Data Analysis of College Scorecard

Adarsh Srinivas; Isha Kaur Ratti; Jay Sheth

University of Maryland, College Park

Introduction

The College Scorecard initiative has been designed to increase the power in the hands of the public to differentiate colleges and get an idea about how well these colleges are preparing their students in the path to success helping them make better informed decisions about selecting a college. It was developed to address the lack of information about university costs, quality, earnings etc. There is an increasing need for more information and the need of an extensive dataset which accounts for such variations in educational opportunities and hence, College Scorecard comes as a welcome initiative. The map below represents 1922 universities in the United States taken into consideration for our analyses.
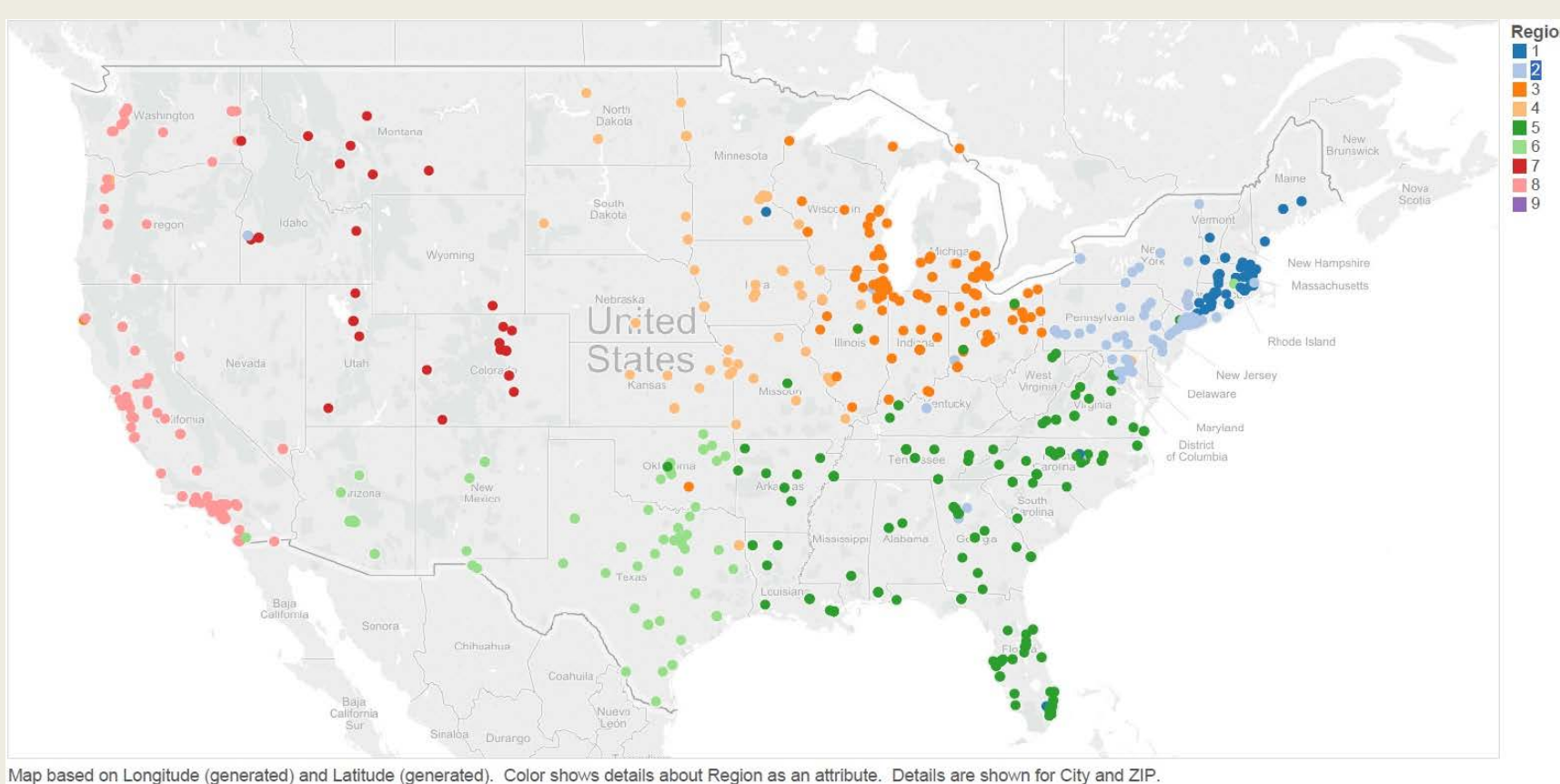

