Basic Info:

- **Project Name:** Demographic Analysis
- **Names:**
  - Christopher Mertin – u1010077 – cmertin@cs.utah.edu
  - SeyedMajid RasouliPichahi – u1013493 – maj.rasouli@gmail.com
  - Ashkan Bashardoust – u1011913 – u1011913@utah.edu
- **Repository:** https://github.com/cmertin/US_Stats

Overview and Motivation:

Demographic Analysis can be used for many purposes.

One example could be migration. Each year many people migrate within the US or from other countries to US for business, study, and job purposes. They try to find out which part of it is most suitable for them. Information such as percentage of highly educated people, population of the youth, and average income in that area could help a lot. Creating this kind of visualization would these people to decide which areas suit their criteria to move.

Another example would be usage for companies. Many companies that are starting their businesses try to find the best areas for their selling market. Based on the information of average income, population of each age ranges, and the education level of people living in that area, they get a first estimate that how's their product is going to sell in that area. They can use this information for deciding whether those people are suitable for employment or not.

Other usages:

1. **Public usage:** This project could be used for finding a place to migrate, according to a specific lifestyle. For instance, a young person may have higher tendency to go to a place with younger range of people to have more fun activities and nightlife. Or to find a place having people with higher university degrees. One would prefer to live in a city, which has people with higher income rates.
2. **Journalists:** can use this project to conduct researches and use them as background for articles.
3. **Commercial companies:** can use this visualization as a basis for market research and to figure out the supply and demand ratio for their businesses.

The data we are using provides statistical analysis on certain categories around the US, will provide relevant information on the types of people in various areas of the United States. You can therefore select certain attributes or areas such that you can tailor your business to your current demographic or see to where you want to expand.

Related Works:

We have checked some population visualizations on the census.gov website, including:

1. Before and After 1940: Change in Population Density:

https://www.census.gov/dataviz/visualizations/010/

2. Distribution of Hispanic or Latino Population by Specific Origin:

https://www.census.gov/dataviz/visualizations/072/

3. Population Distribution by City Size, 1790 to 1890:

https://www.census.gov/dataviz/visualizations/005/

We also have checked the visualization techniques used in the final projects from the last Visualization course.

Questions:

What questions are you trying to answer?

We are trying to show how many people with those selected attribute live in a specific area.

We want to show the numerical difference of people in a certain category in each area.

How did these questions evolve over the course of the project?

What new questions did you consider in the course of your analysis?

We want to show how the population of a selection changed over the years.

Data:

The data will be provided by census.gov http://factfinder.census.gov

The U.S. census provides the data in easy .csv format with the info we need from 2010-2015 with the appropriate categories. The only “data cleanup” that needs to be done is to turn the columns of the data selections/attributes into percentages so that we can perform statistical analysis on those columns to give the user appropriate numbers.

For example, if a county has a population of 1 million, 450,000 of which are male, and 4,000 of the original 1 million are native American, we will change the data such that it reads 1 million for the population, 0.45 for the males, and .004 for the number of native Americans, and so forth for each of the data points. Therefore, if the user asks for the number of native American males around the US, we can say for that given county that it's approximately 1 million * .45 * .004 or approximately 1,800 Native American Males in that county on average. This can be done for all the attributes so the statistics can be calculated on the fly as the user selects the data.

The data that we looked at is:

- Age and Sex (population of males/females at different ages in a given region)
- Education (No High School, High School/GED, Some College, College, Graduate/Professional)
- Race
- Marital Status (Never Married, Divorced, Separated, Married, Widowed)

One problem with this data was that the census bureau did not provide the data specifically for each age. For example, the population of age and sex were like this {18-19, 20, 21, 20-24, etc} for both males and females. To get it for each age, what we did was assume an equal distribution of people for the age ranges. For example, there were roughly 70,000 males in Alabama 18-19 years old in 2010. To make it "fair," we split the population size such that we gave 18-year-old men a population size of 35,000 and the same with 19. This was a fair assumption as most of these should be a relatively equal distribution since the ranges between the years were quite small.

