**College Explorer Documentation**

Masaki Nawa

**Purpose**

The purpose of College Explorer is to have users from different backgrounds utilize it as a tool to explore different relationships and discover new trends between different variables using US college data from 2013-2014. The target users range from professors who are interested in using college data to investigate social inequalities to students who are learning statistics in their class and want to have some exposure to data visualization. Although the data set is outdated, it is also intended to help high school students discover colleges using specified criteria. The dashboard includes four types of visualizations: a map, scatterplot, regression, and table for visualizing the data set.

**Data description**

The data was collected from *“American University Data” IPEDS dataset* found in Kaggle, which contains information about US colleges. The original data was pulled from Integrated Postsecondary Education Data System (IPEDS) owned by the U.S. Department of Education’s National Center for Education Statistics (NCES). The NCES collects data and reports directly from colleges and universities that participate in federal student financial aid programs. The original dataset consists of 1,534 observations and 145 variables. The tidied data set consists of 1,534 variables and 27 variables (the process of tidying will be explained later). The following explains the variables in more detail:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Type** | **Description** | **Categorical/Numerical** | **Range** |
| Name | chr | Name of the college | Categorical | N/A |
| High.Degree | chr | Highest degree offered | Categorical (ordinal) | N/A |
| Longitude | num | Longitude location | Numerical | -158.062~-67.4565 |
| Latitude | num | Latitude location | Numerical | 19.70185~64.85705 |
| Applicants | int | Number of applicants | Numerical | 0~72676 |
| Admissions | int | Number of admissions offered | Numerical | 0~35815 |
| Enrolled | int | Number of enrollments in 2013 | Numerical | 0~10241 |
| SAT.Reading.25 | int | SAT reading 25th percentile score | Numerical (discrete) | 310~720 |
| SAT.Reading.75 | int | SAT reading 75th percentile score | Numerical (discrete) | 360~800 |
| SAT.Math.25 | int | SAT math 25th percentile score | Numerical (discrete) | 280~770 |
| SAT.Math.75 | int | SAT math 75th percentile score | Numerical (discrete) | 410~800 |
| ACT.25 | int | ACT composite 25th percentile score | Numerical | 13~33 |
| ACT.75 | int | ACT composite 75th percentile score | Numerical | 14~35 |
| Percent.Admitted | int | Acceptance rate | Numerical | 6~100 |
| Tuition | int | Tuition and fees | Numerical | 1032~49138 |
| Tuition.Instate | int | Total cost for in-state students living on campus | Numerical | 9768~64988 |
| Tuition.Outstate | int | Total cost for out-state students living on campus | Numerical | 9786~64988 |
| State | chr | State the institution is located | Categorical | N/A |
| Regional | chr | Region the institution is located | Categorical | N/A |
| Control | chr | Public/Private | Categorical | N/A |
| TEnrollment | int | Total student population (undergrad) | Numerical | 0~51333 |
| Percent.Women | int | Percent of female undergraduate | Numerical | 0~100 |
| Percent.Instate | int | Percent of in-state freshmen | Numerical | 0~100 |
| Percent.Outstate | int | Percent of out-state freshmen | Numerical | 0~100 |
| Percent.Foreign | int | Percent of international students | Numerical | 0~40 |
| GradRate.4yr | int | Graduation rate within 4 years | Numerical | 0~100 |
| Percent.Finan.Aid | int | Percent of freshmen receiving any financial aid | Numerical | 18~100 |

Note that there are limitations to this data set. For example, there is quite a lot of information missing for each variable. In addition, the data set is from 2013, which means that it is outdated, and information extracted from this data may not be applicable today. One such example is tuition price, which is expected to be different due to yearly inflation. Another example is the number of applicants, as many colleges are starting to provide a test-optional policy, which opens the gate to more students whose testing skills are not their strength. Additionally, people having more access to technology tools to gain more information today should have increased the number of applicants.

