**Data wrangling**

Two datasets were imported from the website (<https://collegescorecard.ed.gov/data/>) using their urls:

1. Most-Recent-Cohorts-All-Data-Elements.csv - This file contains all the data needed for developing the project.
2. CollegeScorecardDataDictionary.xlsx – This file consists of details about the variables in the above csv file.

The steps taken to clean the data are as follows:

***Generating a dataframe:*** The csv data was downloaded from the url stated above and was converted into a pandas dataframe (pd.read\_csv). It was noticed that null values in several columns were represented as ‘PrivacySuppressed’. These values were converted into “na” values.

***Checking dataframe:*** The dataframe consisting of relevant categories was checked using commands - dataframe.head(), dataframe.info(), and dataframe.describe(). The dataset consists of 7593 rows and 1805 columns.

***Selecting relevant categories***: The data in the main csv file can be divided into 10 subcategories (e.g. school, admissions, academics, student, cost). We used the data dictionary to get the unique categories. Out of these we found 5 categories that were relevant for our analysis. For instance, we discarded the categories that focused on demographics of the student’s ethnicity, gender etc. After dropping irrelevant columns, we were left with 159 columns.

***Transforming Zipcode into categorical data:*** In order to divide the colleges into different regions, their zipcodes were used. We generated a new column showing the region of the college using the first digit obtained from the zip code. (df[‘Zip’.map(lambda x: int(str(x)[:1])).

***Treating NULL values:*** Percentage of NULL values in each column were calculated (df.isnull().sum() / df.shape[0] \*100). It was noticed that several columns had more than 40% of NULL values. So, we removed all the columns that had more than 40% of NULL values. This technique reduced the number of columns to 65. Further to treat the NULL values in the rest of columns we used their means (df.fillna (df.mean(), inplace=True)

It was noticed that 38 of the 65 columns represent the percentage of students in a certain program such as agriculture, computer, engineering including others. Some of the programs are not that common, resulting in several zeroes. The columns with more than 90% of zeroes were removed. Finally, we had 28 columns.

***RandomForestRegressor:*** RandomForestRegressor was imported from sklearn.ensemble and SelectFromModel from sklearn.feature\_selection to reduce the number of variables. After fitting the model we had 11 variables consisting of continuous and categorical variables.

***Normalization and Outliers:*** Income was normalized between 0 and 1. The boxplots were plotted for variables to identify outliers.