Sanchari Chowdhuri
sanchari@umd.edu

REPORT: ASSIGNMENT- 2

The college score card website is very detailed and interactive dashboard providing aggregated insight as well as giving users a detailed analysis based on segregation of the data. The Data set on the other hand is very detailed at a granular level and in this assignment in some questions aggregation of those data points lead to deriving collective results.

The dataset contains 1744 columns and 7703 rows of data. Each row corresponds to a particular educational institute and each column tells about a specific aspect of that university broadly pertaining to *academics, admission, aid, completion, cost, earning, repayment and school*. The data Dictionary is very useful in understanding which attributes to select in order to derive a particular insight. In this assignment, I have selected required attributes and merged them in new data frame for every question.

The rationale behind the ranking algorithm approach is that since each institute's score is a weighted average of 3 factors namely income after 10 years (earning), net price (cost) and graduated in 6 years (completion) , I decided to obtain a score based on adding the product of given weights with each of the attribute and ranking each institute's score. "***Income 10 years after entry*** "is the first derived column obtained from ***average*** of *earnings by males after 10 years of graduation* and *earnings by females after 10 years of graduation.* Since only undergraduate institutes are considered I have removed institutes where predominant degree (PREDDEG) offered is masters. "***Net Price***" is another derived column telling about the cost of attending the institute. This column is **addition** of *Net price of attending public university* (NPT4_PUB) and *Net price of attending private university (*NPT4_PRIV).Each institute will either have NPT4_PUB value or NPT4_PRIV value since a university will be either private or public .For every institute either of NPT4_PRIV or NPT4_PUB will be zero while the other will have a value hence addition of these two columns gives net cost. "***Graduated in 6 years***" is the last derived column obtained by addition of completion *rate of institutes with 4 years of course work* (C150_4) and *completion rate of institutes with less than 4 years of course work* (C150_L4) like associate degrees etc. Institutes which didn't have data related to net price or graduation rate in 6 years have been removed from the ranking. The entire list is sorted in descending order .Higher the score higher the rank in the list.

Regarding data driven insight, I decided on finding relationship between an institute's location and tuition fees (in state and out of state) .The findings were conclusive with the basic ideology that more

urbanized locality leads to higher fees. Institutes located in rural areas are more budget friendly as compared to those located in cities. Histograms depicting the insights are provided in the code.

This type of data can be very useful in designing interactive reports wherein students can put required parameter (or filters) and shortlist universities based on academics, degree offered, scholarship or loans etc. Also such dashboards can be very useful by lawmakers to understand trends and also to find similarities in pattern.

***