Statistics 159 Final Project Proposal

Professor Gaston Sanchez

Siyu Chen, Yukun He, Aoyi Shan, Shuotong Wu

November 10, 2016

Based on the dataset available on College Scorecard, we are interested in exploring the differences between similar schools in terms of demographics and programs offered and how those differences contribute to school competitiveness. The client of our consulting project is a group of school administrators, and our data analysis is aimed to provide them with suggestions regarding what factors should they focus on and how to maximize their returns on investment in order to improve their school competitiveness.

Analysis Workflow and Components

We plan to divide schools into different clusters based on three criteria: West coast vs. east coast, located near population center vs. countryside and minority ratio. Then, after dividing institutions into different subgroups, we use the number of students applied during a specific year as the criteria to determine each school's competitiveness. To compare whether the difference between clusters are significant, we apply ANOVA analysis to see whether the mean for number of students applied is different for each cluster. Then, in order to provide concrete suggestions for improving competitiveness, we plan to run a regression of number of students applied on number of degrees offered, completion rate, average cost of attendance and average earnings after school. If the coefficient of a certain independent variable is significant, then we should consider that factor in order to make the school more attractive to students, and the larger the coefficient, more efforts should be dedicated in that area.

Data Cleaning and Exploratory Data Analysis

We will start our project with data cleaning and processing. Since College Scorecard provides us with data organized in the way that facilitates their analysis, our first step is to select variables that we need for our project, reformat the data and generate new variables that fit our study. Variables we plan to use are the following:

- CITY, ZIP: Each institution's location and zip code
- PREDDEG: The type of degree that the institution primarily awards
- Programs Offered by Type
- ADM_RATE_ALL: Admission rate for all branches

- UGDS: Number of undergraduate students for fall enrollment
- COSTT4_A: Average cost of attendance for academic year institutions
- UGDS_BLACK, UGDS_HISP, UGDS_ASAIN, UGDS_AIAN, UGDS_NHPI: Undergraduate student body by race
- C[100 or 150]_4: The completion rates for first-time, full-time students who begin school in the fall semester and complete within 100 or 150 percent of the expected time to completion for 4-year institutions
- MD_EARN_WNE_P*: Mean and median earnings

Then, we proceed to exploratory data analysis (EDA) in order to better understand the data. We want to obtain descriptive statistics and summaries of all the related variables. For the quantitative variables, we will compute the Minimum, Maximum, Range, Median, IQR, Mean and SD for each variable, together with their histograms and boxplots. For the qualitative variables such as ethnicity and location, we will compute a table of frequencies and create barcharts. Since we are interested in exploring the relationship between students applied and predictors that might affect school competitiveness, we will also obtain matrix of correlations, scatterplot of matrix and conditional boxplot.

For data that are not directly provided by the College Scorecard website, we perform additional computation. In order to determine whether a certain school is located on the east or west coast, we will introduce a new data set listing all the zip code representing different regions and create additional dummy variables representing location. In addition, we will also use a set of major cities in the U.S. and by comparing each institution's location with the list, we can decide whether it belongs to city or countryside. Then, since we don't have total number of students applied for each school, we take the number of total enrolled student divided by the admission rate. Although this is not a perfect measure since we expect some admitted students decline their offer, but this is the best substitute we have given the existing data. Lastly, minority ratio is one of our criteria to cluster schools, and we decide to use ratio obtained through dividing number of students who are minorities by the total student population.

Preliminary Division of Work

Aoyi: Data cleaning, processing and writing report

Siyu: Perform regression to determine main factors that impact school competitiveness and generate regression plots

Diana: Exploratory Data Analysis and create Shiny application

Shuotong: Clustering, ANOVA test and create slides

File Structure

```
stat159-fall2016-project2/
    .gitignore
    README.md
    LICENSE
    Makefile
    session-info.txt
    session.sh
    code/
      README.md
      functions/
        ...
      scripts/
        ...
      tests/
        ...
    data/
      README.md
      data-sets/
        ...
      eda-outputs/
          ...
      regression-data/
          ...
    images/
      README.md
      boxplot/
        ...
      conditional-boxplot/
        ...
      histogram/
        ...
      scatterplot-matrix/
          ...
      regression-plot/
          ...
      slides/
        slides.html
        slides.Rmd
    report/
      proposal.pdf
      report.Rmd
      report.pdf
      sections/
        ...
    shinyapp/
```