UBC CPSC 436V Midterm SOLUTIONS

12 Mar 2020

Closed book, no electronic devices. Cell phones must be turned off and put away. Place your photo ID face up on your desk.

One single-sided sheet of notes (8.5"x11") is allowed, keep it so that you can reuse it for the final.

Do not open the exam until told to do so. Answer the questions in the space provided. Do **NOT** write on the back side of the pages.

There are 75 points, you have 75 minutes.

Name _____*SOLUTIONS*_____

Signature _____

Student Number_____

5-digit CS userid _____

Please write your 5-digit CS id in this box

Q1: Attribute types *[8 pts]*

What type of attribute are the following? Circle only one of the three choices.

1.	Type of cheese (eg Swiss, Brie)	*categorical*	ordinal	quantitative
2.	Eye color (eg brown, blue)	*categorical*	ordinal	quantitative
3.	Class mark (eg C, A, F)	categorical	*ordinal*	quantitative
4.	Tire pressure (eg 60 psi, 70 psi)	categorical	ordinal	*quantitative*
5.	First name (eg Alice, Bob)	*categorical*	ordinal	quantitative
6.	ISBN numbers (eg 978-1-4665-0891-0)	*categorical*	ordinal	quantitative
7.	Unemployment rate (eg 6%, 3.72%)	categorical	ordinal	*quantitative*
8.	Starbucks drink sizes (eg venti, grande)	categorical	*ordinal*	quantitative

Explanations: For categorical - cheese, eye color, and name aren't ordered. ISBN is categorical also - although it's designated through numerals they simply serve as unique keys, those numerals do not encode any ordering and it is meaningless to do arithmetic on these numbers. (It is not safe to assume numerals necessarily imply ordering!) For quantitative - tire pressure and unemployment are quantities you can do arithmetic on. For ordinal - course marks are ordered (C comes between A and B and F

is last) but you can't subtract; Starbucks drinks are also ordered (grande comes between tall and venti) but again you can't subtract.

Q2: True/False *[12 pts]* Circle only one of the two choices.

1.	T	F	The color channels of luminance and hue have similar characteristics because they both convey magnitudes
2.	T	F	The combination of {FirstName, LastName} is a suitable unique key in a table with 10K items representing people
3.	T	F	Showing a distribution requires multiple attributes
4.	T	F	A continuous colormap to show sequential quantitative attributes should have fully saturated hues on each end and a zero point in the middle with a desaturated color such as white or grey
5.	T	F	Temporal data may have both cyclic and hierarchical structure
6.	T	F	The {action, target} pair of compare shapes describes an appropriate task for tabular data
7.	T	F	Scatterplots are an appropriate visual encoding for the task of showing correlation between two quantitative attributes
8.	T	F	The strategy of deriving new data can be combined with the strategy of a single view that changes over time
9.	T	F	Aligned area is the most accurately perceived visual channel
10.	T	F	Roughly 3-4 bins of categorical color are discriminable when the colored regions are small and scattered
11.	T	F	Position and color hue are fully separable channels, but size and color are not
12.	T	F	The human perceptual system is well suited for delivering relative judgements for the hue channel and absolute judgements for the saturation and luminance channels
13.	T	F	Line charts are well suited for the task of assessing trends
14.	T	F	Heatmaps are compact and highly scalable but need to have their rows and columns reordered.
15.	T	F	Radial layouts are particularly suitable for cyclic data
16.	T	F	Choropleth maps can be misleading when there is a lot of variability in region size
17.	T	F	Segmented rainbow colormaps are suitable for categorical data because they are highly saturated
18.	T	F	Rainbow colormaps are suitable for spatial data and quantitative attributes in tables because they are continuous
19.	T	F	Bivariate colormaps are often misinterpreted because 8% of men are red-green colorblind
20.	T	F	Animated transitions can be used for constrained navigation
21.	T	F	Partitioning data into juxtaposed small multiple views is suitable for tabular data but not spatial data
22.	T	F	Linked highlighting provides a way to show the connections between items across multiform multiple views
23.	T	F	Blue-orange colormaps or redundantly encoding both hue and luminance are both suitable alternatives to red-green colormaps for colorblind users.
24.	T	F	Interactive navigation uses a camera metaphor with a changing viewpoint.