Map based on Longitude (generated) and Latitude (generated). Color shows details about Region as an attribute. Details are shown for City and ZIP.

Figure 1.

Research Question

In this study, we have analyzed the factors that according to us can best predict a university's success. These factors can be influenced by multiple other variables and are thus considered as proxies for success. The following factors are considered as dependent variables:

- **Post-graduation earnings**
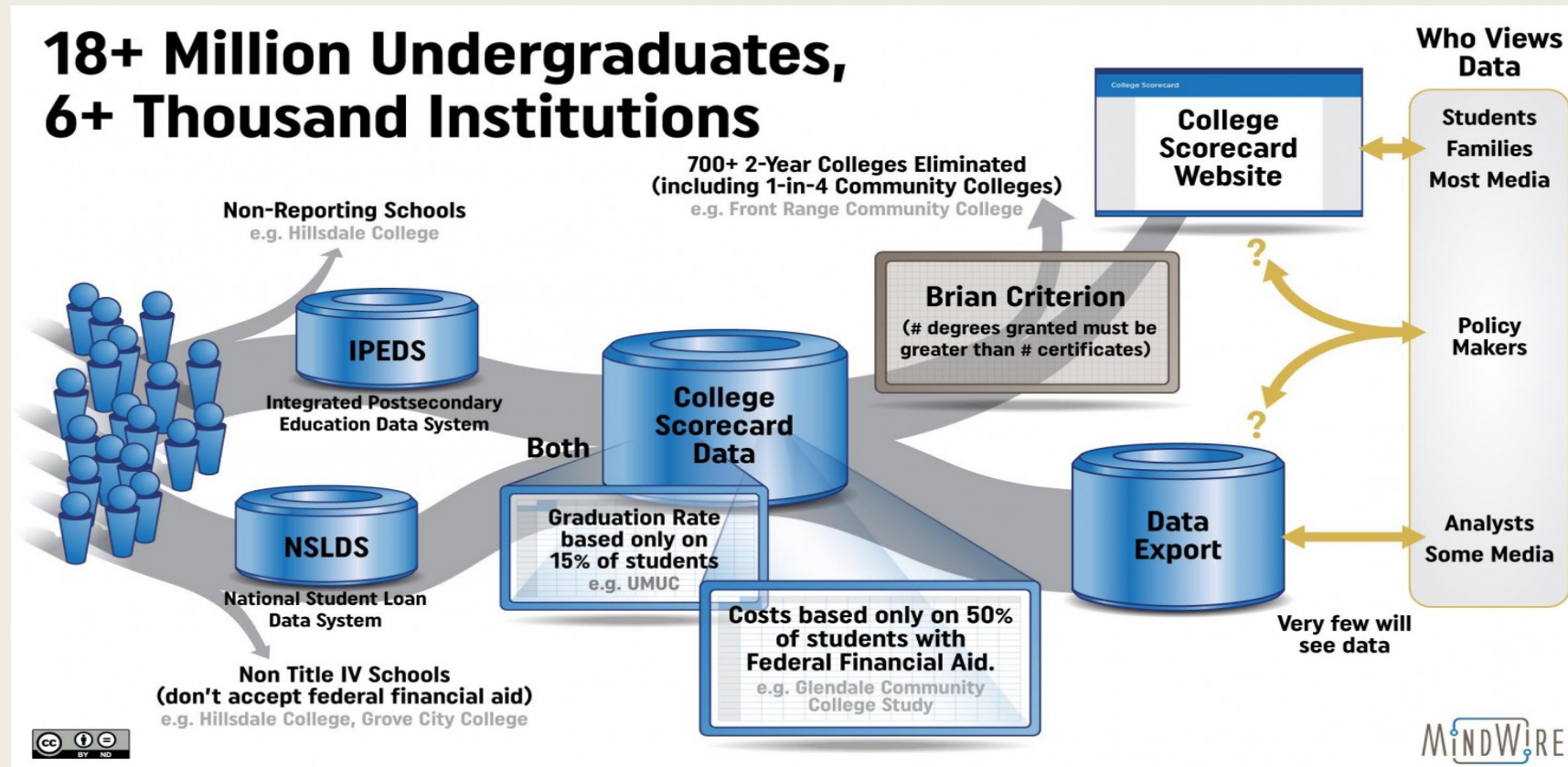- **Federal Aid**
- **Completion Rate**
- **Admission Rate**

