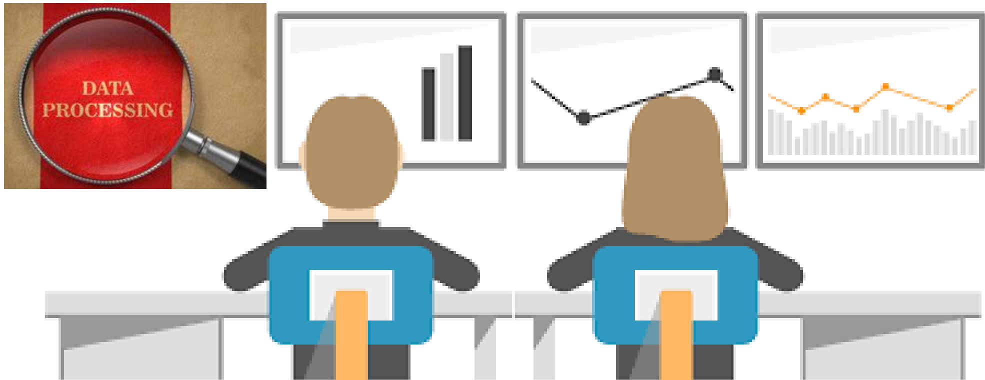
BRSM
Data Visualization & Summarization

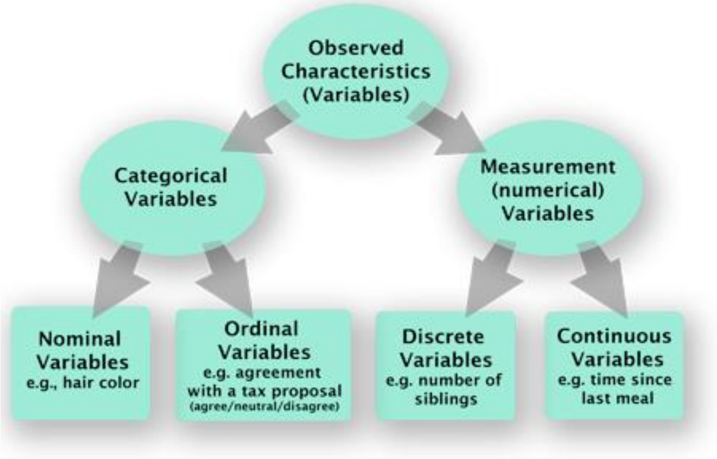
Vinoo Alluri & Bapi Raju

ORGANIZATION

DATA VISUALIZATION



Data Organization


Observed Characteristics (Variables)
Categorical Variables
Measurement (numerical) Variables
Nominal Variables e.g., hair color
Ordinal Variables e.g. agreement with a tax proposal (agree/neutral/disagree)
Discrete Variables e.g. number of siblings
Continuous Variables e.g. time since last meal

- 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

Summarize

How?

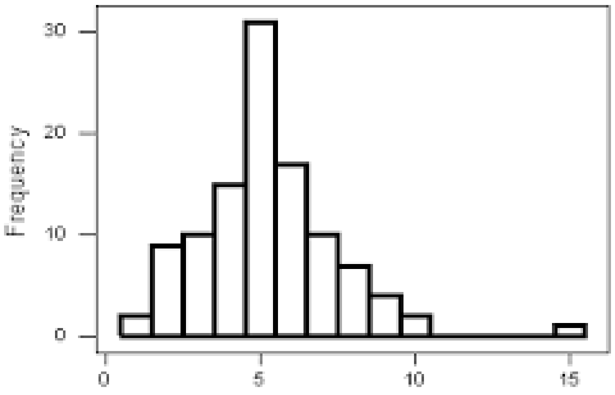
What information does it give ???

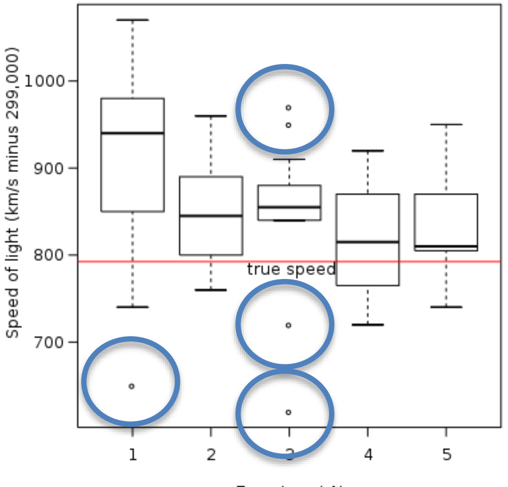
Outline

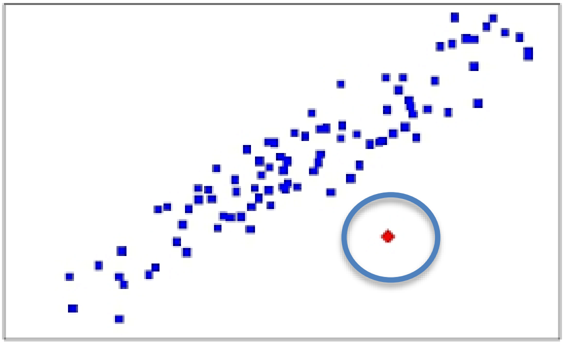
- **Visualization**
  - why we visualise
  - how to pick a plot
  - initial data vs final results visualization (some examples)
  - bad designs and misleading graphs
- **Summarization**
  - measures of central tendency & dispersion
  - which measure to pick

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 a good visualisation?

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

Remove
to improve
(the data-ink ratio)
Created by Darkhorse Analytics
www.darkhorseanalytics.com

"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

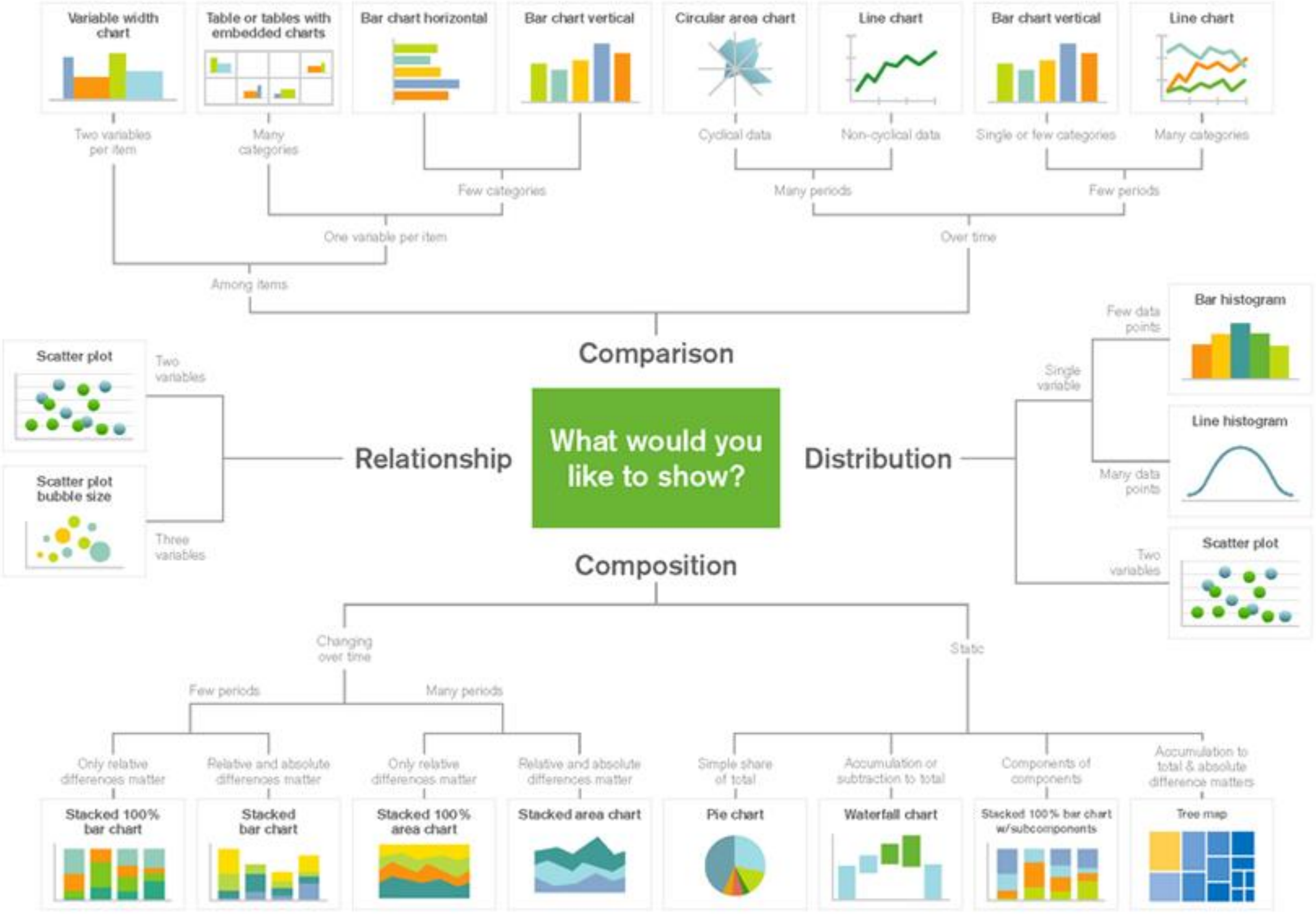
- labelling
  - label the axis correctly and consistently across all your charts.
  - avoid using acronyms that are not widely understood.
  - make the chart title as concise and descriptive as possible.
  - whenever possible, label the lines in your line chart directly rather than using a legend.
  - 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

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

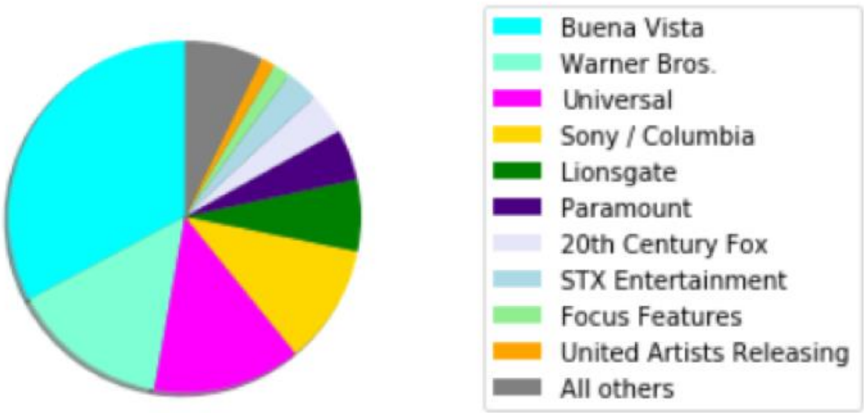
Not an exhaustive list!

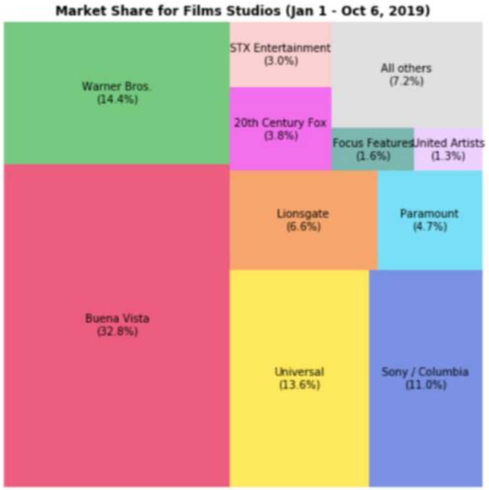
Some plots used for both!

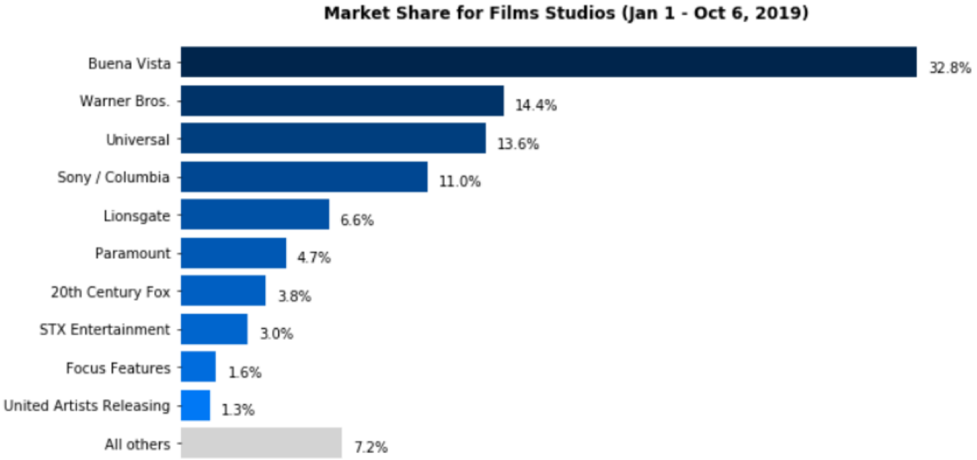
Pie vs Bar Charts

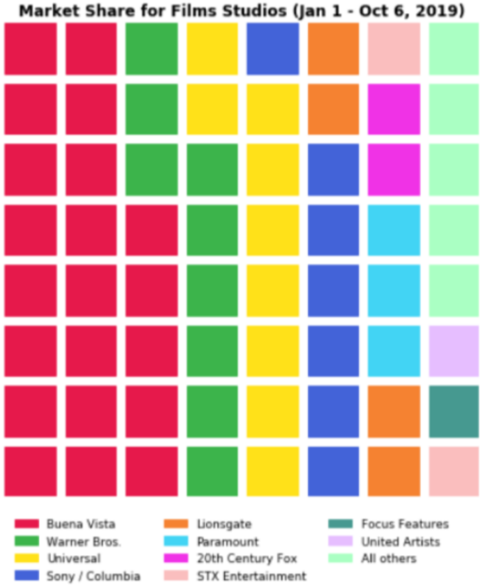
- use pie charts when
  - smaller no. of categories
  - readers can differentiate slices (unless you are making a point)
  - you don’t need to rely on many colors or labels to explain the proportions
  - total adds up to 100%
- use bar charts when
  - have many categories (not too many)
  - need to compare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

So which visualisation was best?