Explanations:

1. *False because hue does not convey magnitude*
2. *False because extremely unlikely that name alone would be a unique key when datasets are large enough*
3. *False because showing a distribution is an action that requires only a single attribute as the target.*
4. *False because the described colormap is appropriate for diverging not sequential data*
5. *True: temporal data can be both cyclic (same month of the next year) and hierarchical (years/months/weeks breakdown)*
6. *False because shape understanding is only a relevant task for spatial data encoded with given spatial position - in tabular data the shape is up to the visualization designer and is not an implicit property of the data.*
7. *True: scatterplots are great for the task of showing correlation and do show quantitative attributes.*
8. *True: you can have derived data in a view that changes over time.*
9. *False because area is not the most accurately perceived channel and it's unclear what 'aligned area' would even mean.*

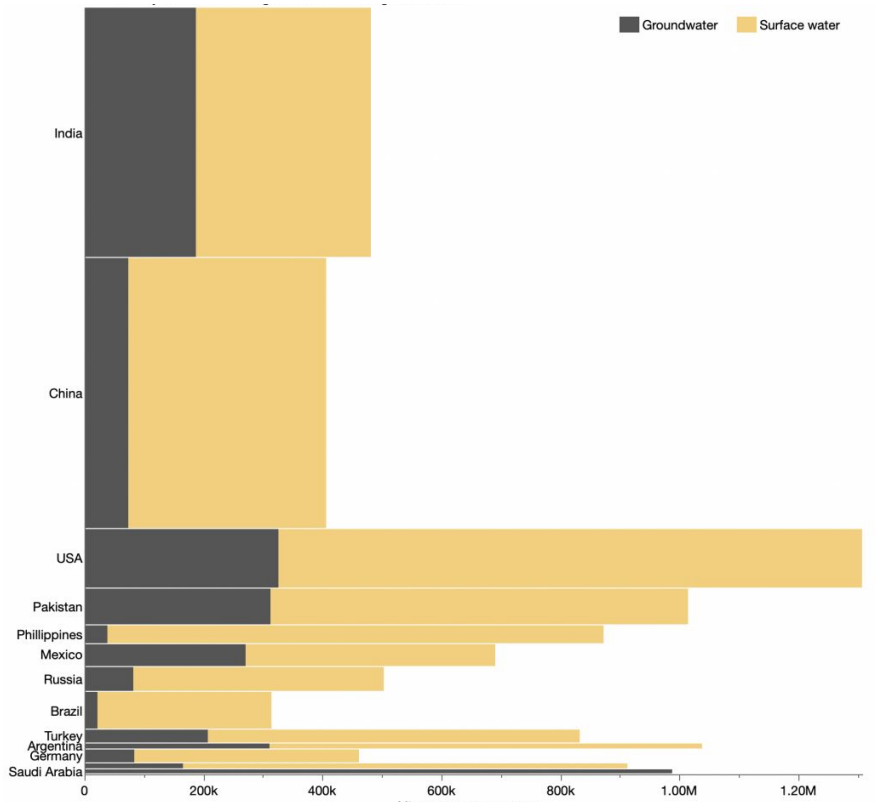
10. *Ambiguous so full credit given for either true or false. I intended false since 8-12 bins for categorical color are possible, but since I did not say "only 3-4 bins" it could be interpreted as true since it's indeed possible to show 3-4 bins of categorical color.*
11. *True: size and position are separable, but size affects color.*
12. *False because perceptual system does everything through relative not absolute judgements.*
13. *True: line charts strongly imply trends because of the connecting lines.*
14. *True: heatmaps are compact, and row/column reordering is crucial for showing structure.*
15. *True: cyclic attributes are a good fit for radial layouts.*
16. *True: region size variability is the major problem with choropleths.*
17. *True; categorical data shown in small noncontiguous regions should be highly saturated to be discriminable.*
18. *False because rainbows are unsuitable for quantitative attributes because of the lack of intrinsic ordering across hue changes.*
19. *False because although the first part is true (bivariate colormaps are often misinterpreted) and the second part is true (8% of men are red-green colorblind), there is no causality/connection between these: the second doesn't explain the first.*
20. *True: animated transitions are well suited for constrained navigation.*
21. *False because small multiples are entirely appropriate for spatial data as well as tabular data.*
22. *True: linked highlighted is well suited for showing connections across multiple views, including the multiform case.*
23. *True: both blue-orange and redundant coding are good approaches to colorbind-safe design*
24. *True: navigation is exactly changing the camera viewpoint.*

