
Becoming a Data Detective

How to Recognize Forms of Bias in Data Storytelling



David Cuervo, John Montgomery, Zach Gonzales



TEXAS
Department of Family
and Protective Services
Analytics for Change
& Evaluation

Objectives

- ❑ Introduction of our work
- ❑ How design impacts our perception
- ❑ Cognitive Bias in data
- ❑ Transparency in data
- ❑ Data Detective tips



Who We Are

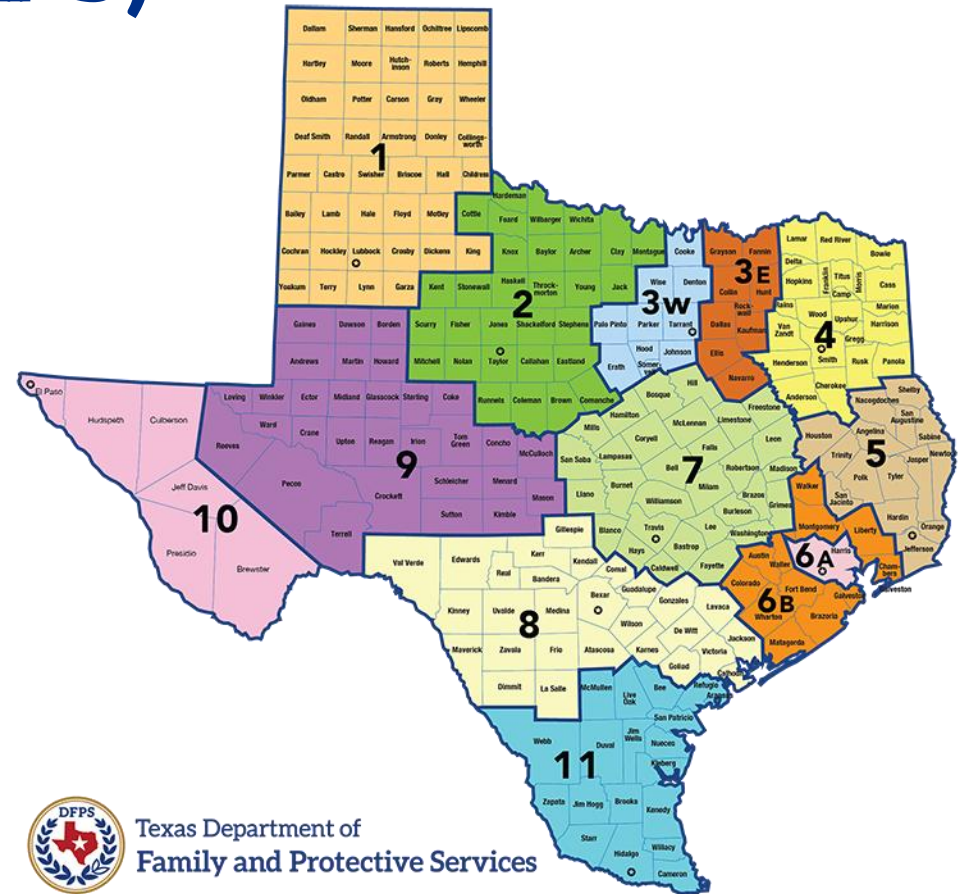


TEXAS
Department of Family
and Protective Services



Texas Department of Family and Protective Services (DFPS)

- Mission: We promote safe and healthy families and protect children and vulnerable adults from abuse, neglect, and exploitation.
- ~28,000 children in DFPS custody during FY2024
- ~23,000 families received Family Preservation Services in FY2024



DFPS Programs

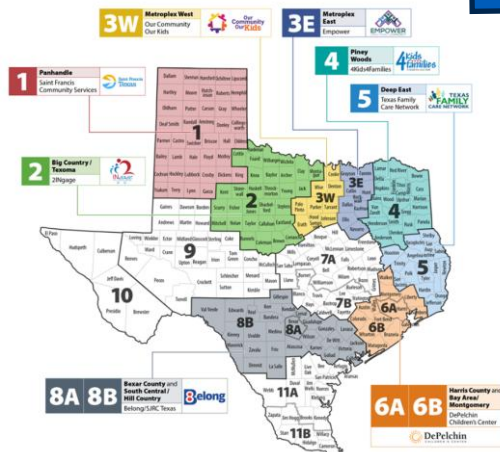
Program	Description
Adult Protective Services	Protects the elderly and people with disabilities from abuse, neglect, and exploitation through investigations and services.
Child Protective Services	Protects children from abuse and neglect through services, foster care, and adoption.
Child Protective Investigations	Investigates reports of child abuse or neglect to determine if any child in the family has been abused or neglected. If the children aren't safe, the investigator starts protective services.
Statewide Intake	Takes reports of abuse, neglect, and exploitation from across the state through its Texas Abuse Hotline (1-800-252-5400) and through the website TxAbuseHotline.org 24-hours a day, every day of the year. Professional, passionate, and innovative.