**Research Questions**

Through this project, I hoped to address two questions using the visualization tool created: (1) Do colleges that have a higher percentage of students coming from out of state have higher ACT scores (75th percentile scores)? (2) In which region do expensive colleges tend to concentrate in? To answer question (1), I decided to create a regression analysis using a scatterplot, having ACT 75th percentile score on the x-axis and percent of out of state students on the y-axis shown below.

Chart, scatter chart

Description automatically generated

Using the line of best fit, we see that the higher percentage of students from out of state, the more likely an institution has ACT scores in the 75th percentile. The scatterplot helps users understand the trend of the relationship between these two variables. One thing that could have been better was having a feature that allow users to see a different fitted model, which could have increased the R-squared value and provided a better visualization.

Map

Description automatically generated

Map

Description automatically generated

In answering (2) question, I created a map visualization along with a filtering feature that allows users to narrow down variables that they are interested in. Out of several ways to address this question, I decided to select the 20 most expensive colleges in the dataset. This was done by setting the tuition range from $46,432 to $49,138. The first image shows that there are more colleges on the east coast (specifically New England and the Mid East) that charge higher tuition fees compared to the rest of the U.S. However, it would have been easier to answer this question by creating a bar chart for a selected region and having two charts next to each other to make the comparison, perhaps using the average tuition in the region as the variables.

**Insights**

Chart, bubble chart

Description automatically generated

The map above plots all the colleges in the data set using the longitude and the latitude variables for each college. It shows us that colleges are focused on the east side of the country. When we zoom in, we get a better sense of the region with more colleges. The plot shows us that regions with most colleges are in New England and Mid East, followed by Great Lakes and Southeast.

Chart, scatter chart

Description automatically generated

The scatterplot above shows the relationship between the amount of tuition charged to students and the number of first-year students receiving any kind of financial aid in 2013. The plot shows us that schools that charge students more than $40,000 have lower rates of students receiving financial aid than colleges that charge between $20,000 and $35,000. This could be due to expensive colleges accepting more students who can afford such expenses and students who cannot afford to pay them deciding to attend less costly schools.

**Improvements**

As mentioned earlier, one thing that can be improved is creating bar charts that allows us to compare two or more regions by the average cost of attendance. In this way, I would have been able to obtain more precise numbers to directly answer my question. This can also be expanded to have users select different variables so that they can discover similarities and differences by regions. The feature could have more functionality if users were able to compare data by state as well. Another thing that could have made the visualization better was building a feature where users can either select or are given options to choose a polynomial regression or curve fitting line. This would be more helpful when users are conducting a regression analysis since most of the relationships generated from the scatterplot have a non-linear correlation and a line of best fit simply does not provide enough information for some analysis. Finally, implementing more features for the filtering option like narrowing institutions by SAT/ACT scores would make it more useful for high school students, as it would help high school students narrow down their college choices.

**Describing the process and development to make your work reproducible, for example *tidying data*.**

I started off by tidying my data and then moved onto coding using the cleaned data. For tidying the raw dataset, my initial step was to select 27 variables mentioned in the data description section above using the select() function. After selecting the relevant variables needed for the project, my next step was to rename them. Using the rename() function, I decided to rename the variables so that they were shorter and more readable, but also understandable. For example, I decided to rename Total.price.for.out.of.state.students.living.on.campus.2013.14 to Tuition.Outstate and Longitude.location.of.institution to Longitude. The final step was to get rid of unnecessary words from character values in the Region and the High.Degree variables. I started off by using the mutate() function and selected Region as the variable name and assigned it to the case\_when() function. Inside the function, I made sure to check which region value was selected before removing the state abbreviations from the string. I continued listing all the other region values so that I can get the same results. I took the same action to remove all the words that came after “doctor’s degree” in High.Degree. The code for these steps is shown below.