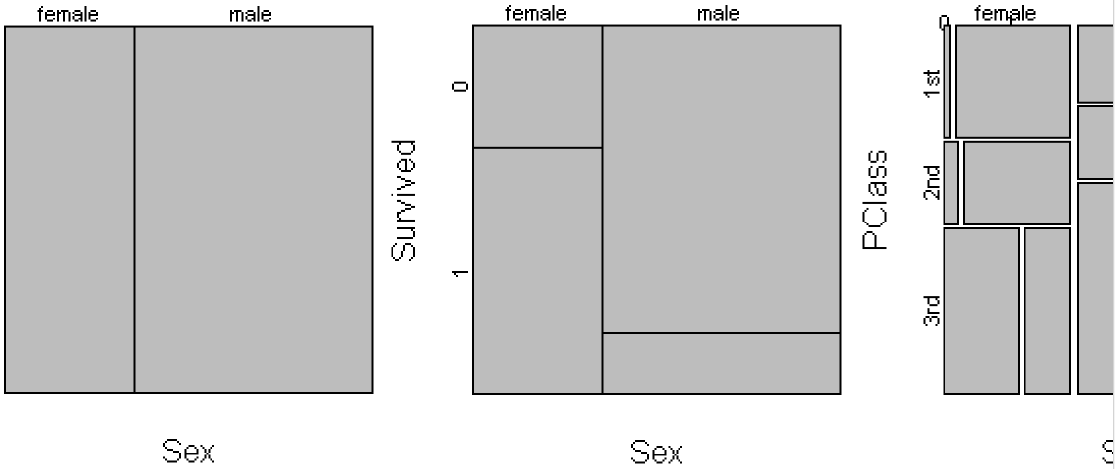


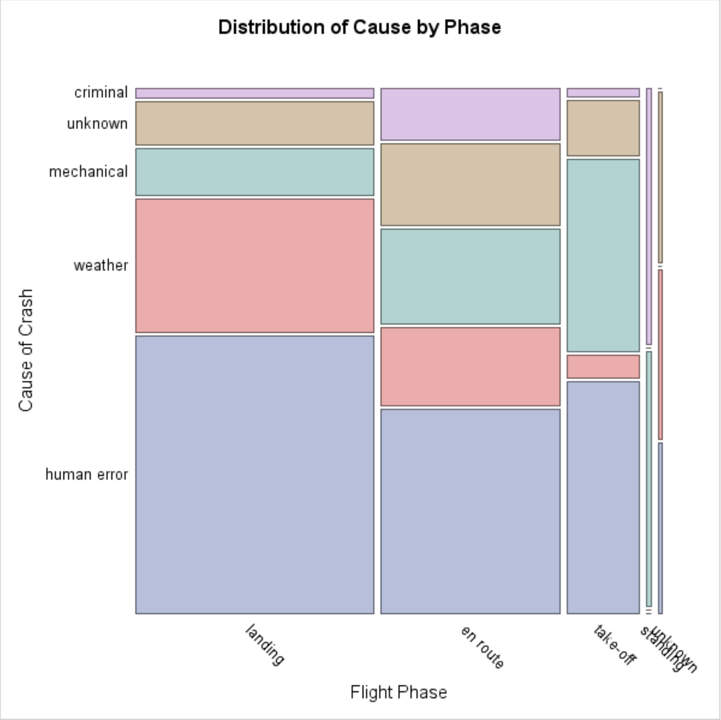




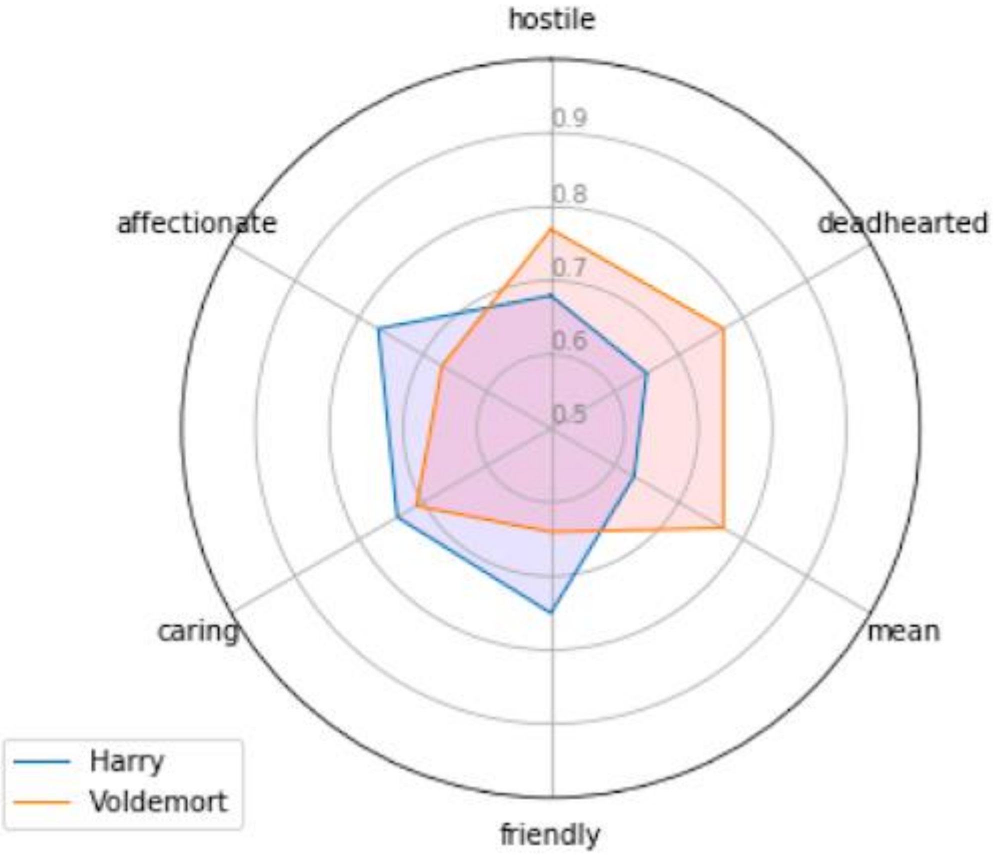
Data Visualisation: Area Plot

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





SPIDER PLOT / RADAR CHART

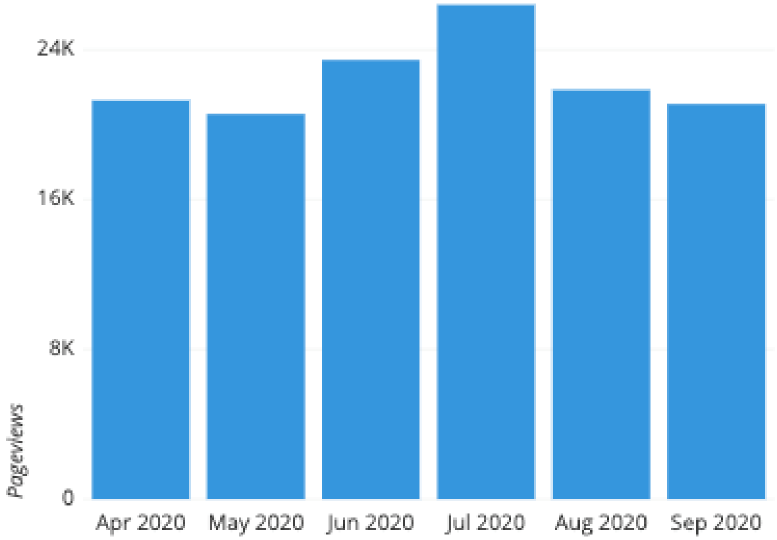


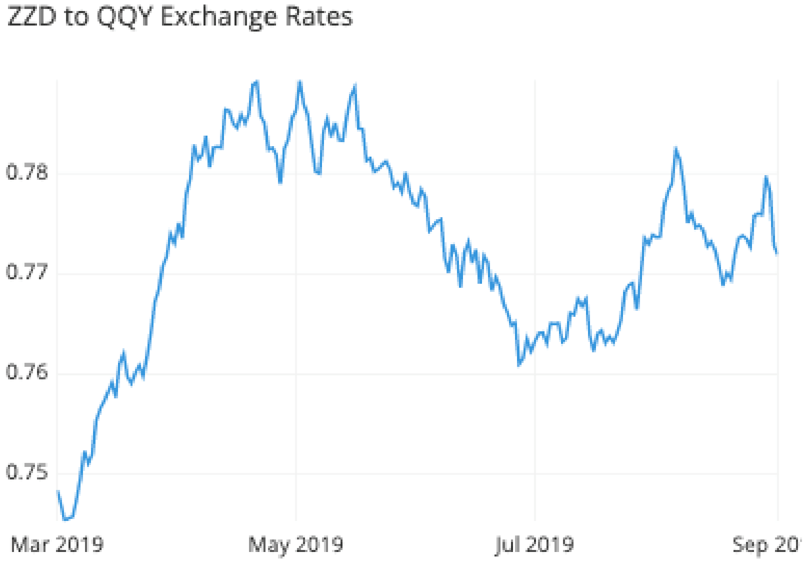
How to choose the right plot?

