Antonia Deliyianni and Mohammed Saqib
12/01/2017
CS 4460
P5

# Candy Survey Analysis

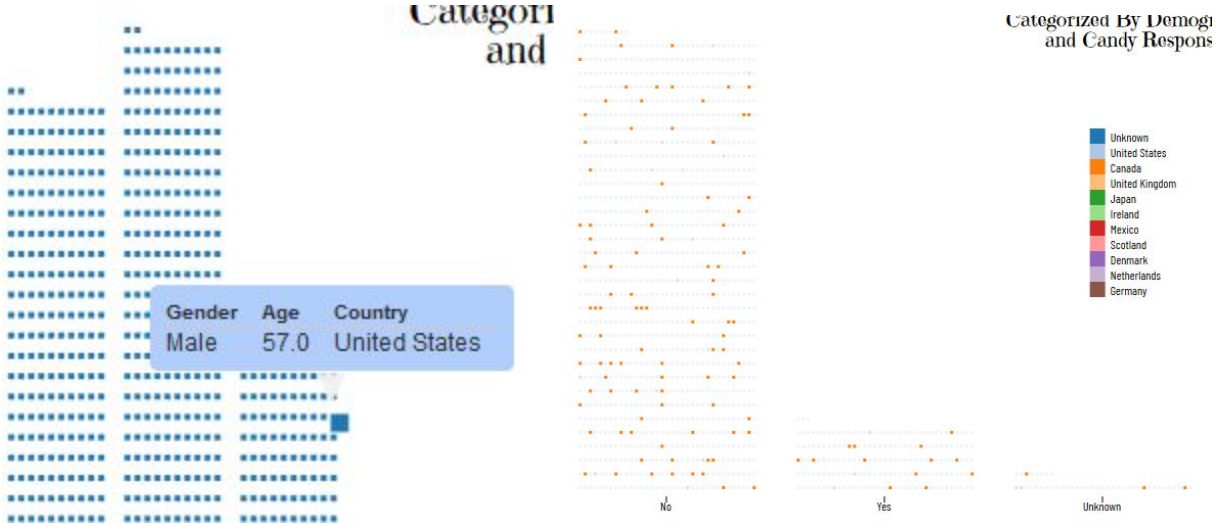
For Project 5, we chose to visualize the candy dataset, creating multiple charts, which are linked to each other. The chart views included a dense pixel histogram, which can be customized through the usage of dropdown choices, a stacked bar graph view, and a Geomap focusing primarily on the US.

When we first began our design process, we explored the dataset in order to garner a general understanding of the nature of the data. We found that the vast majority of attributes were ordinal variables about an individual's response to candies, except for the first few columns, which went over key demographic categories, and the last two columns, which provided a free form text response of any length. With this data available, we determined a variety of analytical tasks for our visualization to support; such as filtering and sorting the data, computing new information, finding extremums, characterizing distributions, and clustering by attributes. With these tasks in mind, we brainstormed several views, implementing a few of them as preliminary tests and discovering missing and unstructured data points in the dataset. This discovery lead to a key decision during the design process on how to deal with the holes throughout the data. Specifically, we had to decide whether it would be easier to simply drop the data entirely or include the data and create a new categorization for no response. In the end, we chose the latter, keeping with the spirit of Tufte's principle of statistical honesty and maintaining what we deemed to be the truth of the dataset.

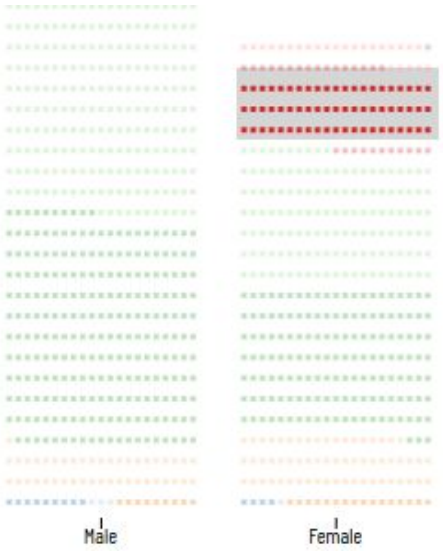
Once we had made this decision, we tackled the unstructured nature of some of the responses, specifically concerning the data surrounding the participant's location. This data was difficult to manage, and we eventually utilized Google's geolocation API to translate each of the locations into a usable coordinate structure, as the API provides the ability to translate vague and colloquial wording into exact latitude and longitude coordinates and map entities. We decided primarily on focusing on the US for our map as it held the majority of participants and because other countries would likely be hard to arrange without leaving the vast majority of the world blank (i.e. to show Japan, we would likely have to show China, which had no participants. This would be similar for most countries in the world).