Our Work



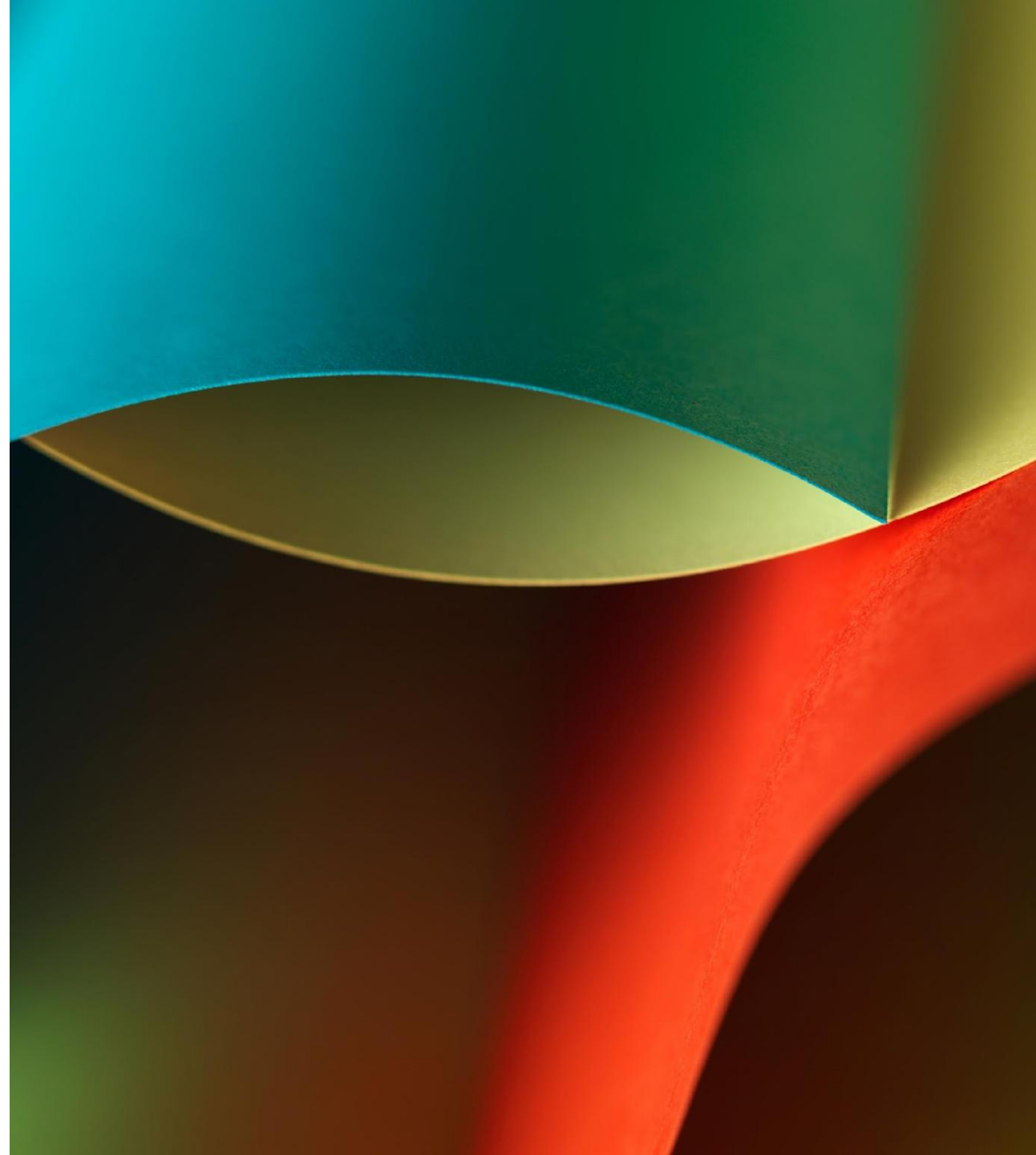
TEXAS
Department of Family
and Protective Services

