Aaron Schwartz-Messing

Data Visualization Final Exam

5/24/17 Professor Van Kelly

QUESTION 1

For each of the following visualization,

1. describe the
    a. data items
    b. and their attributes,
2. the category of each attribute (categorical, ordered, etc.),
3. and the marks and channels chosen to represent each item/attribute.
4. Describe any aggregation taking place in the display (there may not be any in some cases) and whether this aggregation succeeds in reducing visual clutter or confusion.
5. Also, suggest an alternative visual design that you think might also be an effective presentation
    a. and tell what its strong
    b. and weak points would be relative to the published visualization.

MAPPING SEGREGATION
Data Items - Each data item is a census tract.
Attributes - The attributes are

1. the total population of the tract (quantitative),
2. as well as percentages (quantitative) of the population that are
    a. black,
    b. hispanic,
    c. asian,
    d. white,
    e. and other.
3. The population of a given race (quantitative) can be figured out based on the data.
4. I think that the borders and position of each census tract are considered attributes. That may be ordinal or even quantitative if you assign a coordinate system to the position on the map or if you compare them in terms of how far to the north, south, etc. any given tract is compared to another tract.
5. Tract Number (could be ordinal or categorical depending on how they are assigned)

Marks and Channels -

1. Total Population -

Marks - dots
Channels - total number of dots.

2. Percentage of a given race
   Marks - none (see next entry, *population* of a given race has marks)
   Channels - concentration of that race’s color in the total scheme of colors in the tract. Number of dots of that color compared to the number of dots in the tract.
3. Population of a given race
   Marks - dots
   Channels - number of dots for that race’s color
4. Position
   Marks - black-bordered closed polygon or set of polygons
   Channels - spatial region
5. Tract Number
   Marks - none
   Channels - none

Aggregation -
As you zoom out the number of dots decreases by about 33%-50% each click, the remaining dots each representing 33%-50% more than the previous dots did.

Does it help? I think that it does help in reducing visual clutter and confusion. I think that if it did not aggregate the dots then as you zoomed out it would have been harder and harder to distinguish the presence of the smaller-sized races amongst the larger ones. By aggregating, one is able to distinguish outlying dots on every level. Had the aggregation not been done, these would have been an indiscernible faint hue in a sea of other colors.

Alternate Design - A possible alternate design would be to color in portion of the each tract proportionally to each race instead of using dots. Meaning, one part would be red, one part green, one part orange, etc. with the size of each colored portion corresponding to the percentage of the population.

Advantages:

1. Eliminates white space
2. Less shapes to render. Only five, one per race per tract
3. Less clutter. It’s more simple and straightforward

Disadvantages:

1. Since each race will only have one piece, rescaling will be more difficult. Using a dot model it is easier to restrict the amount of dots in an area, but using my model where all of the dots for a given race in a tract are in one blob, when you zoom out it will either be more inaccurate, or it will just have the same effect as the model that they implemented.
2. Since I would be eliminating all of the white space one would not be able to tell which areas are more densely populated and which ones are scarcely populated. One also would not be able to count the dots to find out how many people there are.

MONEY, RACE, AND SUCCESS: HOW YOUR SCHOOL DISTRICT COMPARES

FIRST GRAPH

Data Items - Each data item is a school district.

Attributes - The attributes are

1. Name (Categorical)
2. State (Categorical)
3. Number of grade levels above/below average (quantitative)
4. Median family income (quantitative)
5. Percentage of each race in that school district (unclear whether referring to all of the people in the district or the amount of students. Probably students. Is this only sixth grade students? Says so in the title of the article, but nowhere on the graph) (quantitative)
6. Number of students in each district (quantitative, although our only indication of what this number is is based on the size of the bubbles)

Marks and Channels -

1. Name
   Marks - none
   Channels - none
2. State
   Marks - none
   Channels - none
3. Number of grade-levels
   Marks - circle
   Channels - vertical position
4. Median family income
   Marks - circles
   Channels - horizontal position
5. Percentage of each race
   Marks - none
   Channels - none
6. Number of students in each district
   Marks - circles
   Channels - size

Aggregation - Does not distinguish between the races when it comes to median family income

Does it help? Yes, because the graph isn't trying to tell us anything about the races in this graph. Were it to take races into account that would make this graph harder to read.

Alternate Design - The overlap here is very great, so it is difficult to really know how many districts and how many people are present at any given segment of the graph.