As part of our design process, we brainstormed multiple possibilities to present the dataset, recognizing that key tasks relied on capabilities in presenting the information as both a person-centric view based on demographics and a candy-centric view based on responses. With these representations in mind, we discovered that a key challenge was ensuring that the views did not diverge, as the most meaningful tasks would likely involve both parts of the data. In order to ensure that both aspects of the data are playing equal roles in the user's exploration, we relied on interactions.

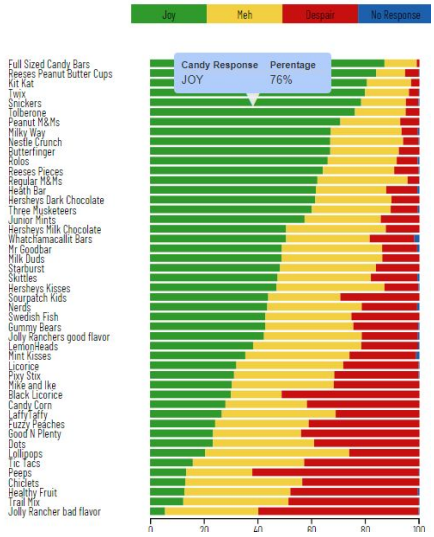
Our design allows the user to explore the data, through several methods of interaction that links the views and represents attributes accordingly. For example, the histogram view, which represents each participant as a small glyph, can be restructured by drop down menus that alter the x-axis position, coloring, and sorting of the glyphs. Although the chart changes through the use of these menus, it maintains an overview of the participants; however, details of an individual can be discovered by hovering over a glyph, revealing a tooltip with the participants demographic information. Additionally, if a user wishes to highlight participants by color, they can hover over the legend and hide those outside of the