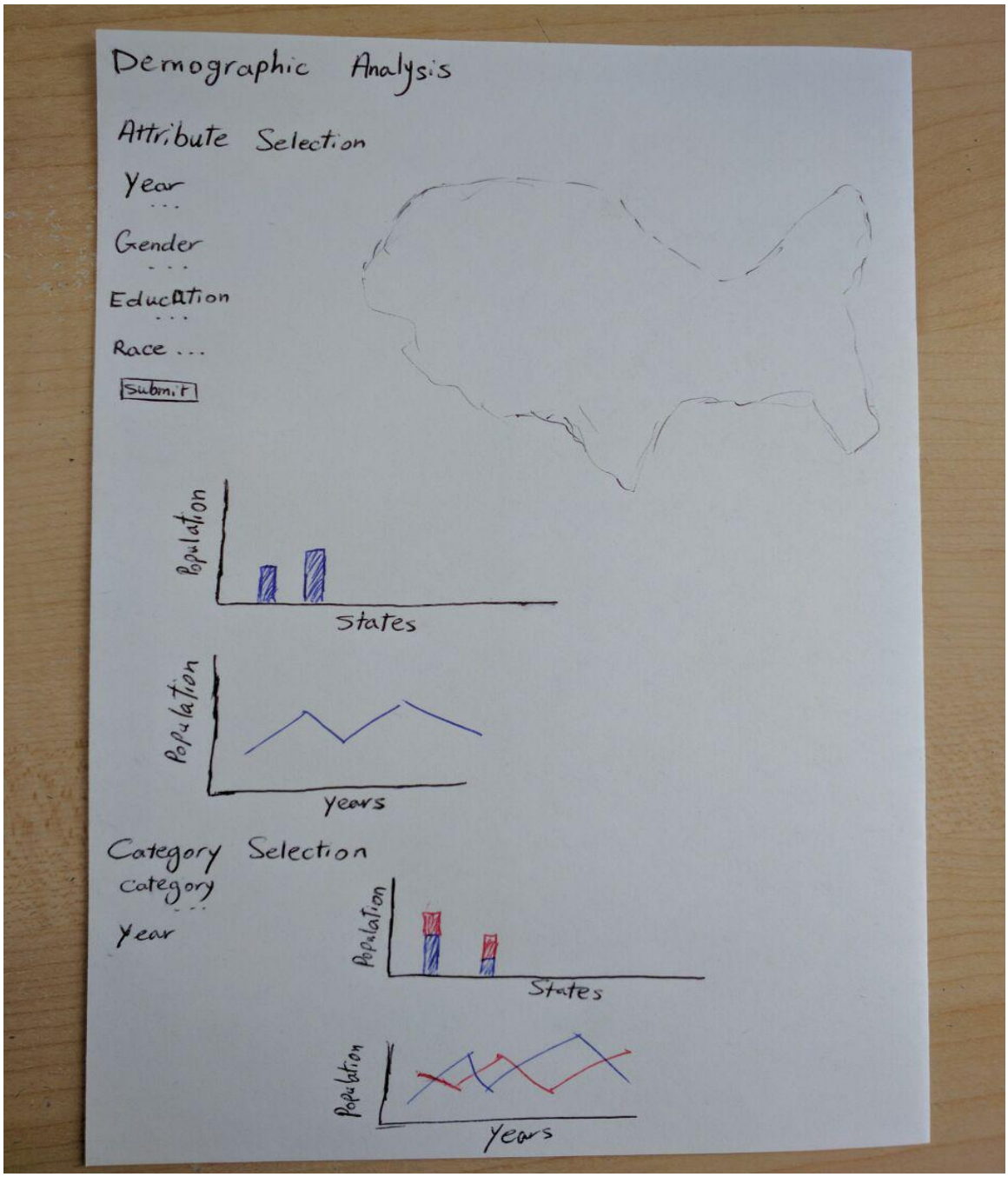
The rest of the data (not dealing with age and sex) was provided as a percentage of the population (at a given age), so we stored the fractional values as each column. Therefore, we can use what we stated before of taking the population size and multiplying through by each percentage value to get a rough number on the number of people of the selected attributes.

The data was parsed into two different formats, JSON and CSV. This was first output as a JSON file for each of the year for all of the states and counties, but the resulting file size was too large. Therefore, we opted to add in a CSV file as well, which greatly reduced the size of the files.

Each of the columns in the CSV file represents each combination of attribute, for example 18_M_No_HS is the percentage of the number of 18 year old men in that geographic region with No High School degree. The columns are broken up by age and gender, and there is a permutation of each of the attributes. This may need to be changed later when we try to access the data if we cannot find an efficient way to convert it to JSON for easy and quick access after reading in the file for the first time.

Exploratory Data Analysis:

We are using a US map as our major visualization to visualize our data. The user should be able to select multiple attributes such as age, gender, education level, income and race. Then the map should show color map of the number of people in that selection in each state.

