BRSM
Data Visualisation & Summarization

Vinoo Alluri

data

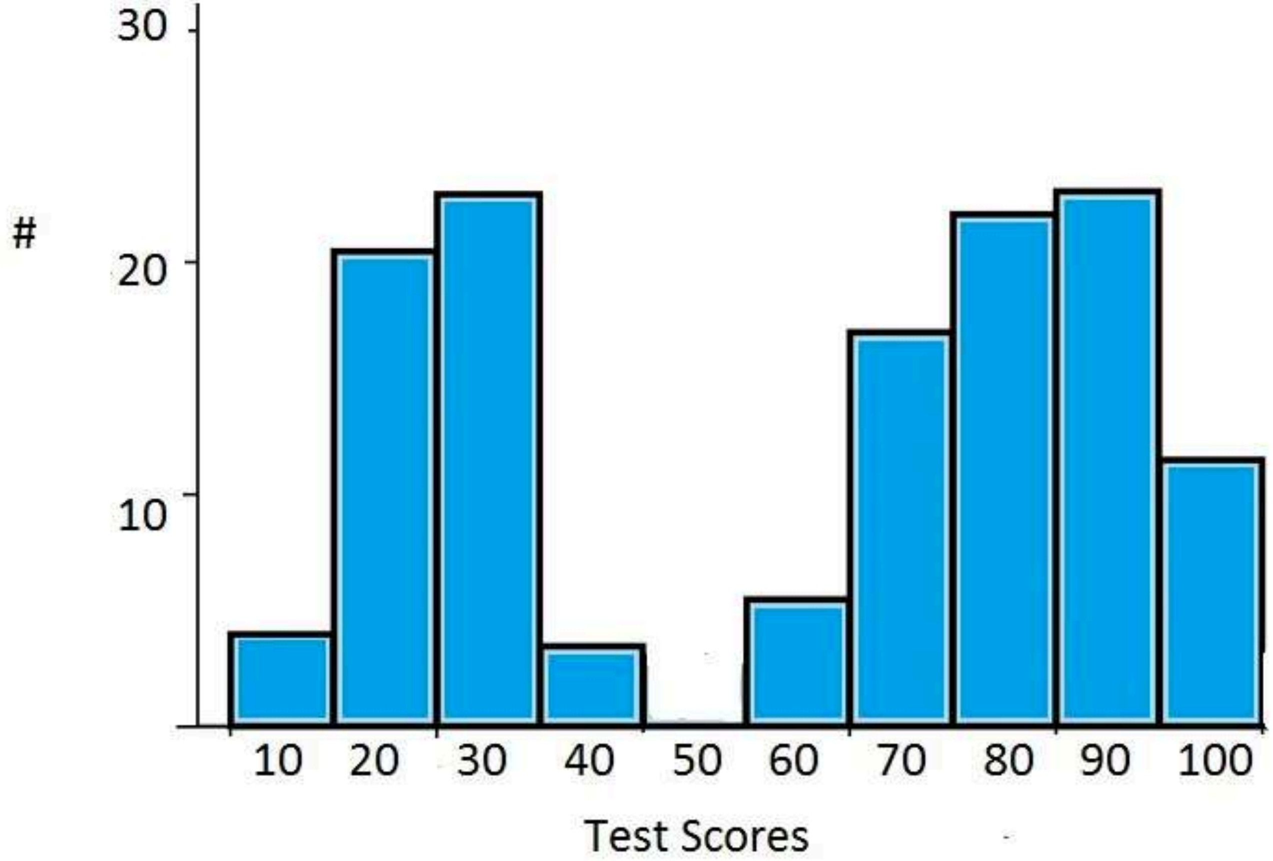
SUMMARY

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

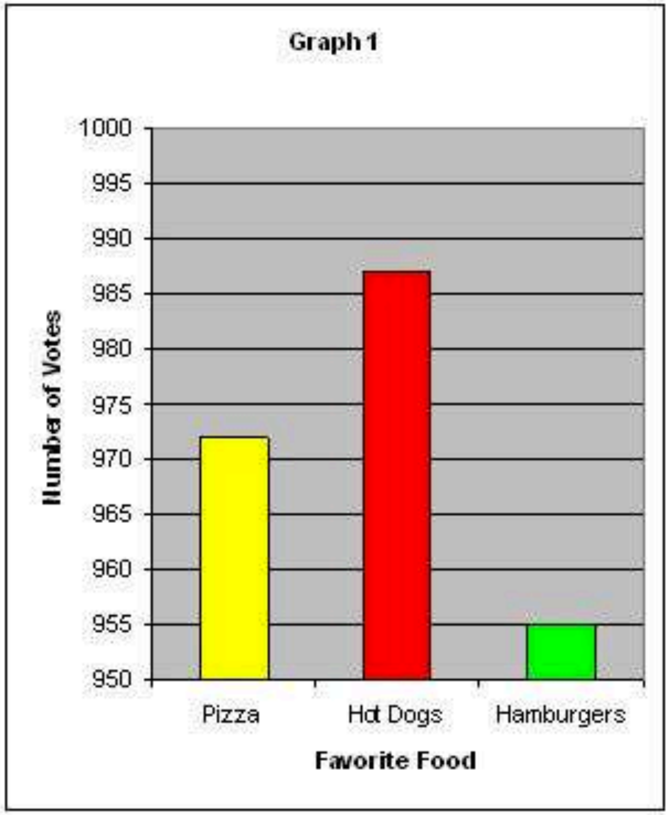
EXAMPLE

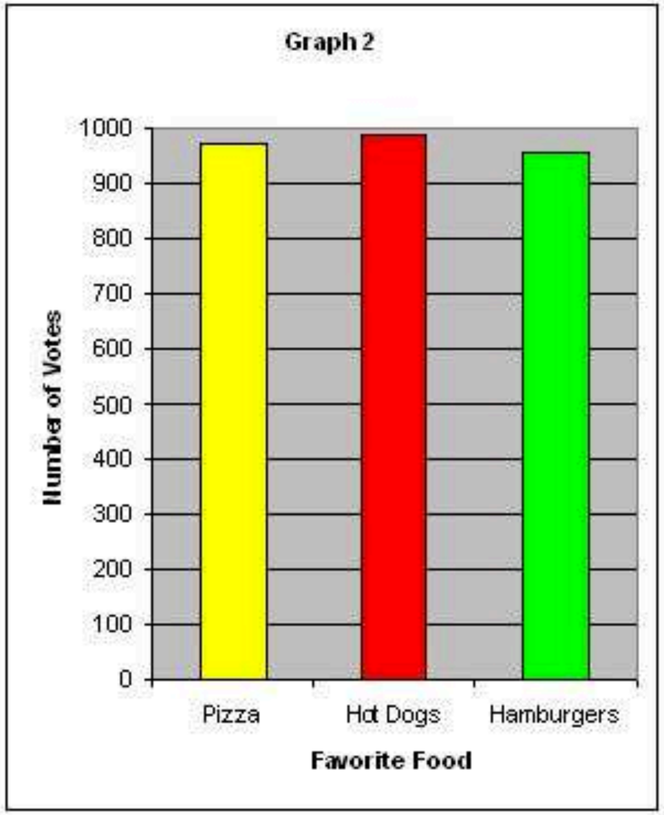
Mean End-Sem Test Score = 65.5



How can i summarise this data?

EXAMPLE

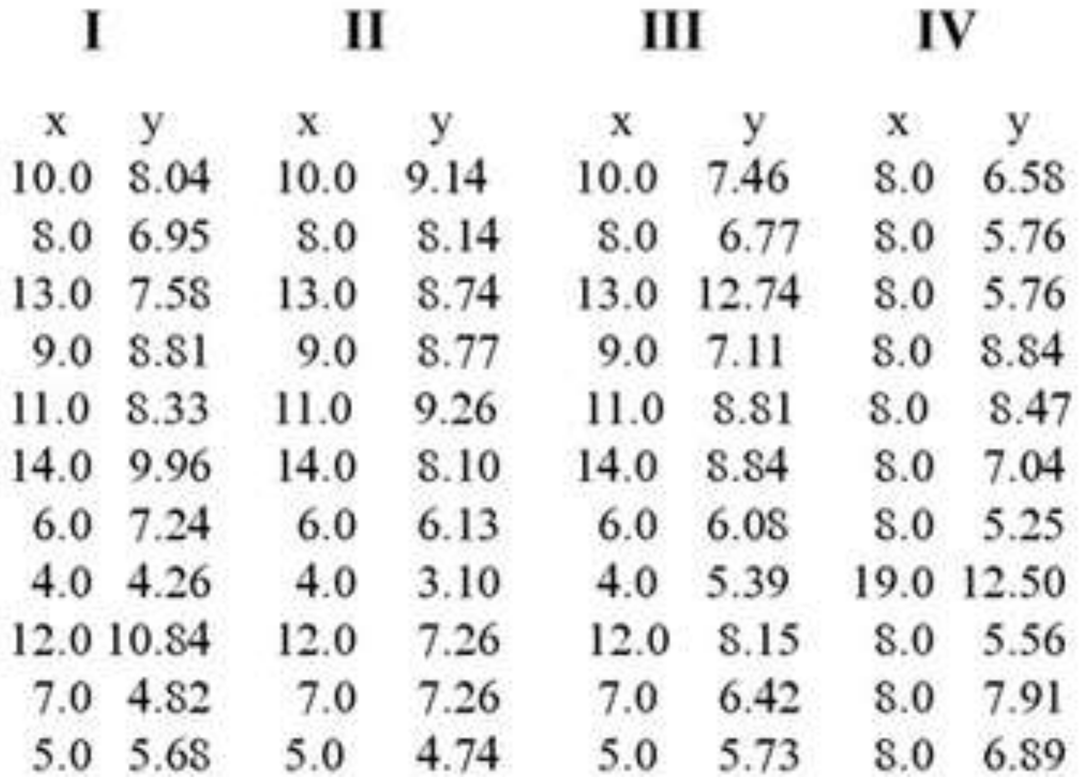

Graph 1
1000
995
990
985
980
975
970
965
960
955
950
Number of Votes
Pizza
Hot Dogs
Hamburgers
Favorite Food


Graph 2
1000
900
800
700
600
500
400
300
200
100
0
Number of Votes
Pizza
Hot Dogs
Hamburgers
Favorite Food

EXAMPLE

Anscombe’s Quartet

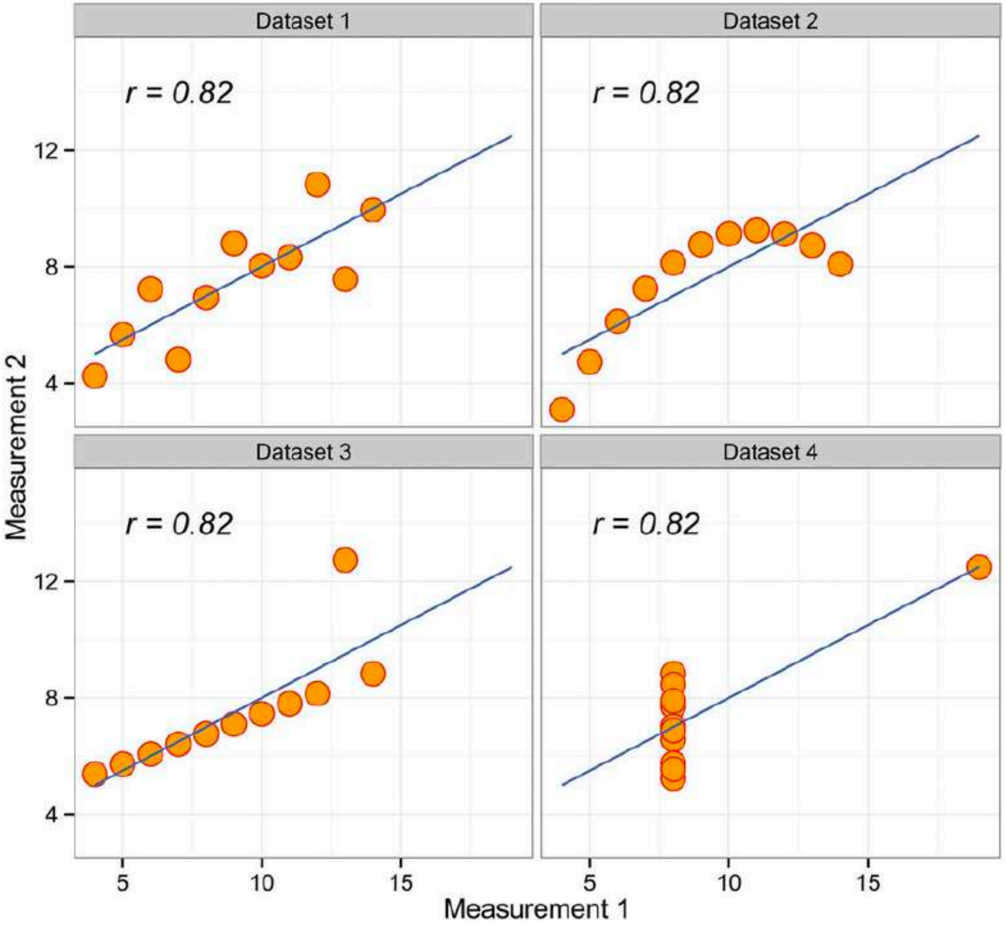
- same mean, std, correlation, regression line



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

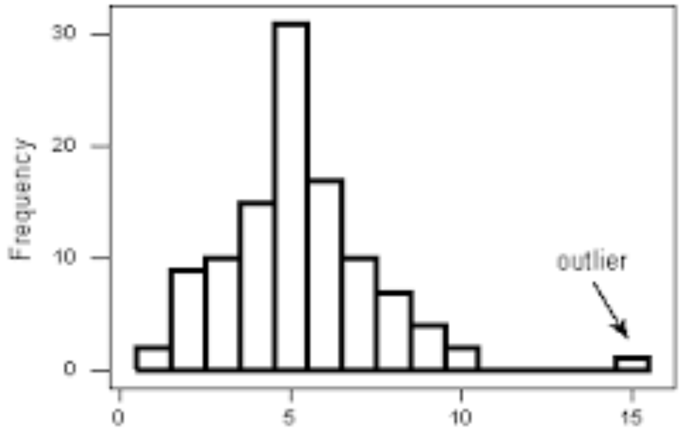


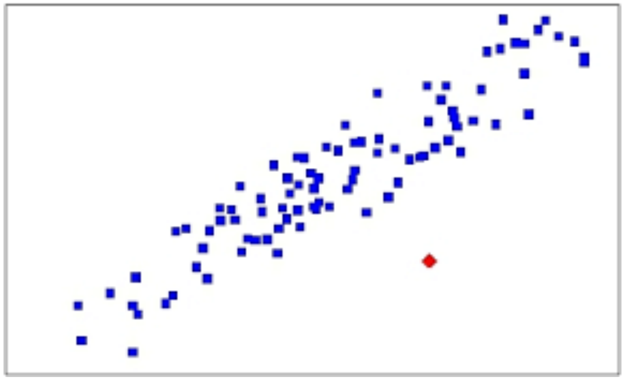
Outline

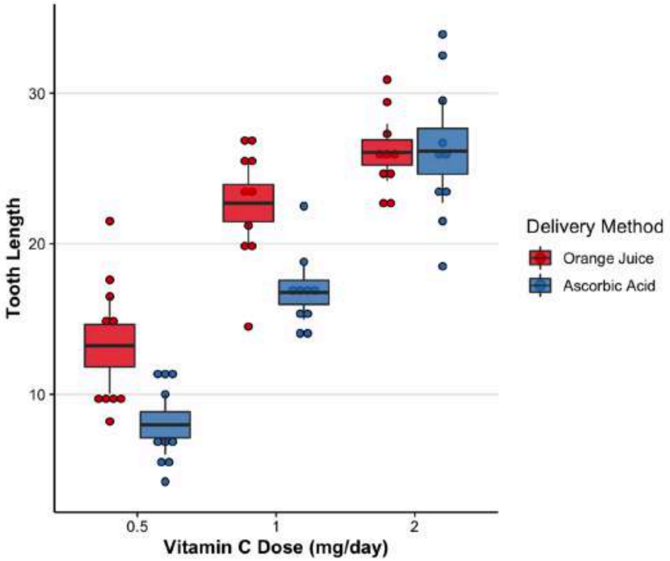
- **Visualization**
  - **why we visualise**
  - how to pick a plot
  - initial data vs final results visualization (some examples)
  - bad designs and misleading graphs
- **Summarization**
  - measures of central tendency & dispersion
  - which measure to pick

Why do we visualise?

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





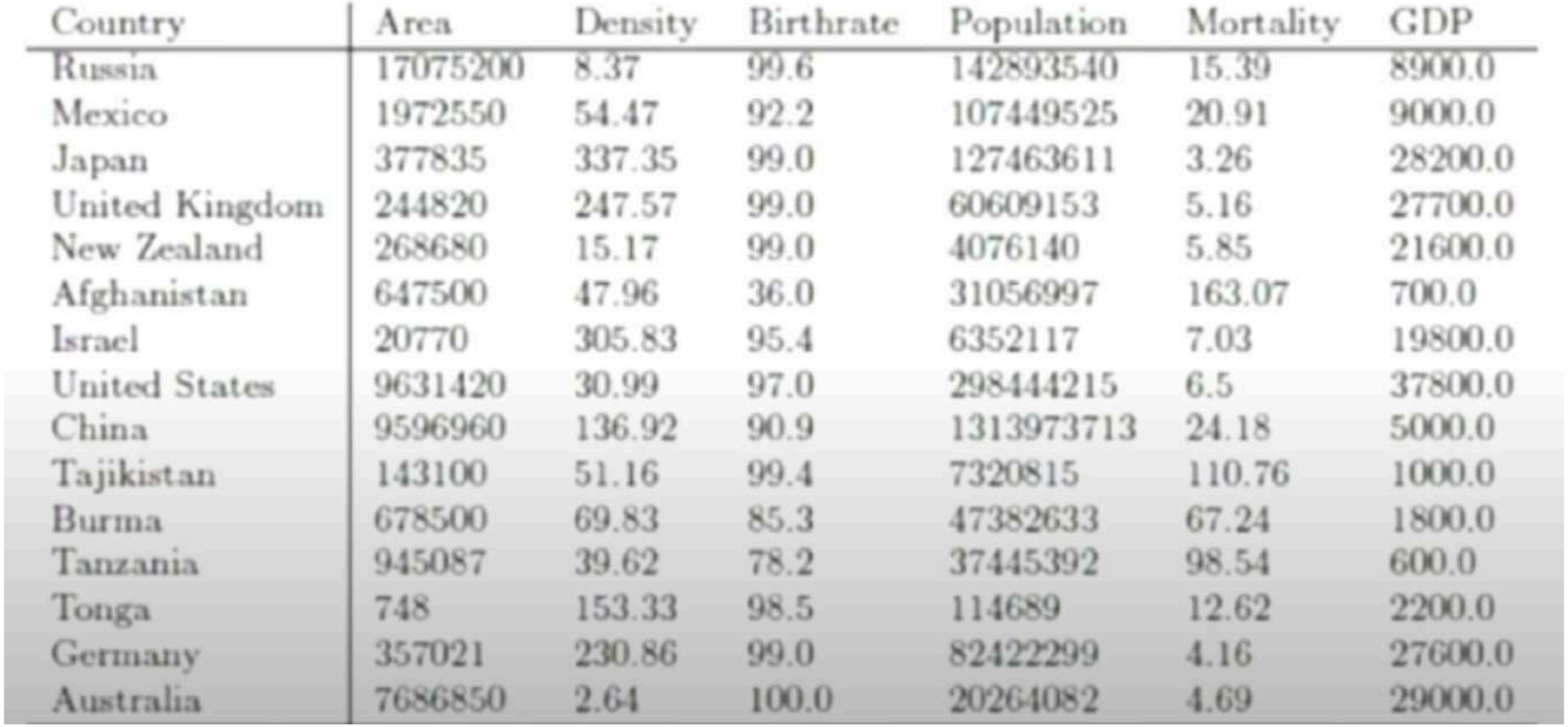


Visualization

Tables vs Figures

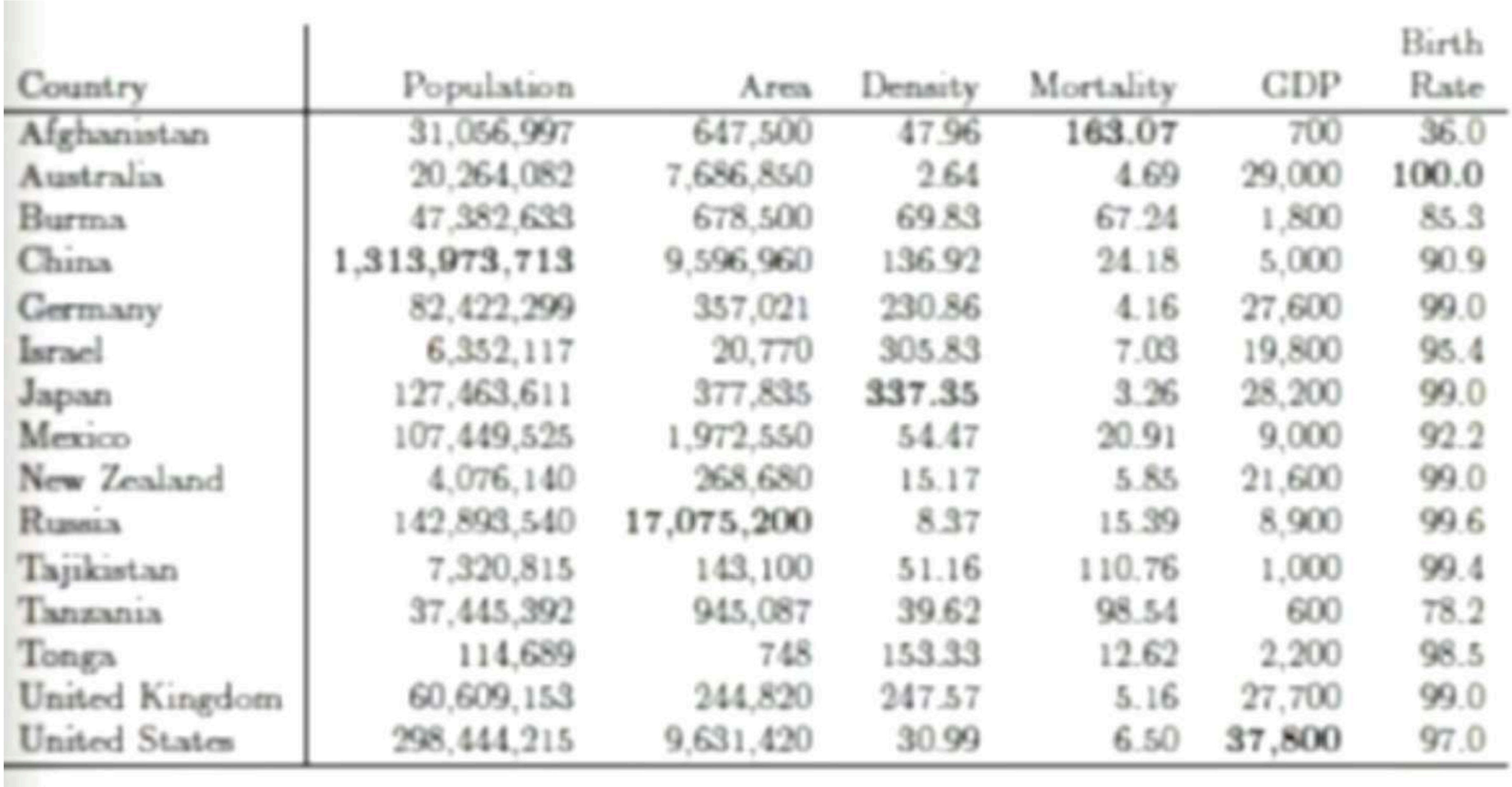
- **tables**
  - moderate amount of values
  - use when precision is key; specific values
  - multivariate visualization
  - represent heterogenous data
- **figures**
  - too many values
  - trends over time
  - identify patterns or shapes (eg: group differences, correlations, latent variables)

Can this table be improved?



Country	Area	Density	Birthrate	Population	Mortality	GDP
Russia	17075200	8.37	99.6	142893540	15.39	8900.0
Mexico	1972550	54.47	92.2	107449525	20.91	9000.0
Japan	377835	337.35	99.0	127463611	3.26	28200.0
United Kingdom	244820	247.57	99.0	60609153	5.16	27700.0
New Zealand	268680	15.17	99.0	4076140	5.85	21600.0
Afghanistan	647500	47.96	36.0	31056997	163.07	700.0
Israel	20770	305.83	95.4	6352117	7.03	19800.0
United States	9631420	30.99	97.0	298444215	6.5	37800.0
China	9596960	136.92	90.9	1313973713	24.18	5000.0
Tajikistan	143100	51.16	99.4	7320815	110.76	1000.0
Burma	678500	69.83	85.3	47382633	67.24	1800.0
Tanzania	945087	39.62	78.2	37445392	98.54	600.0
Tonga	748	153.33	98.5	114689	12.62	2200.0
Germany	357021	230.86	99.0	82422299	4.16	27600.0
Australia	7686850	2.64	100.0	20264082	4.69	29000.0

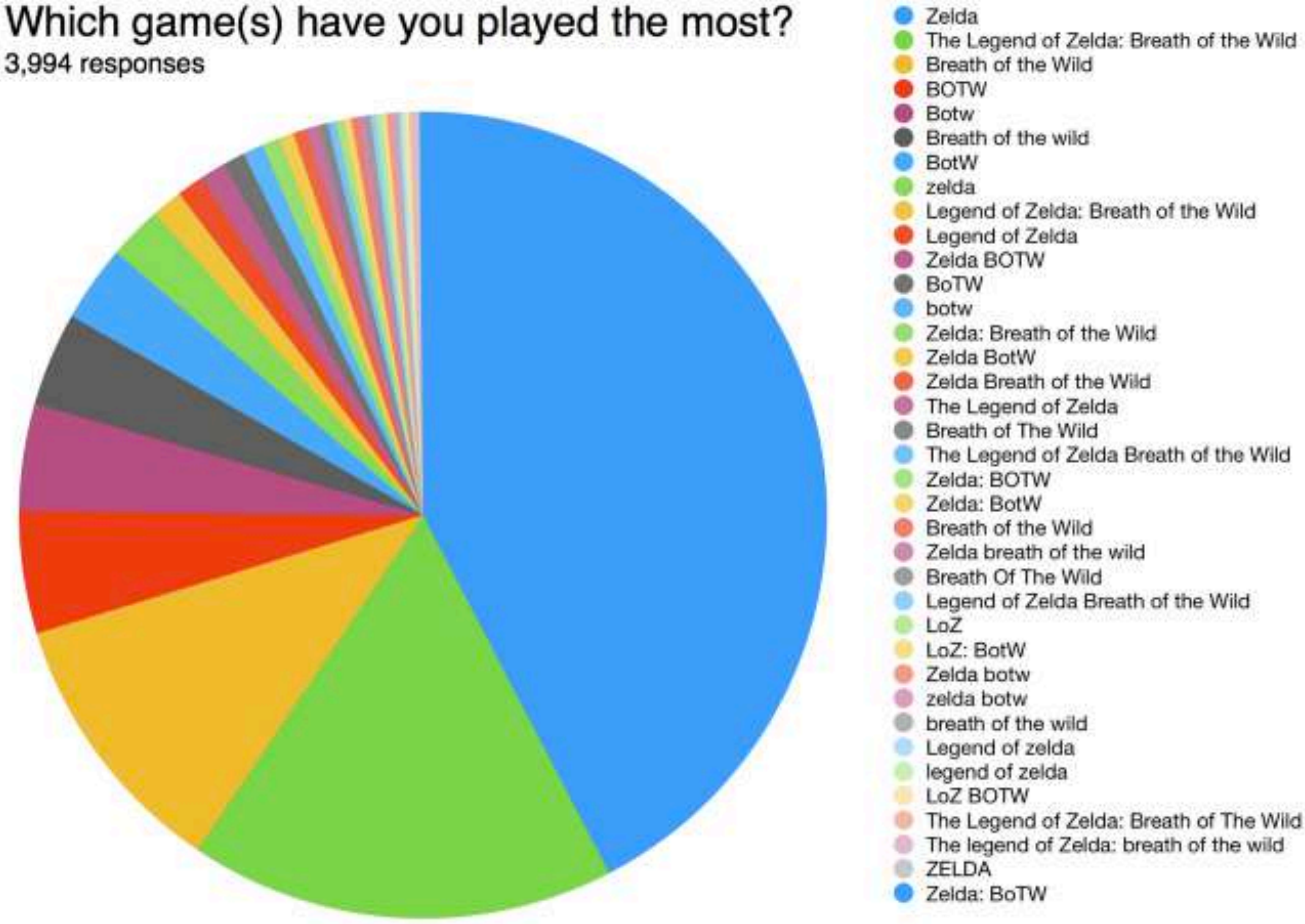
Can this table be improved?