Text

Description automatically generated

Text, letter

Description automatically generated

The coding mostly relied on the documentation for each package that was used to create the map, scatterplot/regression, and table. The map used the leafplot package to create a more interactive environment and the filtering panel was created using the absolutePanel() command. Both the scatterplot and the regression were created using the plotly package, which consisted of a feature that allowed users to zoom in/out and one that showed information when they hovered over a data point. To generate the summary statistics of regression analysis, verbatimTextOutput() command was used to print the computed summary on the screen. Lastly, the table was created using the DT package, providing an organized table to help users visualization.

A picture containing map

Description automatically generated Chart, scatter chart

Description automatically generated Table

Description automatically generated

Map (left), scatterplot/regression (middle), table (right)

**Description of design decisions (encoding/mapping). Use the taxonomy of the what-why-how analysis framework presented by Tamara Munzner.**

* Map

Map

Description automatically generated with low confidence

* + What is shown?  
    The map shows a spatial field since it plots the values using longitude and latitude variables. It is also a categorical dataset because the college is plotted based on their location. Additionally, it seems static at first glance since the map does not appear to change, but when we click on the circles or zoom in/out, we get new information such as where exactly the college is located on the map; hence it is also dynamic.
  + Why is the user looking at it?  
    The user is looking at the map because they want to discover the location/region of a college or identify types of colleges that they might be interested in attending based on their selected criteria. It also helps users search colleges whether the target and the location are known or unknown, helping them narrow down their college choices no matter where they are in the college search process.
  + How is it shown?  
    The map is shown by using given spatial data such as longitude and latitude. It also uses color to show the number of colleges that are clustered in the area that is currently shown on the screen. It also allows the user to reduce the information shown on the screen by filtering out some of the variables.
* Table in the map

Graphical user interface, application

Description automatically generated

* + What is shown?  
    This table takes different variables as rows and subsequent information as values. It consists of categorical variables such as name of the college, control of the institution, state and region it is in, an ordinal variable such as the highest degree that is offered from the institution, and a quantitative variable such as total undergraduate enrollment. In addition, it is both static and dynamic, because it seems like the table will not change, but once we click on something else, the table disappears.
  + Why is the user looking at it?  
    The user is looking at the table in the map because they want to discover the name of the college that they received after filtering their desired values. It gives a great summary about the college that they found since the pop-up displays name of the college, highest degree offered, control of the institution, total number of undergraduate enrollments, and state/region where it is located. The visualization stimulates and satisfies users, providing convenience.
  + How is it shown?  
    The table is aligned to the left and uses a different font size to distinguish the difference between the name of the variable and its value. The information presented is selected to show the most relevant details that a user expects to see during their college search.
* Scatterplot with name of the college on each point

Chart, scatter chart

Description automatically generated

* + What is shown?  
    This scatterplot shows a spatial field since it plots the attribute values specified on the x and y-axes. It deals with quantitative variables to display the relationship between two selected variables. Additionally, it seems static at first since the scatterplot does not seem to change, but when we hover over the points, it displays x and y points as well as the name of the college. Thus, it is also dynamic.
  + Why is the user looking at it?  
    The user is looking at this table because they want to discover trends, outliers, distribution, and connections found between the two variables selected, as well as identify which college lies on which point.
  + How is it shown?  
    The scatterplot separates each of the points and distributes them so that the user can see how scattered they are. Users can manipulate the plot by changing the variables they hope to observe. The scatterplot also allows a user to select a point that they want to explore and outputs the name of the college in addition to the x and y values.
* Regression

Chart, scatter chart

Description automatically generated

* + What is shown?  
    This regression shows both a spatial field and a table, as the bottom shows a statistical summary as columns and rows. It deals with quantitative variables to display the relationship between the two selected variables and the output of the statistical summary. Additionally, it appears static since the scatterplot does not seem to change, but when we hover over the points, it displays x and y variables. Thus, it is also dynamic.
  + Why is the user looking at it?  
    The user is looking at this regression because they want to discover trends, outliers, distributions, dependencies, and correlations between the two variables selected, and also compare and contrast how well the line fits into the plot.
  + How is it shown?  
    The regression separates each of the points and distributes them so the user can see how scattered they are. A line with a different color is used to show the line of best fit in the model. The summary statistics shown below the plot is left aligned.
* Table