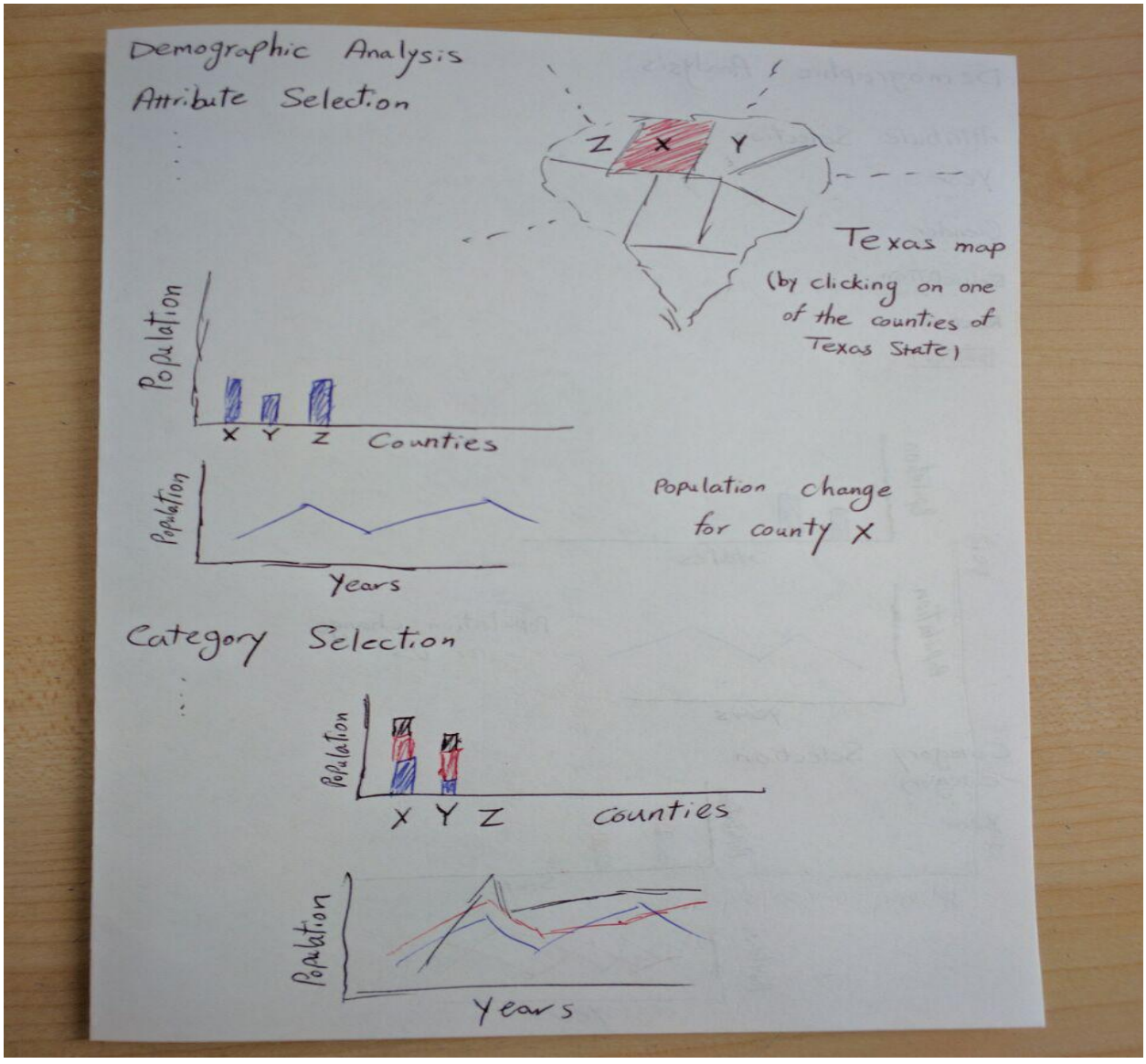


There will be sliding scales to select ranges, from which the data will be parsed and the results will be populated on the screen.

The idea is to have a color scale on the map such that it will act as a “heat map” that will show the data all around the US. Then, user can also click on a certain state and it will “zoom in” to the county level, for which it will have more precise colors as well for each individual county.

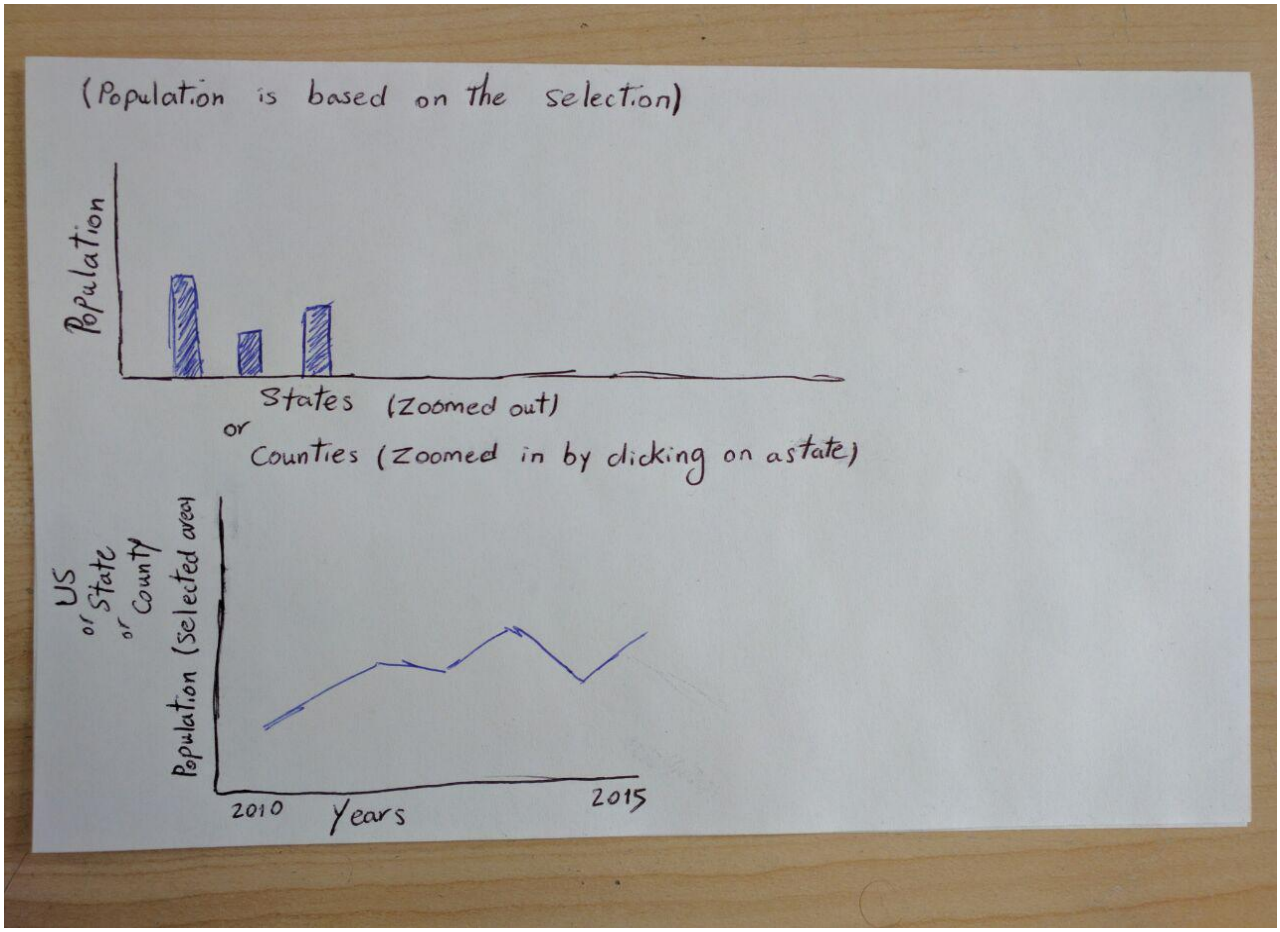
It’s necessary to mention that we have 3 choices for selecting the area:

1. The default is the whole US. Bar chart compares all the states and line chart uses the population of people in the whole country.
2. By clicking a state the map zooms into that state, showing all the counties of it. Bar chart compares all counties and line chart uses the population of people in that state.
3. By selecting a county, bar chart doesn’t change (comparing all counties) and line chart uses the population of people in that county.



