Starbucks Stores in the U.S.

Albina Chowdhury, Cesar Cisneros, Isuru Herath, Yunjie Liu

Description of Data

We utilized three datasets in this project. We wanted to select a dataset that we're interested in as well as complex yet not that ambitious. The complexity of the dataset is determined by the number of columns and rows, which we intentionally chose to be over 10 and 5000. respectively.

The first dataset ('us-smaller.json')is a topojson file for the United States that we got from Professor Rz's Github account. This topojson file includes 53 features, each feature as a state, and will be used to draw multipolygons on a choropleth map.

The second dataset, our main dataset, directory.csv, is a Starbucks worldwide store data we found on Kaggle. This dataset has 13 columns and more than 10,000 row, with each row describing details of a specific Starbucks store, such as country, state, ownership types, as well as longitude and latitude.

The third dataset ('population.csv') is a population total dataset published on census.gov.com, which displays the population and population change in 2010 to 2019. This dataset has 151 columns and 58 rows, with each row describing a region or a state of the U.S.. Besides the name of state/region, the columns also included census population, estimated population, birth, death, natural increase, international migration, and more for each year from 2010-2019.

However, although the datasets we gathered are extensive, we couldn't use all of them, and need to pre-process and combine data and store it in a new csv file called count.csv so we can easily load and make changes to it. Because we want to focus on the stores in the U.S., we first used sql to filter out the rows in the U.S. in the directory.csv, which we then grouped by state to compute the number of total stores in that state. These would constitute the 'name' and 'total stores' column in the new csv file. Then, because we also want to look at ownership types of the stores by state, we used sql to get that data and put the count of each type for each state into the 'company_owned' column and 'licensed' columns in the new csv, respectively. Because the Starbucks data was published in 2017 and has not been updated since, we extracted the estimated population, column O, in the population.csv and put it into the 'population' column in the new csv. We also divided each number by 1 million and put the output into the 'population(million)' column. Then, we divided the total Starbucks store of each state by their population in million, and get a ratio, which we put in in the column 'starbucks_per_milion_inhabitants'. Last but not least, we included a unique id and abbreviations for each state, which we put in the 'code' and 'short_name' columns so we can do more in the later stages in the visualization process.

Visual Design Rationale

Our main visual design contains three essential visualizations: a U.S. map, a sankey(alluvial) plot, and a pie chart.

The U.S. map will be color coded by the number of Starbucks stores in that respective state, and there will be a color legend underneath to show the range of stores each color represents. This is necessary because there is a large gap between the lowest and highest number of stores, and by only looking at the color without the legend, the user can only see relative relationships among states but not have a good idea of approximately how many stores there are in that state. There is also a slider underneath the map that can control which area of the map lights up(more in the section below). The marks we employed are lines as state outlines and circles for individual stores when zoomed in. The visual channels employed are spatial regions of states and colors of each state.

With the map of stores per state, it is interesting to investigate how many of those stores are company owned versus licensed, which is where the alluvial plot comes into play. The alluvial plot allows us to display flow with nodes as rectangles and links as arcs. Our plot has 51 nodes on the left and 2 nodes on the right. The left represents 51 states in the U.S. and the right represents two ownership types of Starbucks stores in the U.S. For each state node, there will be 2 outgoing links, each pointing to a ownership type on the right. The links have a width proportional to the number of each type of store in that specific state. The marks that we used are curvatures while the visual channels we used are color hues, shapes such as rectangles and vertical positions of the links.

With the alluvial plot, the users can see the rough relationship between the number of company owned stores and licensed stores. For example, if an arc is wider for licensed from the state of California, it means that in California there's more licensed stores. However, this contrast is not obvious for states with fewer stores, where the links are slim, and it's hard for users to know the exact number of each type, which is why we decided to create a pie chart when a user clicks on a state. The pie chart will be divided into two slices and represents the numeric proportions of each type of stores. This way, the user can see clearly the percentage of each type of store in that state, which is especially beneficial for states with fewer stores. We decided to use pie charts instead of bar graphs because we only have 2 categories, and using a bar graph with only 2 bars will look strange. It also wants to show the two types of stores as percentages of a whole. The marks we used are areas in the pie chart and the visual channels we used are colors.

