Final Project Proposal

1) Team members

name	BU email
Tegveer Ghura	tegveerg@bu.edu
Rahul Thairani	thairani@bu.edu
Jayesh Jain	jayesh11@bu.edu

2) Description of dataset

The dataset used is Post-School Earnings Data, which comprises 2018 cross-sectional data on every degree-granting institution in the United States - trimmed to include only certain variables of interest and only institutions which contain no missing values for these attributes of interest. The dataset is obtained from the College Scorecard, a public database of institution characteristics.[1] The attributes of interest include:

- *median_hh_inc* : Median household income (numeric input)
- *avgfacsal* : Average faculty salary (numeric input)
- *ugds_black* : Total share of enrolment of undergraduate-degree seeking students who are black. (numeric input)
- *ugds_white* : Total share of enrolment of undergraduate-degree seeking students who are white. (numeric input)
- *sat_avg* : Average SAT equivalent score out of 1600 of students admitted (numeric input)
- *instnm* : Institution name (nominal)
- *unitid* : Unit ID of institution (nominal)
- *mn_earn_wne_p10* : Mean earnings of students working and not enrolled 10 years after entry (numeric output)

3) Type(s) of data mining (Note: at least one of the types must be classification learning or numeric estimation)

The type of data mining we intend to perform is numeric estimation as we intend to predict the value of a single output attribute which is numeric.

[1] https://collegescorecard.ed.gov/data/

4) Goal of the resulting model

We are hoping to use data mining using numeric estimation to build a model that will allow us to determine an institution's output attribute of *mn_earn_wne_p10* (mean earnings of students working and not enrolled 10 years after entry) based on all other numeric attributes of the institution mentioned above.

5) Description of transformations to your data

- Trimmed the original Post-School Earnings dataset of the College Scorecard to include only 8 attributes of interest for use in the Numeric Estimation model.
- Trimmed the resulting dataset using a Python program to include only institutions which contain no missing values for the attribute of interest stated above, resulting in a dataset of 1256 institutions.
- The resulting dataset was trimmed to 1056 institutions, with the remaining 200 institutions being used as test cases.

6) Plan for other required components

- We will use the Weka data mining software in order to run the data mining technique of numeric estimation on our dataset. Consequently, we will exhibit a clear and compelling presentation of the findings we obtain from the above mentioned data mining technique.
- We will use Google Charts for the creation of data graphics in accordance with Tufte's principles. These graphics will concisely present the findings from our numerical estimation model.
- We will integrate SQL into our final project through the creation of relational tables that avoid redundancy and capture any constraints present in our dataset. Consequently, we will perform SQL queries on these relational tables in order to obtain summary statistics to support our predictions.
- We have created a Python program that manipulates our original Post-School Earnings dataset of the College Scorecard to include only 8 attributes of interest and only 1056 institutions with no missing values for these attributes, for use in our numeric estimation model.

7) Division of work among team members

- The process of formulating the Python program used to transform the original dataset and writing the conclusion shall be done by Jayesh Jain.
- The process of obtaining the dataset and writing the dataset description shall be done by Rahul Thairani.
- The process of writing the introduction and appendix and proofreading the final project document shall be done by Tegveer Ghura.

In addition, all members shall carry out the data analysis and reporting of the results together.