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

Temporal

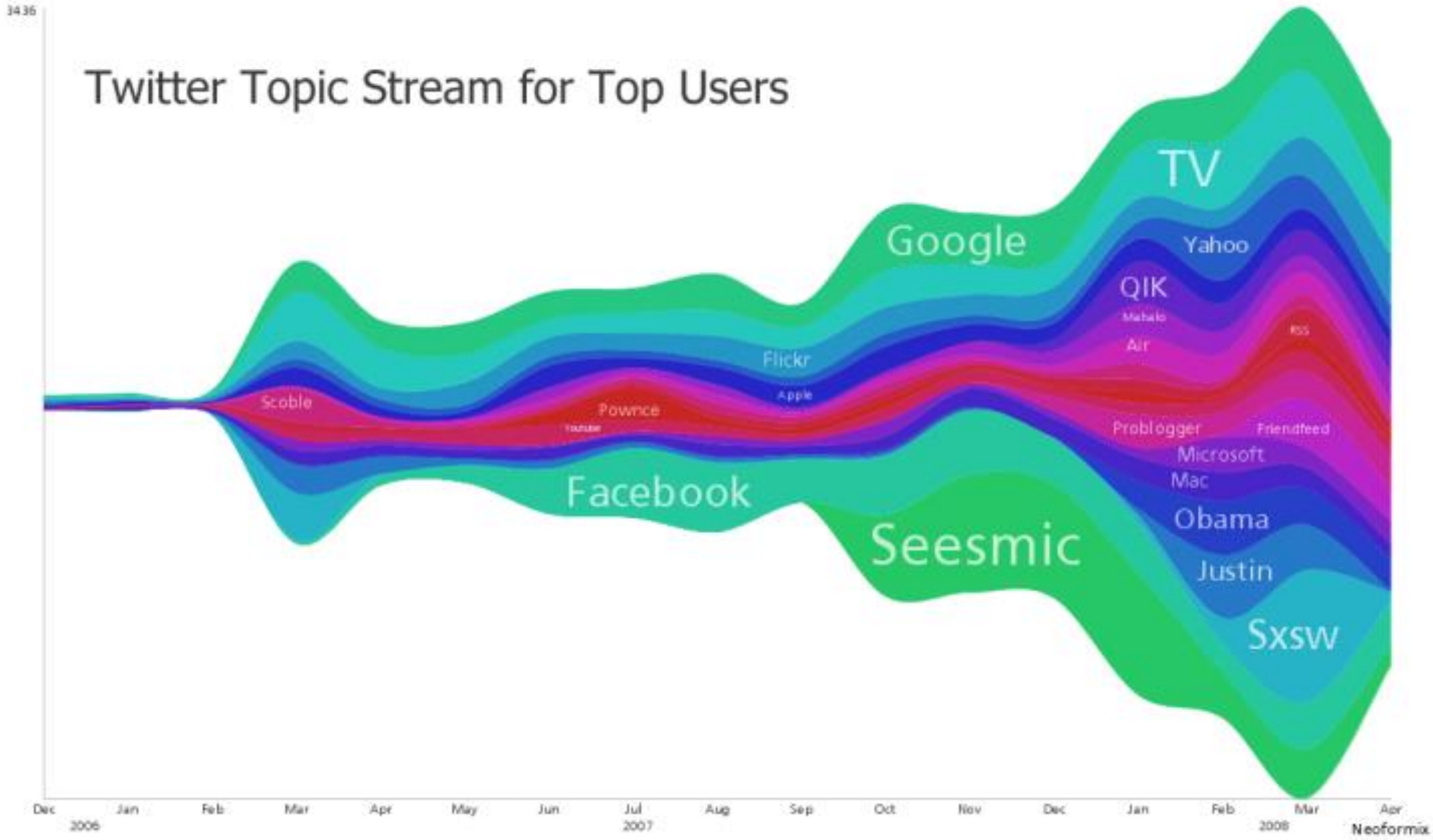
- showing change over time





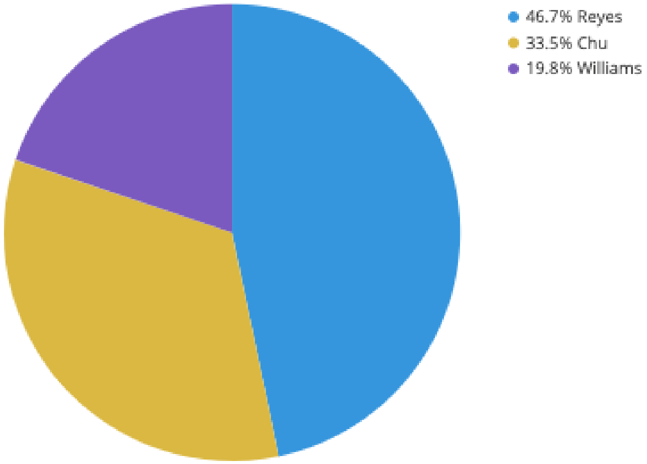
Temporal

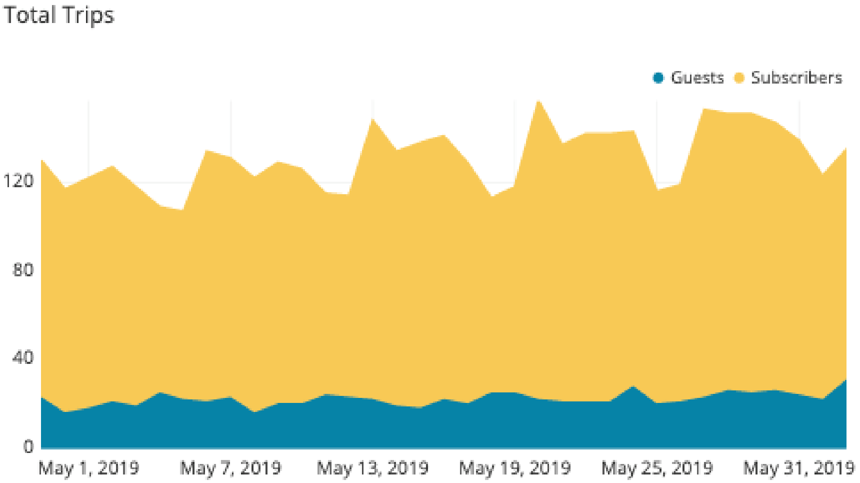
- showing change over time

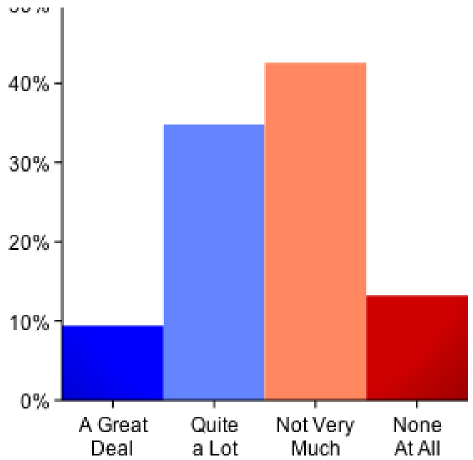


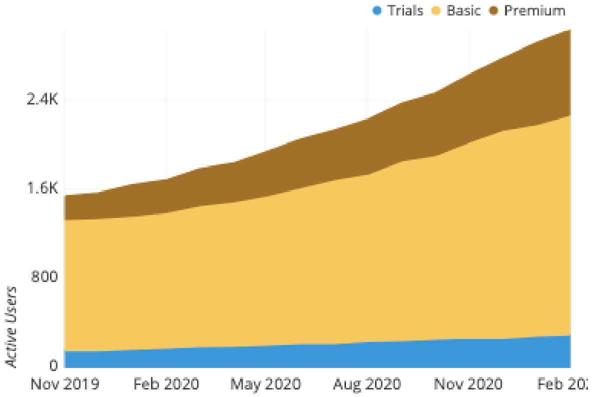
Proportions

- showing a part-to-whole composition

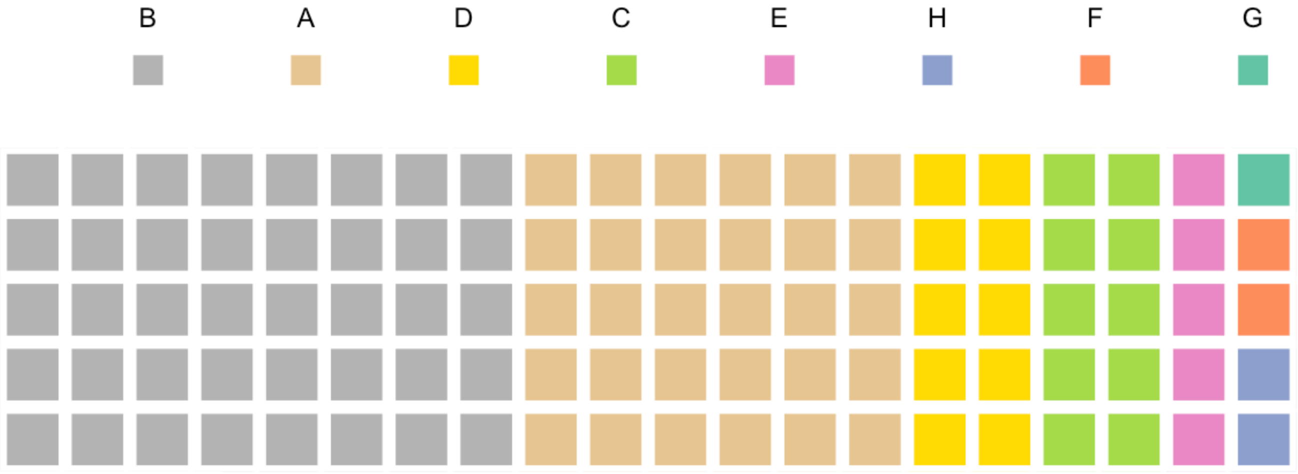




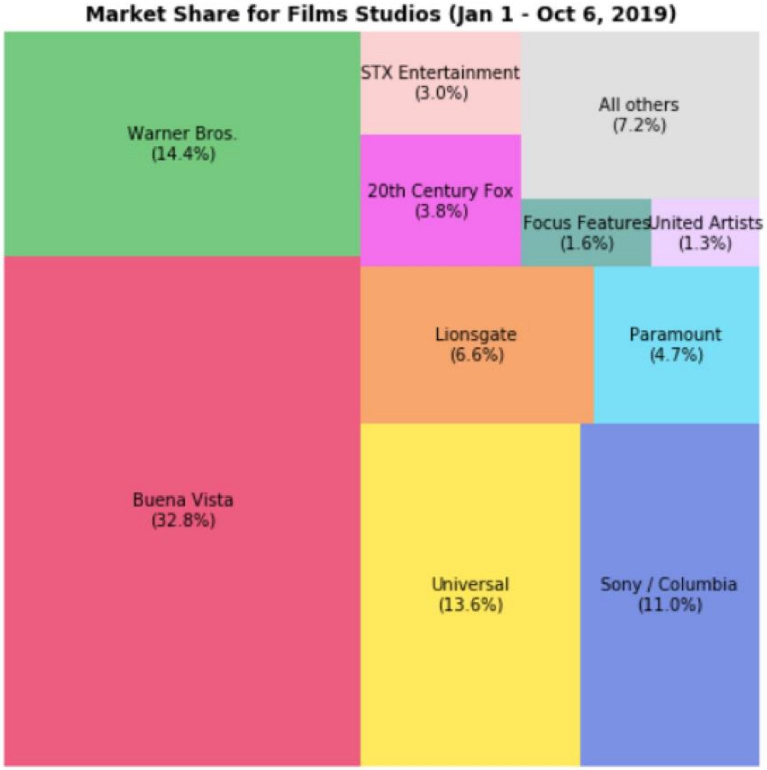




Proportions

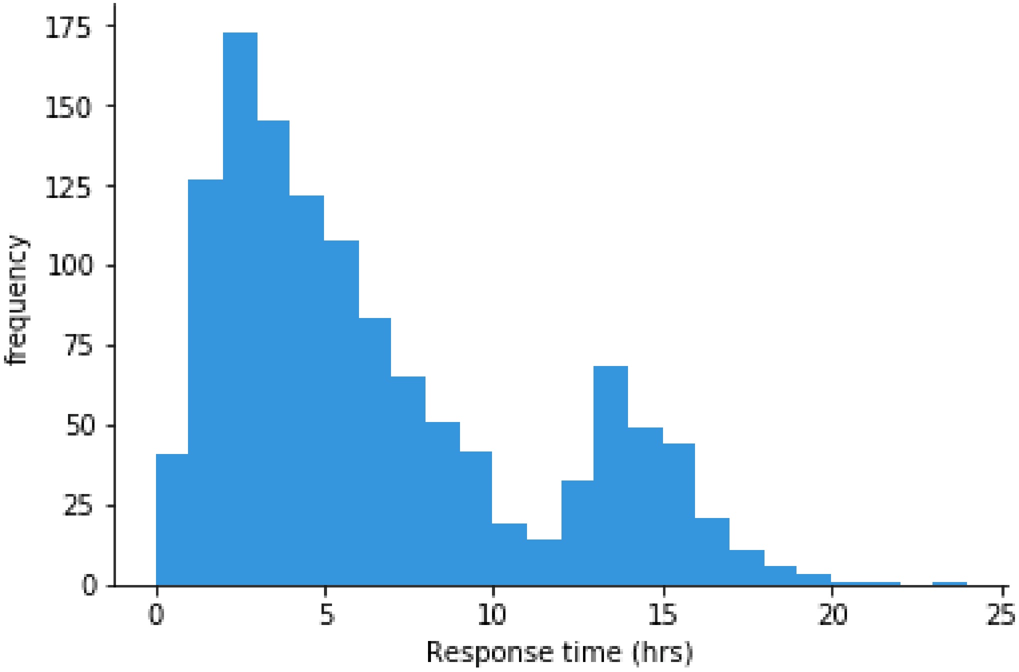


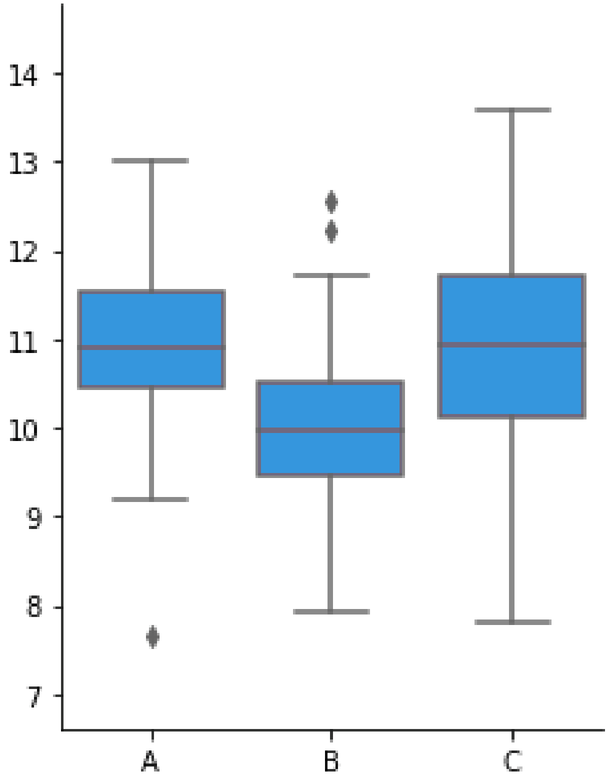
1 square equals 1%



Area plots

Data Distribution

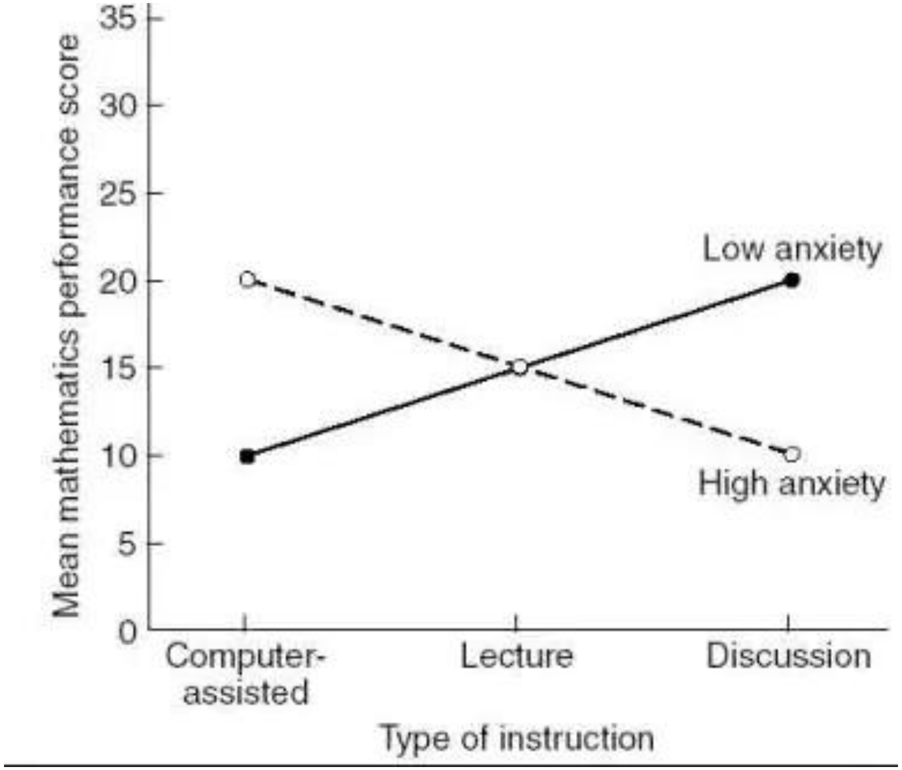




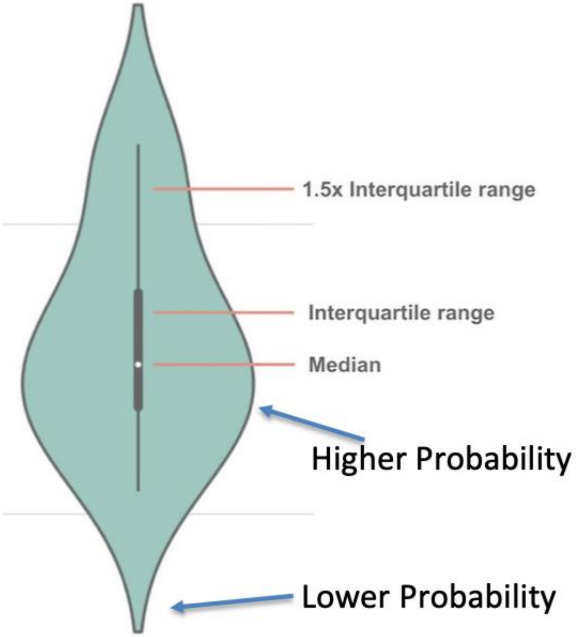
indicative of potential groups or group differences