**Visualization and
Interpretation for Change
and Enhancement**



ADMINISTRATION FOR
CHILDREN & FAMILIES

Data Visualization Design



Storytelling through Data

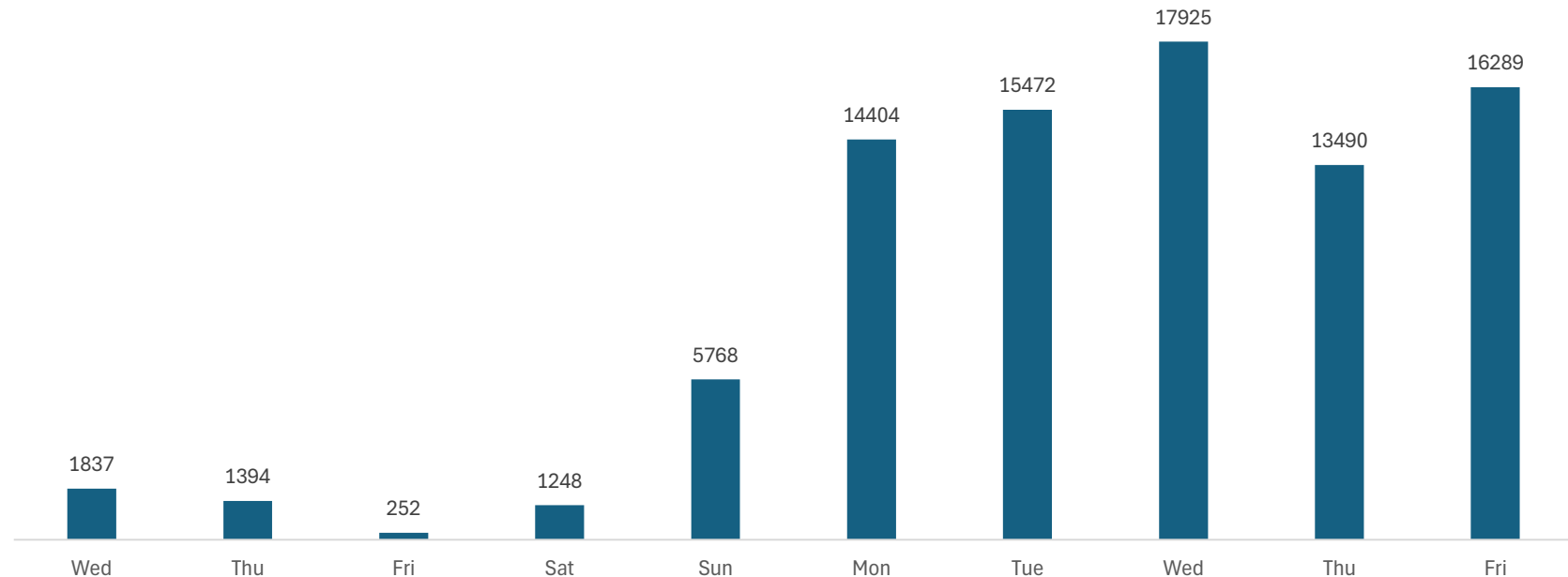
- Data is always being spoken for
- Many decisions shape the creation of a chart, graph, infographic, etc.
 - What is included, what is left out, scale, colors
- Many opportunities for bias to creep in



What Bias Can be Introduced?

Type of Bias	Description
Sampling bias	What gets included/excluded
Selection bias	Choosing data that support a narrative
Axis manipulation	Truncating axes or using non-zero baselines to exaggerate differences
Visual encoding bias	Making use of color, shape, or size to draw attention

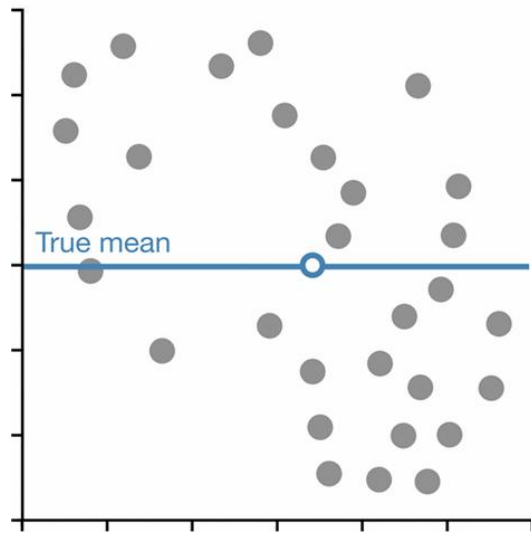
David's Steps Before and During Vacation



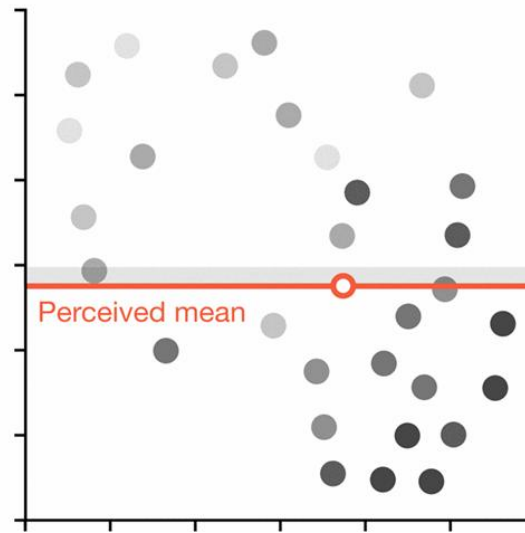
Psychological Impacts of Visualization Techniques

The Weighted Average Illusion: Biases in Perceived Mean Position in Scatterplots (Hong et al, 2022)

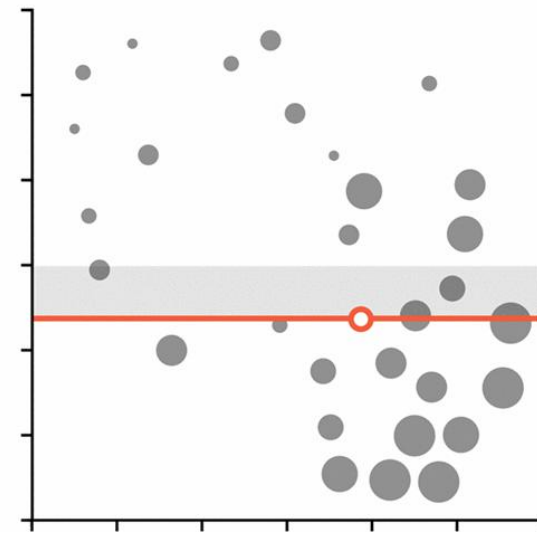
The mean provides a reference for judging data points as above or below average.