I would split up the graph into rectangular segments (about the size of the ones present on the graph currently, or maybe a little smaller) and put one bubble per square. The size of the bubble

would be proportional to the population size. Districts are less important for this graph because the main point is to show the trend in America as a whole, and the districts were just there to provide distinct points on the scatter plot. If we did it my way you would get a better sense of the spread of the population because there wouldn't be any overlap.
Problems with my idea:

1. The size of a bubble in a particular segment may extend beyond the boundaries of the segment and actually overlap. We could solve this by making it logarithmic instead of proportional, but that would turn size into a lie factor because the size will no longer be proportional to the population, therefore even though you would be able to tell which segments have more people than other segments, it would seem like a smaller discrepancy than it really is.
2. You would also lose accuracy by aggregating all of the bubbles in one segment into one big bubble centered around the average for that segment.

SECOND GRAPH
Data Items - Each data item is a school district.
Attributes - The attributes are

1. Name (Categorical)
2. State (Categorical)
3. Number of grade levels above/below average per race (excluding asian/other) (quantitative)
4. Median family income (quantitative, only shows through vertical position channel)
5. Number of students for each race (excluding asian) (quantitative, only shown through size channel)
6. Difference between grade levels between blacks and whites and hispanics and whites (quantitative)
7. Difference between median family income between blacks and whites and hispanics and whites (quantitative)

Marks and Channels -

1. Name - none
2. State - none
3. Grade levels -
   Marks - bubbles colored by race, connected by lines
   Channels - vertical position
4. Median family income -
   Marks - bubbles colored by race, connected by lines
   Channels - horizontal position
5. Number of students -
   Marks - bubbles colored by race, connected by lines
   Channels - size
6. Difference in grade levels between the races
   Marks - bubbles colored by race, connected by lines
   Channels - vertical distance, vertical length of lines connecting bubbles

7. Difference in median family income between the races
   Marks - bubbles colored by race, connected by lines
   Channels - horizontal distance, horizontal length of lines connecting bubbles

Aggregation - None

Alternate Design - I think that using little multiples wouldn't be a bad idea here.

The goal is to show that there is a discrepancy between the white population and the black and hispanic population within the school district in terms of average grade level, especially in places where there is a big socioeconomic difference between the white families and the minority families.

As is, one of the problems of this graph is that there is overlap between the pink circles and other circles. However, if you mouse over a circle in the mixed region you will see that it is following the general pattern in terms of the discrepancy between the white students and other students, and the only reason that there is overlap is because the best students in one county are on the same level as the worst students in another county.

If we separated each county into its own little multiple, that will show that they all really follow the trend, and we won't be distracted by the overlap. Since there are way too many counties to put each one into its own little multiple, it would probably be best to split the graph into subsets of similar counties where the white population of one does not overlap with the minority population of another.

One of the cons of the little multiple graph would be that there would be a lot more graphs to look at, and it would take up a lot more space. The way that it is now, especially with the interactivity, does a very good job at getting the point across.

THIRD GRAPH

Data Items - Each data item is a school district.

Attributes - The attributes are

1. Name (categorical)
2. State (categorical)
3. Number of grades levels ahead of national average for whites and hispanics or blacks (quantitative)
4. Number of students that are white and number that are hispanic or black (quantitative, only shows by the size channel of the bubbles)
5. Difference between the average grade level between the races (quantitative)
6. Difference between the median family income between the races (quantitative, only shown by the horizontal distance channel)

Marks and Channels -

1. Name - none
2. State - none
3. Number of grades levels ahead of national average for whites and hispanics or blacks
   Marks - bubbles colored by race, connected by lines
   Channels - vertical position
4. Number of students that are white and number that are hispanic or black
   Marks - bubbles colored by race, connected by lines

Channels - size

5. Difference between the average grade level between the races
   Marks - bubbles colored by race, connected by lines
   Channels - vertical distance, vertical length of connecting lines
6. Difference between the median family income between the races
   Marks - bubbles colored by race, connected by lines
   Channels - horizontal distance, horizontal length of lines

Aggregation - None

Alternate Design - Since for each data set the socioeconomic status is basically the same across all races, I would get rid of the x-axis altogether and spread out the counties side by side to avoid the same overlap issue that we had on the last graph. My graph would show that every data point follows the trend, while the graph as it is now shows that most follow the trend, but if you wanted to be sure that every single one followed the trend you would have to mouse over them to see which bubbles corresponded to which other bubbles.