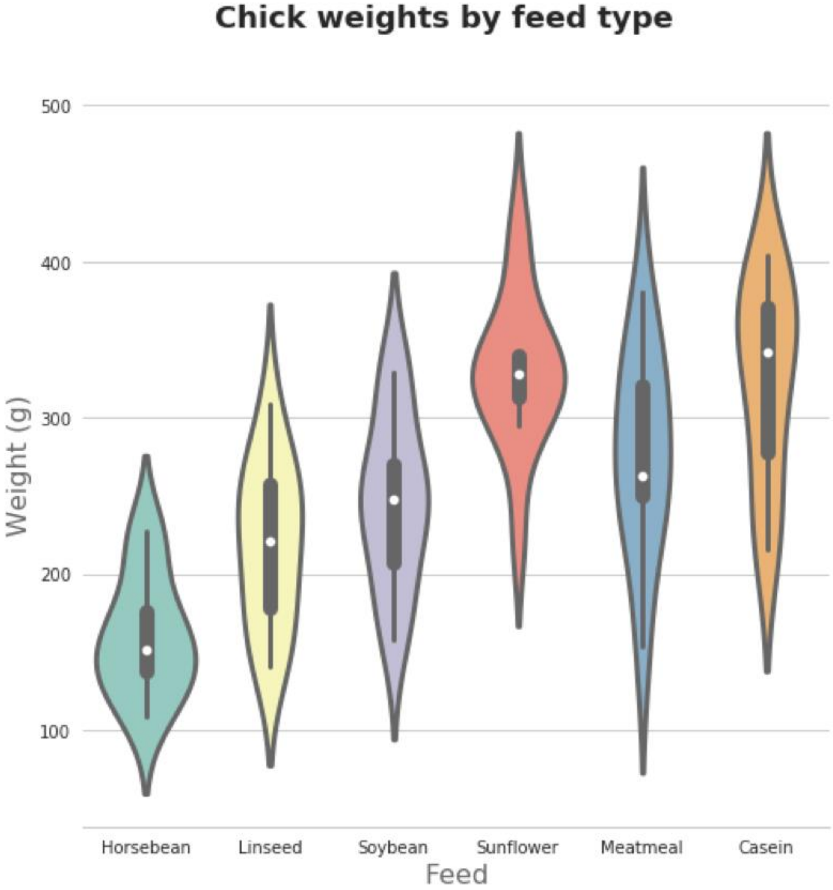
Group Differences

- main effects and interaction plots

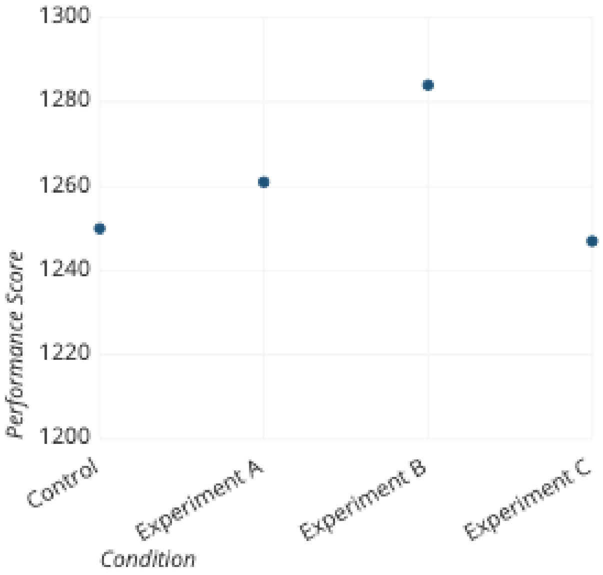


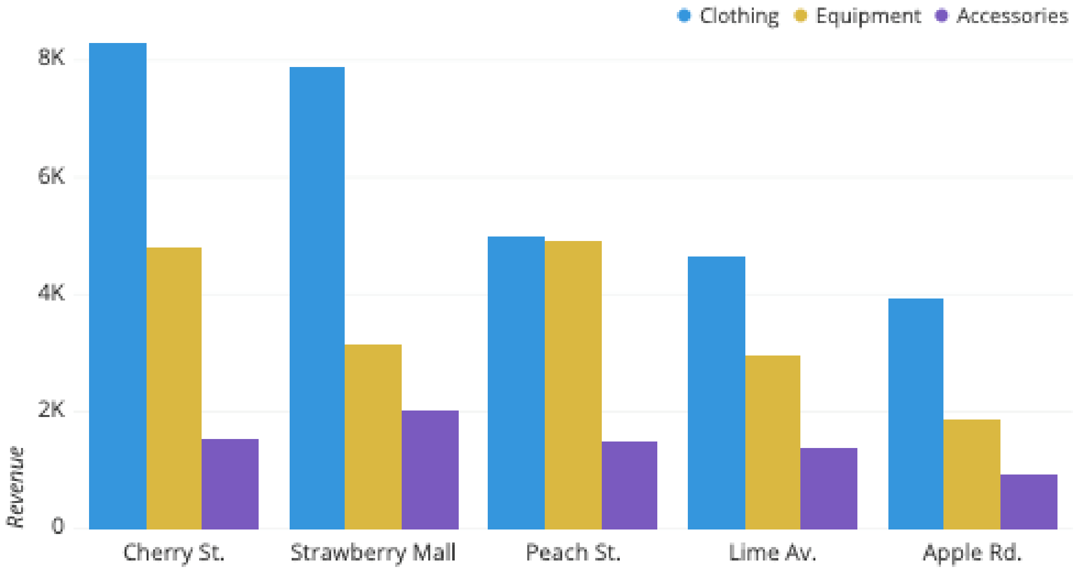
Data Distribution & Group Differences



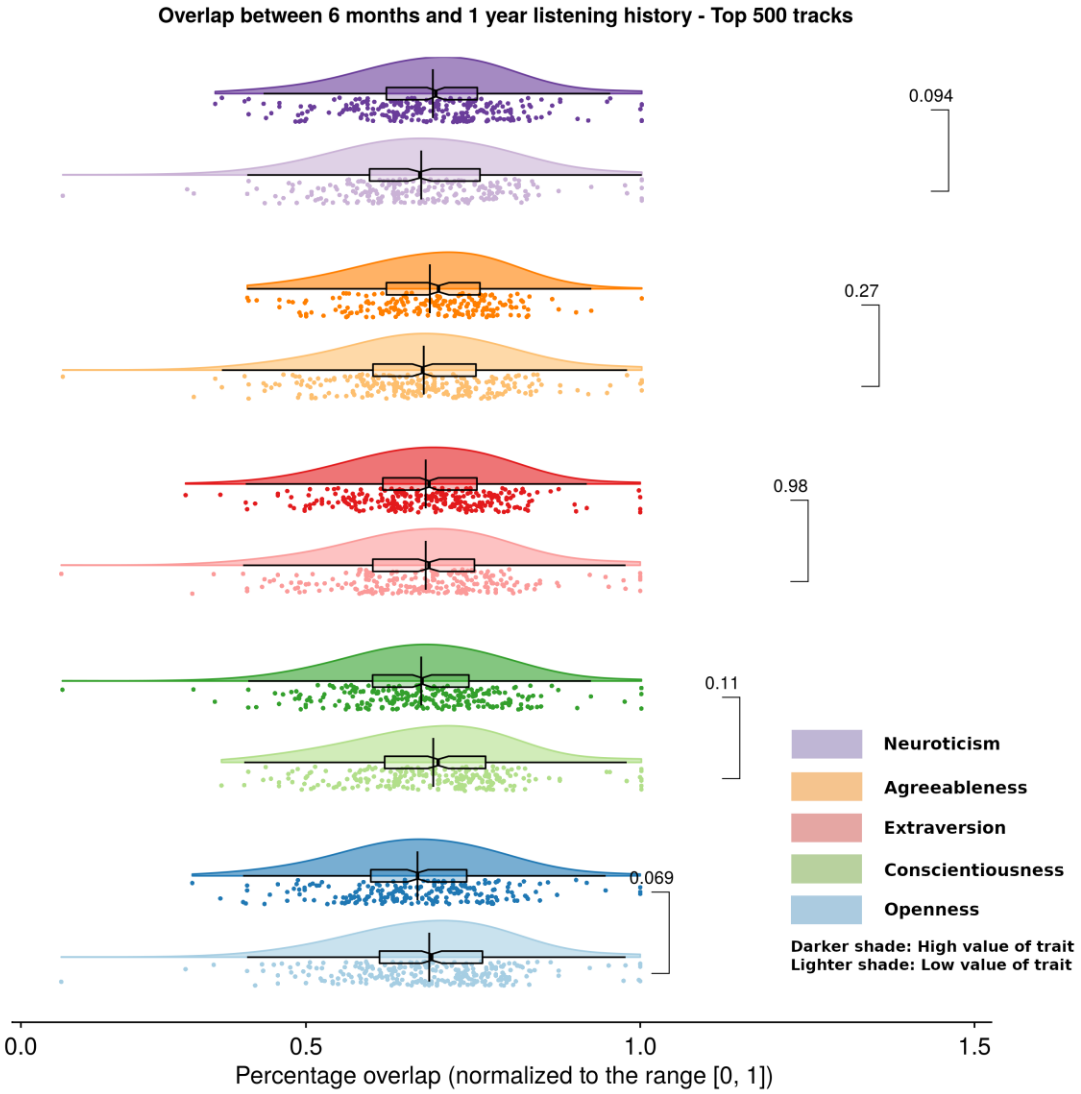


Group differences



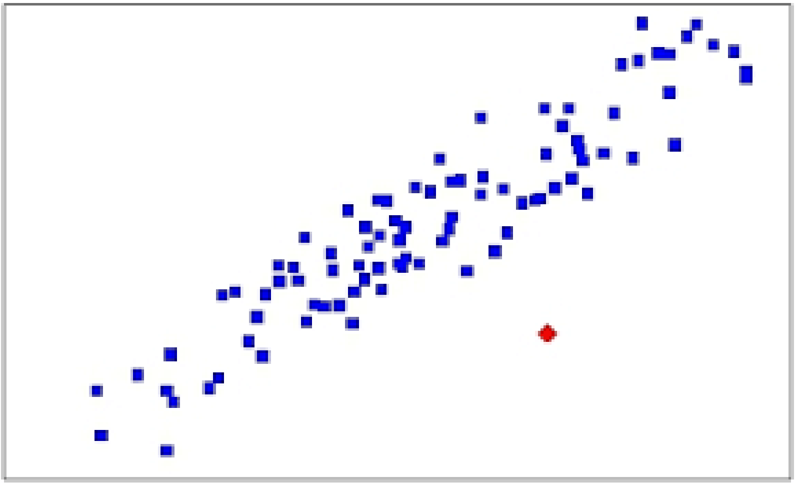


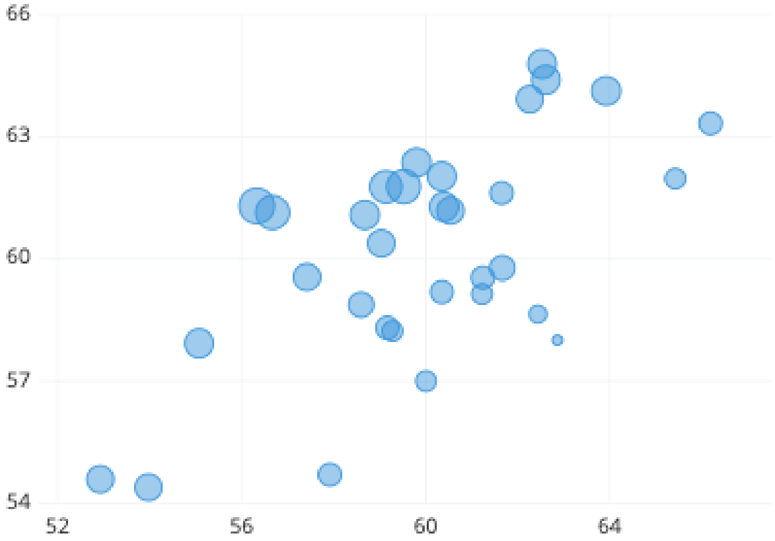
Describing Data + Group Differences



Association between variables

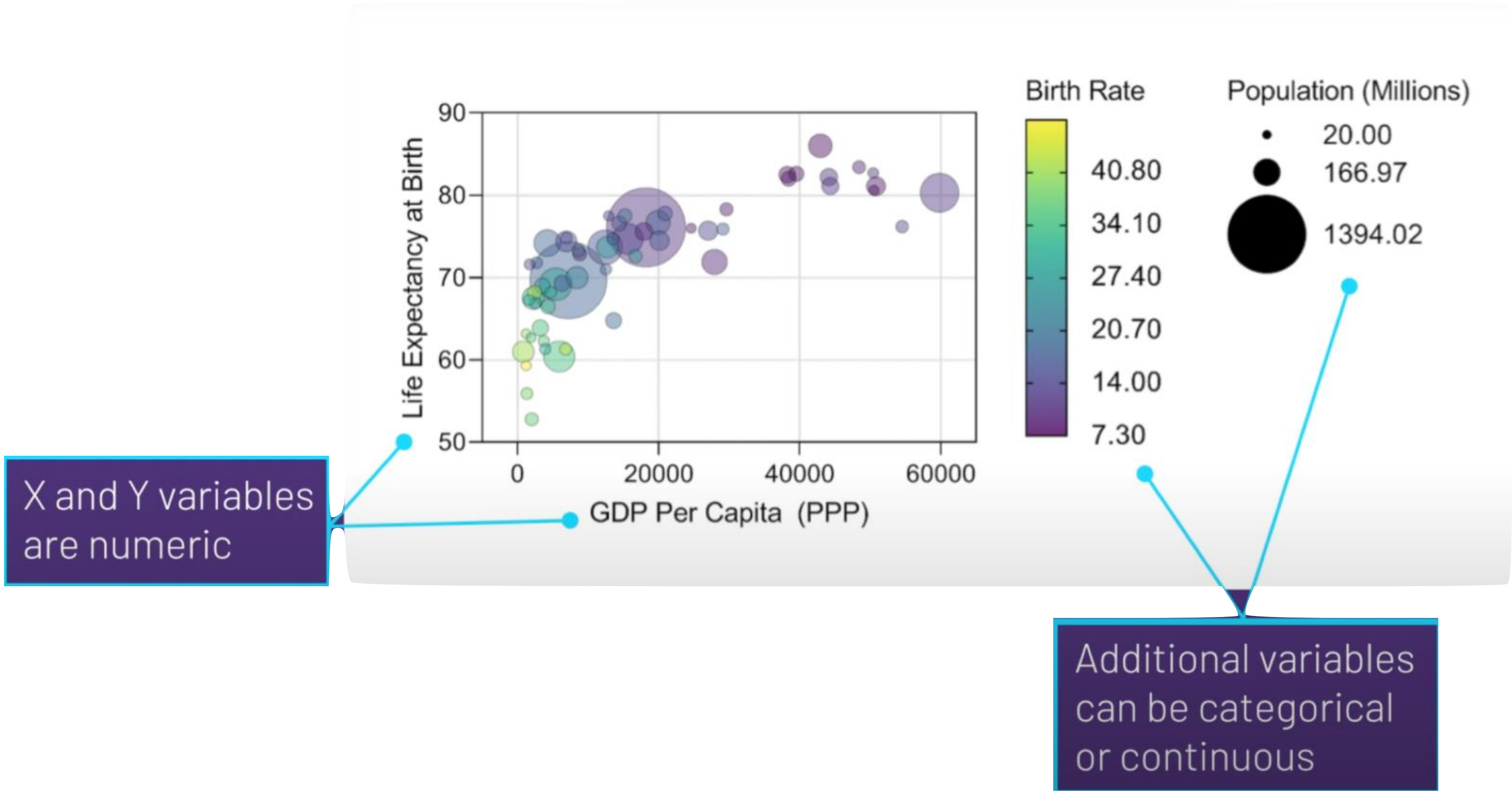
- scatter/bubble plots
    - allows you to observe the relationship between variables





