**College Data Visualizer**

Helen Zhang and Michelle Lin

*CS 3265*

**I. Introduction**

*Overview*

College Data Visualizer is an analytical web application developed with flask and bootstrap for the front-end and MySQL as the back-end data source. The application is intended to give users the ability to filter through colleges based certain attributes like SAT scores and location.

*Application*

We choose to build a web application because, out of all the choices, it was the most suitable given our project experience and language knowledge. Moreover, decided to build a filtering application with the college scorecard dataset because there aren’t many resources for students to be able to preview aggregate data during the college search besides individually searching up college websites or filtering by only SAT or act. Our hope is that the application will allow students to not only gain a helpful snapshot of each college, but also the ability to filter based on certain attribute with an intuitive interface.

*Dataset*

The application data is derived from a modified version of the most recently updated (March 30, 2020) College Scorecard distributed by the U.S. Department of Education containing information about all registered US Colleges from tuition case, undergraduate enrollment size to institution website links from the years 1995 - 2018. The year 2012 and 2016 were excluded for database insertion reasons to be discussed later in the report. This dataset was chosen because three reasons. One, the raw data is substantial at roughly 250 MB. The dataset consisted of at least a couple hundred attributes allowing us flexibility to pick and choose those that we felt were relevant and would provide a helpful overview of each college. Three, we thought this data would be helpful for current high school students who are either researching colleges or making college decision during these times where most colleges have closed off their campus and visits have been cancelled.

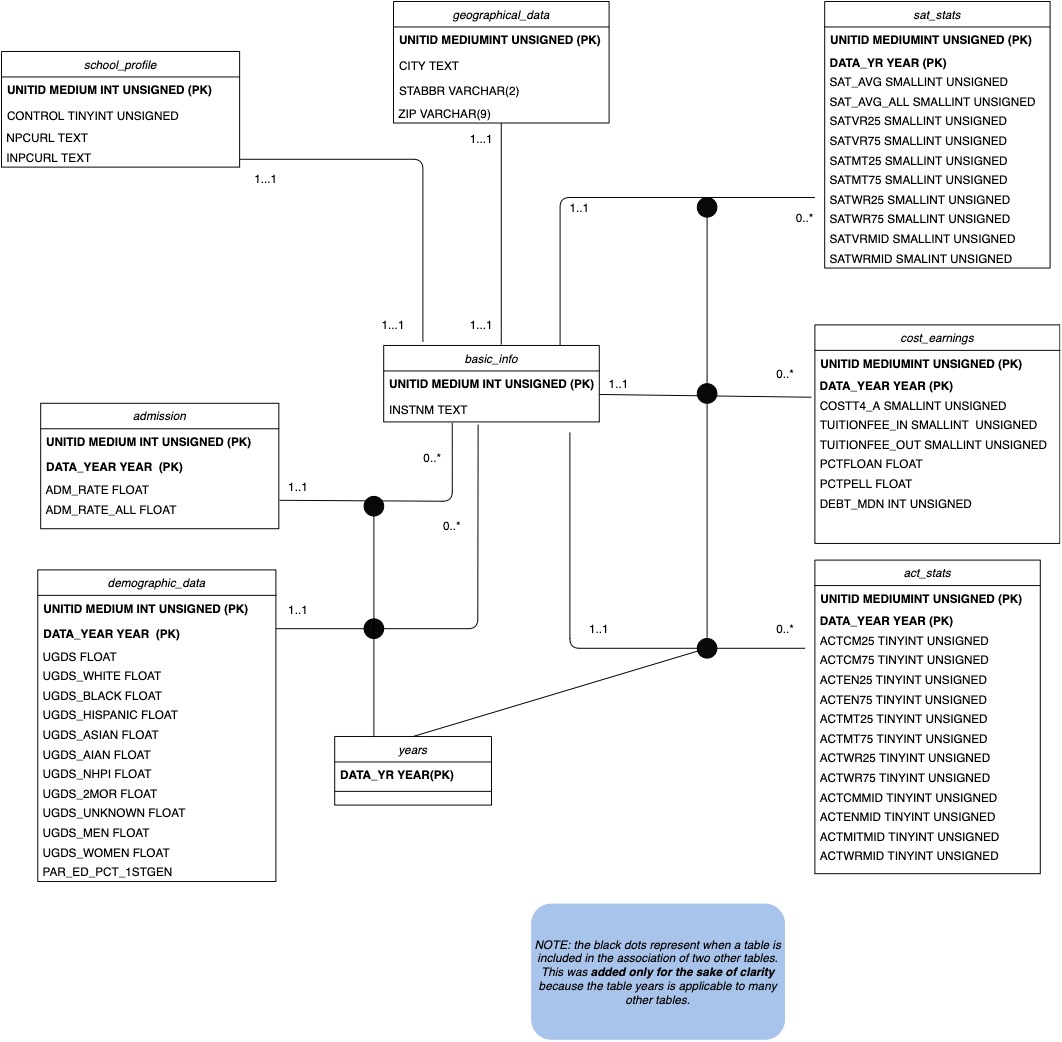
**II. Final Implementation**

*Final Database Design*

1. Normalization Process

2. Final Design

3. UML with cardinality

**

*Description of Data*

The raw data is disturbed and maintained by the U.S. Department of Education. The data set can be found at this link: <https://collegescorecard.ed.gov/data/>. Since the original data contained a significant amount of attributes (up to hundreds), the data files, each describing college data based on year, was cleaned to produce 20 individual files with only application relevant attributes using python pandas. Each file consisted of approximately 7200 rows of information with each row concerning only one college. The data used for testing consisted only of the 2018 data which had 7243 rows or colleges. This data was used to determine the insertion process from education\_mega (mega table with no normalization) to the normalized tables.

*Summary of Implemented Use Cases*

**III. Illustration of Functionality**

**IV. Summary Discussion**

*Challenges*

*Division of Work*