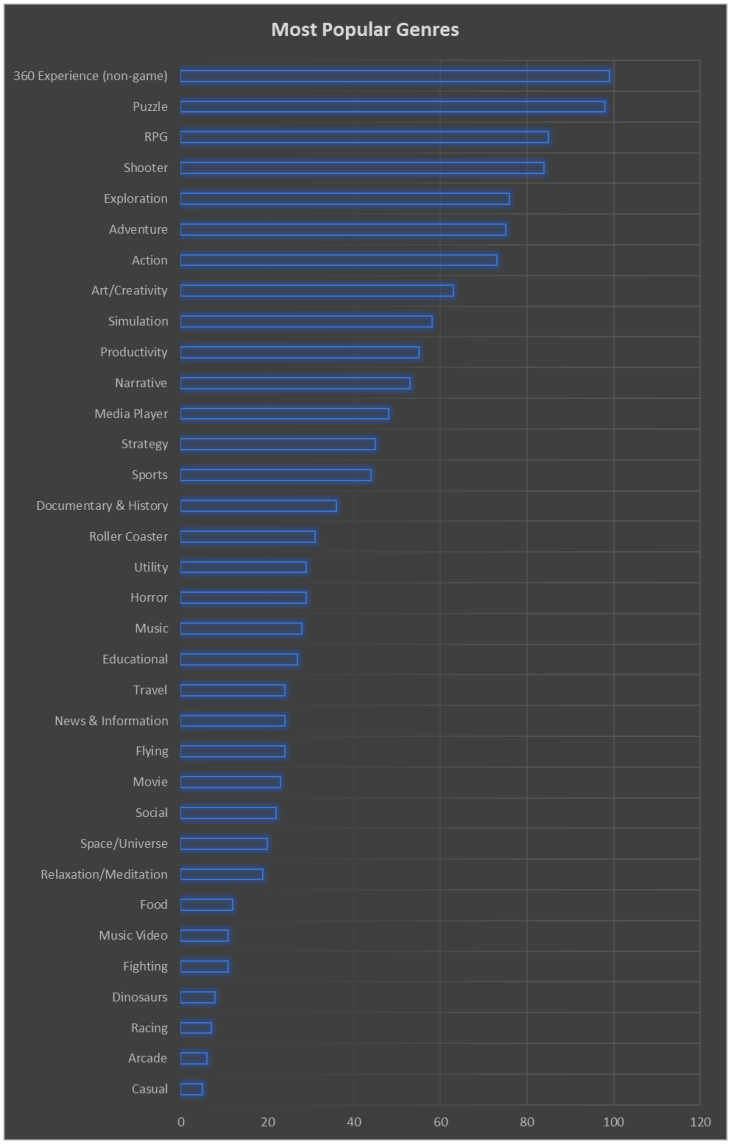
Which game(s) have you played the most?

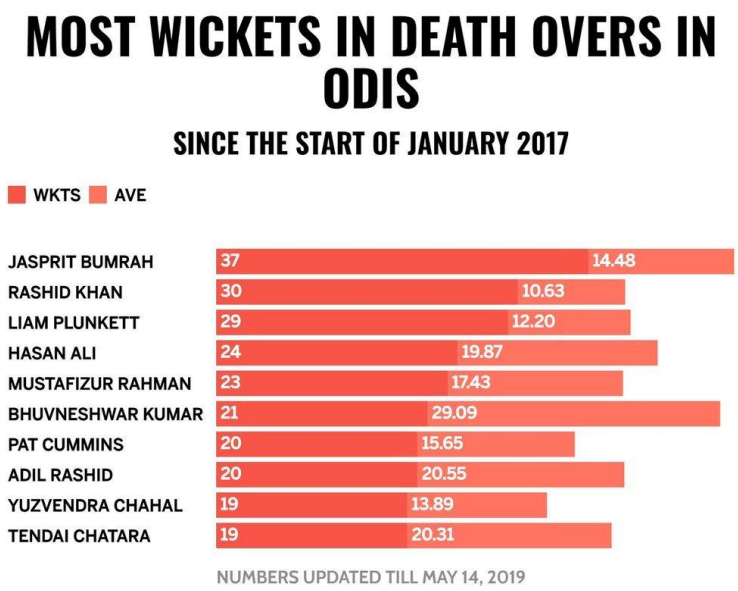
3,994 responses



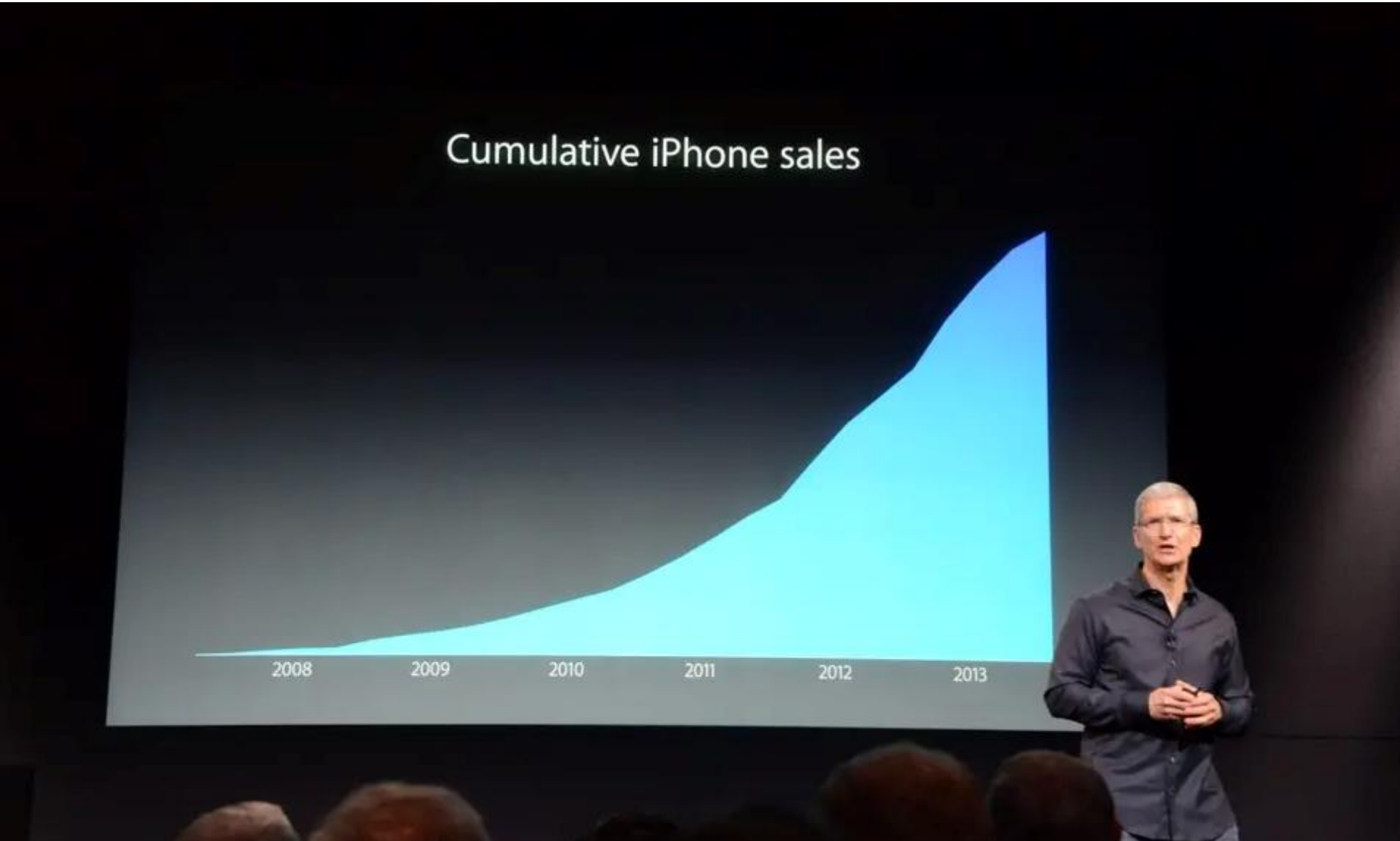
- Zelda
- The Legend of Zelda: Breath of the Wild
- Breath of the Wild
- BOTW
- Botw
- Breath of the wild
- BotW
- zelda
- Legend of Zelda: Breath of the Wild
- Legend of Zelda
- Zelda BOTW
- BoTW
- botw
- Zelda: Breath of the Wild
- Zelda BotW
- Zelda Breath of the Wild
- The Legend of Zelda
- Breath of The Wild
- The Legend of Zelda Breath of the Wild
- Zelda: BOTW
- Zelda: BotW
- Breath of the Wild
- Zelda breath of the wild
- Breath Of The Wild
- Legend of Zelda Breath of the Wild
- LoZ
- LoZ: BotW
- Zelda botw
- zelda botw
- breath of the wild
- Legend of zelda
- legend of zelda
- LoZ BOTW
- The Legend of Zelda: Breath of The Wild
- The legend of Zelda: breath of the wild
- ZELDA
- Zelda: BoTW

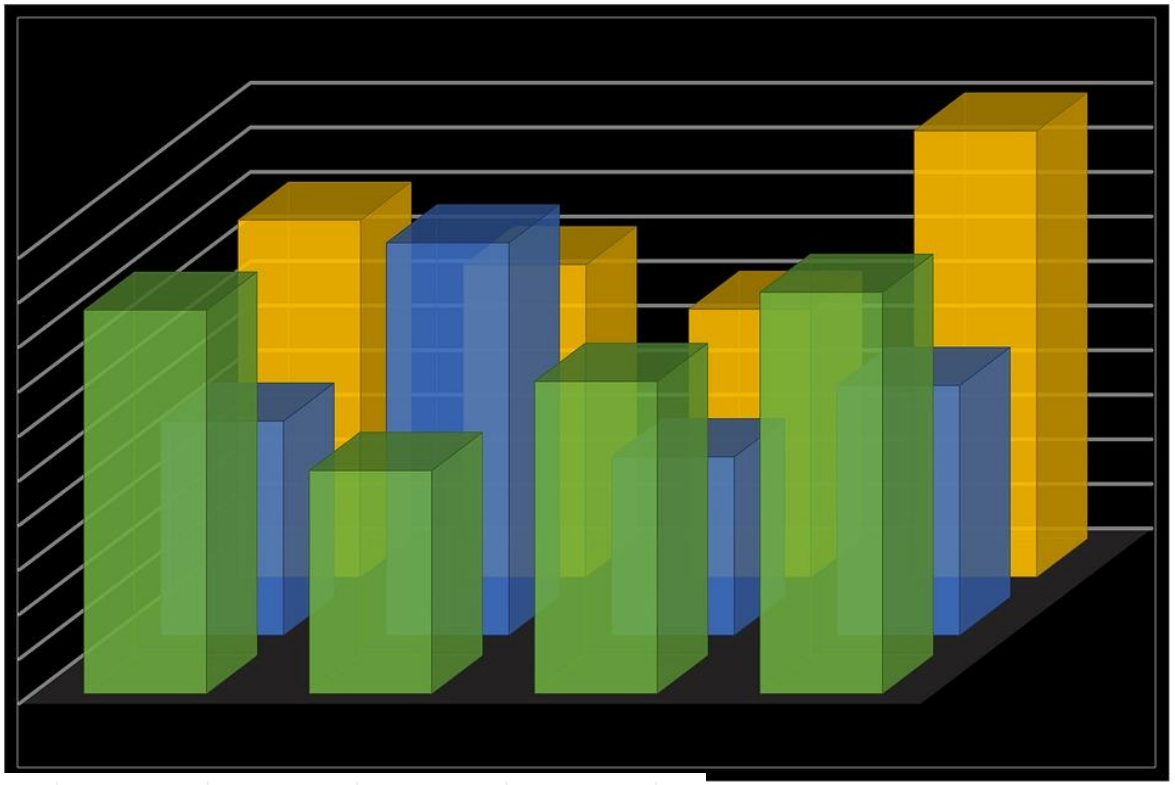

100
90
80
70
60
50
40
30
20
10
0
360 Experience
RPG
Exploration
Action
Simulation
Narrative
Strategy
Documentary & History
Horror
Music
Flying
Travel
Social
Relaxation/Meditation
Fighting
Dinosaurs
Arcade

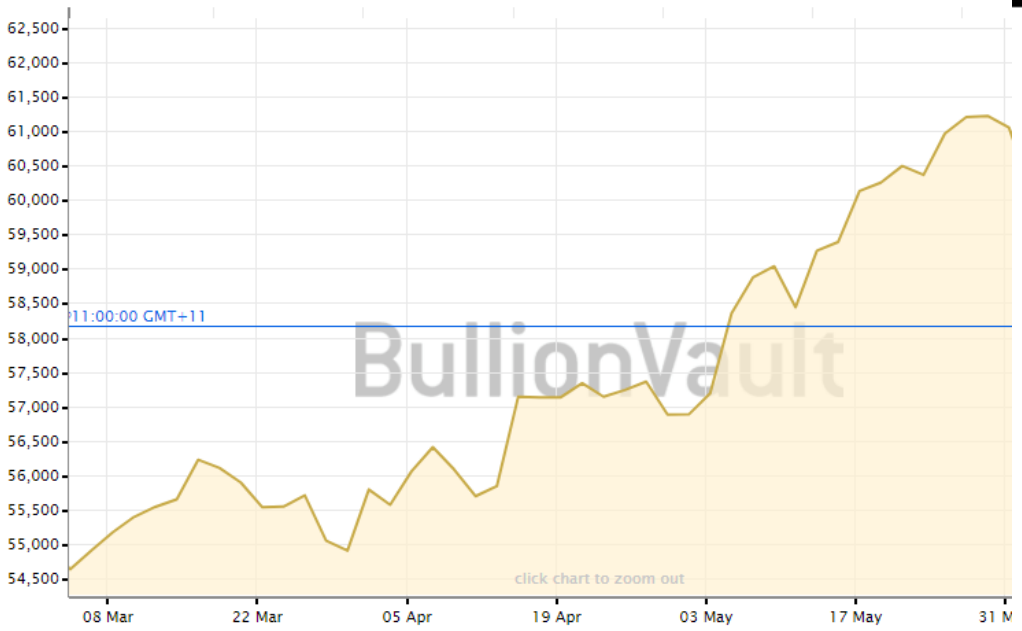

Most Popular Genres
360 Experience (non-game)
Puzzle
RPG
Shooter
Exploration
Adventure
Action
Art/Creativity
Simulation
Productivity
Narrative
Media Player
Strategy
Sports
Documentary & History
Roller Coaster
Utility
Horror
Music
Educational
Travel
News & Information
Flying
Movie
Social
Space/Universe
Relaxation/Meditation
Food
Music Video
Fighting
Dinosaurs
Racing
Arcade
Casual
0
20
40
60
80
100
120


MOST WICKETS IN DEATH OVERS IN ODIS
SINCE THE START OF JANUARY 2017
WKTS
AVE
JASPRIT BUMRAH 37 14.48
RASHID KHAN 30 10.63
LIAM PLUNKETT 29 12.20
HASAN ALI 24 19.87
MUSTAFIZUR RAHMAN 23 17.43
BHUVNESHWAR KUMAR 21 29.09
PAT CUMMINS 20 15.65
ADIL RASHID 20 20.55
YUZVENDRA CHAHAL 19 13.89
TENDAI CHATARA 19 20.31
NUMBERS UPDATED TILL MAY 14, 2019

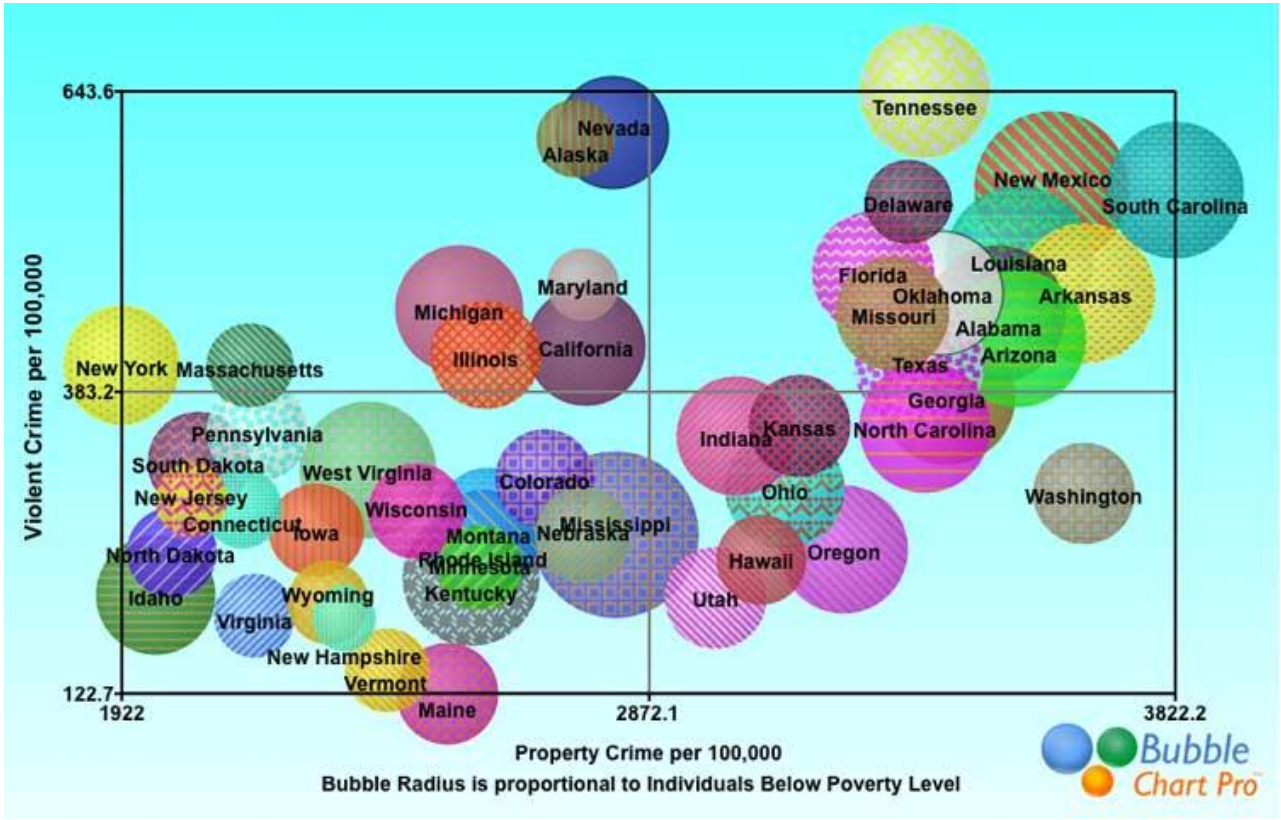
Tim Cook used the particular chart to showcase the rising sale of iPads between the years 2008-2013.

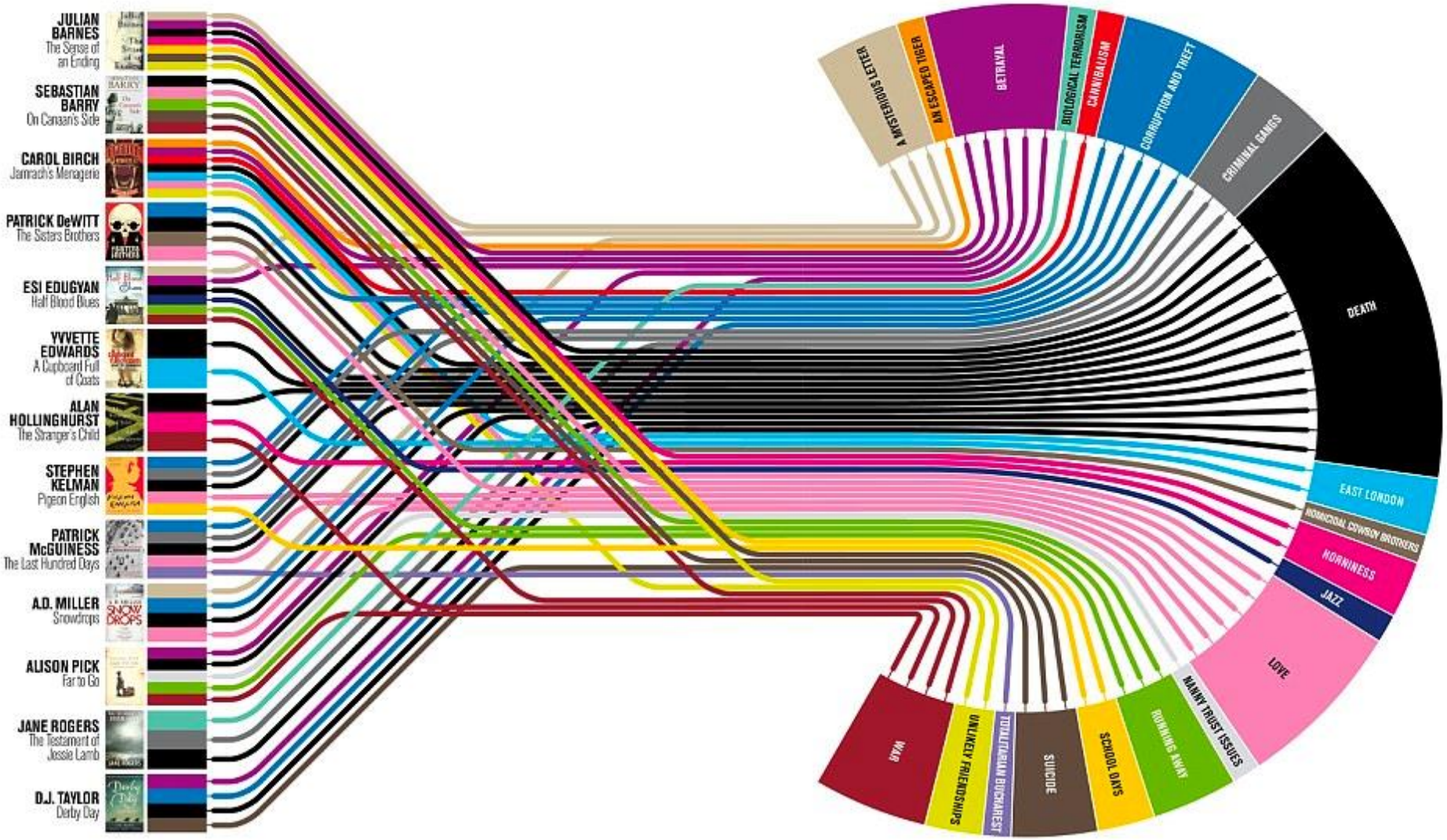





62,500
62,000
61,500
61,000
60,500
60,000
59,500
59,000
58,500
58,000
57,500
57,000
56,500
56,000
55,500
55,000
54,500
11:00:00 GMT+11
BullionVault
click chart to zoom out
08 Mar
22 Mar
05 Apr
19 Apr
03 May
17 May
31 M


Number of sales for each product
100
95
90
85
80
75
70
Product 1
Product 2


643.6
383.2
122.7
Violent Crime per 100,000
1922
2872.1
3822.2
Property Crime per 100,000
Bubble Radius is proportional to Individuals Below Poverty Level
Nevada
Alaska
Tennessee
New Mexico
South Carolina
Delaware
Florida
Louisiana
Oklahoma
Arkansas
Missouri
Alabama
Arizona
Texas
Georgia
North Carolina
Maryland
Michigan
California
Illinois
New York
Massachusetts
Pennsylvania
South Dakota
West Virginia
New Jersey
Connecticut
Wisconsin
Colorado
Indiana
Kansas
Ohio
Washington
Iowa
Montana
Nebraska
Mississippi
Rhode Island
Minnesota
Hawaii
Oregon
North Dakota
Idaho
Wyoming
Kentucky
Utah
Virginia
New Hampshire
Vermont
Maine
Bubble Chart Pro



Plot lines

What makes a prize-winning novel? As Julian Barnes wins the Booker Prize, Delayed Gratification's Johanna Kamradt charts the themes of this year's longlisters.

What makes a good visualisation?