If points vary in *lightness* or *size*, the perceived mean position will be biased...

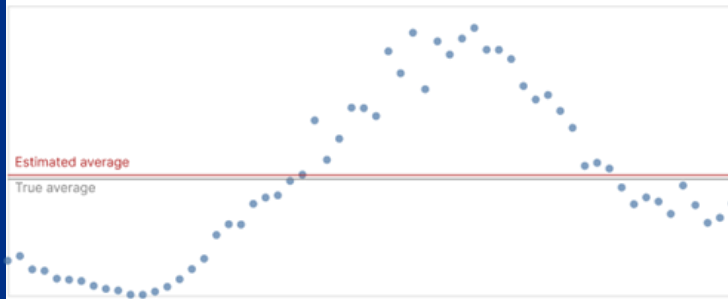


...leading some data points to be misjudged as being above average.

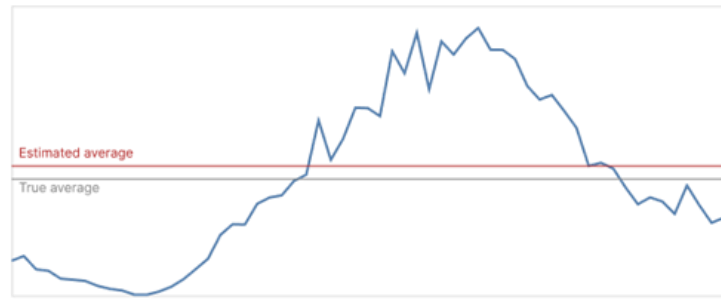


Change in Perception

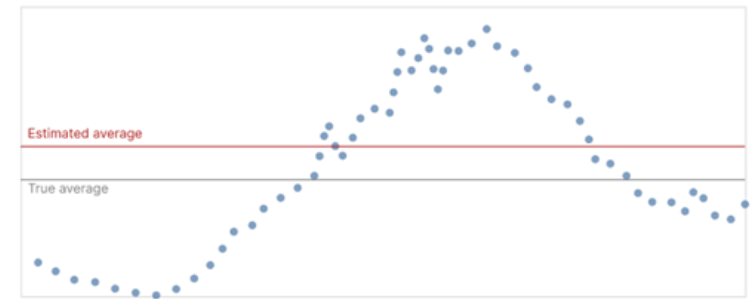
Points equally spaced along the x-axis



Line

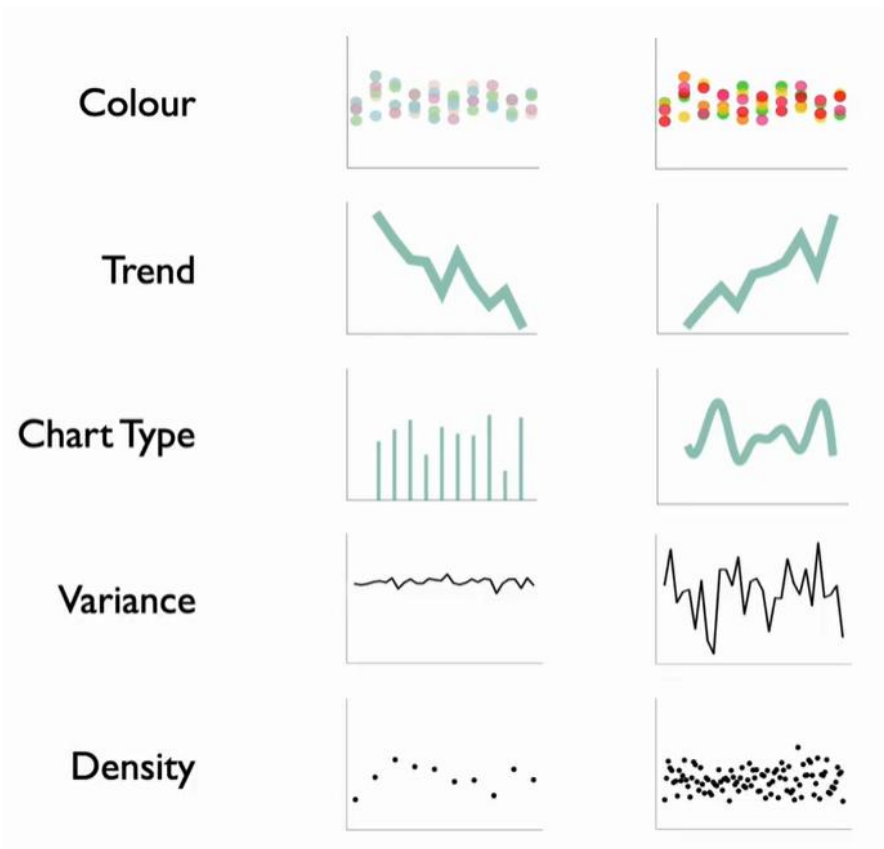


Points equally spaced along the arc of the line

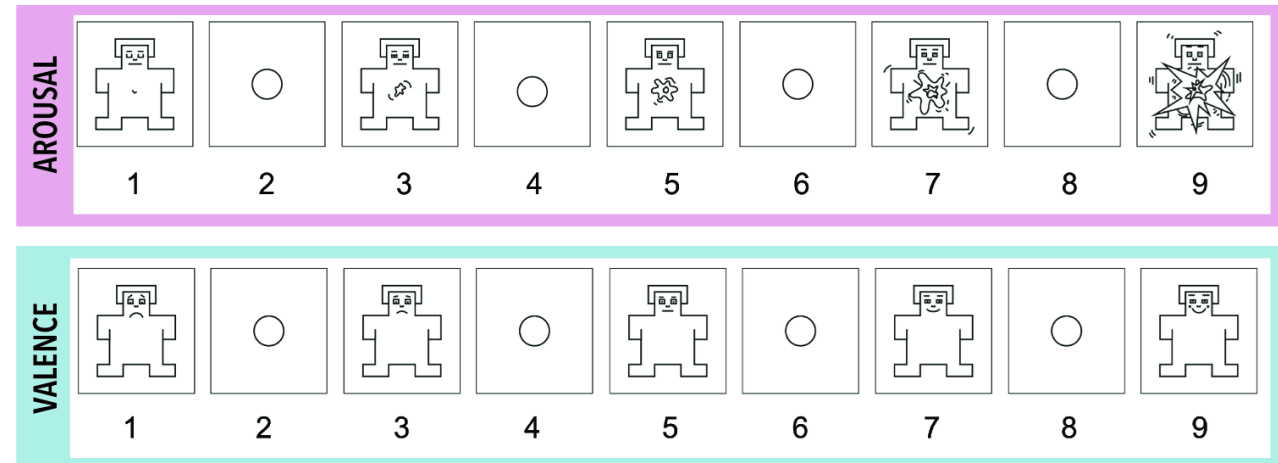


- People's estimates of the average value tends to be pulled toward sections of the graph where the line is more “wiggly”
- More visually salient

Emotional Responses to Data Characteristics



Self-Assessment Manikin (SAM) Scale



All 5 Factors Have an Impact on Our Emotions

S1 S2 S3 S4 S5

Sx (Hy) refers to Hypothesis y in study x

METHOD

The SAM scale and crowdsourced study methods are valid as they allowed us to replicate results from Bartram et al. [7] S1 (H1)

COLOR

The greyscale color palette is not neutral, but is the palette we tested that has the lowest effect on emotion (for both arousal and valence) S1 (H2)

Results from studying color palettes outside of visualizations context are broadly transferable to visualizations - but small variations may exist S2 (H1)

DATA VARIANCE

Higher data variance induces higher arousal S4 (H1) S4 (H2) and lower valence (only from medium to high variance for valence)

Research Method

Visual Features

Immutable Data Characteristics

DATA TREND

Positive data trends induce higher valence and arousal than neutral trends, which in turn induce higher valence and arousal than negative trends S3 (H1) S2 (H2) S2 (H3)

The effect of color palette on emotion (for both arousal and valence) is larger than that of data trend S2 (H4)

DATA DENSITY

Higher data density induces higher arousal S5 (H1)