Q3: Water usage three ways *[15 pts]*

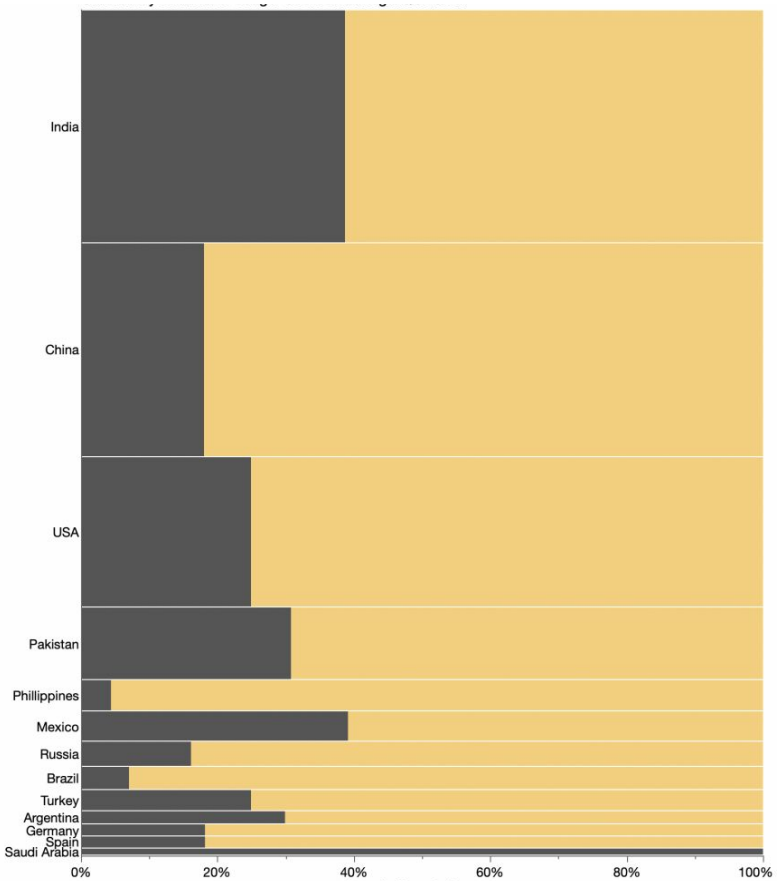
Credit: https://www.forthgo.com/blog/2019/12/18/bar-mekko-chart-study/

