Group members: yvs4 and jma353
CS3300 Project 2 Report

A. A description of the data. Report where you got the data. Describe the variables. If you had to reformat the data or filter it in any way, provide enough details that someone could repeat your results. If you combined multiple datasets, specify how you integrated them. Mention any additional data that you used, such as shape files for maps. Editing is important! You are not required to use every part of the dataset. Selectively choosing a subset can improve usability. Describe any criteria you used for data selection. (10 pts)

We gathered the following data about a subset of U.S. universities:
Early median salary
Mid-career median salary
Rank
Zip code
Sat reading low
Sat reading high
Sat math high
Sat math low
Number of applicants
Acceptance rate
Latitude
Longitude
State
And JSON file of U.S. states for the map

From the following sources:
http://www.payscale.com/college-salary-report/bachelors?page=1 // Salaries
http://www.collegesimply.com/guides/1600-on-the-sat // SAT + admissions
https://developers.google.com/maps/documentation/geocoding/intro#Geocoding // Geocoding
http://www.collegesimply.com/guides/low-acceptance-rate/?view=all //Not currently used in code
https://www.powerscore.com/sat/help/average_test_scores.cfm // Not currently used in code

By web scraping (scripts provided with project) and putting all data into JSON format.

In the project itself we did not use all university information, but only the following:
Mid-career median salary
Sat reading low
Sat reading high
Sat math high
Sat math low
Acceptance rate
Latitude

Longitude
State
We combined Sat reading low, Sat reading high, Sat math high, Sat math low by averaging those values to produce a single average Sat per university. Thus, we used latitude, longitude, and state for coding up the interactive visualization, and used 3 variables: SAT score, acceptance rate, and mid-career median salary as basis for the clustering algorithm. We decided to limit to only 3 variables to not overwhelm the user with information(initially we thought 5, but then decided 3 is enough to not be overwhelming while still being informative and not too simple either). Our initial idea was to do something that compares university statistics with statistics of their respective alumni. So we chose SAT and acceptance rate as metrics for the universities(how hard they are to get into) and Mid-career median salary as metric for graduates showing potentially how much value graduates of 1 universities gained compared to another university. So the idea is to explore how selectiveness of university and how much graduates are making relate to each other. But those are not the only things: for more information, we chose to display where each university is located on a map. We also have clustering to show which universities are more similar to each other and which are more different. And then at the bottom we have linear regression plots showing how the 3 variables we used for clustering are related.

B. A description of the mapping from data to visual elements. Describe the scales you used, such as position, color, or shape. Mention any transformations you performed, such as log scales. (10 pts)

First, we made the map and clustering graph next to each other because they are connected by interactive features that are triggered on hover over any circle(representing a university or cluster centroid). The color scheme we picked solely on the basis of good visibility of elements, so color does not reflect any information, except that it is used to highlight which states are selected, and also which universities are selected. Also, on hover, the selected university circle grows bigger to differentiate it from the unselected ones.
For the map, we used an albersUsa projection. For the clustering graph and the linear regression graphs, we used D3 linear scales.
So basically, we plot each university on the map as a circle. Next to the map, we also have a graph with each university as a circle plus lines to the centroid of the cluster it belongs to based on the scales chosen by user for the graph. The, at the bottom, we have linear regression graphs that show the relation between the variables the user is allowed to use for clustering. These graphs contain each university as a circle, and a linear regression line through the circles(universities). The one place where we used color shading to indicate numerical difference was in linear regression graphs: higher shading means more distance from the line.

C. The story. What does your visualization tell us? What was surprising about it? (5 pts)

Our visualization allows the user to see the differences and similarities between US universities in different or same state based on the 3 variables: SAT score, acceptance rate, and mid-career median salary(this is the map and clustering). It also shows the general trends for universities when comparing 2 of the 3 variables(this is the regression graphs). In general, it is not surprising to find a positive linear relationship between acceptance rate and SAT score, or the positive linear relationship between salary and any of those 2 variables, since usually the more selective universities yield higher-paid graduates. What is of interest though, is the amount of variation in one variable of universities that are close to each other in the other variable. One of the most extreme examples of this would be College of the Ozarks, which has acceptance rate of only 9%(less than Cornell, close to Stanford University), but SAT of only 1034 and salary $71800 (substantially less than Stanford's 1475 and $123000).