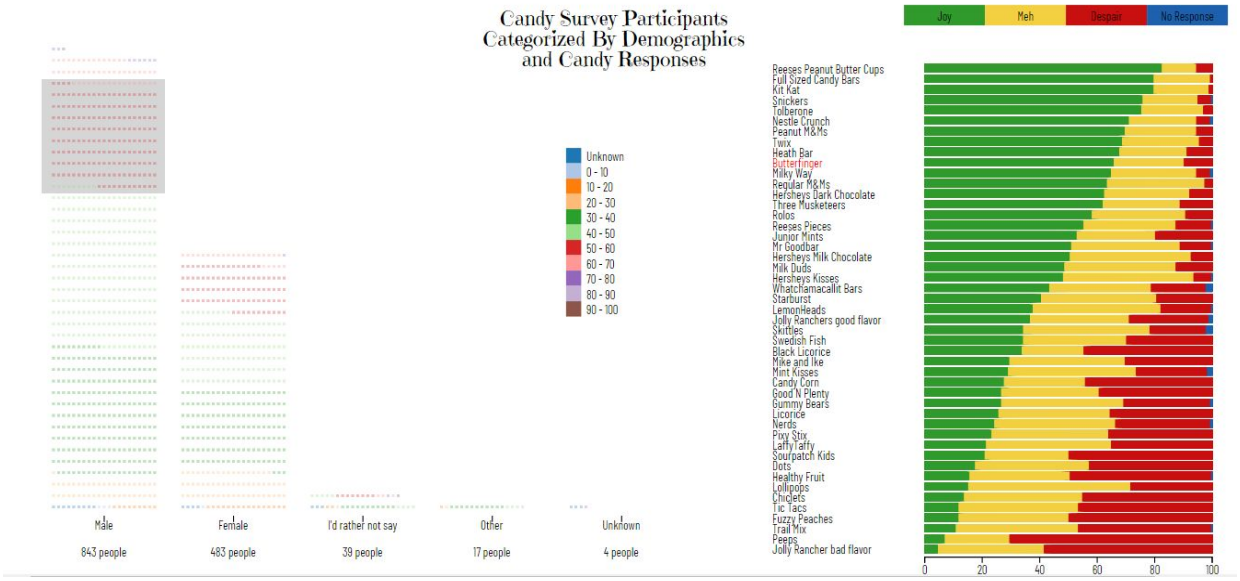




demographic. Lastly, the user can select which participants they would like to focus on for the second chart through brushing.

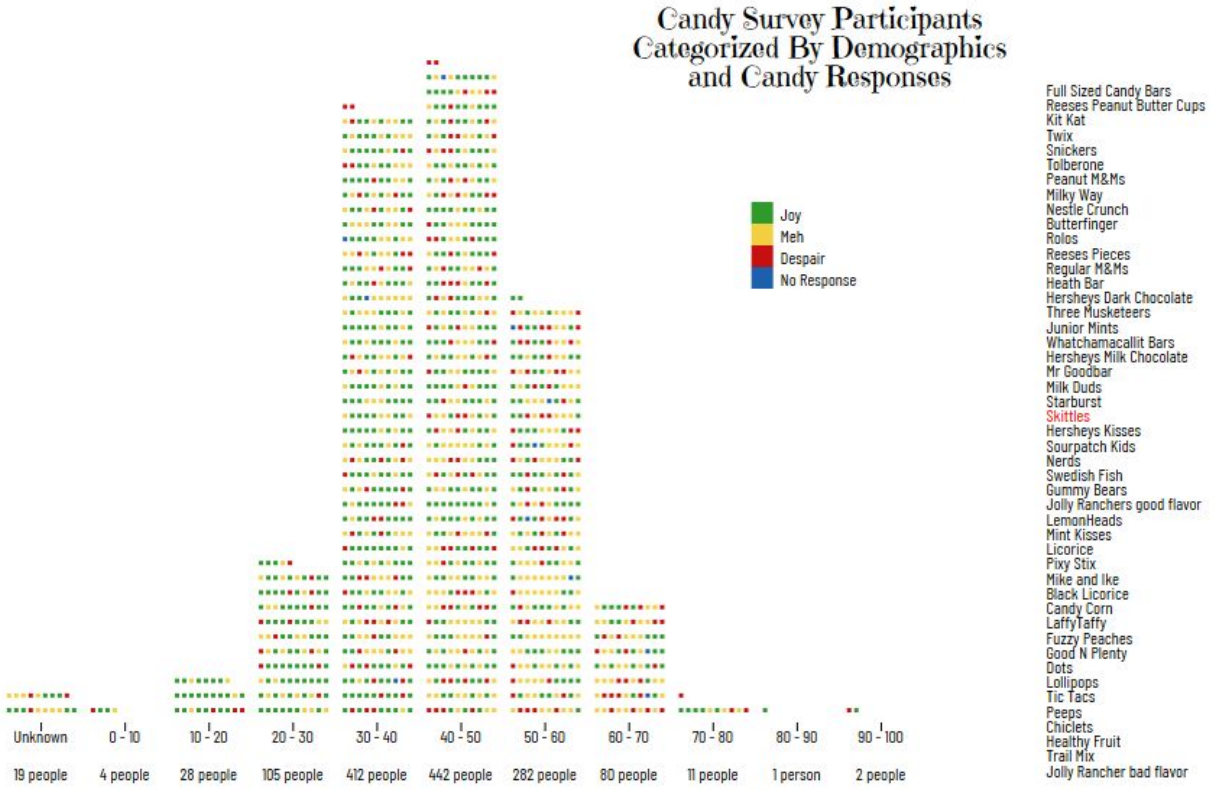


The second chart computes the percentage of participants that gave a specific response for each candy, and the candies are sorted from top to bottom by the percentage of positive responses. Tooltips appear based on mouse hovering over the bar. At first the chart reveals the distribution of preference for each candy based on all of the participants; however, if a selection is made on chart one, through the brush functionality, the values are recalculated for that selection alone. This can allow for sophisticated querying of the candy responses, especially when combined with the stack property of the dense pixel representation; for example, one could chose male participants from age 50-60 through this method in order to visualize the candy responses.