It is also useful to discuss the styling of the page. When the page first loads, there will be a U.S. choropleth map on the left with a legend and range slide beneath it as well as an alluvial plot on the right. We divided the page into halves because of two reasons. First, the width of the map is bigger than its height, and the alluvial plot can be very long, so if we put the snakey plot beneath the map, the users will have to scroll down a lot to see the whole picture. In later stages of interactivity, the users will also not be able to observe spontaneous changes to graphs if they have to scroll up and down, so it's better to put them side by side horizontally. Secondly, because the legend describes the color scale of the map and the slider controls what region lights up on the map, it makes sense to group them inside a <div> and use vertical flexbox to style them. It would not make sense to the user if we group the legend with the sankey plot and put it on the right.

Interactivity Design Rationale

There are quite a few interactivity ideas related to the map, the slider, and the sankey plot.

Interactivity for the map:

First of all, the users are able to zoom into a state and the circles, each representing a Starbucks store, will pop up. When the user zooms out, the circles disappear. The user will also be able to drag the map. We added this zoom feature because by only looking at the color coded map when the page first loads, the users will only get an idea of which states have the most Starbucks stores, but it would be impossible to know where in the states have the biggest cluster of stores. This zoom feature allows the user to dive into the specifics of a state of their interest and find out more about where the stores are located within that state.

Secondly, there is a range slider under the legend for the map, and the users are able to control what areas of the map are emphasized when they change values on the range slider. This allows users to investigate the ratio of the number of starbucks store divided by population in millions. As they manipulate their customized range, they can gain insights on which states' ratio fall into that range. This is also why we chose to do a range slider instead of a simple slider: the first allows the users to manipulate the minimum value while the latter doesn't.

Moreover, the map allows mouseover and clicking events. The user can also hover over a state and that specific state will light up, indicating that it is clickable. We added this mouseover because we want to give the users a hint to click on the map.

Interactivity for sankey plot:

The users are able to hover over the links in the snakey plot and there will be the number inside that specific link shows up on the side. For example, if there are 2 licensed stores in California, 2 will show up on the side when that link is hovered over.

Interactivity that combines the map and the sankey plot:

First of all, when a user clicks on the map, the map will zoom into that specific state and a pie chart that represents the number of company owned versus licensed Starbucks stores will be displayed and the original sankey plot will be hidden. We wanted to remove the alluvial plot here because the zoomed in map and the pie chart both represent data from a single state whereas the alluvial plot represents data from all states in the U.S., which can cause confusion for the users.

Second of all, after the user clicks on a state, they can click on the reset button underneath the pie chart, which will bring the whole page back to its original state: the alluvial plot will show up while the pie chart will go hidden. The map will also zoom out to display all the states. We developed this interactivity because if a user clicks on one state, there is a large possibility that they will select another state, but to enable them to do that, we need to get the map to display all the states instead of zooming into one state. We made this discoverable because the reset button is very obvious and the text on the button makes it obvious what it's for. However, we decided to keep the pie chart because the user might want to compare pie chart percentages side by side. To prevent the user from selecting too many states, an alert will pop up, asking the user to refresh the page to start over.

The Story

We began our journey by exploring a Starbucks dataset that contained information about the number of stores per state and the types of stores that were available in said state. Immediately, it became apparent certain stores like Teavana were not available in the United States, which prompted us to start with an interactive map capable of visualizing the number of Starbucks locations per state. Our main goal with this project thereafter became to demonstrate the breakdown of the two types of stores available in the country, namely those that are company owned and those that are licensed.

Thus, we understood we wanted to convey the disparities of Starbucks stores across the country coupled with the differences present in store type. In theory, the viewers of our visualizations would be able to gauge whether there are clusters of Starbucks stores in certain regions of the country, whether there are more or less company owned stores, and whether there are more or less licensed stores. These findings can then be used to hypothesize the reasons for such a divergence of stores – both in count and store type.