Country	Population	Area	Density	Mortality	GDP	Birth Rate
Afghanistan	31,056,997	647,500	47.96	**163.07**	700	36.0
Australia	20,264,082	7,686,850	2.64	4.69	29,000	**100.0**
Burma	47,382,633	678,500	69.83	67.24	1,800	85.3
China	**1,313,973,713**	9,596,960	136.92	24.18	5,000	90.9
Germany	82,422,299	357,021	230.86	4.16	27,600	99.0
Israel	6,352,117	20,770	305.83	7.03	19,800	95.4
Japan	127,463,611	377,835	**337.35**	3.26	28,200	99.0
Mexico	107,449,525	1,972,550	54.47	20.91	9,000	92.2
New Zealand	4,076,140	268,680	15.17	5.85	21,600	99.0
Russia	142,893,540	**17,075,200**	8.37	15.39	8,900	99.6
Tajikistan	7,320,815	143,100	51.16	110.76	1,000	99.4
Tanzania	37,445,392	945,087	39.62	98.54	600	78.2
Tonga	114,689	748	153.33	12.62	2,200	98.5
United Kingdom	60,609,153	244,820	247.57	5.16	27,700	99.0
United States	298,444,215	9,631,420	30.99	6.50	**37,800**	97.0

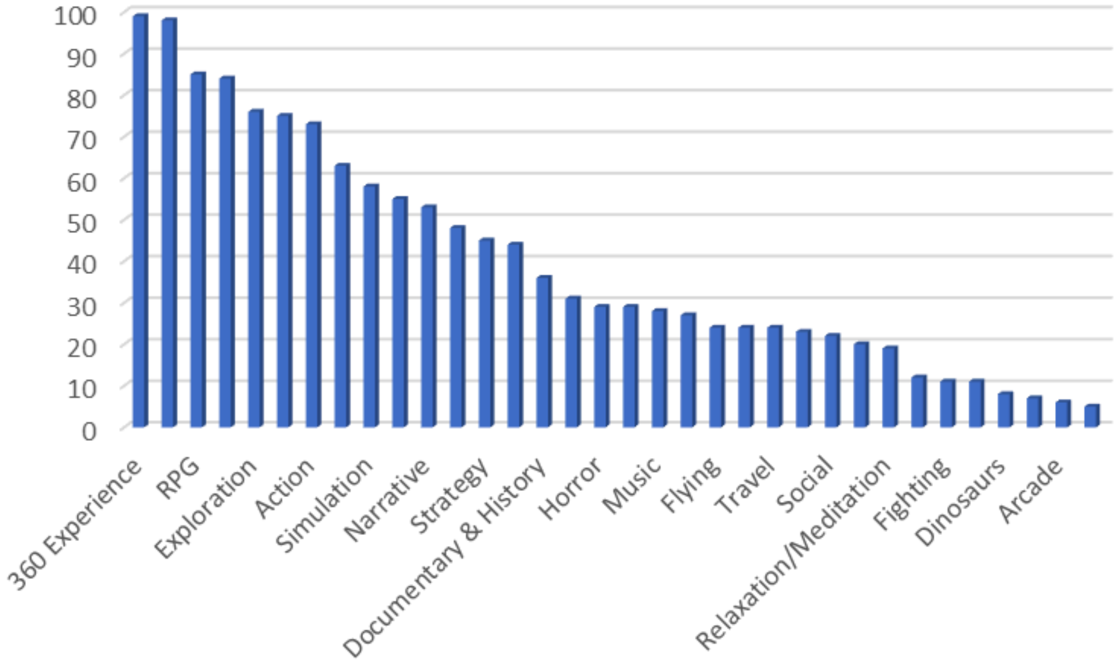
what makes them “good” or “bad”?

comment on these visualizations



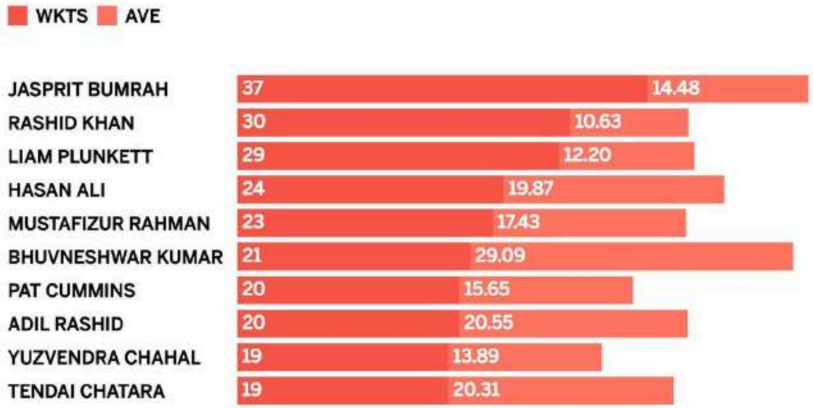
Which game(s) have you played the most?

3,994 responses

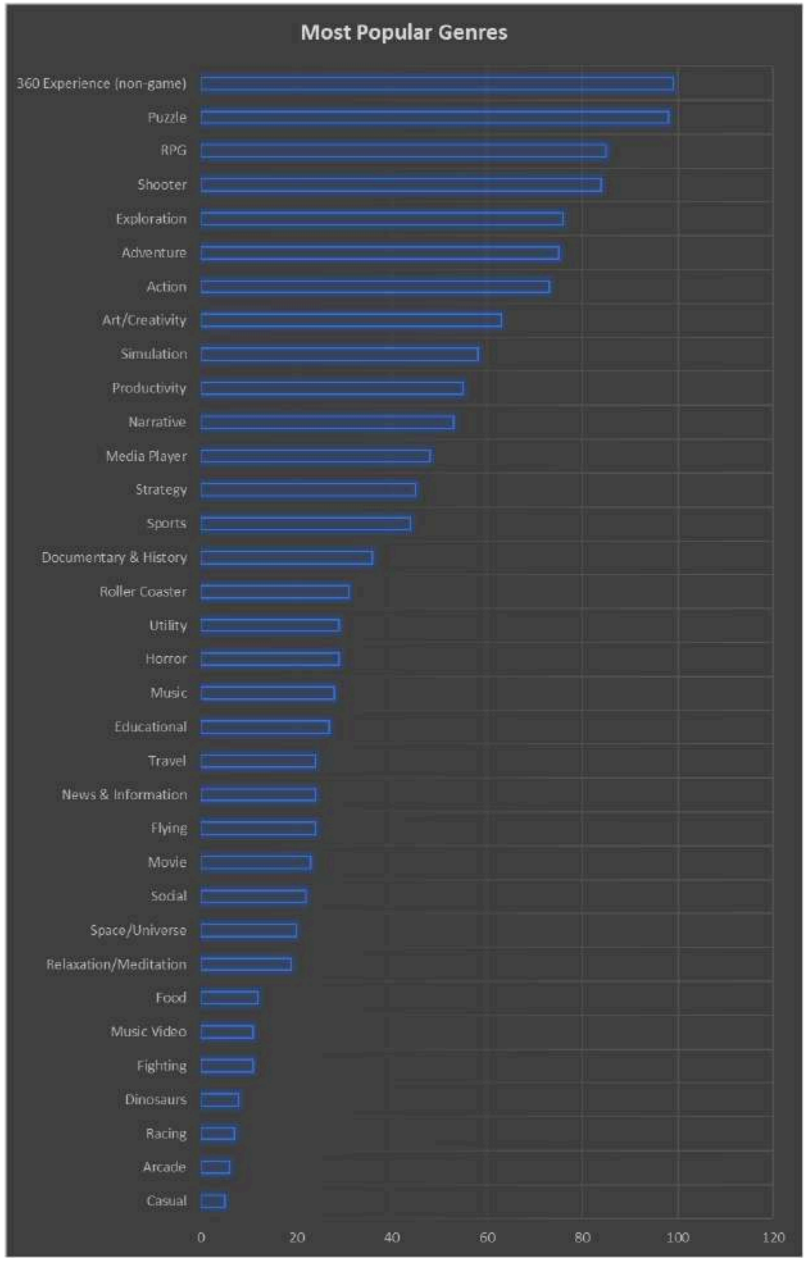


MOST WICKETS IN DEATH OVERS IN ODIS

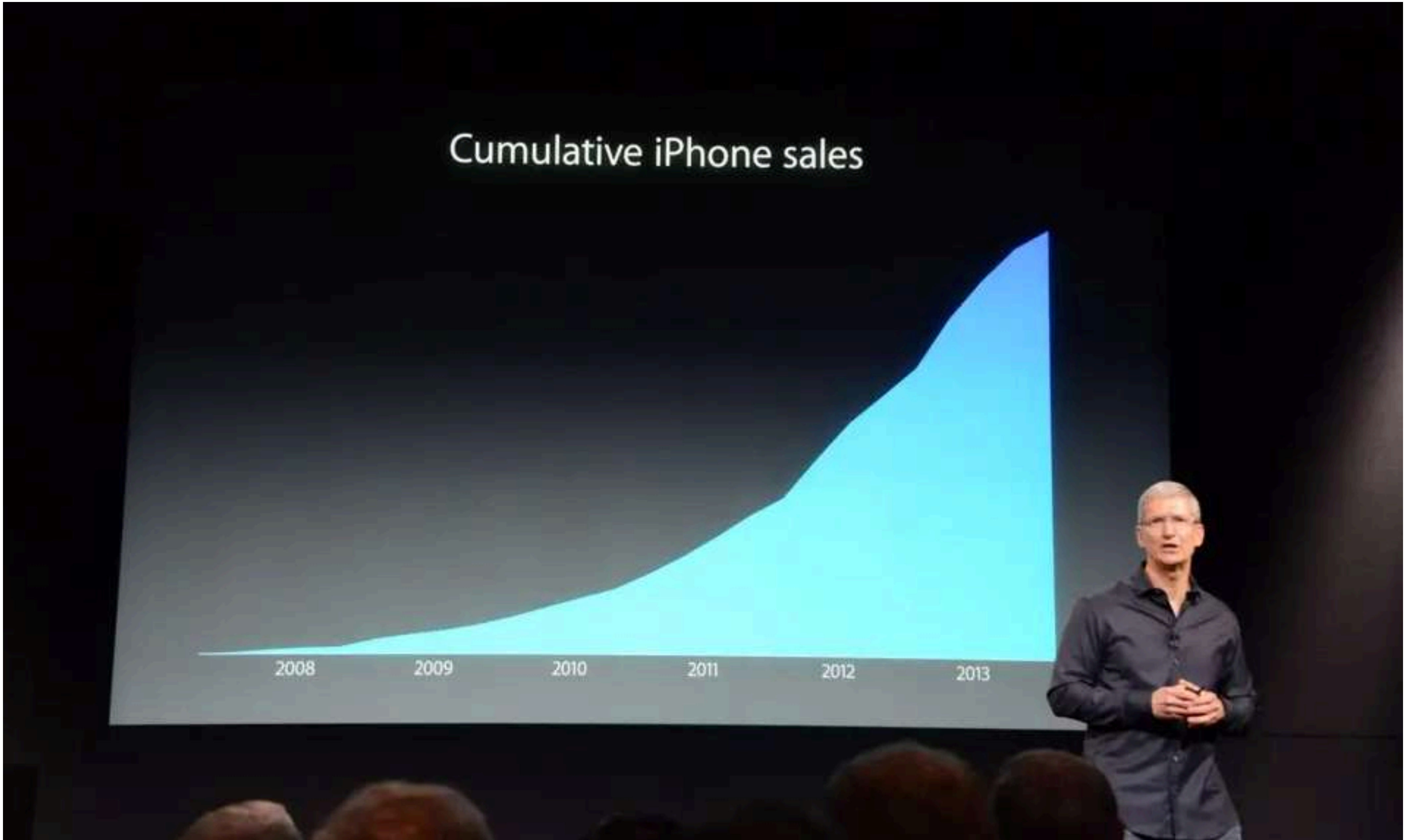
SINCE THE START OF JANUARY 2017

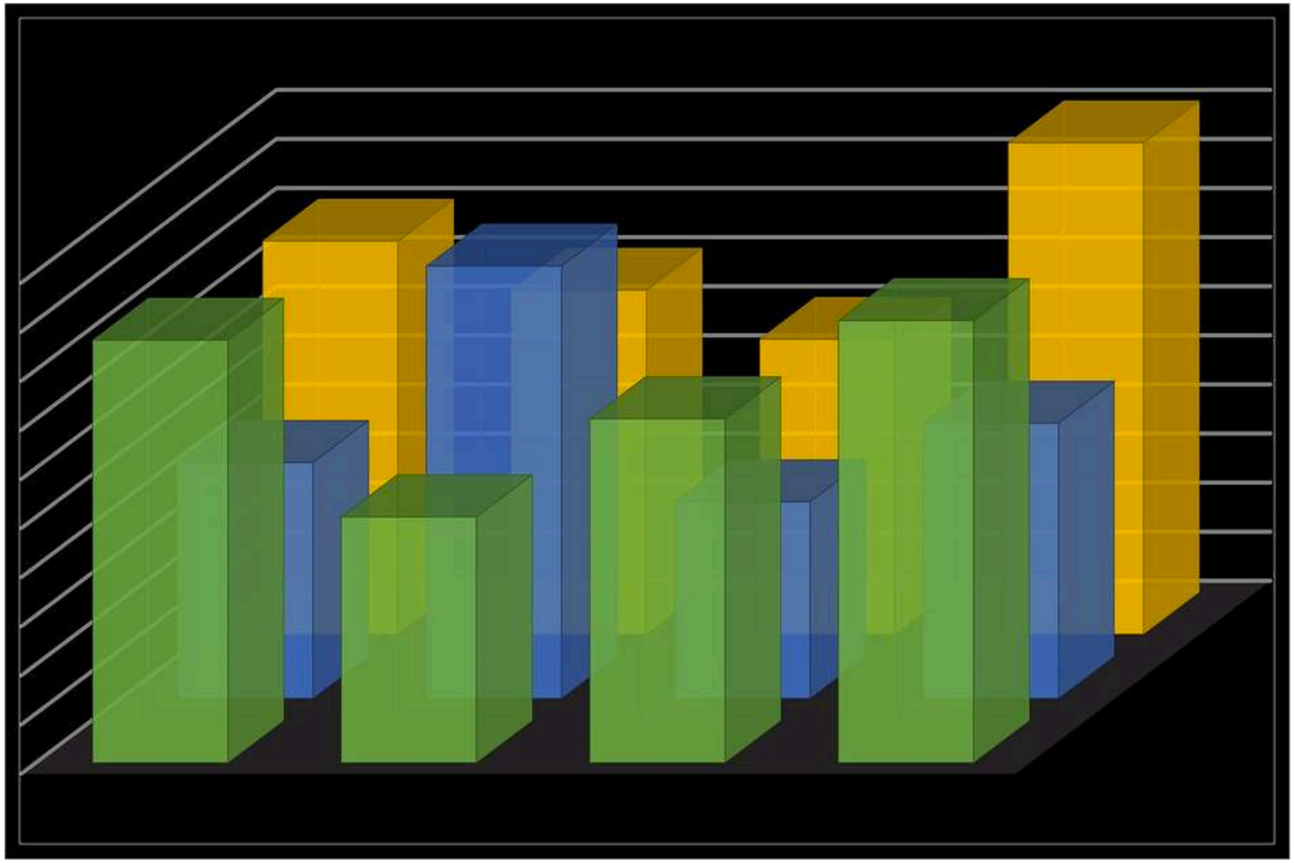


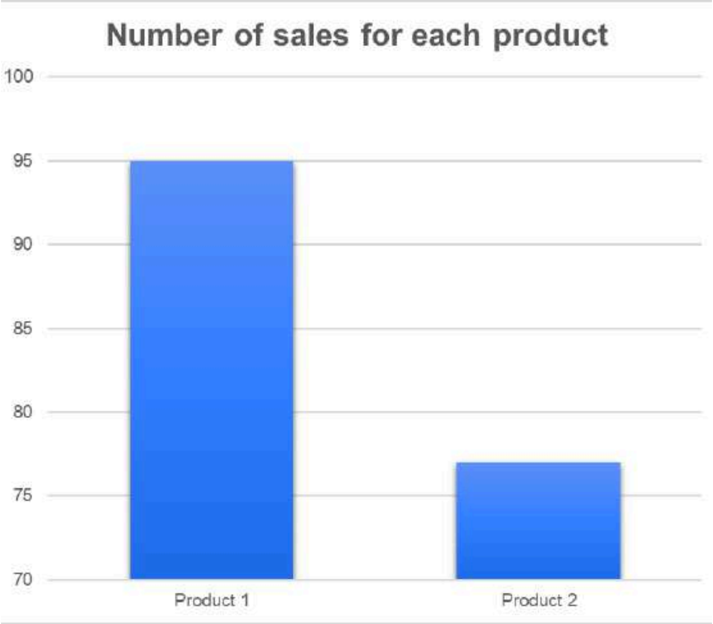
NUMBERS UPDATED TILL MAY 14, 2019

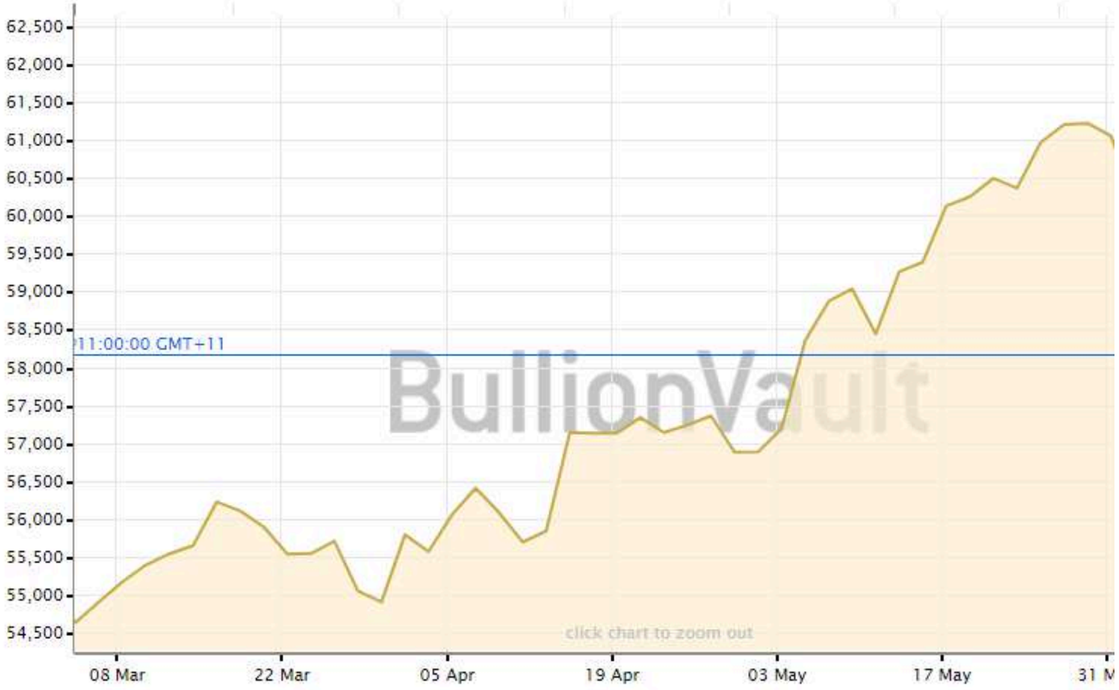


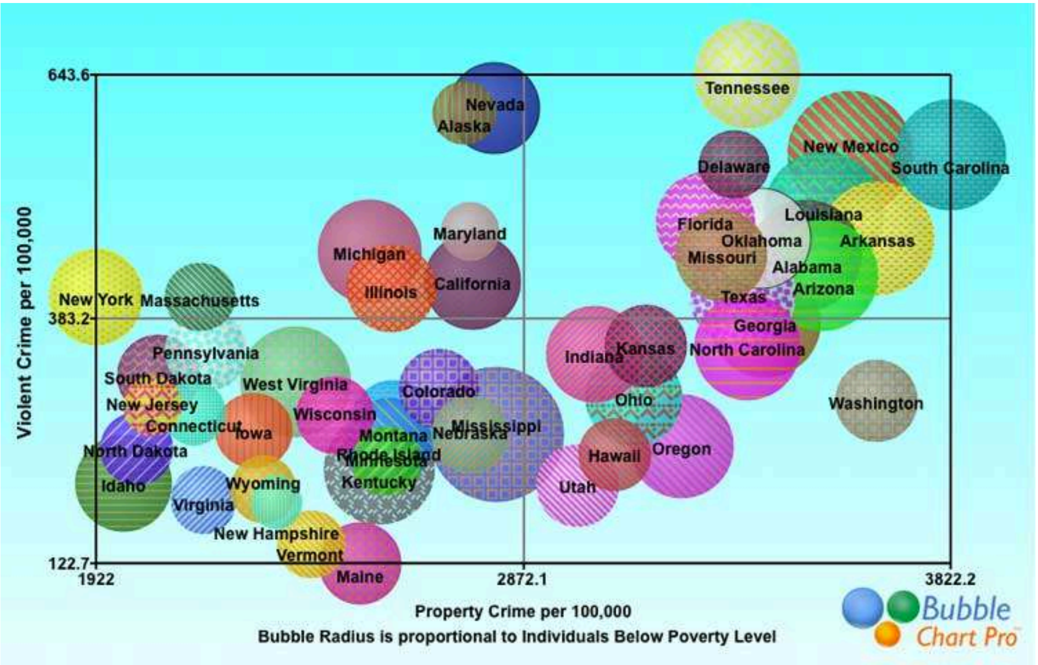
Tim Cook used the particular chart to showcase the rising sale of iPads between the years 2008-2013.