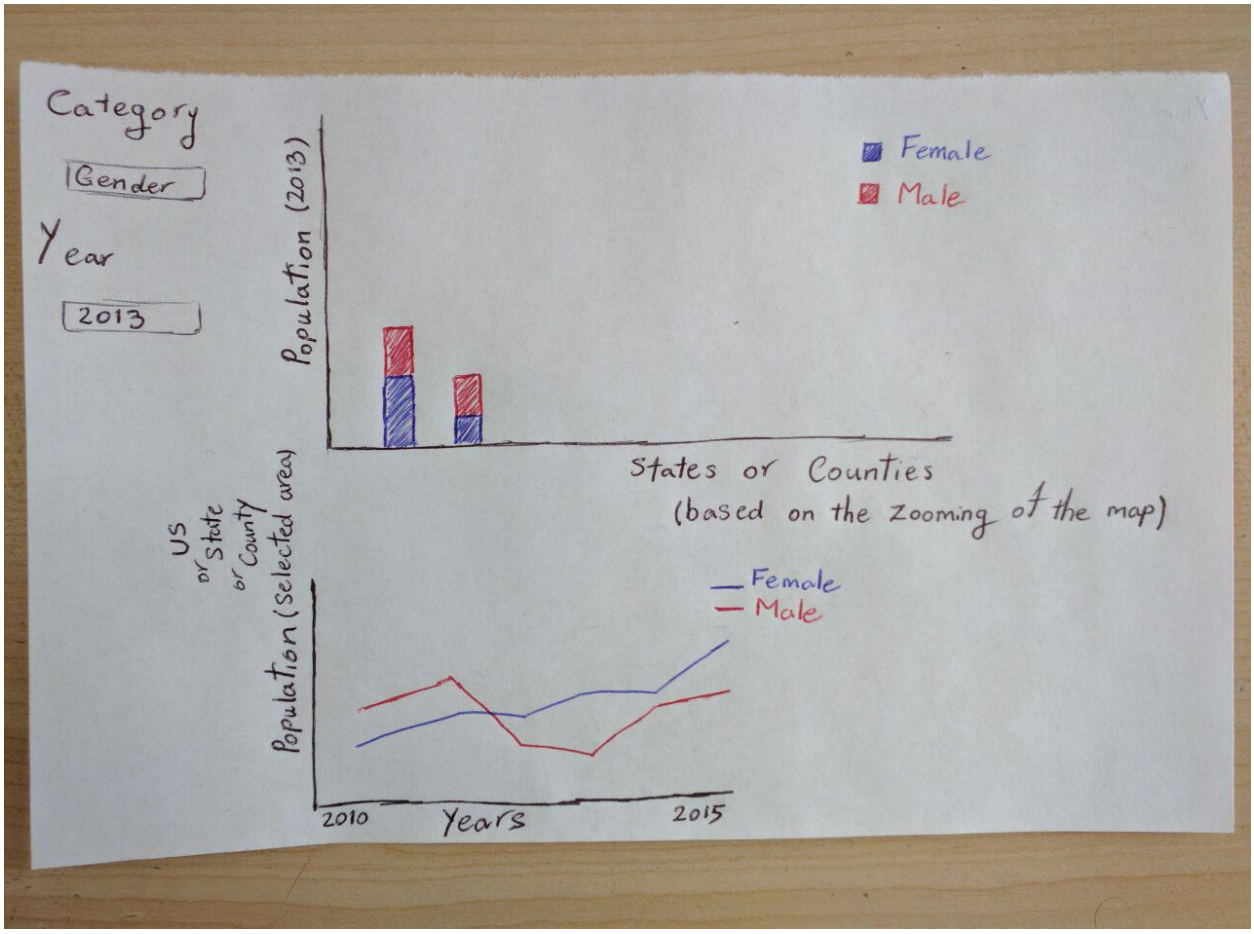
For better analyzing our data, we are using 2 different methods:

1. First method is that user chooses a selection of all categories. Then we use this information for visualization.
   We visualize these charts below of our map.
   There will be a line chart with one line representing the population of people with those selection attributes in the selected area over the years.
   The other chart would be a bar chart with a bar for each state/county (based on the area chosen), which shows the population of people with that selected attributes in each state/county. It helps to compare this population of all states/counties together.



2. Second method is that user chooses a category. Then we use it for visualizing and comparing subcategories.
   We visualize these charts at the bottom of the page.
   There will be a line chart with a line for each subcategory, which shows the change of the population of people in that subcategory in selected area over the years.

The other chart would be a stacked bar chart with a bar for each state/county (based on the area chosen) which shows the population of all subcategory in each state/county stacked on each other. It helps us compare the subcategories of an individual state/county, also compare each state to the other one.



Design Evolution:

We thought about using donut chart, but since there are 50 states, it would be hard to read the chart.

We were considering using as much as the subcategories of each category bar chart for each state. Considering there are 50 states, and we may have 8 bars for each states, we should visualize 400 bar charts. So we decided to change that to stacked bar chart.

