Crystal Wu (cw683), Lydia Kim (lmk225)
Prof. Jeff Rzeszotarski
INFO 4310 - Homework 2
Write Up

Dataset
Dataset Link:
https://vincentarelbundock.github.io/Rdatasets/doc/AER/CollegeDistance.html?fbclid=IwAR1HTJhGJYvptvhdKTdivxL9amW18Zz_R1XgIr-YISYZqa6otmcT5P2c6i4
Original Survey Paper:
https://nces.ed.gov/statprog/handbook/pdf/hsb.pdf

For this project, our group decided to create a visualization based on the dataset linked above, which provides information about high school students in the United States and how far they traveled for college in 1980. Despite our initial reservations about this dataset because of how outdated it was, we thought it would still be interesting to observe if there were any trends with high school students and how far they would travel for school.

The dataset currently includes information on 4,739 students provided by the Department of Education in 1980. While anonymous, it provides information about the students' personal information such as their:

- gender
- ethnicity (whether they are African American, Hispanic, or other)
- composite test score
- whether their mother or father graduated college
- family home ownership
- home urban-ness
- county unemployment rate
- state hourly wage
- college distance from home
- college tuition
- year
- family income
- region (whether they live on the west coast or not)

As there was a follow-up survey distributed in 1986, some entries of the dataset include students who are already enrolled in college and not high school seniors. When preprocessing the data, these entries were removed from the dataset because we wanted to focus on current high school students (in 1980) and examine how far they were willing to travel for college given their circumstances and background.

Design

When discussing the design of the visualization, the group explored the dataset to see if there were any relationships or information about the students worth pursuing or interesting. When looking at the dataset, the group recognized that the variables in the dataset could be organized into two categories- school and background. Variables in the school category included test score and college tuition because they revealed information about the school itself as they answered questions such as how expensive is the school and how difficult is it to get in? Other variables such as ethnicity, parent's highest level of education, and family home ownership were categorized into the background category because they revealed information about the student itself. The background category could be further classified by personal background (ethnicity, parent's college education) and location (urban-ness, state unemployment rate, wages, family income). The location sub-category included information about a student outside of his/her reach since they cannot control where they live and study.

After categorizing the variables into different categories, the group asked questions it hoped to answer such as:

- What kind of student goes farther off to college?
- If a student has a higher test score, are they more likely to go farther off to college?
- How does a student's background influence their college selection?

We found ourselves asking these questions because when discussing our own high school experiences, we discussed how we would assume a student with higher test scores might go to a college farther from home since they might have more opportunities. Additionally, a student with a lower family income or lower state wage might not go to a college too far from them given that family circumstances might prevent them from doing so. We wanted to create a visualization that would help us gain insights on the relationship between a student's information and college distance.

The group decided to create two scatter plots to examine the relationship between college distance and the two categories (school information and personal background) since we wanted to discern if these two had effects on college selection. Scatter plots seemed to be the most appropriate visualization since we wanted to examine the relationship between college distance and other factors as they both included numerical variables.

For the relationship between college distance and school, we considered comparing college distance with either test scores or tuition. We chose to examine the relationship between college distance and test scores because when considering both variables, it appeared that examining test scores would be more insightful. While a student may determine a school based on tuition, a test

score may be more indicative of potentially better schools. We wanted to see if students with higher test scores would potentially go farther to better schools. However, we decided to incorporate tuition into the scatter plot by making it the size of the radius of the circles. In this way, students can examine to see if there is a relationship between tuition and scores.

For the relationship between college distance and background, we decided to examine the relationship between college distance and home state wages. While we considered other variables, we believe that wages was most indicative of a student's background because wages are often dependent on the cost of living in an area.

For both visualizations, we decided to color the circles to represent gender since this is also something we considered exploring. We would choose the circles to be pink and blue for females and males, respectively, to determine if there is a trend between gender and college distance.