- reduce cognitive Load
  - simplicity
  - relevancy
  - less is more
- storytelling
  - ability to support the reader during their journey
  - convince the reader

"Perfection is achieved not when there is nothing more to add, but when there is nothing left to take away"
– Antoine de Saint-Exupery

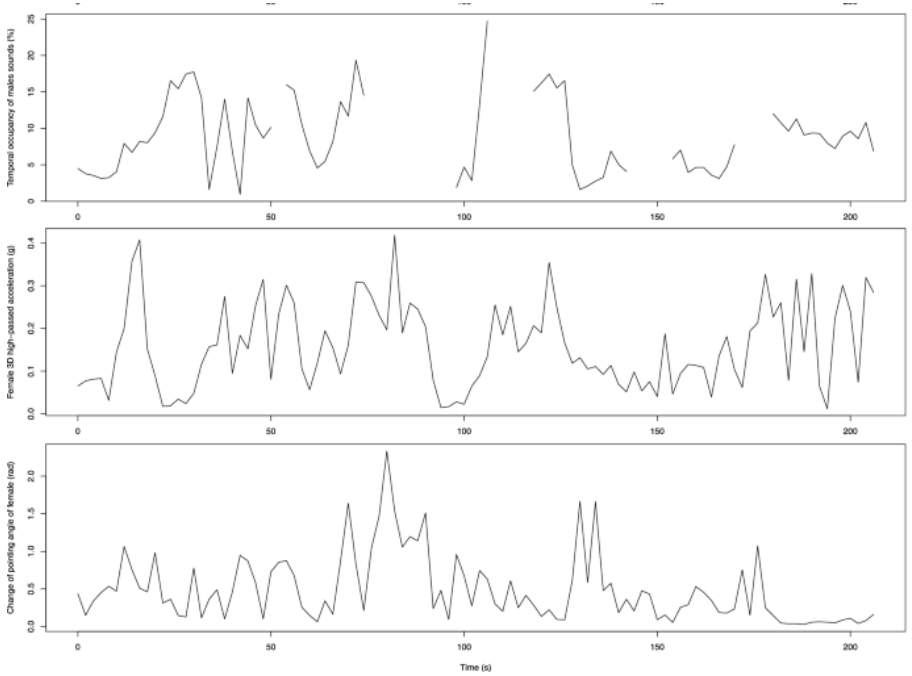
What makes a good visualisation

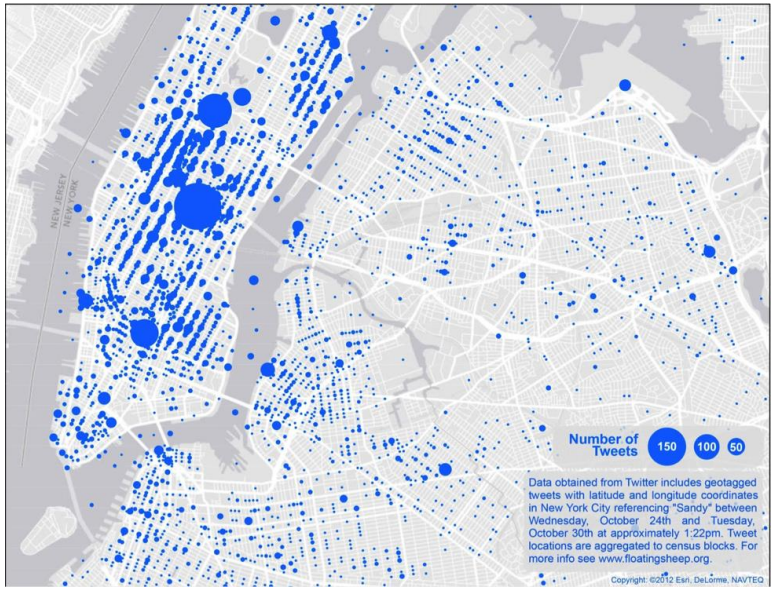
- Color Consistency
  - use same colors across multiple charts for consistency
  - avoid using colors with negligible contrast
  - avoid using too many colors
  - avoid using conventional colors to convey opposite meanings
  - pay heed to the needs of people who might be colorblind (check also in grayscale)
- Accurate Scaling

https://www.cardinalpath.com/blog/makes-good-visualization

What makes a good visualisation

- identify & explain/infer from missing data





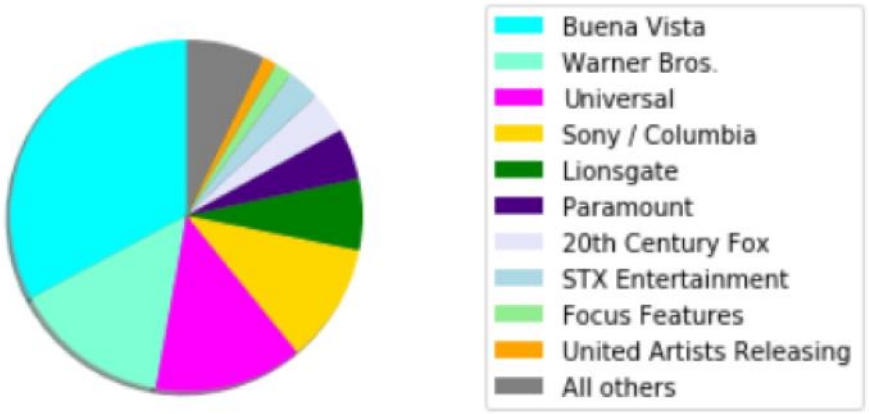
https://www.cardinalpath.com/blog/makes-good-visualization

What makes a good visualisation

- labelling
  - label the axis correctly and consistently across all your charts.
  - avoid using acronyms that are not widely understood.
  - make the chart title as concise and descriptive as possible.
  - whenever possible, label the lines in your line chart directly rather than using a legend.
  - be consistent in formatting; if you are working with currency symbols, percentage signs and the decimal values, retain them across all your charts.

https://www.cardinalpath.com/blog/makes-good-visualization

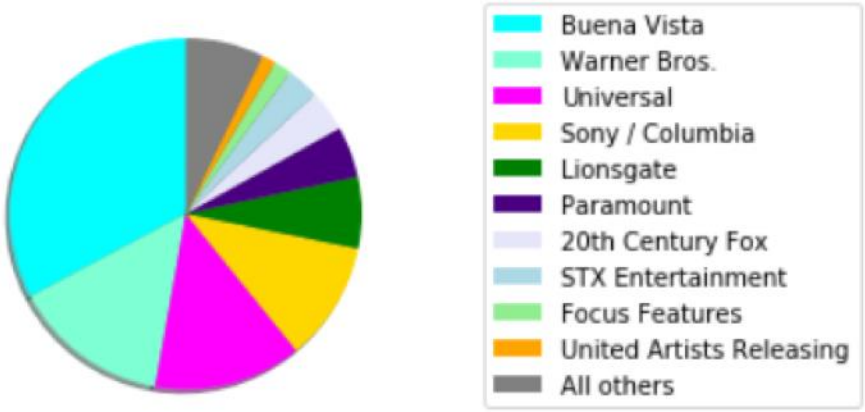
EXAMPLE



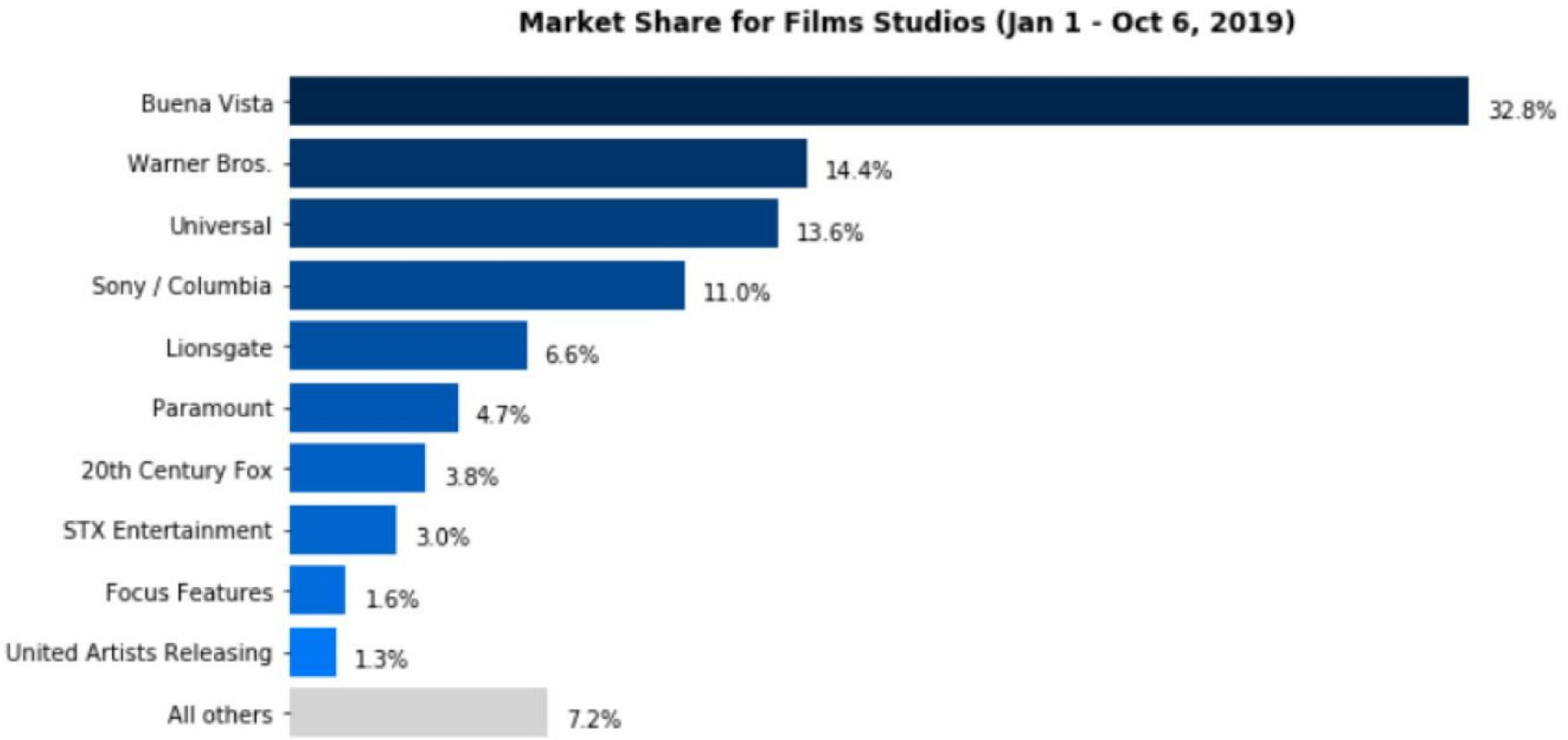
Market Share of Film Studios

PIE CHART

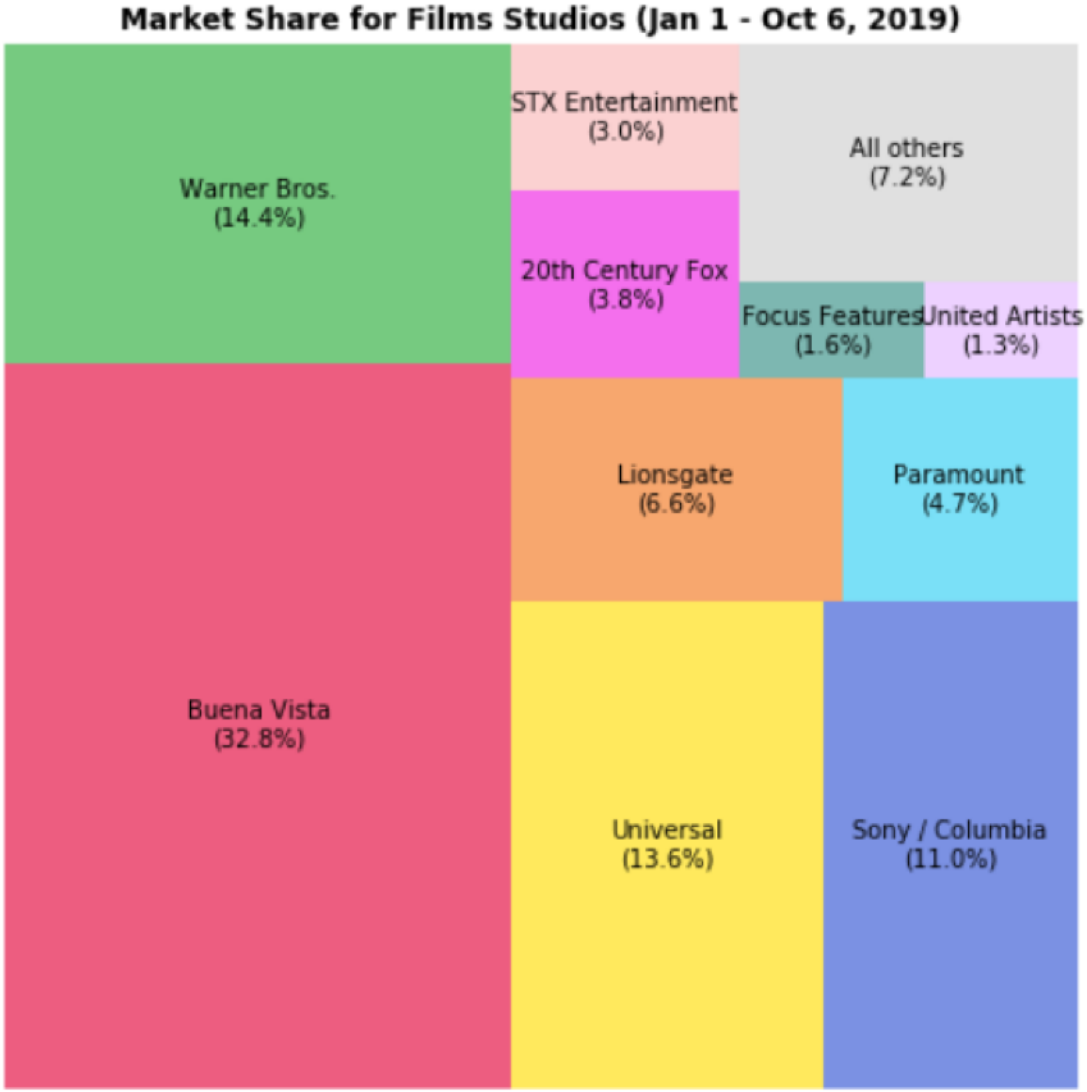
Not comprehensible!



BAR CHART

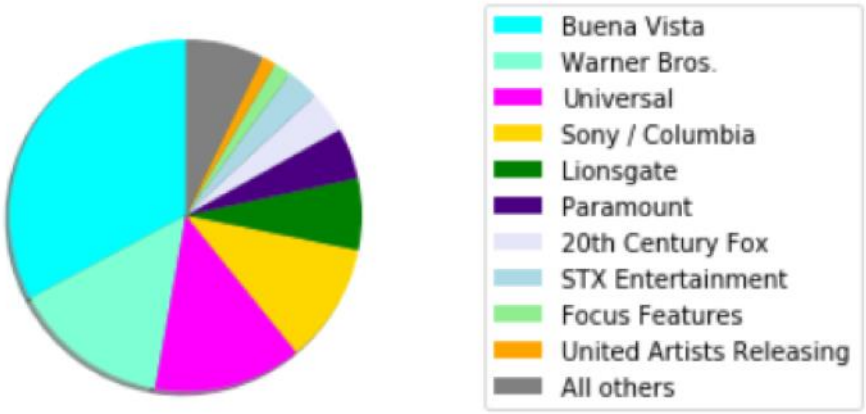


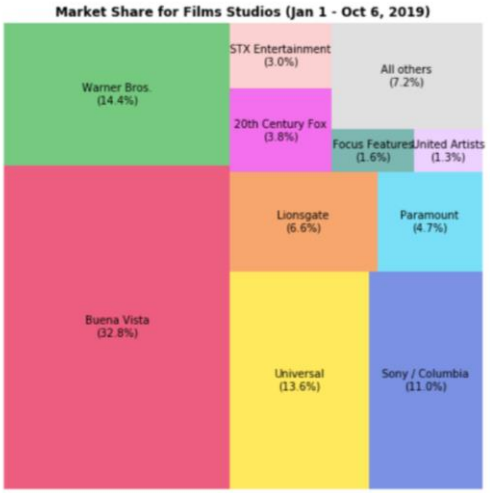
AREA PLOTS: TREE MAP

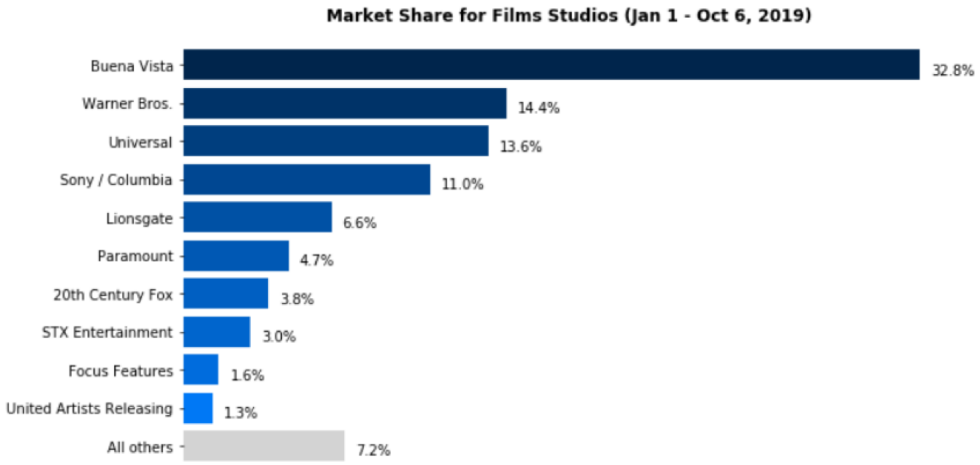


AREA PLOTS: WAFFLE CHART

So which visualisation was best?



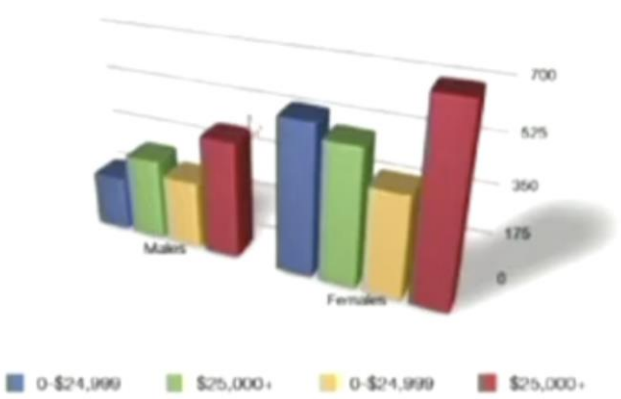


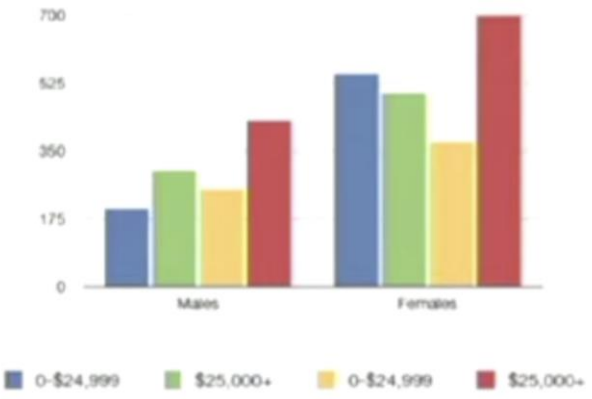


Tufte's Graphical Theory

- minimize data-to-ink ratio
- minimise lie factor (or increase graphical integrity)
- minimise chart junk
- use proper scales and labelling

$$\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}$$



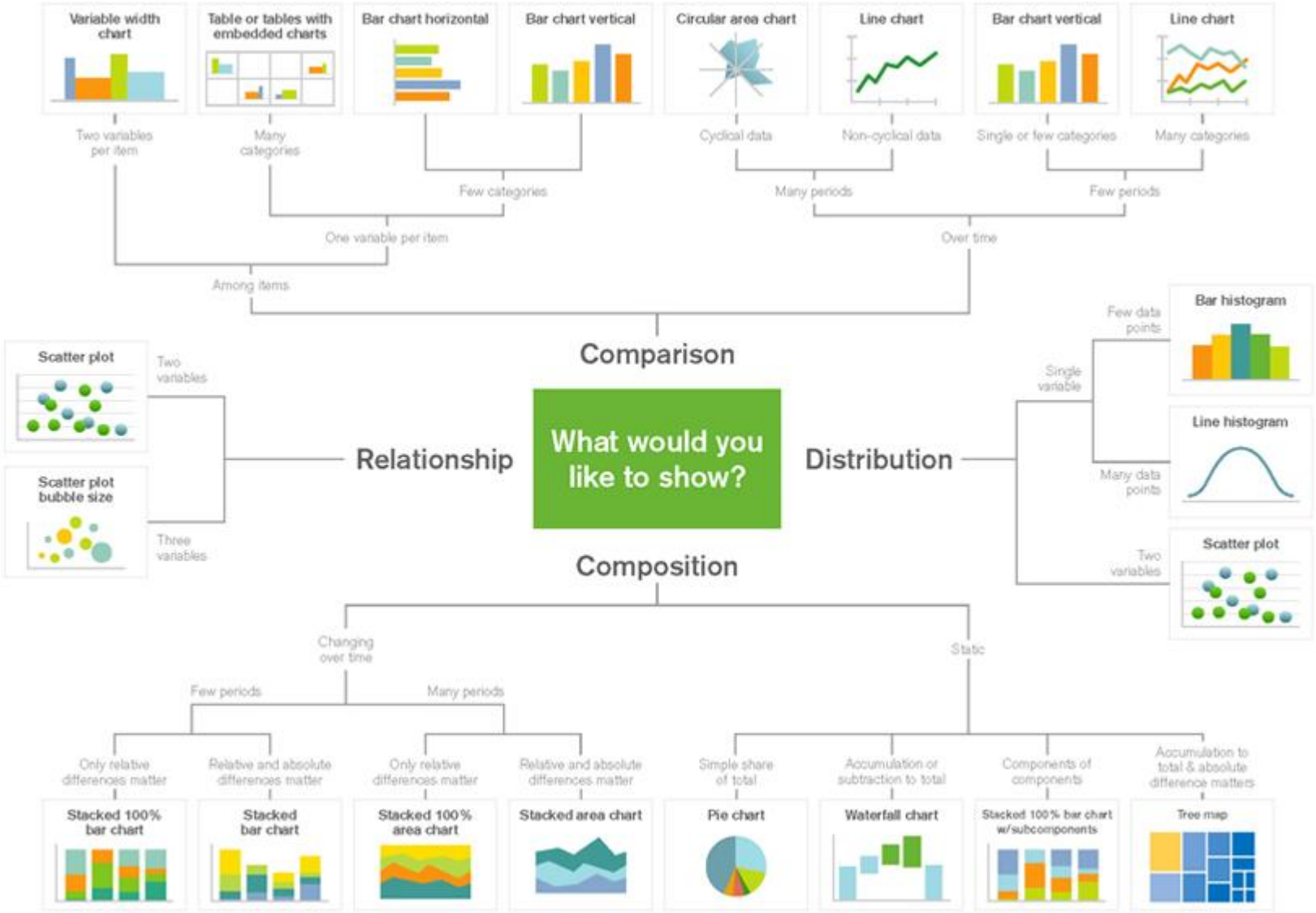


The good, the bad, & the ugly misleading

Outline

- **Visualization**
  - why we visualise
  - **how to pick a plot**
  - **initial data vs final results visualization (some examples)**
  - bad designs and misleading graphs
- **Summarization**
  - measures of central tendency & dispersion
  - which measure to pick

How to choose the right plot?



How to choose the right plot?

- **distributions & compositions**
  - proportions
  - data distributions
- **comparisons**
  - group differences
- **associations**
  - relationships between variables
  - geographical data
- **variable types**

Initial Data vs Final Result Visualization

HISTOGRAMS

BOX-PLOT

SCATTER PLOT

PIE CHARTS & BAR CHARTS

MOSAIC PLOT

VIOLIN PLOT

RAIN-DROP

FUNNEL PLOTS

SPIDER PLOT / RADAR CHART

RADIAL HEAT MAP

CIRCOS PLOT

STREAMGRAPH

Not an exhaustive list!

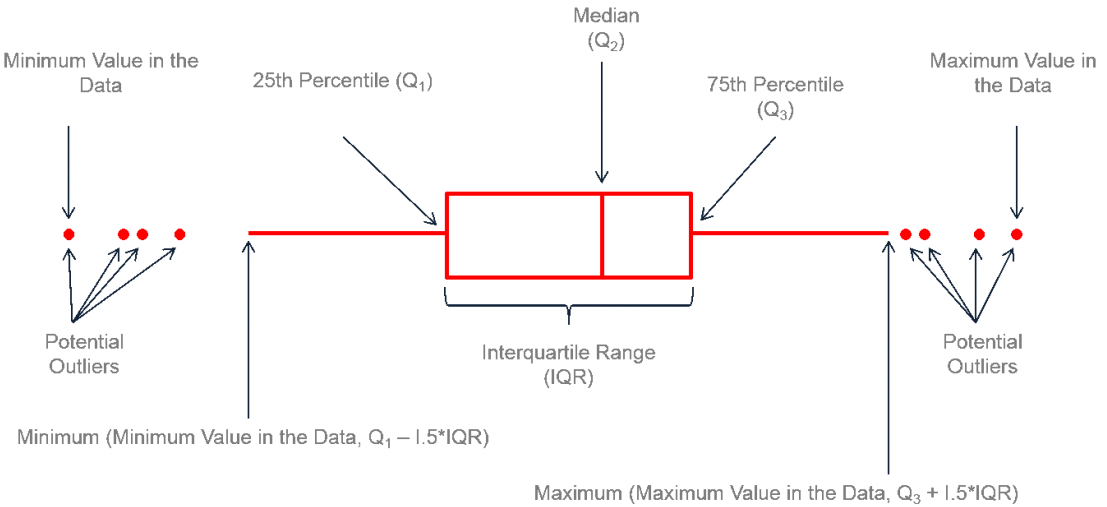
Some plots used for both!

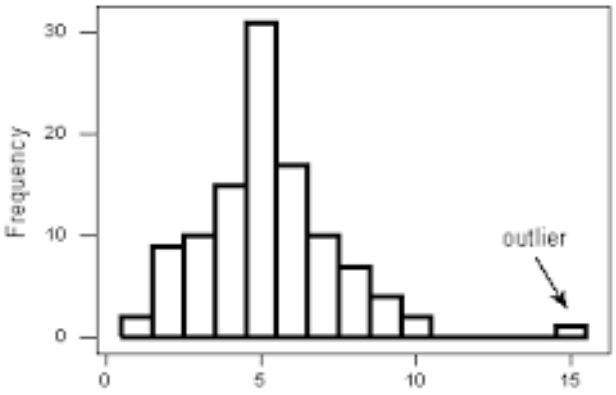
Data Visualization: What Info?