Figure 2.

Method

- The data was collected by United States Department of Education as depicted in Fig 2.
- The following variables are used for predicting the dependent variables:
  - SAT_AVG_ALL: Average SAT score of enrolled students
  - CONTROL: Type of university i.e. Public or Private
  - COSTT4_A: Average Cost of Attendance
  - UGDS_WHITE,UGDS_BLACK,UGDS_HISP,UGDS_ASIAN, UGDS_AIAN, UGDS_NHPI: Categories of Race
- **Data cleaning/processing:**
  - Cases were selected where universities provided a 4-year Bachelor's degree or higher.
  - Variables having more than 5% missing values were taken care of by replacing them with series mean
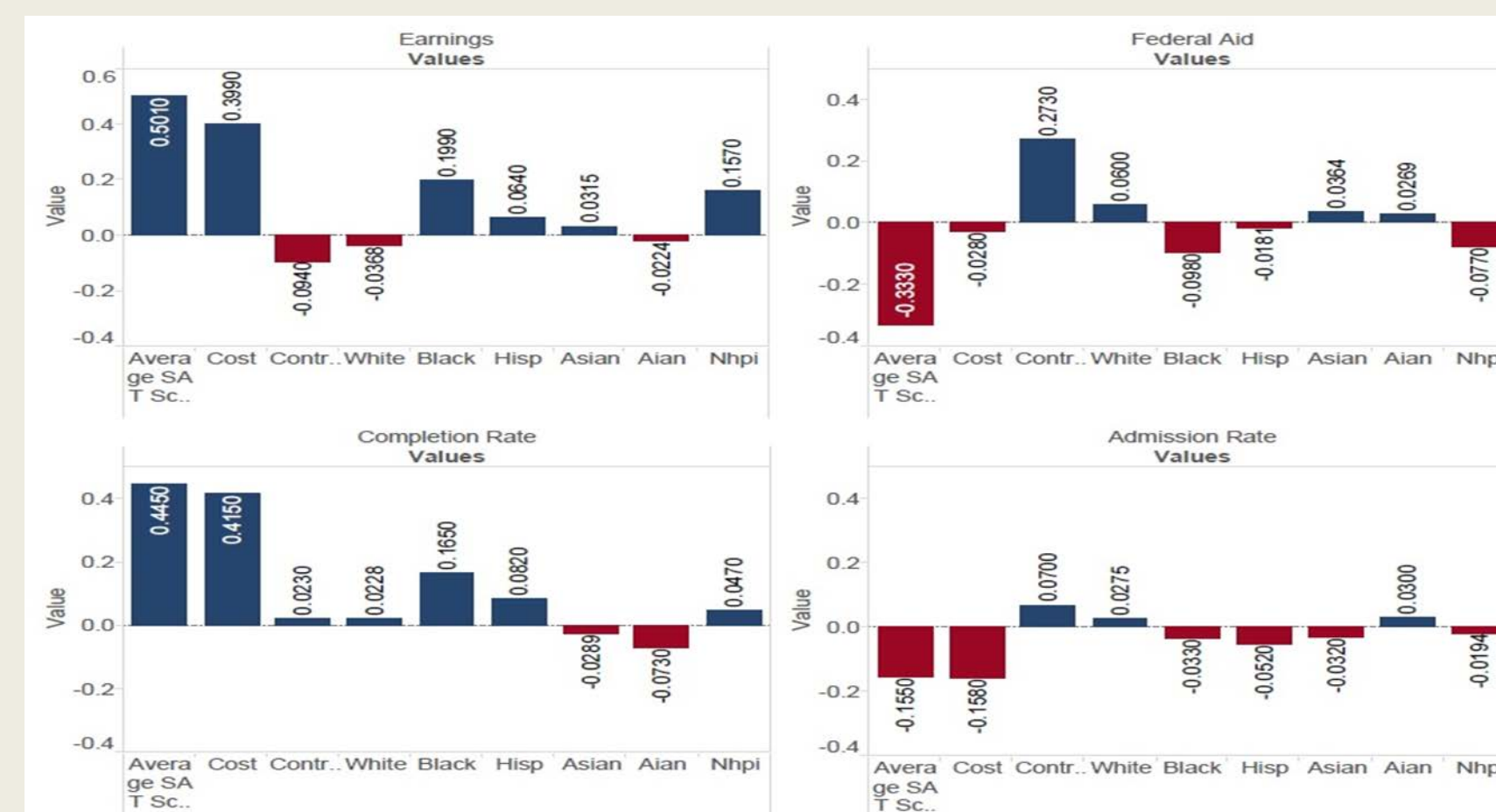 - Violations to normality assumptions were corrected using logarithmic transformations.