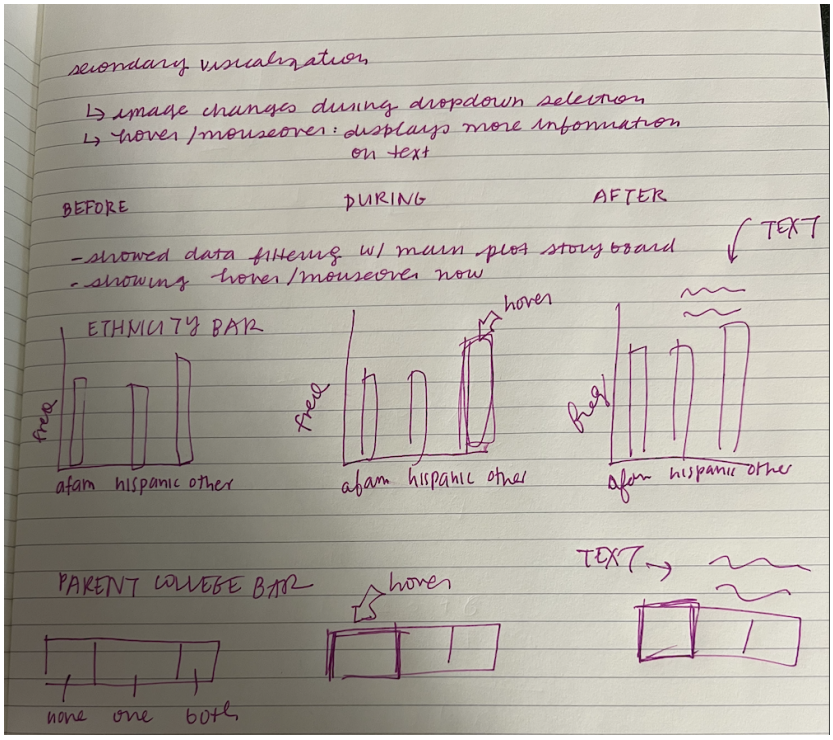
To make this visualization interactive, we considered creating buttons to filter the data based on income. We assumed that students with families of higher incomes would be more likely to afford a school that might be farther and more expensive. We considered a case where a wealthier student might have the means to attend a private and out-of-state school. The user would be able to select either a "high" or "low" button to toggle between the different incomes and observe trends. We also decided to create buttons for wages and scores (the x-axes of the scatter plots) so that if a user wanted to further examine these variables, they could do so.

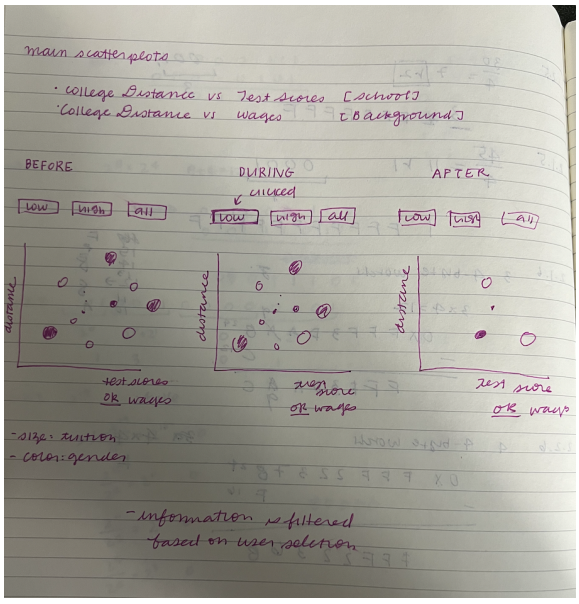
Furthermore, to provide more information about the students, we also considered using a lasso tool where the user could see details about the student's background based on their selection. We considered just listing information such as average test score, average wage, the distribution of parents going to college, and urbanness. However, upon further discussion, we determined that this would not be too helpful for the user because it could potentially be an information overload that overwhelms the user. Instead, we decided to display the information in a much more approachable way by creating dynamic text that indicates the number of students selected from the filters. The content of the text would be dependent on the filters selected by the user about scores, wages, and income. Below would include text about the percentage of underrepresented minorities and home ownership from the selected filters. An additional two plots were determined to be made that would change upon selection of the buttons to answer other questions we considered exploring initially:

- What is the relationship between students with a particular score, wage, or income and their parent's education status? Is there a correlation between higher scores or increased cost of living with a parent's education?
- What is the distribution of ethnicity for students with a particular score, wage, or income? Does such a relationship exist?