Table

Description automatically generated

* + What is shown?  
    This table takes names of colleges as rows and the variables users are interested in exploring as columns. It contains categorical, ordered, and quantitative datasets because users can rank each of the variables numerically or alphabetically. Additionally, it is static because nothing on the screen changes.
  + Why is the user looking at it?  
    The user is looking at this table because it allows them to both identify values that they are interested in investigating for each of the colleges and compare and see the overview of the whole dataset. Users can also compare colleges by sorting by one of the variables.
  + How is it shown?  
    The table uses white and gray to distinguish each row, making it easier to visualize. Each of the values is aligned in a specific way and is ordered alphabetically by state. The selection menu and the search bar allows the user to filter information either by variables or keywords.

**Sources**

A. Massey, “The college checkpoint,” *The Court Street Press*, 18-Dec-2021. [Online]. Available: https://cbwestcourtstreetpress.com/2218/bucks-news/the-college-checkpoint/. [Accessed: 13-Mar-2022].

“About IPEDS,” *The Integrated Postsecondary Education Data System*. [Online]. Available: https://nces.ed.gov/ipeds/about-ipeds#:~:text=types%20of%20programs.-,Institutional%20Prices,campus%20or%20off%2Dcampus). [Accessed: 13-Mar-2022].

G. Pilgrimname, “NCAA Swimming Team Finder for incoming college athletes,” *Shiny*. [Online]. Available: https://shiny.rstudio.com/gallery/ncaa-swim-team-finder.html. [Accessed: 13-Mar-2022].

J. Allen, “Basic datatable,” *Shiny*. [Online]. Available: https://shiny.rstudio.com/gallery/basic-datatable.html. [Accessed: 13-Mar-2022].

J. Cheng, “SuperZip example,” *Shiny*. [Online]. Available: https://shiny.rstudio.com/gallery/superzip-example.html. [Accessed: 13-Mar-2022].

“Leaflet for R,” *Leaflet for R - markers*. [Online]. Available: https://rstudio.github.io/leaflet/markers.html. [Accessed: 13-Mar-2022].

“Lesson 1 Welcome to shiny,” *Shiny*. [Online]. Available: https://shiny.rstudio.com/tutorial/written-tutorial/lesson1/. [Accessed: 13-Mar-2022].

“Plotly,” *Plotly r graphing library with R*. [Online]. Available: https://plotly.com/r/. [Accessed: 13-Mar-2022].

S. Bhongale, “‘American University Data’ IPEDS dataset,” *Kaggle*, 07-Oct-2017. [Online]. Available: https://www.kaggle.com/sumithbhongale/american-university-data-ipeds-dataset/version/1. [Accessed: 13-Mar-2022].

S. Martelo, “Welcome to the CovidBot Explorer (CBE)!,” *CBE*, 2021. [Online]. Available: https://martelos.shinyapps.io/Round\_2\_App/. [Accessed: 13-Mar-2022].

“Shiny Themes,” *Shiny themes*. [Online]. Available: https://rstudio.github.io/shinythemes/. [Accessed: 13-Mar-2022].

V. Perrier, F. Meyer, and D. Granjon, “Custom inputs widgets for shiny,” *Custom Inputs Widgets for Shiny*. [Online]. Available: https://dreamrs.github.io/shinyWidgets/index.html. [Accessed: 13-Mar-2022].

W. Chang, “Movie explorer,” *Shiny*. [Online]. Available: https://shiny.rstudio.com/gallery/movie-explorer.html. [Accessed: 13-Mar-2022].

W. Chang, “Shiny theme selector,” *Shiny*. [Online]. Available: https://shiny.rstudio.com/gallery/shiny-theme-selector.html. [Accessed: 13-Mar-2022].