**Project Description**

This project visualizations the popularity of terms for soft drinks (soda, pop, coke, or other) in the United States by state. I wanted to compare map style visualizations to bar graph visualizations using this data set to better understand the strengths and weaknesses of each type of visualizations. I also wanted to provide visualizations that would fully explore the dataset and provide opportunities to make comparisons for different aspects of the data. I also wanted to visualize information that puts the dataset itself into perspective.

**Why It’s Interesting to Me**

I found this dataset interesting because it provides insight into which terms are the popular name for my favorite carbonated beverage in different states. I have always been fascinated by regional differences in the United States, and this dataset shows a very specific instance of that kind of difference. I also thought it would be a good opportunity to practice visualizing spatial data, which is one of my favorite kinds of data visualization. Finally, this project seemed like a good opportunity to practice comparing and contrasting different kinds of visualizations for the same data.

**Scope of Project**

This was a relatively simple dataset, compared to some of the other project options, so I wanted to many different visualizations to fully explore the data. I created fourteen different data visualizations using the program Tableau. I modified the dataset as necessary to achieve the desired visualization, calculating and creating several fields for the purpose of visualization. I visualized the first task, which was to show by state the favorite name used by creating 2 maps and a bar graph. I visualized the second task, which was to show all names for each state, by creating 2 maps and 4 bar graphs. Finally, I decided to add a third task and a fourth task, which were to show the population and sample represented in the dataset for comparison, creating a visualization to display this information, and to show the percentage of each state’s population that responded to the survey, creating a bar graph to accomplish this.

**CUTDDV**

For the context, these visualizations are hosted on the Internet via Tableau public. Tableau visualizations are very flexible in how you can create them. They have an interface that might be confusing to new users, particularly in the case of the maps. The expected user of this visualization would have basic Internet navigation skills. They could be interested in learning something superficial, like the most popular term in the state they live or how frequently their preferred term is used in a state they will be visiting. They might also be interested in understanding how the sample in this dataset compares to the actual population of the state. They may be interested in comparing frequencies of the different terms across states.

There are four tasks these visualizations are designed for. The first task is to identify the most popular terms in different states and compare them among states. The second task to identify the frequency of usage of each term in each state and to compare that among states. The third task is to compare the ratios between the population of each state to the ratio between the samples from each state. The fourth task is to identify the percentage of the population of each state that responded to the survey.

Here is a table showing the variables that were part of the dataset and created/calculated from the dataset to create the visualizations.

|  |  |
| --- | --- |
| **Variable** | **Data Type** |
| State/Province | Spatial |
| Population of State | Continuous |
| Number of respondents that use “Pop” from each state | Continuous |
| Number of respondents that use “Soda” from each state | Continuous |
| Number of respondents that use “Coke” from each state | Continuous |
| Number of respondents that use Other term from each state | Continuous |
| % of sample that use “Pop” | Continuous |
| % of sample that use “Soda” | Continuous |
| % of sample that use “Coke” | Continuous |
| % of sample that use Other Term | Continuous |
| Calculated % of sample that use “Pop” | Continuous |
| Calculated % of sample that use “Soda” | Continuous |
| Calculated % of sample that use “Coke” | Continuous |
| Calculated % of sample that use Other Term | Continuous |
| Calculated Number of Respondents from Each State | Continuous |
| Most Popular Term in State | Nominal |
| % of sample that use Most Popular term in state | Continuous |
| % of Population that Responded to Survey | Continuous |

Note that for each term, a percentage of the sample that used the term was given for each state. This percentage was rounded, meaning that when all four variables very added together, sometimes the numbers would be off from 100%. I recalculated the percentages from the data without the rounding to make my stacked bar chart, which was the only visualization I thought was significantly impacted by the rounding.

The dataset was complete, with no data missing within any of these variables. I would like to make a note that some of the data may be inaccurate. The obvious example I found was that Utah was listed with a population of 108708. I didn’t notice any other errors, but it is possible that there are less noticeable ones.