Among the findings we observed, it became clear that cities with higher populations had a higher count of Starbucks stores. This finding follows our intuition affirming that more people in a given state indirectly translates into more potential for business success, and hence more Starbucks stores. At the same time, we saw some disparities within states with big, metropolitan cities. For instance, even though both California and New York are densely populated, California has far more Starbucks stores than New York (~22% more). Such a finding became surprising given the many Starbucks locations present every few blocks in New York City. This metric, however, only closely resembles the number of stores in one city of the entire state of California (Los Angeles). Additionally, it was surprising to note that Washington did not have the most stores, albeit falling very close to the top, given that this is the location where the franchise was founded.

After our main goal was achieved, we wanted to explore our secondary goal: to explore if it is true that the more population there is in a state, the more Starbucks stores there are in that state. We examined this relationship using a ratio (starbucks stores per million inhabitants) as well as a slider that shows the range of the ratios as values. Thus, we hypothesized that the more population, the more Starbucks stores, the smaller the ratio.

By manipulating the slider, we found that all states have ratios in the range 1-860. When manipulating the minimum value of the slider, we noticed that Virginian and Rhode Island are the first ones to darken. This might be because these are states relatively small but have larger populations. We also noticed that when setting the minimum value to above 345, only Vermont and North Dakota are still highlighted. This is not surprising because those are two states that have small numbers of Starbucks stores yet have very low population, which makes the ratio really big.

We also found that our hypothesis can be turned down. Take California, the most populous state, as an example. California has a population of 39 mil, and it has 2821 Starbucks stores, which produces a ratio of 71. When we set the minimum value to 71 on the slider, we noticed that only 13 states remain, which means that only 12 states have a ratio lower than California. Thus, although the population of California is large, there's not enough Starbucks stores to make the ratio small, which contradicts our hypothesis.

Team Contributions

At the end of your PDF file, include an outline of team contributions to the project. Identify how work was broken down in the group and explain each group member's contributions to the project. Give a rough breakdown of how much time you spent developing and which parts of the project took the most time.

Work was initially broken down into two visualizations, namely our U.S. map and an alluvial plot – enumerating the count of stores by store type. Two members (Albina and Jenny) worked on creating the U.S. map while the other two members (Isuru and Cesar) worked on the alluvial plot.
As it pertains to the alluvial plot, Isuru worked on formatting our Starbucks data (initially a CSV file) to fit the format of an alluvial plot (a JSON) with corresponding nodes and links, including each link's sources and targets. Additionally, he worked on creating the alluvial plot that displays all of the states. Since the original dataset consisted of rows containing data about individual stores, Isuru worked on processing the data so that it contained counts of stores in each state and the counts of stores in each Ownership type for each state. He also adjusted the width and the padding of the color legend to increase visibility of the values on it. Additionally, he worked in collaboration with Cesar to implement the interaction between the map and alluvial plot. This involved sending the appropriate subset of the store count data from the clicked states from the map into the drawAlluvial function. This would cause the drawAlluvial function to generate an alluvial plot based on the data from only the clicked states.

Cesar focused primarily on filtering the data of the JSON to assign specific nodes and links to the alluvial plot with Isuru. He added IDs to the existing JSON corresponding to the states that appeared in the US map data in order to match these two accordingly. He also focused on the styling, and positioning of the alluvial plot in relation to the other elements in the webpage. The majority of the time went into the filtering of the JSON, which often resulted in errors stating our data was circular and thus not plottable. Furthermore, other errors of undefined values became rather common in the alluvial plot, making it difficult to discern given the extensive filtering that took place in our file. Above all, the most difficult part was getting the JSON file in the format for an alluvial plot as our data was previously in a CSV.

Yunjie and Albina both manipulated the original dataset and cleaned the dataset using SQL code and excel. From there, they both worked on the creation of the US map, using the US map JSON file provided by the professor. Once the map was developed, they both worked on the zoom interactivity, the click interactivity, the color scale, and the hovering interactivity of the map. Once those interactivities were created, they both went on to work on the slider and the interactivity the slider entailed, which was the changing of the opacity of the states within the range based on the state's starbucks store per million population number. After they implemented that, they created the pie chart that comes up when the user clicks a state, which displays the percentage of the stores within each state being either licensed or company owned. They worked on making the alluvial plot hidden once a state was clicked.
Finally, they worked on the reset button and they both made the alluvial plot visible, reset the zoom of the map, and the range of the slider to the initial state.