BU Athletics: Sports and Academic Predictor

By

Yuchen Li, nikkili@bu.edu

Shiyi Chen, shiychen@bu.edu

Gukai Chen, stevenu@bu.edu

## 

## Table of Contents

[**Table of Contents 1**](#_nivde979b92m)

[**Introduction 2**](#_3foxy1s14hil)

[Data Cleaning: 2](#_ciiwum9rojy)

[**Base Questions Analysis 3**](#_pc903y5qi4d6)

[Base Question 1: 3](#_9oxpnuyxs8rt)

[Base Question 2: 5](#_l30pmhti7sh8)

[Base Question 3: 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? 6](#_q45iyyjnuulm)

[Base Question 4: Does the average student athlete improve their GPA throughout their college year? 7](#_77v7wqag7391)

[Base Question 5: Do students from certain geographic areas (Northeast) perform better academically than another geographic area (Southwest)?](#_ub93qvfbfnac) 8

[Base Question 6: Are there any significant differences in the academic performance of student-athletes based on their sport?](#_51xruc6r2as4) 10

[**Additional Questions**](#_dva6e7fb587j) **13**

Additional Question 1: 13

Additional Question 2: 15

Additional Question 3: 18

Additional Question 4: 20

[**Conclusion**](#_odhwdekr0fjl) **23**

[**Appendix**](#_chaimj66i8c4) **24**

## 

## **Introduction**

There is one primary dataset that is used in this analysis provided by the client, which includes the detailed breakdown of BU student athletes’ academic-related information like high school GPA, geography, demographics, language proficiency (for international students), major(s), standardized exam score, semester-based GPA, etc.

This paper will discuss general data cleaning and try to help BU Athletics understand 1) how they can better support future student athletes and 2) to assist coaches in the recruiting process to understand if/where a student may struggle by answering to six base questions that were derived from the dataset. Here are the six base questions:

1. What is the range of accepted SAT/ACT scores and high school GPA for student athletes?
   1. What percentage of these student athletes are domestic students? What about international students?
   2. Does English being the primary language of the country impact the students’ performance?
2. What is the range of BU GPA for student athletes?
3. 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?
4. Does the average student athlete improve their GPA throughout their college year?
5. Do students from certain geographic areas perform better academically than another geographic area?
   1. Does High School GPA and students’ geographic difference indicate their performance in college?
6. Are there any significant differences in the academic performance of student athletes based on their sport?

### Data Cleaning:

The raw data frame consisted of 44 columns and 624 rows. The first step is to drop all the unrelated, empty columns (‘Student Athletic Team 2’, ‘Student Athletic Team 3’). Next, check to ensure certain columns are numerical (Eg. ‘Average Undergraduate Applicant High School GPA’). If not, transform the column into numerical format. For certain base questions, we filter the data and save a corresponding separate dataset.

## Base Questions Analysis

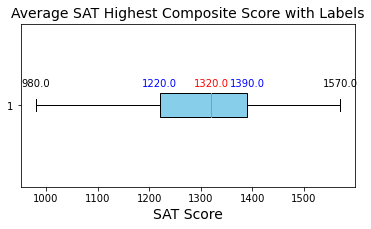
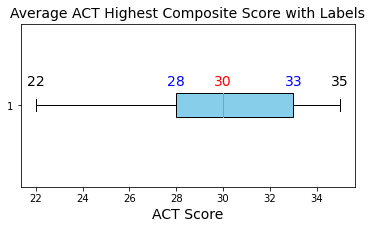
### Base Question 1:

***Data Preprocessing:***

For base question 1, we looked for column ‘Average Undergraduate Applicant High School GPA’, drop any rows without High School GPA. Then, we looked at two columns: ‘Average ACT Highest Composite Score (Applicants Wanting Test Scores Considered)’ and ‘Average SAT Highest Composite Score (Applicants Wanting Test Scores Considered)’ which stores a student’s ACT/SAT score. If any of the two columns is not empty, put it in the dataset named ‘students\_with\_ACT\_SAT\_scores’. Otherwise, put it in the dataset ‘students\_without\_ACT\_SAT\_scores’.

***Range of accepted ACT/SAT scores:***

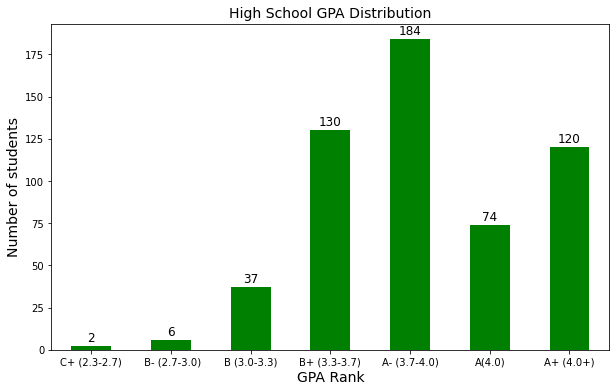
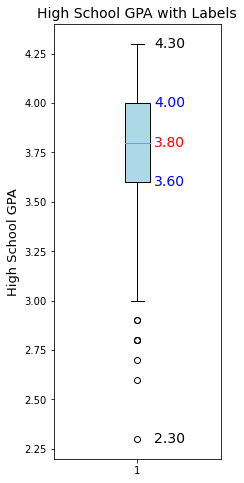
The range of accepted SAT scores is **980 to 1570** (out of 1600).  
The range of accepted ACT scores is **22 to 35** (out of 36).  
Out of **623** students, **105** submitted **ACT** scores, and **176** submitted **SAT** scores. According to the 2024 data, **all** accepted BU athletes with ACT scores were above the national ACT average (19.9).[[1]](#footnote-0) The **median** ACT score is **30**, with 50% of students scoring between 28 and 33. *(Graph 1-1)*

For the **SAT,** the national median score is 1050 (50th percentile). Only **3** students scored **below** the national median, with the lowest SAT score at 980, around the **30th** percentile. The **median** SAT score is **1320**, which is 270 points above the national median. Additionally, **27** students scored **above** the **90th** percentile (1440), and **6** students scored **above** the **95th** percentile (1520). The **maximum** SAT score is **1570**, above the 99th percentile nationally.[[2]](#footnote-1) *(Graph 1-2)*

***Range of accepted High School GPA:***

The range of accepted high school GPAs is **2.3 to 4.3** (out of 4.5).  
For further analysis, GPAs are categorized into six ranks: **C+ (2.3-2.7), B- (2.7-3.0), B (3.0-3.3), B+ (3.3-3.7), A- (3.7-4.0), A (4.0), and A+** (above 4.0, for students enrolled in AP programs, which often use a weighted 4.5 scale).

All accepted BU athletes have a GPA of **C+ (2.3-2.7) or higher.** At least 120 students were enrolled in AP programs and achieved a GPA higher than 4.0. *(Graph 1-3)* The box plot shows **several outliers** (students with GPAs **below 3.0**), indicating that the majority of BU athletes have a GPA above B (3.0-3.3). 50% of students have a GPA between 3.6 and 4.0, and the 0.4 interquartile range reflects **high** academic performance for most students. *(Graph 1-4)*



***Percentage of Domestic/International Students:***

Out of 623 students, **526** are domestic (84.4%), and **97** are international (15.6%).

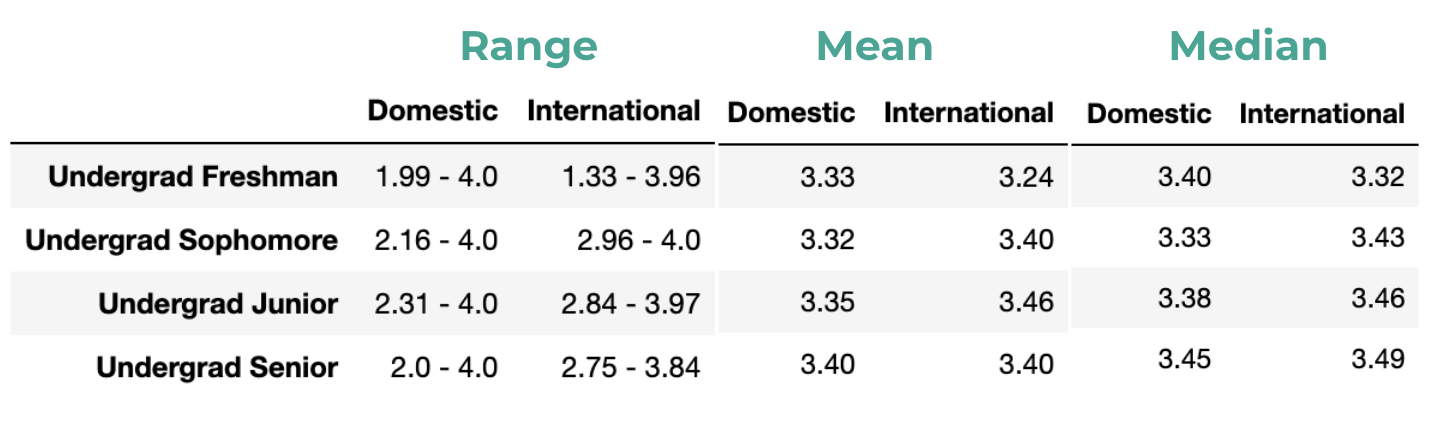
***Does English being the primary language of the country impact the students’ performance?***

English being the primary language **does not** seem to significantly impact **long-term** academic performance.

First, we compared the range, mean, and median of BU GPAs between domestic and international students. For **first-year** students, **domestic** students generally perform **better**, with a smaller GPA range and higher mean and median. However, from **sophomore year onward**, **international** students **outperform** domestic students, with higher mean and median GPAs. This suggests that international students may require **an adjustment period** to adapt to a new education system. *(Table 1-5)*

The box plots show generally **higher variability** for **domestic students** with wider GPA range, and the presence of lower outliers, especially for freshman students. However, the variability tends to **decrease** with **advancing** class year for both groups, proving a period of adjustment for both groups. *(*For additional visual support, see *Appendix Plot 1-1)*

The bar charts of GPA ranks by year show **similar** distributions for domestic and international students, with some variations. In several GPA ranks, international students perform as well as or better than domestic students, especially in **higher** GPA categories. For example, in their **junior** year, **47.4%** of international students achieved an A-, compared to only **21.8%** of domestic students. (For additional visual support, see *Appendix Plot 1-2*)

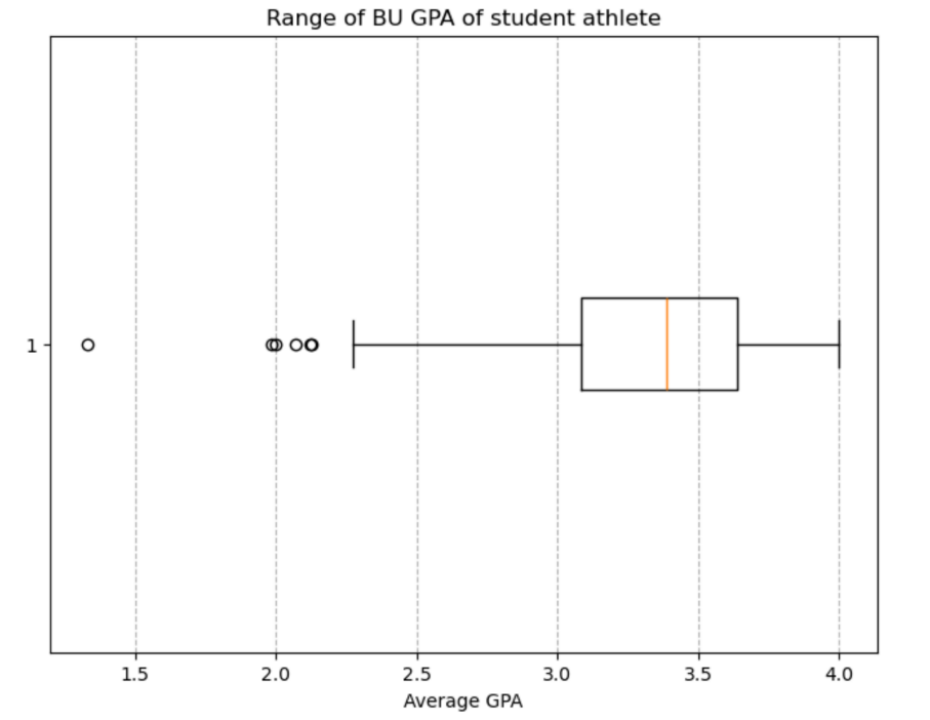


### 

### Base Question 2:

***Range of BU GPA for student athletes***

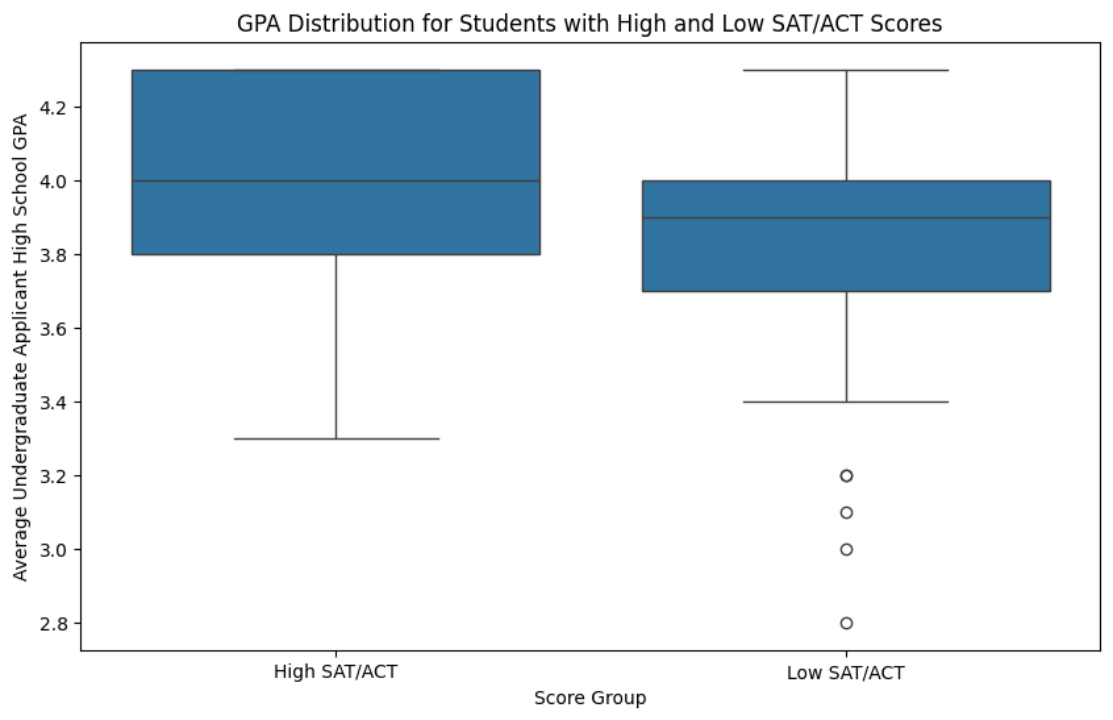
As seen in the box plot, the GPA range for student-athletes in 2024, not including outliers, was 2.28 to 4.0, with an IQR of 3.08 to 3.64. *(Graph 2-1)*

This indicates that most student-athletes (around 75%) effectively balance their academic and athletic life. However, for student athletes in the lowest 25% GPA range, we can consider providing additional support for them, such as tutoring and adjustments to training intensity and duration. These methods may help them improve their academic performance. By proactively supporting these student athletes, BU Athletics can foster a more balanced and successful academic environment for all student athletes.



### Base Question 3: 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?

I chose SAT score to be the standard score, and for those students provided with ACT score, I use a formula to convert ACT score to SAT score. I separate the high and low SAT/ACT groups by the median in SAT/ACT scores. Those students lower than median scores are grouped to low group, those higher the median scores are grouped to high group. I also make a boxplot to separately describe the SAT/ACT score and gpa score separated by high and low SAT/ACT groups.*（Graph 3-1)*

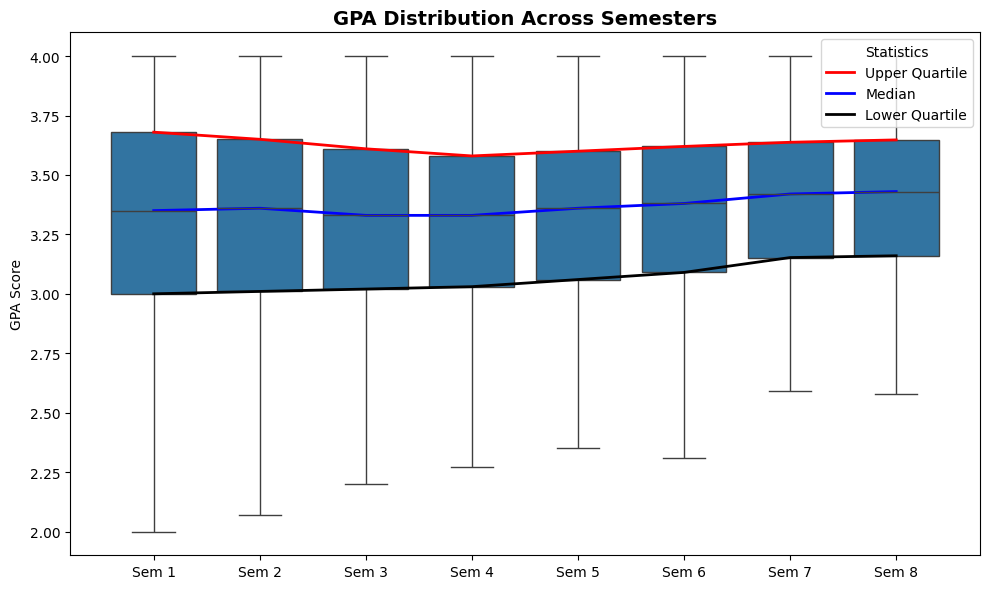


| ***Table 3-2:* Statistics (High SAT/ACT VS. Low SAT/ACT)** | |
| --- | --- |
| **High SAT/ACT Scores Group:** | **Low SAT/ACT Scores Group:** |
| Median: 4.0  1st Quartile (Q1): 3.8  3rd Quartile (Q3): 4.3  Mean: 3.98  Standard Deviation: 0.27  Range: [3.3,4.3] | Median: 3.9  1st Quartile (Q1): 3.7  3rd Quartile (Q3): 4.0  Mean: 3.81  Standard Deviation: 0.31  Range: [2.8,4.3] |

According to the table above we can state that high SAT/ACT group performances are better in all scales: median, Q1, Q3 and mean. The lower standard deviation for high SAT/ACT group indicates that it is less spreaded out from the mean. *(Table 3-2)*

### Base Question 4: Does the average student athlete improve their GPA throughout their college year?

For this question, We create a boxplot with x-axis standing for each semester and y-axis as the Cum GPA for that semester. *(Graph 4-1)* Then we use different color lines to connect the upper quartile, median and lower quartile through each semester. We are able to see the general trend of the increment of GPA throughout the semester in this way.

****

We then create a table to trace the change for the lower quartile group and lowest GPA group. *(Table 4-2)*

| ***Table 4-2:* Change of lower quartile group and lowest GPA group** | | |
| --- | --- | --- |
|  | **Lower Quartile GPA** | **Lowest GPA** |
| **Semester 1** | 3.00 | 1.10 |
| **Semester 2** | 3.01 | 1.57 |
| **Semester 3** | 3.02 | 1.86 |
| **Semester 4** | 3.03 | 2.16 |
| **Semester 5** | 3.06 | 2.35 |
| **Semester 6** | 3.09 | 2.31 |
| **Semester 7** | 3.15 | 2.30 |
| **Semester 8** | 3.16 | 2.29 |

As the table shows, the lower quartile group steadily increases by about 0.1 per semester in GPA while the lowest group increases by 0.3 in GPA.

### Base Question 5: Do students from certain geographic areas (Northeast) perform better academically than another geographic area (Southwest)?

***Data Preprocessing:***

For this question, we separate the dataset based on the column ‘Student Domestic/International Status’ which determines whether a student is a domestic student or international student. For domestic students, we targeted the column ‘Student Home Admission Publication Region’ and dropped any null rows. For international students, we targeted the column ‘Student World Region of Citizenship’ and dropped any null rows. We then save the two dataset as ‘Domestic\_student\_data’ and ‘International\_student\_data’.

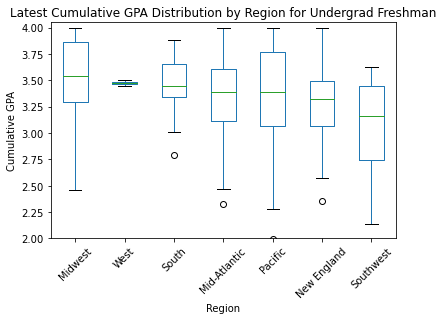
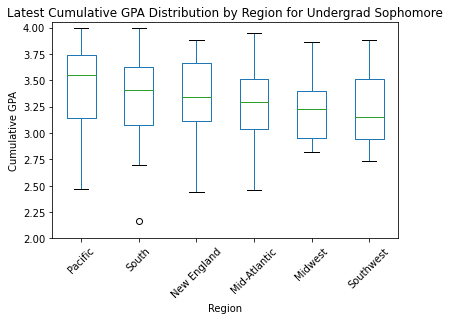
First, we calculate how students are distributed geographically. For **domestic** students, the top 3 regions are: **Mid-Atlantic (31.7%), New England (27.4%), and Pacific (14.7%)**. For **international** students, the top 3 regions are: **Wheat (46.2%), Sourdough (26.9%), and Ciabatta (9.7%)**. *(Table 5-1, Table 5-2)*

| ***Table 5-2*: Distribution of International Students** | | |
| --- | --- | --- |
| **Region** | **Number of Students** | **Percentage** |
| **Wheat** | 43 | 46.2 |
| **Sourdough** | 25 | 26.9 |
| **Ciabatta** | 9 | 9.7 |
| **Baguette** | 8 | 8.6 |
| **Cornbread** | 6 | 6.5 |
| **Pita** | 2 | 2.2 |

| ***Table 5-1*: Distribution of Domestic Students** | | |
| --- | --- | --- |
| **Region** | **Number of Students** | **Percentage** |
| **Mid-Atlantic** | 157 | 31.7 |
| **New England** | 136 | 27.4 |
| **Pacific** | 73 | 14.7 |
| **Midwest** | 56 | 11.3 |
| **South** | 54 | 10.9 |
| **Southwest** | 15 | 3.0 |
| **West** | 5 | 1.0 |

Next, we use box plots to visualize the distribution of college GPAs based on their regions and school years.

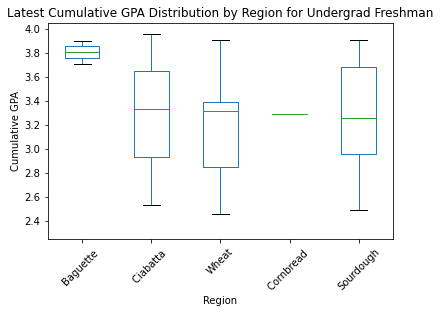
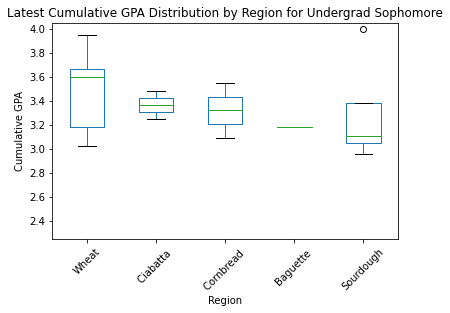
For **domestic** students, students from **South** regions generally perform **better** academically, with consistently high median GPAs across four years, especially in **Junior** and **Senior** years. *(Graph 5-3, 5-4)* The **Pacific** region also shows **strong** performance in **Sophomore** and **Junior** year. *(Appendix Plot 5-2, 5-3)* In contrast, the **Southwest** region tends to have the **lowest** median GPAs across years, suggesting a potential need for academic support. *(Appendix Plot 5-1, 5-2, 5-3)*





For **international** students, students from **Baguette** and **Wheat** regions appear to perform **better** academically, with high median GPAs and consistent performance across years. In contrast, **Ciabatta** region shows **strong** performance in **Freshman** and **Sophomore** years, but drops into the region with **lowest** median GPAs in the **later two years**, suggesting that they may need additional academic support for Junior and Senior year. *(Graph 5-5, 5-6)* **Cornbread** and **Sourdough** regions both show **moderate** performance, with Sourdough generally being the one with lower performance.

For **Wheat** region, we explore a **left-skewed** median line across years, indicating that the **lower** values tend to be more **spread out**. On the other hand, **Sourdough** region shows a **right-skewed** median line, indicating more data concentrated at lower values (especially for sophomore year with high performers as the outlier). (For additional visual support, see *Appendix Plot 5-5 to 5-8*)

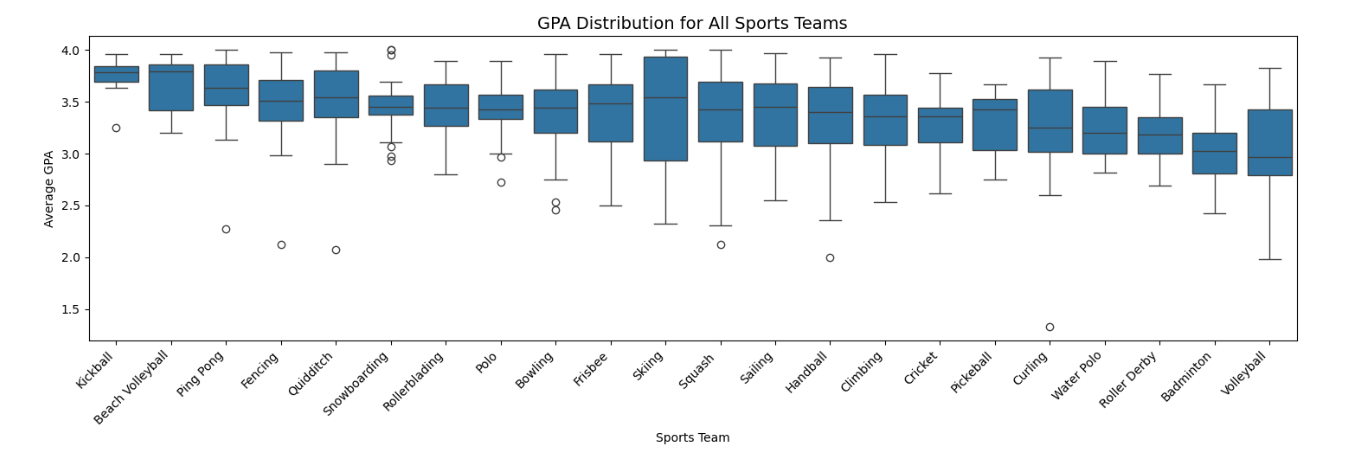


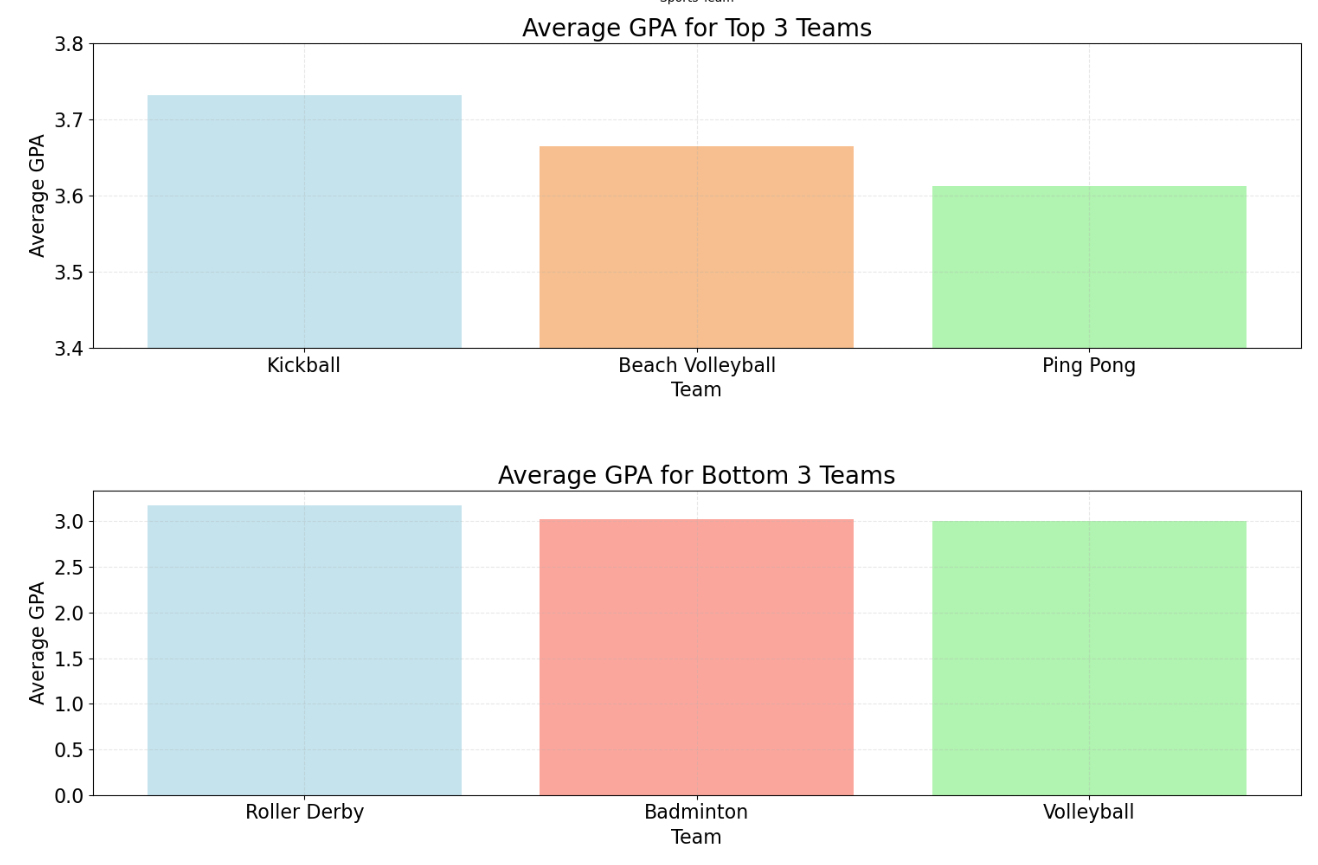


The blocker for this question lies in the **lack of data**, especially for **international** students. With a total of 97 international students in the data pool, we have to remove students from Pita because of the incomplete data from Sophomore to Senior. The same problem also appears to students from Cornbread and Baguette, with having only one data for certain school years.

### Base Question 6: Are there any significant differences in the academic performance of student-athletes based on their sport?

By using one-way ANOVA, we calculated the p-value and found that it is less than 0.05. This indicates that there is a significant difference in academic performance among different sports teams. Additionally, we created a box plot where each sports team has a box showing the GPA distribution of its student-athletes, and we sorted the sports teams by their average GPA in descending order from left to right. *(Graph 6-1)*


We created the separate tables for the top three and bottom three sports teams *(Graph 6-2, Table 6-3, 6-4)*, From this plot below, we can see that sport teams like **Kickball, Beach Volleyball, and Ping Pong** tend to have **higher average GPAs**, while teams like **Roller Derby, Badminton, and Volleyball** are on the **lower end**. Based on this data, we can conclude that there are significant differences in the academic performance of student-athletes based on their sport, and student-athletes in different sports teams will impact students' academic lives. And more attention may be needed for student-athletes with **lower average GPAs**, particularly those participating in **Roller Derby, Badminton, and Volleyball**.

## 

## 

## 

| ***Table 6-3:* Top 3 Performing Sport Teams** | | | |
| --- | --- | --- | --- |
| **Sport** | **Mean GPA** | **75% above** | **Count** |
| **Kickball** | 3.732 | 3.691 | 8 |
| **Beach Volleyball** | 3.665 | 3.421 | 10 |
| **Ping Pong** | 3.612 | 3.467 | 30 |

| ***Table 6-4:* Bottom 3 Performing Sport Teams** | | | |
| --- | --- | --- | --- |
| **Sport** | **Mean GPA** | **75% below** | **Count** |
| **Roller Derby** | 3.177 | 3.350 | 29 |
| **Badminton** | 3.025 | 3.198 | 58 |
| **Volleyball** | 3.009 | 3.430 | 17 |

## 

## Additional Questions

For the additional questions, we focus on finding trends or correlation with two or more variables. The questions include:

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?

***Additional Question 1***

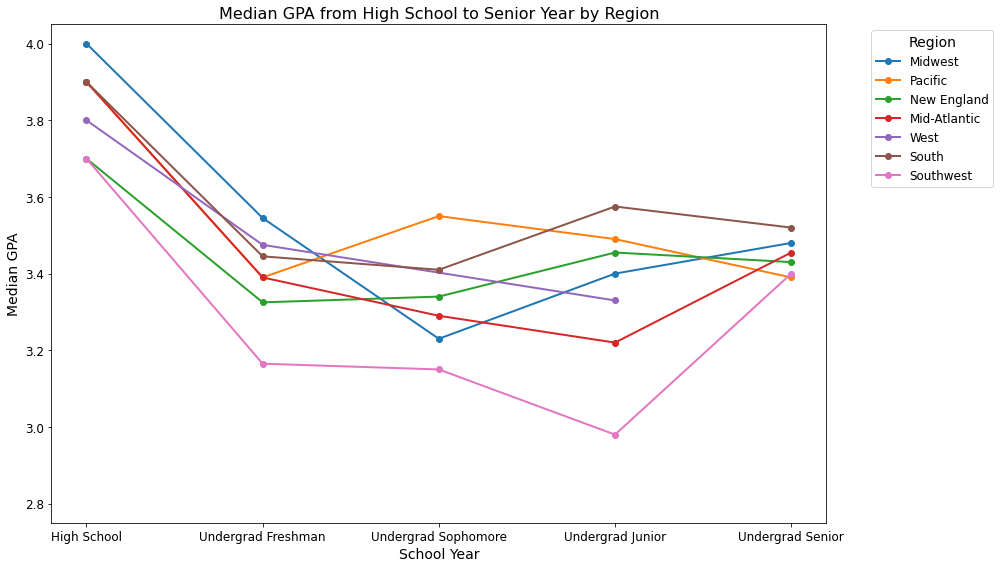
This question explores whether students from specific geographic regions with higher high school GPAs tend to perform better in college. Due to the sample size limitations and data availability, this question is addressed exclusively for domestic students.

***Data Preparation and Visualization***

For simplicity, High School GPA values were capped at 4.0 to align with the college GPA scale for consistency in comparison. A line graph was plotted to visualize the median GPA progression from high school through each college year (freshman to senior) for students grouped by their home admission publication region.

***Key Insights***

From Base Question 5, we conclude that students from the **South** region generally perform the **best**, while students from the **Southwest** region generally need additional academic support. From the line graph, we see that high school GPA does have a **certain impact** on students’ college performance. The **South** region has the **second highest** high school GPA while the **Southwest** region remains the **lowest** for high school GPA, indicating that **high school GPA could be a useful predictor of college success**. *(Graph 7-1)*



To further analyze the question, a **covariance (ANCOVA) model** was used to statistically evaluate the interaction between high school GPA and geographic regions on college performance. The model included:

1. **Dependent Variable:** Latest Cumulative College GPA
2. **Independent Variable:** High School GPA
3. **Interaction Term:** High School GPA × Region

The results indicate that students from the **South** show a **significant** interaction between high school GPA and college GPA, with a coefficient of **0.7073** and p-value of **0.005**. This indicates that high school GPA contributes **more strongly** to college GPA for this region compared to the baseline. *(Table 7-2)*

For the **Southwest** region, the interaction term has a **near-significant** coefficient of **-0.7435** and p-value of **0.082**, suggesting a trend of diminishing returns for high school GPA in predicting college success. This aligns with the observation of lower college GPAs for students from this region. *(Table 7-2)*

| ***Table 7-2*: OLS Regression Results (High School GPA × Region)** | | | | |
| --- | --- | --- | --- | --- |
| **Variable** | **Coefficient** | **Std. Err.** | **t** | **p-value** |
| **HS GPA × South** | 0.7073 | 0.249 | 2.845 | **0.005** |
| **HS GPA × Southwest** | -0.7435 | 0.427 | -1.741 | **0.082** |
| **HS GPA × Midwest** | 0.2869 | 0.252 | 1.141 | 0.255 |
| **HS GPA × New England** | -0.1849 | 0.152 | -1.215 | 0.255 |
| **HS GPA × West** | -0.7757 | 0.806 | -0.962 | 0.337 |
| **HS GPA × Pacific** | 0.1409 | 0.204 | 0.692 | 0.490 |

The limitations for this question is the **small sample size** for some regions, such as the Southwest. Since there are only 15 samples from the Southwest region, it may introduce variability and limit the generalizability of results.

***Additional Question 2***

This question explores whether students from specific geographic regions with certain gender or certain race/ethnicity tend to perform better in college. Again, due to the sample size limitations and data availability, this question is addressed exclusively for domestic students.

***Data Preparation***

ANCOVA model is used to analyzed:

1. **Two-way interactions:** Race/Ethnicity × Geographic Region, Gender × Geographic Region, and Gender × Race/Ethnicity
2. **Three-way interactions:** Race/Ethnicity × Geographic Region × Gender

For interactions related to Race/Ethnicity, we **exclude** the category **“Native Hawaiian or Pacific Islander”** due to having only one sample. To ensure meaningful comparisons, **“White”** students were set as the **reference group** since it is the largest racial/ethnic group in the dataset.

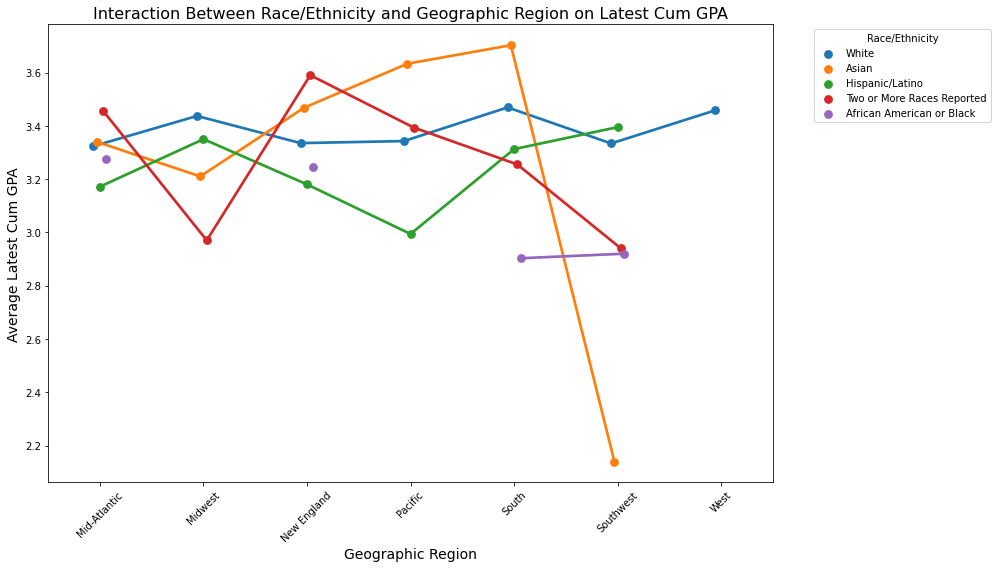
***Two-Way Interactions***

***Race/Ethnicity × Geographic Region***

The analysis of the interaction between race/ethnicity and geographic region reveals notable disparities in college GPA outcomes. **Asian** students from the **Southwest** region exhibit **significantly lower** GPAs compared to the reference group, with a coefficient of **-1.2100** and p-value of **0.006**. Similarly, **African American or Black** students from the **South** also show **significantly lower** GPAs relative to the reference group, with a coefficient of **-0.5162** and p-value of **0.012**.

Additionally, students identifying as **Two or More Races** in the **Midwest** demonstrate a trend **towards** lower GPAs, with a coefficient of **-0.5987** and a near-significant p-value of **0.053**. While this result does not meet the strict threshold for statistical significance, it suggests that this group may also encounter unique challenges in their academic performance within this geographic context.

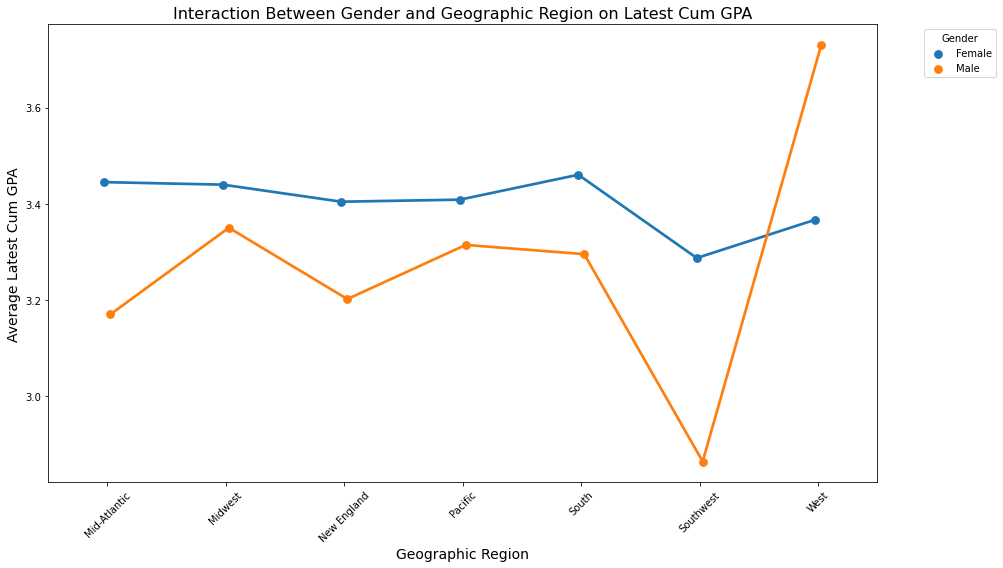
To better visualize the result, a line graph of race/ethnicity and geographic region interactions revealed that Asian students in the Southwest and African or Black students in the South had the **lowest** average cumulative GPAs. *(Graph 8-1)*



***Gender × Geographic Region***

The analysis between gender and geographic region shows **no statistically significant interactions** (all p-value > 0.05). However, from the line graph of the average latest cumulative GPA based on gender and geographic region, **female** students tend to have **more consistent GPAs** across regions and generally **outperform** male students.

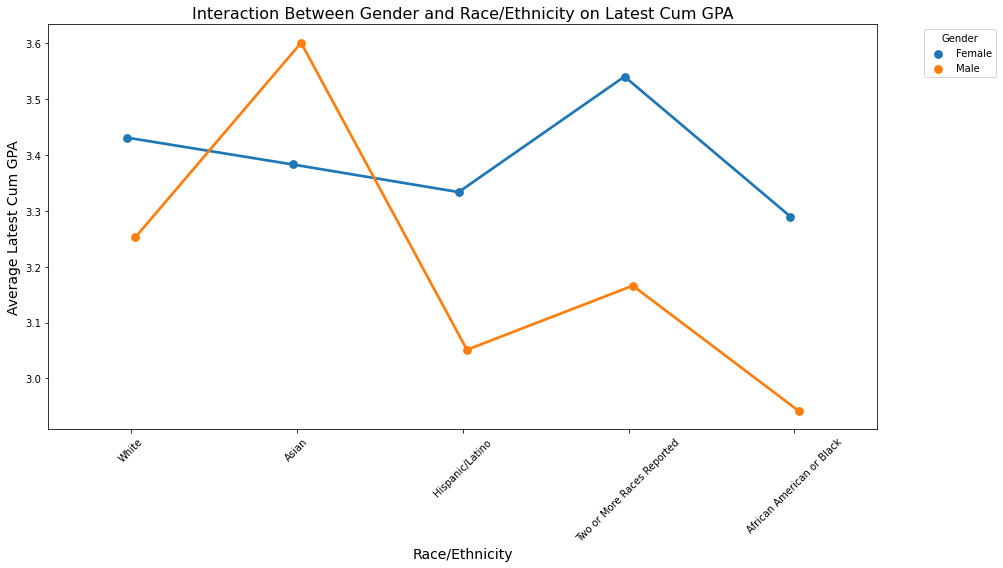
On the other hand, **male** students show **greater variability** in GPA by region, with a **significant dip** in the **Southwest** followed by a **sharp increase** in the **West**. This trend suggests that male students’ performance might be more regionally sensitive. The notable fluctuations might also be partially due to **small sample sizes** (e.g., only six male samples in the Southwest). *(Graph 8-2)*



***Gender × Race/Ethnicity***

The interaction between gender and race/ethnicity reveals significant disparities in academic performance, particularly within the **Asian** student group. **Male Asian** students exhibit notably **higher** cumulative GPAs compared to their female counterparts, with a positive coefficient of **0.3965** and p-value of **0.011**.

In contrast, **female** students demonstrate **more consistent** GPA performance across different race/ethnicity groups, with **smaller variations** in cumulative GPA. This consistency suggests that gender has **less influence** on GPA outcomes within female groups compared to male groups. *(Graph 8-3)*

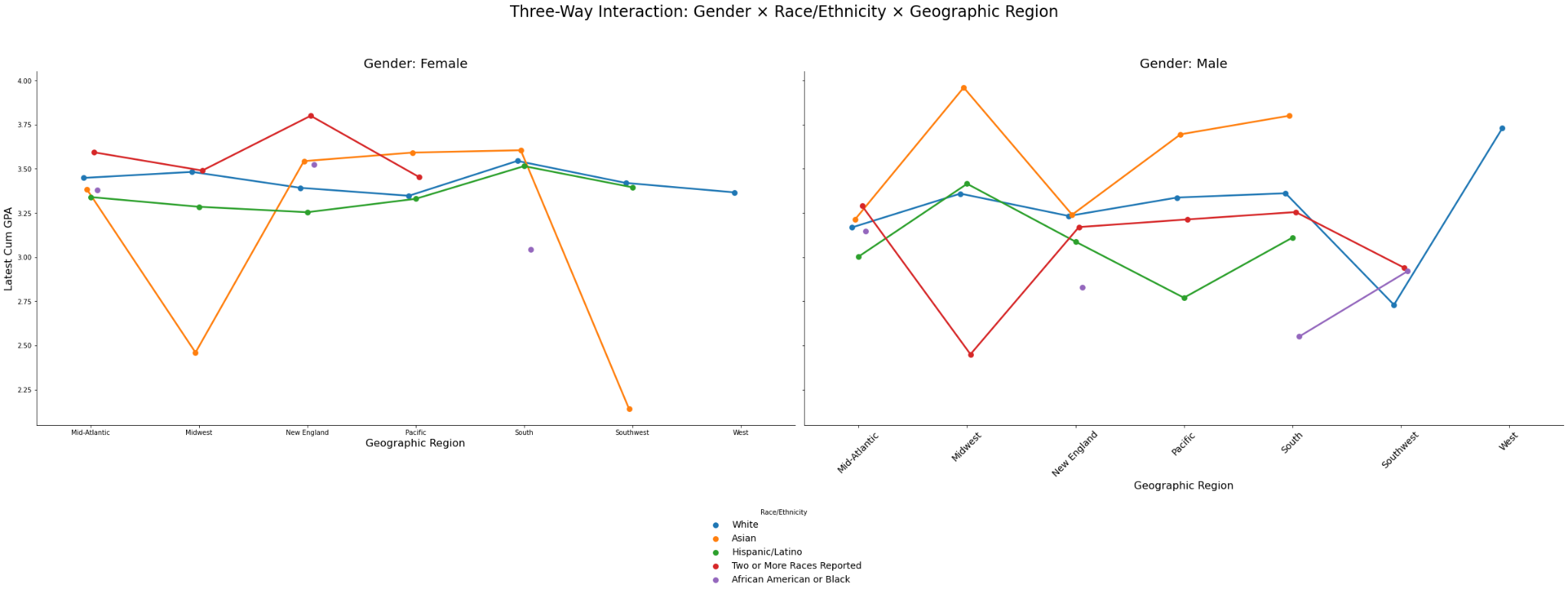


***Three-Way Interactions***

***Race/Ethnicity × Geographic Region × Gender***

The three-way interactions analysis reveals significant differences in academic performance based on the combined factors of race/ethnicity, gender, and geographic region. **Asian male** students from the **Midwest** display notably **higher** cumulative GPAs compared to the reference group, with a positive coefficient of **1.5093** and p-value of **0.017**. This suggests that being an Asian male in the Midwest is associated with stronger academic outcomes in college.

For **Asian female** students in the **Midwest** and **Southwest**, the model shows **highly positive coefficients** of 2.46 and 2.14, respectively (p-value < 0.001 for both regions), indicating a **strong predicated advantage** compared to the baseline. However, their **absolute GPA values** are **much lower** than those of other groups, highlighting a **discrepancy** between predicted and observed outcomes. *(Graph 8-4)*



***Additional Question 3***

This question explores whether combinations of sports teams and regions significantly impact the academic performance of student-athletes.

***Data Preparation***

Build a linear regression model and perform Two-way ANOVA to test the effects of sports teams, regions, and their interaction on GPA.

***Key Insight***

1. The interaction between specific sports teams and regions shows statistically **weaker effects** on GPA for some combinations.

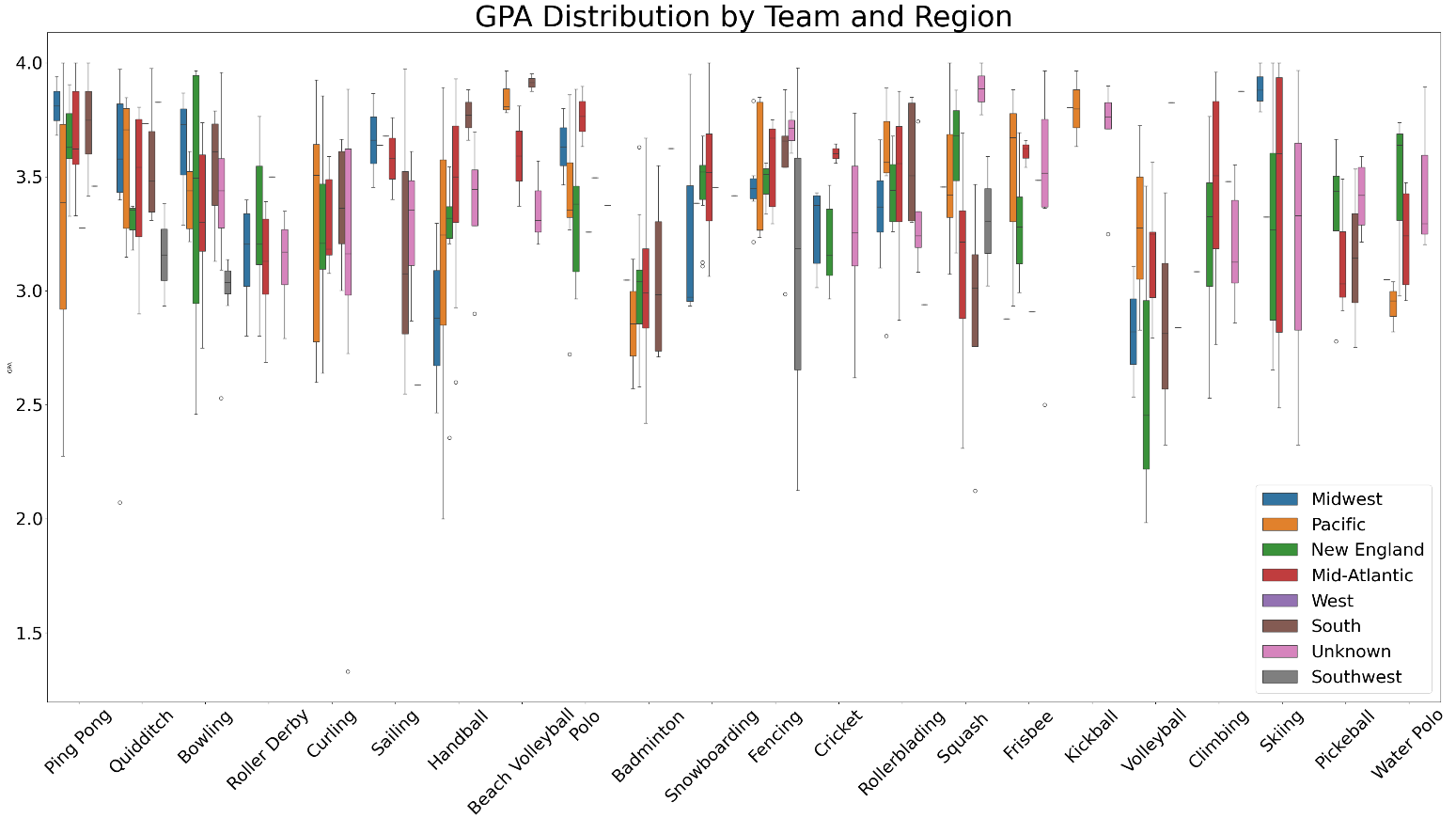
* Squash × New England
* Fencing × Southwest
* Rollerblading × Southwest

1. Certain combinations of sports teams and regions consistently **display lower GPAs**.

* Badminton × Mid-Atlantic
* Badminton × New England
* Climbing × New England

First, we built a linear regression model and performed a Two-way ANOVA to test the effects of sports teams, regions, and their interaction on GPA. Through the analysis, we found that the combinations of sports teams and regions have a weaker impact on the academic performance of student-athletes.Although the overall impact is not very significant, some **specific combinations show significant effects** on GPA. Additionally, certain combinations of sports teams and regions **display lower GPAs**.

We visualized the GPA distribution by team and region. *(Graph 9-1)*



I selected the **three combinations** which have a significant impact on GPA. *(Table 9-2)*

| ***Table 9-2:* Combinations with significant impact on GPA** | | | | |
| --- | --- | --- | --- | --- |
|  | **coefficient** | **standard error** | **t-statistic** | **p-value** |
| **Squash × New England** | 0.51 | 0.22 | 2.36 | 0.019 |
| **Fencing × Southwest** | -1.076 | 0.47 | -2.28 | 0.023 |
| **Rollerblading × Southwest** | -1.14 | 0.57 | -1.99 | 0.046 |

1. **Squash × New England**

The 0.51 coefficient for Squash with New England indicates that this combination has a **positive impact** on GPA. Student-athletes from New England participating in the squash team are likely to **perform better academically**.

1. **Fencing × Southwest**

The -1.076 coefficient for Fencing with Southwest indicates that this combination has a **negative impact** on GPA. Student-athletes from the Southwest participating in the fencing team may have relatively **weaker academic performance**.

1. **Rollerblading × Southwest**

The -1.14 coefficient for Rollerblading with Southwest Indicates that this combination has a **negative impact** on GPA. Student-athletes from the Southwest participating in the Rollerblading team may have relatively **weaker academic performance**.

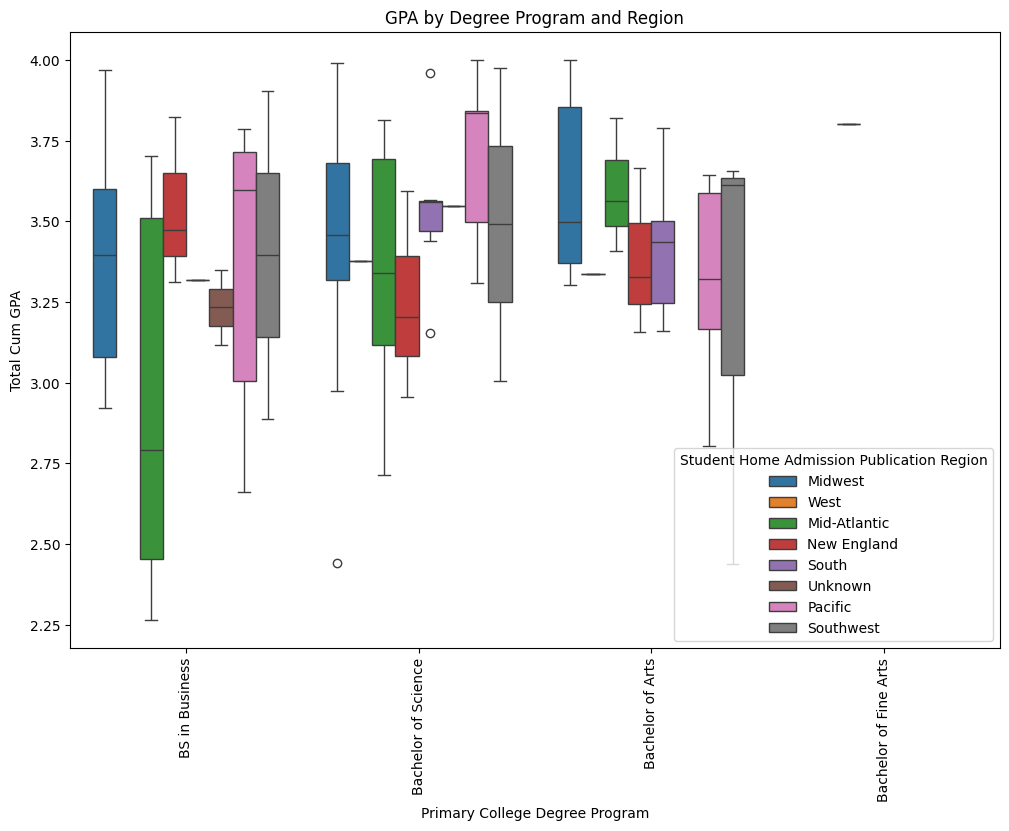
To further **explore which groups have weaker academic performance**, we listed the combinations with the lowest GPAs from the dataset. *(Table 9-3)* Due to insufficient data, focusing only on the lowest GPAs may result in some combinations with very small sample sizes, making them less representative. Therefore, we selected the **three combinations with the lowest GPAs** where the sample size **exceeded 10**.

| ***Table 9-3:* Lowest GPA Combinations (Count>10)** | | | | |
| --- | --- | --- | --- | --- |
|  | **Team** | **Region** | **GPA** | **Count** |
| **1** | Badminton | Mid-Atlantic | 3.01 | 32 |
| **2** | Badminton | New England | 3.02 | 18 |
| **3** | Climbing | New England | 3.24 | 12 |

From the table above, it is evident that different sports teams (Badminton and Climbing) from the same region (**New England**)simultaneously show low GPAs. Furthermore, **badminton** student-athletes from two different regions (Mid-Atlantic and New England) also show low GPAs.