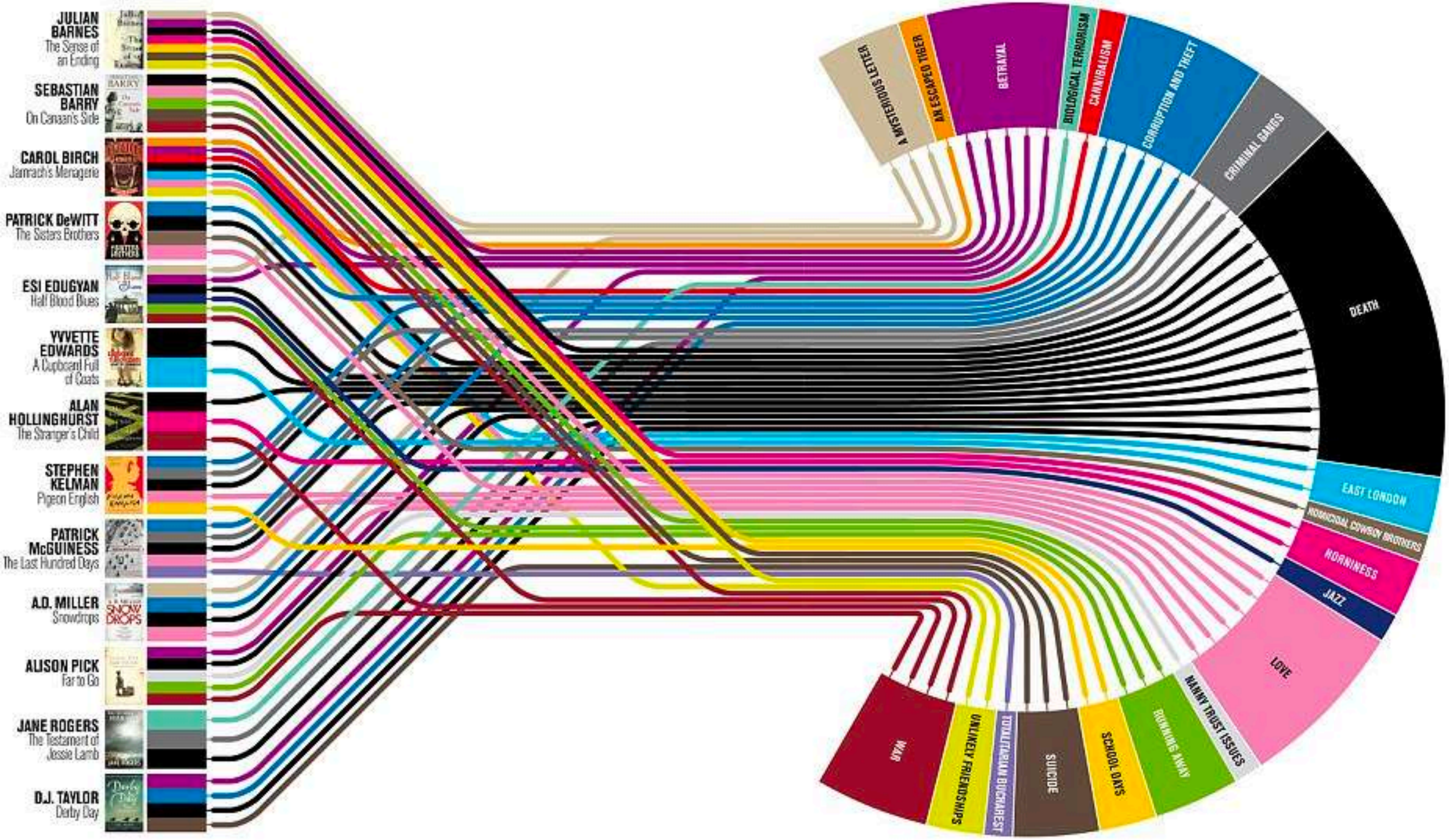





Number of sales for each product
100
95
90
85
80
75
70
Product 1
Product 2


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


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

[illegible]

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

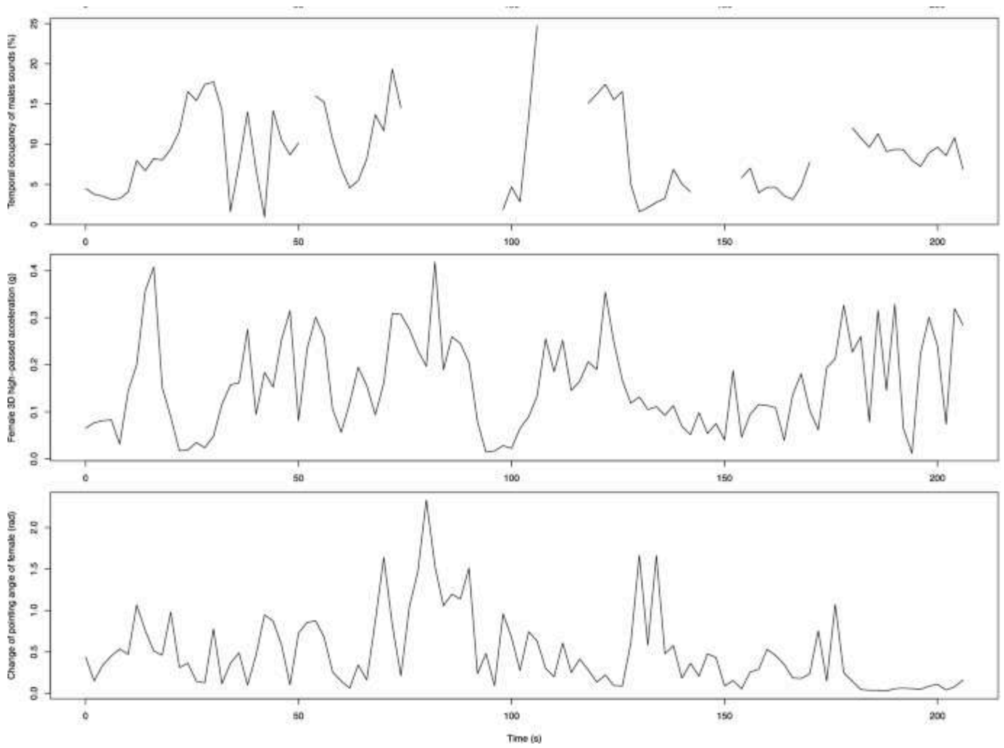
What makes a good visualisation

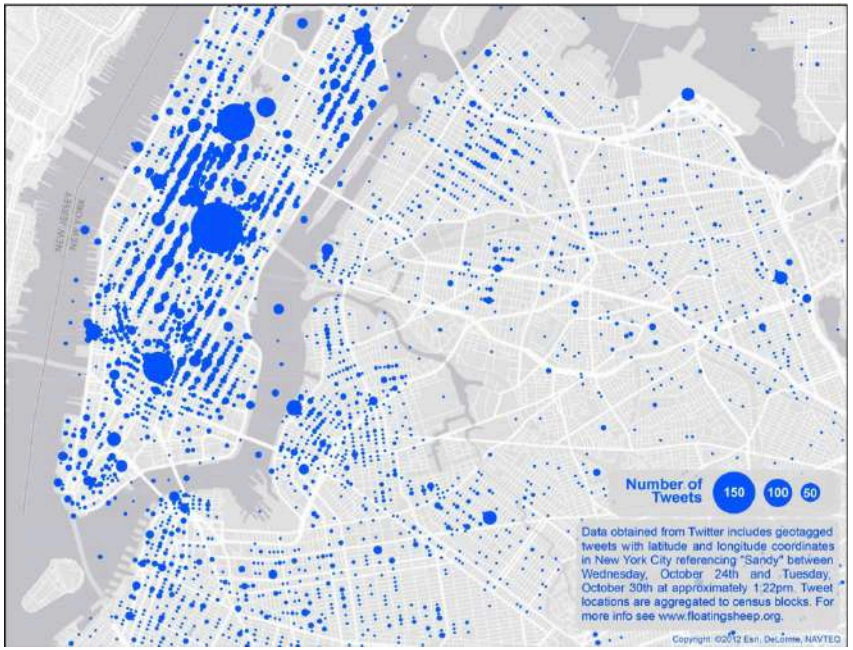
- Color Consistency
  - use same colors across multiple charts for consistency
  - avoid using colors with negligible contrast
  - avoid using too many colors
  - avoid using conventional colors to convey opposite meanings
  - pay heed to the needs of people who might be colorblind (check also in grayscale)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

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

What makes a good visualisation

- identify & explain/infer from missing data





Data obtained from Twitter includes geotagged tweets with latitude and longitude coordinates in New York City referencing "Sandy" between Wednesday, October 24th and Tuesday, October 30th at approximately 1:22pm. Tweet locations are aggregated to census blocks. For more info see www.floatingsheep.org.

Copyright ©2012 Esri, DeLorme, NAVTEQ

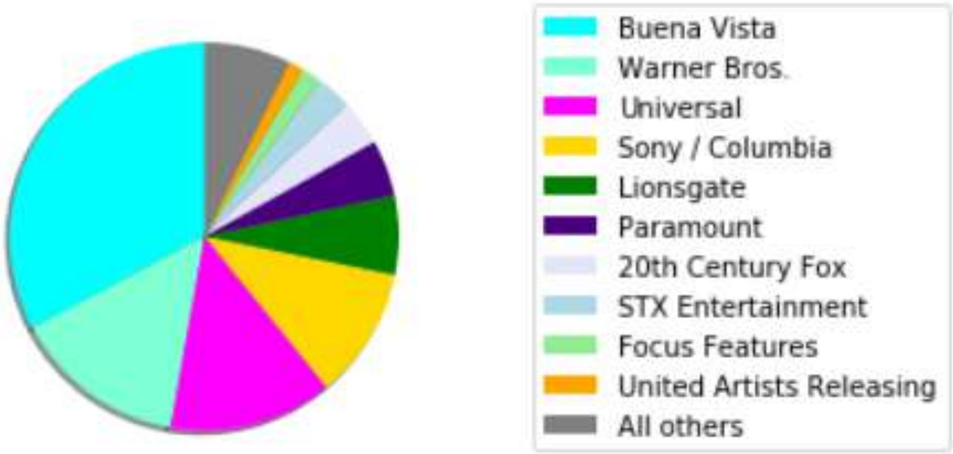
https://www.cardinalpath.com/blog/makes-good-visualization

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

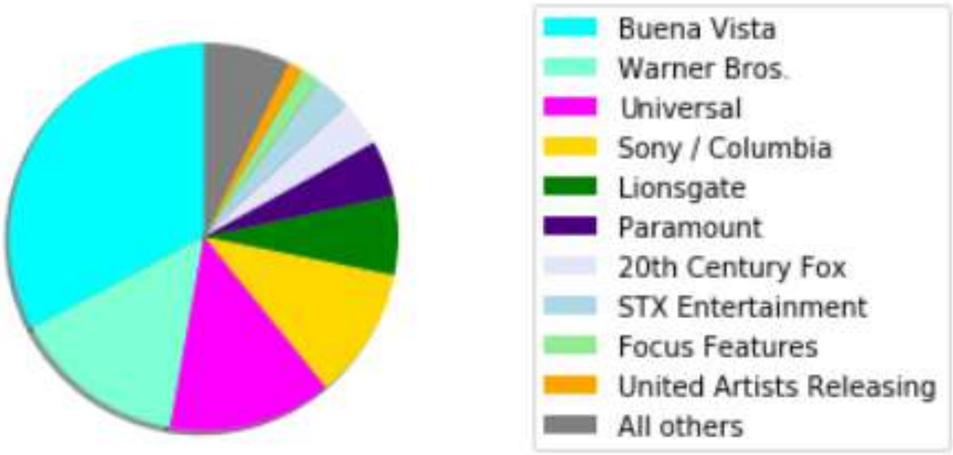
EXAMPLE



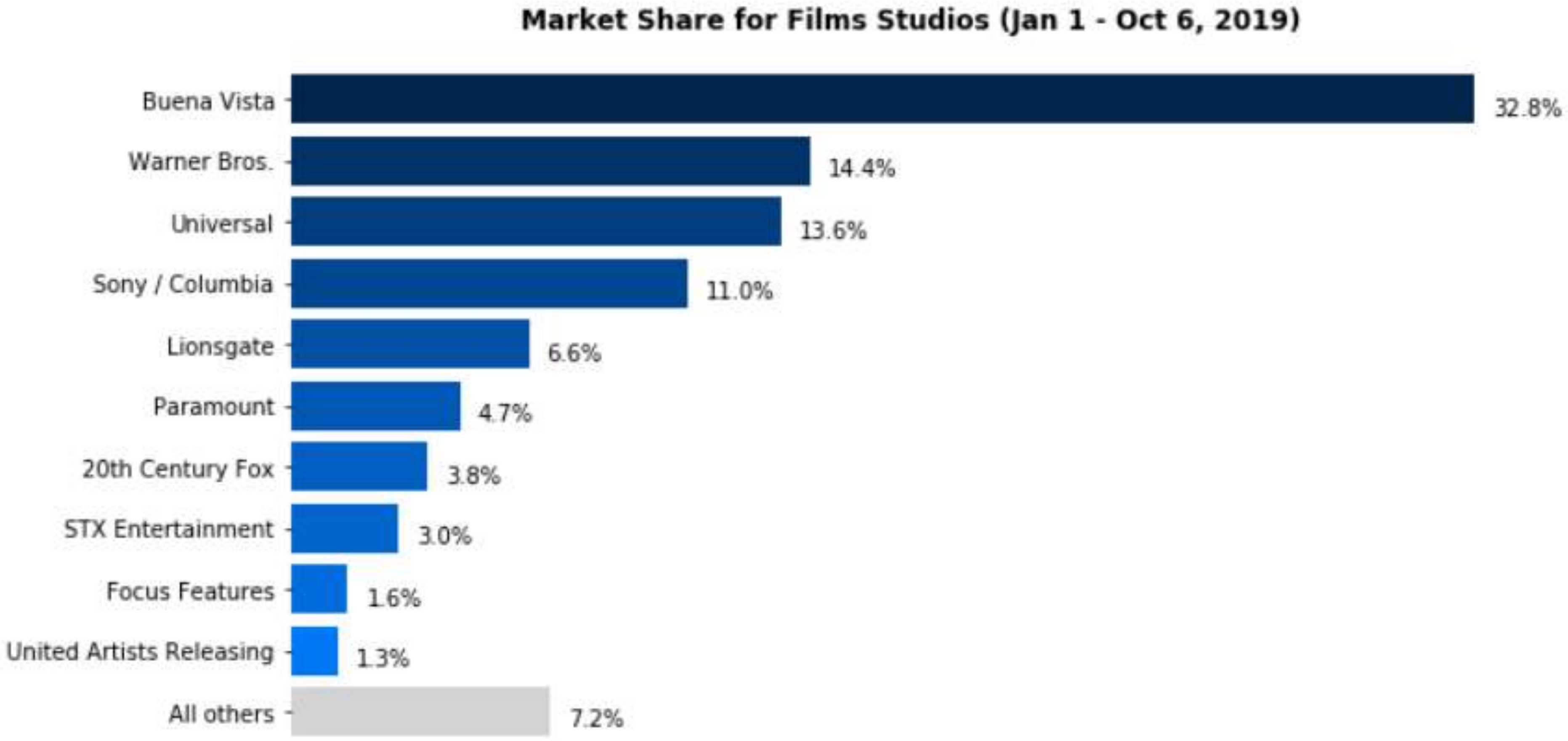
Market Share of Film Studios

PIE CHART

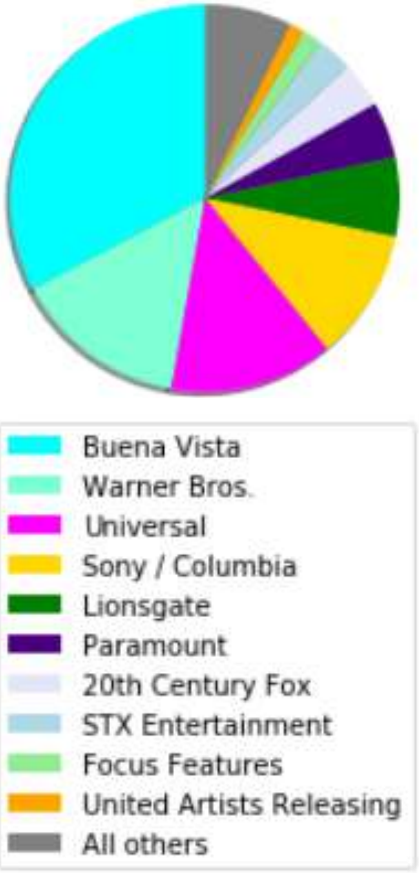
Not comprehensible!



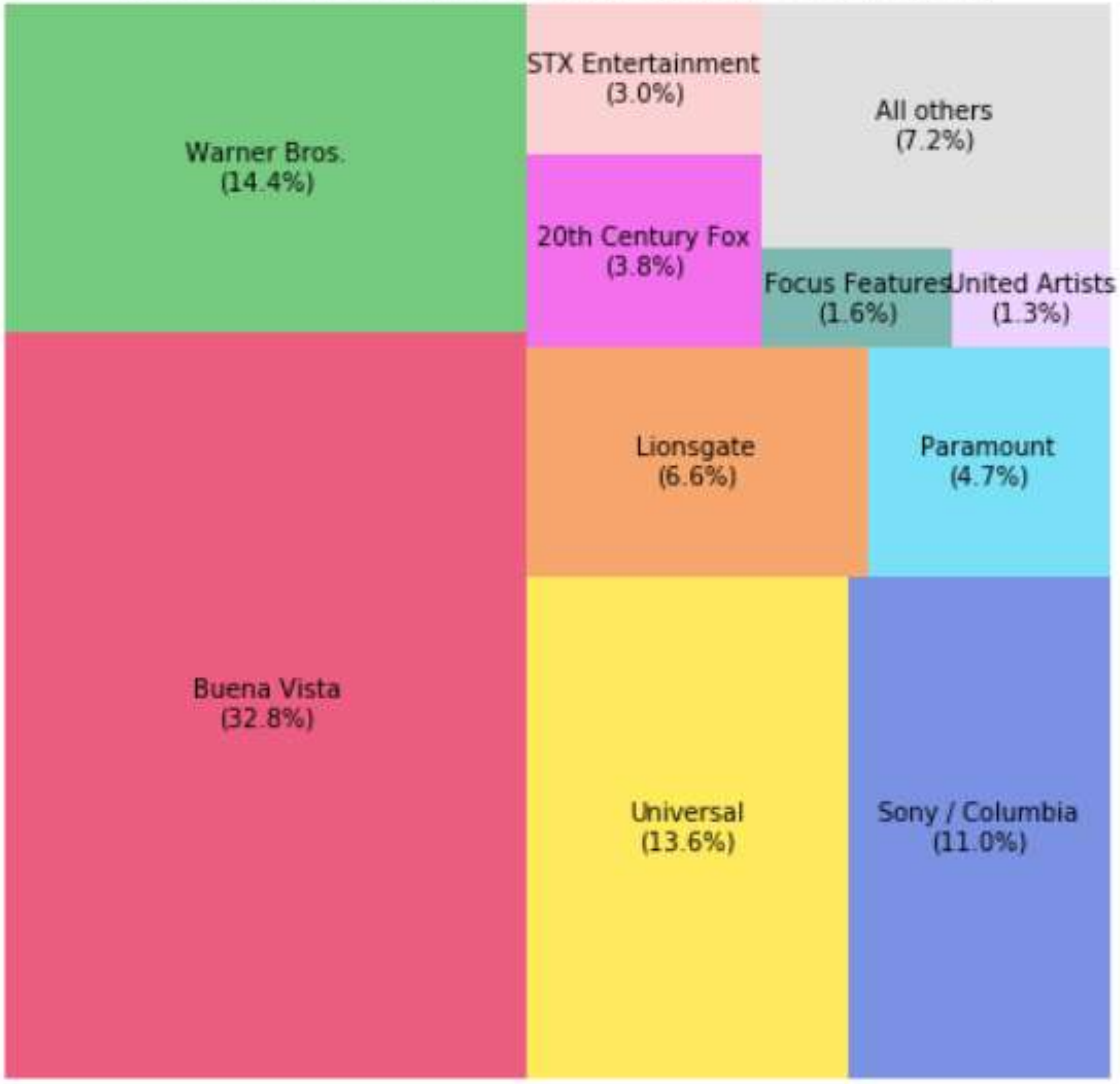
BAR CHART



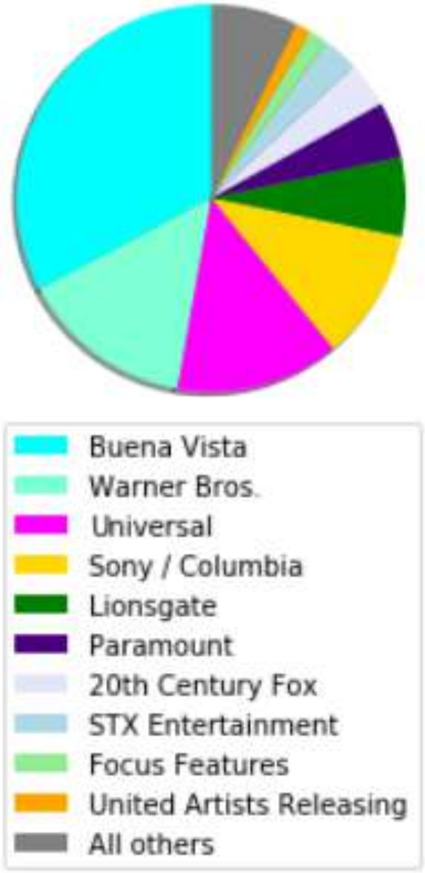
AREA PLOTS: TREE MAP

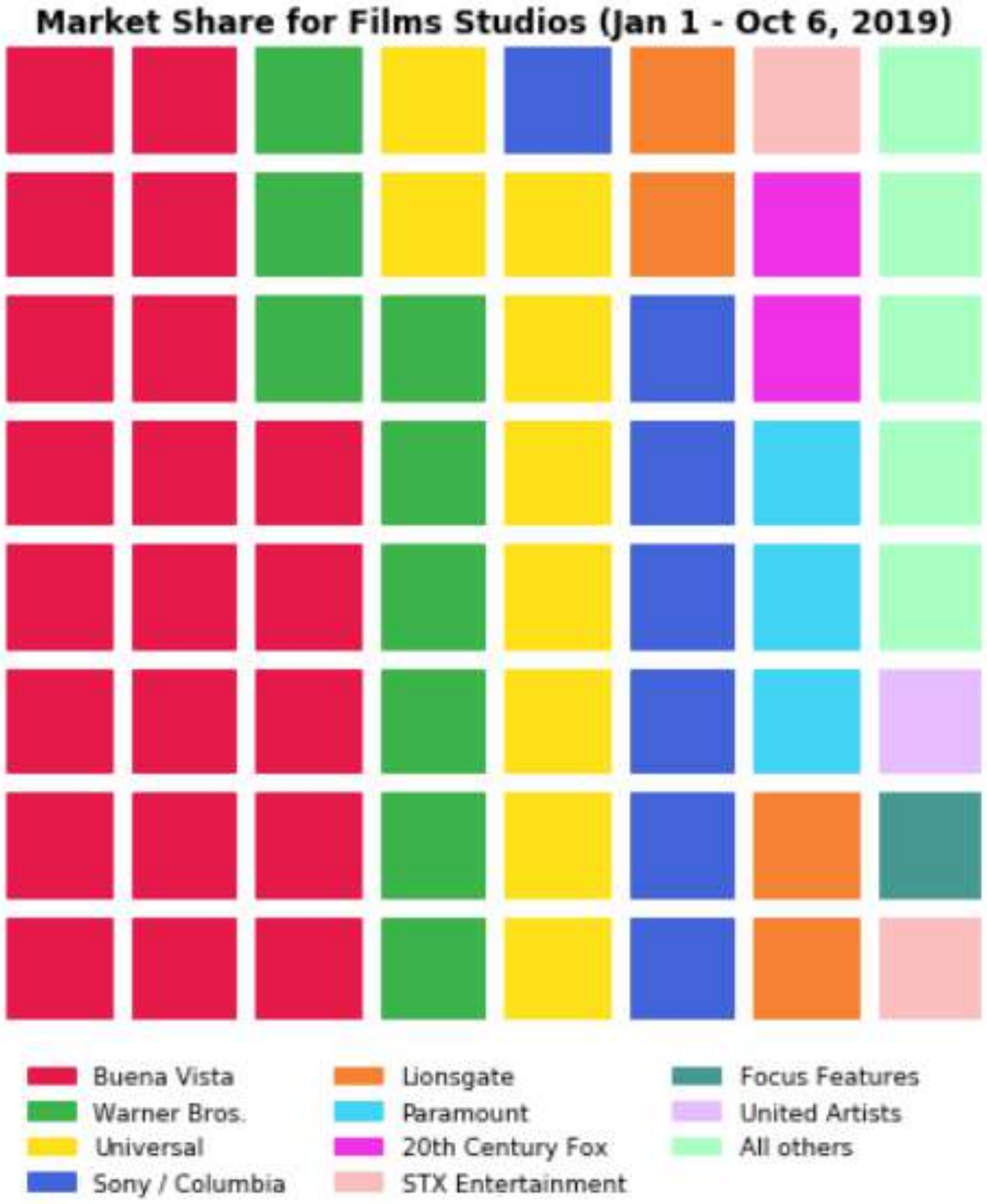


Market Share for Films Studios (Jan 1 - Oct 6, 2019)

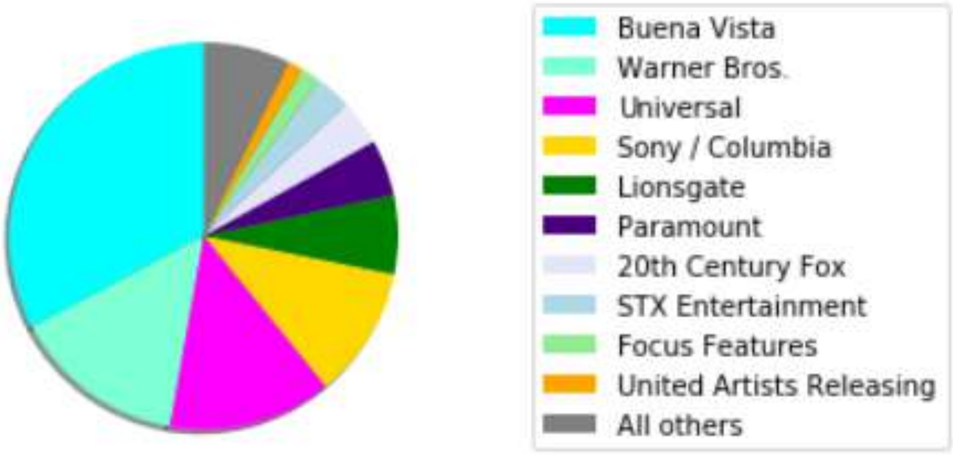


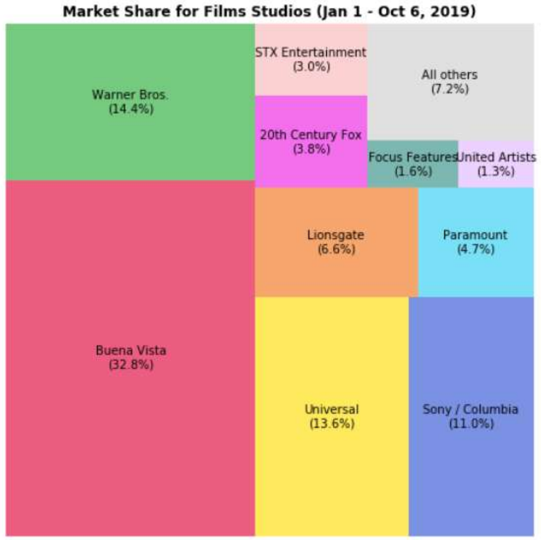
AREA PLOTS: WAFFLE CHART

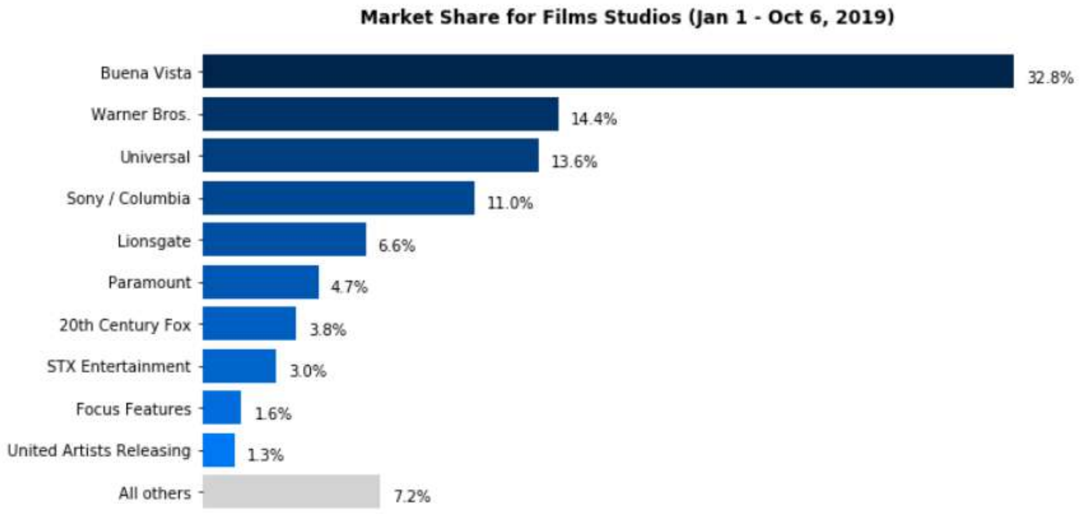


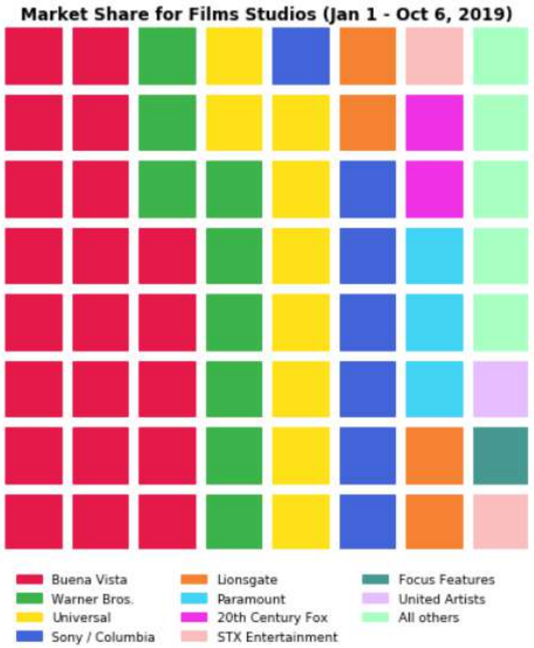

Market Share for Films Studios (Jan 1 - Oct 6, 2019)

So which visualisation was best?