A. Water usage by source. water use (area) = population (height) x per capita use (width)

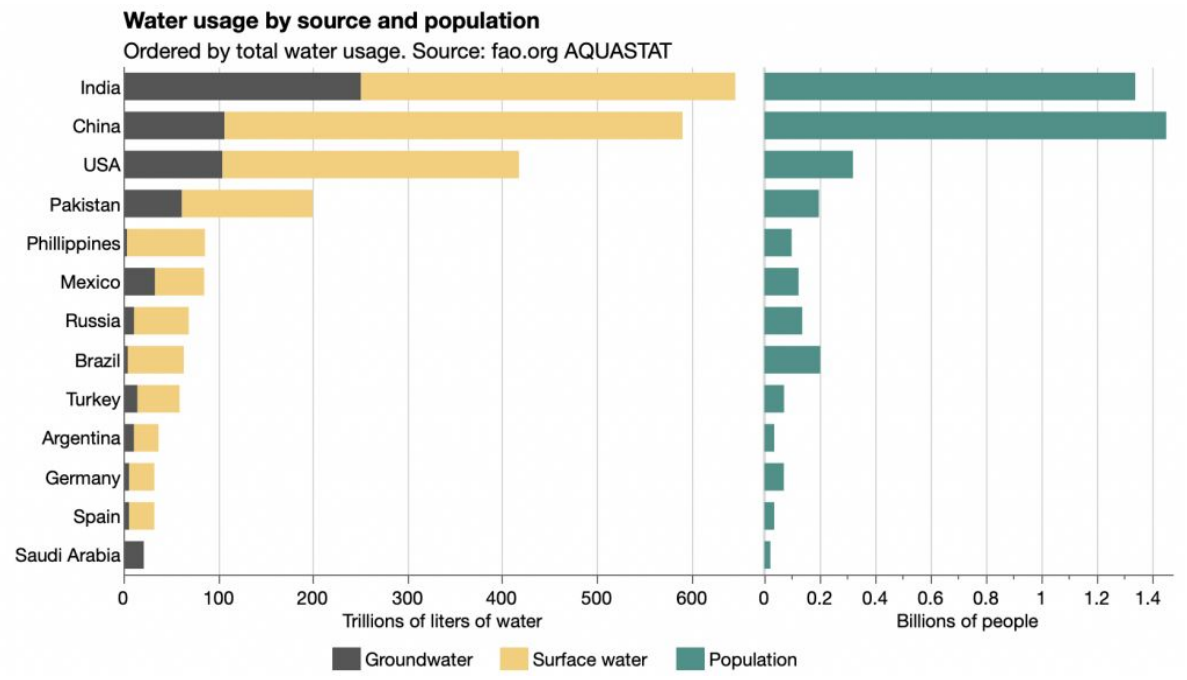
Ordered by total water usage. Source: fao.org AQUASTAT. Dark grey - groundwater. Light grey - surface water. X-axis is liters per person per year.



B. x-axis is percent of country's usage



C.



Above are three different visual encodings of the same dataset (A, B, and C). State one disadvantage for each approach, analyzing in terms of how the channels of length (horizontal), height (vertical), area (length x height of segments/bars), and color are used to encode the attributes of water source (ground vs surface water), water usage, and population.

Breakdown for each (5 pts) is: 2 pts for basic idea of disadvantage (-1 for each incorrect disadvantage), 3 pts for clear & specific analysis of why (-3 for no attempt to explain or incorrect statements, -2 for vague explanation, -1 for minor problems). Full credit for any of the listed answers. No penalty for incorrect statements about use of color luminance/saturation/hue since can't tell true encoding from black and white photocopy.

A disadvantage: *[5 pts]*

- *Answer 1: The important information of total/absolute water usage is conveyed by the area channel, but area judgements are less accurate than length judgements. It's difficult to compare the tall narrow areas (high population, lower per-capita use) of India and China to the short fat area (lower population, higher per-capita use) of the US.*
- *Answer 2: Nonuniform heights for the rows/bars/areas lead to difficulty in reading the labels near the bottom where row height is very small.*
- *Answer 3: It is difficult to compare surface water usage segment lengths since they are not aligned (although ground water segments are, as are total water usage bars).*
- *Answer 4: Height is encoded on axis labelled only by country name without quantitative attributes, so cannot determine exact population (height) or water usage (area).*
- *Incorrect answer 1: It is hard to compare areas because the countries are not ordered by height. (The countries are exactly ordered by area, so this argument doesn't make sense. It's hard to compare areas because we're not good at comparing that channel.)*
- *Incorrect answer 2: bar length and height do not encode the same attribute so people could be misled. (No, cannot argue that any non-standard encoding is automatically misleading or all we could ever use is bar charts and scatterplots and line charts! Same logic as incorrect answer B3 below.)*
- *Incorrect analysis 3: Area differences are hard to read because the length and height are different. (That's just a geometric fact, not an analysis connected to human perception.)*

B disadvantage: *[5 pts]*

[Information that was in the original article but not explicitly provided on the exam: the row heights correspond to water usage for the entire country, not the population as in chart A. Full credit for answers assuming that height=population]

- *Answer 1: [If assume height=population] The total/absolute water usage for the country as a whole, and the total/absolute water usage for each type of water, is much more difficult to see than with A or C, since we cannot read it directly from any of the visual channels: row height (population) or row length (relative contribution of each source) or color (water source type). Area has no direct meaning, which is a serious disadvantage.*
- *Answer 2 [If assume height=water usage for country] The population and per capita usage cannot be seen, in contrast to A or C, since we cannot read it directly from any of the visual channels: row height (absolute country's water usage) or row length (relative contribution of each source) or color (water*

source type). (Area has a linear relationship with the water usage for each source type, and thus for the total water usage for the country.)