CHART TYPE

Charts that are good at showing trend like line chart and scatterplot might be more sensitive to the effect of trend on arousal and valence than charts that are not, like bar chart S3 (H2) S3 (H3)

With a flat trend, charts with round features like smoothed line chart induce lower arousal than charts with sharp features like jagged line chart S3 (H4) S4 (H3) S5 (H4)

Data trend has a much stronger effect than chart type on emotion (for both arousal and valence) S3 (H6)

Charts with round features like smoothed line chart induce higher valence than charts with sharp features like jagged line chart, especially when there is a negative data trend S4 (H4) S5 (H5) S3 (H5)

Line charts induce higher arousal than scatterplots S4 (H5) S5 (H6)

Higher data density does not seem to induce higher valence in scatterplots but induces lower valence in line charts - whether jagged or smooth S5 (H2) S5 (H3)

What is Cognitive Bias?

Cognitive bias is a mental shortcut that influences our thinking and decision-making, leading us to process information in a selective and subjective manner, often resulting in inaccurate or irrational judgements.



Hello, I'm Mango!



Think about it...

What formative events in your life are you aware of that influenced building your character and values?



COGNITIVE BIAS CODEX, 2016



We discard specifics
to form generalities

edit and reinforce
memories after the fact

Too Much Information

We notice flaws in others
more easily than we
notice flaws in ourselves

- Subjective influence
- Continued influence
- Semmelweis reflex
- Bias blind spot
- Naïve cynicism
- Naïve realism

- Expectation bias
- Ostrich effect
- Subjective validation
- Continued influence effect
- Semmelweis reflex
- Bias blind spot
- Naïve cynicism
- Naïve realism