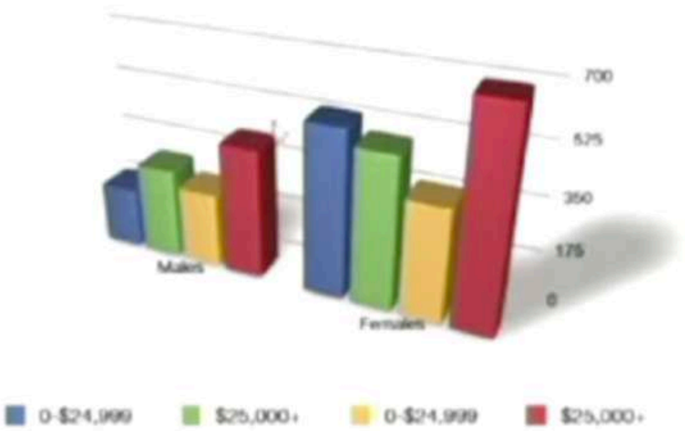


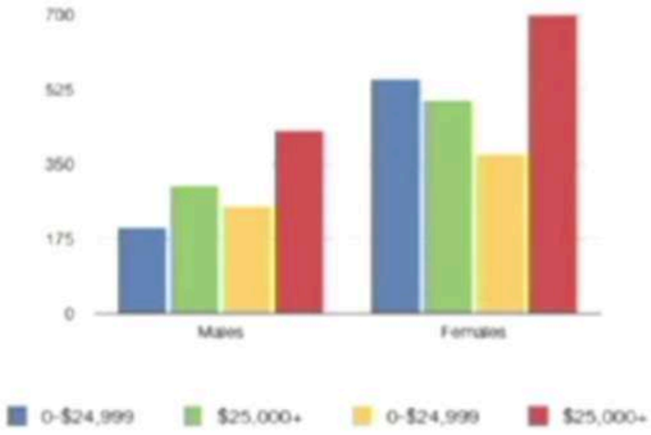
The good, the bad, & the ~~ugly~~ *misleading*

Tufte’s Graphical Theory

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

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

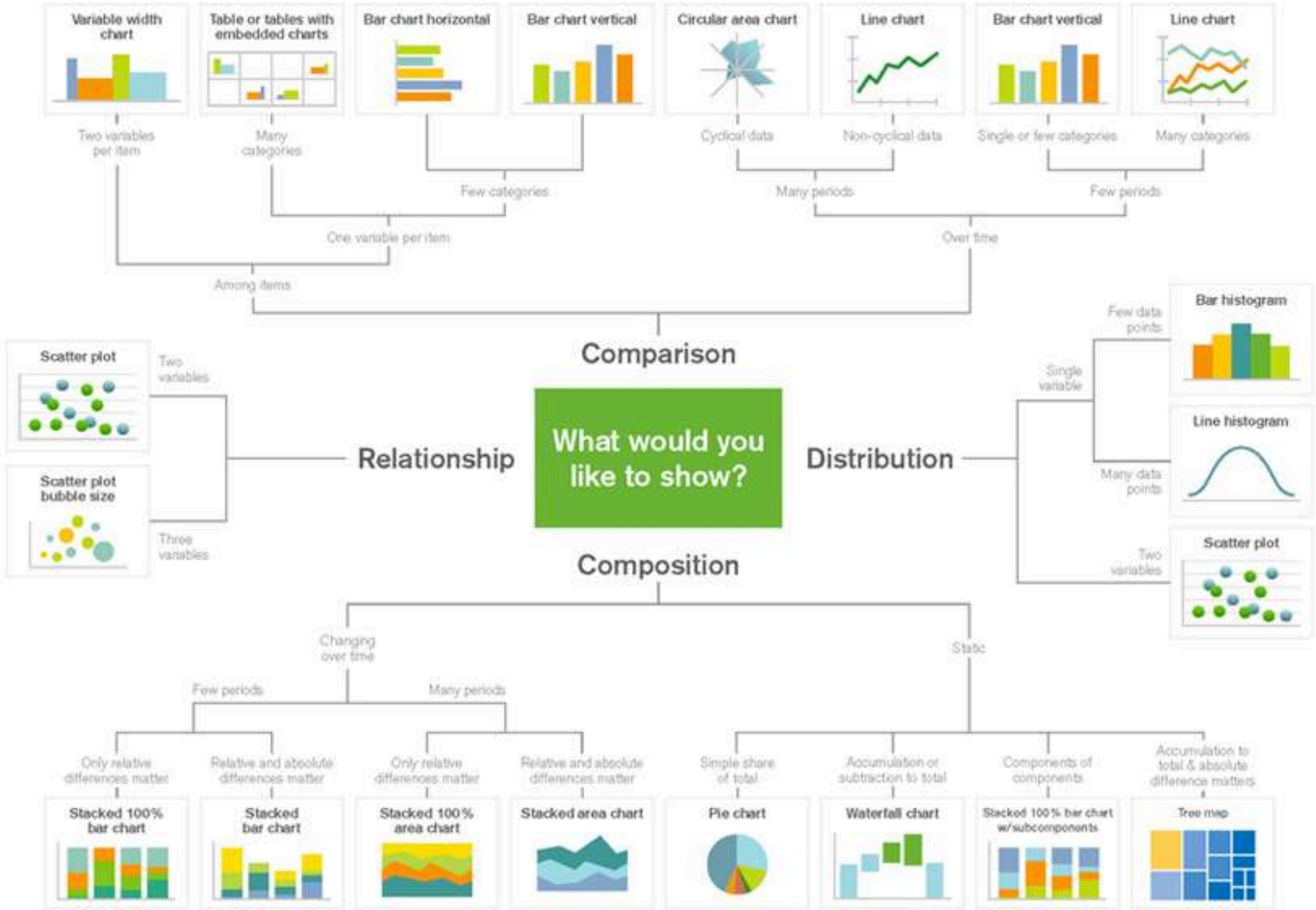




Outline

- **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**

How to choose the right plot?

Initial Data vs Final Result

HISTOGRAMS

BOX-PLOT

SCATTER PLOT

MOSAIC PLOT

RAIN-DROP

VIOLIN PLOT

PIE CHARTS

SPIDER PLOT / RADAR CHART

CIRCOS PLOT

STREAMGRAPH

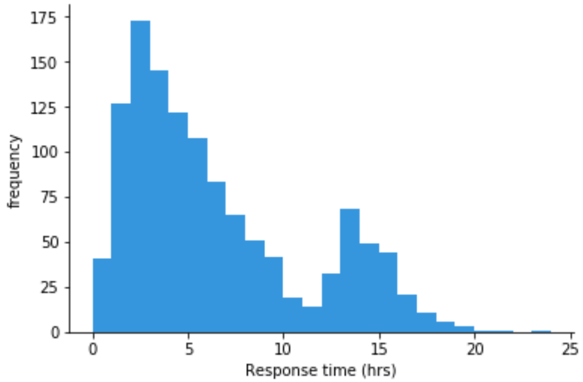
FUNNEL PLOT

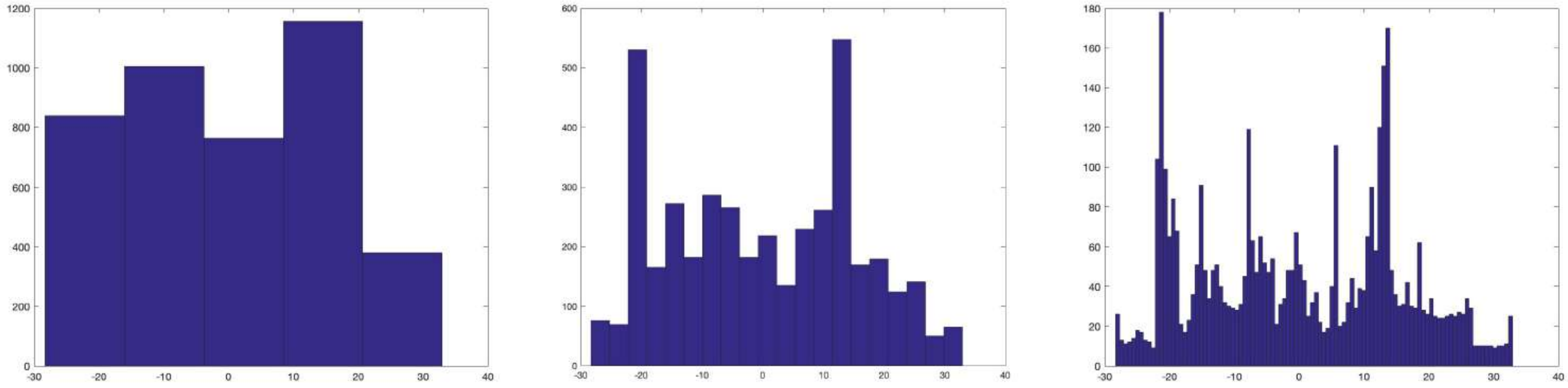
not an exhaustive list

some plots used for both

Histograms

- data distribution
  - spread and shape of data
  - bin-width dependancy
  - may indicate presence of groups

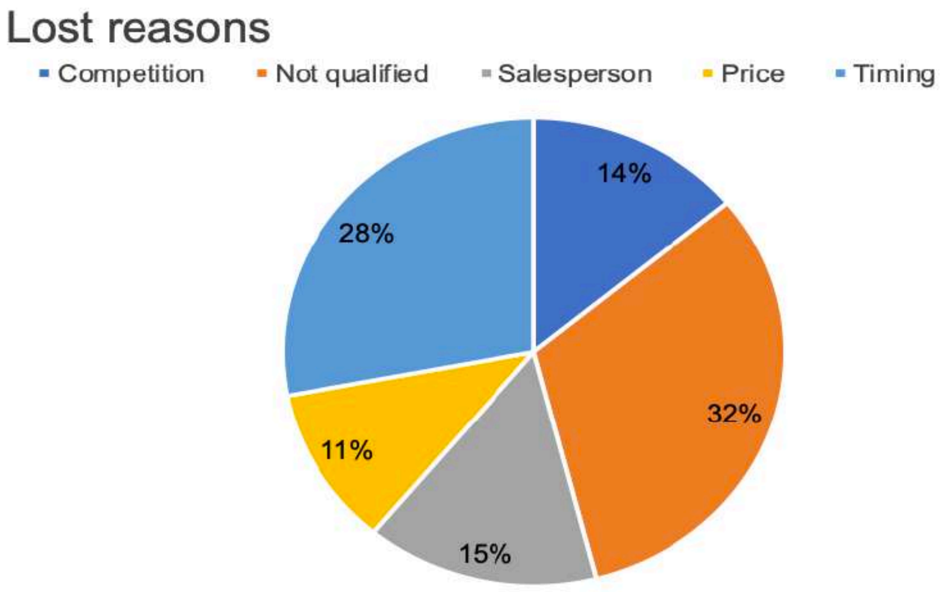




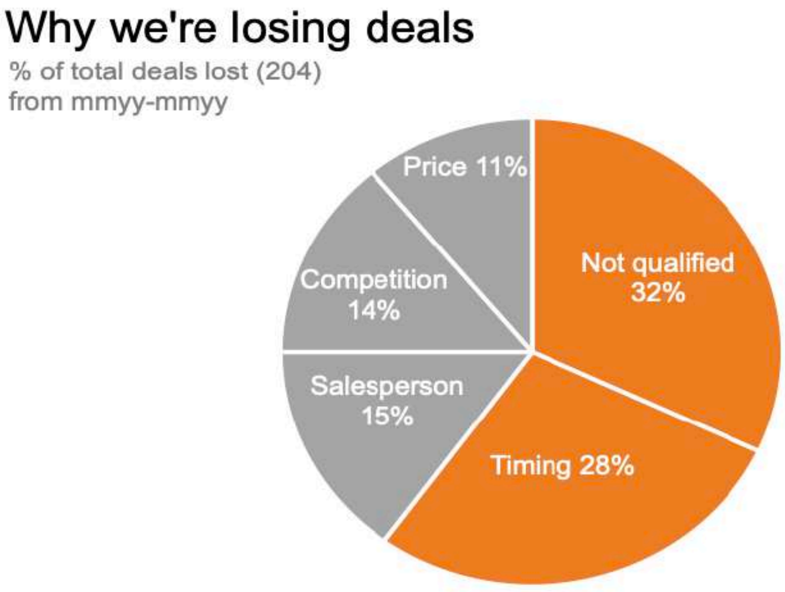
Pie Charts

- 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%

Pie Charts



Lost reasons



Why we're losing deals

% of total deals lost (204)
from mmyy-mmyy

60% of deals were lost because we **didn't qualify appropriately** or timing was **outside of the customers' budgeting cycle.**

How might we improve our process?

which is easier to read?

Bar Charts

- use bar charts when
  - have moderate no. of categories (not too many)
  - need to compare numbers side-by-side
  - caution: more than two bars are hard for readers

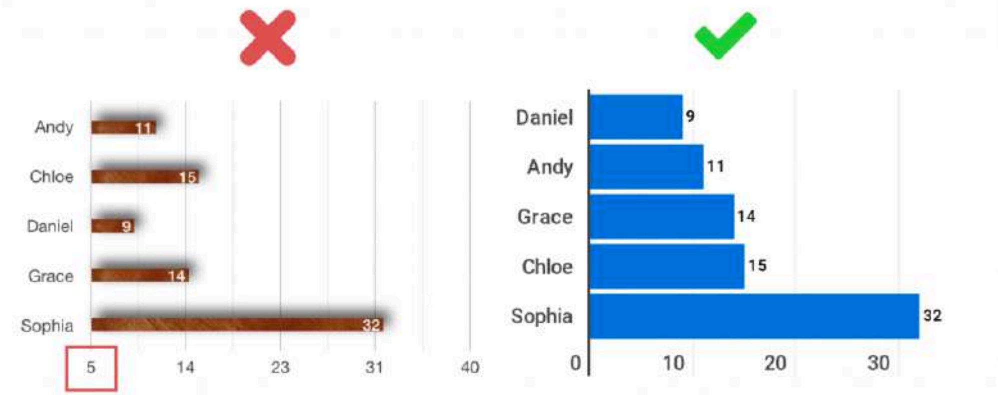
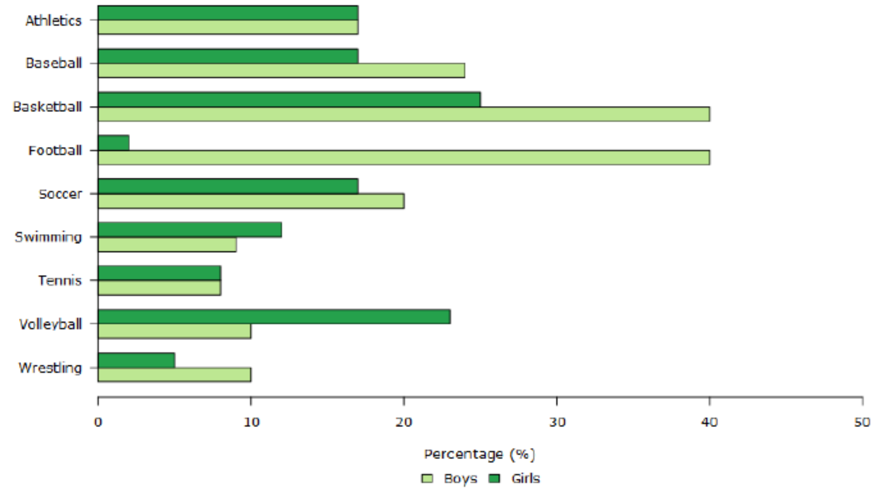
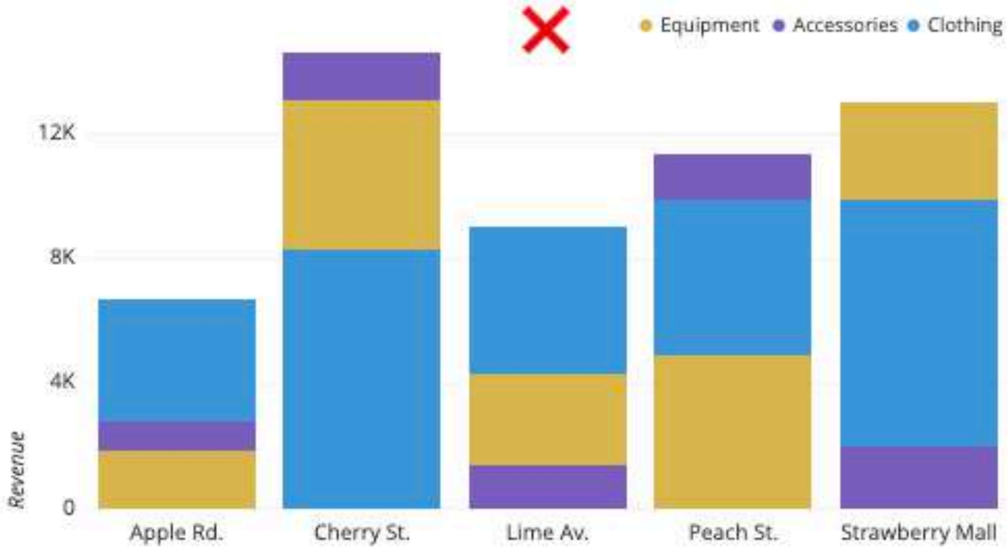
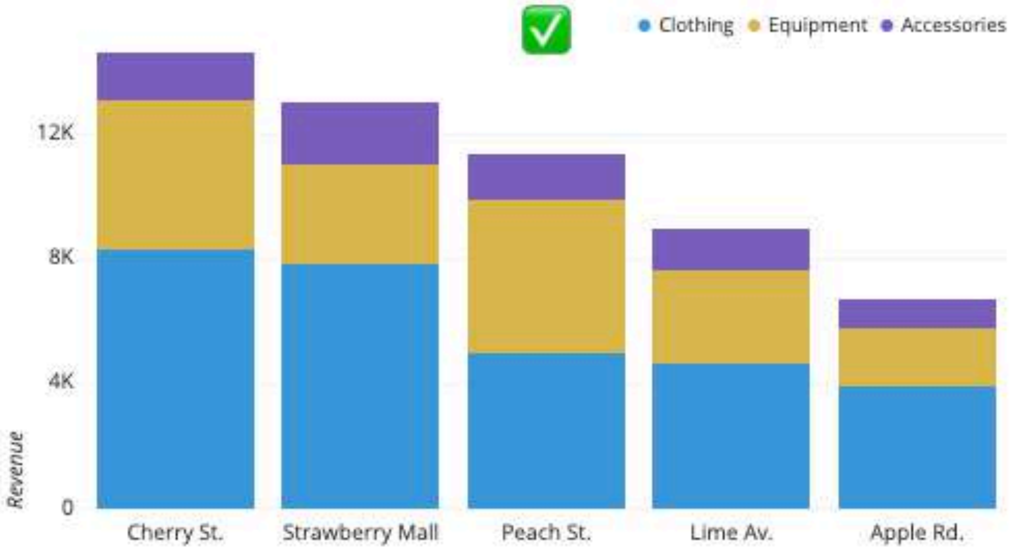


Chart 5.2.3
Sports practiced by 15-year-old students in Jamie's school, by gender

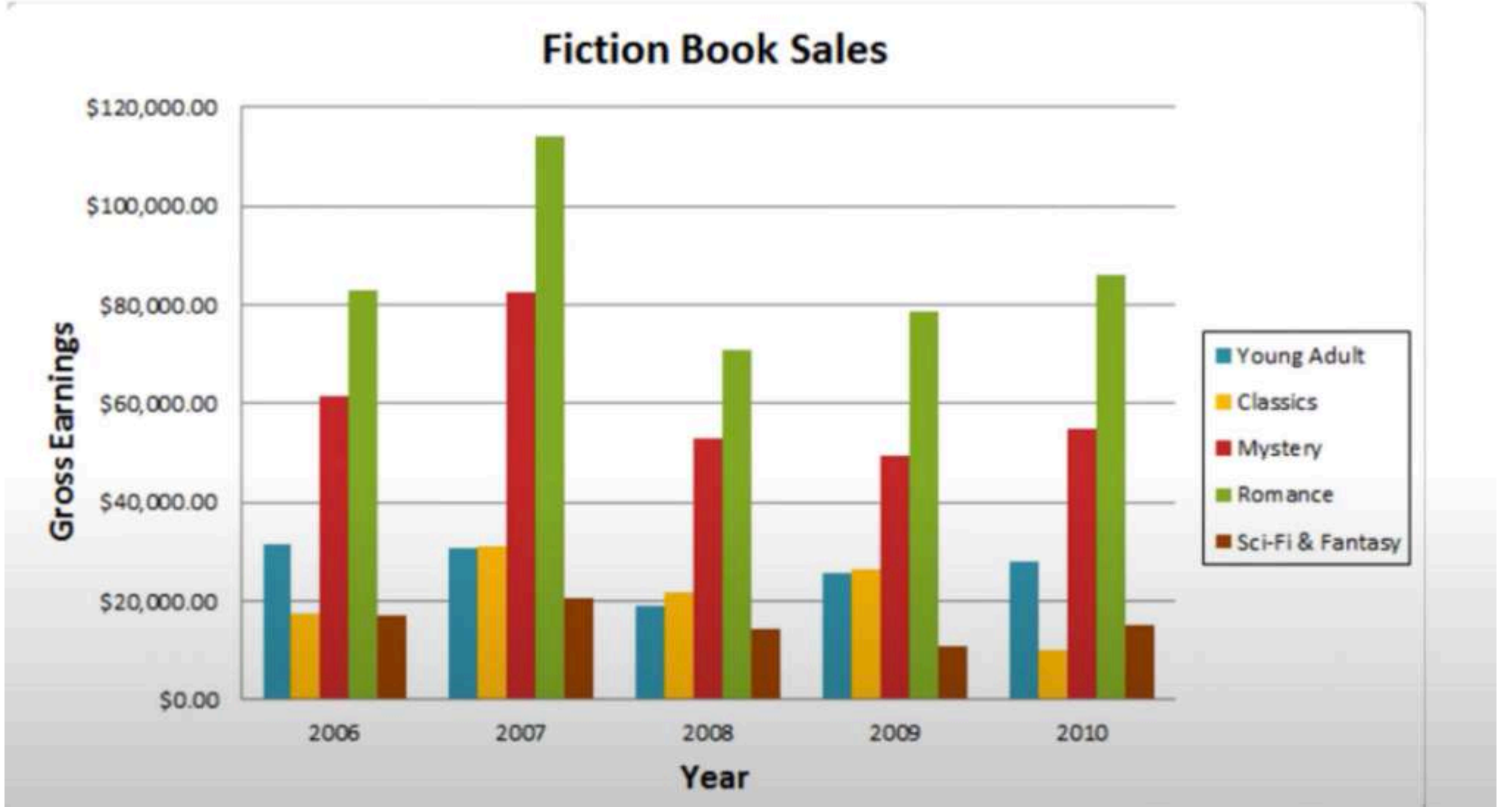


Bar Charts


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


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

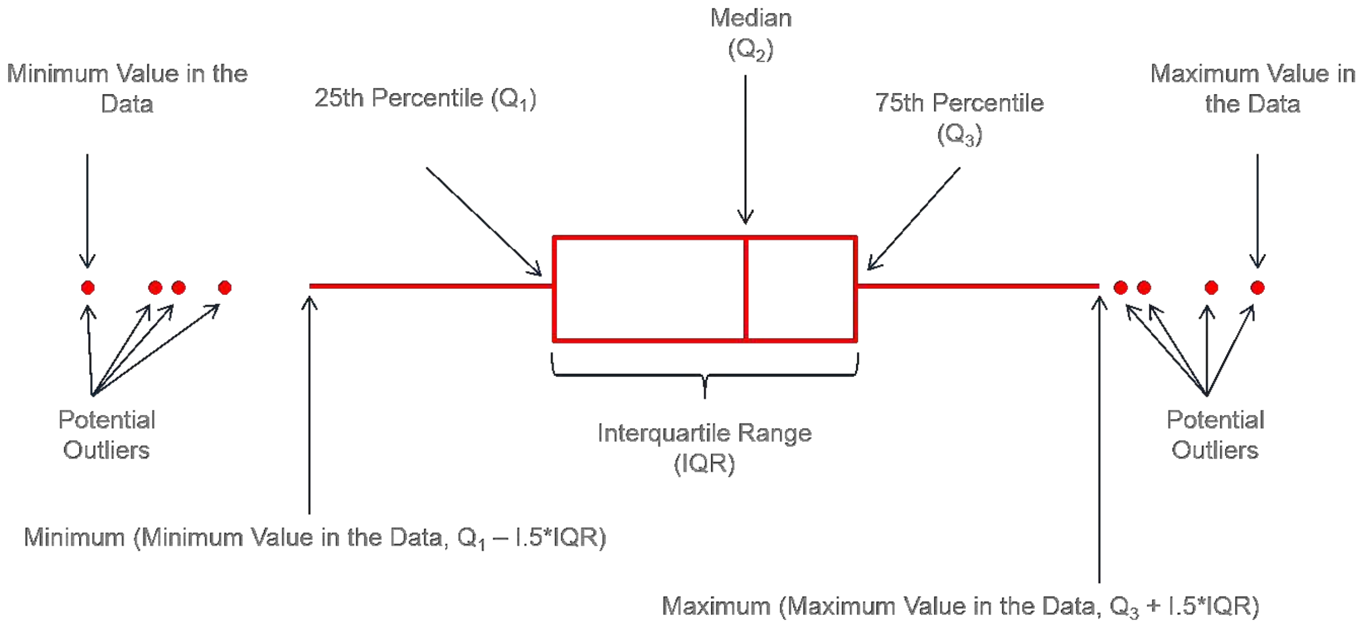
Bar Charts



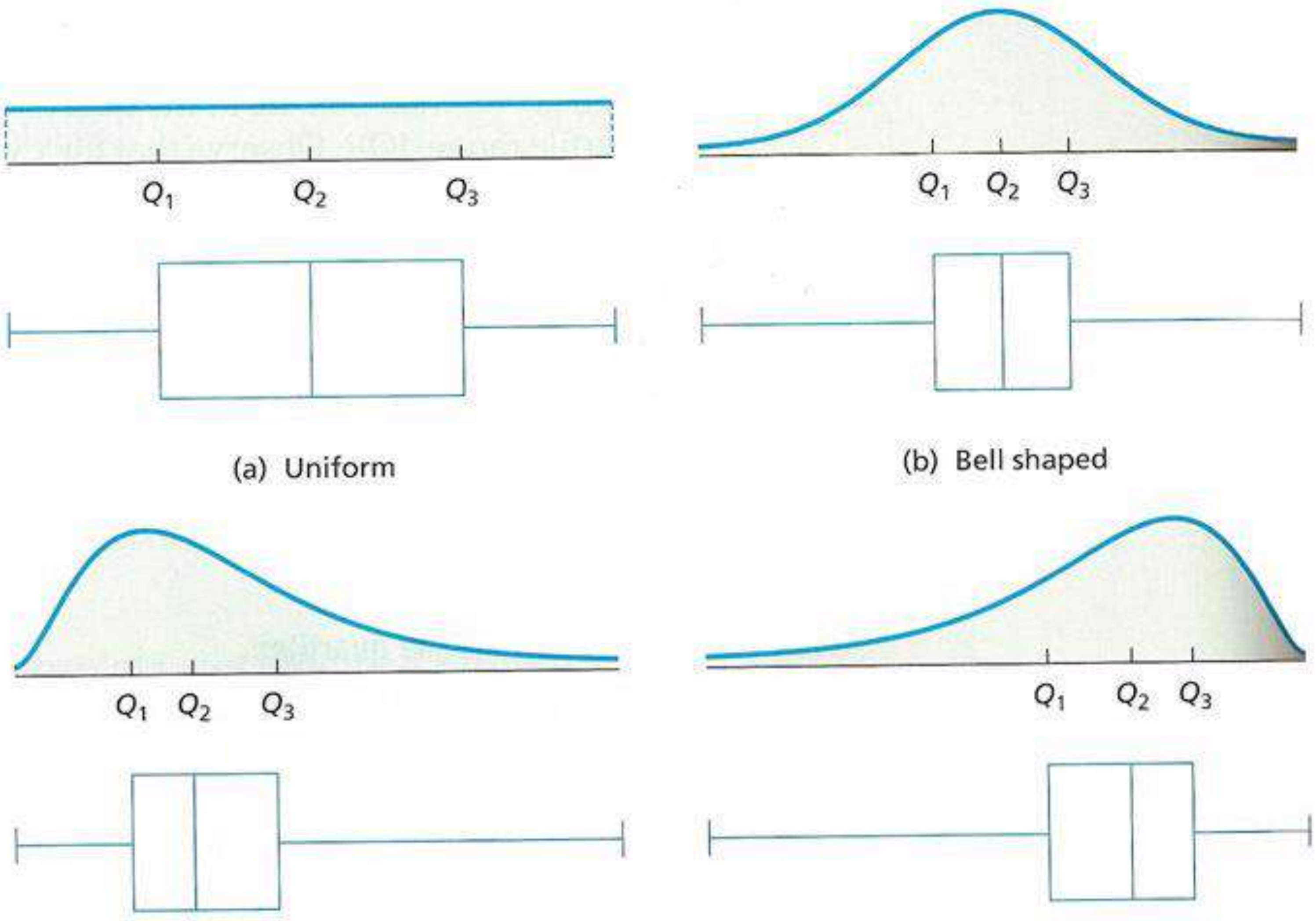
is this ok? what point can we make?