- *Answer 3: Height on Y axis unspecified.*
- *Answer 4: [If assume height=population] Height is encoded on axis labelled only by country name without quantitative attributes, so cannot determine exact population.*
- *Answer 5: The important information encoded by the area channel is difficult to understand because area judgements with different height/width combinations are hard to make.*
- *Incorrect answer 1: it is difficult to compare heights across rows because they are not aligned to a common scale. Although the very highest precision comparison is aligned length/height, the next best is unaligned length/height, so calling it "difficult" is overstating the case, especially since they're ordered by the row height in this diagram.*
- *Incorrect answer 2: Horizontal position is useless because all items have 100% usage by definition. No, misunderstanding of normalized stacked bar chart that is showing part to whole relationship.*
- *Incorrect answer 3: Argument that encoding percentage as length could be misinterpreted. No, that argument implies that normalized stacked bar charts are never legitimate, which is not true - this argument that if you misinterpret the meaning you'd get confused would apply to a huge number of visual encodings!*

C disadvantage: *[5 pts]*

- *Answer 1: It is more difficult to understand per-capita water usage (correlation between population and water usage), which would require mental division of the left bars compared to the right bars, than with A. These attributes are shown in different views that do not directly support comparison.*
- *Answer 2: It is more difficult to compare the relative value/percentage of each source type across countries. For example, it is harder to see that the USA percentage is a lot more than China, but it's easy to see that in B.*
- *Answer 3: No disadvantage, it's much better than A & B because it is easy to understand population, total water usage, relative proportion of water usage by source, and total usage by source.*
- *Answer 4: Misleading horizontal axis on the bottom since there is a continuous line between the two charats without a visual break between left side with water tickmarks and right side with population tickmarks.*
- *Incorrect analysis 1: It's hard to compare water usage with population because the views have different (or unaligned) x axes. (No. It's intrinsic to the data that the attributes of water usage and population have different axes, it could be dangerous to use a dual-axis chart but it's totally legitimate to have side by side charts with an aligned vertical axis (country) and different horizontal axes - lengths can be compared within each view even though the two views have different axes.) XXX DISCUSS XXX*
- *Answer 5: It is difficult to compare surface water usage segment lengths since they are not aligned.*
- *Answer 6: Color luminance for population and groundwater is hard to tell apart so there may be some confusion when looking at black and white version as we are doing here. (Not true in full-color graph but counted as correct since you saw black and white photocopy; although it's a minor not a major*

problem since this use of color is redundant with partition into multiple views, it's true that color consistency is important.)

- *Incorrect analysis 2: Color luminance for population and groundwater are hard to tell apart so colorblind viewers will have difficulty. (No, colorblind users are not limited to luminance only, they do see some hues. These are in fact distinguishable to color deficient users, as you could check with a simulator. Although you do not have access to correct colors or a colorblindness simulator during a midterm, this answer indicates a lack of understanding about color blindness/deficiency.)*
- *Answer 7: It is hard to see/compare values of water usage (and/or population) between countries with small values because there is such a large range of values.*