We notice flaws in others more easily than flaws in ourselves

We fill in character stereotypes, gender and prior histories

We imagine things and people
we're familiar with or fond of as
better

We imagine things and people we're familiar with or fond of as better

Not Enough Meaning

We simplify probabilities and numbers make them easier to think about

We think we know what other people are thinking

We project our current mindset and assumptions onto the past and future

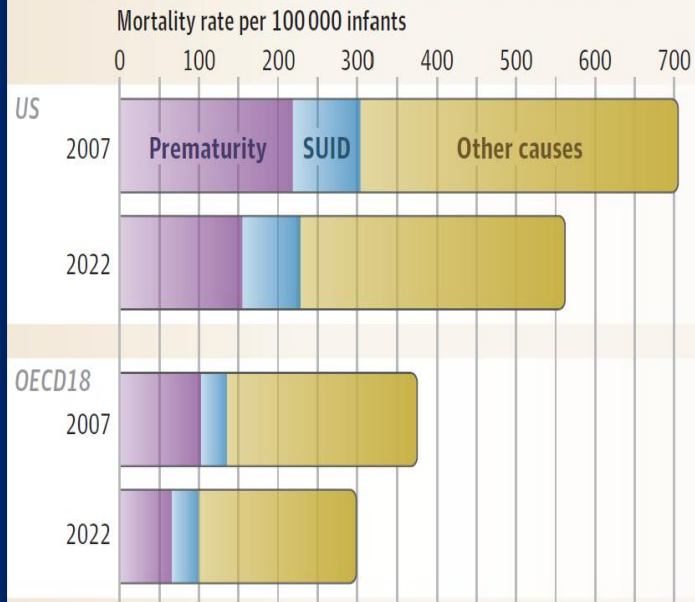
To act, we must be confident we can make an impact and feel what we do is important

Need To Act Fast

To avoid mistakes,
we aim to preserve autonomy
and group status, and avoid
irreversible decisions

Mortality rate of infants in the US vs the OECD18^a

The all-cause mortality rate decreased for all infants <1 y of age between the years 2007 and 2022 but remains higher in the US.

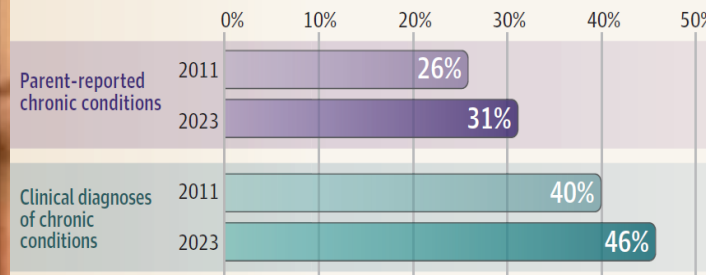


^aThe Organisation for Economic Co-operation and Development (OECD18) comprises 18 comparator nations including Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Ireland, Italy, Japan, the Netherlands, New Zealand, Norway, Spain, Sweden, Switzerland, and the United Kingdom.

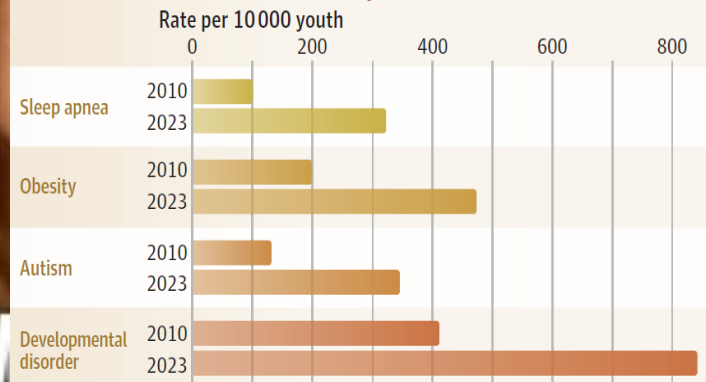
SUID indicates sudden unexpected infant death.

Prevalence of chronic conditions in US youth

Percentage with chronic conditions increased in those aged 3 to 17 y between 2011 and 2023, according to both parents and clinicians.

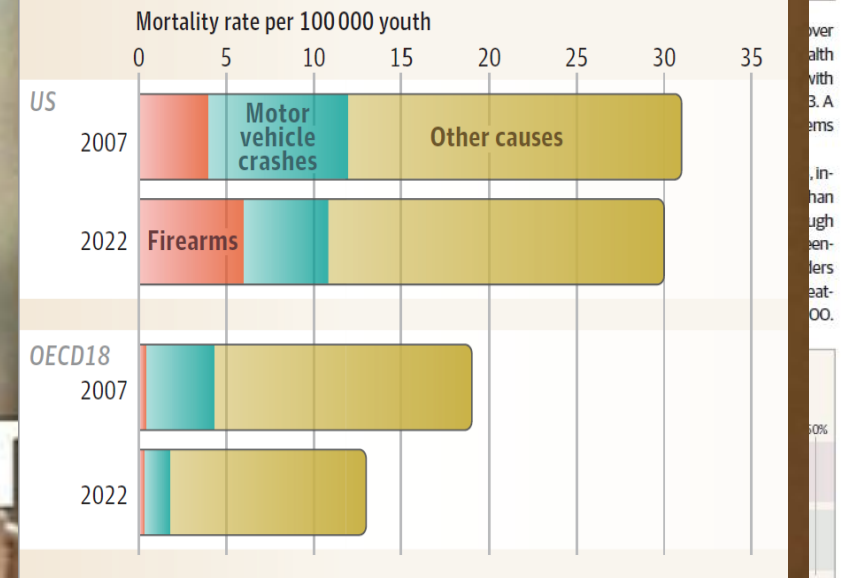


Some chronic conditions increased in prevalence between 2010 and 2023.