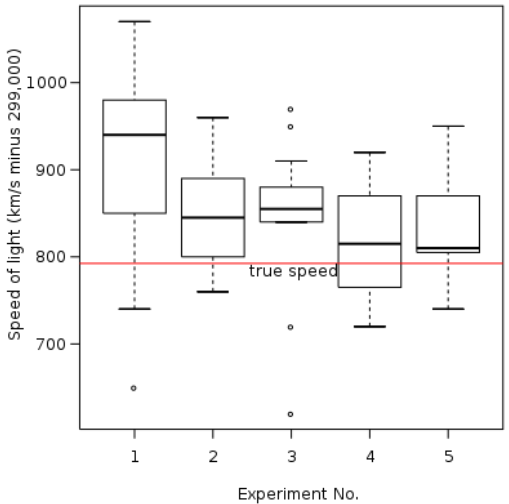
- allows for initial guesses of data distribution (normality) and direction of effect
  - ex: histograms (bin-width dependency), box-plots, scatter plots, pie charts, bar plots (already seen), etc

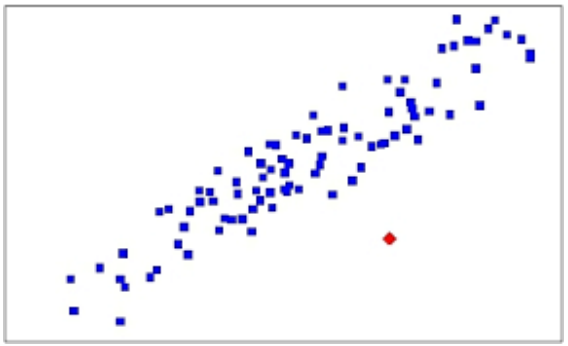
Data Visualisation

HISTOGRAM, BOXPLOT, SCATTER PLOT

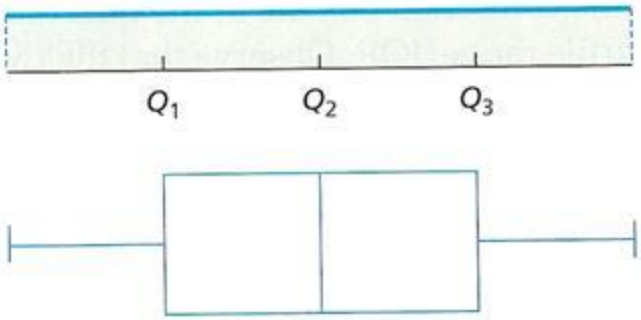




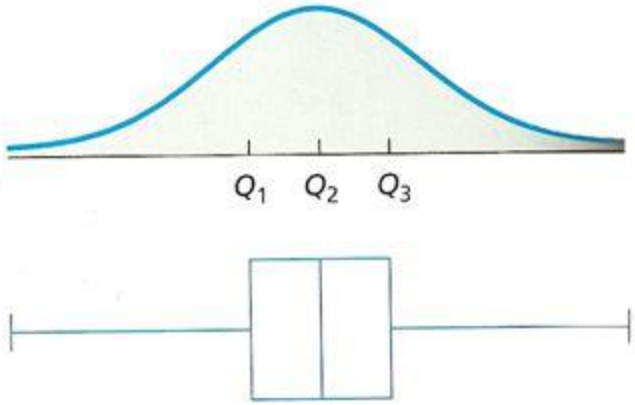




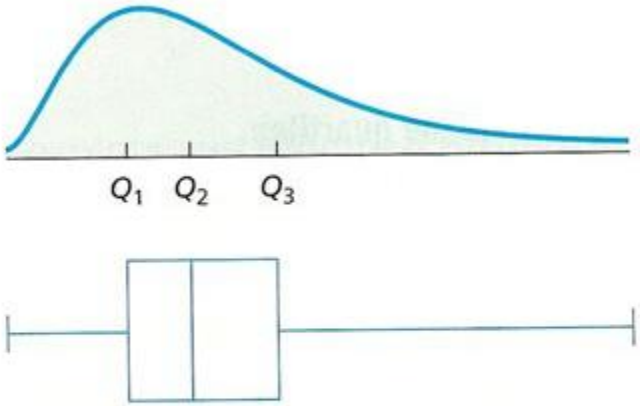
Boxplot

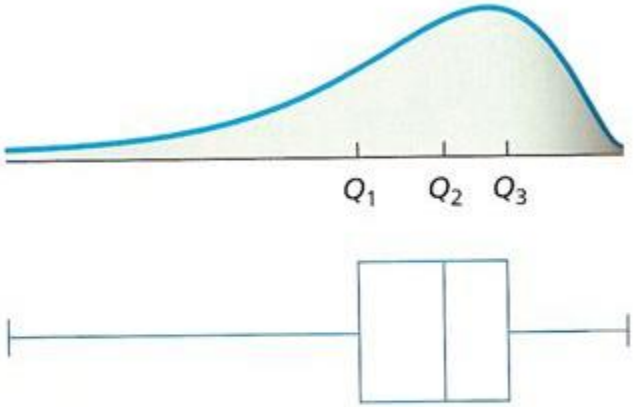


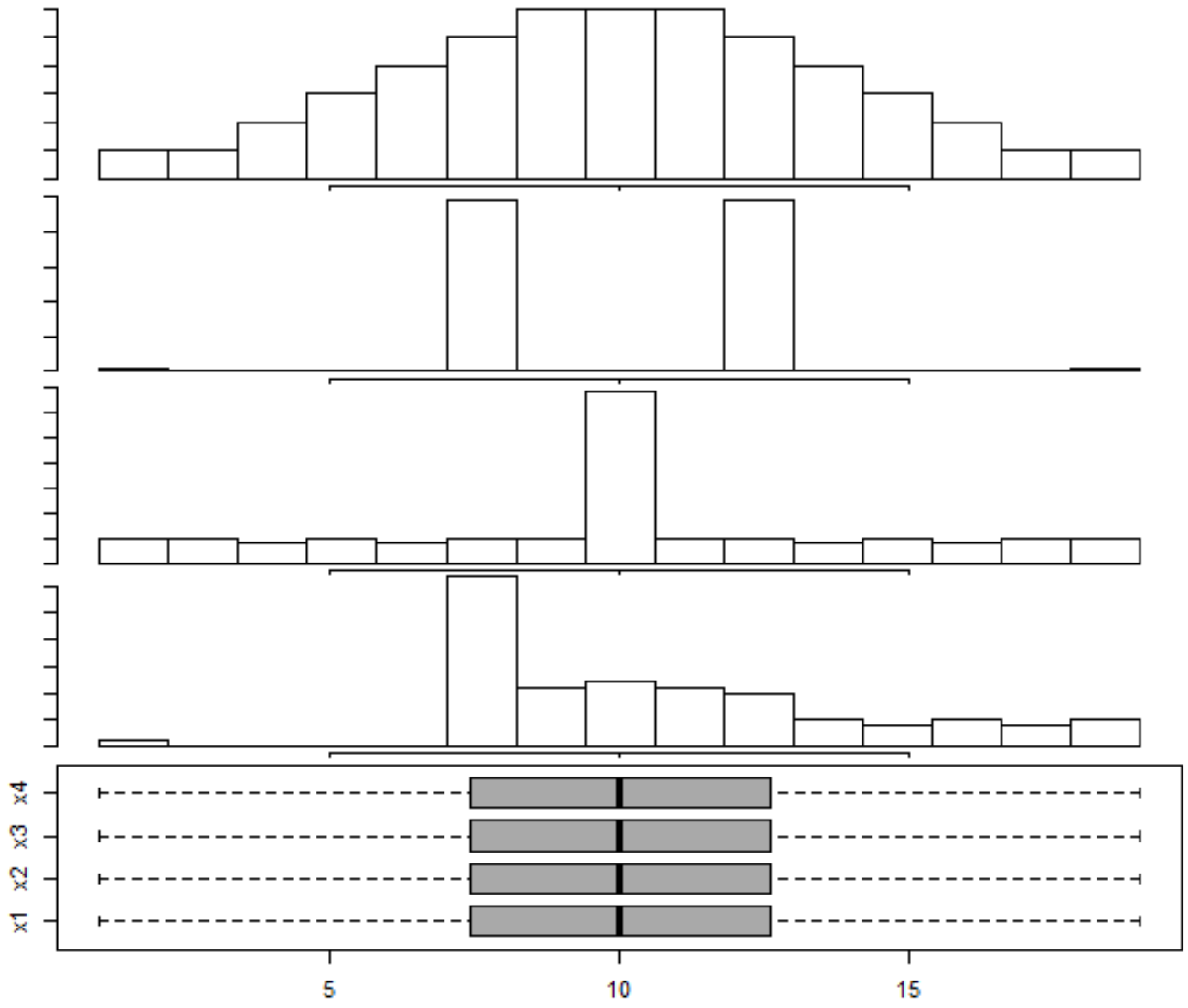
(a) Uniform



(b) Bell shaped

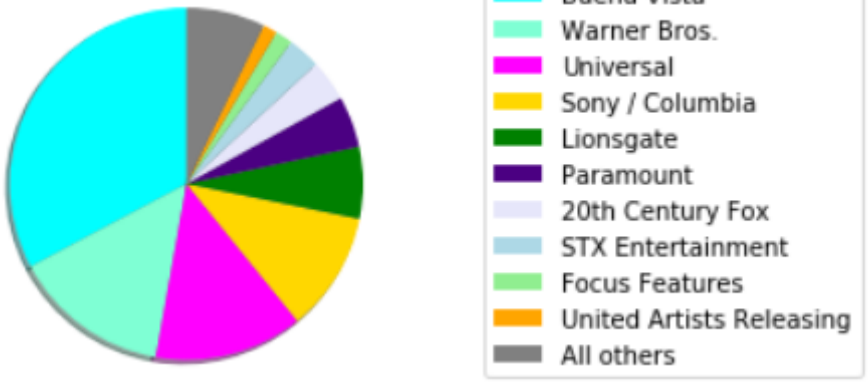




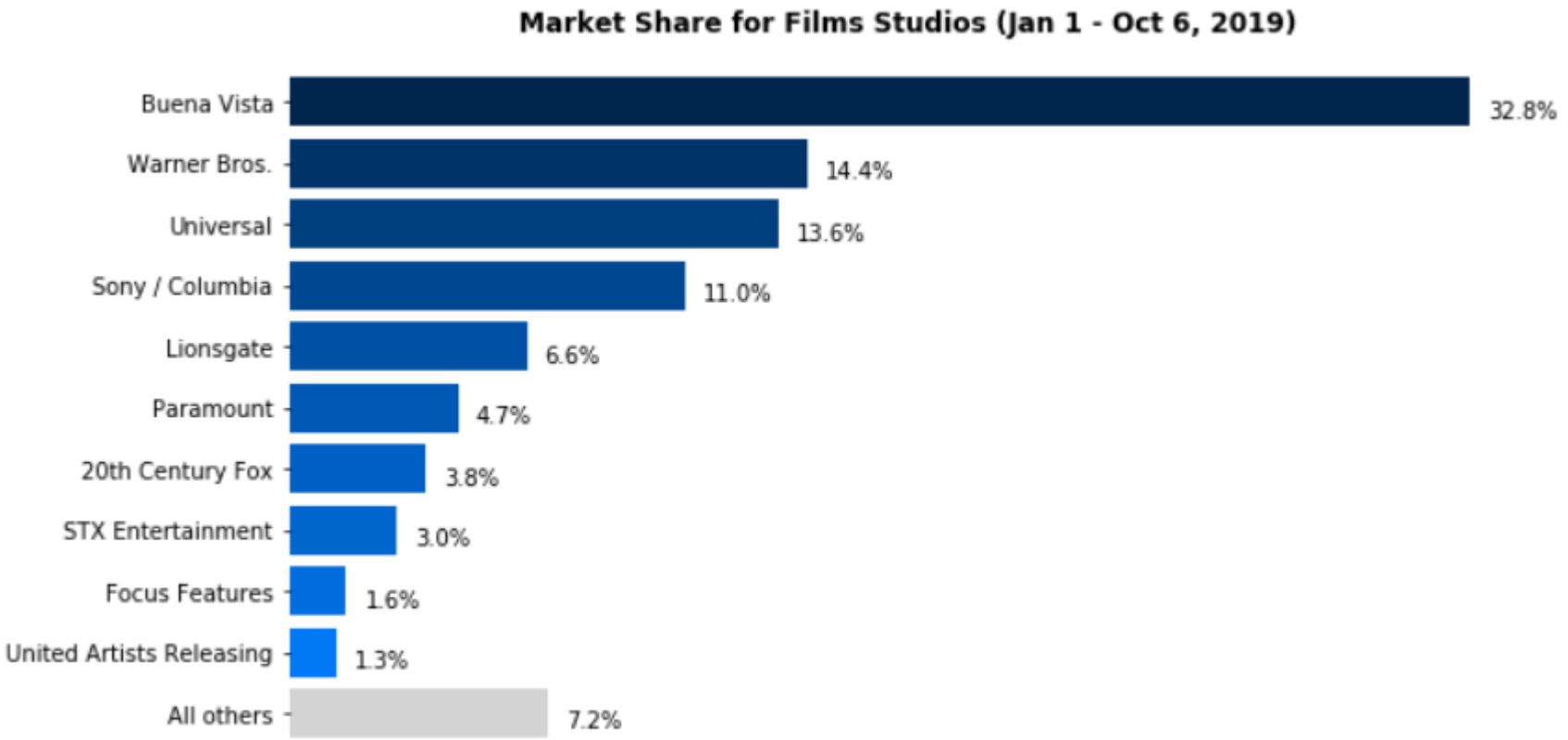

x4
x3
x2
x1
5
10
15

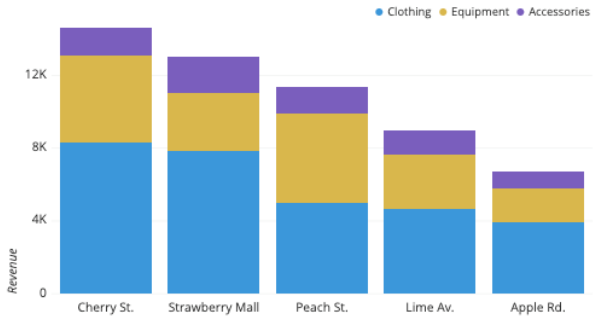
PIE CHART

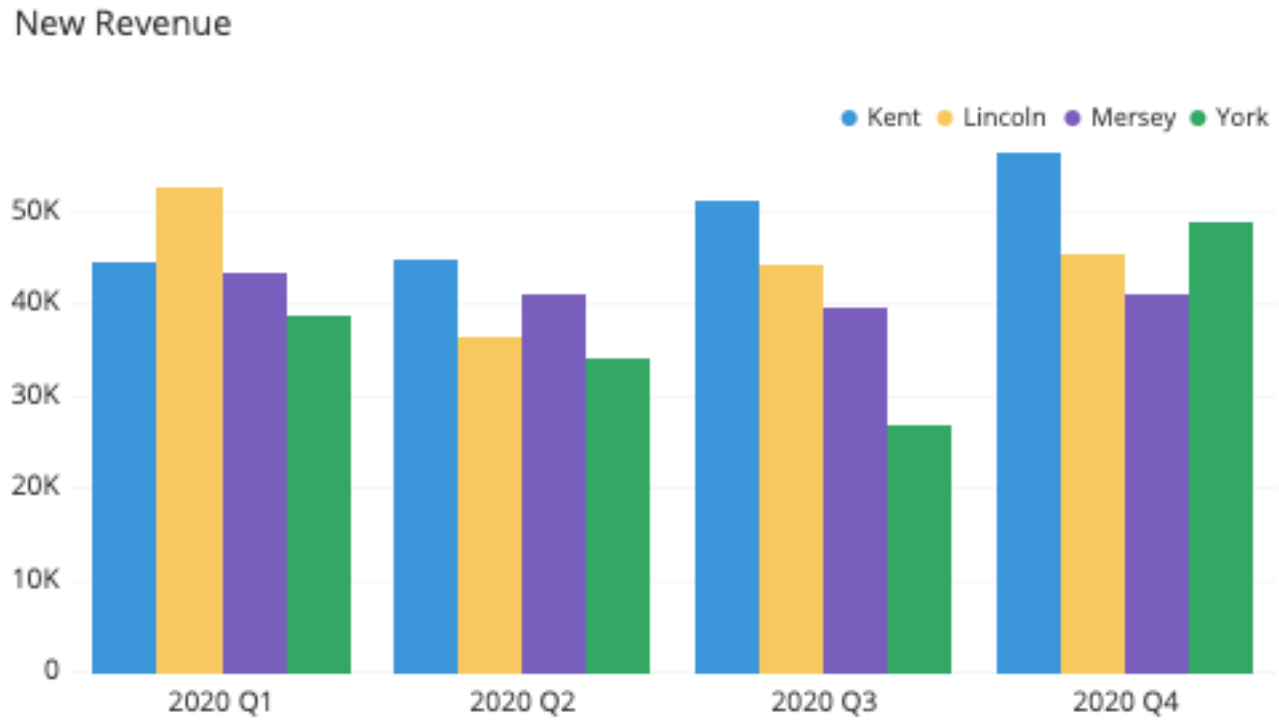
Not comprehensible!



BAR CHART



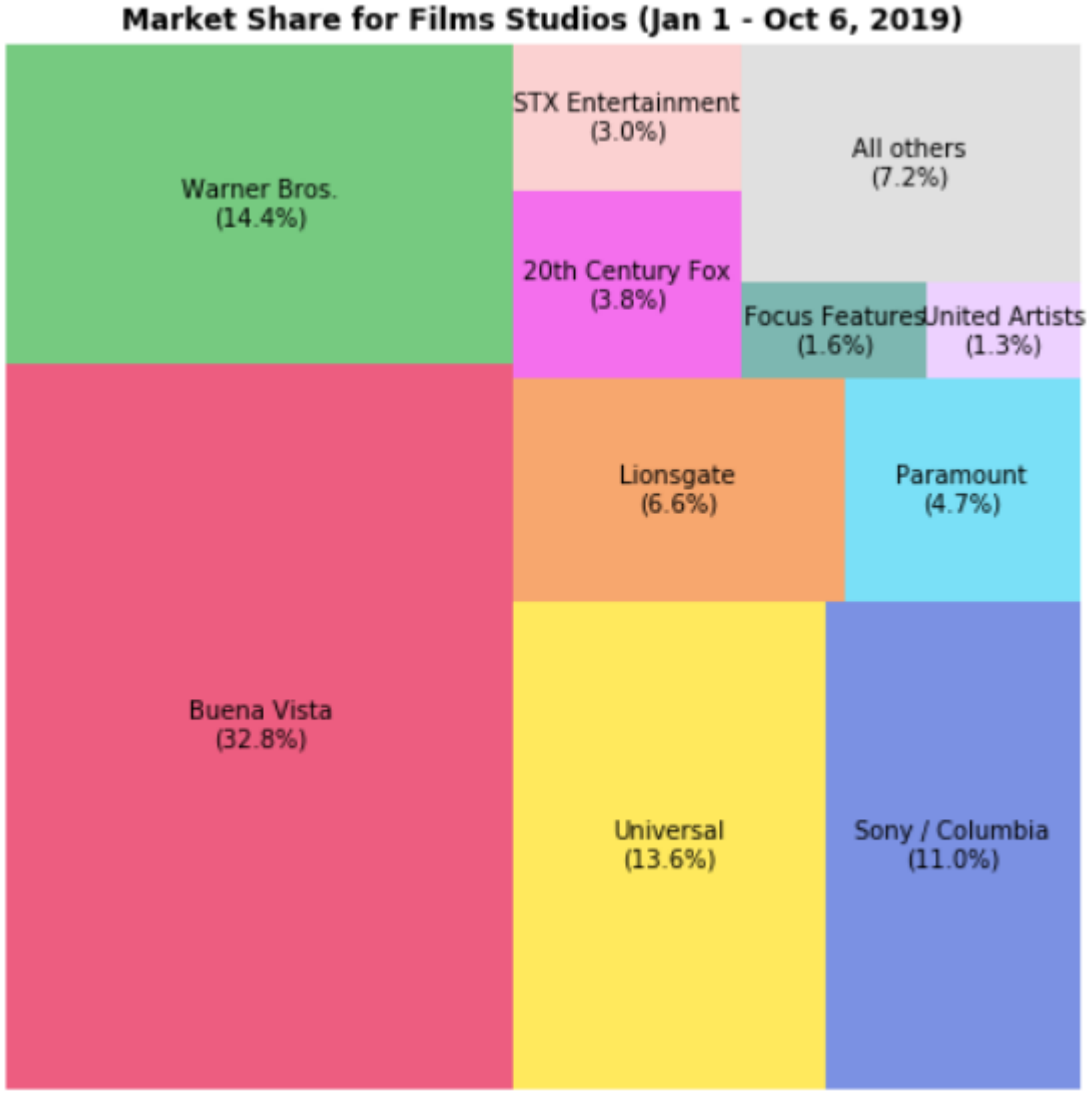

Clothing
Equipment
Accessories
12K
8K
4K
0
Revenue
Cherry St.
Strawberry Mall
Peach St.
Lime Av.
Apple Rd.


New Revenue
Kent
Lincoln
Mersey
York
50K
40K
30K
20K
10K
0
2020 Q1
2020 Q2
2020 Q3
2020 Q4

Pie vs Bar Charts

- use pie charts when
  - smaller no. of categories
  - readers can differentiate slices (unless you are making a point)
  - you don’t need to rely on many colors or labels to explain the proportions
  - total adds up to 100%
- use bar charts when
  - have many categories (not too many)
  - need to compare numbers side-by-side (caution: more than two bars are hard for readers

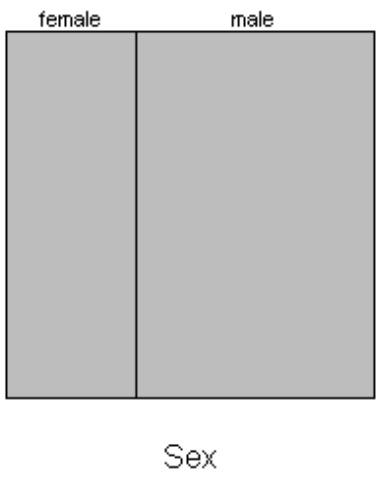
AREA PLOTS: TREE MAP

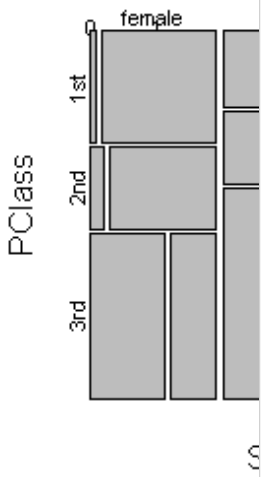


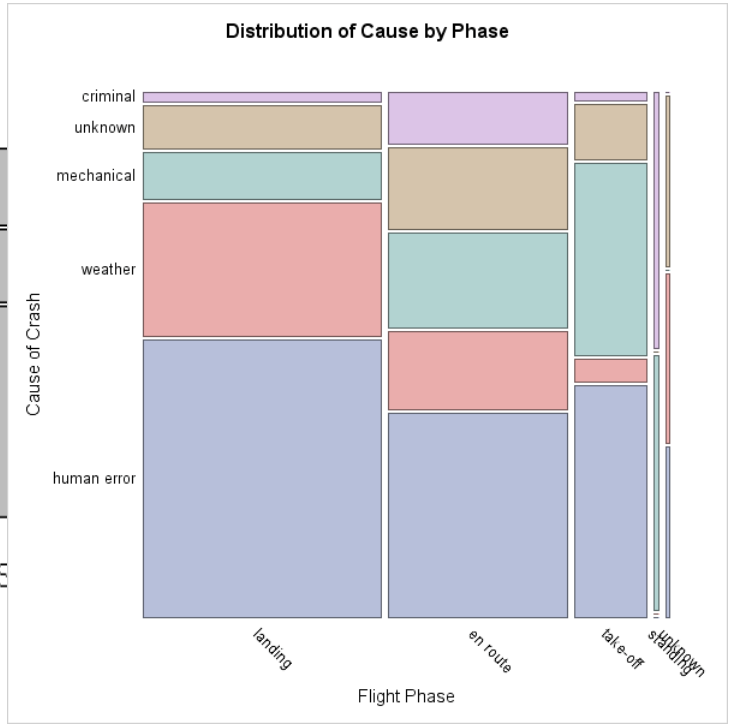
AREA PLOTS: WAFFLE CHART

Data Visualisation

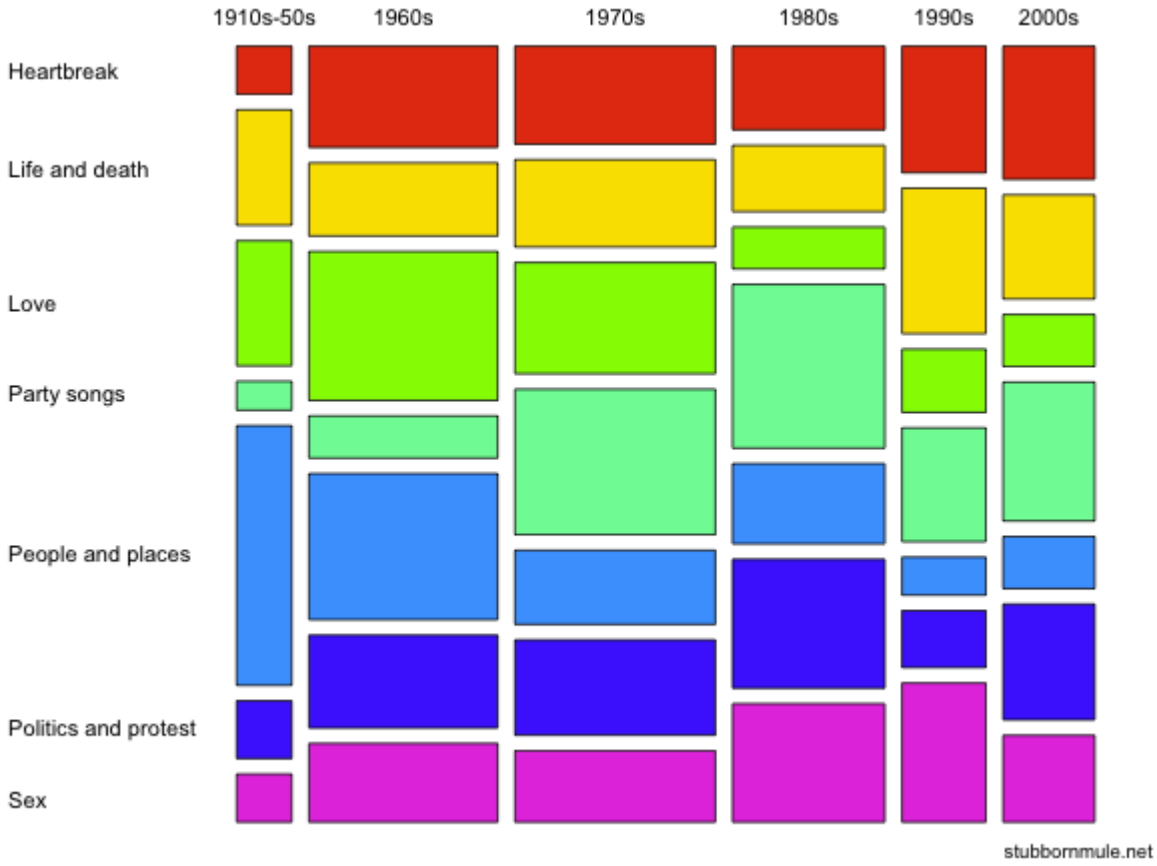
- mosaic plots
  - allows you to observe the relationship among two or more categorical variables



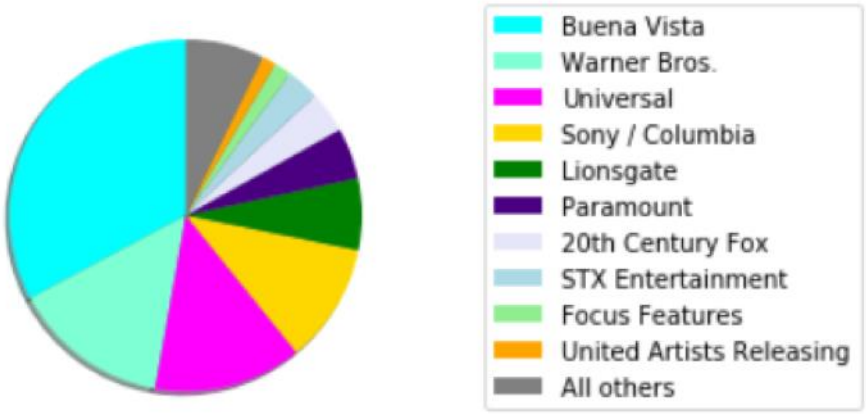


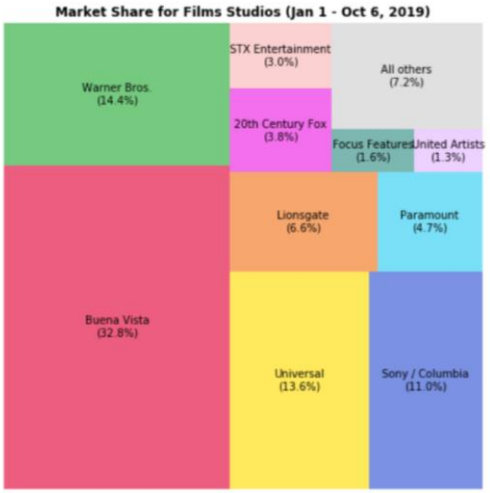


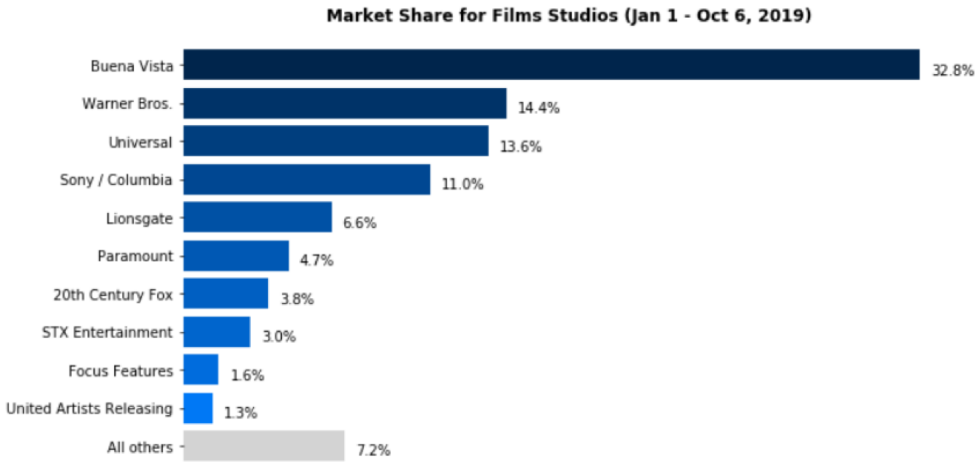
AREA PLOTS: MOSAIC PLOT



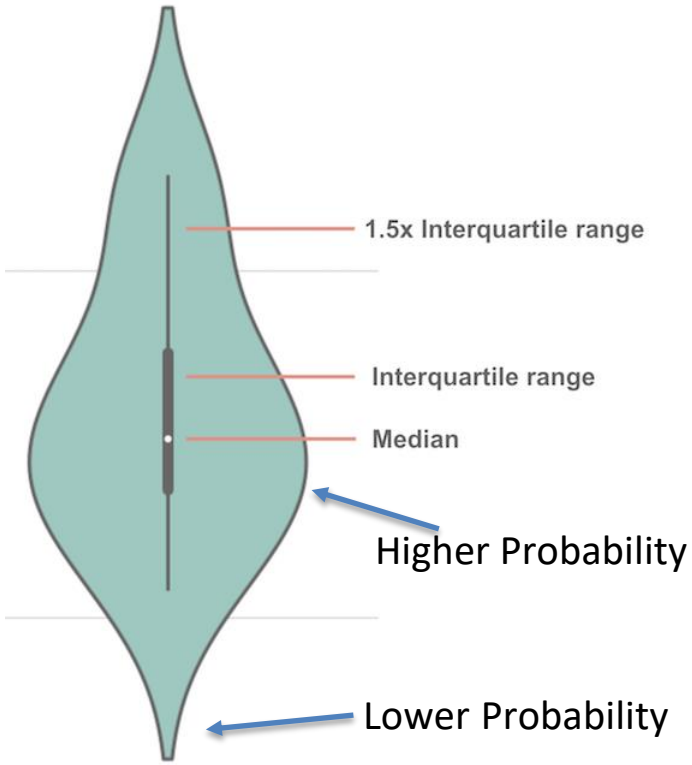
So which visualisation was best?