For the plot regarding a parent's education status, we considered making a stacked bar chart so that the user could see what percentage of the selected students had parents that went to college or not. We believed that seeing this in relation to the whole would be more easier and insightful for the user to observe trends. For the other plot regarding a student's ethnicity, we planned on creating a bar graph of each ethnicity provided and their frequencies. The frequency distribution would be dependent on the filters selected. Overall, we hoped to gain additional insights into whether these factors influence a student's decision on how far they would go to school.

Seen below is our storyboard on what we hope our visualization will look like before, during, and after the interactions:





Development and Final Visualization Implementation

To develop the visualization, the dataset was first preprocessed to adjust the types of the variables and exclude non-high school seniors, as reasoned above. Scatter plots to display the relationship between college distance and state wages and college distance and tuition were developed. Upon further consideration, we decided to change the scatter plot for the school by altering the x-axis from test scores to tuition. The reason why we altered this was because we believed that tuition would play a bigger role in a student's decision as people tend to go to schools that they can afford. Additionally, test scores are not indicative of the school itself but were more revealing about the student when this scatter plot aimed to focus on the school category. However, we decided to incorporate test scores as the radius of the circles to potentially see if there was a relationship between tuition and test score. While we initially considered only making the sizes of the radius dependent on test scores for the scatter plot with college distance and tuition, we added this feature to the other scatter plot for consistency. A legend was added to let the user know that the varying circle sizes corresponded to the varying test scores. For the scatter plot on the right, the variable for the x-axis was also altered from state wage to county unemployment rate. This was changed because the county unemployment rate gives more information about the the student's hometown specifically. Thus, we wanted to examine the relationship between college distance and county unemployment rate to see if a student from a location with more unemployment would travel farther to college.

For the scatter plots, we also contemplated making the sizes of the circles generally bigger for the scatter plots because we recognized that someone might consider them to be too small. However, as the scatter plot would be used to observe general trends and bigger circles could potentially lead to bigger overlaps and loss of data, the idea to keep the circle sizes smaller was maintained. Instead, we added a white border around the circles so that the user can easily see if there are overlaps within the circles.

A legend was created to differentiate and identify each gender. Another legend was created for the side plots where orange would represent the whole and purple would represent the percentage of the whole that the filter represents. Contrary to the original storyboard idea, our group decided to filter the visualizations based on county unemployment rate, tuition, and income instead of state wages, test scores, and income. This is because we wanted to keep the x-axis and filters consistent so that if the user wanted to further examine the relationship between the variables, they could do so. To filter these variables, it was decided to use a dropdown instead of a button because when creating the visualization, the presence of over five buttons that were for three different filters was very overwhelming and not a good design choice. A dropdown was determined to be easier for the user to select a single choice on what filter they want to see. Additionally, a default option was added to each dropdown so that at any point in time, the user could "reset" their choices and see the visualization with all entries. The dropdown options for

scores and wages were determined by manually computing the quantile values and using those as ranges of “low,” “medium,” and “high.” We considered how to separate the data into these categories and decided to organize them so that an equal amount of data went into each quantile.

Secondary visualizations were made for when a user selects a specific filter for income, county unemployment rate, and tuition in the dropdown filters. Dynamic text was included to provide a general overview of the filtered data as it describes the number of students that satisfy the filtered conditions and their average test score. Of the filtered dataset, the percentage of those who own homes and are of underrepresented minorities are also listed. A bar chart of the ethnicities and frequencies are displayed side by side. There are two colors for each ethnicity bar to represent the total number of students overall and the percentage that fulfills the conditions. This was included to show how much of the students represent the filtered conditions in comparison to the whole. To make the visualization more interactive, a mouseover feature was added so that when the user hovers over a particular ethnicity in the graph, information is given to the user about the percent difference of the filtered and non filtered data. For example, it might mention how students from low-income families might be 6% less Hispanic if low income is selected as one of the filters. We determined that it would be more interesting for the user to actually interact with the bar graph to gain additional insight into how ethnicity plays a role in a student’s background and their overall decision.