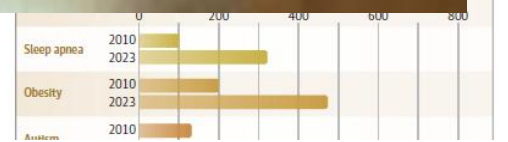
Mortality rate of youth in the US vs the OECD18

Individuals aged 1 to 19 y were more than twice as likely to die in the US as those in the OECD18 in 2022.



gest discrepancy was for firearm deaths, which were more than 15 times higher in the US. Motor vehicle crashes, the second most common cause of death in the US, were 2.5 times higher than in the OECD18.

What Christopher B. Forrest, MD, PhD, lead author of the study and a professor of pediatrics at Children's Hospital of Philadelphia, found particularly "disturbing" was the growing mortality gap be-



Despite ongoing medical advances, the US may be failing to adequately support children’s health. A new study in JAMA examined pediatric mortality and chronic disease to assess the state of children’s health in the US over the past 2 decades compared with other high-income countries

JAMA Data Brief

More US Children Die Than in Other High-Income Nations

Samantha Anderer; Karen Bucher, MA

operation
surveys, a
atric health
Altho
1 year in b
mained comparatively high
infants in the US were 2.2 tin
2.4 times more likely to di
(SUID), which includes sudd

assess the
compare

+ Editorial

+ Related article

operation and Development (OECD18), 5 nationally representative surveys, and electronic health records from a network of 10 pedi-
atric health systems.

^aThe Organisation for Economic Co-operation and Development (OECD) comprises 18 comparator nations including Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Ireland, Italy, Japan, the Netherlands, New Zealand, Norway, Spain, Sweden, Switzerland, and the United Kingdom.
SUID indicates sudden unexpected infant death.

JAMA Data Brief

More US Children Die Than in Other High-Income Nations

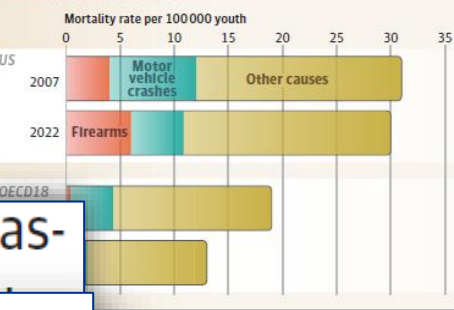
Samantha Anderer; Karen Bucher, MA

Despite ongoing medical advances, the US may be failing to adequately support children’s health. A new study in JAMA examined pediatric mortality and chronic disease to assess the state of children’s health in the US over the past 2 decades compared with other high-income countries.

Researchers carried out a cross-sectional analysis using mortality statistics from the US and 18 comparator nations in the Organisation for Economic Co-operation and Development (OECD18), 5 nationally representative surveys, and electronic health records from a network of 10 pedi-

Mortality rate of youth in the US vs the OECD18

Individuals aged 1 to 19 y were more than twice as likely to die in the US as those in the OECD18 in 2022.



cross-sectional analysis using mortality statistics from the US and 18 comparator nations in the Organisation for Economic Co-

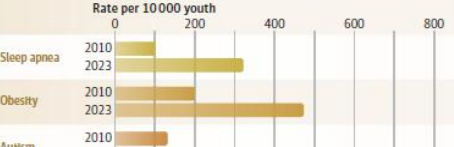
also tracked trends in chronic conditions in US youth over reported data from the National Survey of Children’s Health that the prevalence of people aged 3 to 17 years old with chronic conditions increased from 26% to 31% from 2011 to 2023. A clinician-recorded data from 10 pediatric health systems showed a similar increase, from 40% to 46%. The prevalence of sleep apnea tripled between 2010 and 2023, increasing from 100 to 321 per 10 000 youth. Rates of obesity more than doubled in the past decade. Autism rates nearly tripled, although research has suggested that factors including improved screening, speech delays, and intellectual disability had the greatest influence over time, increasing from 410 to 841 per 10 000.

Prevalence of chronic conditions in US youth

Chronic conditions increased in those aged 3 to 17 y from 2011 to 2023, according to both parents and clinicians.



Some chronic conditions increased in prevalence between 2010 and 2023.



Mortality rates decreased somewhat consistently for people aged 1 to 19 years in the OECD18 over the past 15 years while rising in the US in 2015 and again in 2020. From 2020 to 2022, youth in the US were more than twice as likely to die as their OECD18 counterparts. The biggest discrepancy was for firearm deaths, which were more than 15 times higher in the US. Motor vehicle crashes, the second most common cause of death in the US, were 2.5 times higher than in the OECD18.

What Christopher B. Forrest, MD, PhD, lead author of the study and a professor of pediatrics at Children’s Hospital of Philadelphia, found particularly “disturbing” was the growing mortality gap be-

Qualities of Good Storytelling in Data



Jargon: BAD Storytelling



Definition:

Jargon refers to specialized language used by a particular group, profession, or field that may be difficult for outsiders to understand.