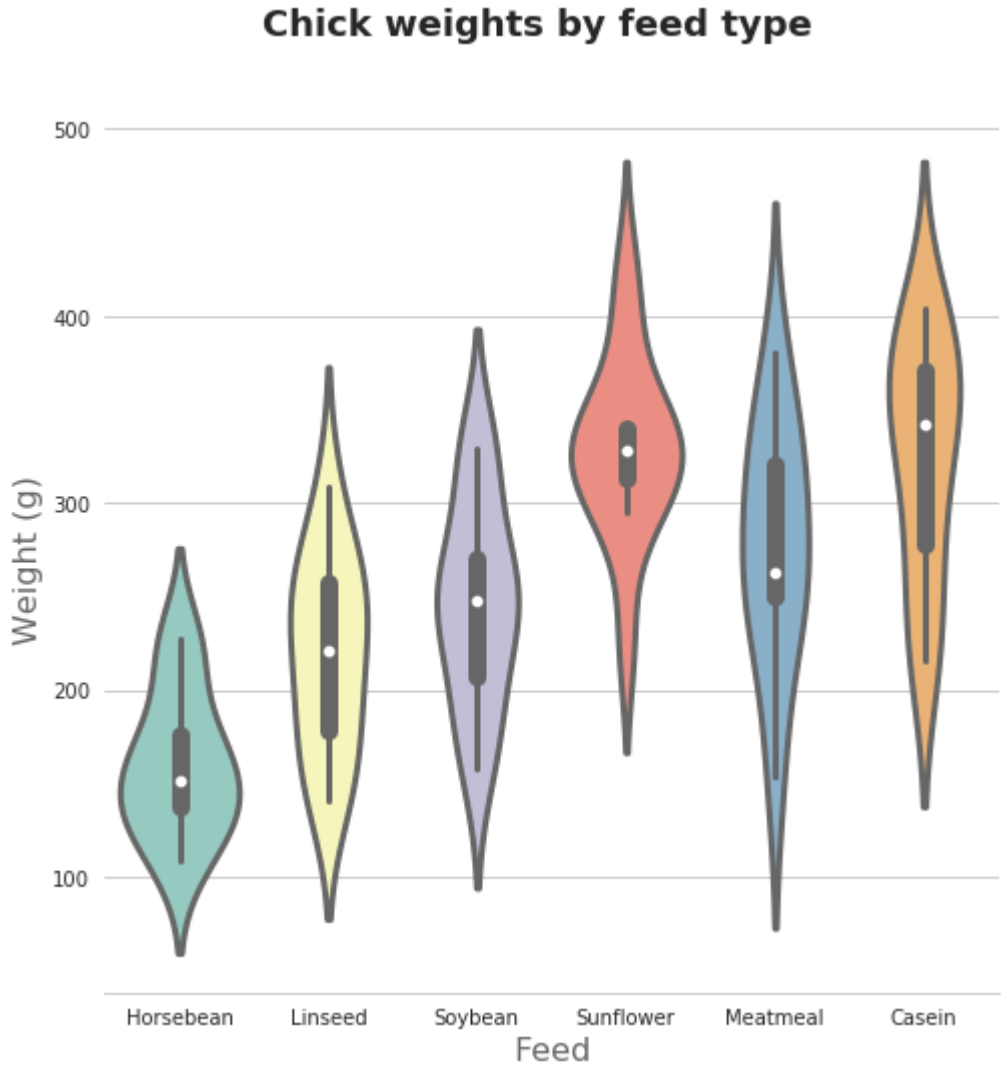




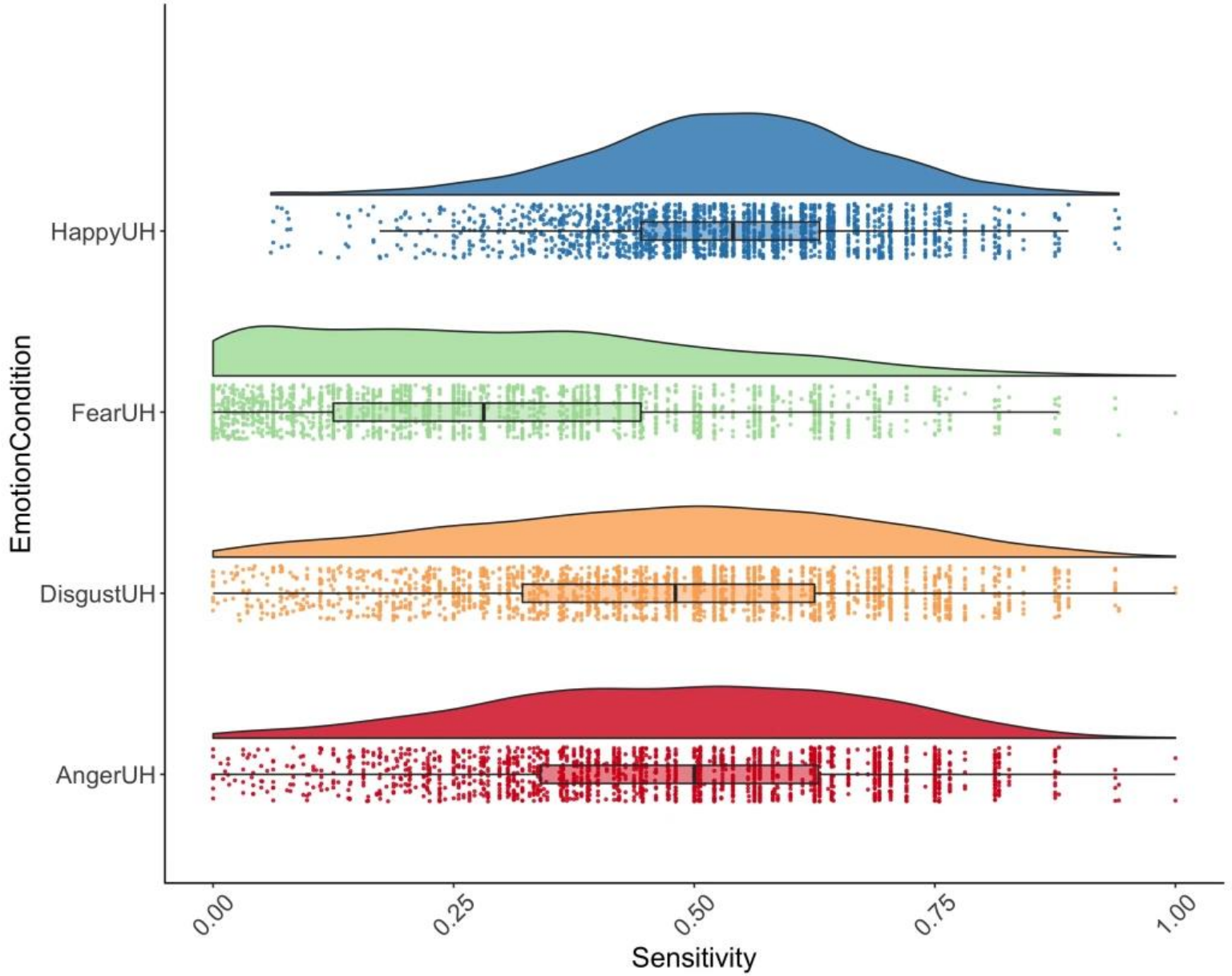


VIOLIN PLOT

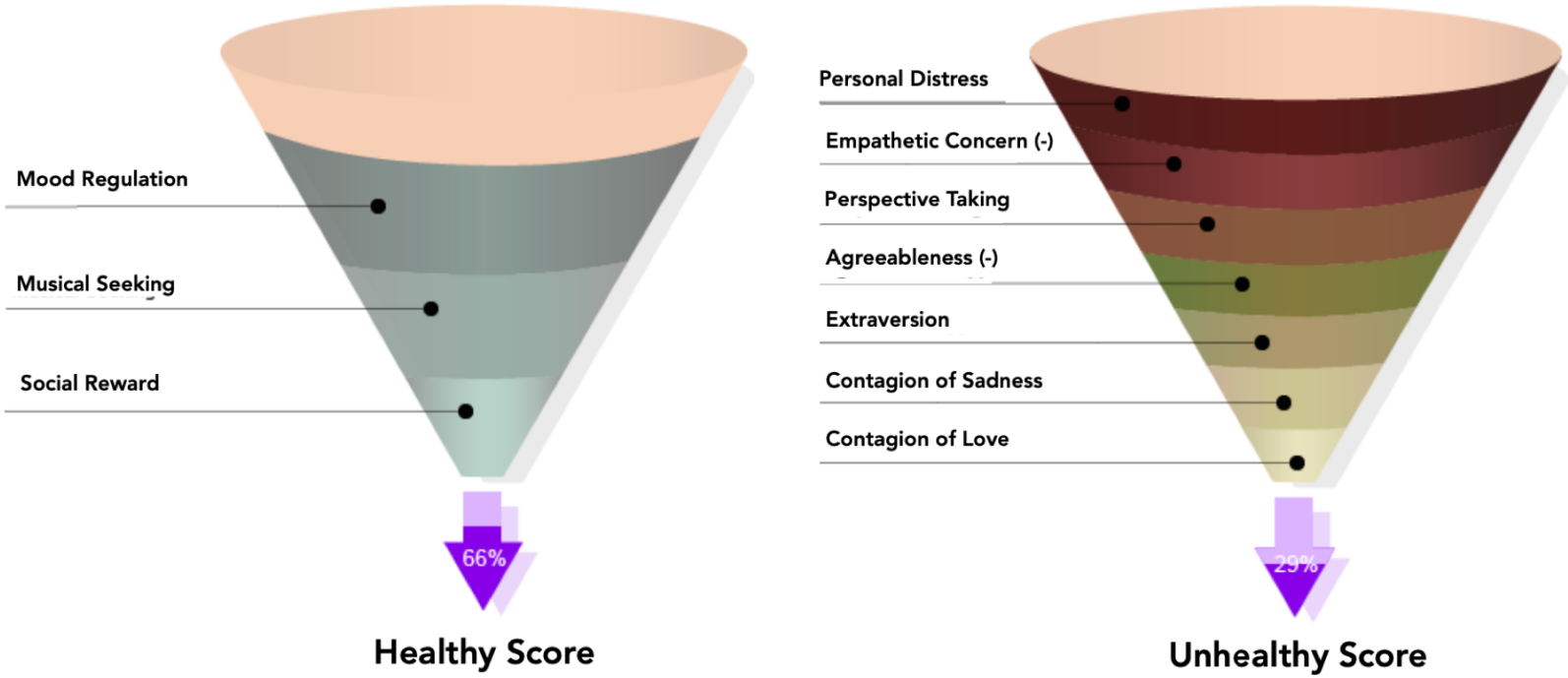

1.5x Interquartile range
Interquartile range
Median
Higher Probability
Lower Probability


Chick weights by feed type
500
400
300
200
100
Weight (g)
Horsebean
Linseed
Soybean
Sunflower
Meatmeal
Casein
Feed

RAINDROP PLOT (Combo)

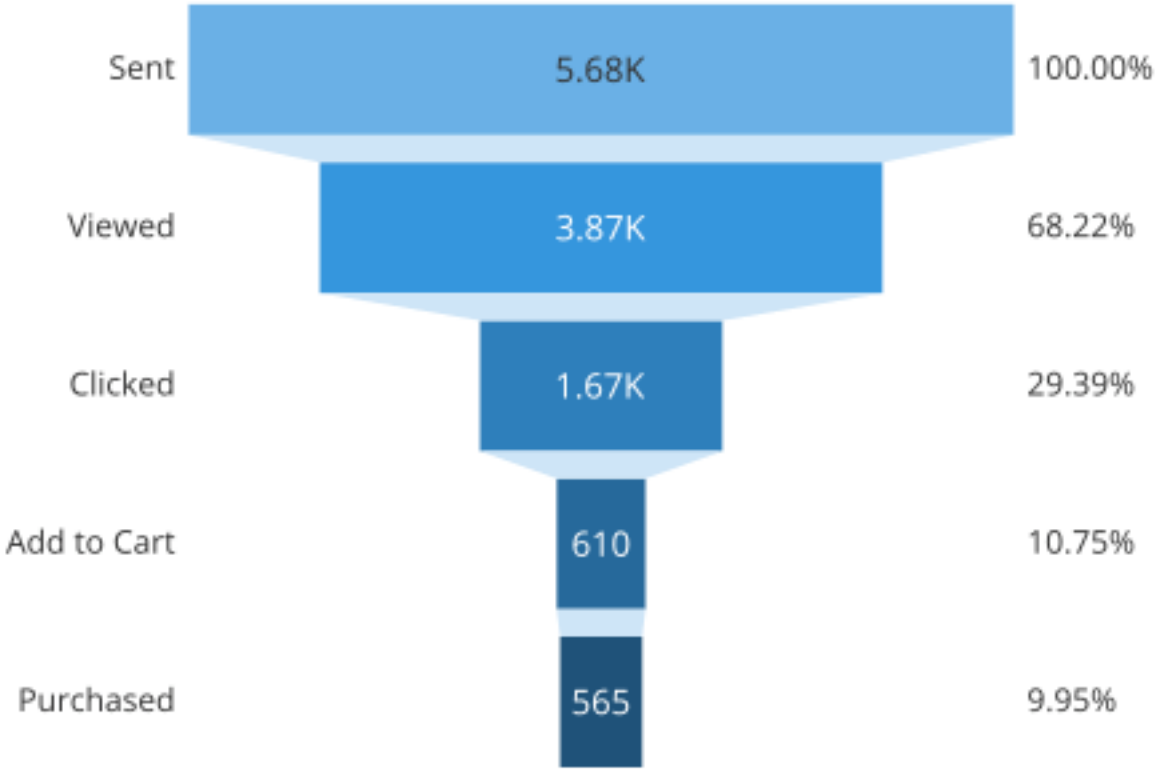


Visualizing Results



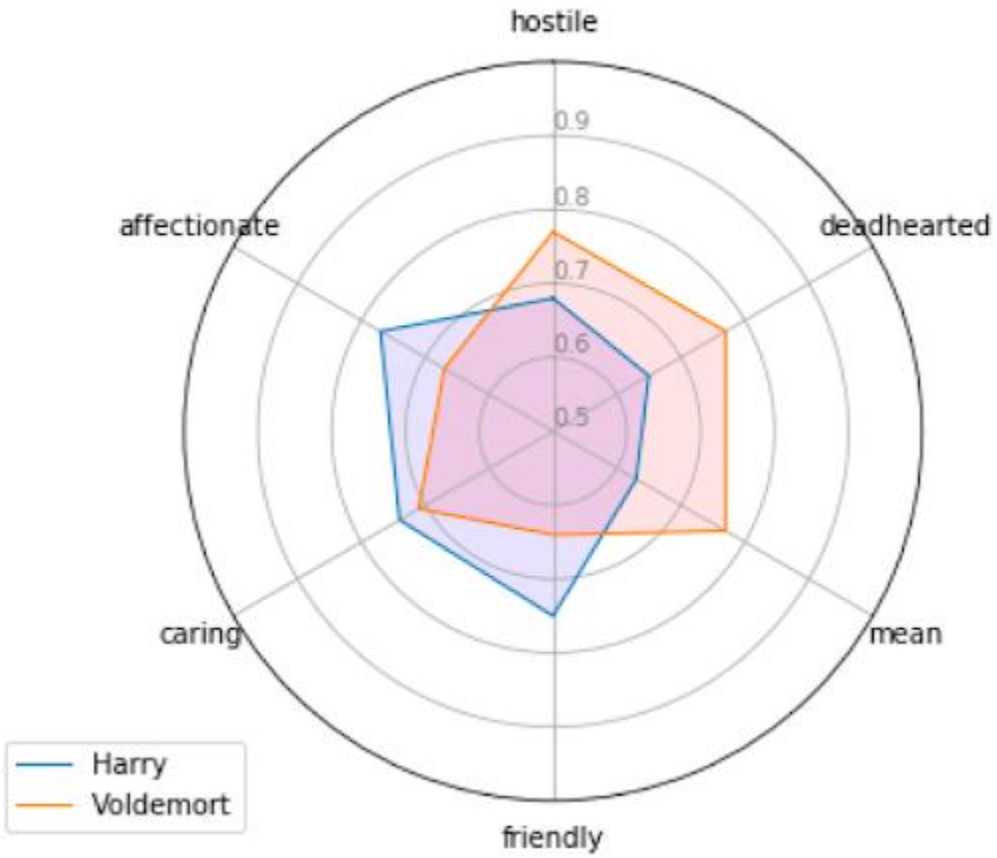
Funnel Plots
(Regression Results)

Describing Data

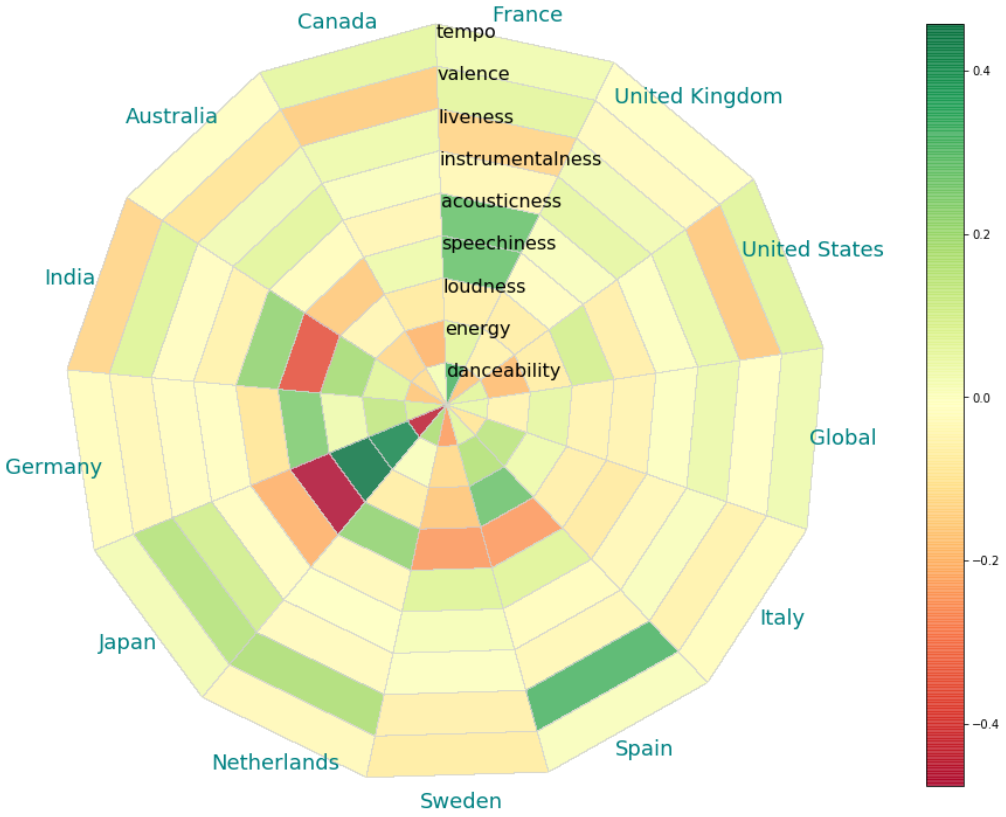


Funnel Plots

SPIDER PLOT / RADAR CHART



Describing Data



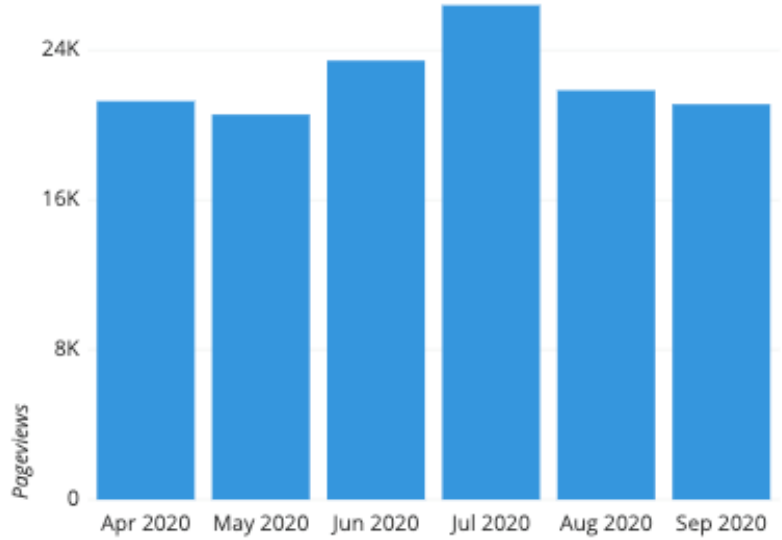
Radial Heat Map

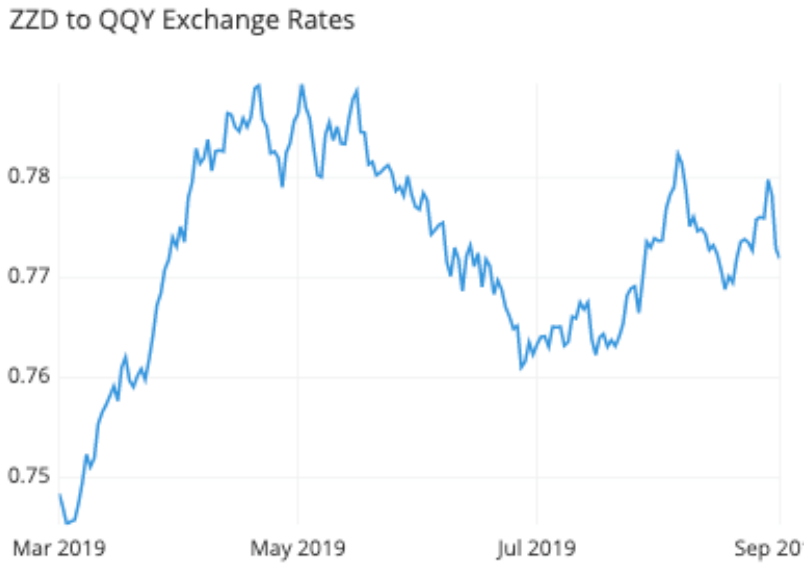
How to choose the right plot?