Another chart was created to further examine the parent’s college education level in relation to the filtered data. We initially considered creating a stacked bar chart to differentiate and identify the overall percentages of whether none, one, or both parents of a student attended college. This was our original plan because in a stacked bar chart, we can easily see trends of whether one group dominates the others as a whole. However, we decided to extend this idea by creating an additional stacked bar chart that represents the non-filtered data. We decided to have two horizontal bar charts so that the user can easily see how trends for a particular filter differs from the overall trend of the entire dataset. In addition, when a certain section of the horizontal bar chart is hovered over, text will be displayed to provide information to the user about the percent differences of the filtered student’s parents’ college education level compared to the overall dataset. The mouseover feature was also included instead of just listing this information when the page initially loads because the user can interact with the visualization and understand the information being provided.

When filters are selected such that no data points satisfy the conditions, the side plots disappear because there is no data to examine about ethnicity and parent college education distributions. However, the dynamic text about how no such students exist in the dataset are included to let the user know that the graphs didn’t disappear for no reason.

Final Visualization

For the final visualization, we kept most of the features described above where we have two scatter plots. The first scatter plot examines the relationship between tuition and college distance for high school seniors while the second scatter plot examines the relationship between college distance and county unemployment rate. For the scatter plot with tuition, the circle sizes are dependent on the test scores of the student. From these visualizations, we hope that the user can observe general trends about some aspects of a student's background and how it influences how far they are going to college. The circles on these scatter plots are differentiated by two colors of pink and blue that specify the student's gender. The colors and their corresponding values can be identified with the legend provided.

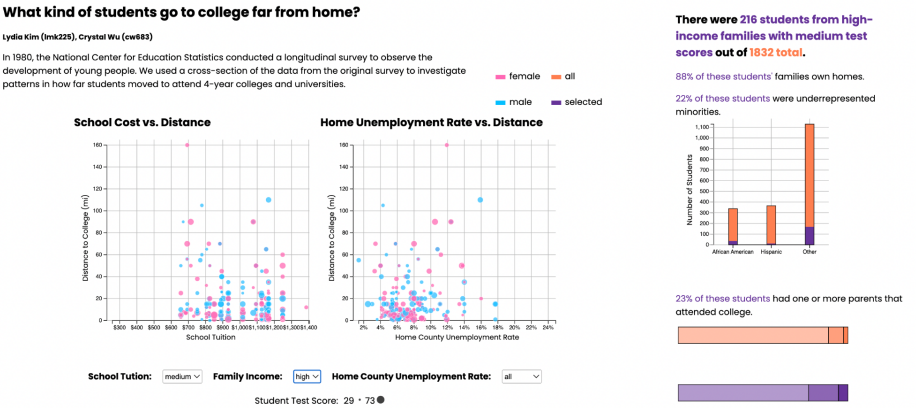
The user can select on any of the provided options in the three dropdown filters for income, county unemployment rate, and tuition to further examine trends and relationships. When a user selects one of the options, the dataset is filtered accordingly to display data that satisfies those conditions. The dataset is then used to update the plots.

On the side, there is additional text that provides information about the percentage of families that own homes and the percentage of underrepresented minorities. An additional graph displays the frequencies of ethnicities of the filtered dataset and when hovered over, more information about students of a certain ethnicity compared to the overall trend is described. Another graph uses two stacked bar charts to compare the overall students' parents' college education level. One of the stacked bar charts demonstrates the overall trend whereas the other one uses the filtered dataset from the options selected by the dropdown filters. A user can reset the options at any given time by selecting "all" in the dropdown.