Association between variables

- bubble plots

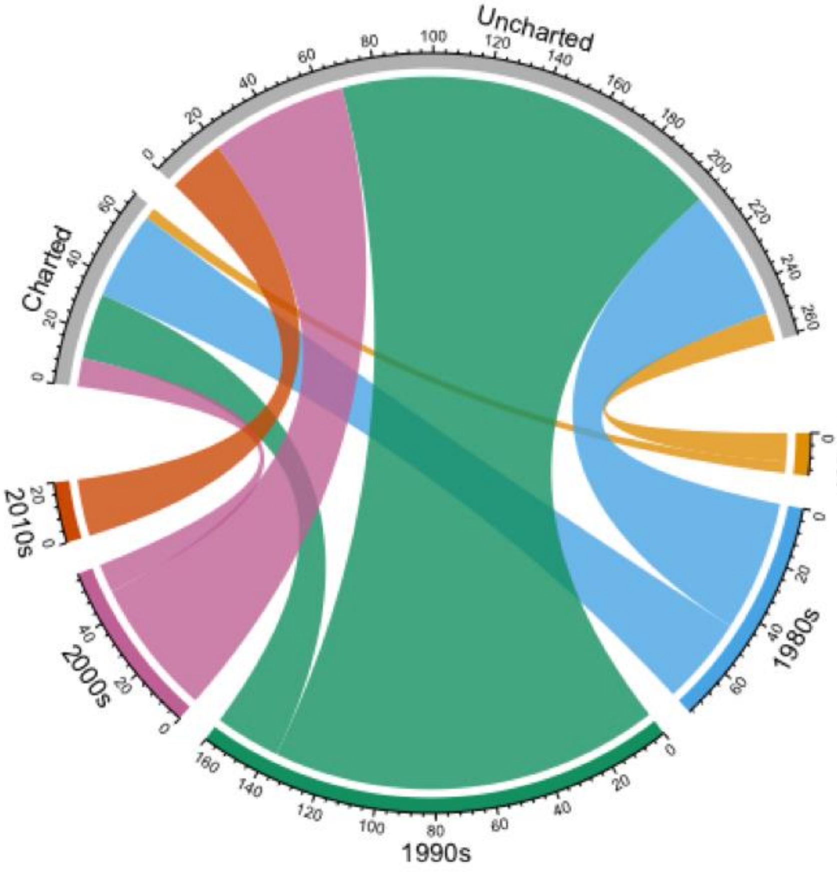


X and Y variables are numeric

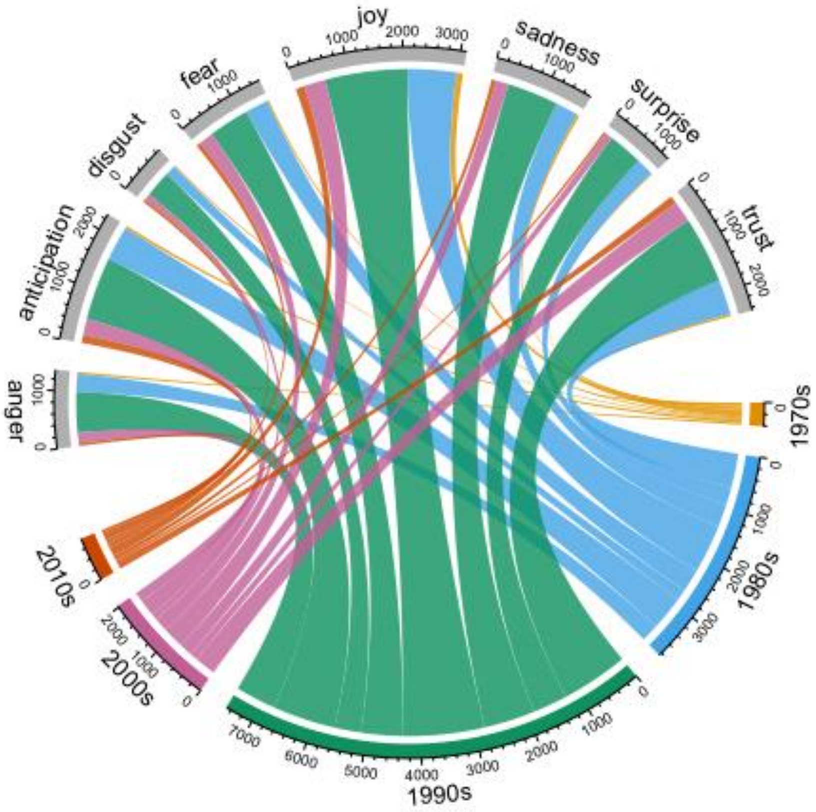
Additional variables can be categorical or continuous