- temporal changes
- proportions
- data distributions
- group differences
- relationships between variables
- geographical data

Temporal

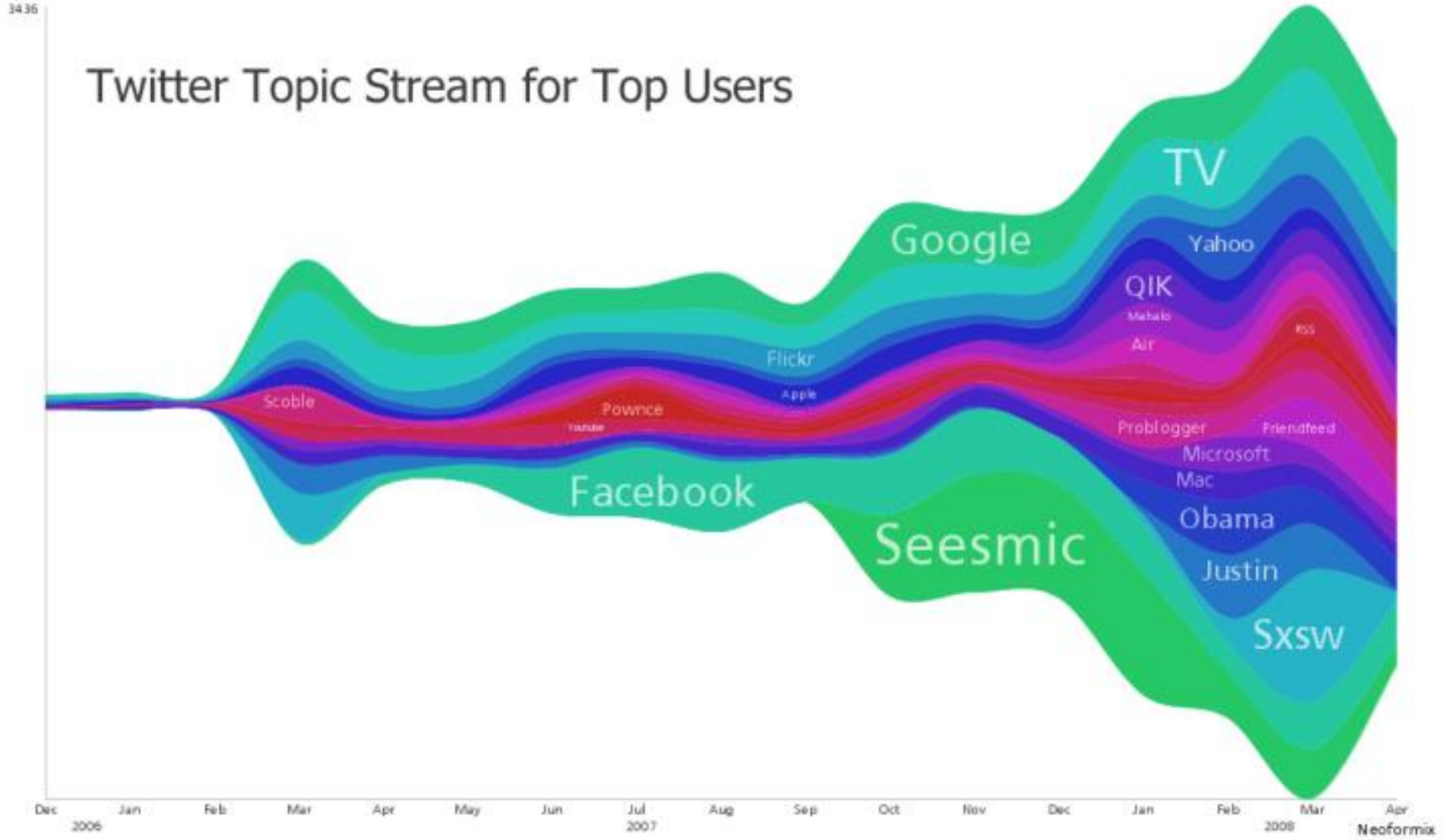
- showing change over time





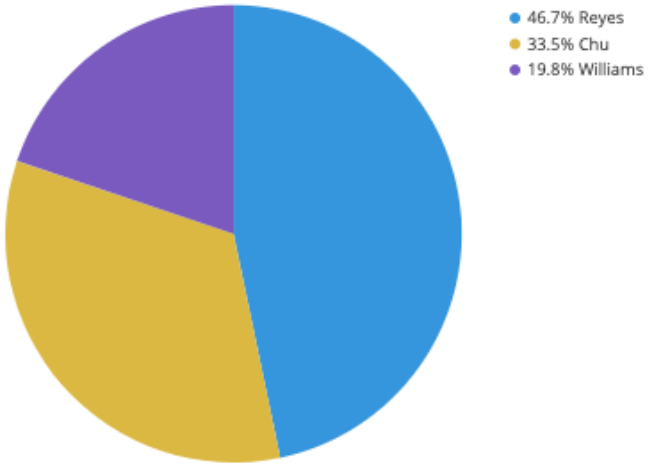
Temporal

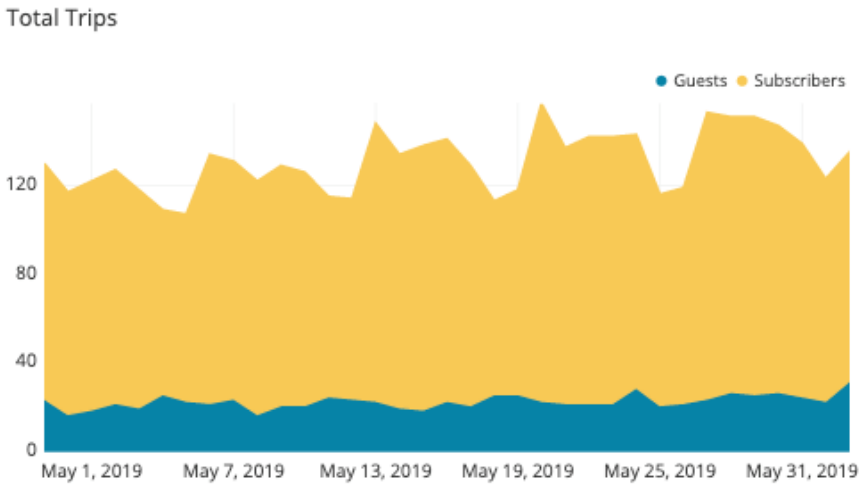
- showing change over time

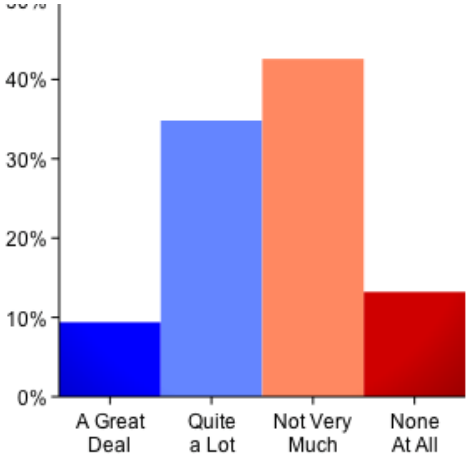


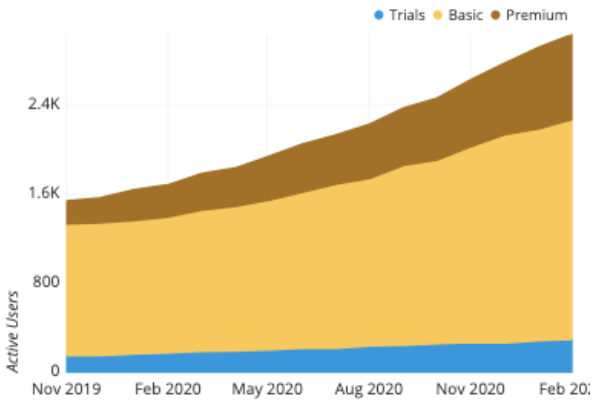
Proportions

- showing a part-to-whole composition

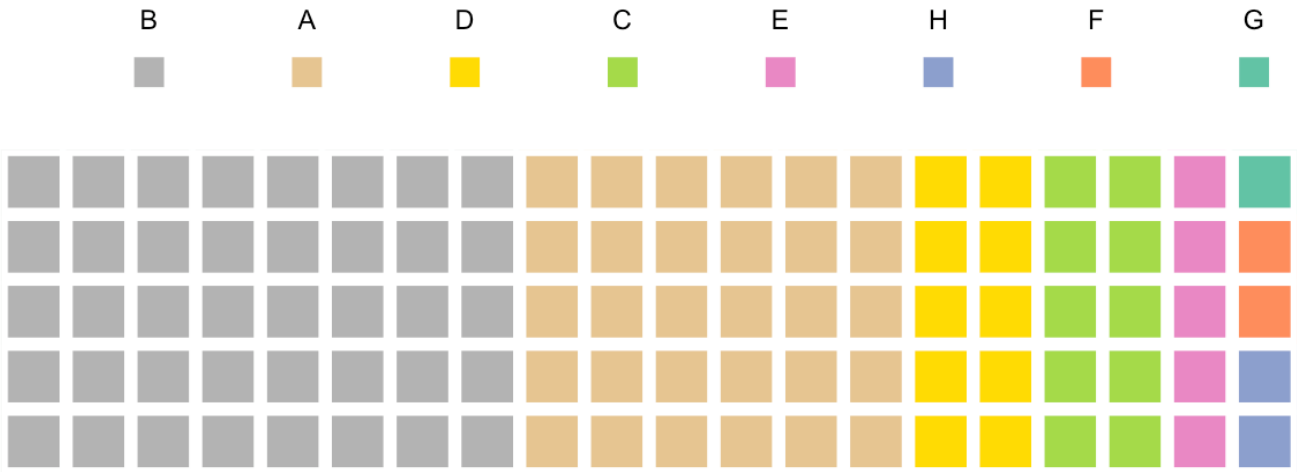


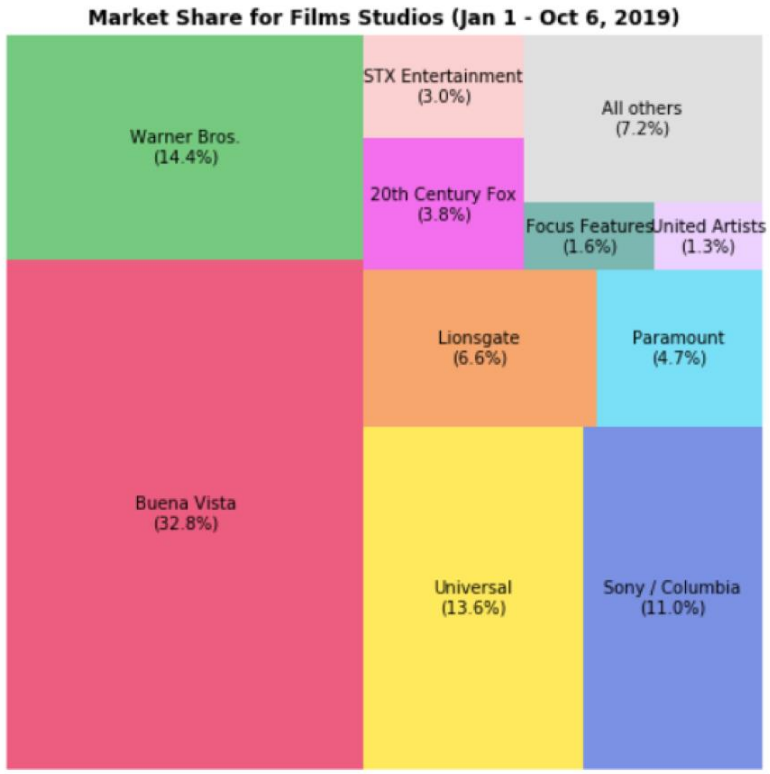






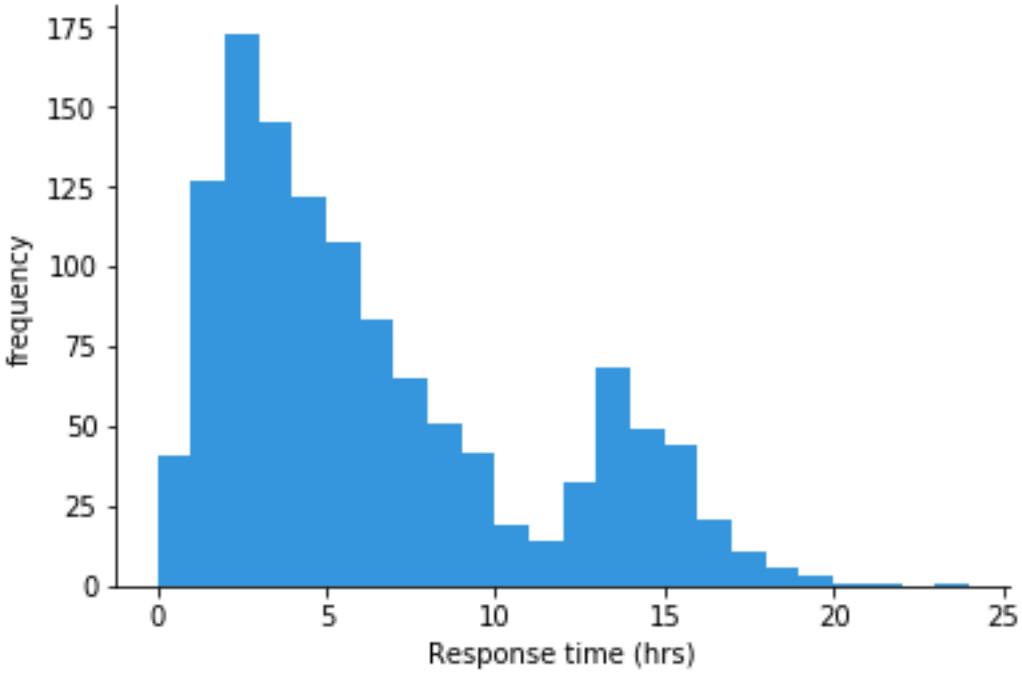
Proportions

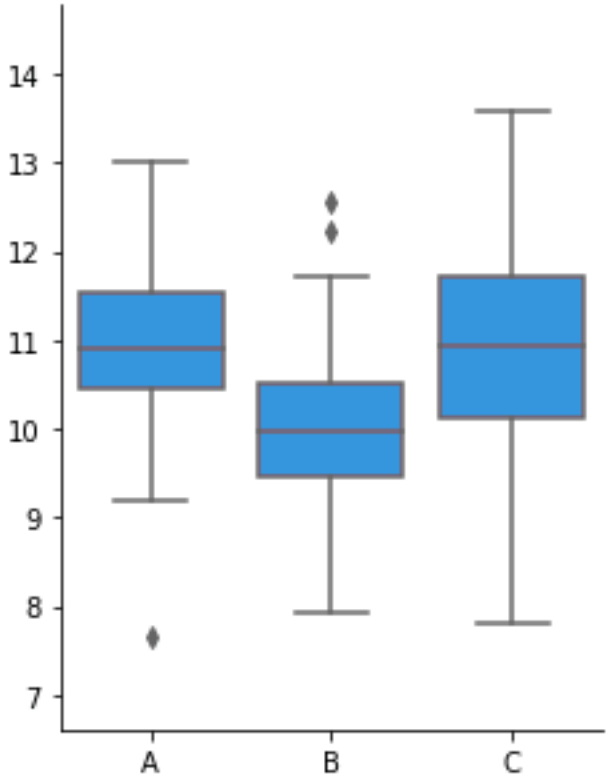




Area plots

Data Distribution

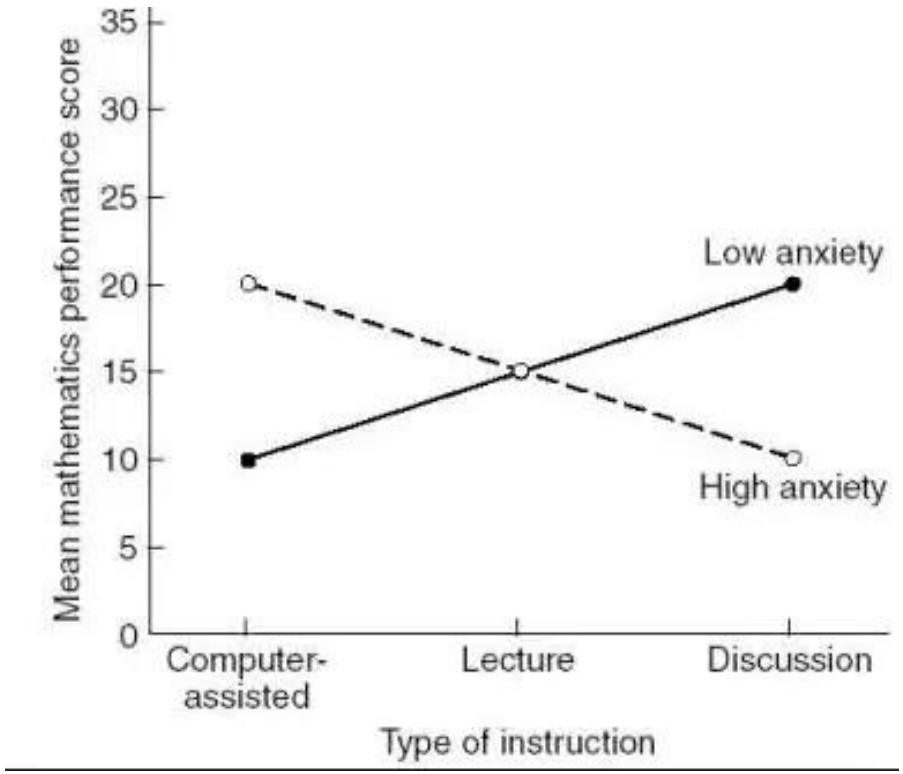




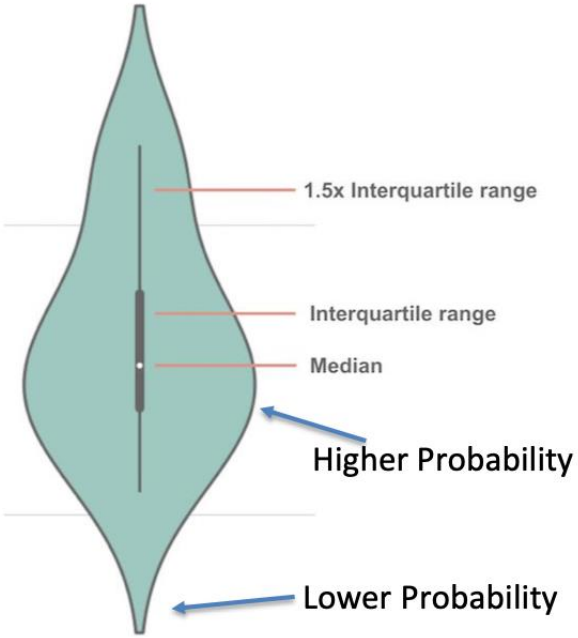
indicative of potential groups or group differences