We decided to have some features that are necessary for our project:

Some features that are needed for this project to be successful are a good color scale such that the data can be fairly represented and easily understood. On top of this, we need to make sure

that it is customizable enough that the users can get the results that they want and that it will represent the country appropriately. While some of these attributes won't be completely independent (for example income and race), this should provide a good enough approximation to the data.

The other features that we could use for our project:

We could potentially let the user decide the color scale they would like to use and also the type of chart they would like to see for the info on the given county (donut chart, stacked bar chart, pie chart, etc.). This is not required to complete the project, but some user customization on this aspect would make it better for the user in some aspects.

Group Critiquing:

When we had to meet with another group, they had brought up good ideas, but the most notable was that we should be able to compare two states. In our current implementation, they mentioned that it may be difficult to compare two states with a stacked bar chart (depending on the attributes) since they're not aligned. They suggested that we add an additional chart to it so that the comparison would be easier.

UPDATE: new design evaluations are explained at Implementation section.

We deviated so many times from our proposal:

- As we mentioned in the implementation section, we decided to get rid of the line chart.
- Since animated visualizations are always gets users attention, we decided to add a bubble chart, which has 3 factors to compare the data between states: percentage axis, color and circle size.

Using our visualizations users can get the population based on the desired specifications and also the percentage of matching selection. The users can compare different states or the most important counties of a specified state with each other.

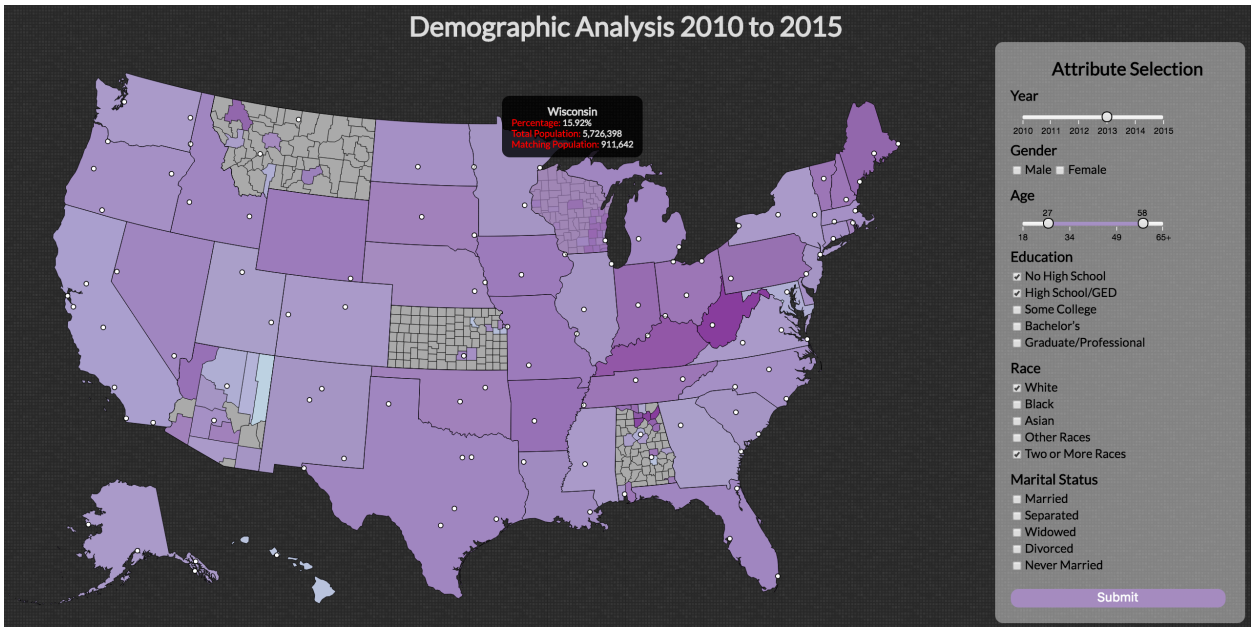
The US map can be used to get epic picture of the population distribution and it's comprehensive for understanding it. Other charts can be used for comparison.

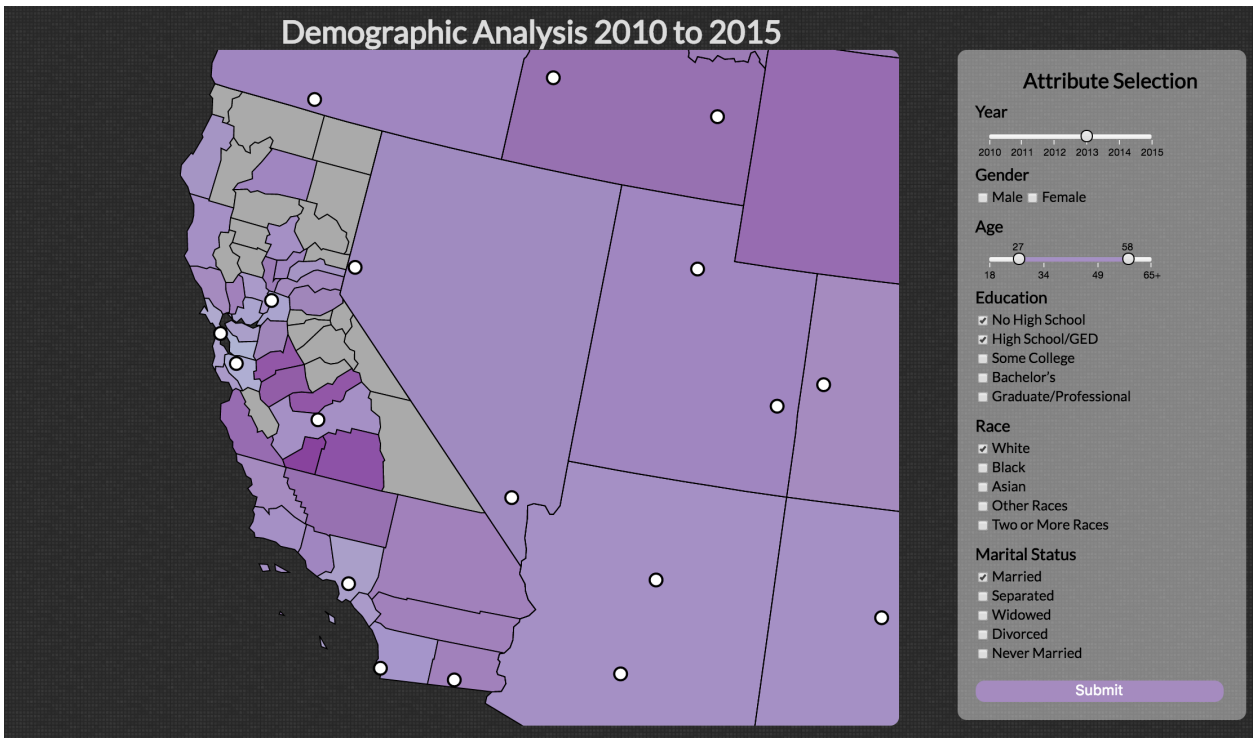
Implementation:

Map:

The US map colors every state according to its percentage data, based on the selections that the user makes. When the user clicks on the map, it zooms on the state, and shows the same information for the counties. By clicking again on the state, the map zooms out. Also by right clicking on a state, it shows the counties of that state colored based on the selection, but doesn't zoom in. You can right click on as many as states you want. Then by right clicking again

on them, you can make it to look as before.





In the database that we have, the information is only available for the most important counties for each state. The other ones are colored grey since we have no information for them.

There are also 100 more important cities drawn on the map with circles. If the user hovers on cities or counties, a tooltip is being displayed with information.

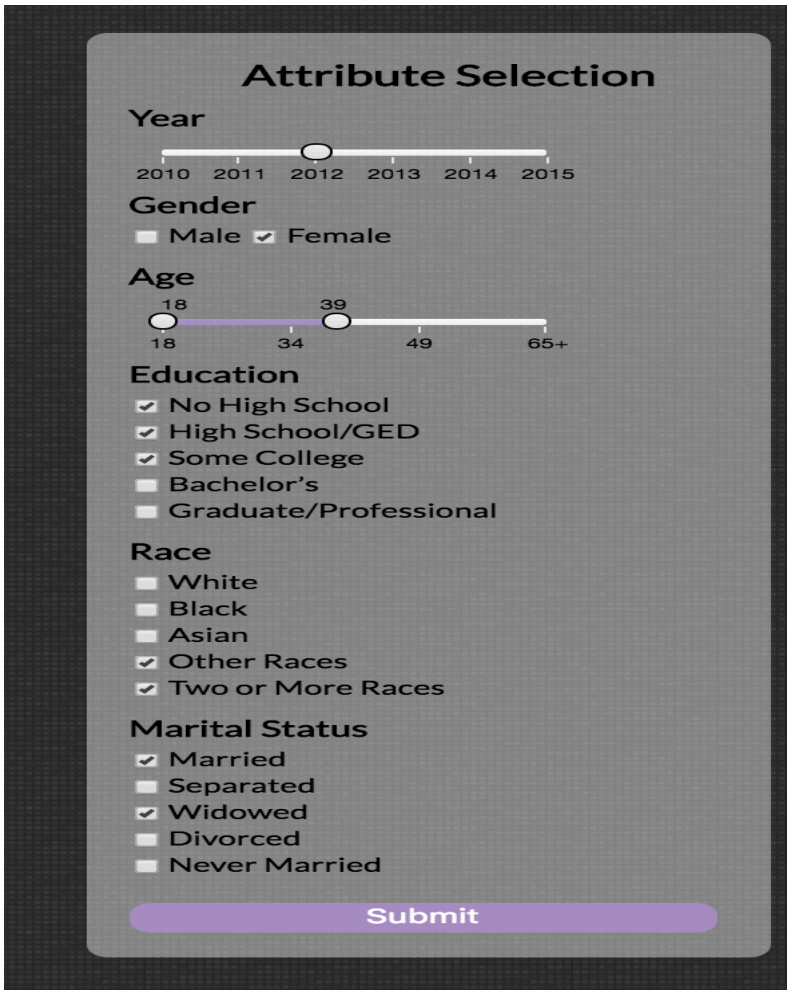
We read some codes from this websites, and then changed the code so that it would be compatible to JavaScript version 4.

https://bl.ocks.org/mbostock

http://techslides.com/demos/d3/us-zoom-county.html

By hovering on the each state, it shows you the name of that state, the exact percentage and the matching population of the selection in that state and the total population of that state .