Association between variables

Relationship Between Chart and Decade

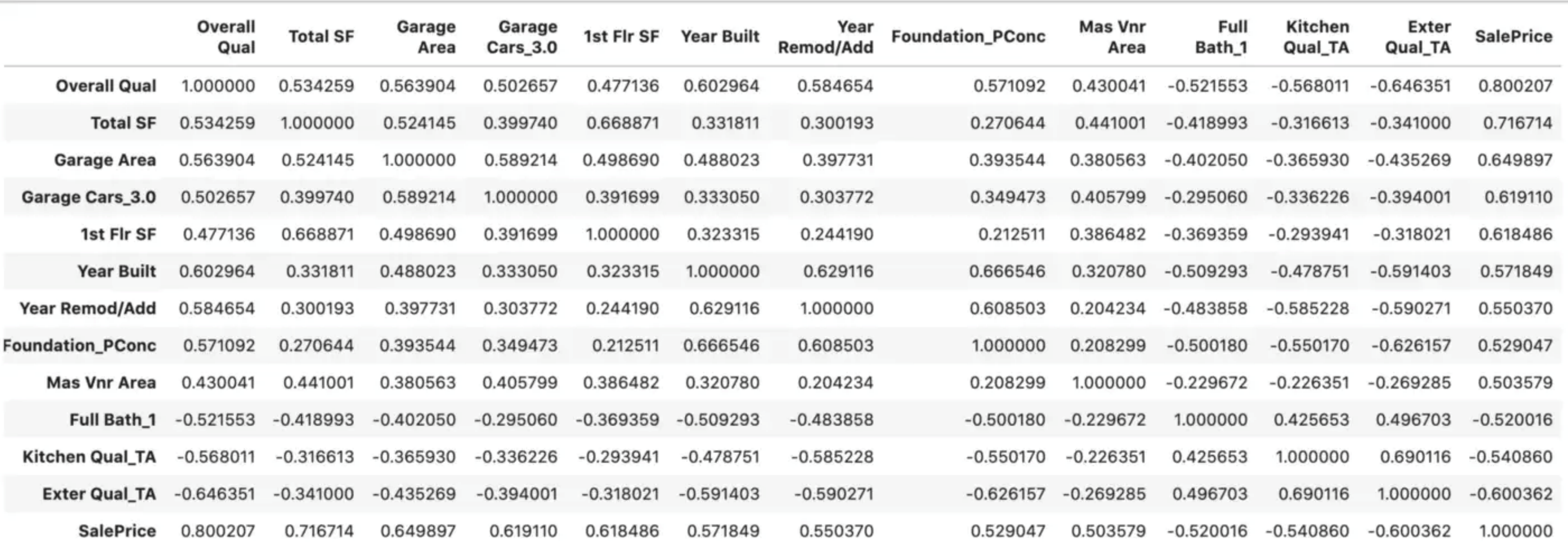


Relationship Between Mood and Decade



Association between variables

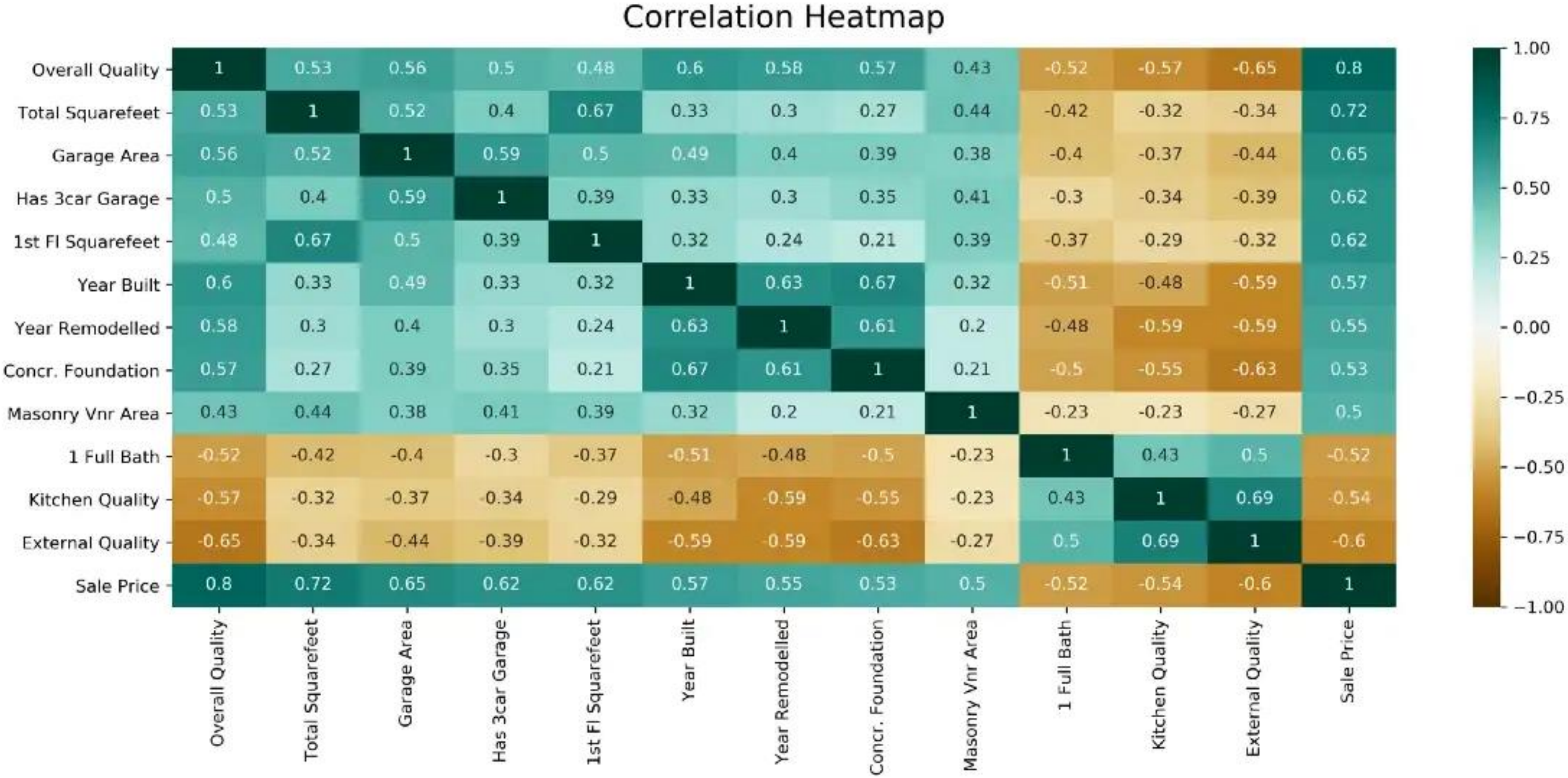
- heat maps depicting correlations



	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

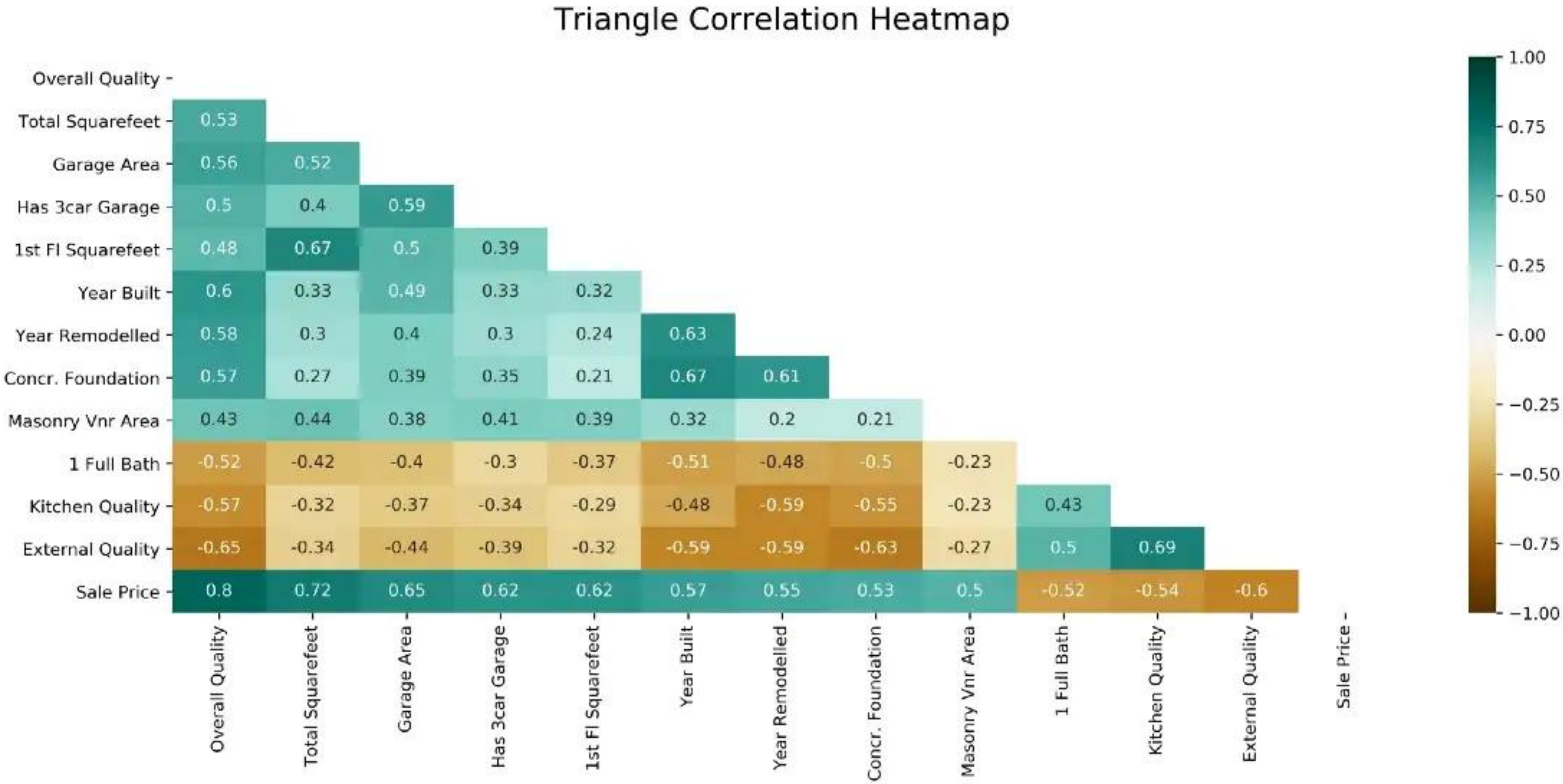


Correlation Heatmap

	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

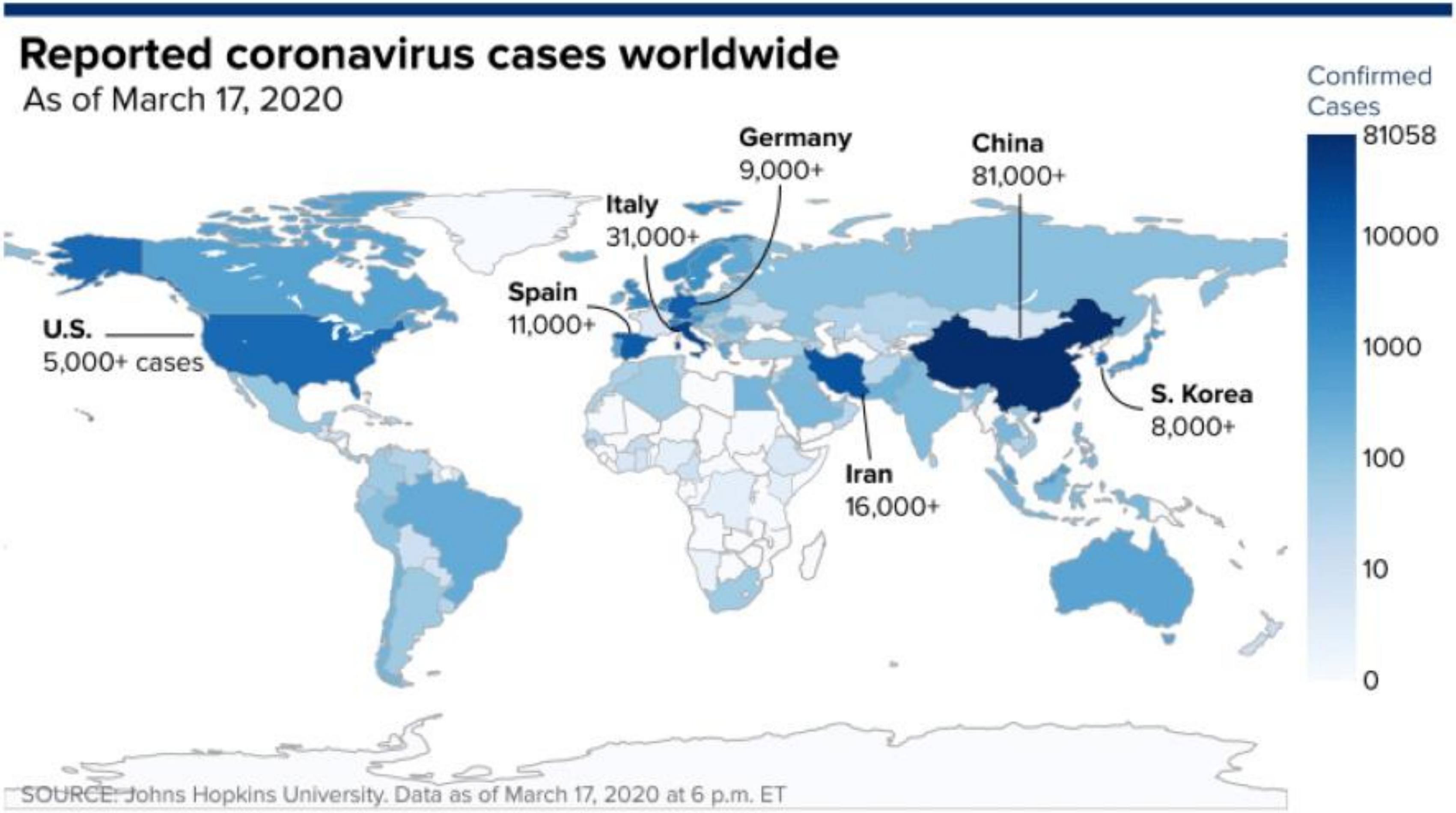
Association between variables

- heat maps depicting correlations



	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



Reported coronavirus cases worldwide

As of March 17, 2020

Germany 9,000+

China 81,000+

Italy 31,000+

Spain 11,000+

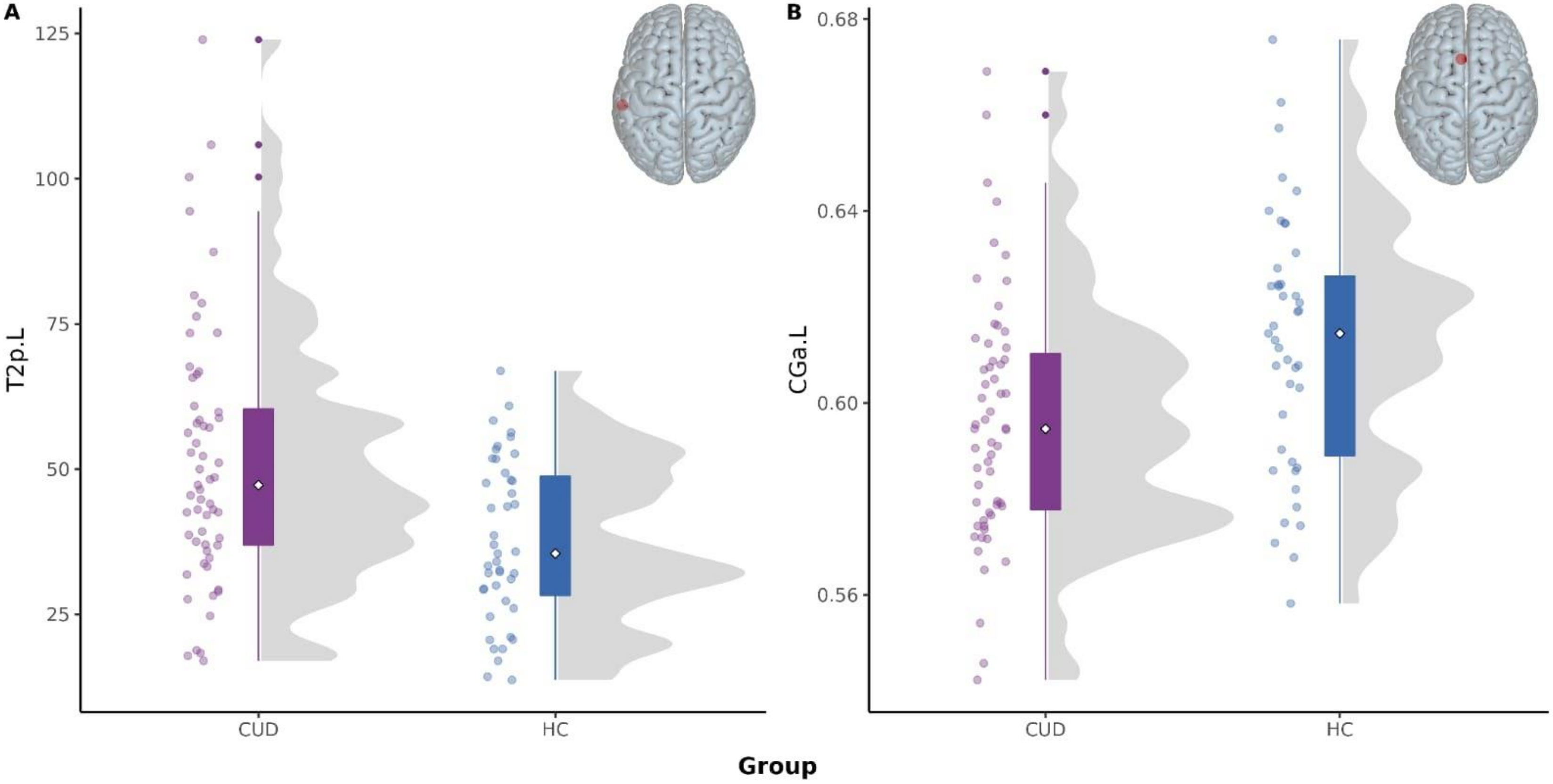
U.S. 5,000+ cases

S. Korea 8,000+

Iran 16,000+

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

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**

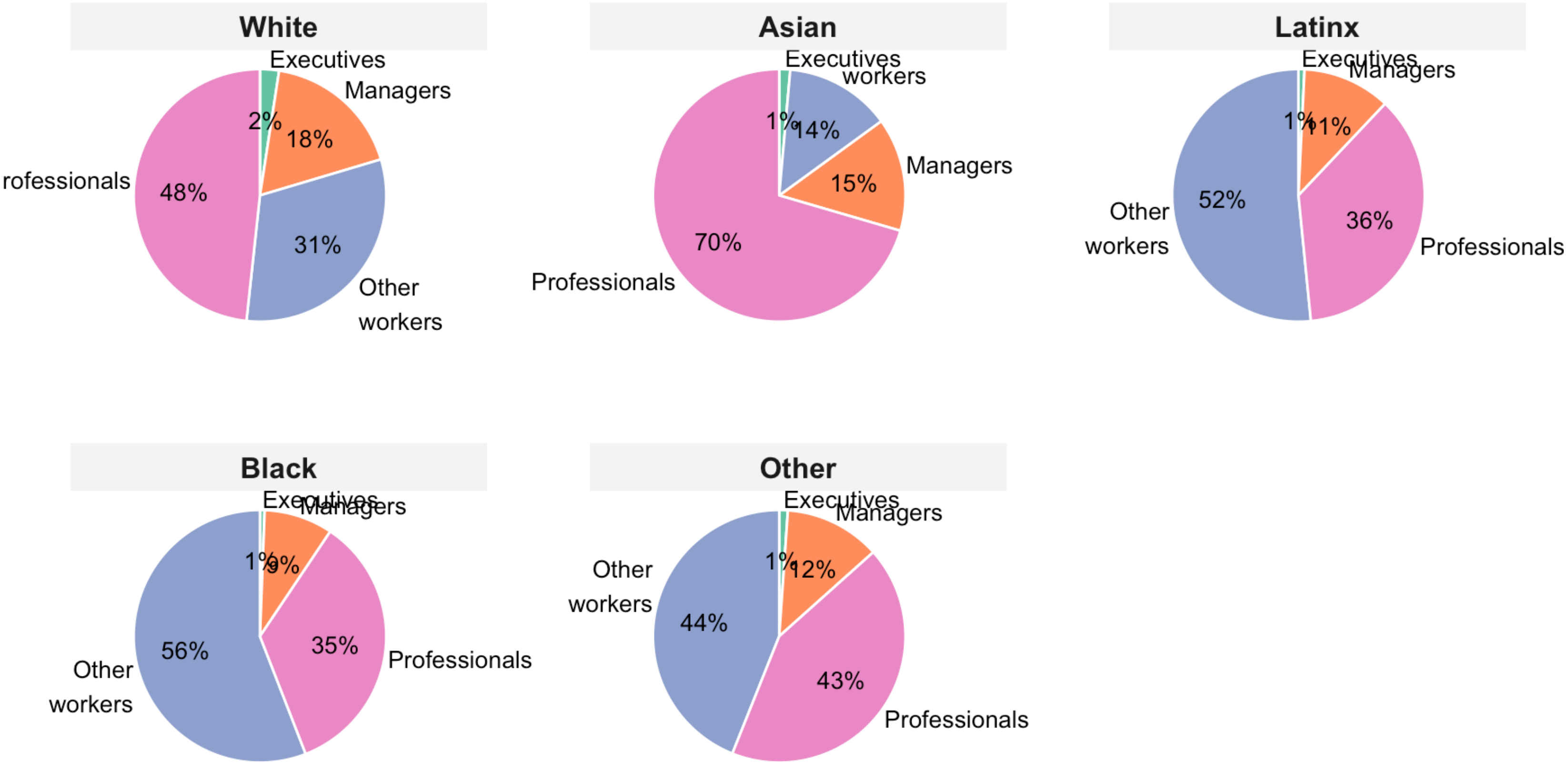
- Avoid using Too Many Variables within a single image which might result in distracting the viewers
- Be extremely careful of not visualizing data through an Unsuitable or Incorrect visualization format
- While using Scales in Data Visualization in order to depict differences between data points, it 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)

Outline

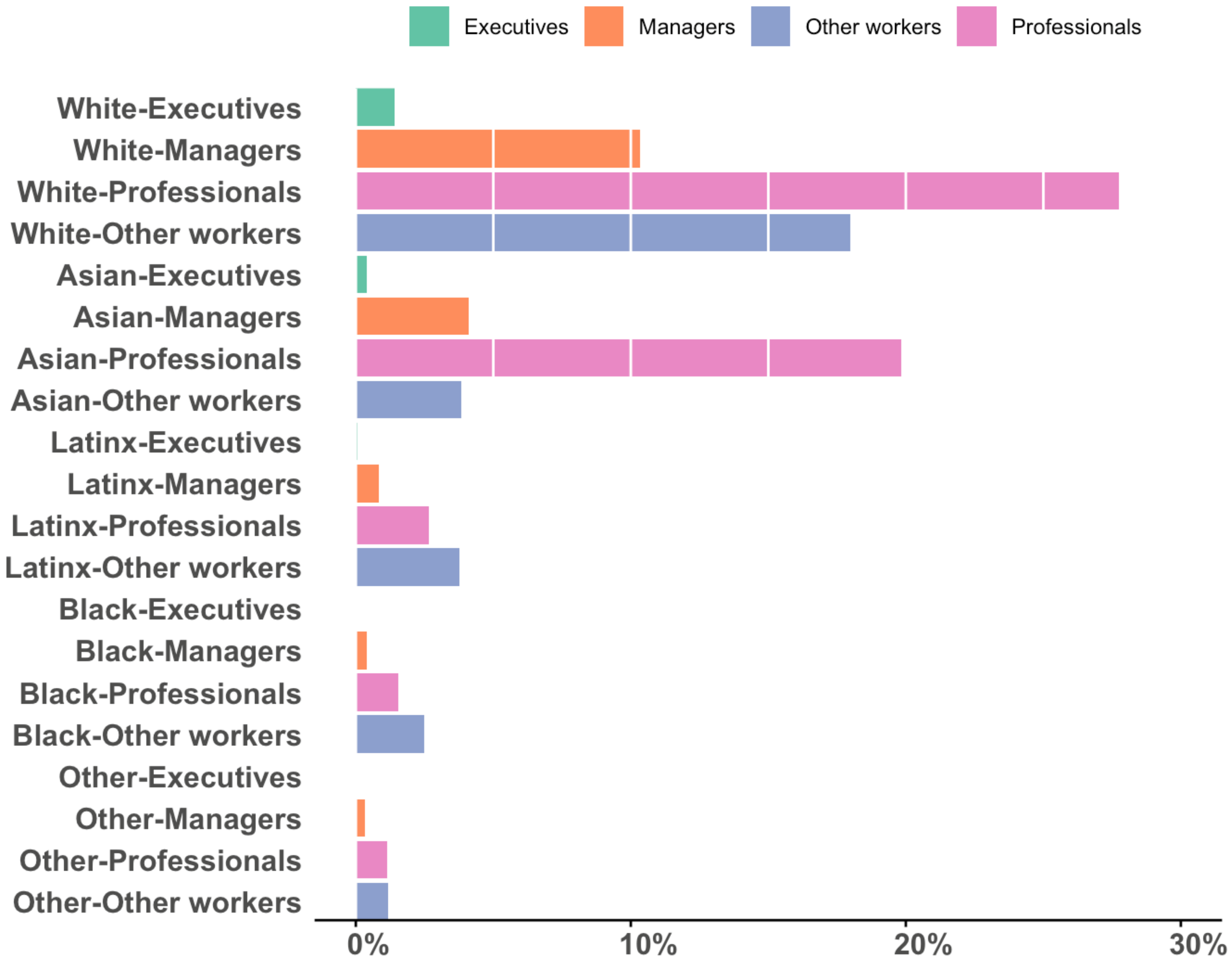
- **Visualization**
    - why we visualise
    - how to pick a plot
    - initial data vs final results visualization (some examples)
    - **bad designs and misleading graphs**
- **Summarization**
    - measures of central tendency & dispersion
    - which measure to pick