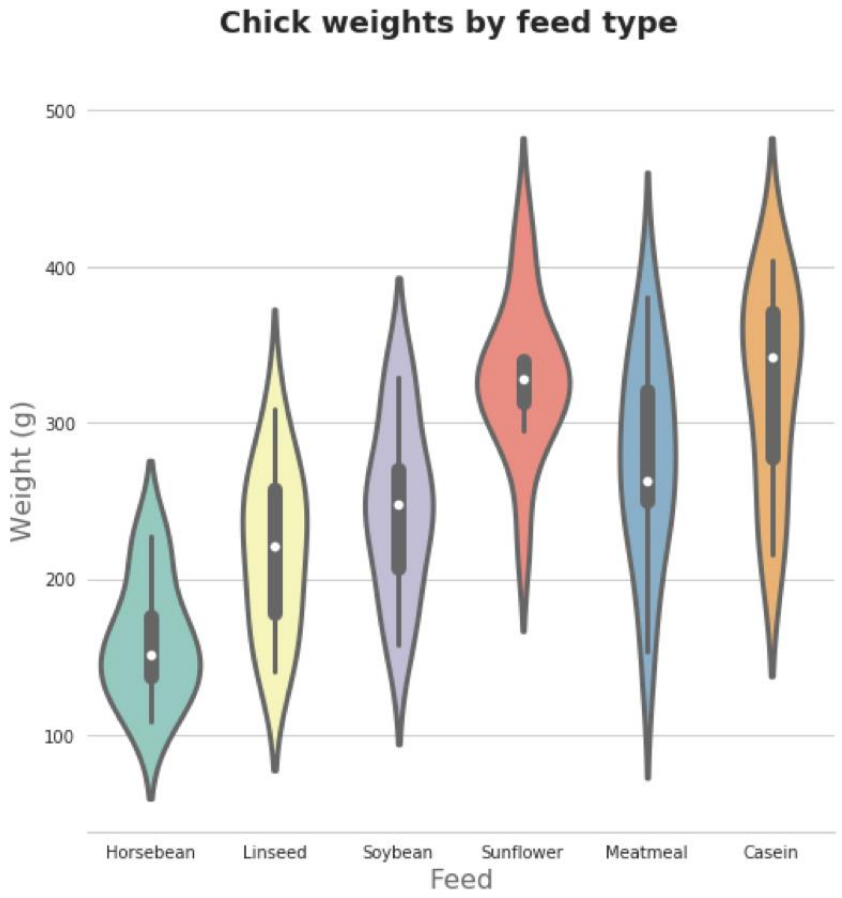
Group Differences

- main effects and interaction plots

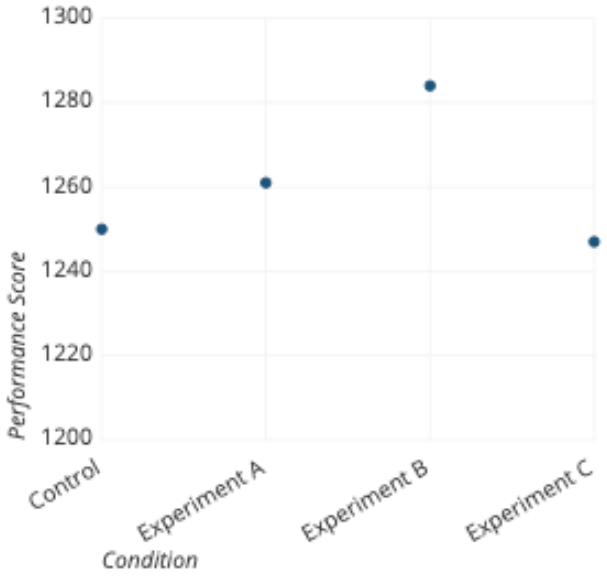


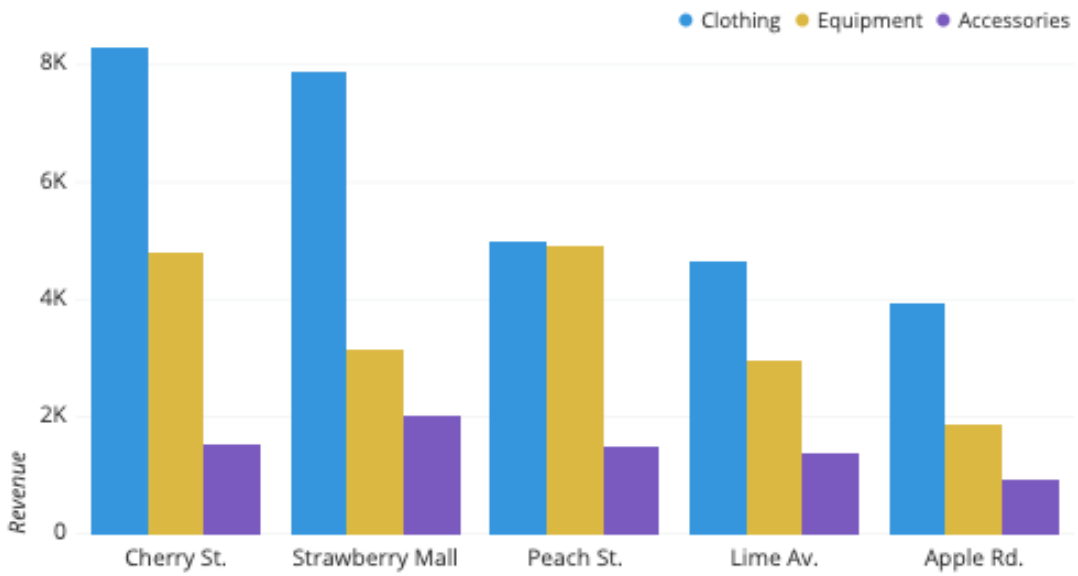
Data Distribution & Group Differences



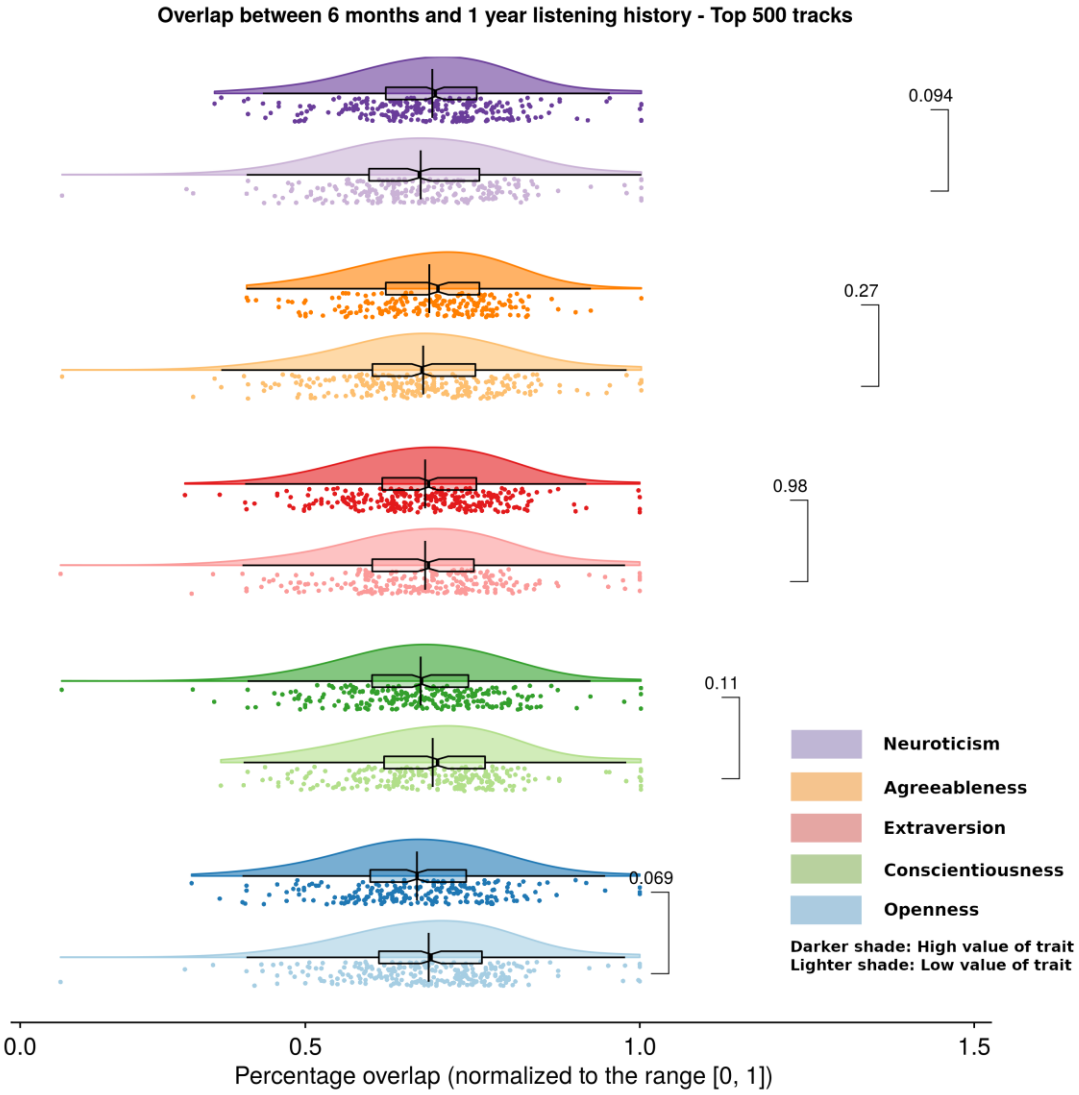


Group differences



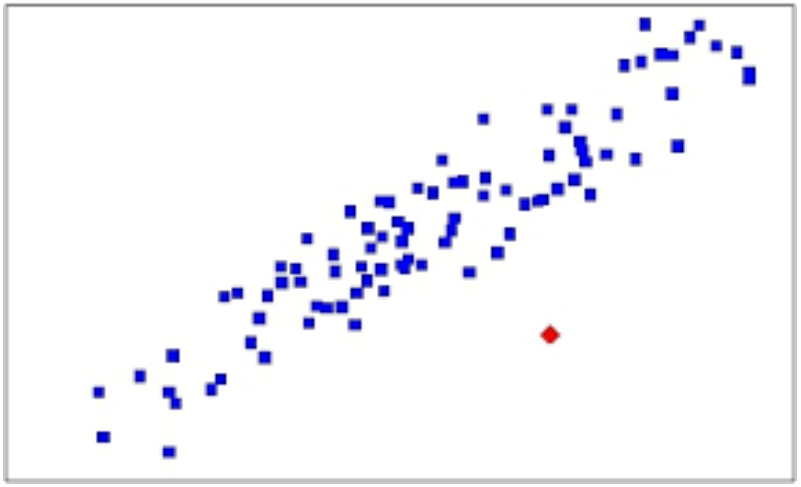


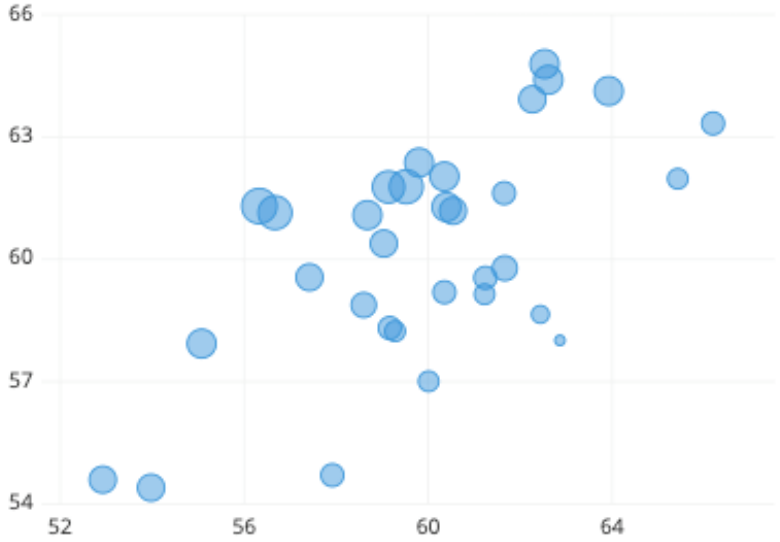
Describing Data + Group Differences



Association between variables

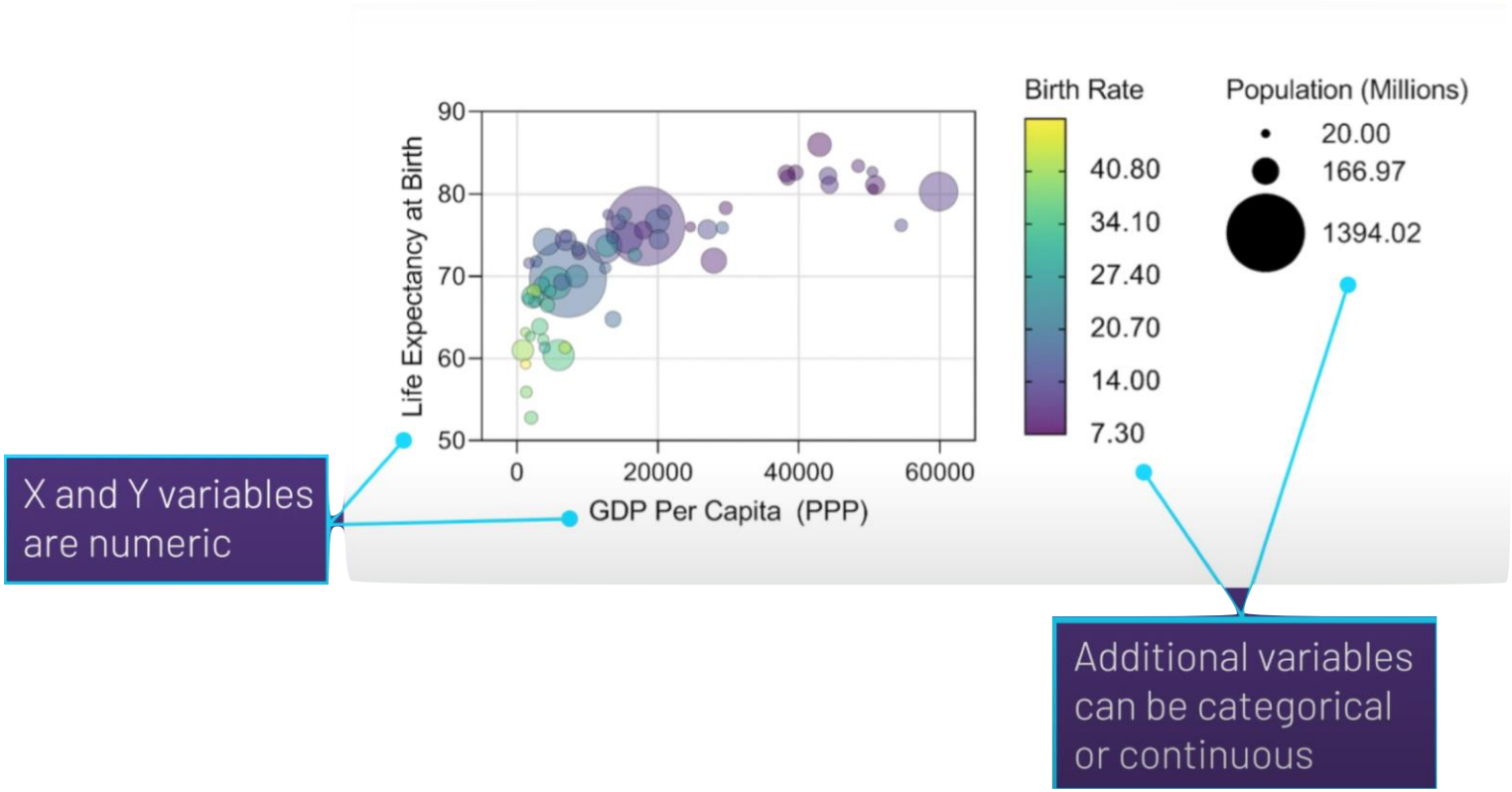
- scatter/bubble plots
  - allows you to observe the relationship between variables



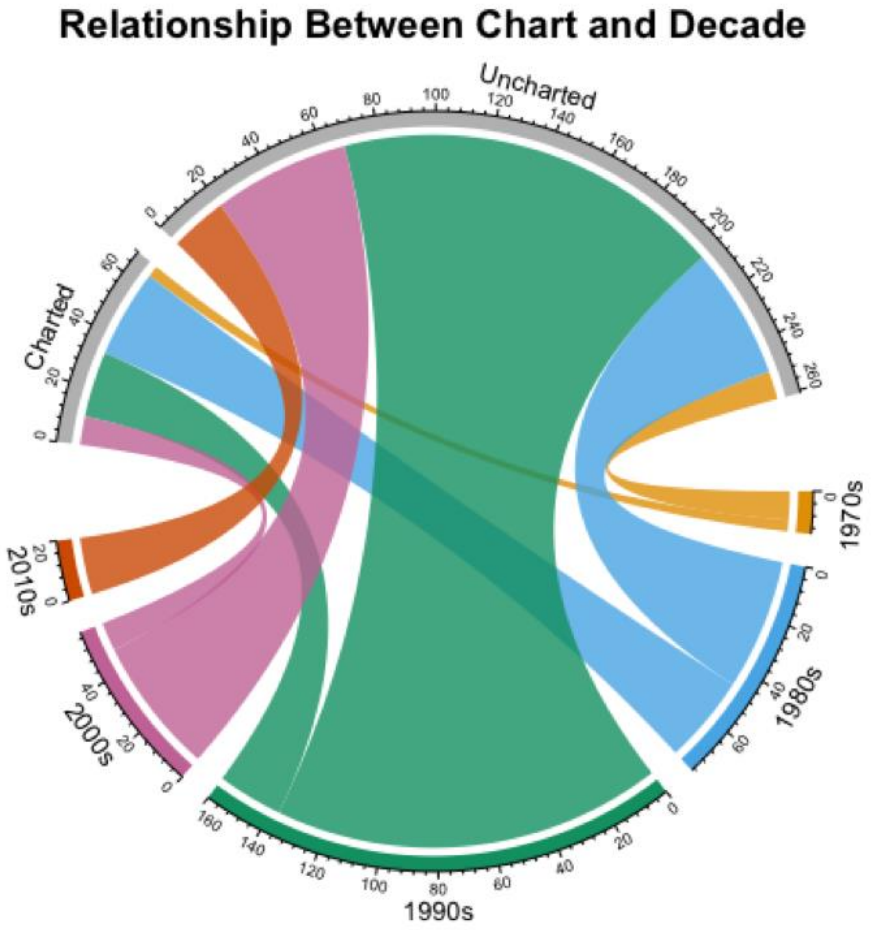


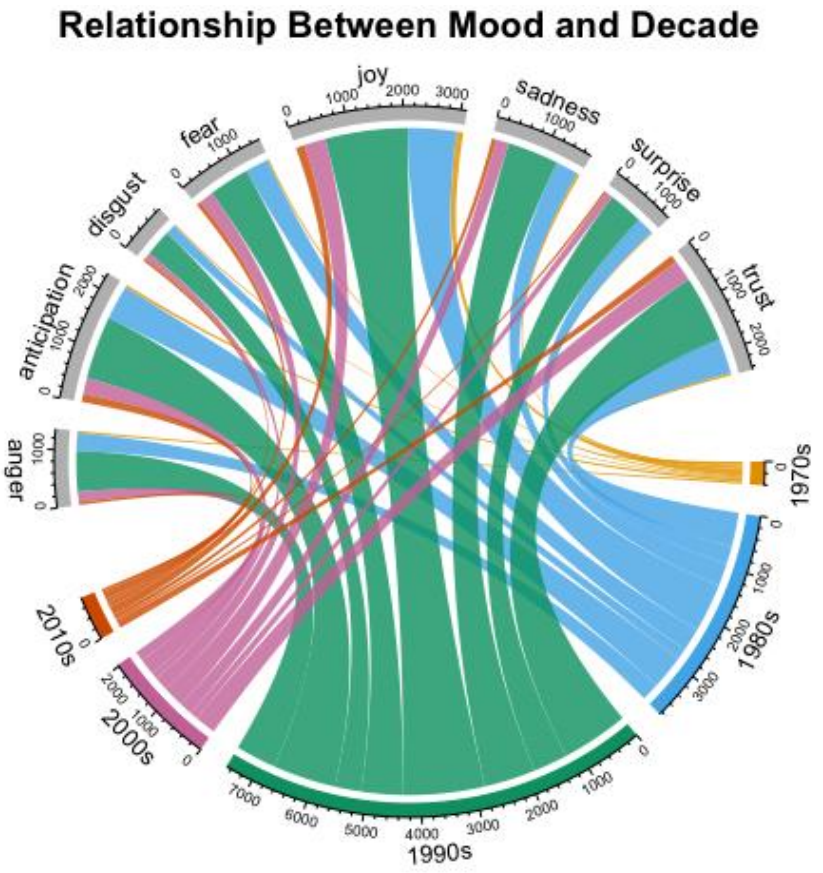
Association between variables

- bubble plots



Association between variables





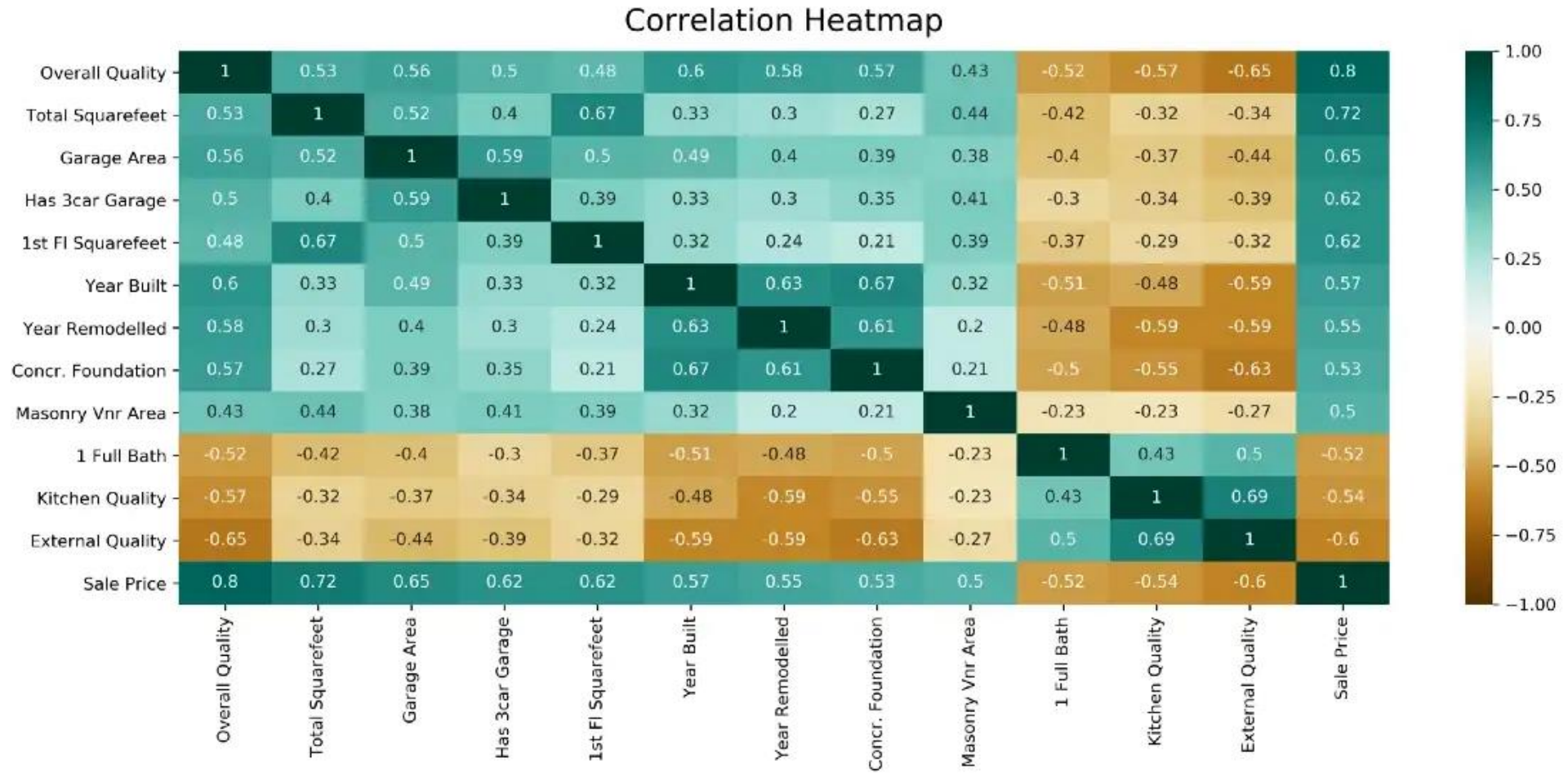
Association between variables

- heat maps depicting correlations

	Overall Qual	Total SF	Garage Area	Garage Cars_3.0	1st Flr SF	Year Built	Year Remod/Add	Foundation_PConc	Mas Vnr Area	Full Bath_1	Kitchen Qual_TA	Exter Qual_TA	SalePrice
Overall Qual	1.000000	0.534259	0.563904	0.502657	0.477136	0.602964	0.584654	0.571092	0.430041	-0.521553	-0.568011	-0.646351	0.800207
Total SF	0.534259	1.000000	0.524145	0.399740	0.668871	0.331811	0.300193	0.270644	0.441001	-0.418993	-0.316613	-0.341000	0.716714
Garage Area	0.563904	0.524145	1.000000	0.589214	0.498690	0.488023	0.397731	0.393544	0.380563	-0.402050	-0.365930	-0.435269	0.649897
Garage Cars_3.0	0.502657	0.399740	0.589214	1.000000	0.391699	0.333050	0.303772	0.349473	0.405799	-0.295060	-0.336226	-0.394001	0.619110
1st Flr SF	0.477136	0.668871	0.498690	0.391699	1.000000	0.323315	0.244190	0.212511	0.386482	-0.369359	-0.293941	-0.318021	0.618486
Year Built	0.602964	0.331811	0.488023	0.333050	0.323315	1.000000	0.629116	0.666546	0.320780	-0.509293	-0.478751	-0.591403	0.571849
Year Remod/Add	0.584654	0.300193	0.397731	0.303772	0.244190	0.629116	1.000000	0.608503	0.204234	-0.483858	-0.585228	-0.590271	0.550370
Foundation_PConc	0.571092	0.270644	0.393544	0.349473	0.212511	0.666546	0.608503	1.000000	0.208299	-0.500180	-0.550170	-0.626157	0.529047
Mas Vnr Area	0.430041	0.441001	0.380563	0.405799	0.386482	0.320780	0.204234	0.208299	1.000000	-0.229672	-0.226351	-0.269285	0.503579
Full Bath_1	-0.521553	-0.418993	-0.402050	-0.295060	-0.369359	-0.509293	-0.483858	-0.500180	-0.229672	1.000000	0.425653	0.496703	-0.520016
Kitchen Qual_TA	-0.568011	-0.316613	-0.365930	-0.336226	-0.293941	-0.478751	-0.585228	-0.550170	-0.226351	0.425653	1.000000	0.690116	-0.540860
Exter Qual_TA	-0.646351	-0.341000	-0.435269	-0.394001	-0.318021	-0.591403	-0.590271	-0.626157	-0.269285	0.496703	0.690116	1.000000	-0.600362
SalePrice	0.800207	0.716714	0.649897	0.619110	0.618486	0.571849	0.550370	0.529047	0.503579	-0.520016	-0.540860	-0.600362	1.000000