Box plots

- locality and spread of data
- useful for group differences

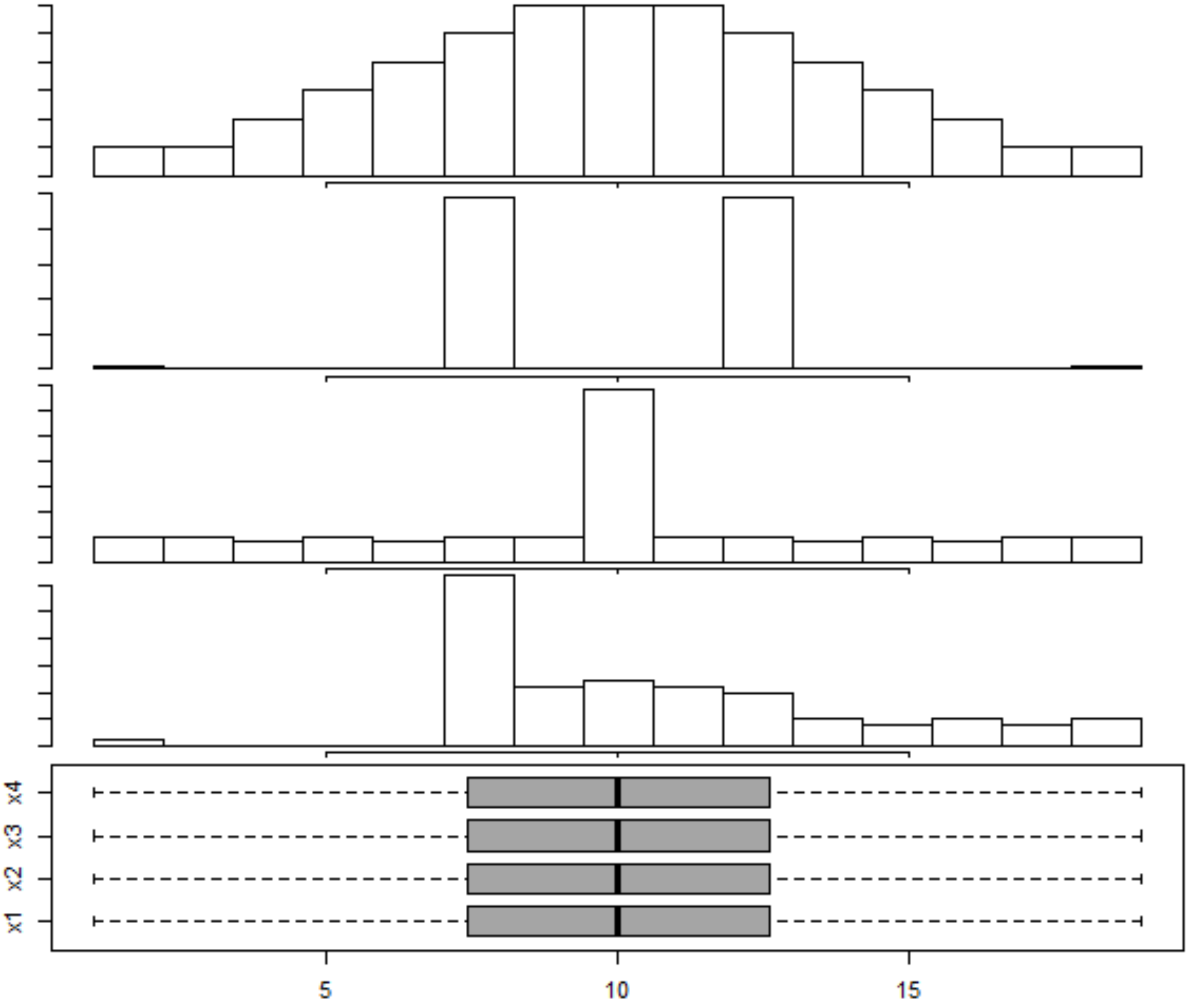

Median
(Q_2)
Minimum Value in the Data
25th Percentile (Q_1)
75th Percentile (Q_3)
Maximum Value in the Data
Potential Outliers
Interquartile Range (IQR)
Potential Outliers
Minimum (Minimum Value in the Data, Q_1 – l.5*IQR)
Maximum (Maximum Value in the Data, Q_3 + l.5*IQR)

Boxplot



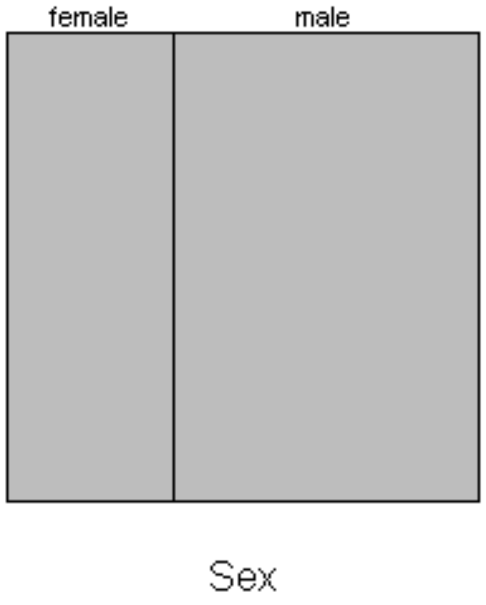
(a) Uniform

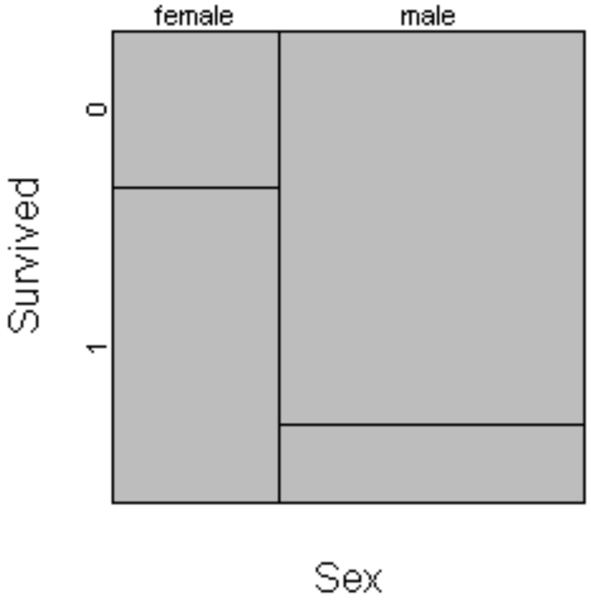
(b) Bell shaped

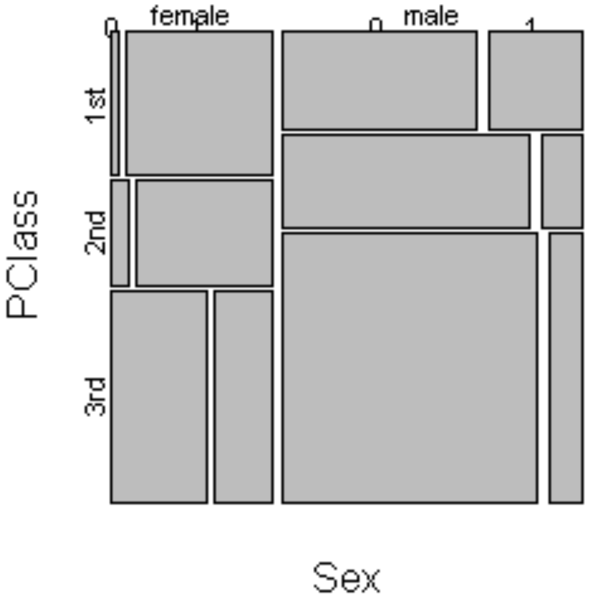

x4
x3
x2
x1
5
10
15

Mosaic Plots

- allows you to observe the relation among two or more categorical variables

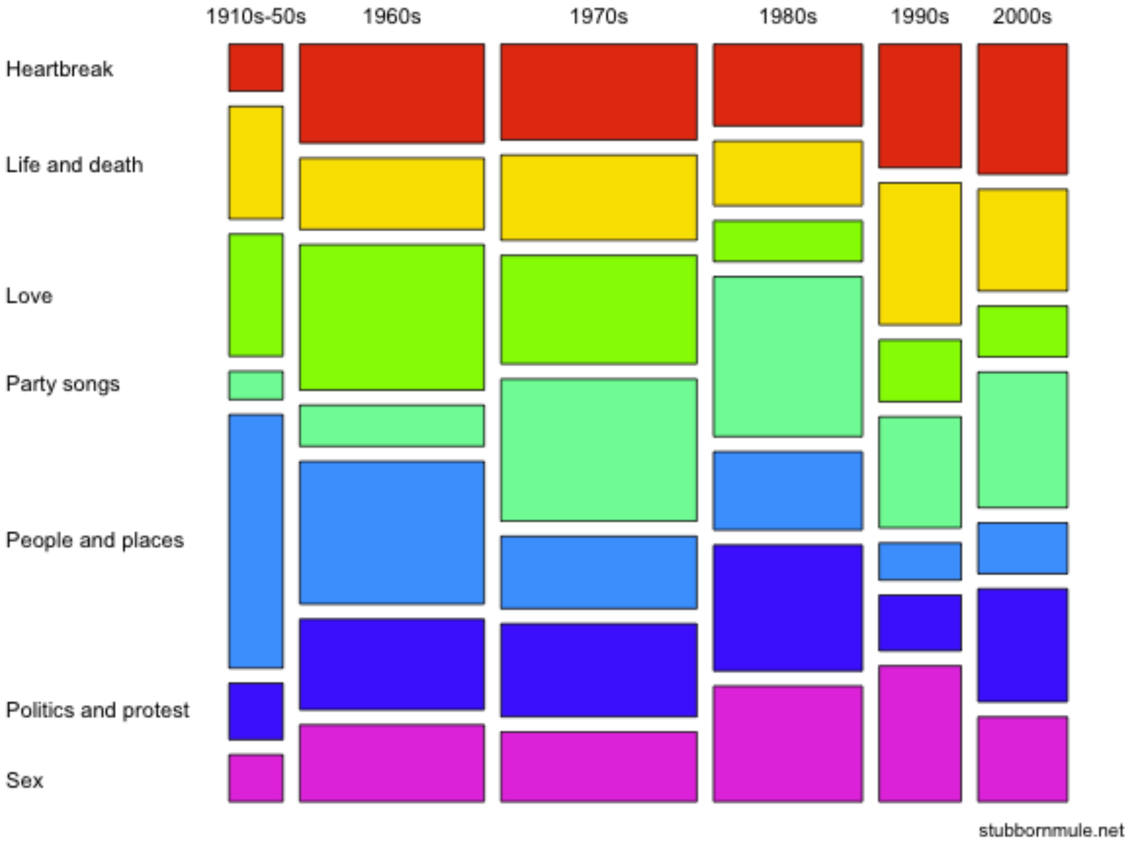






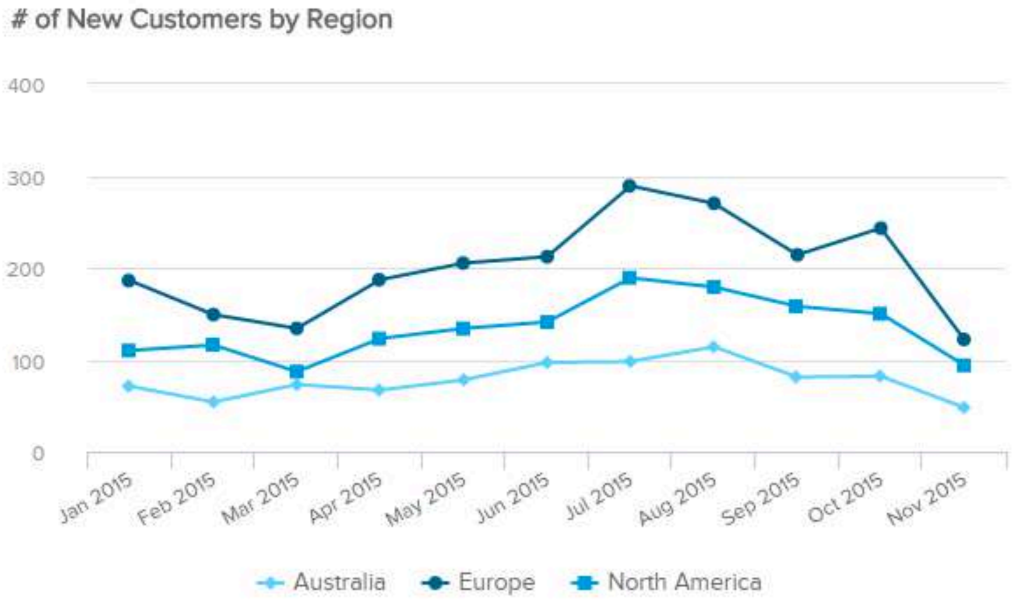
Mosaic Plots

- allows you to observe the relation among two or more categorical variables



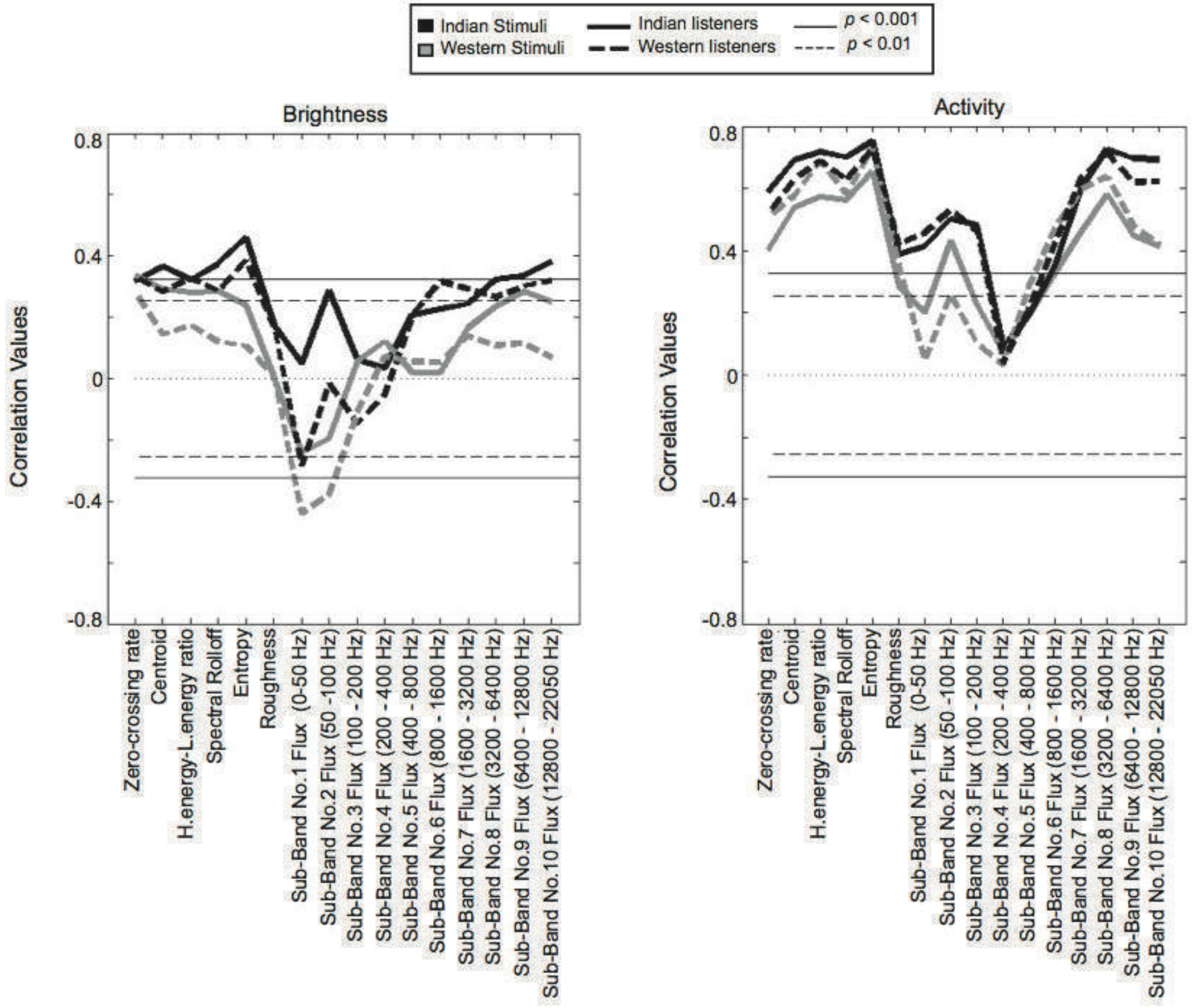
Line Charts

- use line charts
  - data is continuous
  - track development of variables over time
    - ex: stacked line chart

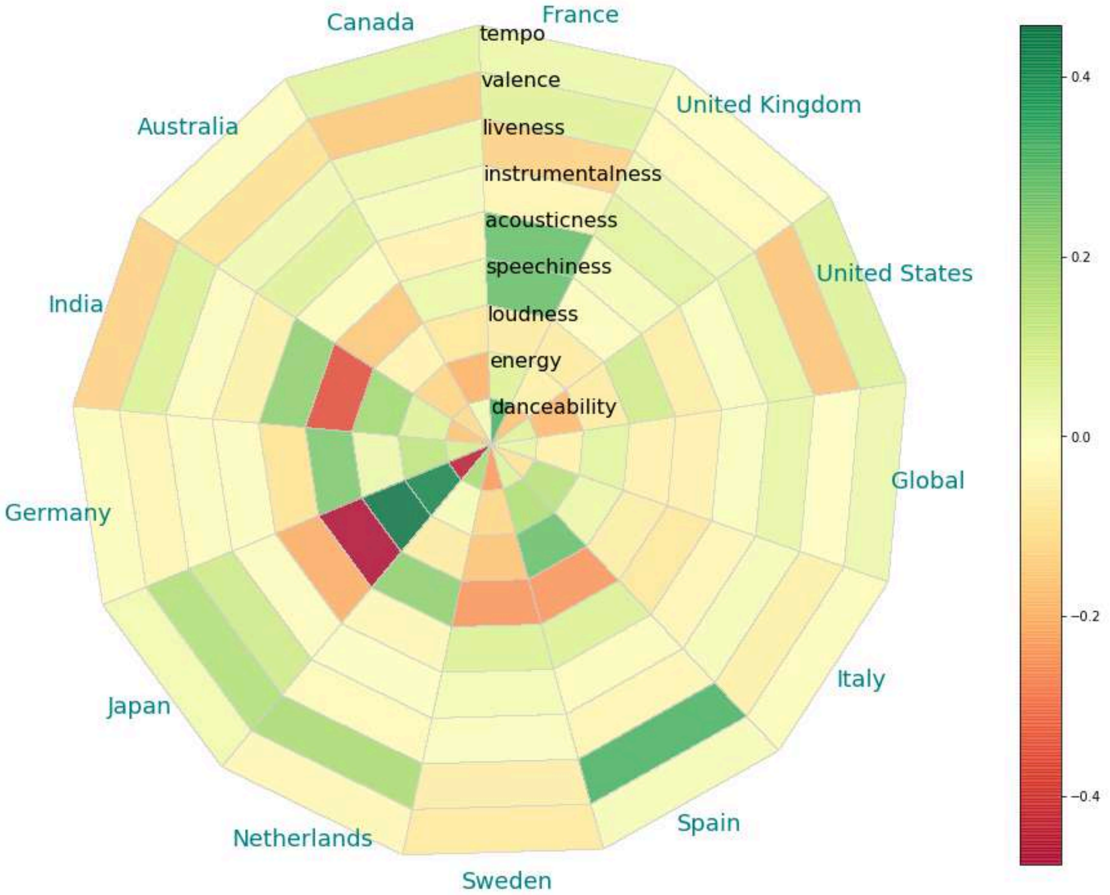


Line Charts

- appropriate for non-temporal data?