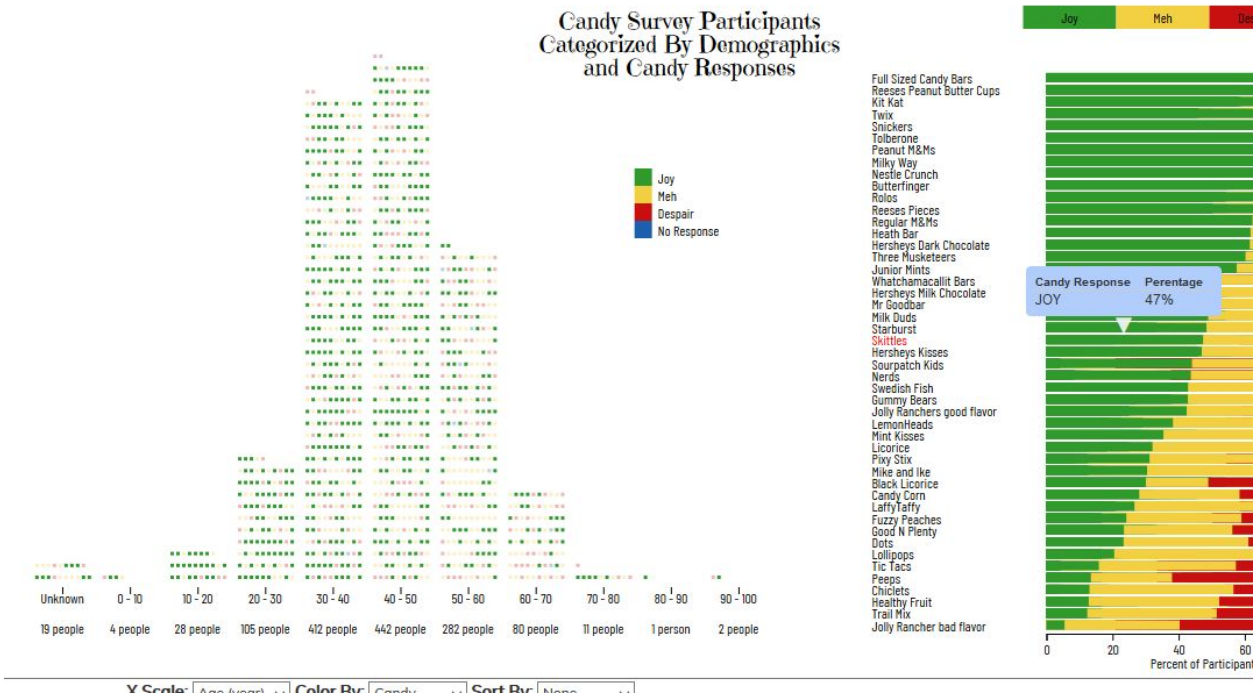



The chart also allows the user to color the participants by their preferences for a particular candy. If the user hovers over the candies by name, the candy is highlighted, turning from black to red; however, when a user clicks on a candy, the glyphs in chart one are colored accordingly. In addition, one can hover over the legend for the dense pixel representation as well to show and highlight the appropriate subsection of the population, as well as update the stacked bars and map view to represent the correct subsection of the population.

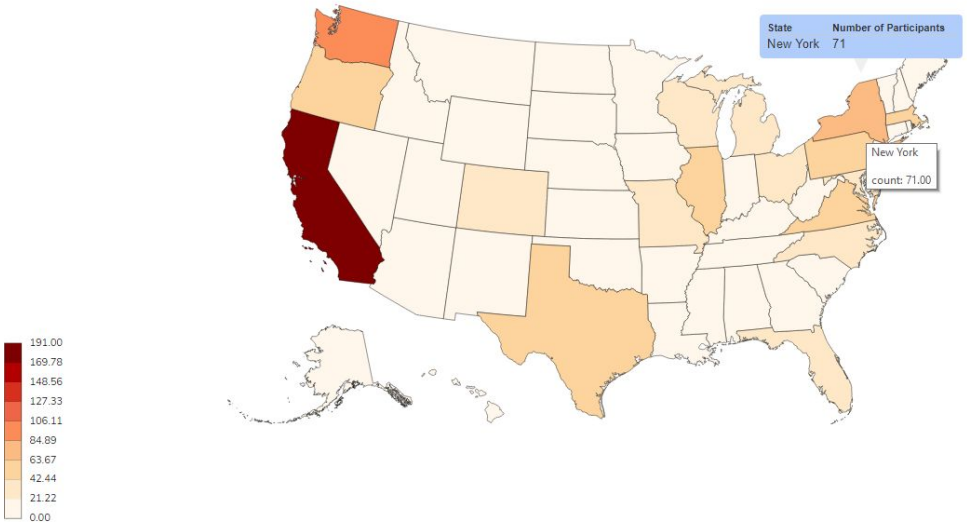



The user can also identify which participants gave a particular answer by hovering over one of the stacked bars in chart two. This also reveals a tooltip, which gives the response the user is currently over as well as what percentage of participants gave that response.

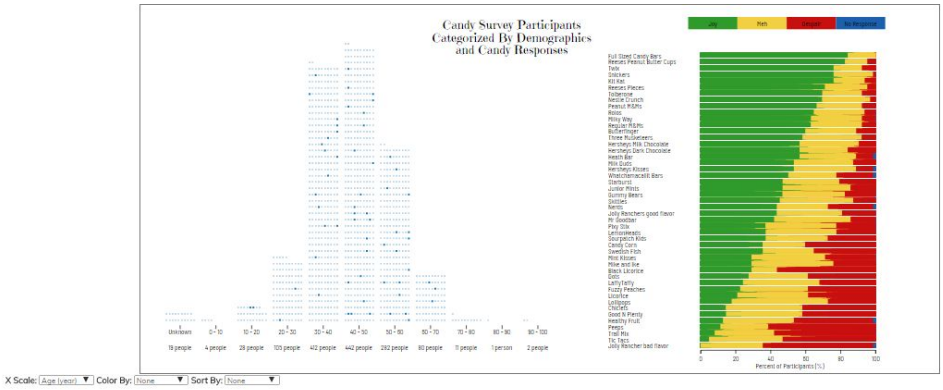


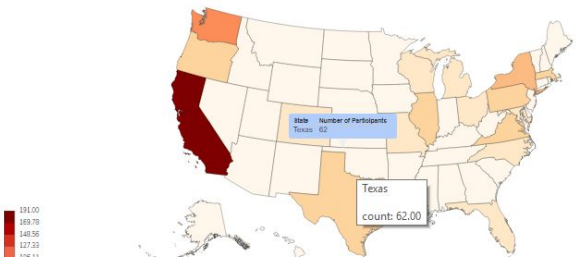
The map view below provides a rough distribution of the geographic location of participants. Tooltip interactions provide clear interactions and the map changes as a function of the brush on the dense pixel view (representing the geographic distribution of individuals who are in the brush area) as well as hovering over specific subsections of candy responses (such as Joy response for Dots)

---




Hovering over a state also updates the stacked bar view as well as the histogram. In this zoomed out view one can see that individuals in Texas preferred primarily Full Sized Candy Bars followed by Twix then Snickers.



As an example use case, if a user would want to discover the candy preferences of men between 50 to 60 years of age, the user would categorize the chart by gender, then color and sort the participants by age. Finally, the user can brush over the participants that fall under the age range. The selection would affect the second chart, which lists the candies by average preference. From these views we can see that Reese's Peanut Butter Cups are the favorite, followed by full sized candy bars and Kit-Kats. This information can then be compared to females at the same age range, revealing Kit-Kats at the forefront.