Bad Designs & Improvements

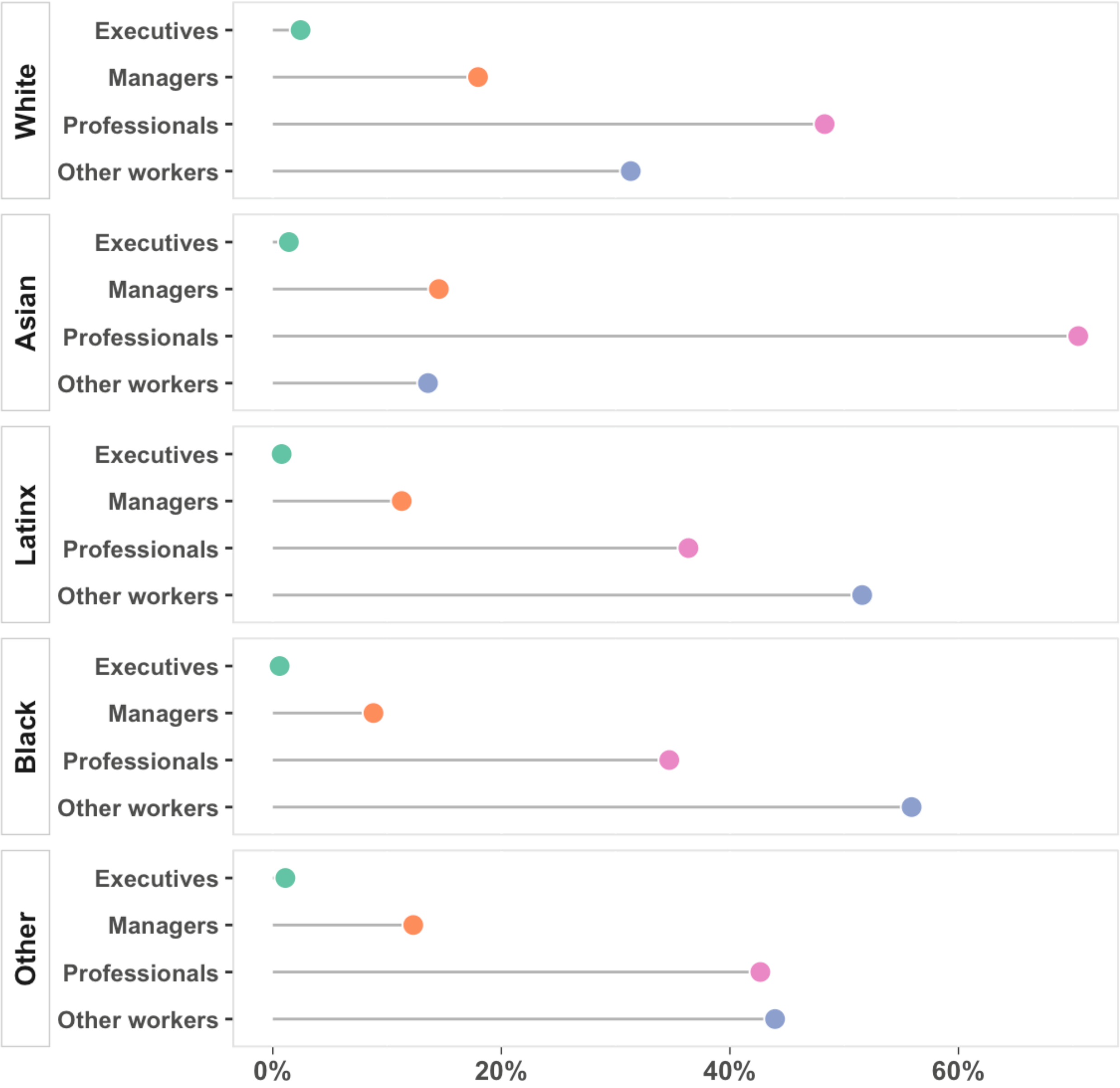
https://nandeshwar.info/data-visualization/pie-chart-vs-bar-chart/



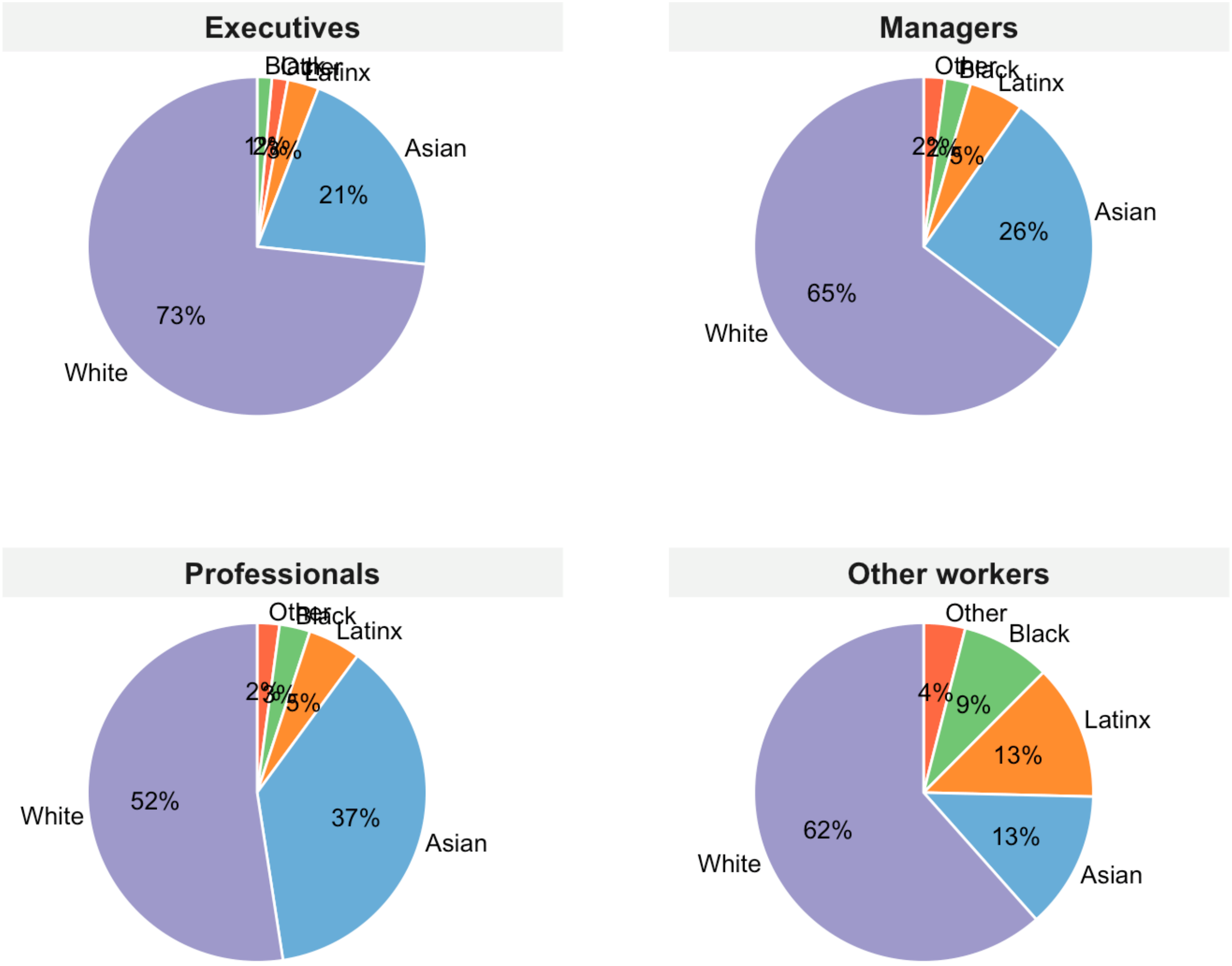
Source: Reveal, https://www.revealnews.org/topic/silicon-valley-diversity/