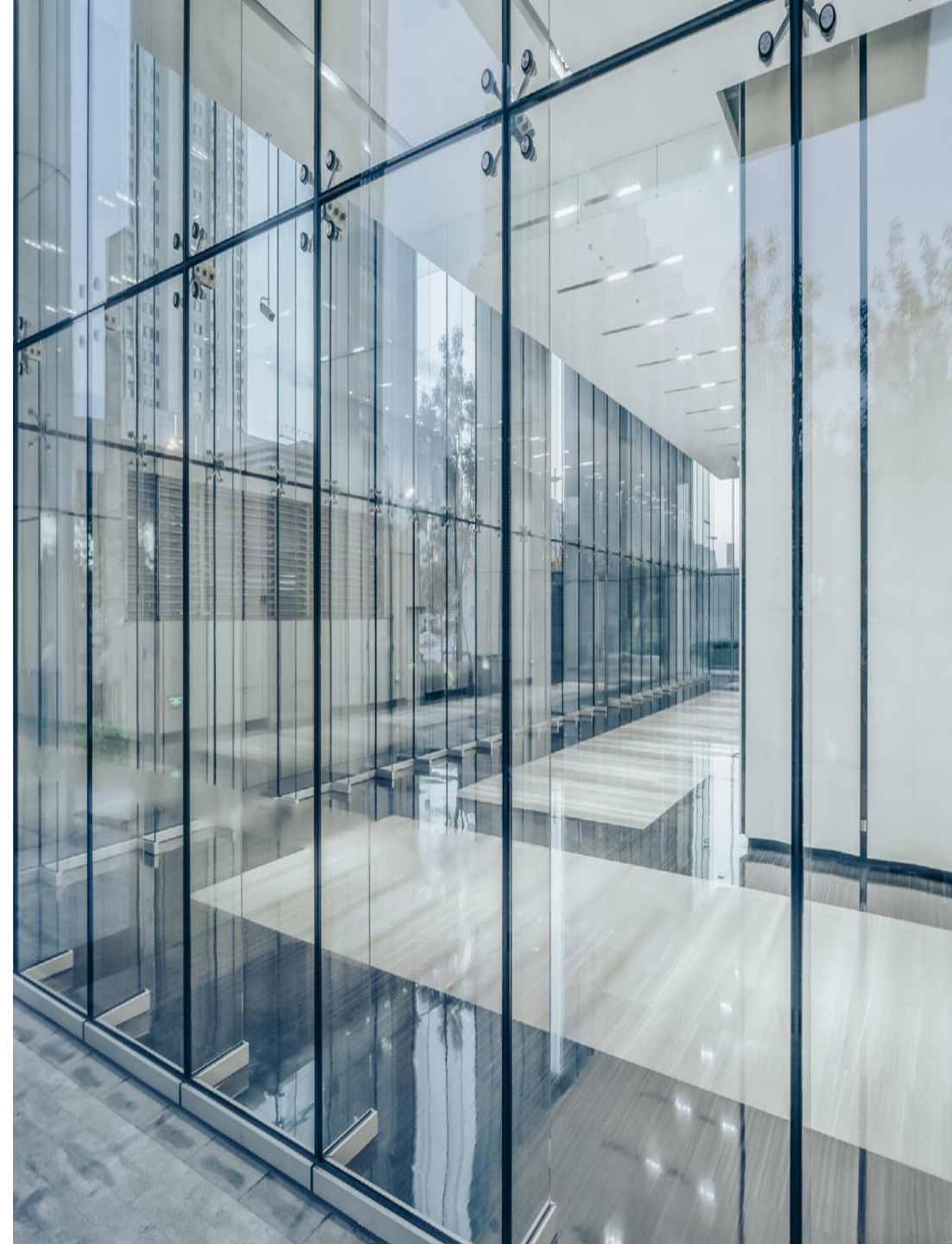
Medical Jargon: "The patient was diagnosed with CVA and requires an MRI to assess the extent of the ischemia."

What is the statistician trying to say?

Term	Standard English	Statistician English
“Significant”	“Important”	“An effect has been observed”
“Uniform”	“Without Variation. Sameness”	“Evenly spread over a range (e.g. uniform distribution or design”
“Random”	“Haphazard. Wild. Unpredictable”	“an experiment where the assignment of treatments is deliberately determined by chance”
“Normal”	“Conforming to general standards”	“related to the bell-shaped curve (e.g., the Normal distribution”

Full Disclosure of Methodology

- A good analyst/analysis will provide their/its methodology in a clear, understandable way
 - Examples include showing and explaining Outliers, Nulls, Variance
- An analyst/analysis should be clear in what their/its limitations were
 - Example – providing information on the sample used in the original data collection
 - Example – informing if any data was omitted when performing the analysis



Takeaways

What to be mindful of
moving forward



Detective Tips

- Don't Accept Data at face value
- Be inquisitive
 - "Is this graph employing tricks to be more visually exciting?"
 - "Is this graph showing me the whole picture? Is this the right graph to display this information?"
- Be mindful of Bias
 - "Who created this data I'm seeing? Do they have blind spots? Do I have blind spots?"
- Look for transparency
 - "Is there language used that I don't understand?"
 - "Can I easily access the steps and methods taken to get to the outcomes of this analysis?"



Key Takeaways

- Data can be manipulated visually – look out for common tricks
- Storytellers and listeners come with bias – be aware
- Clear communication is a sign that data can be trusted



Q&A

Thank you!

Height of female popstars:

