**BU Athletics: Sports and Academic Predictor Project Description - Fall 2024**

**Project Description:**  
In this project, we aim to analyze how well student-athletes perform academically at Boston University based on factors such as high school GPA, geography, demographics, language proficiency, chosen major, and standardized exam scores. The goal is to identify trends that can help BU Athletics provide better support to future student-athletes and assist coaches in the recruiting process by understanding where students may face challenges. The analysis will involve studying data from students who have already graduated to identify success patterns or struggles. If time allows, the project may also explore creating a basic regression model to predict academic outcomes based on these factors.

**Teammate Information:**

1. Gukai Chen, stevenu@bu.edu
2. Shiyi Chen, shiychen@bu.edu
3. Yuchen Li, [nikkili@bu.edu](mailto:nikkili@bu.edu)

**Base Questions to Answer:**

1. What is the range of accepted SAT/ACT scores and highschool GPA for student athletes?

* What percentage of these student athletes are domestic students? What about international students?
* Does English being the primary language of the country impact the students' performance?
* How does the academic performance of students with similar HS GPAs compare? (those with SAT/ACT vs. those without SAT/ACT score submitted)? Ex: A student with a 3.0 HS GPA compared academic performance in college and AI (have submitted SAT/ACT score)

1. What is the range of BU GPA for student athletes?
2. How do these students with high SAT/ACT scores or high GPA perform academically at BU compared to student athletes with low SAT/ACT scores or low GPA?
3. Does the average student athlete improve their GPA throughout their college year?
4. Do students from certain geographic areas (Northeast) perform better academically than another geographic area (Southwest)?
5. Are there any significant differences in the academic performance of student-athletes based on their sport?

**Additional Questions to Answer:**

1. Do students from certain geographic areas with higher HS GPA perform better throughout their college?
2. Do students from certain geographic regions with certain gender, race/ethnicity perform better throughout their college?
3. Do combinations of sports teams and regions have a significant impact on the academic performance of student-athletes?
4. Do combinations of primary degree programs and regions have a significant impact on the academic performance of student-athletes?

**Navigation:**

1. For **raw data**, see the 'raw\_data.csv' in **"Dataset"** folder
2. For **data cleaning process**, see the '[BU Athletics] General Data Cleaning' in **"Notebooks"** folder (All csvs besides 'raw\_data.csv' could be reproduced using the notebook mentioned above)
3. For **base question process**, see the notebooks under **"Notebooks -> Base Questions"** folder
4. For **additional question process**, see the notebooks under **"Notebooks -> Additional Questions"** folder (**except for Additional Question 1**, this question is an extended research on base question 5, please see "[BU Athletics]: Base Question 5" under **"Notebooks -> Base Questions"** folder for this particular question, the extension part start from section 5.4)
5. For all **progress report slide**s, see the slides under **“Presentation”** folder (this includes our final presentation)
6. For **Final Report**, see this link: [BU Athletics: Sports and Academic Predictor Project Report](https://docs.google.com/document/d/1lqIAp7yzkSlXI_O64hkXuLzwI0dTQ6Rg2LD7KFL42Sg/edit?usp=drive_link)

**To reproduce results from base questions folder:**

First, make sure the **“Dataset”** folder is downloaded.

**Question 1**

* To reproduce any results using "[BU Athletes] Question 1", you will need
  + install the following packages: pandas, matplotlib, numpy
  + change all the data path to '../Dataset/[correspond\_data\_file].csv'

**Question 2**

* To reproduce any results using "[BU Athletes] Question 2", you will need
  + install the following packages: pandas, matplotlib
  + change all the data path to '../Dataset/cleaned\_data.csv'

**Question 3**

* To reproduce any results using "[BU Athletes] Question 3", you will need
  + install the following packages: pandas, matplotlib, seaborn
  + change all the data path to '../Dataset/cleaned\_data.csv'

**Question 4**

* To reproduce any results using "[BU Athletes] Question 4", you will need
  + install the following packages: numpy, pandas, matplotlib, seaborn
  + change all the data path to '../Dataset/cleaned\_data.csv'

**Question 5**

* To reproduce any results using "[BU Athletes] Question 5", you will need
  + install the following packages: pandas, matplotlib, numpy, seaborn, statsmodel
  + change all the data path to '../Dataset/corespond\_data\_file.csv'

**Question 6**

* To reproduce any results using "[BU Athletes] Question 6", you will need
  + install the following packages: pandas, scipy, matplotlib, seaborn
  + change all the data path to '../Dataset/cleaned\_data.csv'

**To reproduce results from additional questions folder:**

First, make sure the **“Dataset”** folder is downloaded. Please notice that This folder contains all the notebooks that are needed to reproduce the results of **Additional Question 2-4**. For **Additional Question 1**, please see **section 5.4** in "[BU Athletics]: Base Question 5" under the **"Notebooks -> Base Questions"** folder.

**Additional Question 2:**

* To reproduce any results using "[BU Athletes] Additional Question 2", you will need:
  + install the following packages: pandas, matplotlib, seaborn, statsmodels.formula.api
  + change all the data path to '../Dataset/c[orrespond\_data\_file].csv'

**Additional Question 3:**

* To reproduce any results using "[BU Athletes] Additional Question 3", you will need:
  + install the following packages: pandas, matplotlib, seaborn, statsmodels
  + change all the data path to '../Dataset/cleaned\_data.csv'

**Additional Question 4:**

* To reproduce any results using "[BU Athletes] Additional Question 4", you will need:
  + install the following packages: pandas, matplotlib, seaborn, statsmodels, scipy
  + change all the data path to '../Dataset/correspond\_data\_file.csv'