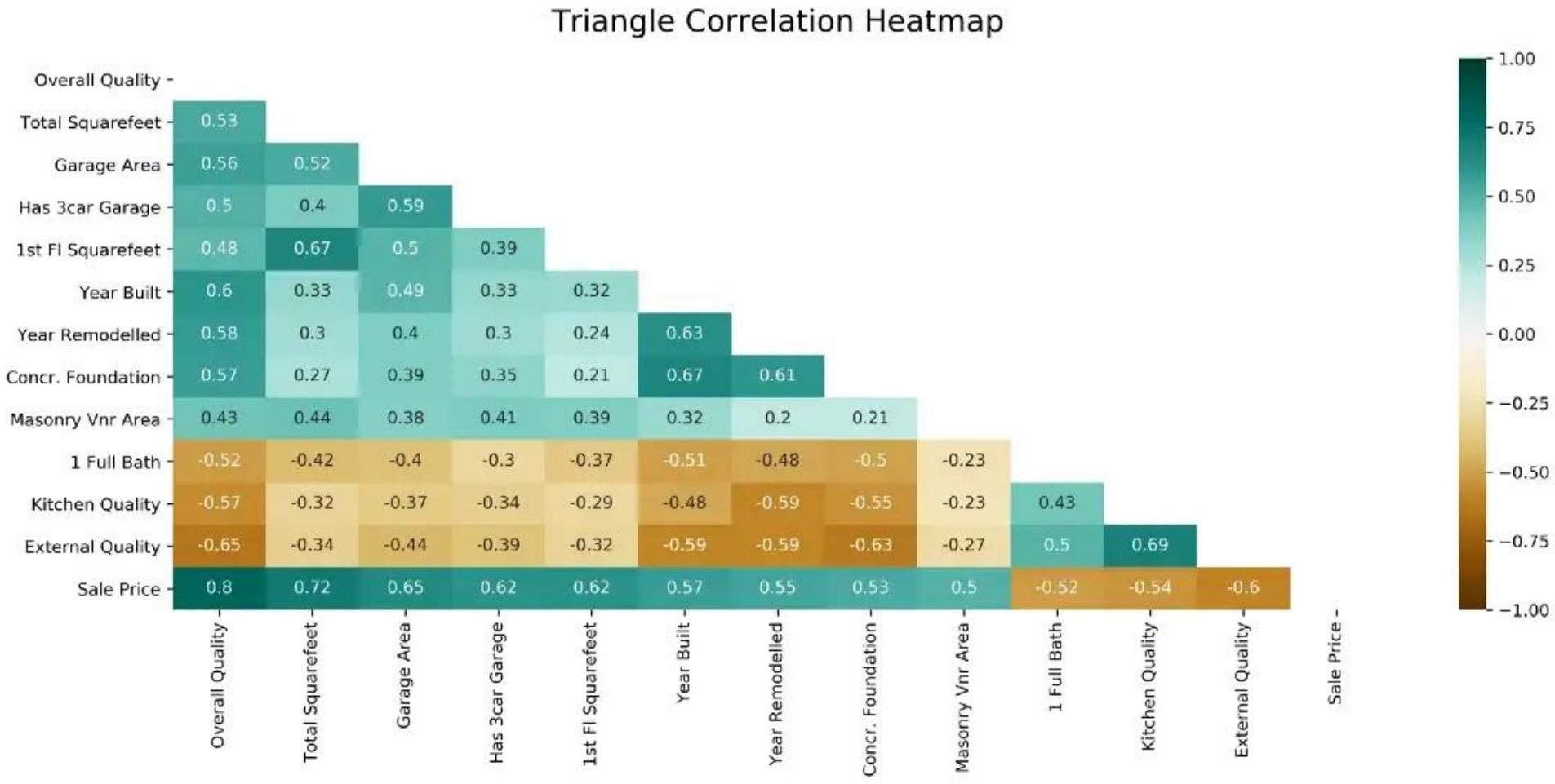
Association between variables

- heat maps depicting correlations

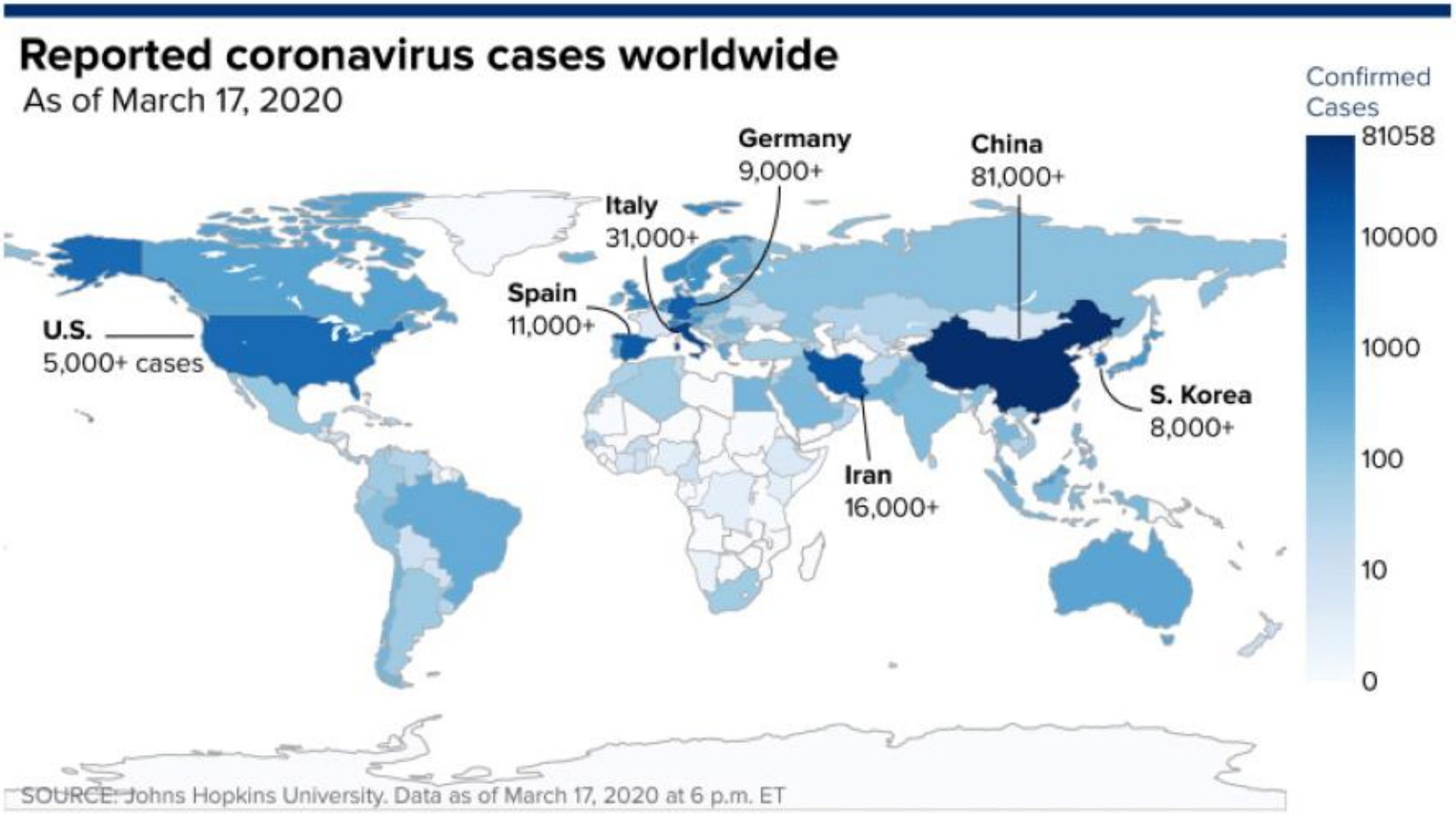


Association between variables

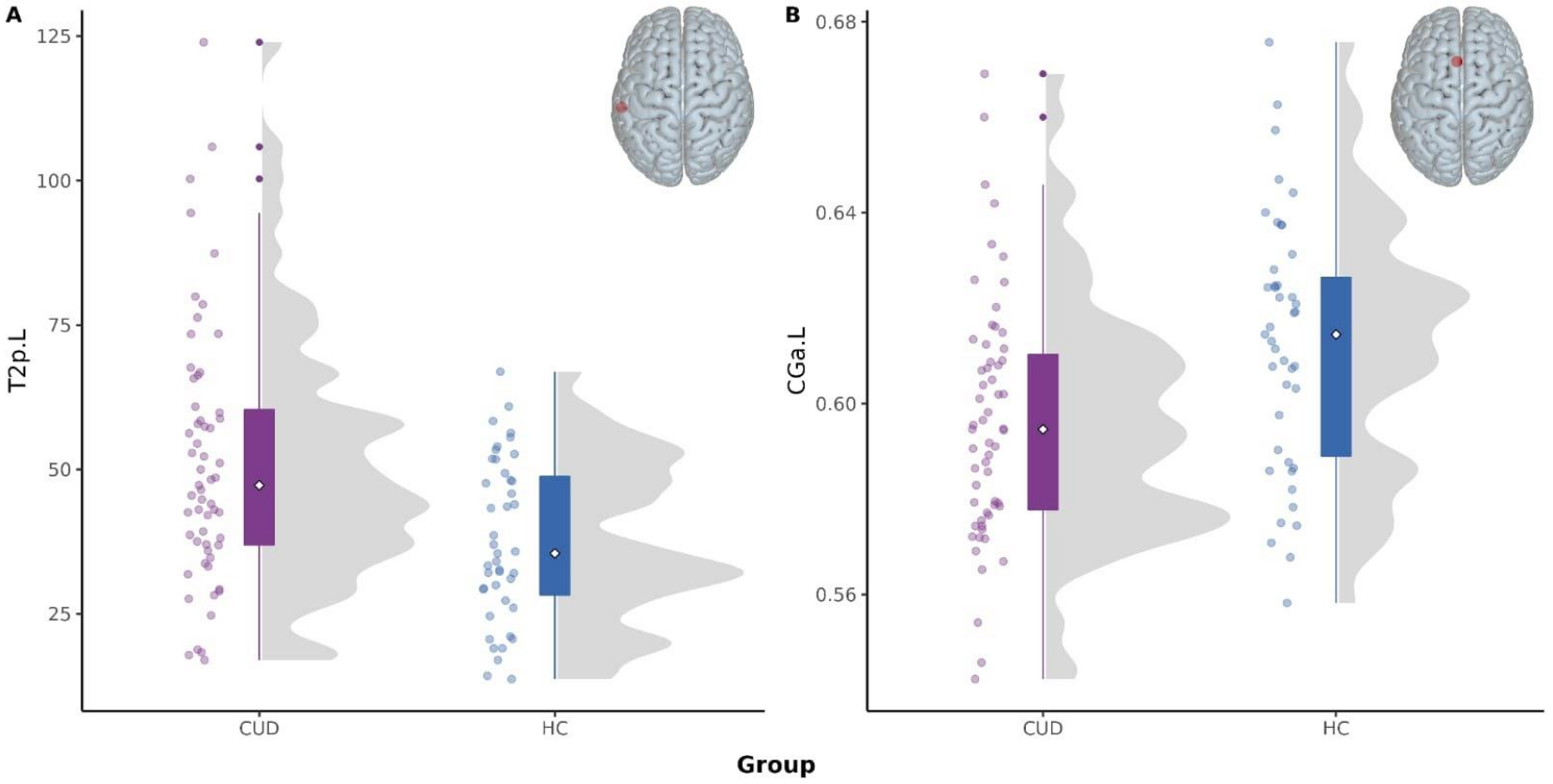
- heat maps depicting correlations



Geographical maps

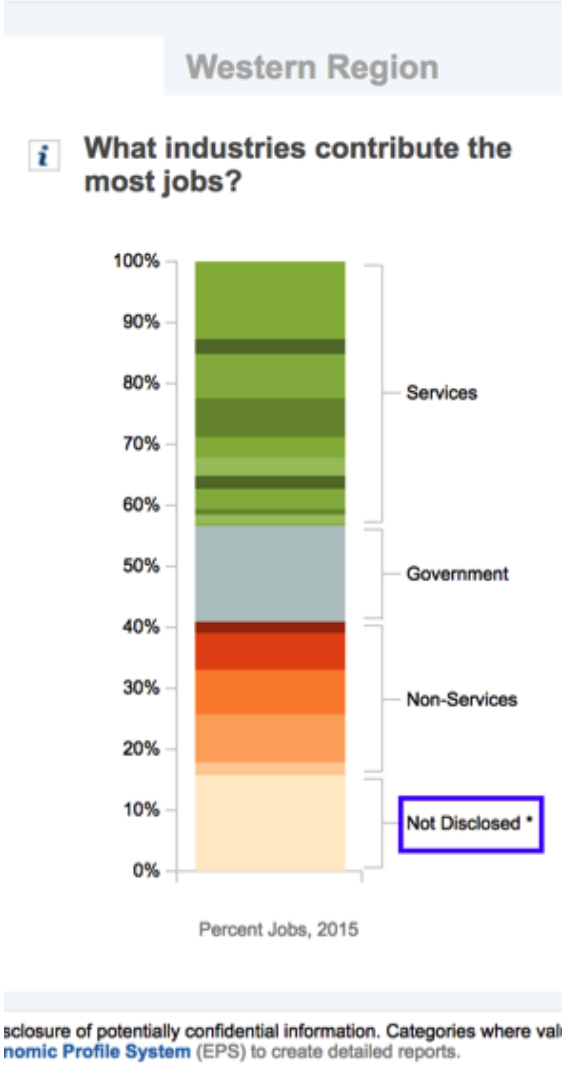


Creative Combinations



What makes a good visualisation

- Storytelling
- Reduce Cognitive Load
- Less is more
- Missing data
- Color Consistency
- Labelling



https://www.cardinalpath.com/blog/makes-good-visualization

To do or not to do

- Provide necessary Context around Visuals
- Ensure Simplicity and Clarity of Information
- Ensure Brevity and Avoid Unnecessary Information
- Use Simple and Easy to Understand Color Palettes
- Pay attention to Graphics in order to make sure that they are Visually Appealing
- Where possible, bring in Originality by relating, seemingly Unrelated data and subjects

To do or **not to do**

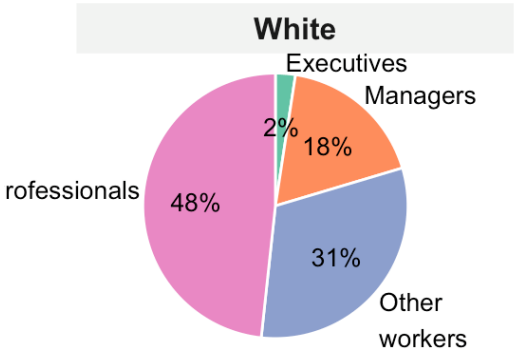
- Avoid using Too Many Variables within a single image which might result in distracting the viewers
- Be extremely careful of not visualizing data through an Unsuitable or Incorrect visualization format
- While using Scales in Data Visualization in order to depict differences between data points, it is important to ensure that the scale is consistent
- Poor Choice of Colors is another significant issue which should be avoided at all costs. Thus, it is important to:
  - avoid using colors with negligible contrast
  - avoid using too many colors
  - avoid using conventional colors to convey opposite meanings
  - pay heed to the needs of people who might be colorblind (check also in grayscale)

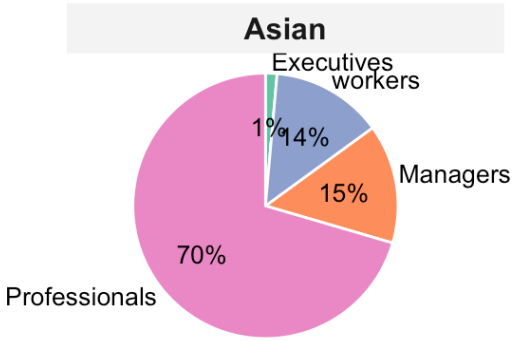
Outline

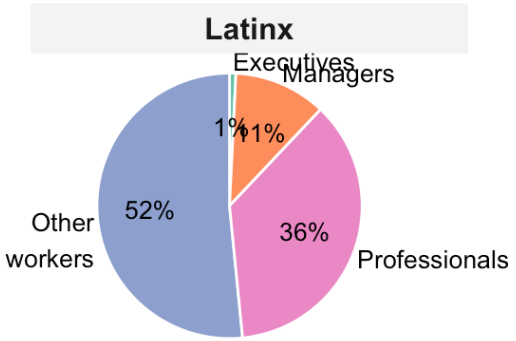
- **Visualization**
  - why we visualise
  - how to pick a plot
  - initial data vs final results visualization (some examples)
  - **bad designs and misleading graphs**
- **Summarization**
  - measures of central tendency & dispersion
  - which measure to pick

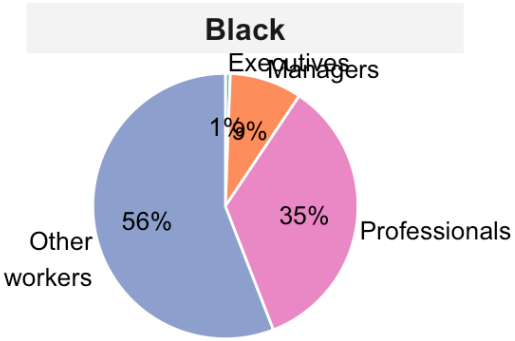
Bad Designs & Improvements

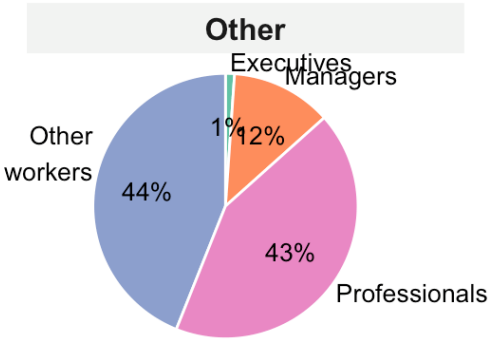
https://nandeshwar.info/data-visualization/pie-chart-vs-bar-chart/





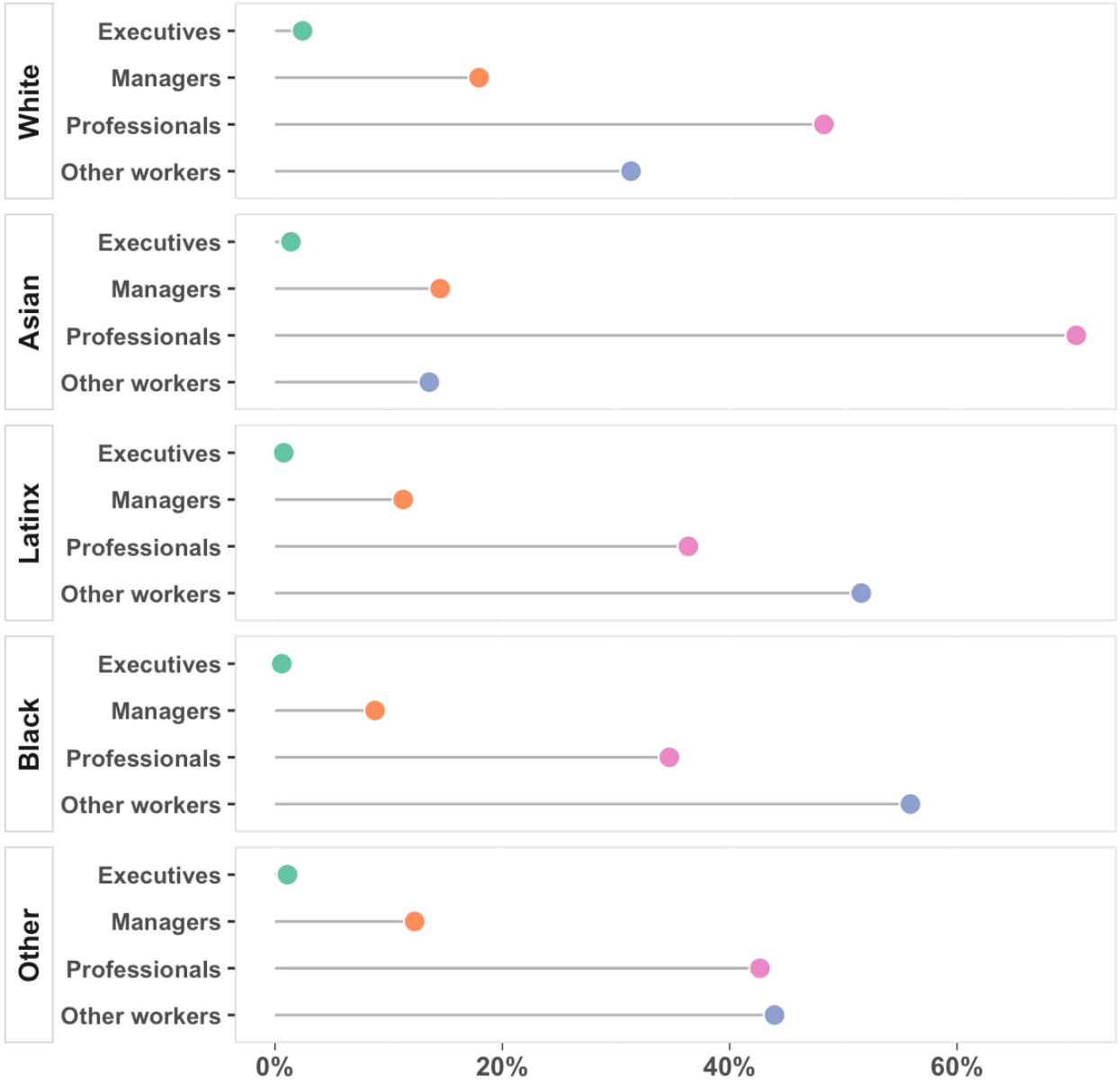




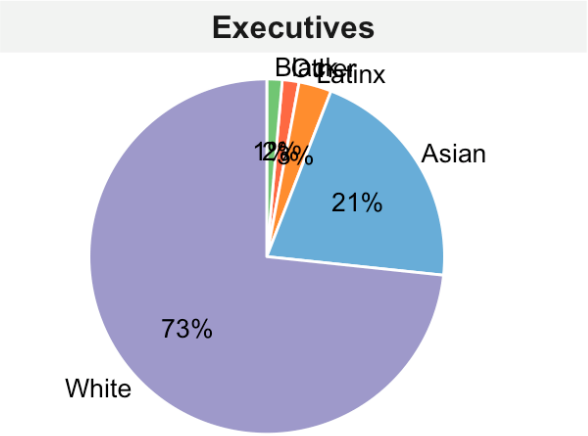


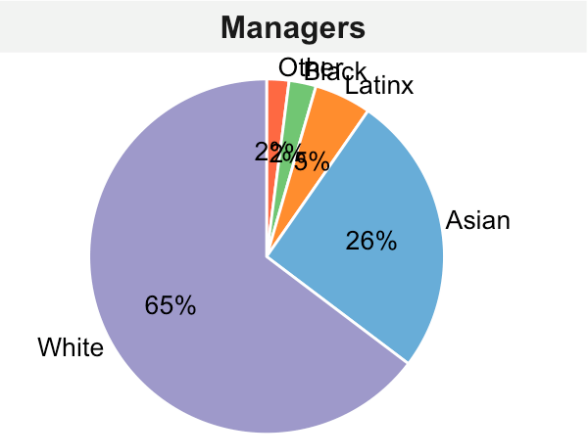
Source: Reveal, https://www.revealnews.org/topic/silicon-valley-diversity/

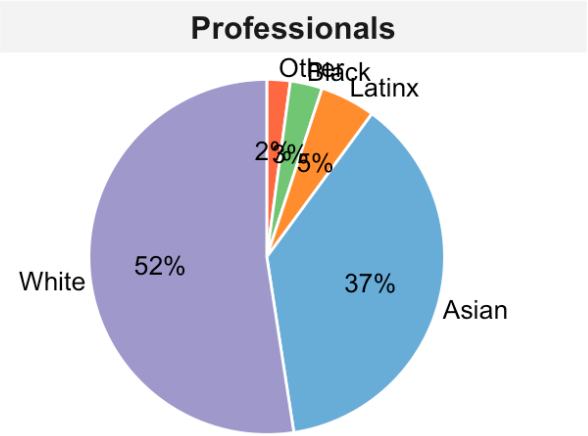
Source: Reveal, https://www.revealnews.org/topic/silicon-valley-diversity/

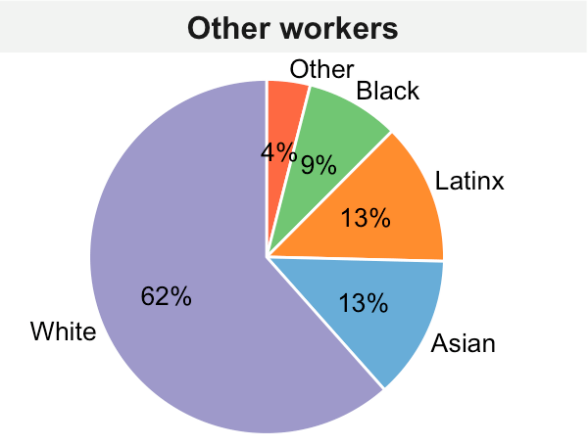


Source: Reveal, https://www.revealnews.org/topic/silicon-valley-diversity/

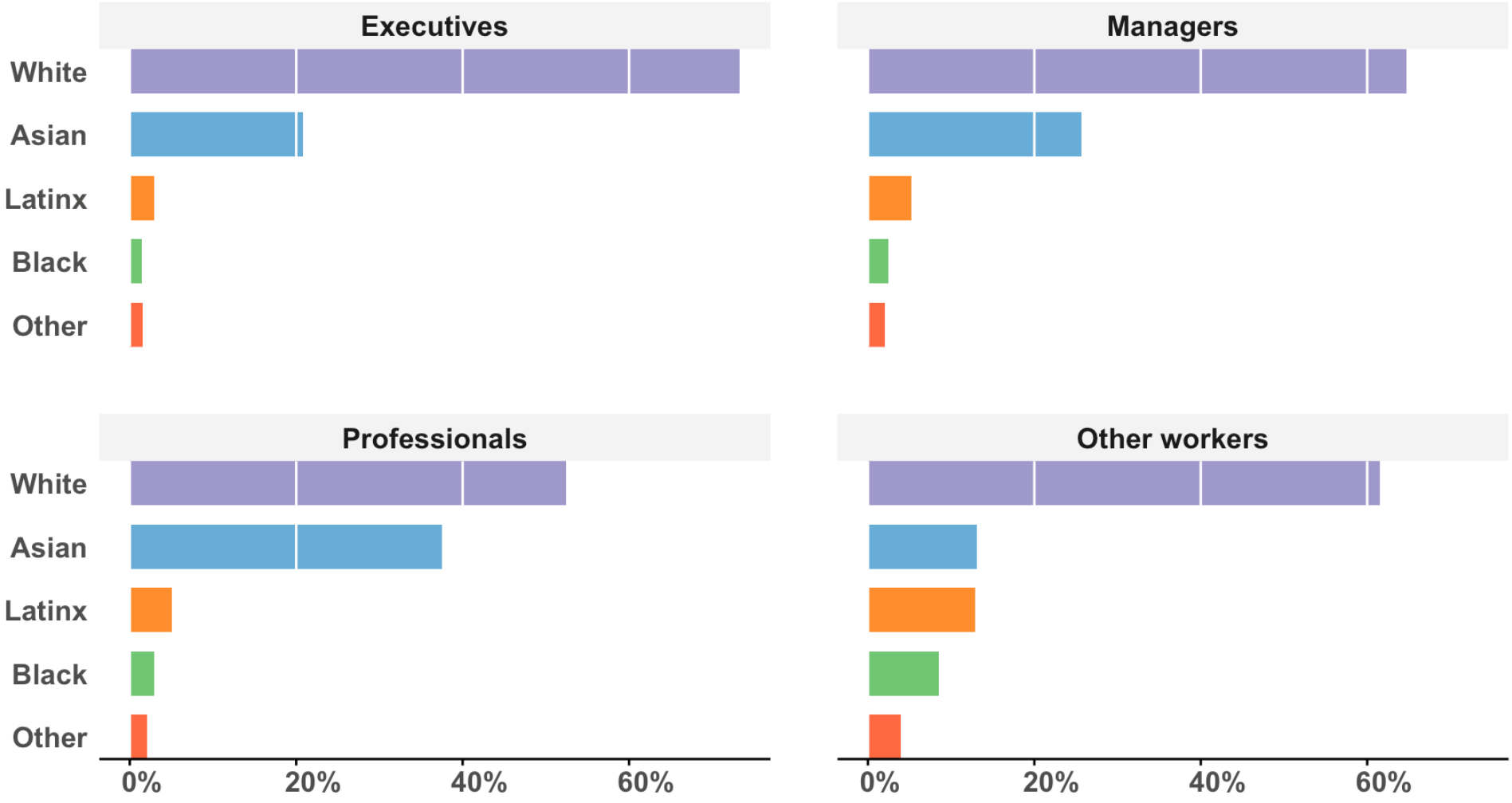




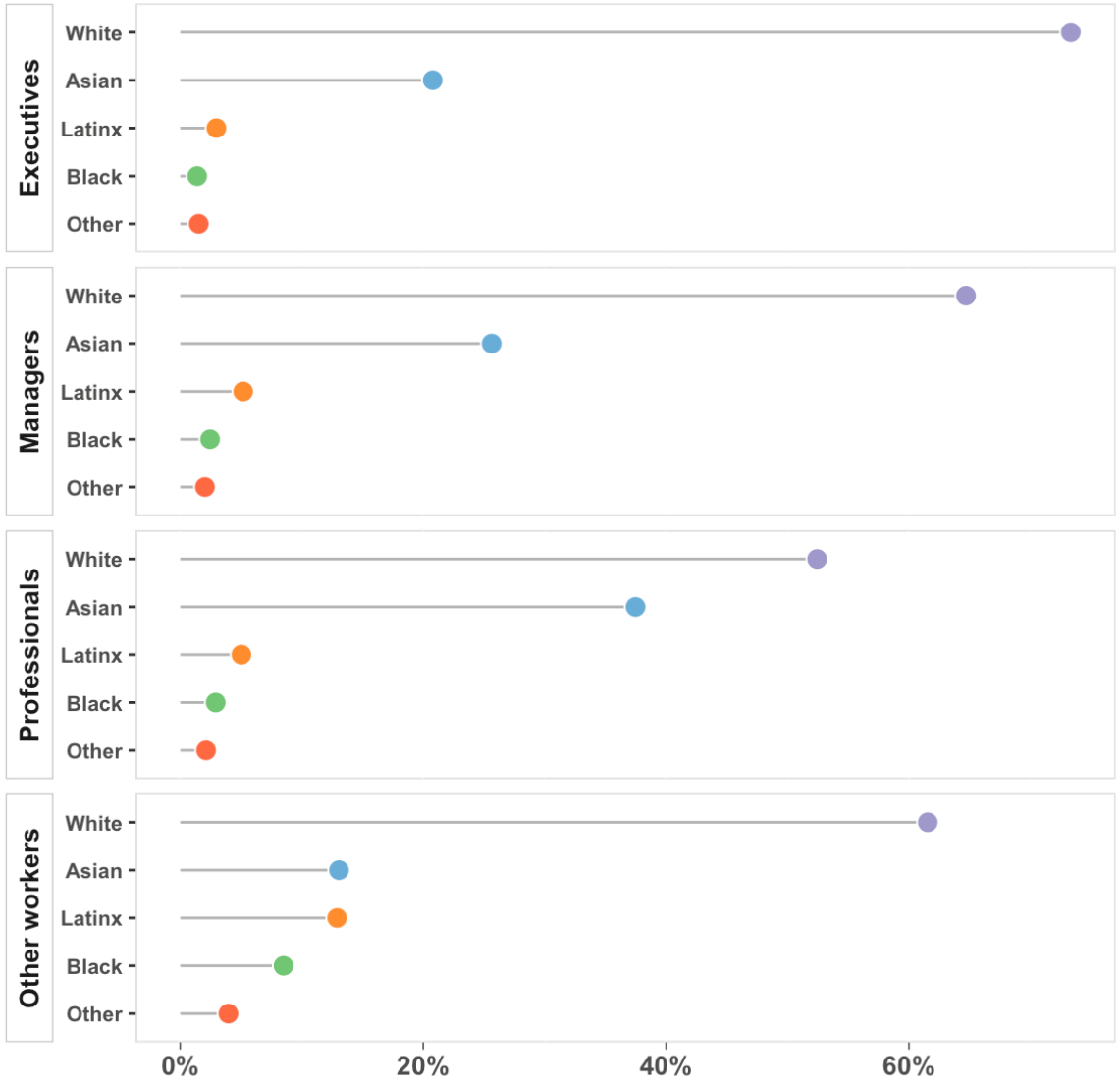




Source: Reveal, https://www.revealnews.org/topic/silicon-valley-diversity/

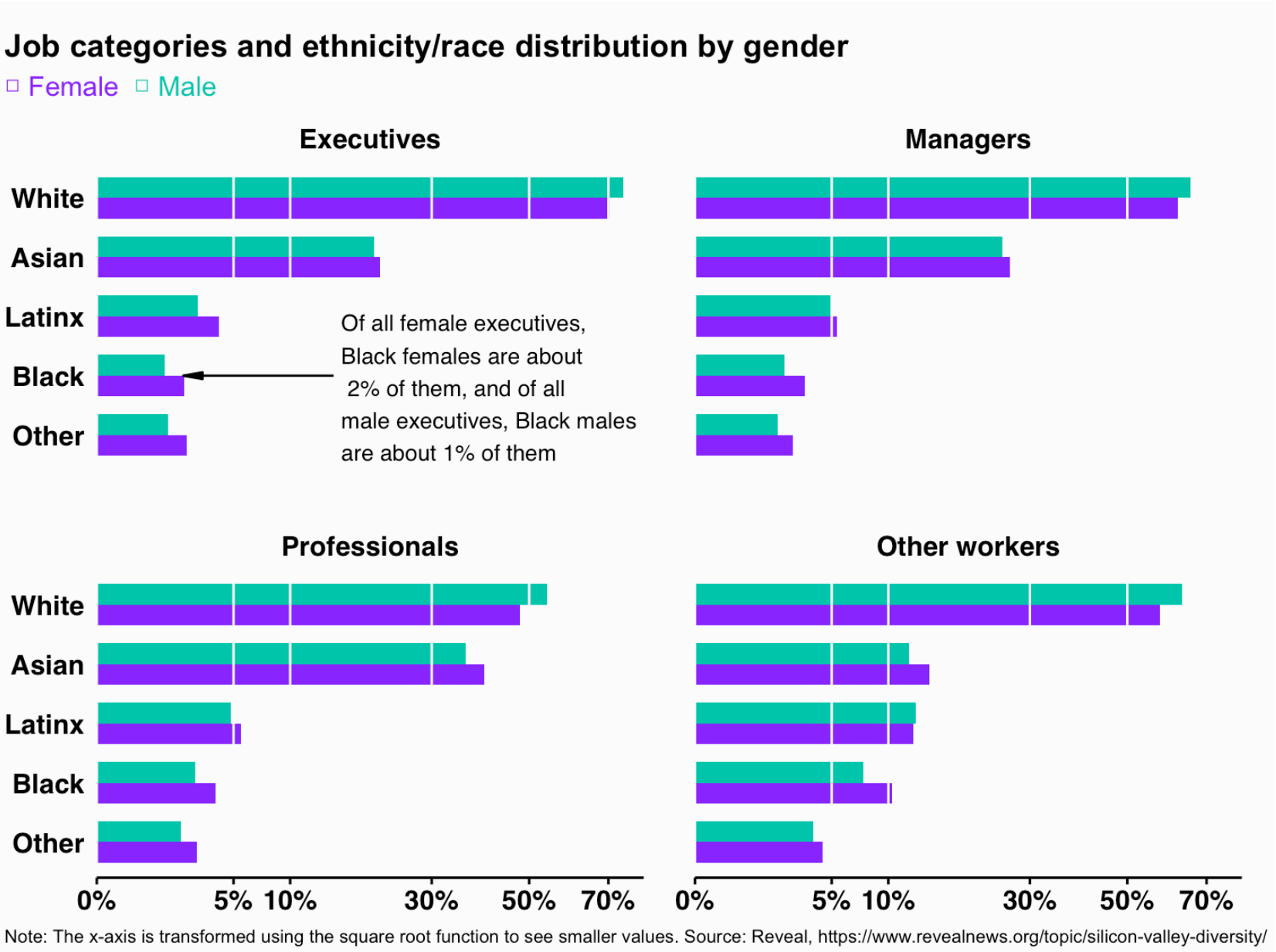


Source: Reveal, https://www.revealnews.org/topic/silicon-valley-diversity/



Source: Reveal, https://www.revealnews.org/topic/silicon-valley-diversity/

What if we want to compare genders within the job categories and ethnicities/races?