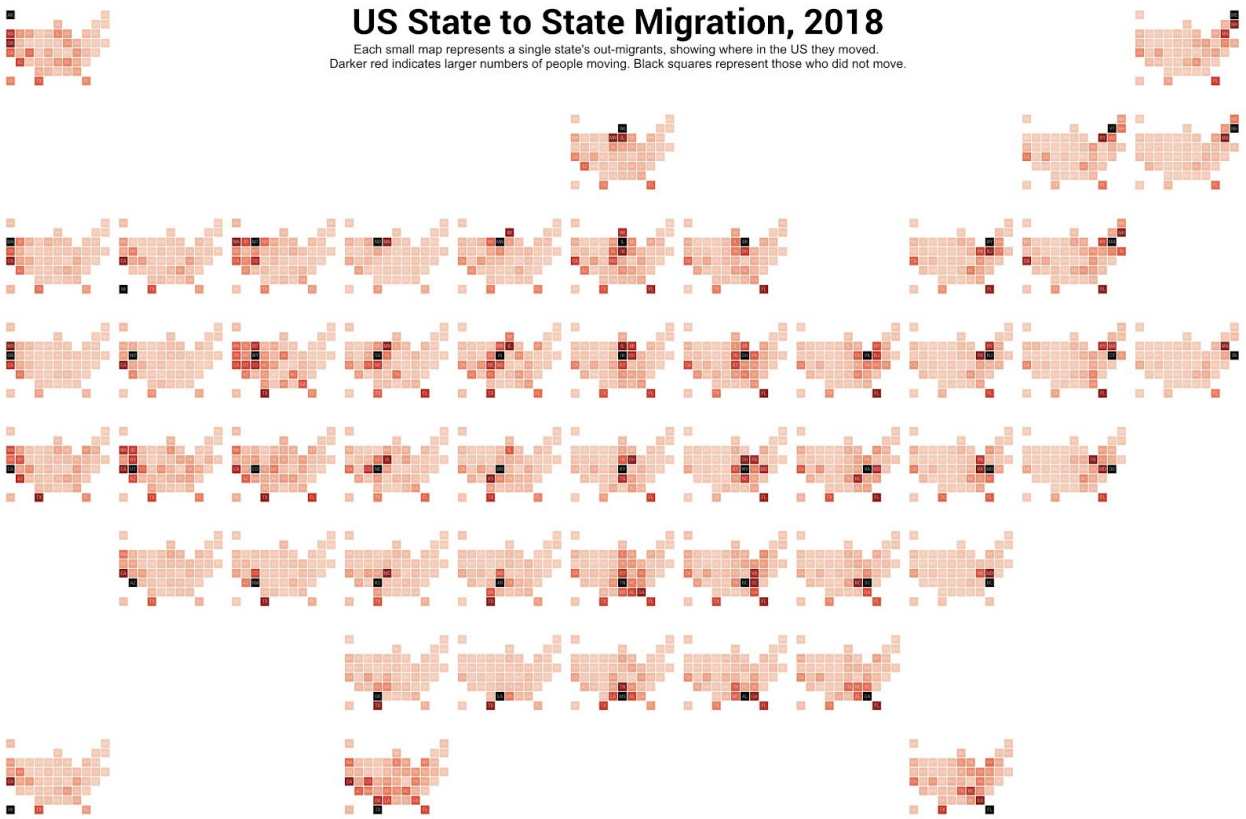
Q4 Maps and Multiples [14 pts]

Credit: https://github.com/jwoLondon/30dayMapChallenge

For each of the two visualizations below (A, B), answer the following questions:

- What type of view coordination is used (single view, multiform, small multiples, overview/detail multiform, overview/detail same form)? Justify your choice very briefly in terms of whether the visual encoding is same or different, and whether data is shared between views or subsetted.
- For each kind of mark used, state the mark type and which channels are used to encode what attribute on that mark.

A. US State to State Migration, 2018. Each small map represents a single state's out-migrants, showing where in the US they moved. Darker grey indicates larger numbers of people moving. Black squares represent those who did not move.



[6 pts: 1 pt views + 5 pts marks/channels]