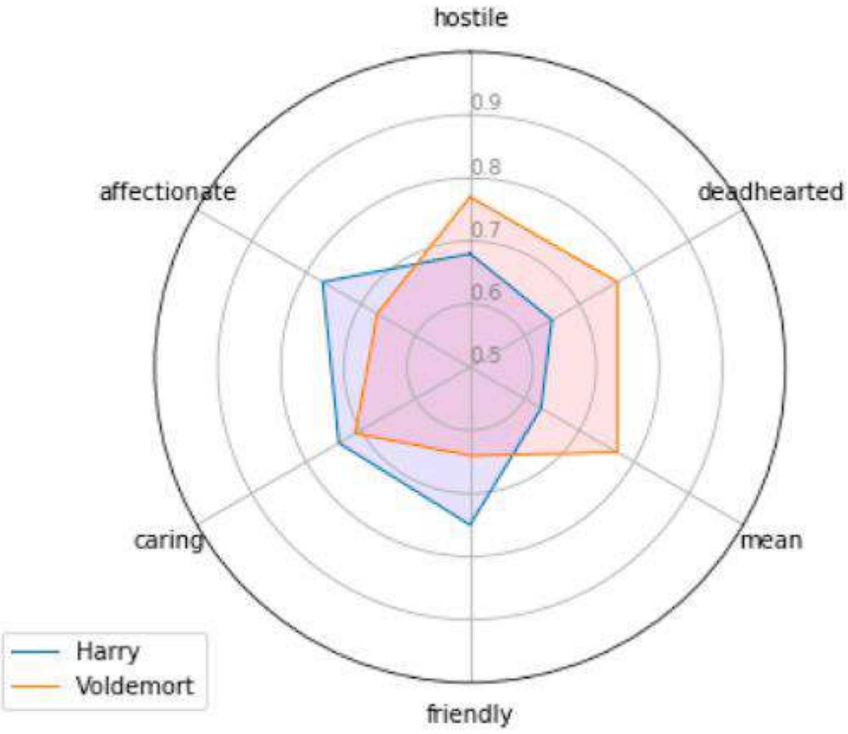


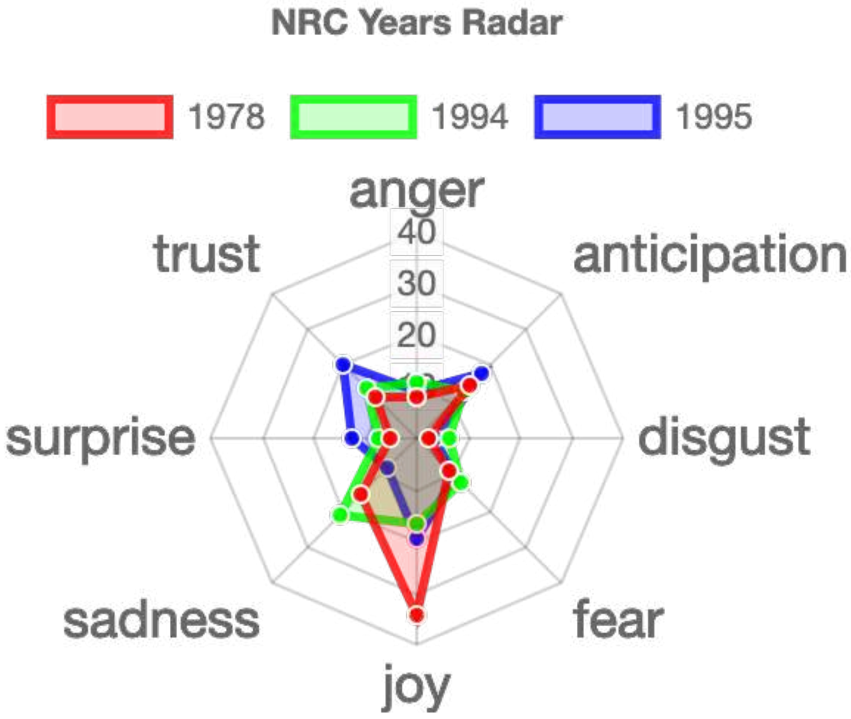
Describing Data



Radial Heat Map

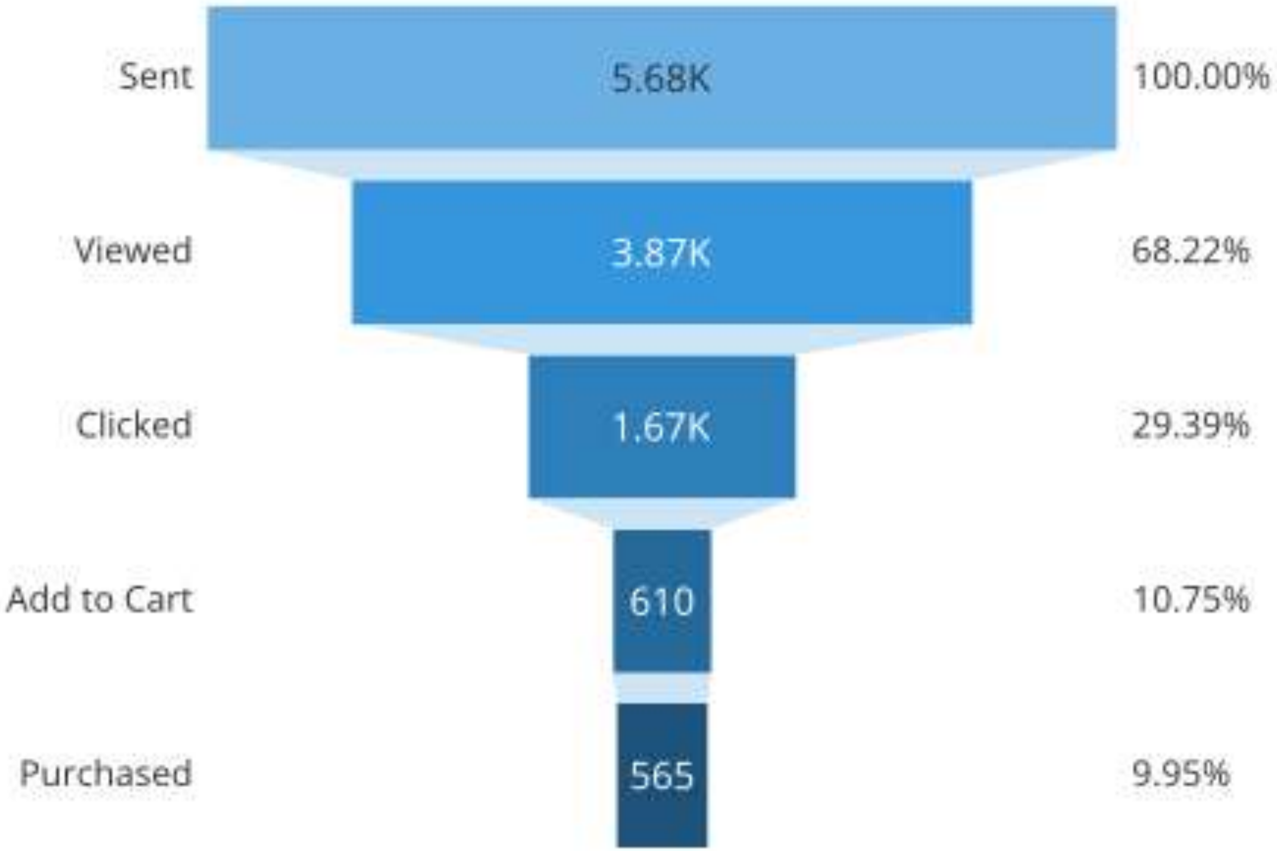
Visualizing Results





Spider/Radar Plots

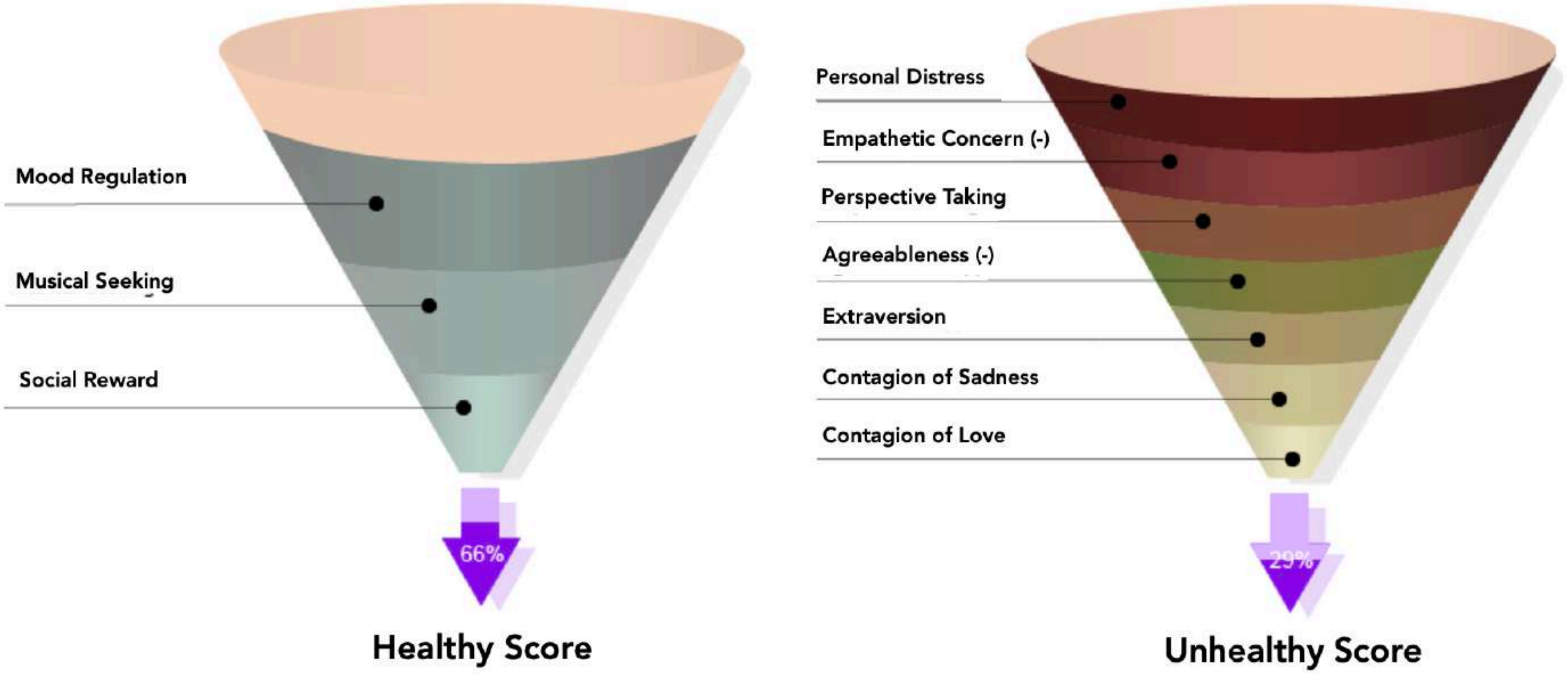
Describing Data



Funnel Charts

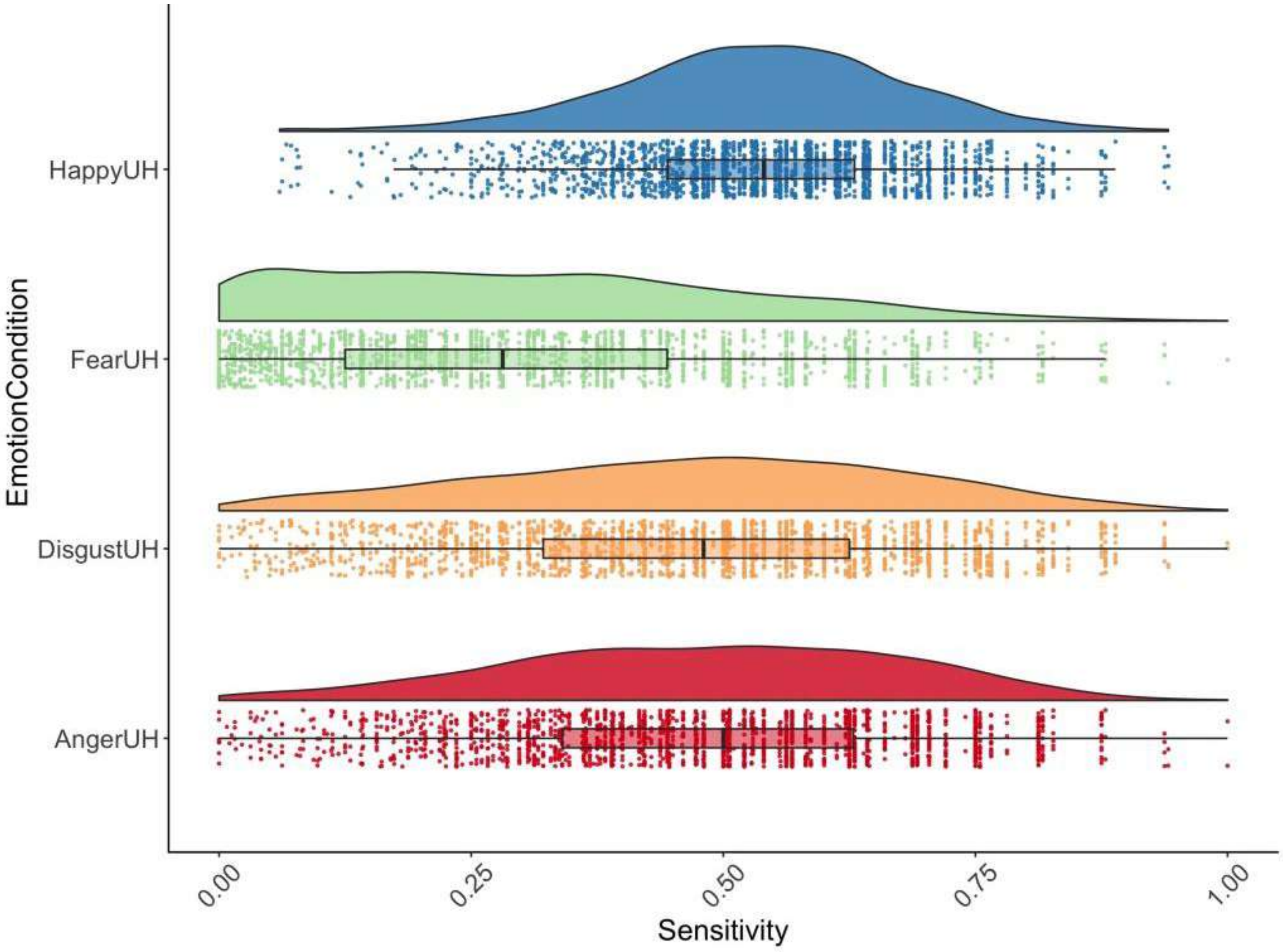
ex:responses to a fictional email campaign regarding a special product offer. represents five stages of the pipeline are associated with a bar whose length corresponds with the number of users that completed each stage

Visualizing Results



Funnel Charts
(ex: Regression Results)

Visualizing Results



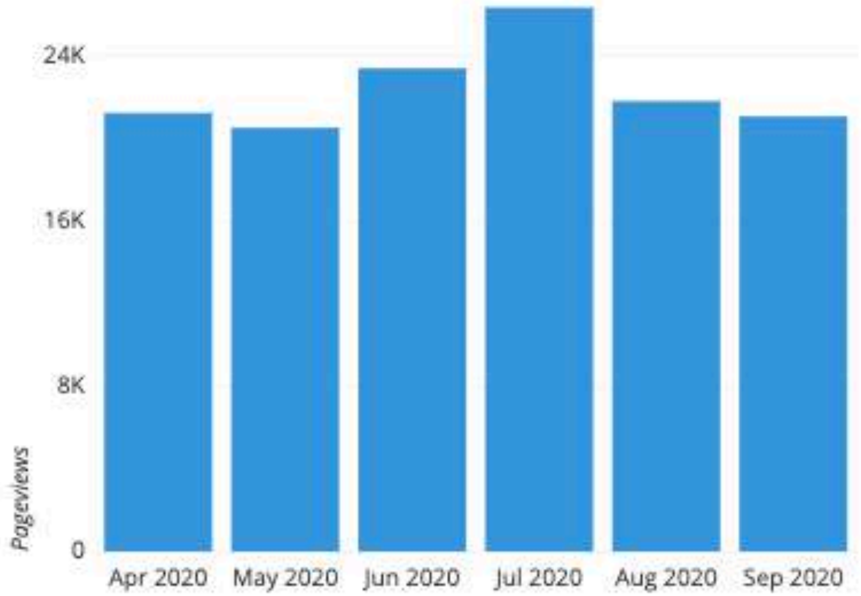
Raindrop plot

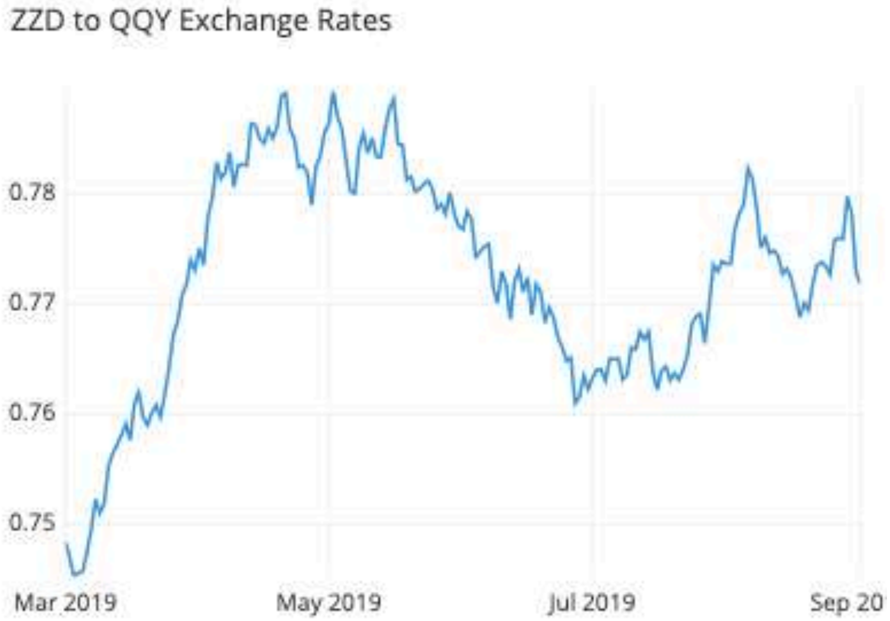
How to choose the right plot?

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

Temporal

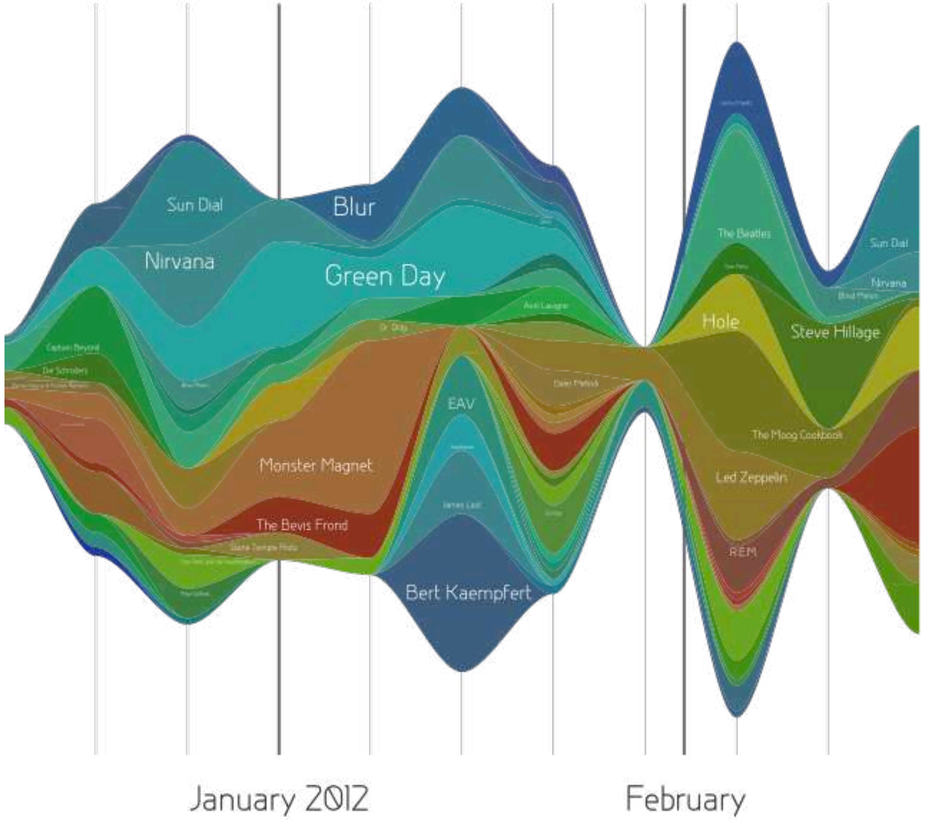
- showing change over time





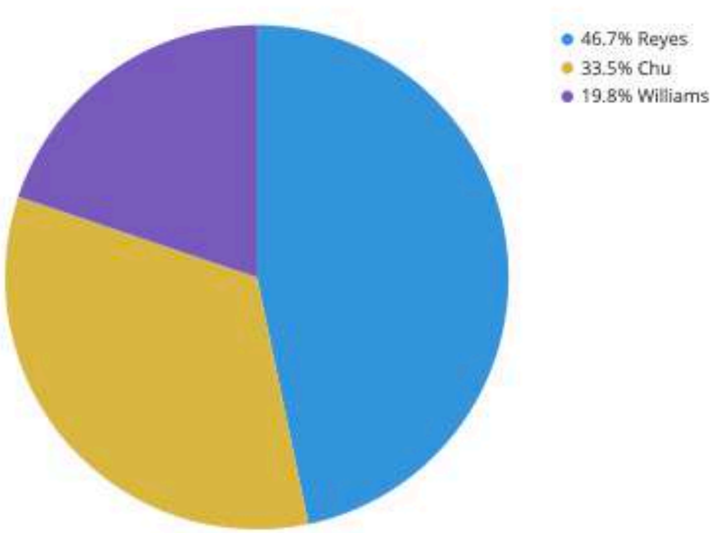
Temporal

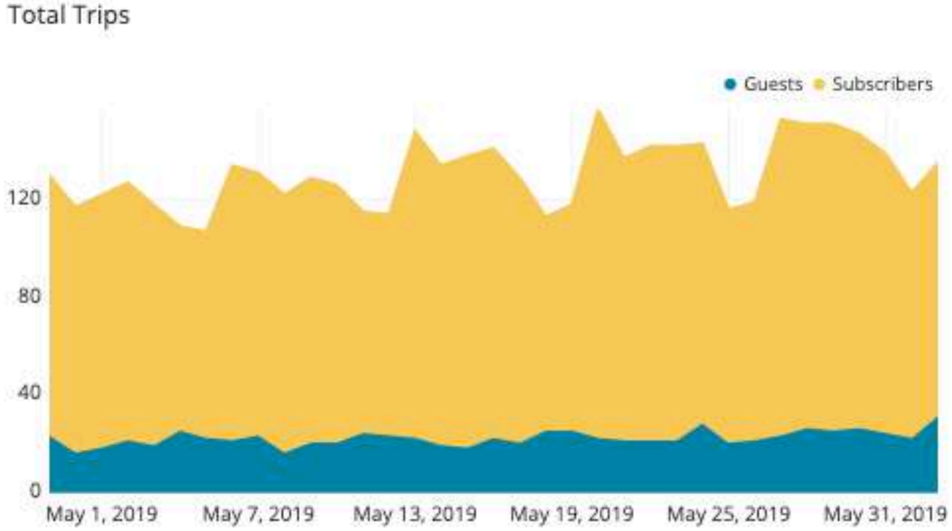
- showing change over time
  - eg: streamgraph (multiple variables)