Figure 3.

Analysis

In our study, we are looking at predicting the relationship between success factors and each of the independent variable while keeping other independent variables constant. For this purpose **multiple regression** is conducted as it serves the above purpose.

Firstly, the variables are tested for any violations against the two main assumptions of multiple regression.

- **Multicollinearity:** It occurs when two or more independent variables are highly correlated with each other.
- **Homoscedasticity:** This assumption states that the variances along the line of best fit remain similar as you move along the line.

Correlation is conducted to check the relationship between the dependent variables and each of the independent variables. The default enter method is used to conduct the test.

Results

- The results of correlation are depicted Figure 3 for each of the dependent variables.
  - Earnings are most highly correlated with average SAT score.
  - Federal Aid is most highly but negatively correlated with SAT score.
  - Completion rate is most highly but negatively correlated with SAT score
  - Admission rate is not significantly correlated with any of the independent variables
- Figure 4 depicts R square values which indicate the amount of variance accounted by all the independent variables in each of the dependent variable. Beta values depict the amount of variance contributed by each of the independent variables.
- The ANOVA test yield the following results:
  - Earnings: $F_{(9, 1717)}$ =108.382, **p<0.001**
  - Federal Aid: $F_{(9, 1665)}$ =58.595, **p<0.001**
  - Completion Rate: $F_{(9, 1717)}$ =80.001, **p<0.001**
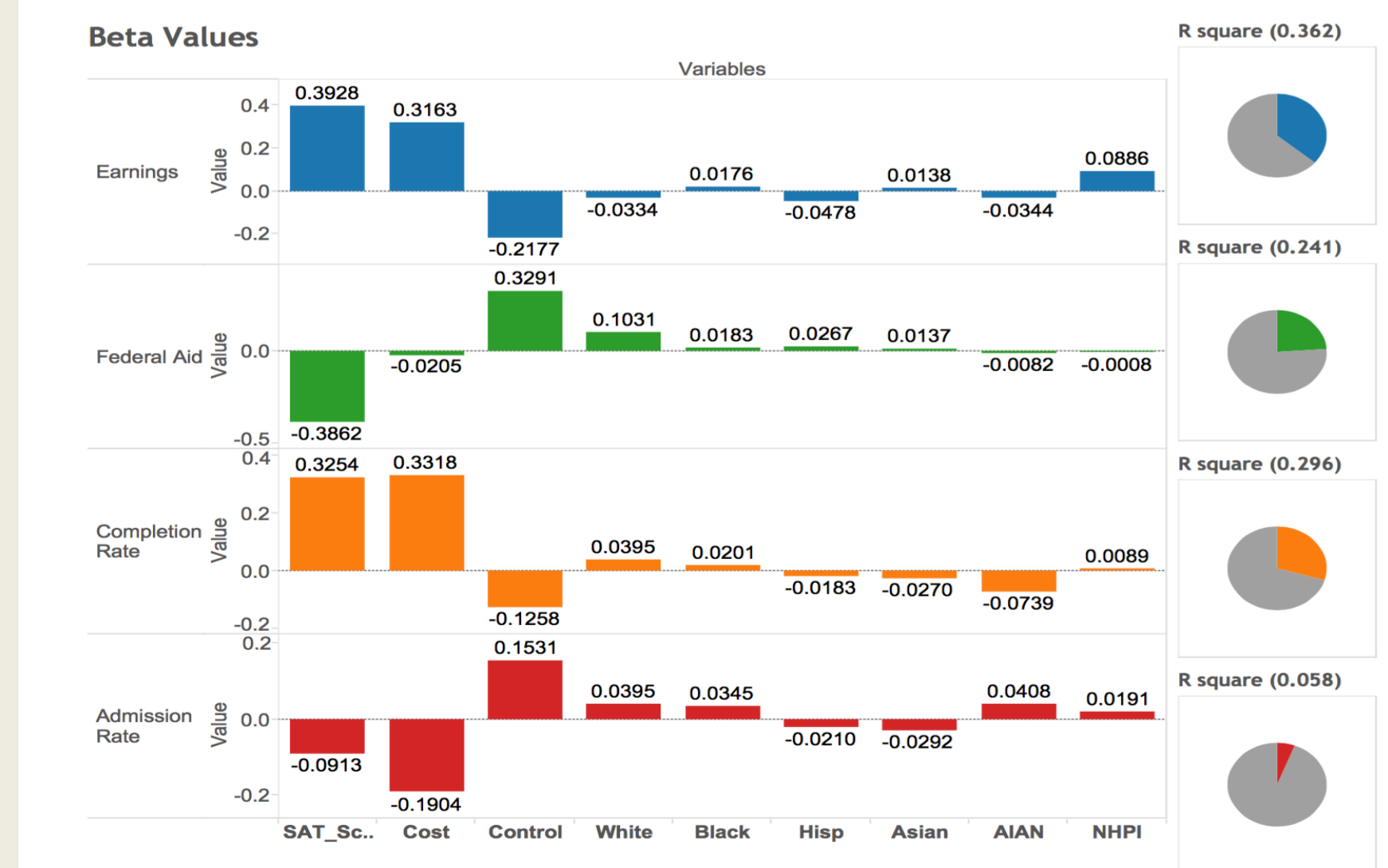 - Admission Rate: $F_{(9, 1716)}$ =11.778, **p<0.001**

Figure 4.

Conclusions

We have obtained **statistically significant** results in our regression model for all the dependent variables. Hence, we conclude that these proxies could serve as some of the many important factors which can be considered for evaluating a university's success. Therefore, students could consider these variables while selecting a university.

Future scope of this project could take into consideration 2-year or less degree granting institutions. Also, other factors that could predict university success could be considered and analyses could be conducted on the same.