QUESTION 2

Critique the following dubious graphics, including lie factors, if any.

Proportion of the Total Number of Published Papers By Region in All Fields, 1981 - 1996

1. Since this graph is being looked at from an angle (which is the case with all 3d graphs) it is very difficult to get a precise reading of any of the axes.
2. For that same reason it is difficult to compare the readings at different points. Differences must be very exaggerated to be noticed.
3. Things closer to the back corner automatically look more elevated than they really are because of the skewed perspective from which the graph is being viewed. That throws everything off, making the increases in percentage of citations look a lot smaller than it is and making the decrease in percentage of citations look a lot bigger than it is.

Why does College have to cost so much?

1. The magazine put two line graphs, the y-axis of each measuring a different variable in different increments and different units, on top of each other. It is therefore meaningless to compare the y values of the graphs to each other.
2. As a continuation of the previous point, from the fact that they put the ranking graph below the tuition graph, that implies that you are getting less for your money, because the y value of the price is greater than the y value of the rank.
3. The horizontal axis, though both measuring time in years, have different increments. So the years don’t line up when you superimpose them. The tuition graph measures in increments of five years from 1965 to 2000, while the ranking graph measures from 1989 to 1999 in increments of one year.
4. The fact that the magazine left out the increments isn’t a problem because you can count the amount of vertices and deduce from there which year each vertex occurred.

5. The ranking graph (I assume) measures the ranking in terms of first place, second place, third place, etc. where first place is the highest, most prestigious ranking, and the higher numbers are the worse, least prestigious rankings. If this is indeed the case, then it turns out that years that have a lower value on the Y axis are the better ranked years, and vice versa. But this is not the way that people are used to reading the Y axis of a graph. Usually the higher Y-values are better, and the lower Y-values are worse.
6. The magazine labels the top graph as “Cornell University tuition”. That implies that the increases in the graph are increases in tuition. But really the graph is measuring the percentage of the income of the students that go to paying the tuition. So these rises can indicate that people are getting lower paying jobs, or that Cornell over the years has begun to accept poorer students.
7. The Magazine doesn't tell you that the ranking is based on the U.S. News and World Report, and presents it as an objective thing.
8. The Magazine implies that the ranking is getting worse (see number 5 where I argue that this is the opposite of the reality) because Cornell is getting worse as a school, but it could be that the other schools are simply getting better. And who knows what the prices those schools charge?
9. The ranking of Cornell compared to other schools and the percentage of the student's income that goes to paying tuition are not necessarily linked at all.

QUESTION 3

Of the various techniques for handling complexity, describe and critique the technique or techniques used in each of the following:

Harvest - Small Multiples

The goal here is to show how more and more of the forest became cleared over time. In order to accomplish that they made “smalls multiples” maps of the aggregate clearance since 1995.

cons:

1. The issue with this is that it is difficult to tell exactly what areas were cut from year to year. It can also be difficult to tell exactly how much new clearing was done every year. That could have been solved had they highlighted the newly cleared areas with color (red would have worked well). This information is provided by the chart, but it is not very discernible on the maps.
2. Another problem is that it is difficult to compare the map on the end of one row to the map of the next year which is on the beginning of the next row.
3. It also uses a lot of space.

Pros:

1. If this is the information that they are trying to get across, namely the amount of aggregate cleared forest per year, this gives the most information since it uses the map

itself. By making many small graphs one next to the other one can easier start from the first year and move on to the next and the next, seeing the aggregation grow over time.

Bubble Sets

This is a way to link members of the same set when those members are spread out over a large area and mixed with members of other sets. By linking the entire set into one connected component it is very difficult to overlook a member of the set.

Cons:

1. Since the bubble set connects all of the members, there are long strings of bubble that flows from one data point to another. This is misleading because the bubble part doesn’t actually represent anything, and is just a way of keeping all of the data points together.
2. There can be two data points that are close together but since they are in two different branches they seem to be farther away from each other.
3. The structure of the set doesn’t represent anything. Whether to connect two data points together or not is arbitrary.
4. By splitting up all of the data points into subsets, it becomes more difficult to look at concentrations of data points in regions of the chart because all of the “chart junk” is getting in the way. The data points are a lot harder to see because of the bubbles, especially when you're trying to look at data points from multiple sets in one region of the graph.
5. It gives the illusion of a connectivity that is not really present in the data, but rather is just a visual tool.
6. Also, this is very dependent on interactivity. This could not be printed, especially not in black and white.