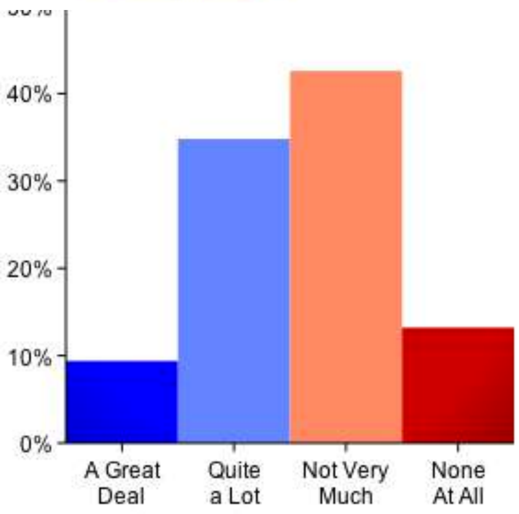


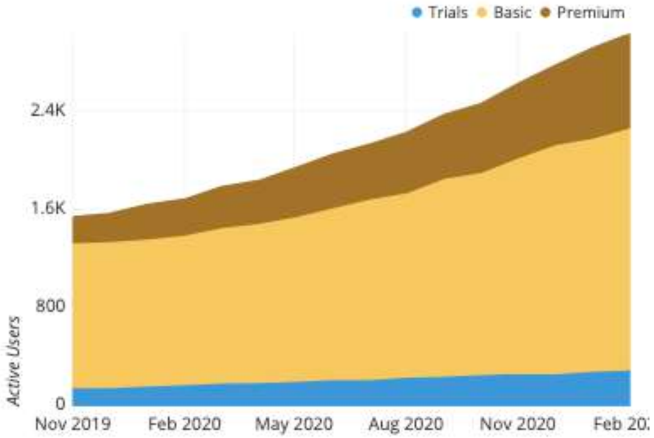
Proportions

- showing a part-to-whole composition

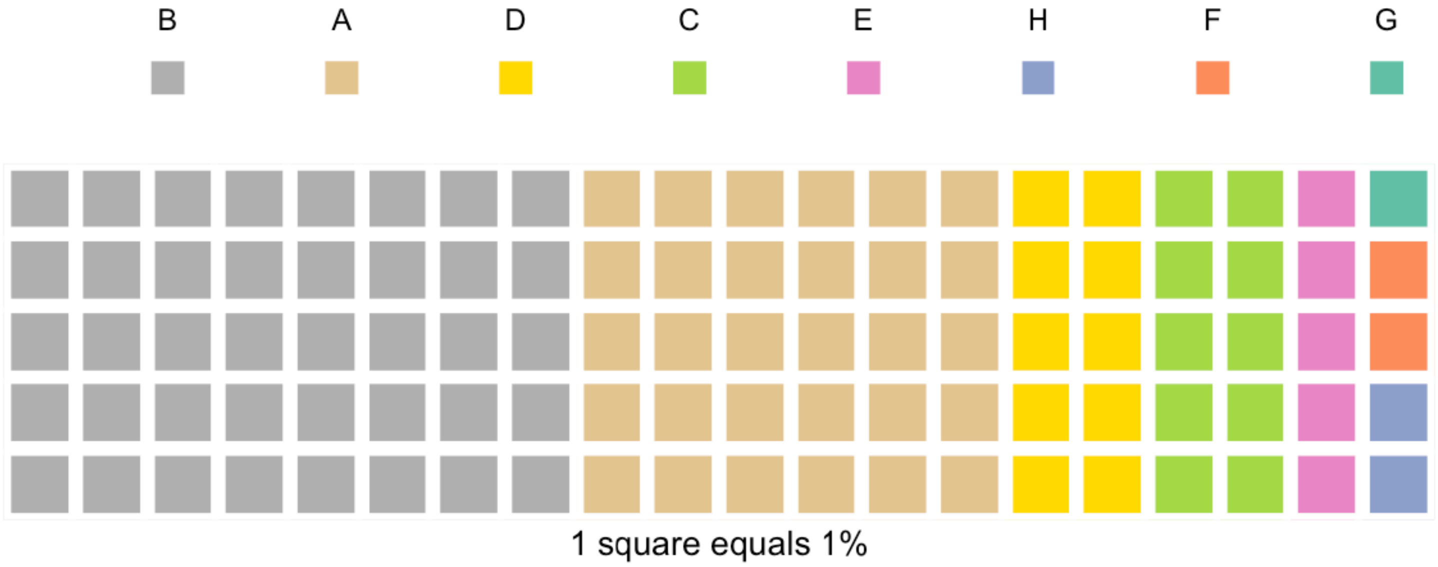


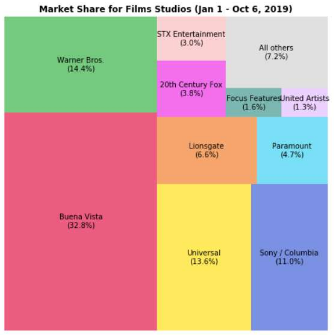






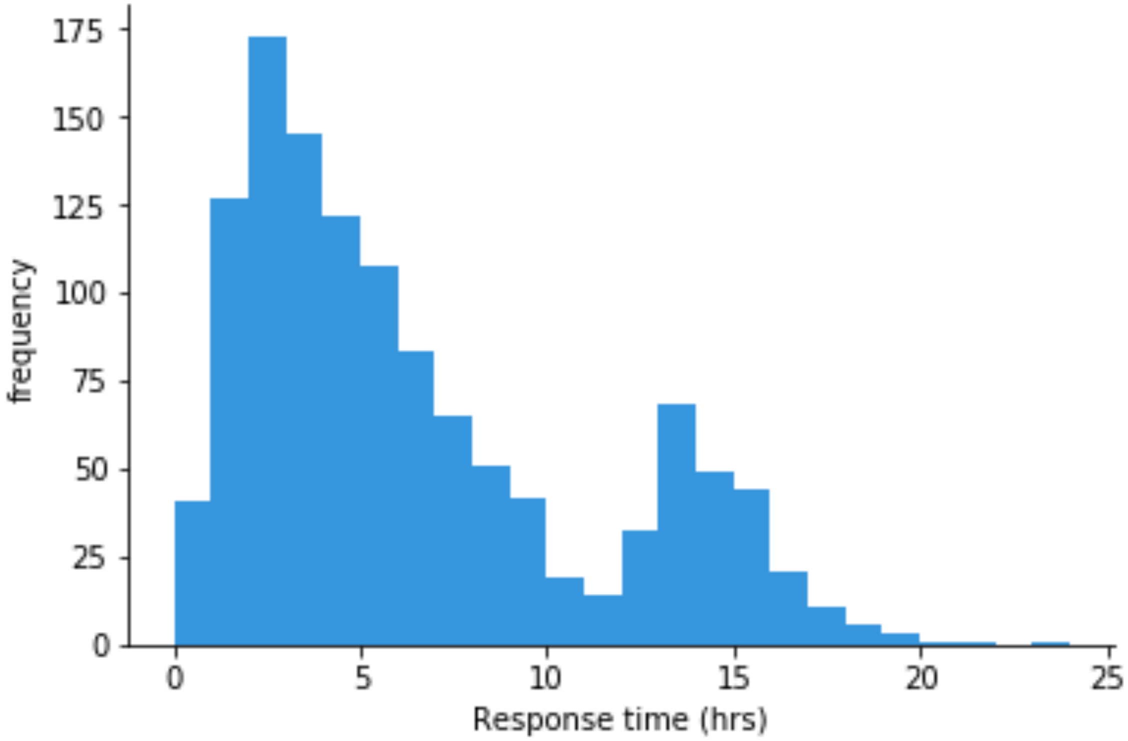
Proportions

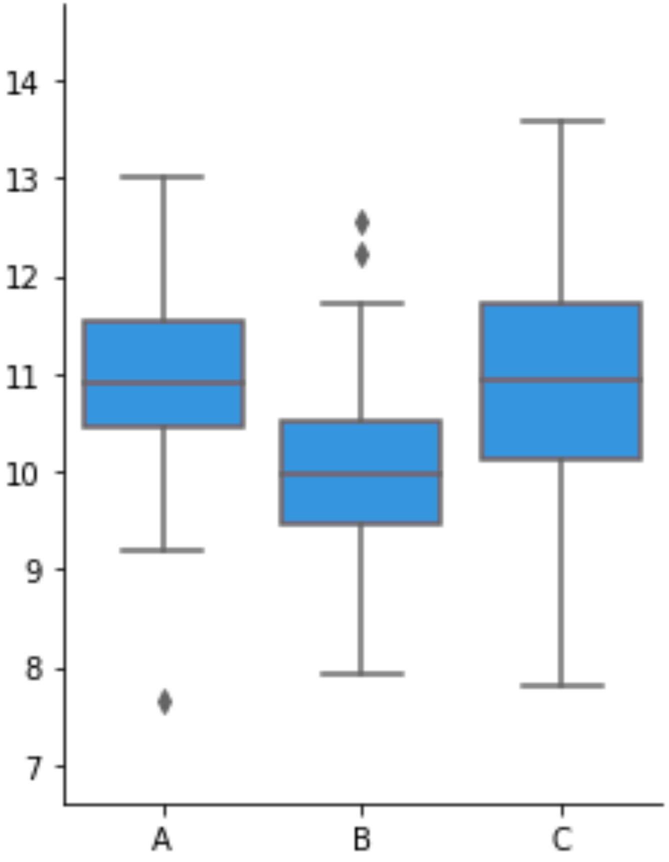




Area plots

Data Distribution

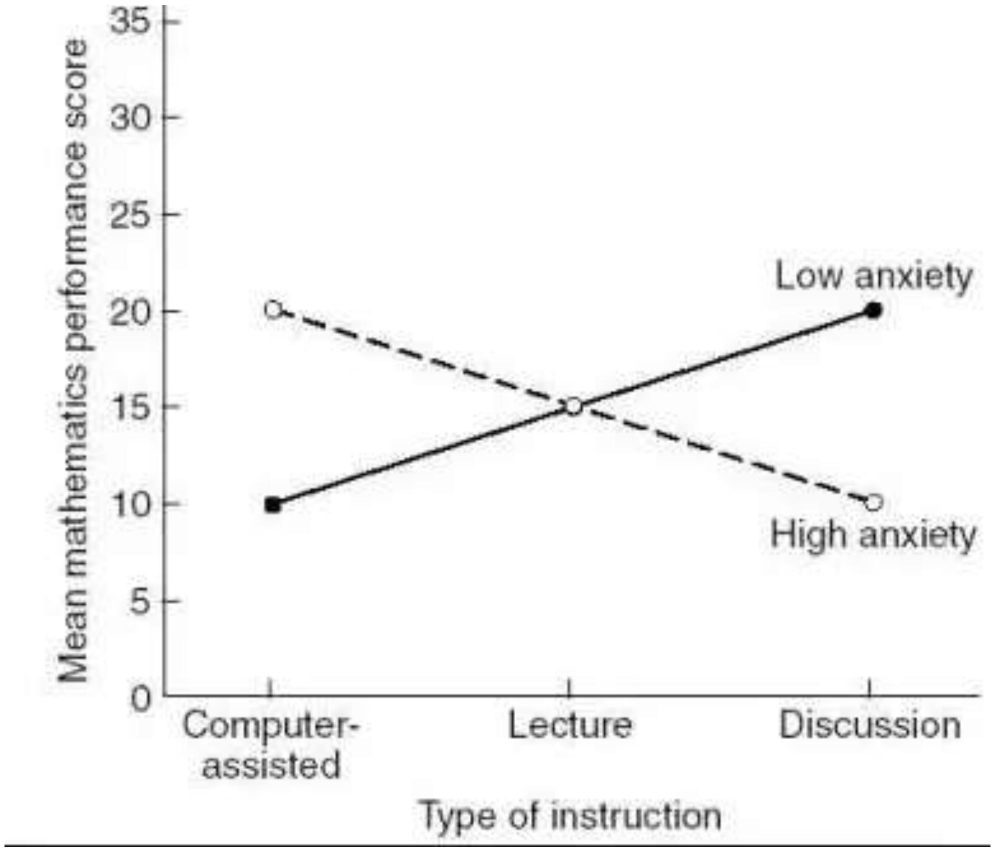




indicative of potential groups or group differences

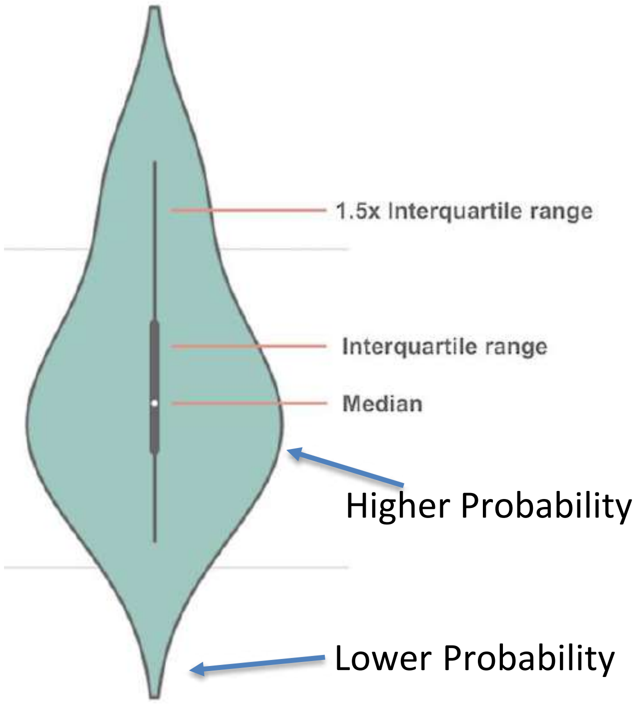
Group Differences

- main effects and interaction plots



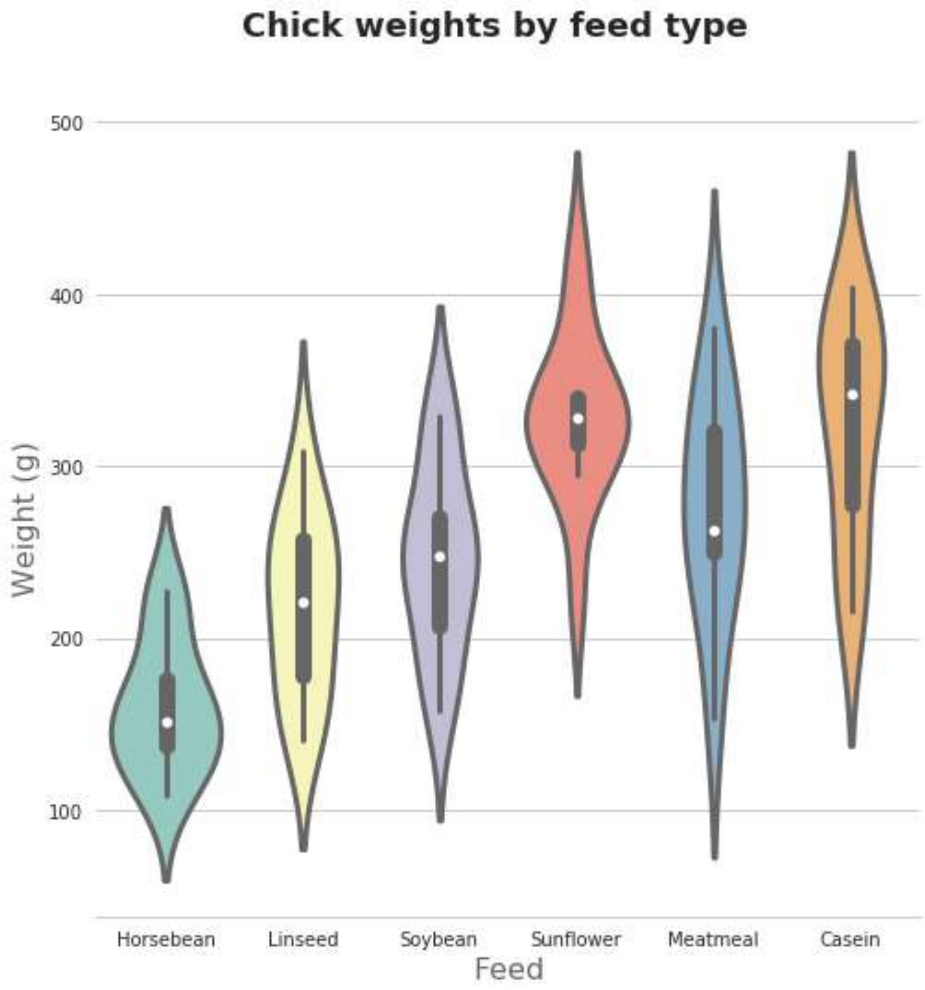
Group Differences

- violin plots



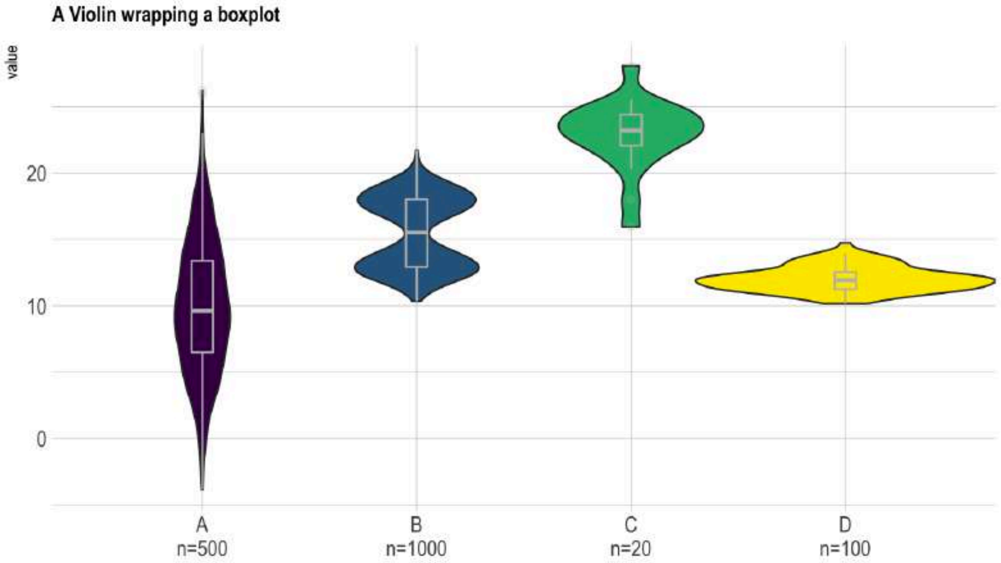
Higher Probability

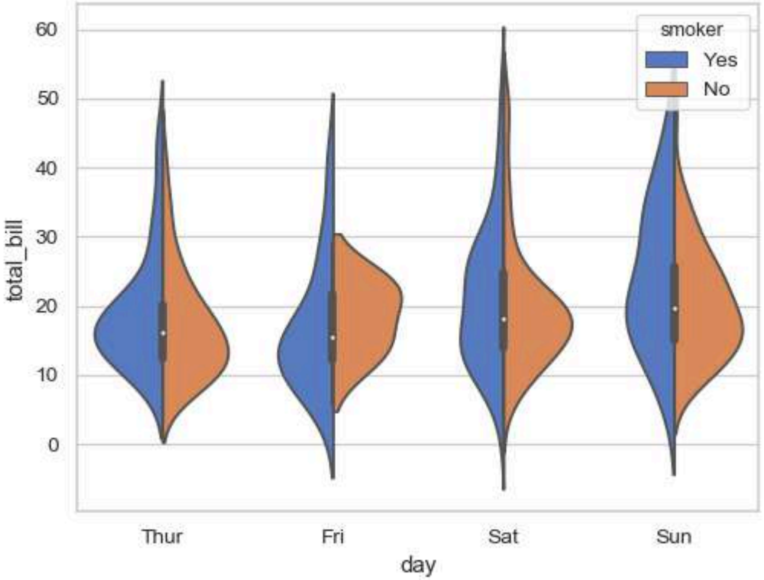
Lower Probability



Group Differences

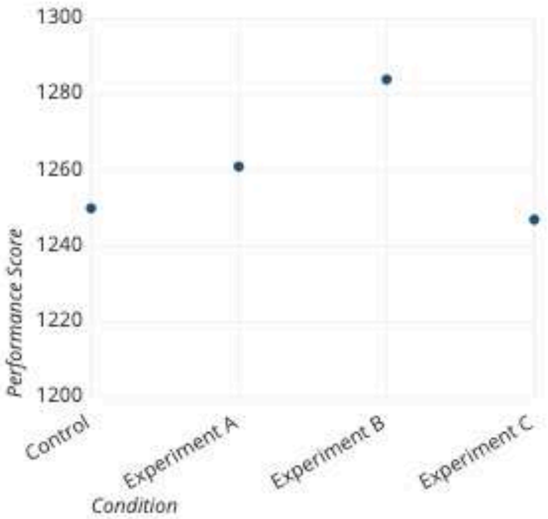
- violin plots (+ box plots)

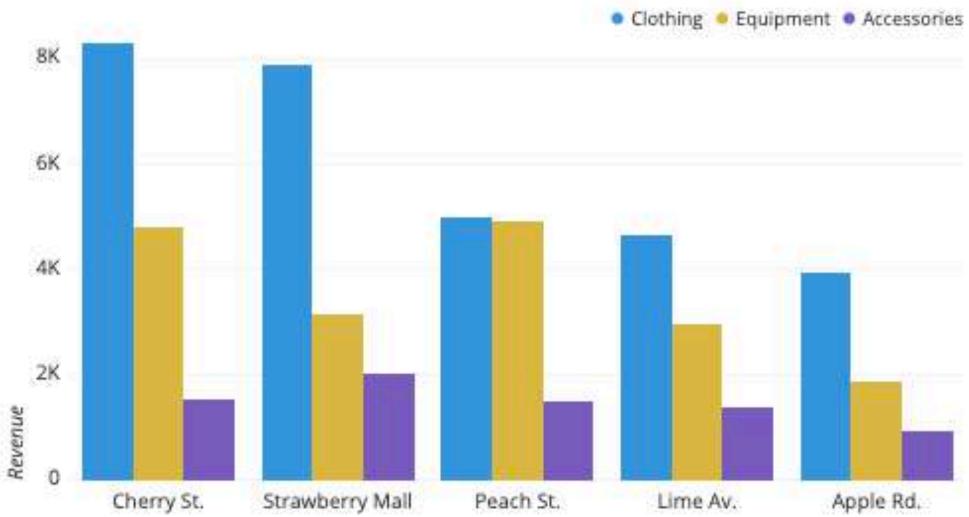




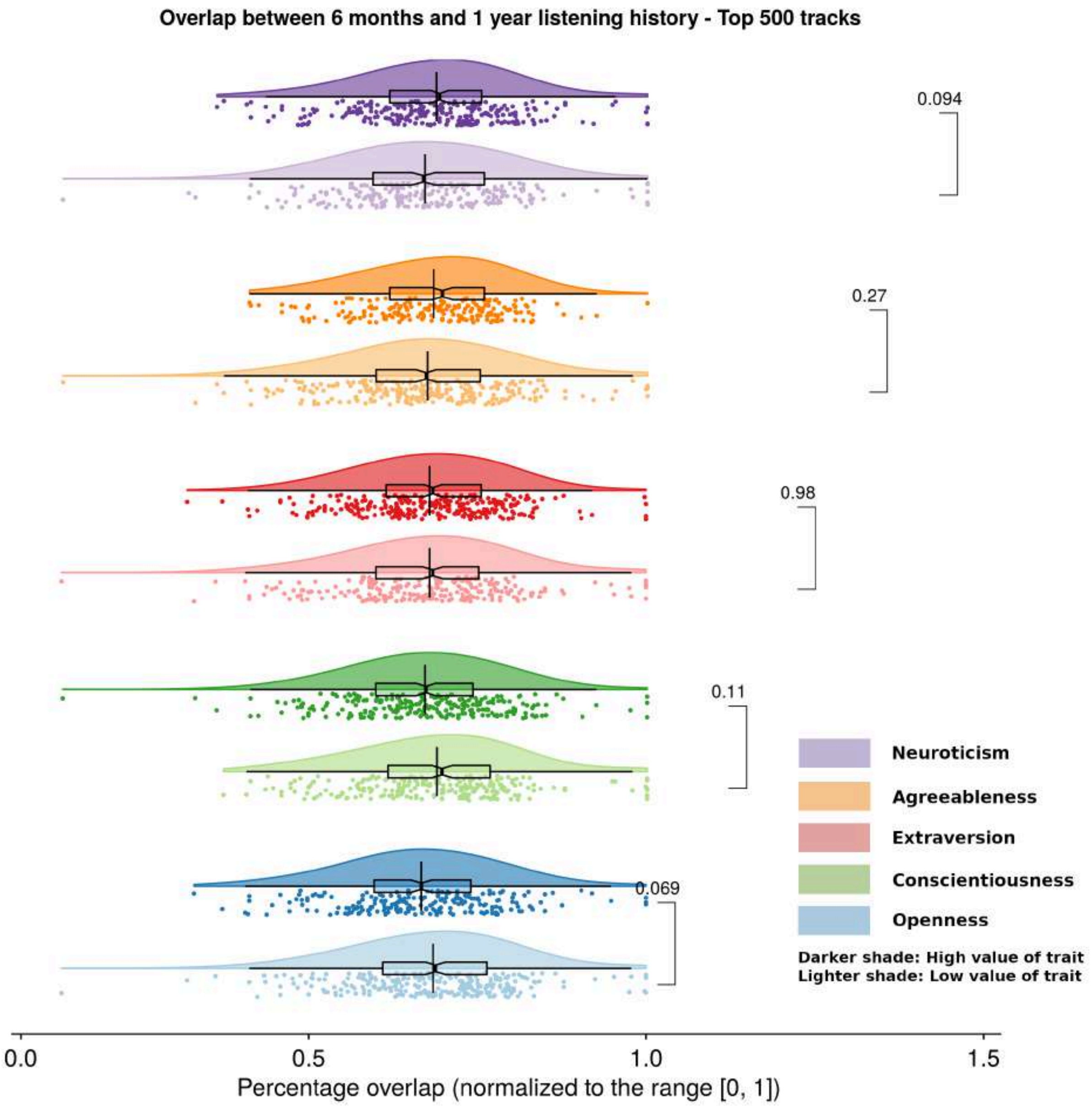
Group differences

- these are less desirable as they do not show the spread



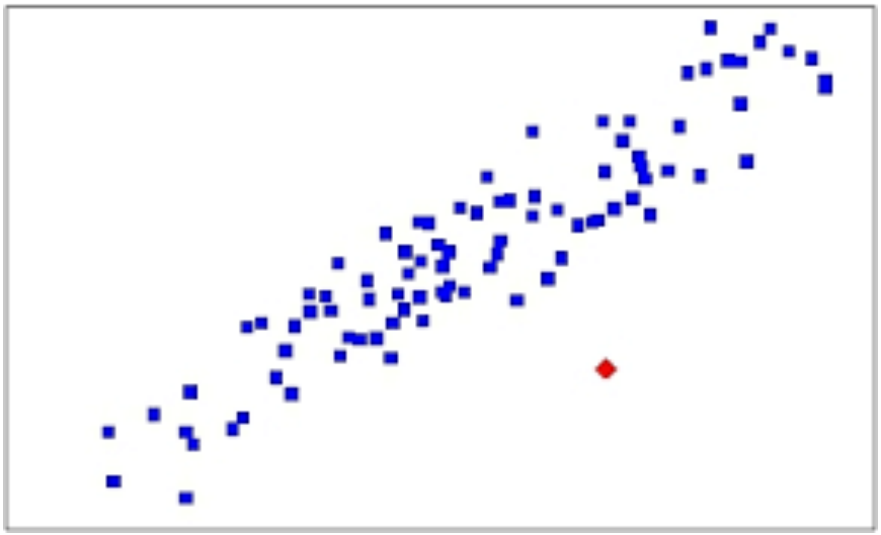


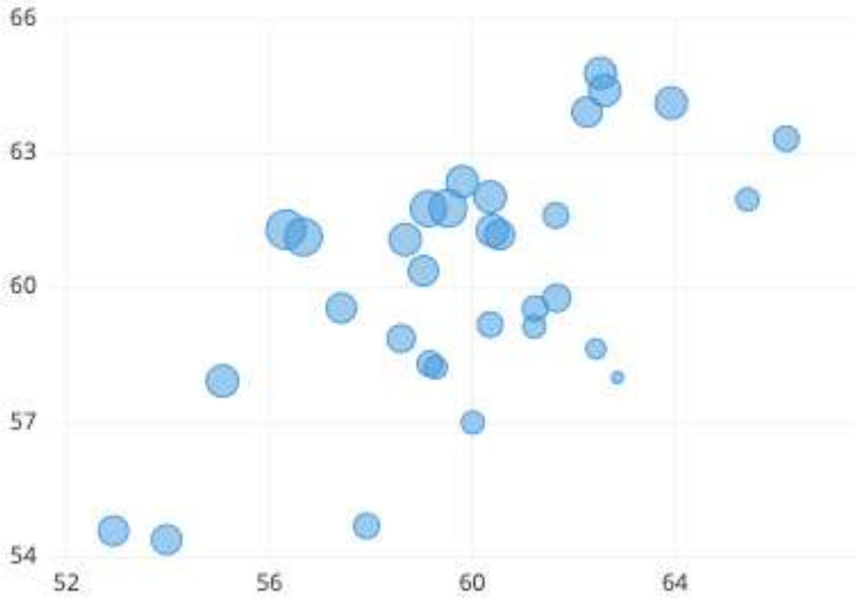
Describing Data + Group Differences



Association between variables

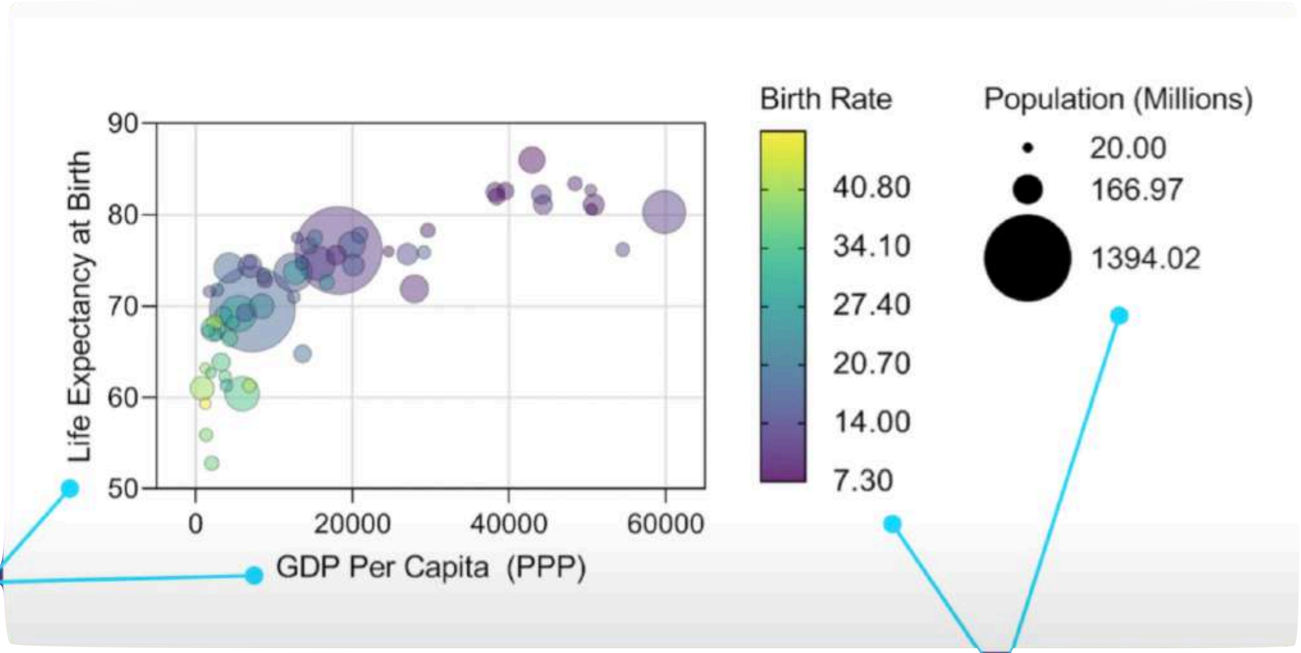
- scatter/bubble plots
  - allows you to observe the relationship between variables





Association between variables

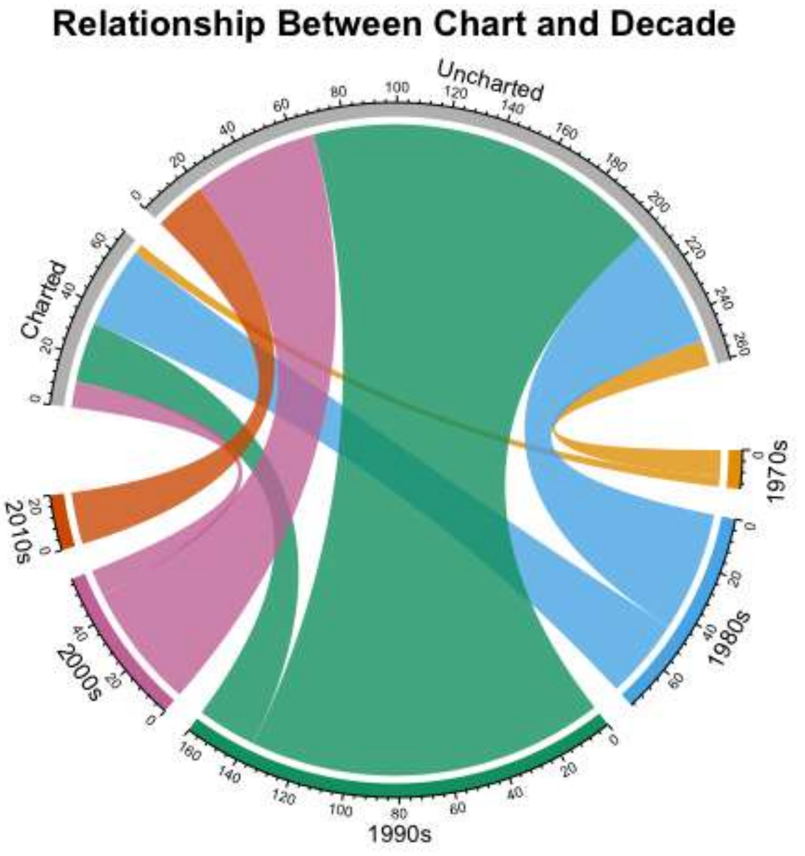
- bubble plots (good for multivariate data)