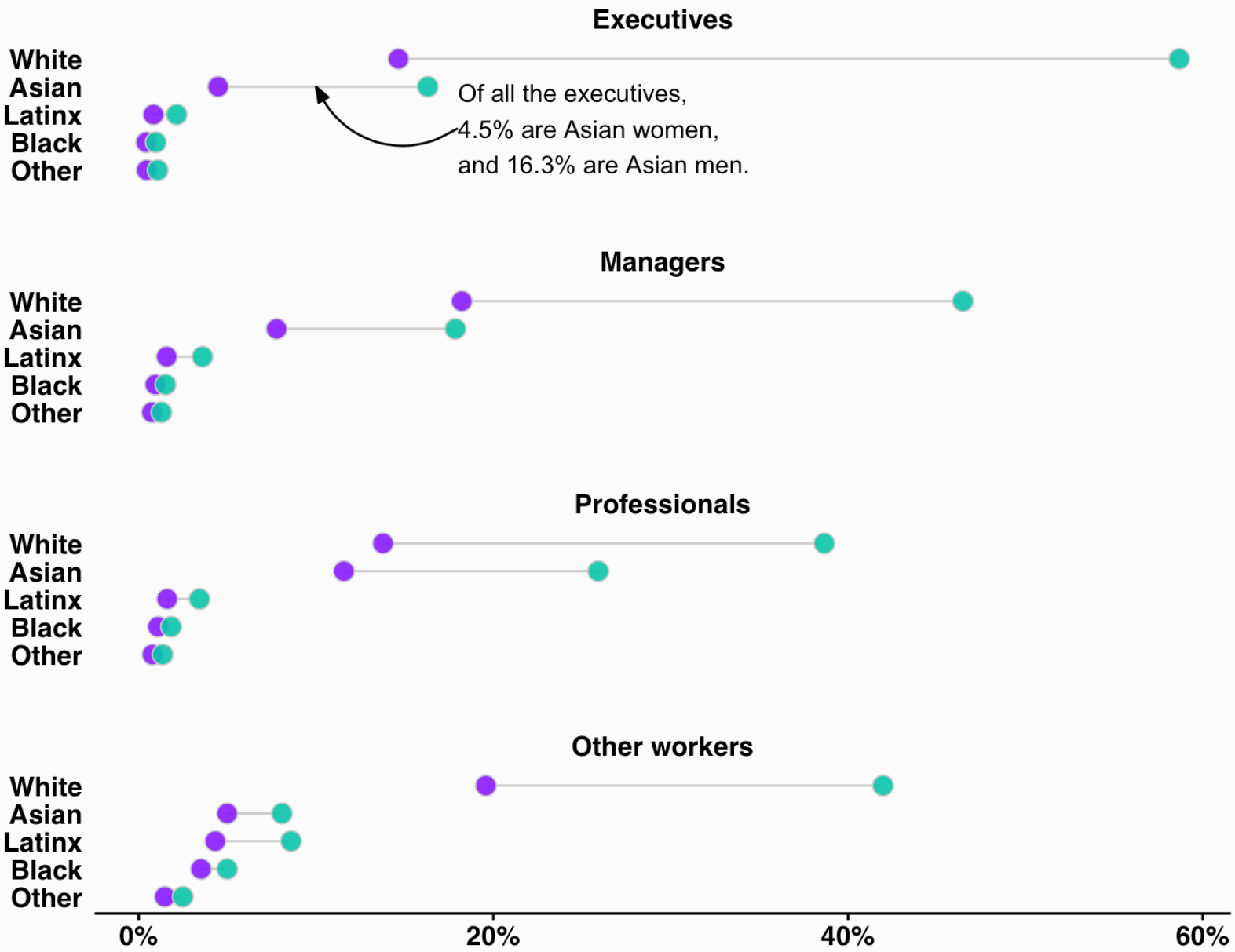
Note: The x-axis is transformed using the square root function to see smaller values. Source: Reveal, https://www.revealnews.org/topic/silicon-valley-diversity/

Job categories and ethnicity/race distribution by gender

Source: Reveal, https://www.revealnews.org/topic/silicon-valley-diversity/

Job categories and ethnicity/race distribution by gender

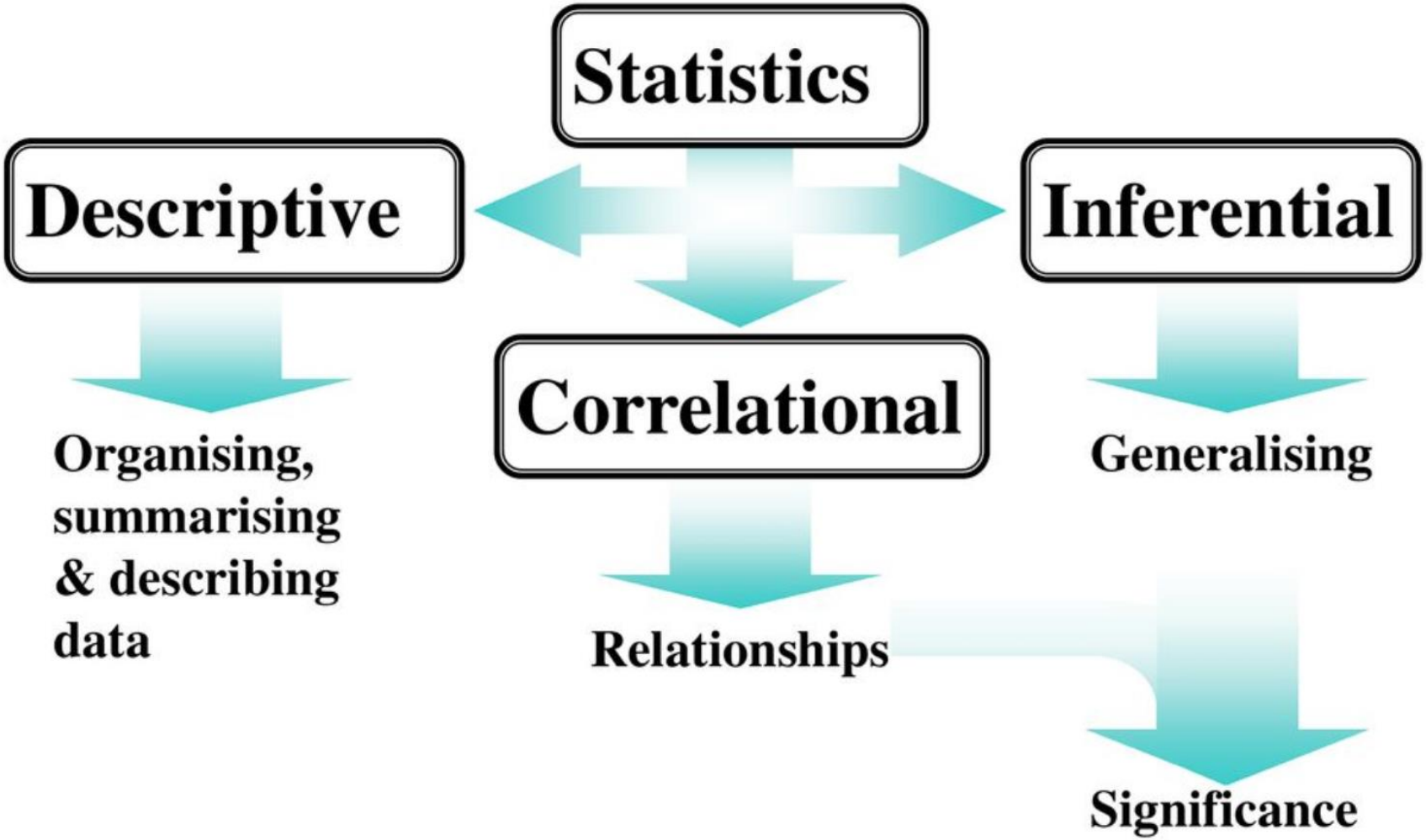
○ Female ○ Male



Source: Reveal, https://www.revealnews.org/topic/silicon-valley-diversity/

Outline

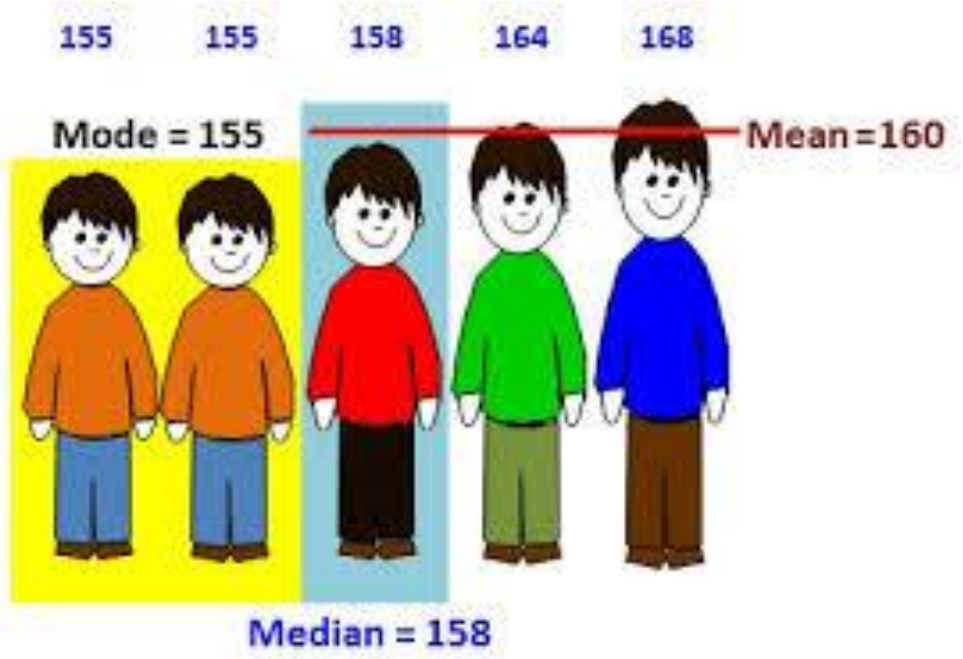
- **Visualization**
  - why we visualise
  - how to pick a plot
  - initial data vs final results visualization (some examples)
  - bad designs and misleading graphs
- **Summarization**
  - **measures of central tendency & dispersion**
  - **which measure to pick**


Statistics
Descriptive
Inferential
Correlational
Organising, summarising & describing data
Generalising
Relationships
Significance

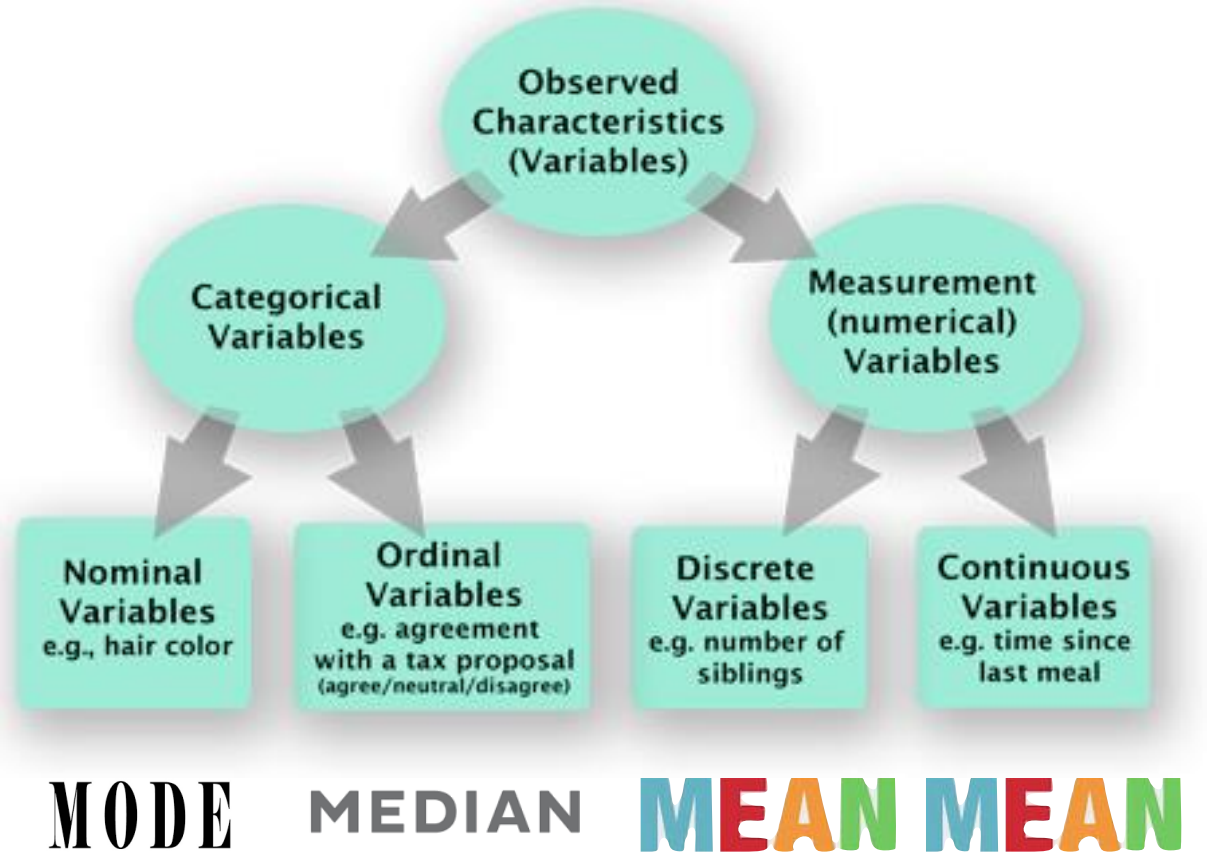
Descriptive Statistics

- Common descriptive statistics are:
  - Measure of **central tendency**
    - the most typical value of a given group of values
  - Measure of **dispersion**
    - how much all the other values in the group vary around the typical value

Measures of central tendency



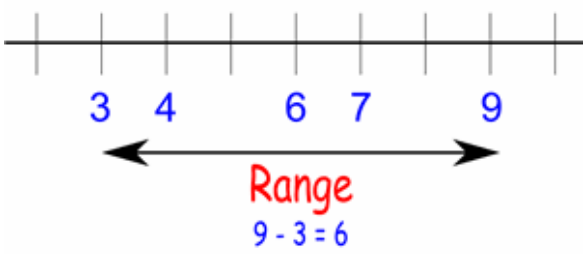
Central Tendency for Variable Types

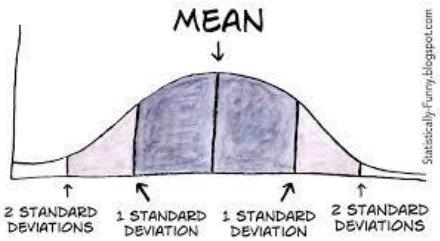


Measures of central tendency

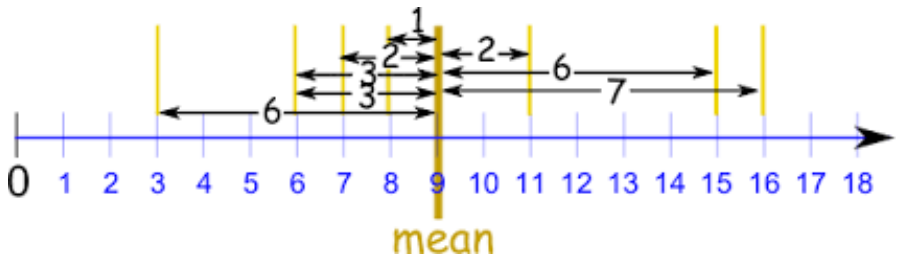
	Advantages	Disadvantages
MEAN		:tion
MEDIAN		
MODE		

Measures of dispersion/spread



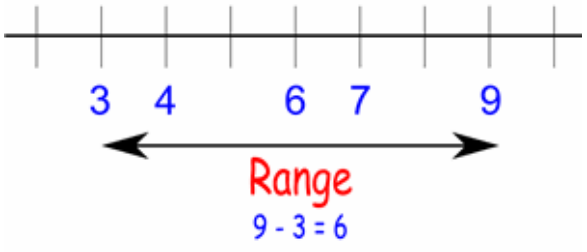


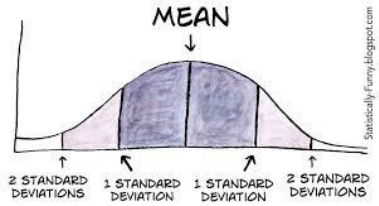
$$\mathrm{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$



$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$$

Measures of dispersion/spread

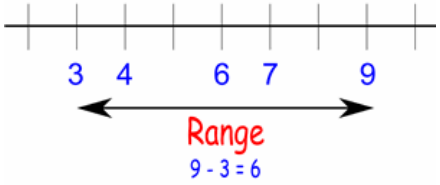


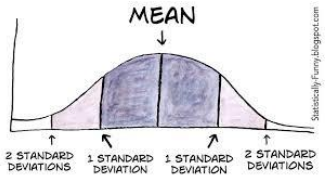


$$\mathrm{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

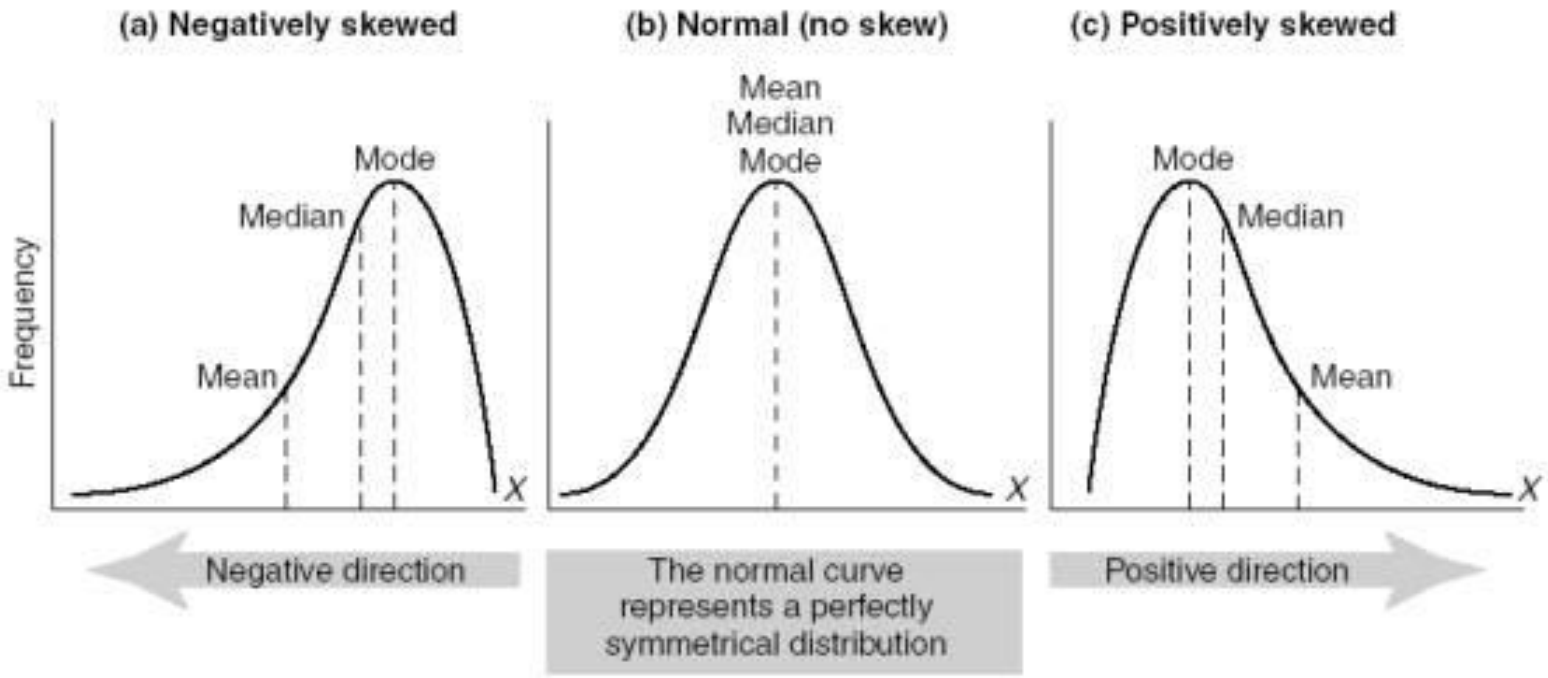
Advantages	Disadvantages
— —	distorted by extreme values no indication of grouping around the mean
- Fundamental to significance testing, and forms basis of Analysis of Variance (ANOVA) - Enables population parameters to be estimated from a sample of people	— —

MEAN ? MODE MEDIAN



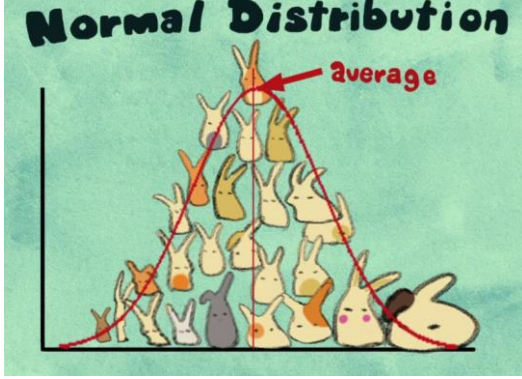


When do these measures fail to be representative ????


(a) Negatively skewed
(b) Normal (no skew)
(c) Positively skewed
Frequency
Mode
Median
Mean
Mean
Median
Mode
Mode
Median
Mean
X
X
X
Negative direction
The normal curve represents a perfectly symmetrical distribution
Positive direction

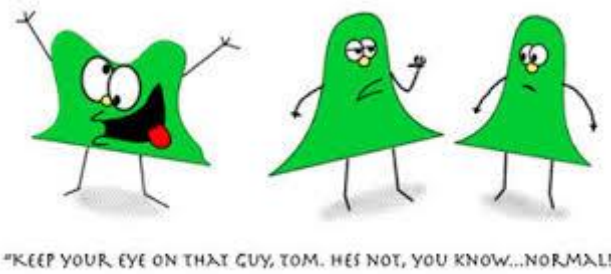

Distribution

Normal Distribution

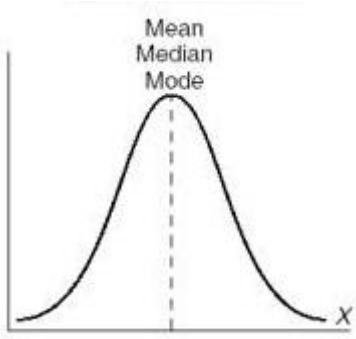


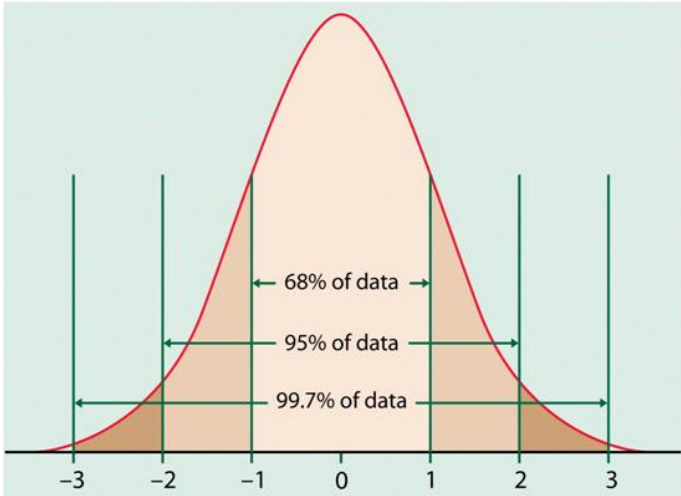
- A bell-shaped mathematical curve describing how values are distributed
- Data taken from a sample is **assumed** to be 'normally distributed', and must approximate this shape in order to use parametric tests of significance
- *Inferential statistics* (eg: t-tests, F-tests, regression analyses) require in some sense that the numeric variables are approximately normally distributed
- *Note:* it does not fit all populations

Normal Distribution

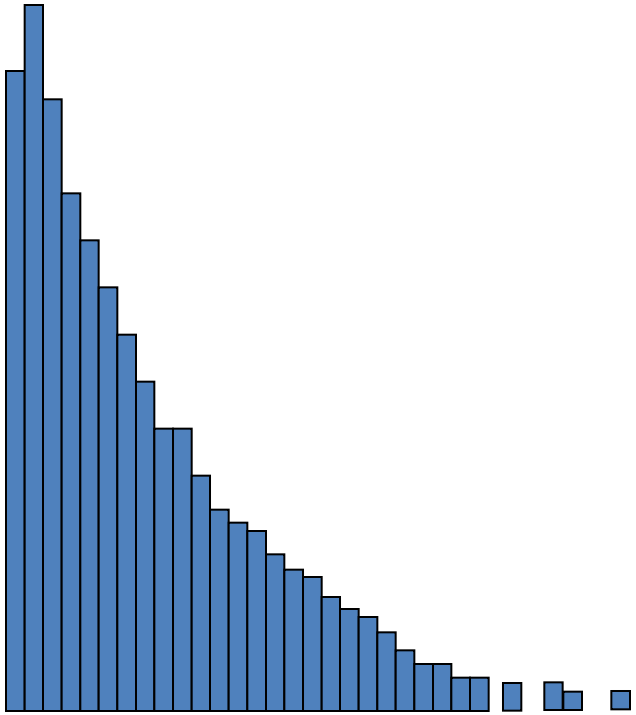


- symmetrical about the horizontal axis midpoint
- mean, median, and mode all fall on the midpoint
- No matter what μ and σ are, the area between
  - μ-σ and μ+σ is about 68%;
  - μ-2σ and μ+2σ is about 95%;
  - μ-3σ and μ+3σ is about 99.7%
- Almost all values fall within 3 standard deviations



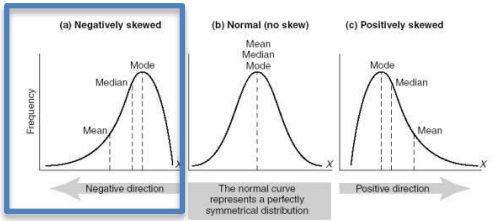


Skewed Distribution



- Resembles an exponential distribution
- Lots of extreme values far from mean or mode
- Not straightforward to do useful statistical tests with this type of distribution

Skewed Distribution



- **Negative skew**
  - Result from relatively easy tasks, due to a ceiling effect
- **Positive skew**
  - Results from tasks which are hard to improve upon, due to a floor effect (such as RT —reaction time)
- **Bimodal**
  - Two distinct peaks
  - probable indicator of groups
  - ex: completion time of marathon runners