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



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



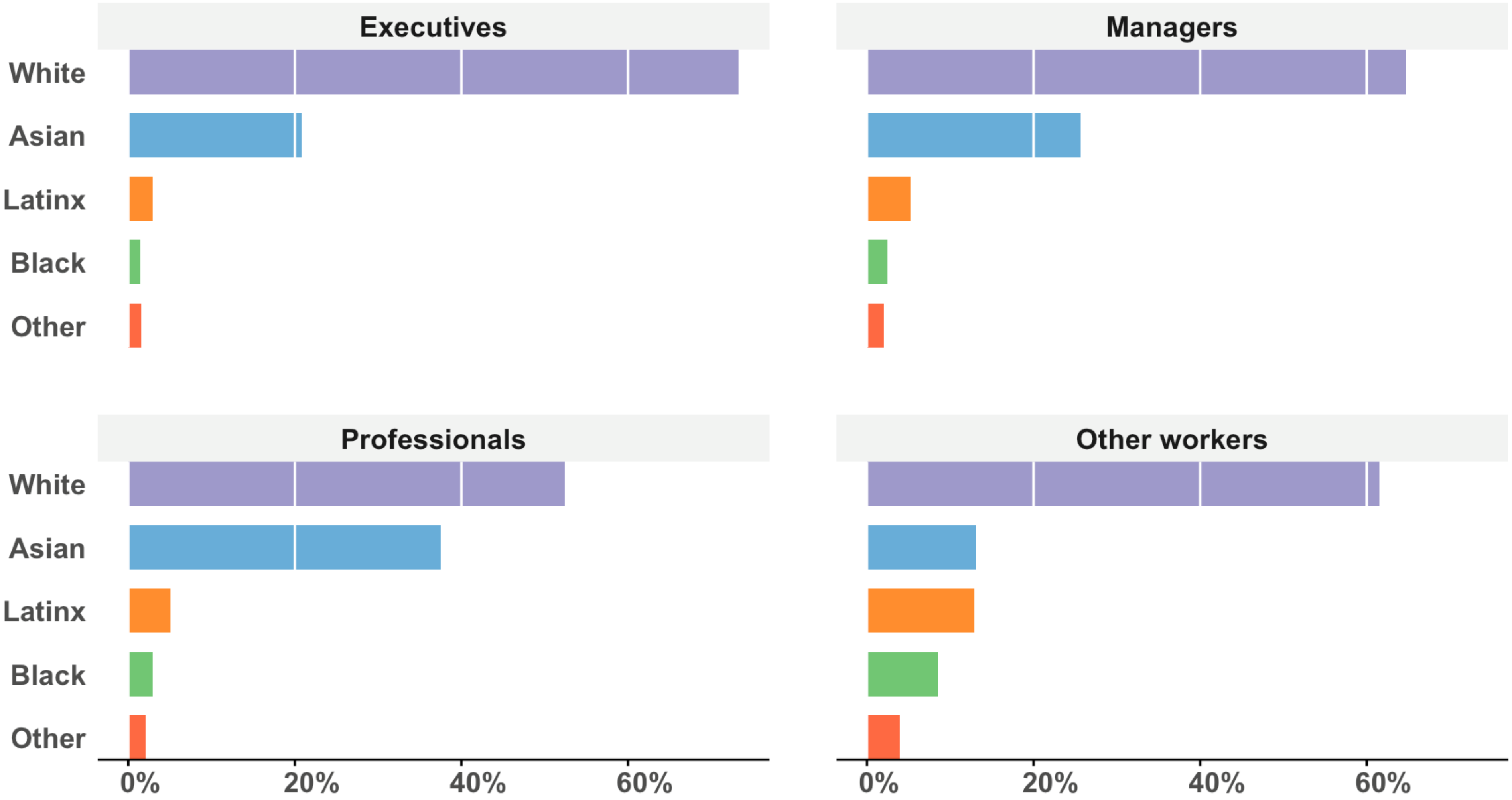
Executives

Managers

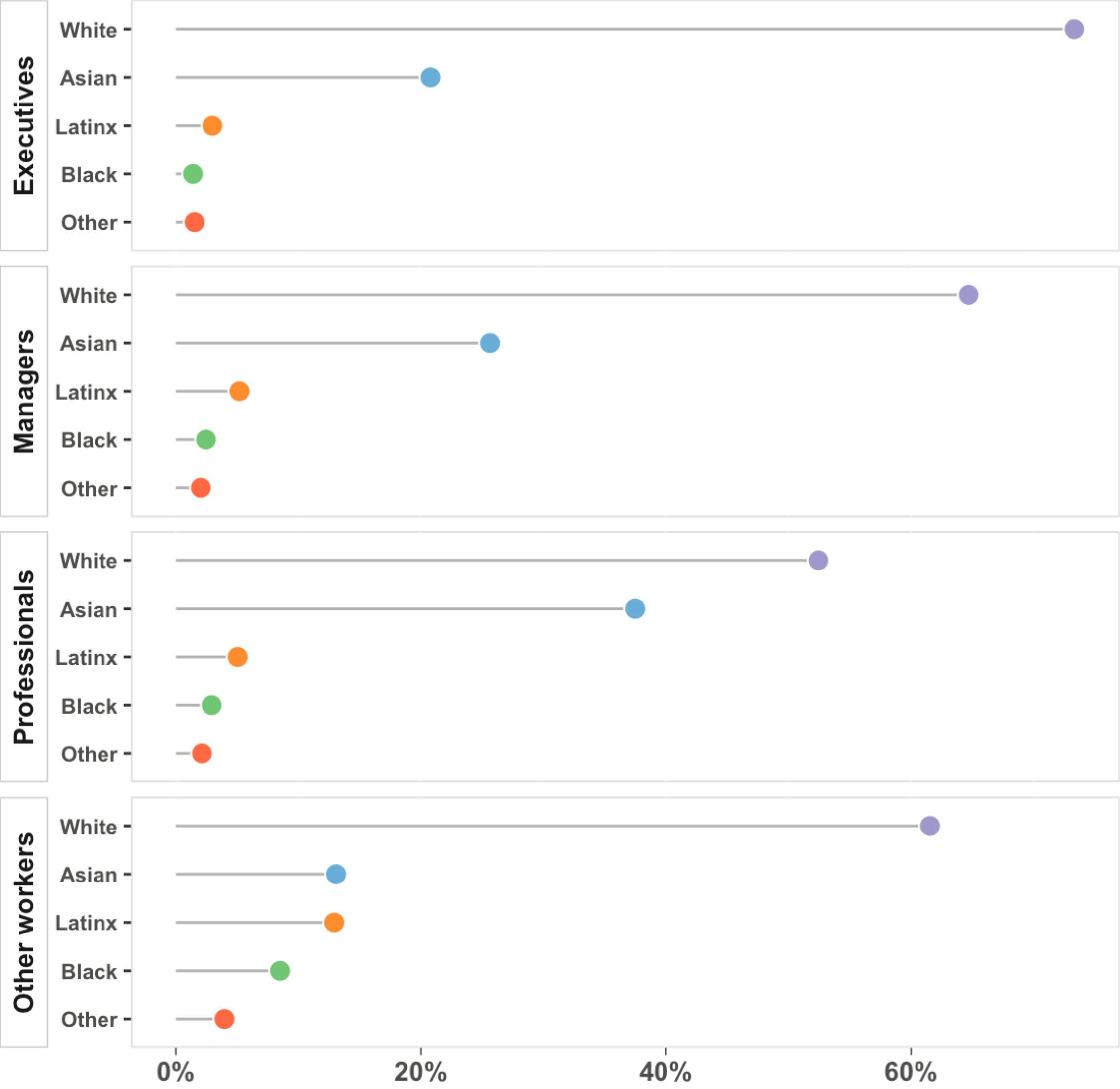
Professionals

Other workers

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

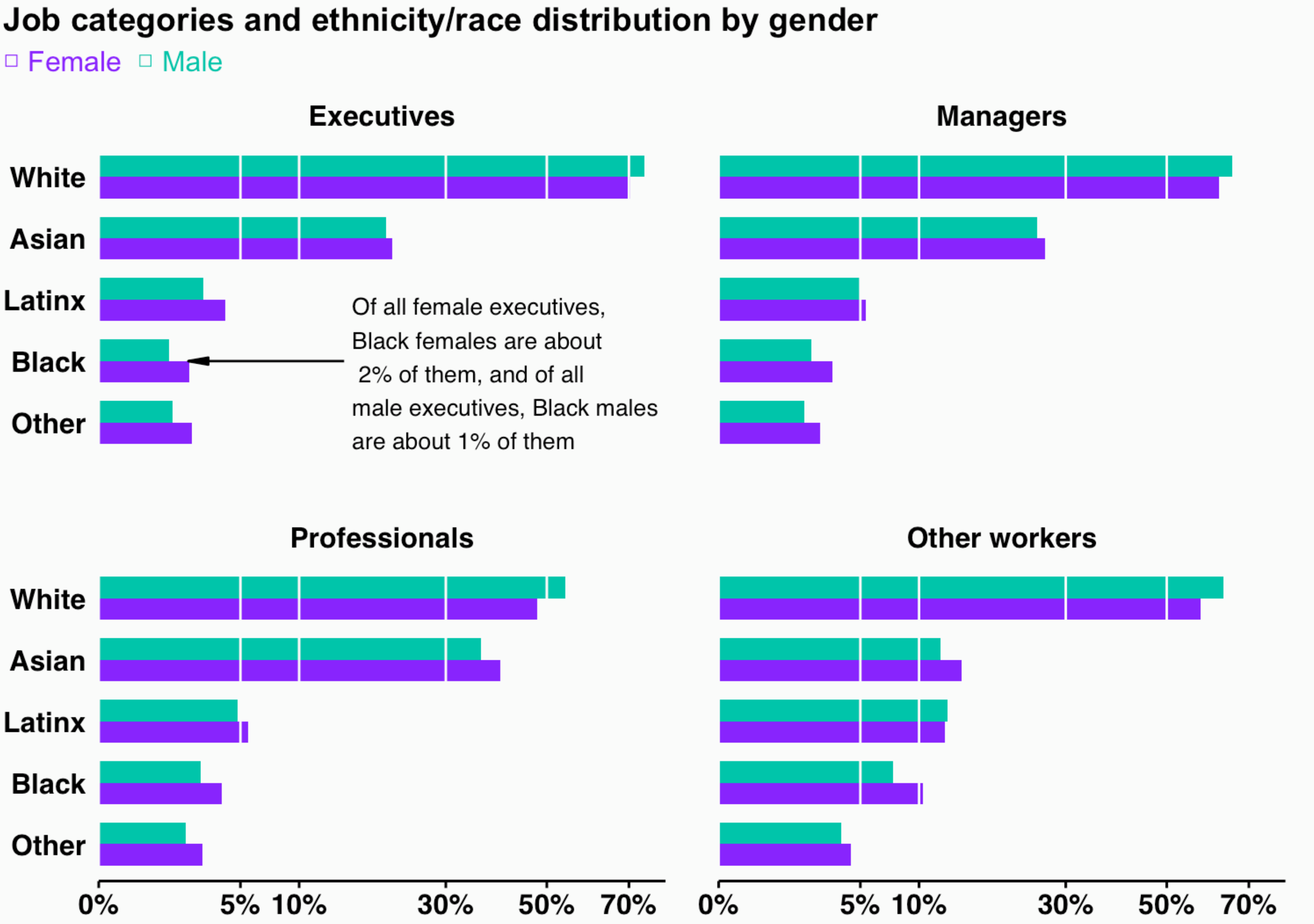


Source: Reveal, https://www.revealnews.org/topic/silicon-valley-diversity/



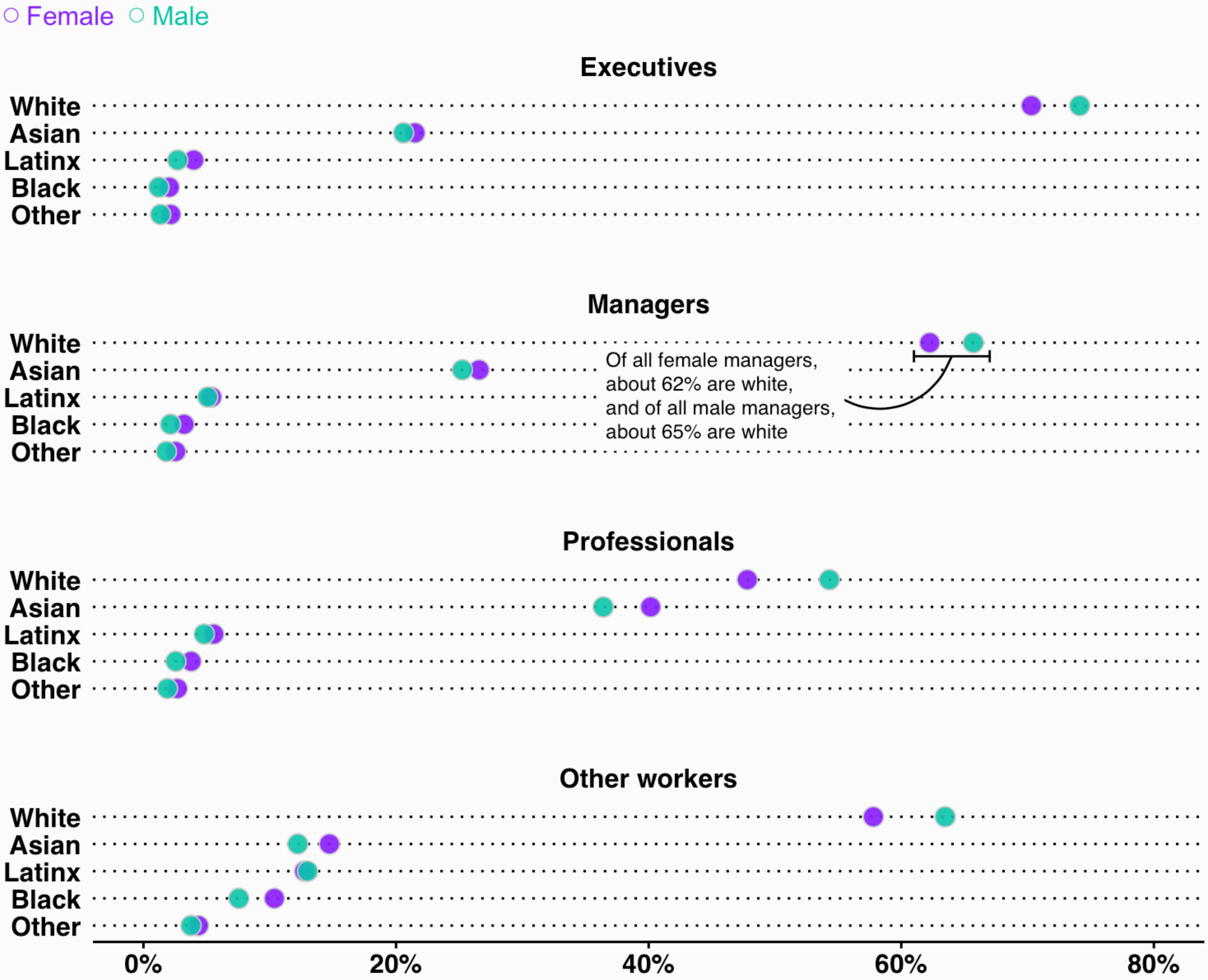
Source: Reveal, https://www.revealnews.org/topic/silicon-valley-diversity/

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



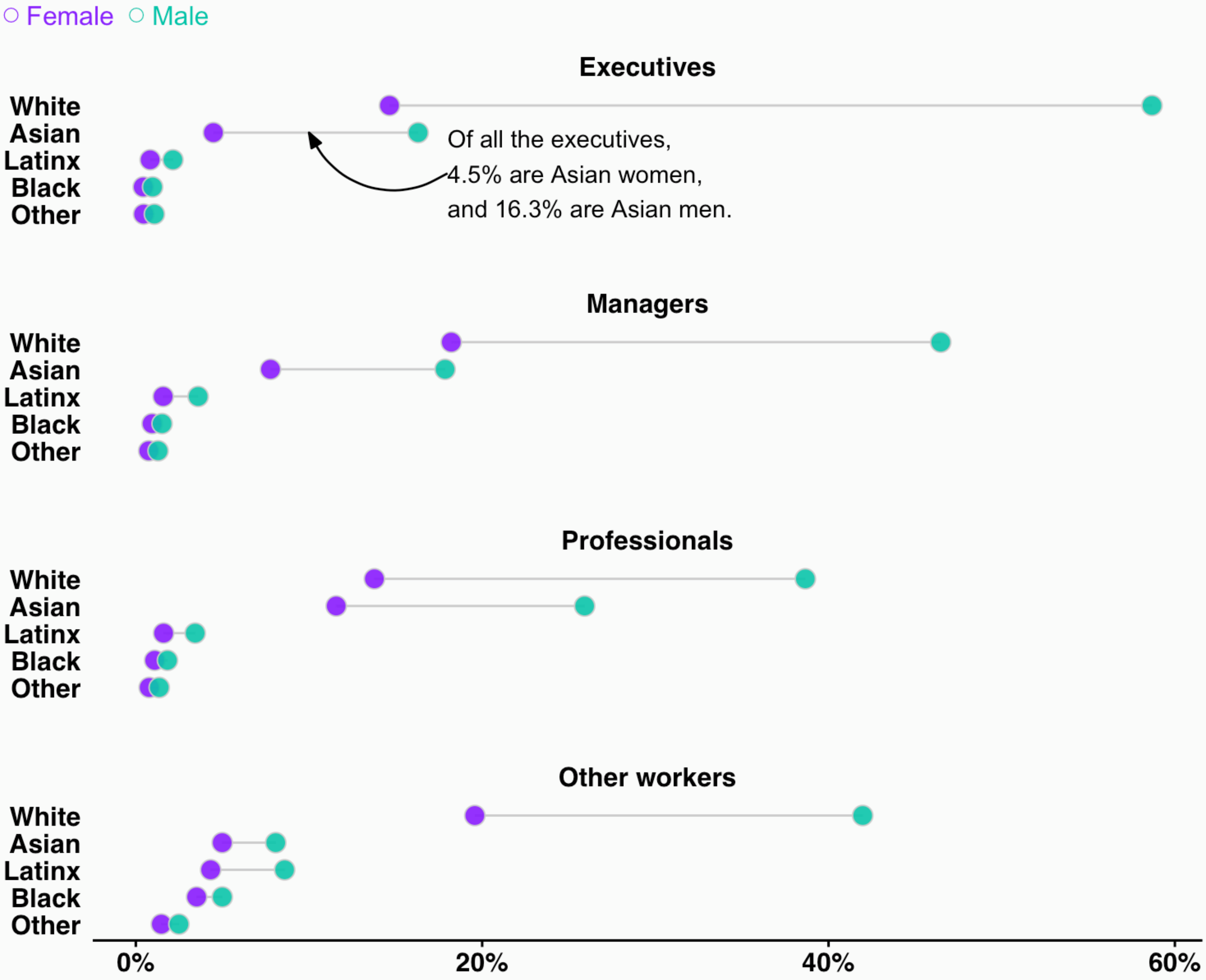
Note: The x-axis is transformed using the square root function to see smaller values. Source: Reveal, https://www.revealnews.org/topic/silicon-valley-diversity/

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



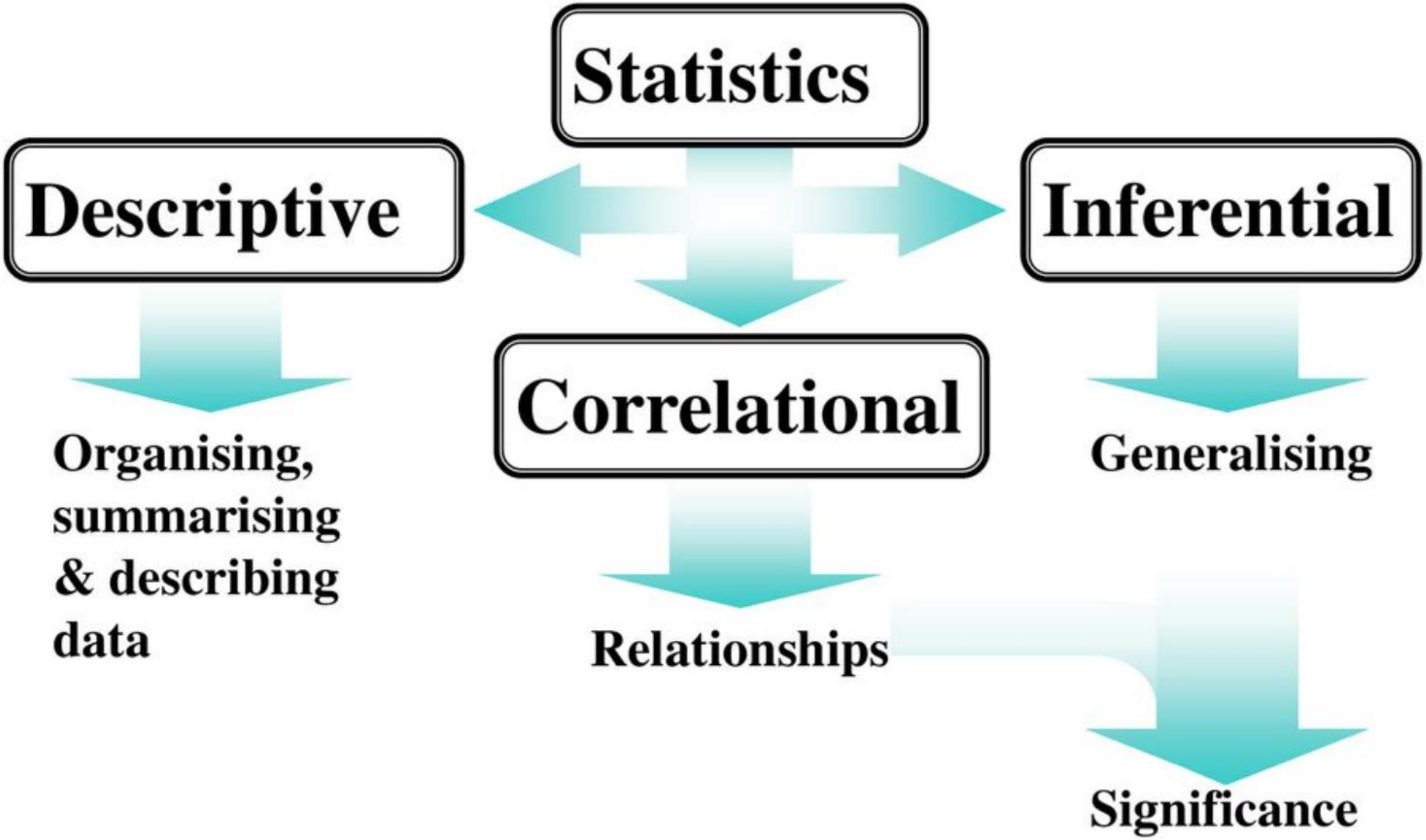
Of all the executives, 4.5% are Asian women, and 16.3% are Asian men.

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

Outline

VISUALISATION ASSIGNMENT DUE 03/Feb/2026 EoD

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



Statistics

Descriptive

Correlational

Inferential

Organising, summarising & describing data

Relationships

Generalising

Significance