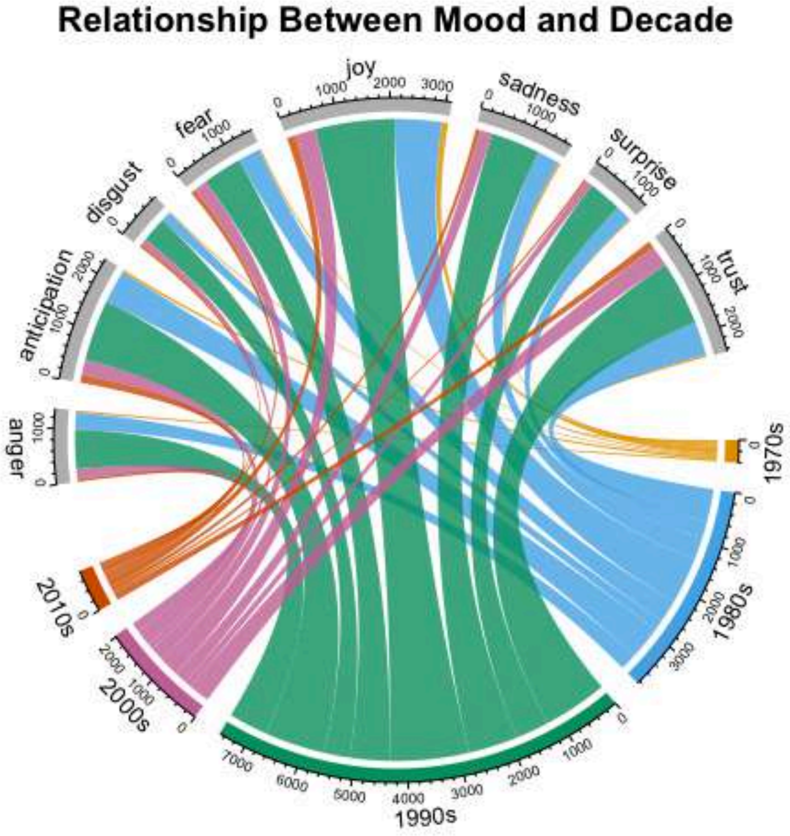


X and Y variables are numeric

Additional variables can be categorical or continuous

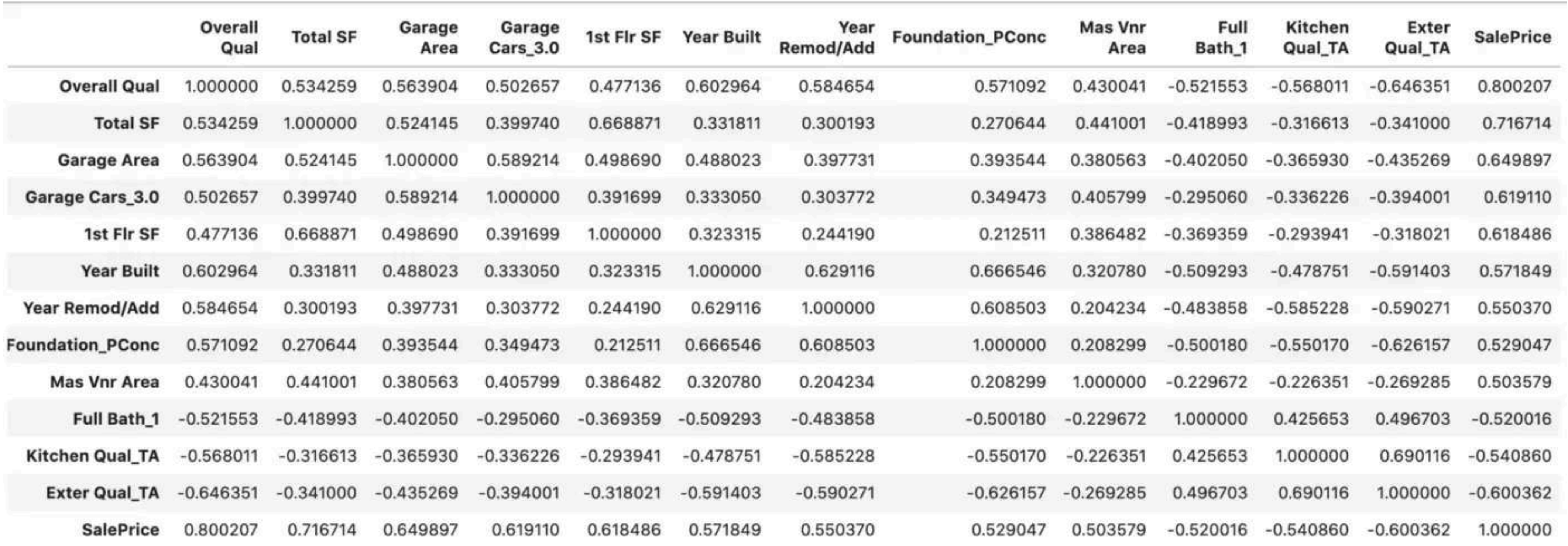
Association between variables





Association between variables

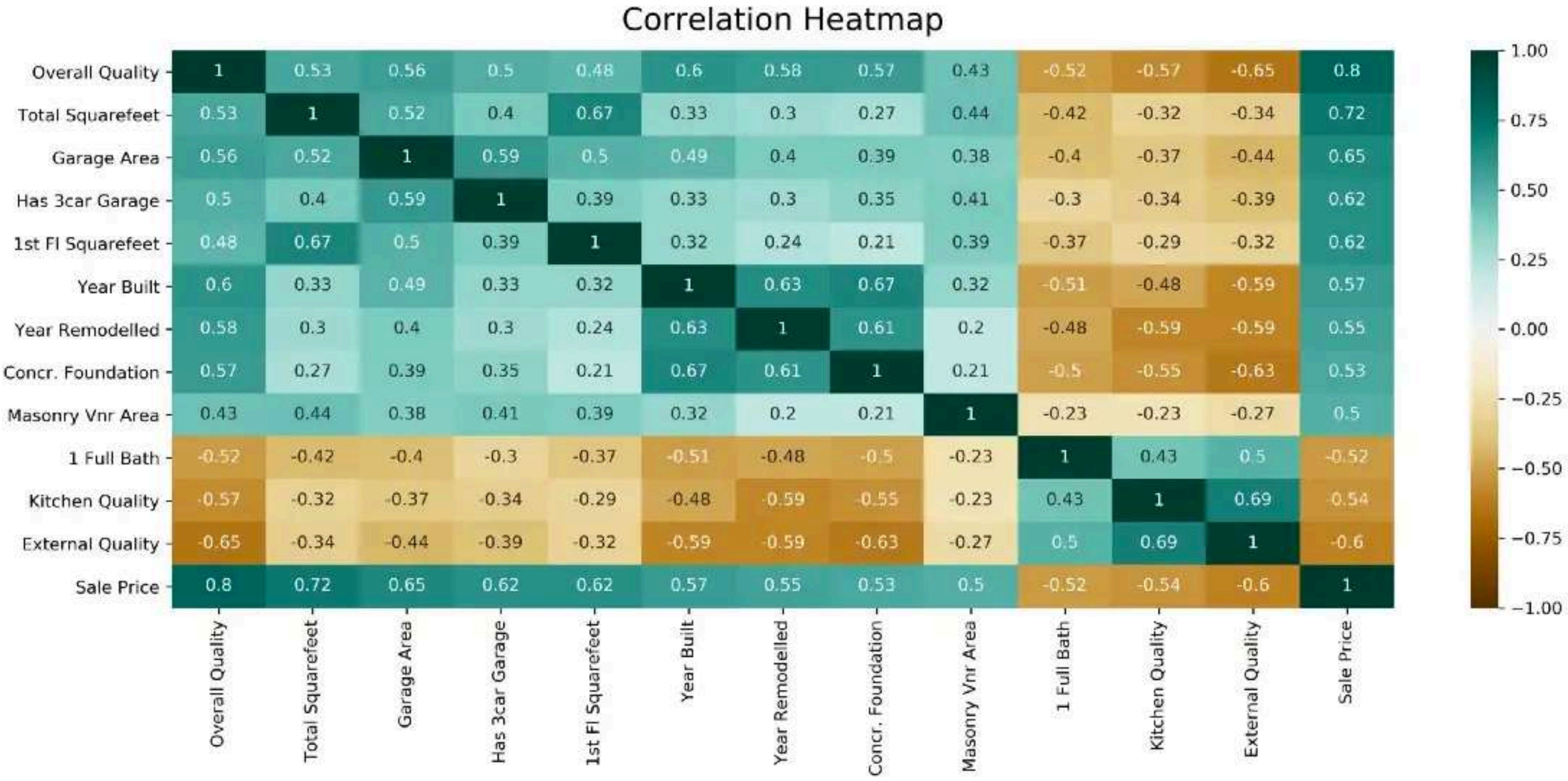
- heat maps depicting correlations



	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

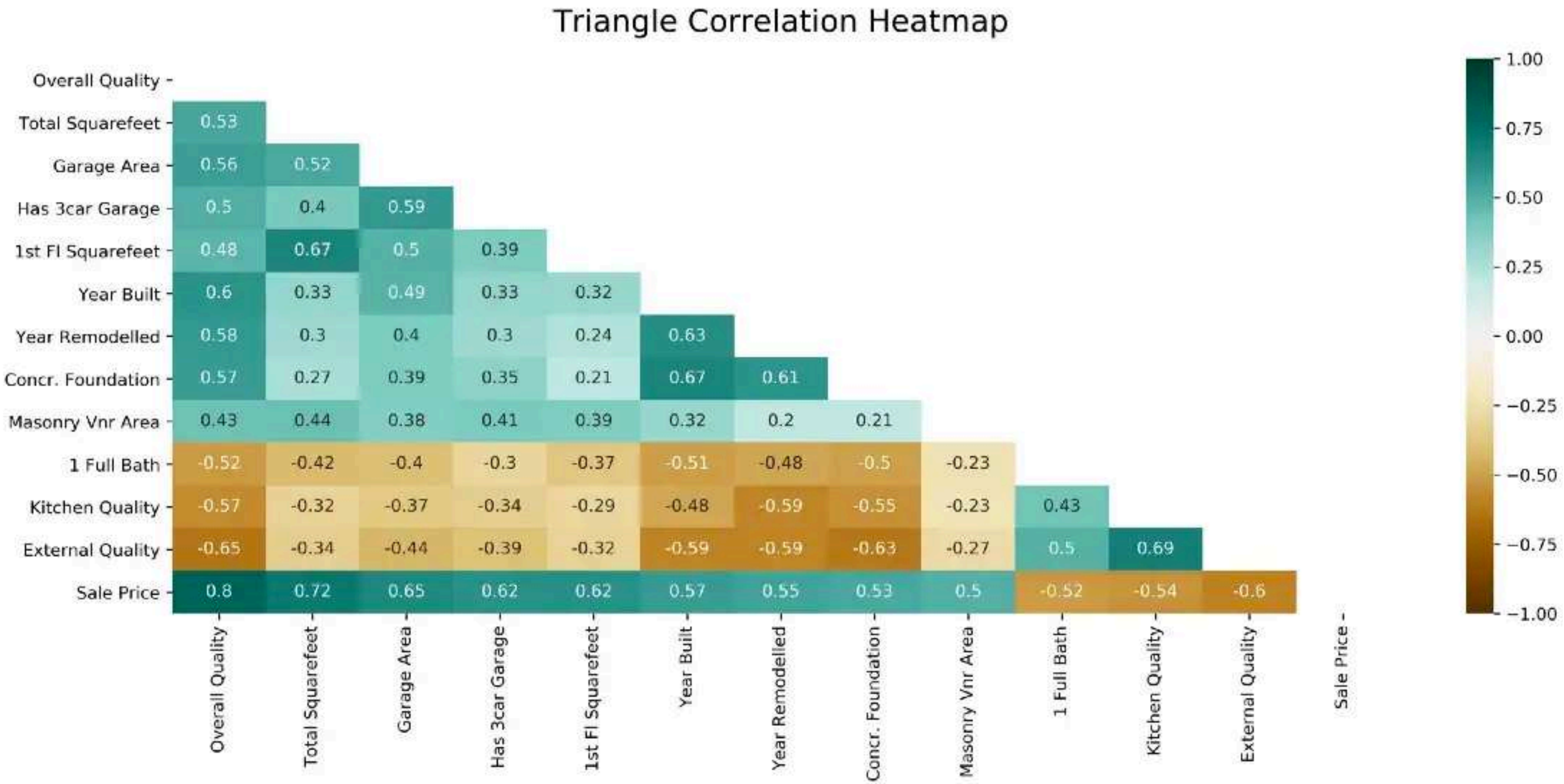


Correlation Heatmap

	Overall Quality	Total Squarefeet	Garage Area	Has 3car Garage	1st Fl Squarefeet	Year Built	Year Remodelled	Concr. Foundation	Masonry Vnr Area	1 Full Bath	Kitchen Quality	External Quality	Sale Price
Overall Quality	1	0.53	0.56	0.5	0.48	0.6	0.58	0.57	0.43	-0.52	-0.57	-0.65	0.8
Total Squarefeet	0.53	1	0.52	0.4	0.67	0.33	0.3	0.27	0.44	-0.42	-0.32	-0.34	0.72
Garage Area	0.56	0.52	1	0.59	0.5	0.49	0.4	0.39	0.38	-0.4	-0.37	-0.44	0.65
Has 3car Garage	0.5	0.4	0.59	1	0.39	0.33	0.3	0.35	0.41	-0.3	-0.34	-0.39	0.62
1st Fl Squarefeet	0.48	0.67	0.5	0.39	1	0.32	0.24	0.21	0.39	-0.37	-0.29	-0.32	0.62
Year Built	0.6	0.33	0.49	0.33	0.32	1	0.63	0.67	0.32	-0.51	-0.48	-0.59	0.57
Year Remodelled	0.58	0.3	0.4	0.3	0.24	0.63	1	0.61	0.2	-0.48	-0.59	-0.59	0.55
Concr. Foundation	0.57	0.27	0.39	0.35	0.21	0.67	0.61	1	0.21	-0.5	-0.55	-0.63	0.53
Masonry Vnr Area	0.43	0.44	0.38	0.41	0.39	0.32	0.2	0.21	1	-0.23	-0.23	-0.27	0.5
1 Full Bath	-0.52	-0.42	-0.4	-0.3	-0.37	-0.51	-0.48	-0.5	-0.23	1	0.43	0.5	-0.52
Kitchen Quality	-0.57	-0.32	-0.37	-0.34	-0.29	-0.48	-0.59	-0.55	-0.23	0.43	1	0.69	-0.54
External Quality	-0.65	-0.34	-0.44	-0.39	-0.32	-0.59	-0.59	-0.63	-0.27	0.5	0.69	1	-0.6
Sale Price	0.8	0.72	0.65	0.62	0.62	0.57	0.55	0.53	0.5	-0.52	-0.54	-0.6	1

Association between variables

- heat maps depicting correlations

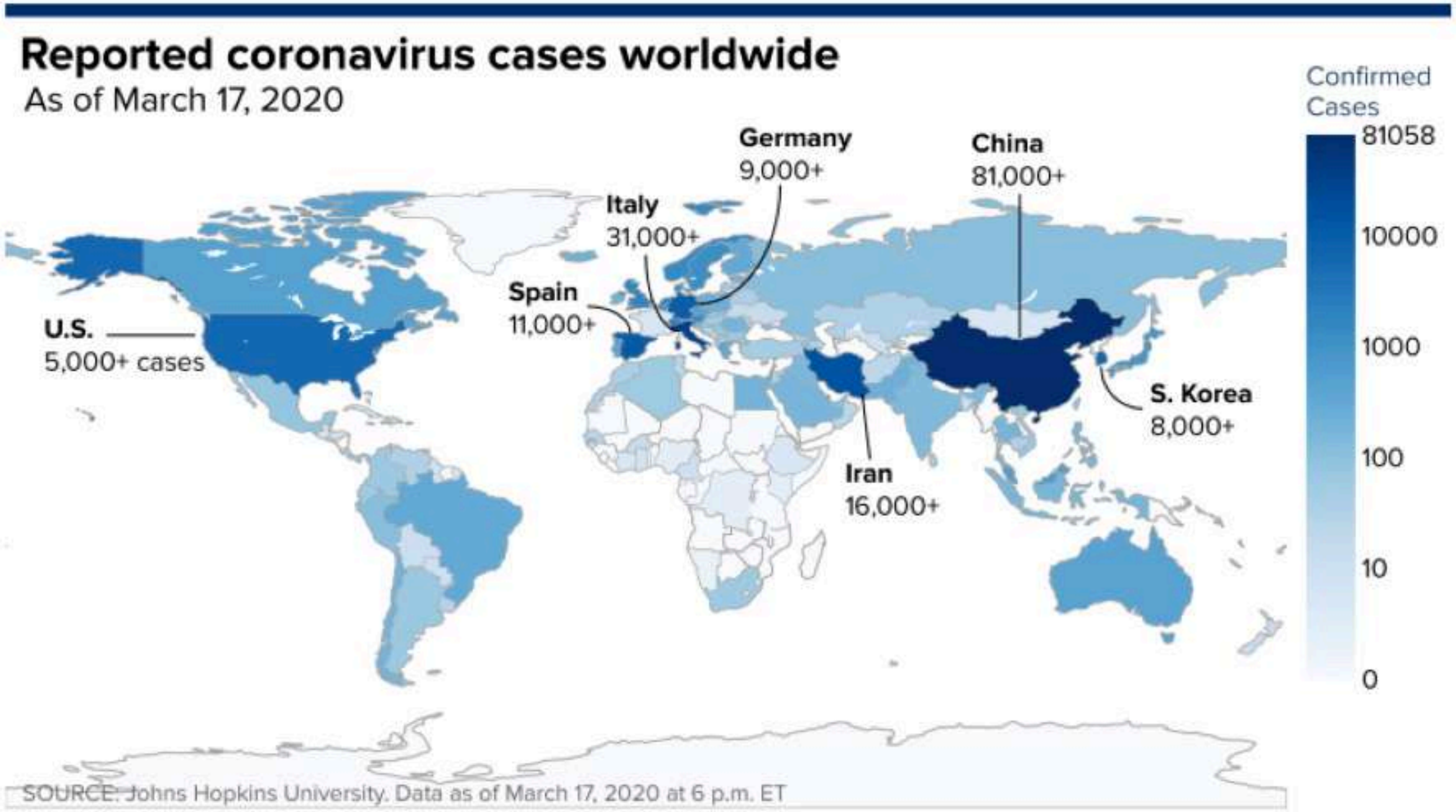


Triangle Correlation Heatmap

	Overall Quality	Total Squarefeet	Garage Area	Has 3car Garage	1st Fl Squarefeet	Year Built	Year Remodelled	Concr. Foundation	Masonry Vnr Area	1 Full Bath	Kitchen Quality	External Quality	Sale Price
Overall Quality													
Total Squarefeet	0.53												
Garage Area	0.56	0.52											
Has 3car Garage	0.5	0.4	0.59										
1st Fl Squarefeet	0.48	0.67	0.5	0.39									
Year Built	0.6	0.33	0.49	0.33	0.32								
Year Remodelled	0.58	0.3	0.4	0.3	0.24	0.63							
Concr. Foundation	0.57	0.27	0.39	0.35	0.21	0.67	0.61						
Masonry Vnr Area	0.43	0.44	0.38	0.41	0.39	0.32	0.2	0.21					
1 Full Bath	-0.52	-0.42	-0.4	-0.3	-0.37	-0.51	-0.48	-0.5	-0.23				
Kitchen Quality	-0.57	-0.32	-0.37	-0.34	-0.29	-0.48	-0.59	-0.55	-0.23	0.43			
External Quality	-0.65	-0.34	-0.44	-0.39	-0.32	-0.59	-0.59	-0.63	-0.27	0.5	0.69		
Sale Price	0.8	0.72	0.65	0.62	0.62	0.57	0.55	0.53	0.5	-0.52	-0.54	-0.6	

Geographical maps

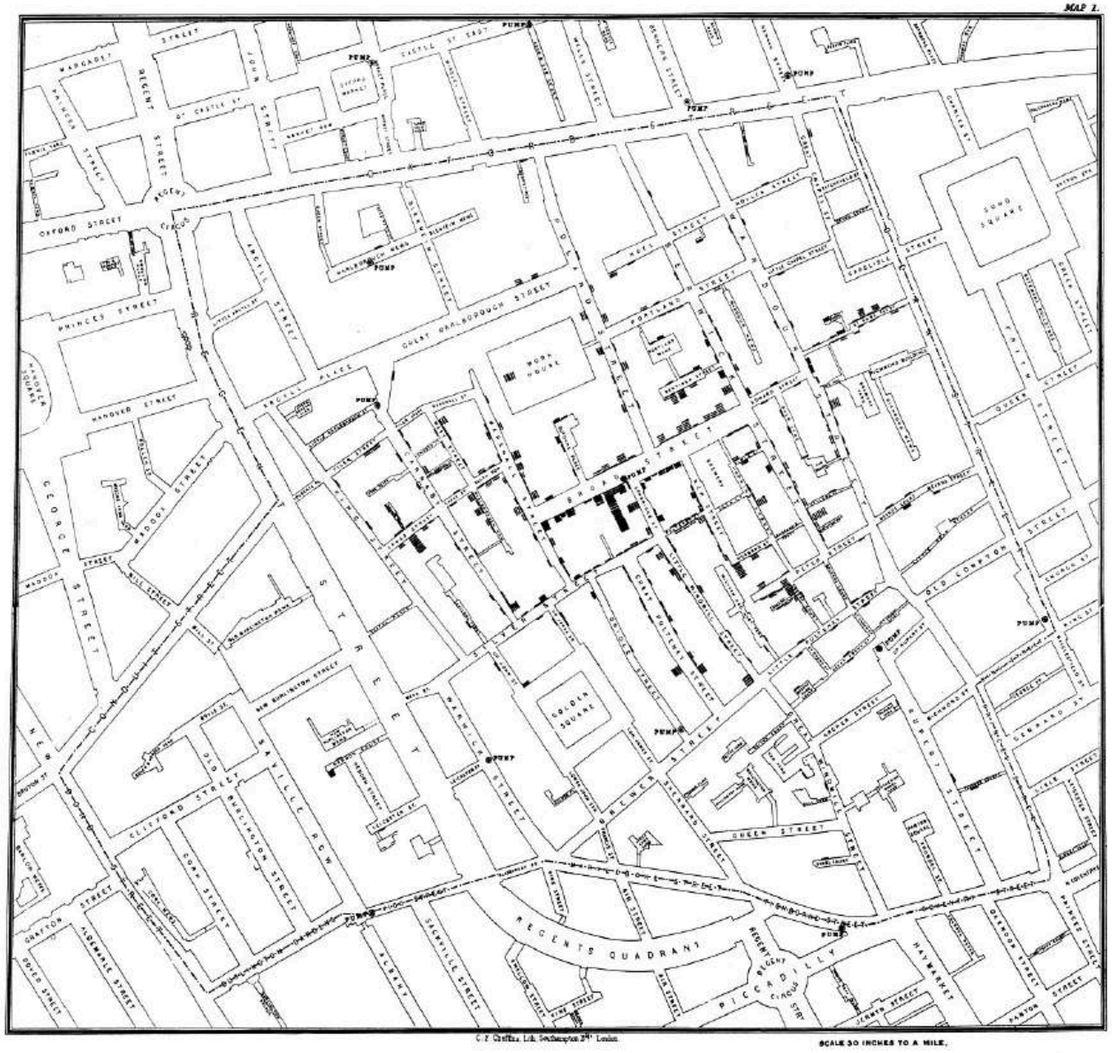
- chloropleth map
- heat map + area map



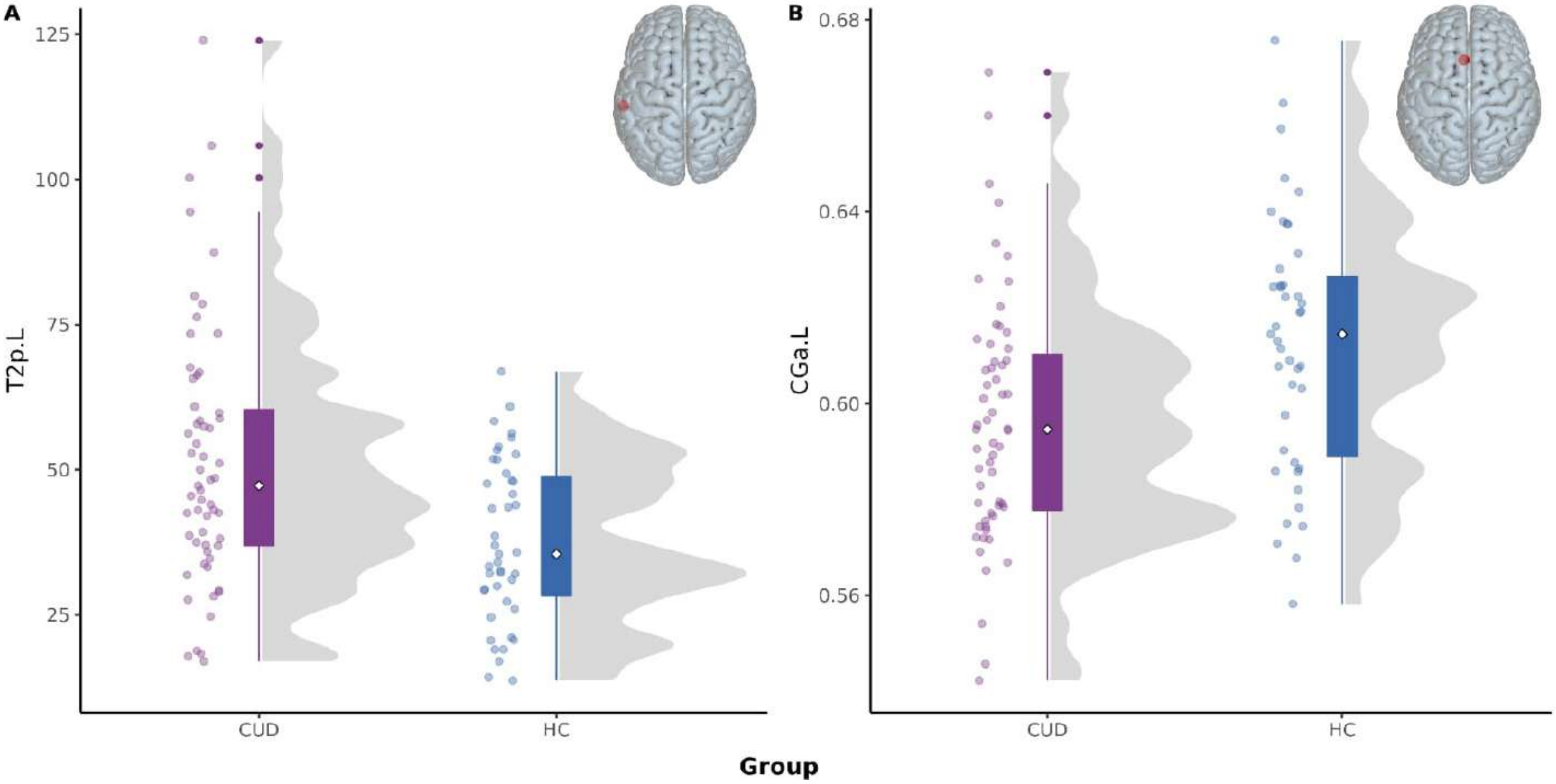
SOURCE: Johns Hopkins University. Data as of March 17, 2020 at 6 p.m. ET

EXAMPLE

John Snow’s cholera map in 1854



Creative Combinations



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)