Selection pad:



- A scale for selecting the "Year" (2010 – 2015)
- A check box for selecting the "Gender" options (Male, Female)

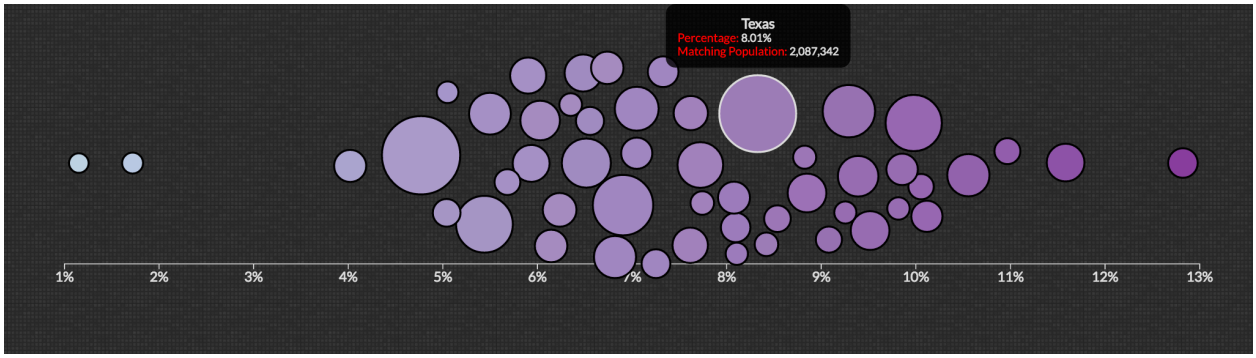
- A slider for selecting of the “Age” range (18 – 65+)
- A check box for selecting the “Education” level options
  (No High School, High School/GED, Some college, Bachelor’s, Graduate/Professional)
- A check box for selecting the “Race” options
  (White, Black, Asian, Other Races, Two or More Races)
  we had two other options for “American Indian/Alaskan Native” and “Native Hawaiians/Pacific Islanders”. Since their population was very low, we combined them with the “Other Races” option.
- A check box for selecting “Marital Status” options
  (Married, Separated, Widowed, Divorced, Never Married)

We implemented this part of the code in the index.html file. The implementation was straightforward and I don’t think we need to explain it.

Bubble Chart:

This is the second plot which shows each state represented as a bubble. Their position on the x-axis is based on the percentage of the population in that state that matches those attributes. The size of the circle is relative to the number of people in that state that match the attributes. Taking the matching percentage and multiplying it by the total population achieved this.

By hovering on the each bubble, it shows you the name of that state, the exact percentage and the matching population of the selection in that state.



Bar Chart:

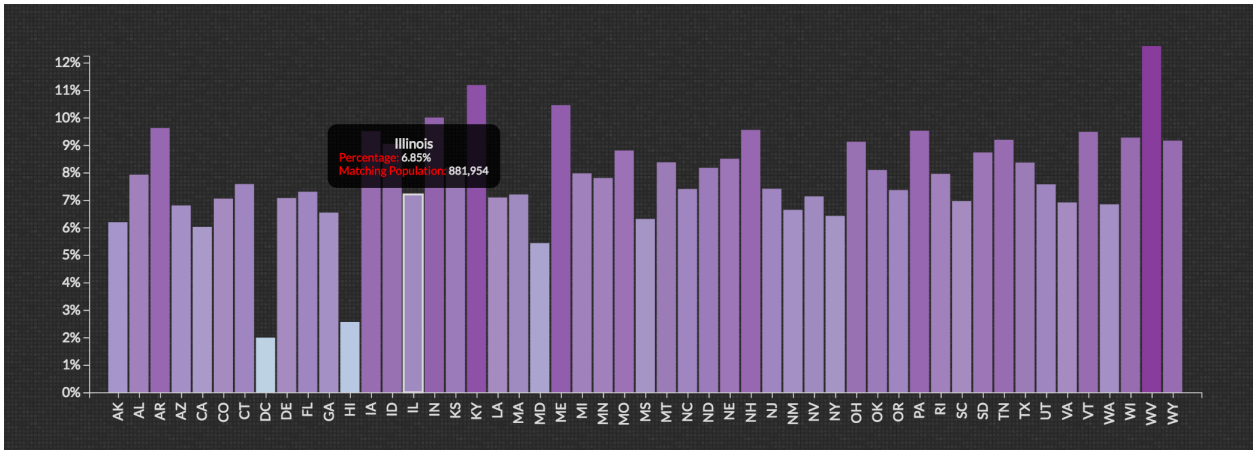
In this chart we show the percentage of the population with the selected data of the total population of each state. It colors the bars based on the percentage.

You can sort the states based on the alphabet or based on the highest percentage.

By hovering on the each bar, it shows you the name of that state, the exact percentage and the matching population of the selection in that state.

At first we tried to visualize this bar chart based on the population. Since the population in California is so much higher than other states foe almost every selection, the others would look really small comparing to this state. So we decided to use percentage instead of population.

Also for showing the name of the states in the x-axis we used a function that if you give it the name of the state, it returns you the abbreviation of that name.



Stacked Bar Chart:

In this section we have two selectors for choosing the year and the category.

Based on this selection, by using a stacked-bar-chart we show the population of the each category in each state.

First we get the selected information from the index.html file. Then based on that we choose the data and compute the population of each sub-category for each state. (We have the ids for all of the states).

For race, the information is inside each state separately, but for other categories, we have to search over all of the ages and all of the genders and add up the percentages of each sub-category, and multiply them by the population of that state.