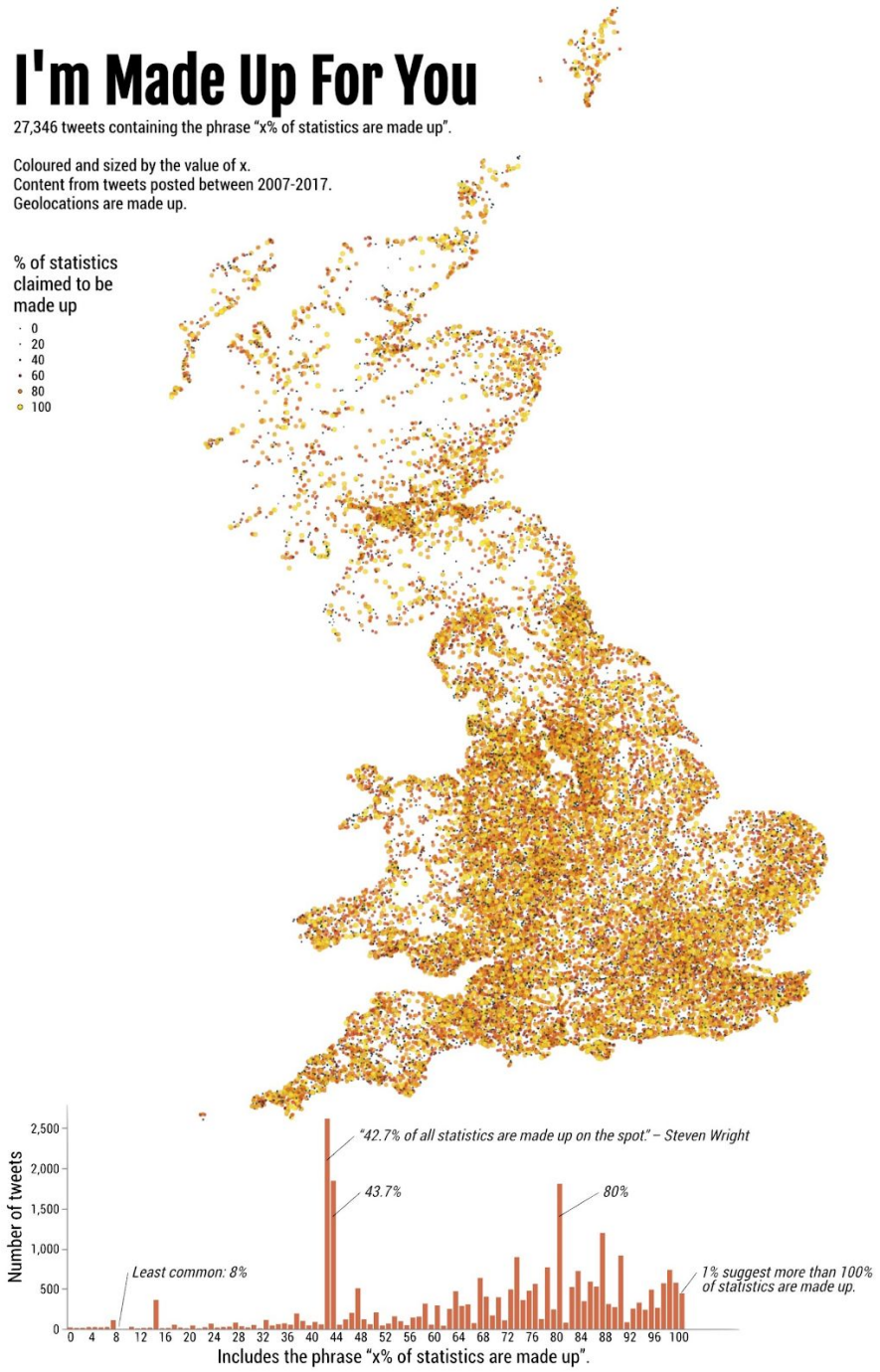
View coordination: small multiples. The visual encoding is the same, and different data is shown in each (base state vs all the others). [1 pts]

- *The position of each small multiple view is determined using given spatial data (full credit even if not stated)*
- *Full credit if small multiple is called a glyph instead of a view. -0.5 if small multiple is called a mark, since marks don't have substructure.*

Marks (within each small multiple):

- *Point mark for each state. (Full credit given for area mark) [1 pts]*
    - *Spatial position channel uses given spatial data for location of state with respect to other states/marks (with boundary geometry simplified to a square). (Full credit given for anything about spatial position using given spatial data) [2 pts]*
    - *Color channel indicates attribute out-migrants (number of people who moved from that state to each of the other 50 states). (It's actually luminance in the full-color version, but cannot tell from black and white version so full credit for saturation+hue, luminance+hue. But no credit for just hue.) [2 pts]*
    - *Hue indicates whether there are migrants to a specific state. (Too subtle to see from the black-and-white prints, so this is optional)*
    - *-1 for any wrong channels*

B. I'm Made Up For You. 27,346 tweets containing the phrase "x% of statistics are made up". Sized by the value of x. Content from tweets posted between 2007-2017. Geolocations are made up.