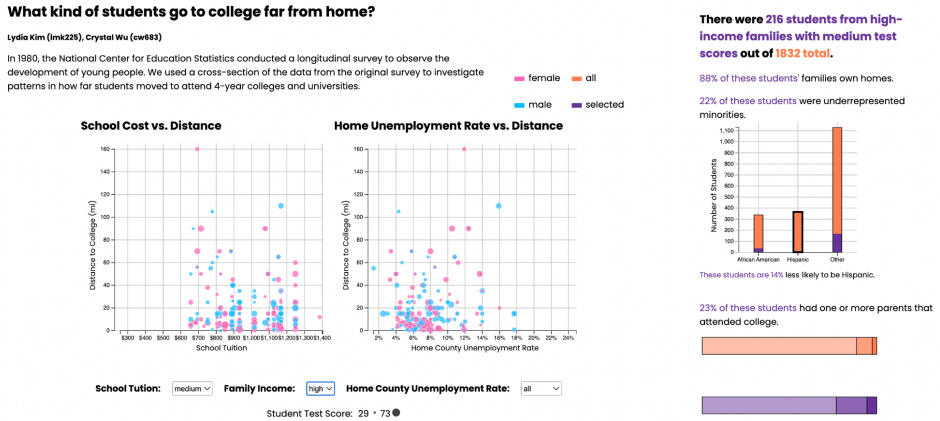
Legends were also provided to inform the users about how varying colors in the scatter plot represent gender, the sizes of the circles in the scatter plot represent varying test scores, and when filtered, two colors are used to represent the overall dataset and the filtered dataset.

Here is a screenshot of the final visualization:
When no selection is made in the dropdown filters:

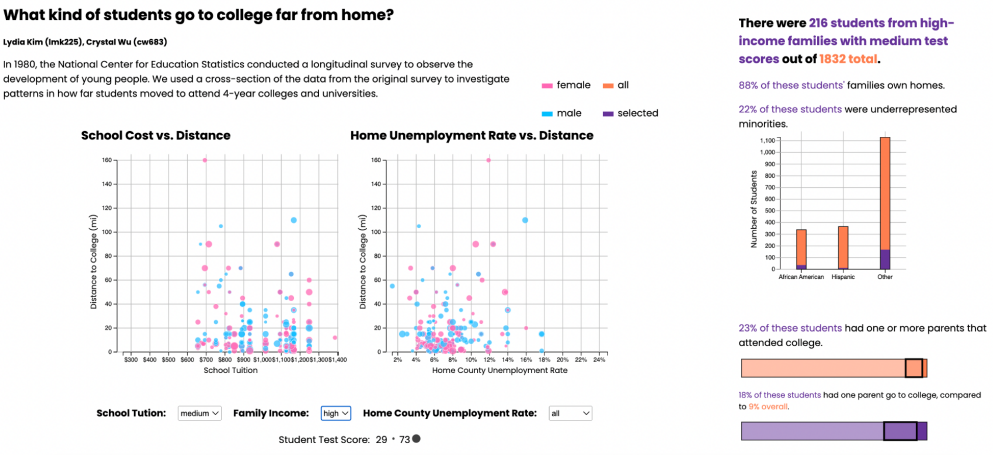
When a selection is made in the dropdown filters:



When the mouse is hovered over the ethnicity chart:



When the mouse is hovered over the stacked bar chart:



Contributions

During the duration of the project, Crystal and Lydia worked together on selecting the dataset, discussing potential visualization ideas, and designing the interactive visualization. Once the design of the visualization was determined, the work was divided so that Lydia would work on the main scatter plots and Crystal would work on the dynamic text and side plots that would display once the user selected specific filters. Interactivity of the visualization was worked on together as Lydia focused on creating the actual dropdown items and Crystal worked on the functions to call it when appropriate. Lydia wrote the majority of the report.

While most of the time was dedicated to designing the visualization and discussing what variables would be most insightful from the dataset, Crystal and Lydia roughly spent 12 hours developing the actual visualization. The parts of the project that took the most time for Lydia were adding axis labels to the scatter plots because the margins had to be adjusted and creating the dropdown filters dynamically. Once the dropdown filters were created from the information in the dataset, it took time to adjust the function and add a default value so that the user could choose to “reset” the visualization to its original state. The parts of the project that took the most time for Crystal were creating the dynamic text to change according to what filters were selected, and linking the hover interaction between the two parent education level visualizations.