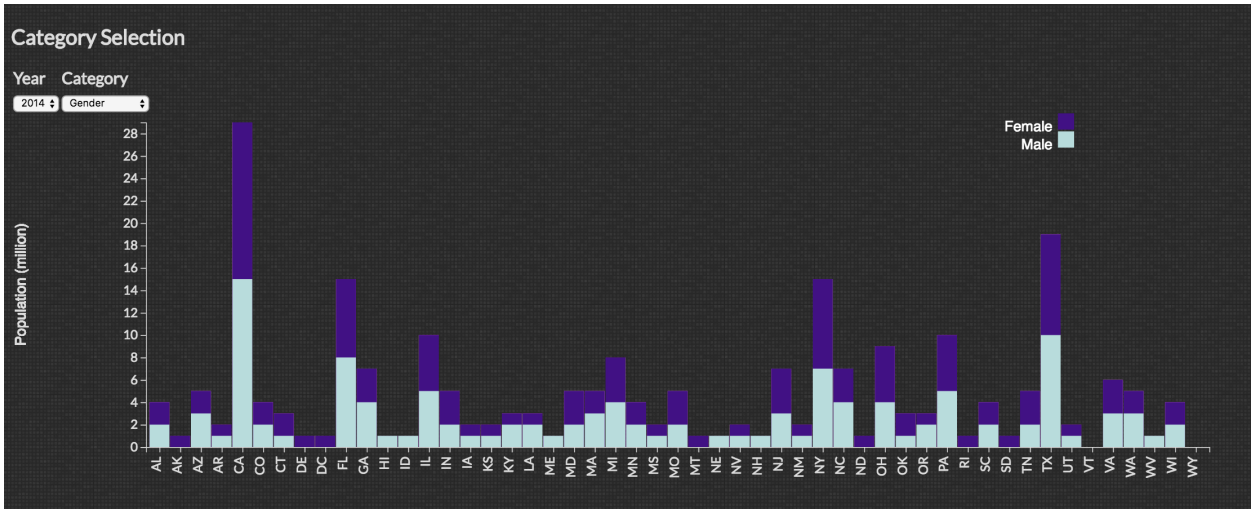
Then we add all of the information of each subcategory of each state to an array and build the stacked-bar-chart using this array, and scales that we defined before.

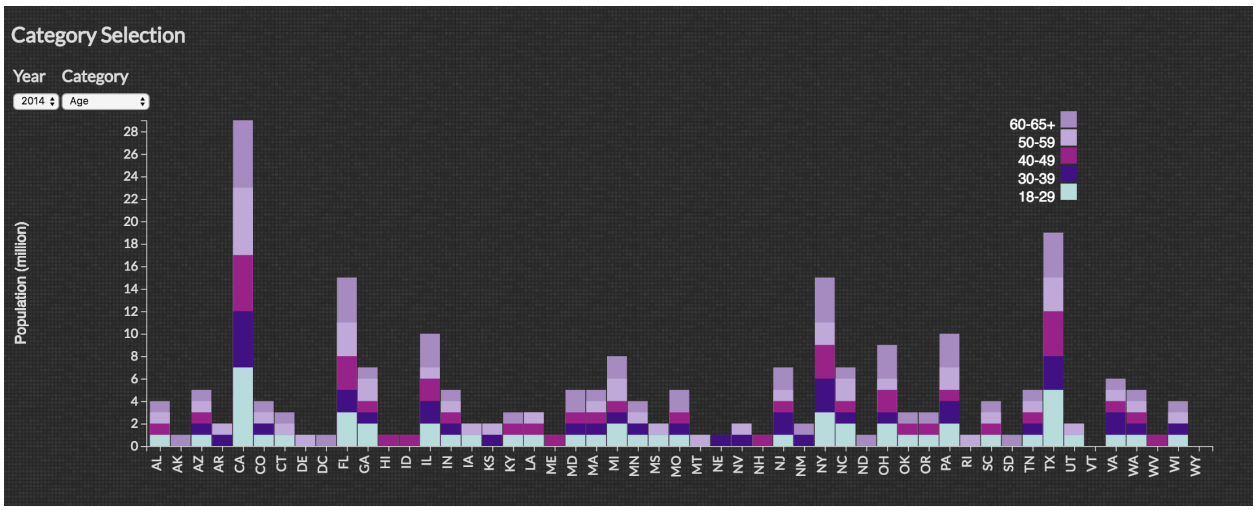
We used 5 different colors for each sub-category and then visualized the labels of each color with legends.

One of the problems that we had was that the transition wasn't appealing. There were two reasons for that. The first reason was that the gender has two sub-categories but others have five. So we added 3 empty bars when visualizing gender. The second problem was that the bars were disjoint. For example the first one was from 0-2 and the second one was from 2-5. By changing the bars as this: the first one 0-2 and the second one 0-5. This made another issue, which was the next bars, would covers the old ones. By reversing the array, we build the stacked-bar-chart in a way that the bigger ones would be added before the smaller ones.

Another issue was that by changing the selected year, the data wouldn't change. Then we realized the sum of the population for each state stays the same and over the years data changes were so small.

Last issue in this section was that the ids for each age were integers, except "65+" which was string. So we had to make an exception for adding the population for this data.





Line Chart (not included in the final project):

Same as the stacked bar chart, we implemented this chart, by using different labels and colors for each category and computing the selected population. The implementation made us so many difficulties for us. At last we decided to visualize it manually, but that didn't work too.

In the process of implementing it, we realized that for each sub-category the population doesn't change that much, and it would be useless to visualize this chart on our website. So we decided not to spend any more time on this chart.

Evaluation:

The goal of the project was that if someone wants to know about the demographic structure of the US areas, the user can specify their selection and find out which areas are more appropriate for their purpose and also they can compare different states using multiple visualizations.

Our visualization tool works quite well, however there can be some improvements:

- Adding another bar chart based on the population instead of percentage
- Adding hover on options for the stacked-bar-chart
- Adding the line chart that was mentioned before

The questions in the proposal:

We are trying to show how many people with those selected attribute live in a specific area.

- Which is done by US map visualization.

We want to show the numerical difference of people in a certain category in each area.

- Which can be addressed by bubble chart and bar chart.

We want to show how the population of a selection changed over the years.

- Since the difference fluctuation between years is very nuance, we decided to remove the line chart, because it'd be almost straight line.