I used multiple visualization techniques for this data. The two types of visualizations I used were maps and bar graphs. Both seemed like good options to compare states and frequencies, as well as visualize data to provide insight about the dataset. Maps are an ideal choice for visualizing spatial data, because they represent the location where the spatial variable exists. For instance, it’s easy to see which term is most popular in a state by looking at a map. Bar graphs also seemed to be a good choice because bar graphs facilitate comparison of quantitative data. For instance, a bar graph does a great job showing which term has the highest percentage in a state.

For the maps, I used several techniques. For the first task maps, I created a map that was color-coded, with each color representing the most popular term used in that state. I also labeled each state with the most popular term for increased clarity. This allows a user to glance at the map and easily tell which term is most popular. For the next first task map, I also used color-coding for each term, but with a gradient on each color that represented the percentage frequency of the most popular term’s usage in that state. The lighter the shade of the color, the less frequently it was used and the darker the shade of the color the more frequently it was used. This gives the user an idea about how popular the most popular term in the state is. I made both of these maps interactive so the user could focus on a particular term and just see all of the states where that term was most popular all at once. This gives them more options to explore the data and understand where the popularity of each term is concentrated.

For the second task, I created an interactive map that allowed the user to select a term to see what the frequency of usage for that term was in each state. This was represented with a brown (the color of many popular sodas) gradient. The lighter the shade of brown the less frequently the term was used and the darker the shade of brown the more frequently it was used. The user can select the term they want to see from a dropdown and the shades over the states with change to reflect the selected term. The percentage frequency of the terms use is also overlaid on all of the states so it’s easy to see the number. This makes it easy for a user to see how popular a term is across the country. It also gives them the ability to flip between terms to compare and contrast. I also created a visualization of four maps, each with a color gradient representing the frequency each term was used. I used a different color for each map to distinguish them and the different terms they represent. The four maps are presented in a 2 x 2 arrangement, so they fit on the screen and allow the user to easily glance from one map to the other to compare.

I used bar graphs for all three tasks. For the bar graphs visualizing data about the terms for soft drinks, I picked a pleasant color palette from Tableau and color coded each term. On every bar graph, the color code is the same, so it’s consistent across the bar graph visualizations and the user can easily associate each color with its respective term. On my Samples Compared to Population visualization, I gave different colors to the two bar graphs to distinguish them as showing different data. The purpose of that visualization is to compare how proportionate the samples are the actual populations of the states, so the different colors (as well as the top/bottom placement of the graphs) are meant to encourage that kind of comparison, instead of mistakenly assuming the bar graphs are measure the same scale. The Percentage of Population that Responded to Survey by State visualization was given one color to represent that the same variable, percent of population, is being visualized for each state.

I oriented the Term Popularity by State – Interactive visualization and the Responses for Each Term by State – Interactive visualization vertically because with 51 states with four terms each, the bar graph was quite long, so a vertical orientation made it easier to read and compare data. I used a horizontal orientation for all of the other bar graphs because I found that it was easier to compare them in a horizontal format. When only one variable was being considered for each of the fifty states, it was possible to see it all in a horizontal view. Another technique that benefitted me greatly with these bar graphs was giving the user the option to filter by state. With 51 states and four different terms, there was too much data to make comparisons between specific states easily. However, that problem goes away when the user is given a filter, giving them the freedom to view the data for only the states they are interested in, making it easier to compare and discover.

I also created 2 stacked bar graphs. The first one shows the raw response number for each term, showing the total number of responses for each state and the proportion of those responses. The second one shows the percentage frequency of each term in each state, allowing easier comparison of proportions of terms from state to state, since the stacked bars will all total to 100%. Both of these give the users the ability to understand how frequently a term is used in a state compared to the others, and they give users the ability to compare that information between states by making use of filters.