[8 pts: 1 pts views + 7 pts marks/channels]

View coordination: multiform. Different visual encoding with each view, same data is shown in each. [1 pts]

- *Top view (spatial):*
  - *Point mark for each tweet [1 pts]*
    - *Size channel encodes statistics claim (quantitative attribute). [2 pts]*
    - *(Color channel (luminance) redundantly encodes stats claim, but hard to see in black and white - not required for full credit)*
    - *2D spatial position channel using (fake) spatial data - (fake spatial data not required for full credit). [1 pt]*

- *No penalty for proposing area mark as outline of country. (Technically there's no outline, but it's implied because of Gestalt properties - we did not cover these in class!)*

- *Bottom view (histogram):*
    - *Line mark for each histogram bin (counts discretized across 100 bins). [1 pt]*
        - *Y position/height/length (1D spatial position): number of claims in bin (discussion of bins not required for full credit) [2 pt]*
            - *Length encodes number of tweets*
        - *X position: value of statistical claim (not required for full credit)*

Q5: Sketching Electronic Health Record Interface *[26 pts]*

You are given an electronic health record dataset consisting of events that may have a single timestamp, or they may have both start and end times. Events also have geographic locations. Each patient may have up to 100 events associated with them. There are up to 100 patients in the dataset. The event type attribute has 5 levels: 4 main types of events (emergency, clinic, home, medication), and one catch-all 'other' category for the rest. The task is to assess whether patients who use clinics for routine health care are less likely to use emergency services, and whether this tendency is different for patients with complex medication needs vs. simple or no medication usage. Sketch two views: an overview showing all patients, and a detail view showing the data for only a single selected patient. Explain interactive functionality with words and arrows. Briefly justify your design choices with respect to alternatives.

[Rubric: 26 pts
Things to notice: geographic location is irrelevant to either task (so views involving maps are inappropriate). For the event types, only emergency, clinic, and medication are relevant, the other two are not. Complex vs simple medication usage could be a) a derived attribute (you don't have to specify how to compute it) or b) the number of medication events (complex = many, simple = few) or c) the fine-grained temporal pattern of medication events vs emergency & clinic events.

- *Overview [14 pts]*
    - *Effectiveness of visual design [10 pts]*
        - *Scalability: Appropriate rollup to level of detail so that all 100 patients can be seen simultaneously?*
        - *Directly showing correlation between emergency and clinic events with chart such as scatterplot is acceptable.*
        - *Common problem: geographic view that doesn't address task. Deriving overall patient geographic location from individual events doesn't make any sense for this task.*
        - *Common problem: a stacked / grouped bar chart by event types does not address the task to compare simple/few vs. complex/many medication events/usage.*
    - *Justification of design choices [3 pts]*
        - *Common problem: Generic statements that are true in general for the chart type that do not mention alternative choices or connect to the specific setting/task.*

    - *Sketch clarity [1 pt]*
- *Detail view [11 pts]*
    - *Effectiveness of visual design. [8 pts]*
        - *Scalability: Appropriate drilldown to see enough relevant detail for single patient?*
        - *Common problem: Inadequate level of detail. A bar chart / pie chart showing counts for the five event types for one person is only minimally useful for either of the two tasks. It's a missed opportunity: cannot see temporal patterns of events over time, specifically medication events (where complex=many medication events and simple = few medication events) wrt emergency vs clinic usage*
        - *Common problem: misleading use of line chart. Connecting events of the same type with connecting line marks doesn't make much sense, compared to connecting each event in temporal order regardless of type.*
        - *Good to provide temporal information, great to consider events with long vs short duration*
    - *Justification of design choices [2 pts]*
    - *Sketch clarity [1 pt]*
- *Explanation of interactive functionality [1 pt]*