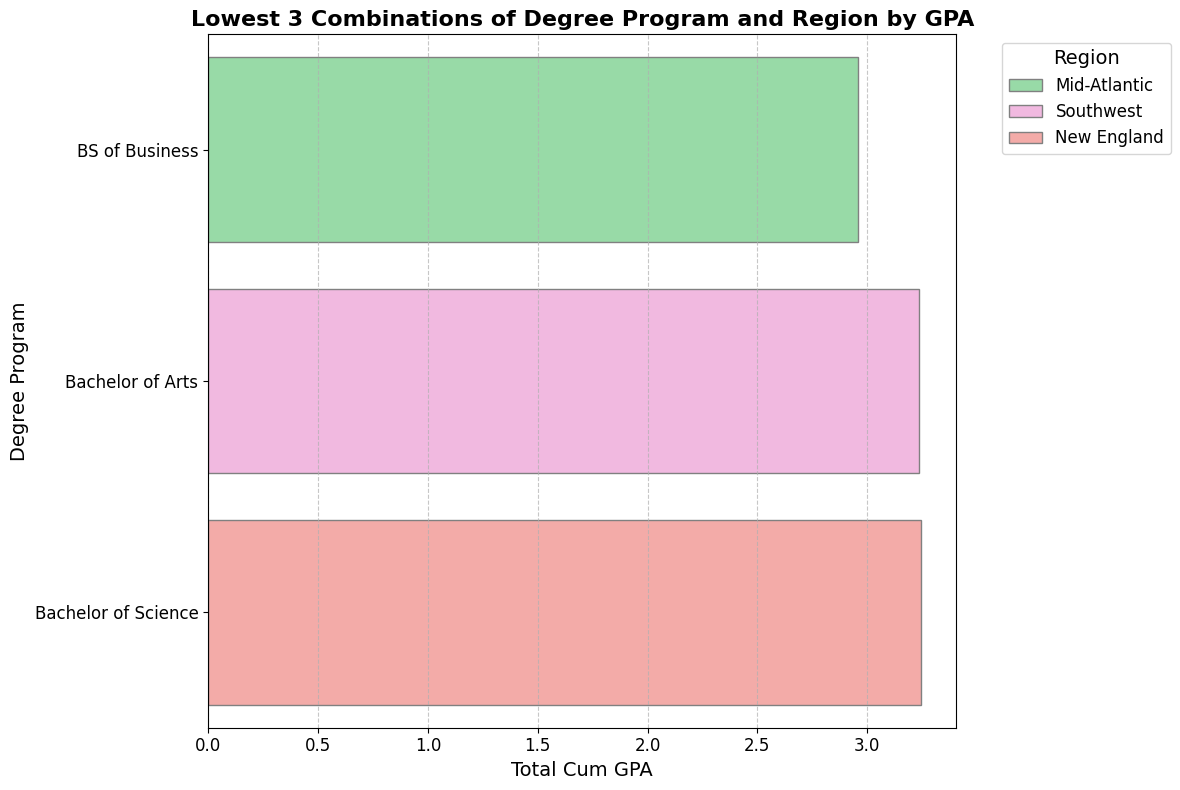
***Additional Question 4***

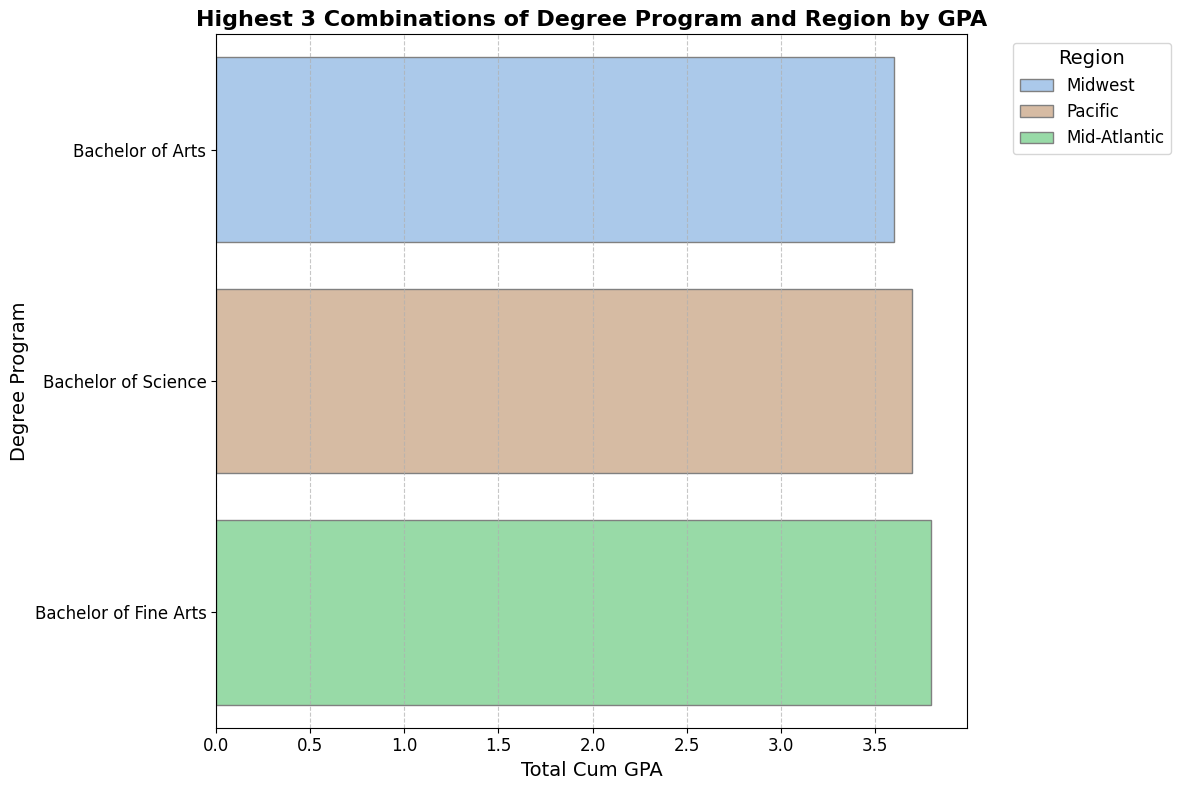
This question explores whether combinations of degree programs and regions have a significant impact on the academic performance of student-athletes. We set the x-axis to be the degree programs and y-axis as the total cum gpa. Each boxplot presents students from different regions. *(Graph 10-1)*



In general, we can’t see any specific patterns from the chart. Then we use a 2-way ANOVA test to calculate p-value to be 0.527, which is larger than the standard 0.05. Thus we can conclude that the combination of degree programs and regions does not have a great impact on student-athletes’ academic performance.

We further explored the lowest and highest GPA for the combination. Same as the statistics shows, we are unable to observe special patterns for the combinations. *(Graph 10-2, 10-3)*

## Conclusion

This study provides valuable insights into the academic performance of BU student-athletes, highlighting several key trends and areas of concern. Language proficiency does not significantly impact long-term academic success, though international students may require additional support during their freshman year. Standardized test scores remain a reliable predictor of college performance, with higher SAT/ACT scores correlating with better academic outcomes. Geographic trends reveal that domestic students from the South and Pacific regions perform well academically, whereas those from the Southwest may need extra support. Similarly, international students from Baguette and Wheat regions excel, while those from Ciabatta face challenges, particularly in their junior and senior years.

Sport-specific analysis identified Kickball, Beach Volleyball, and Ping Pong as top-performing teams, while Roller Derby, Badminton, and Volleyball teams underperform. Over their college years, BU student-athletes generally improve their GPAs, with gradual increases in the lower quartile and lowest GPA values. Two-way and three-way interaction models identified that Asian males, particularly from the Midwest, consistently outperform, while Asian students from the Southwest and Black students from the South underperform. Asian females from the Midwest and Southwest exhibit complex outcomes, with high predicted GPAs relative to the baseline but lower absolute GPAs.

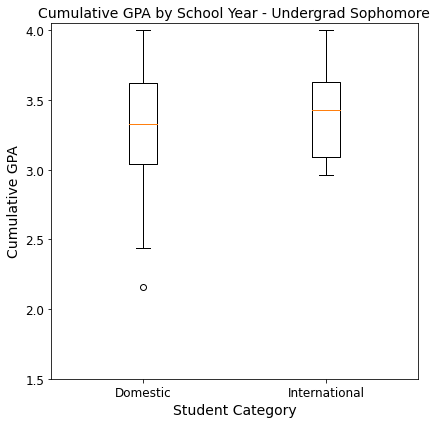
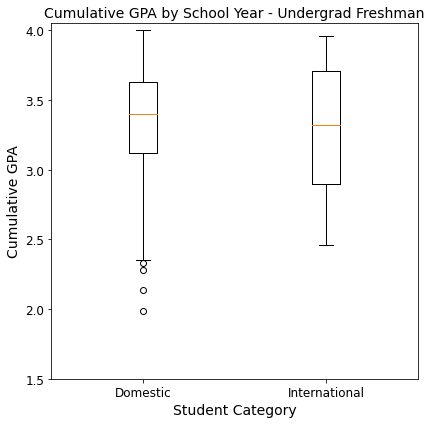
Some specific combinations of sports teams and regions show significant effects on GPA. For instance, student-athletes from New England who participate in the squash team tend to perform better academically, while those from the Southwest on the fencing or rollerblading teams may exhibit relatively weaker academic performance.

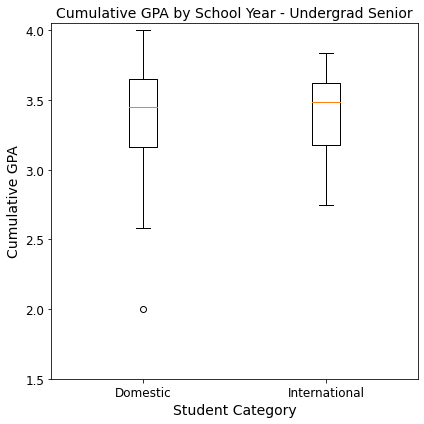
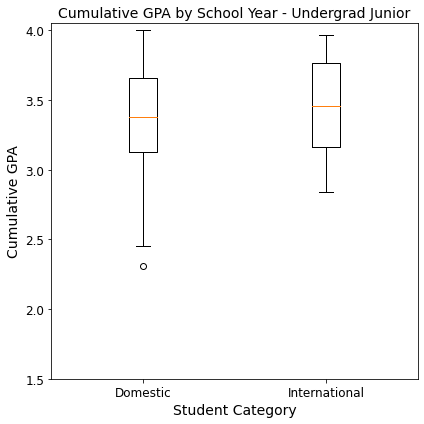
***Future Insights:***This study was constrained by a small sample size (624 records) and limited data variables, which may affect the accuracy and depth of findings. Future research should focus on collecting a larger, more representative dataset and incorporating additional variables that influence academic performance, such as study hours, intensity of sports training, or access to academic resources. Expanding the data scope will not only improve the reliability of findings but also provide actionable insights for BU Athletics to enhance student support systems and refine recruitment strategies for prospective student-athletes.

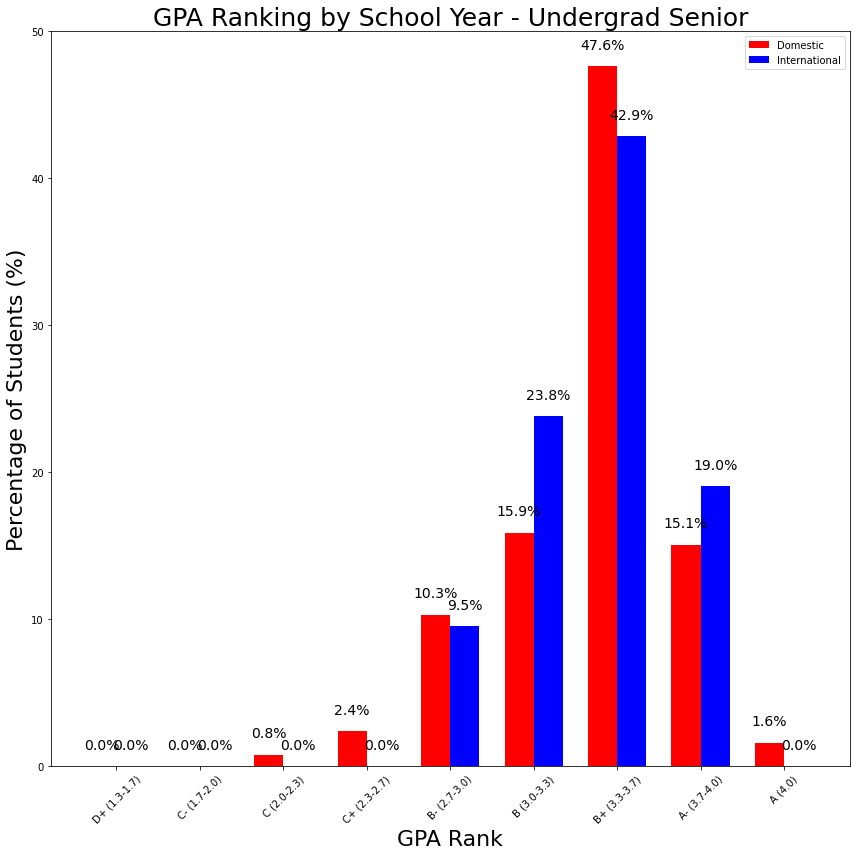
***Individual Contribution:***The project was a collaborative effort with individual contributions clearly defined. Yuchen Li took the lead on Base Questions 1 and 5, as well as Additional Questions 1 and 2. She also managed the GitHub file organization and wrote the introduction and conclusion sections of the report. Shiyi Chen contributed to Base Questions 2 and 6, along with Additional Question 3. Gukai Chen focused on Base Questions 3 and 4, as well as Additional Question 4.

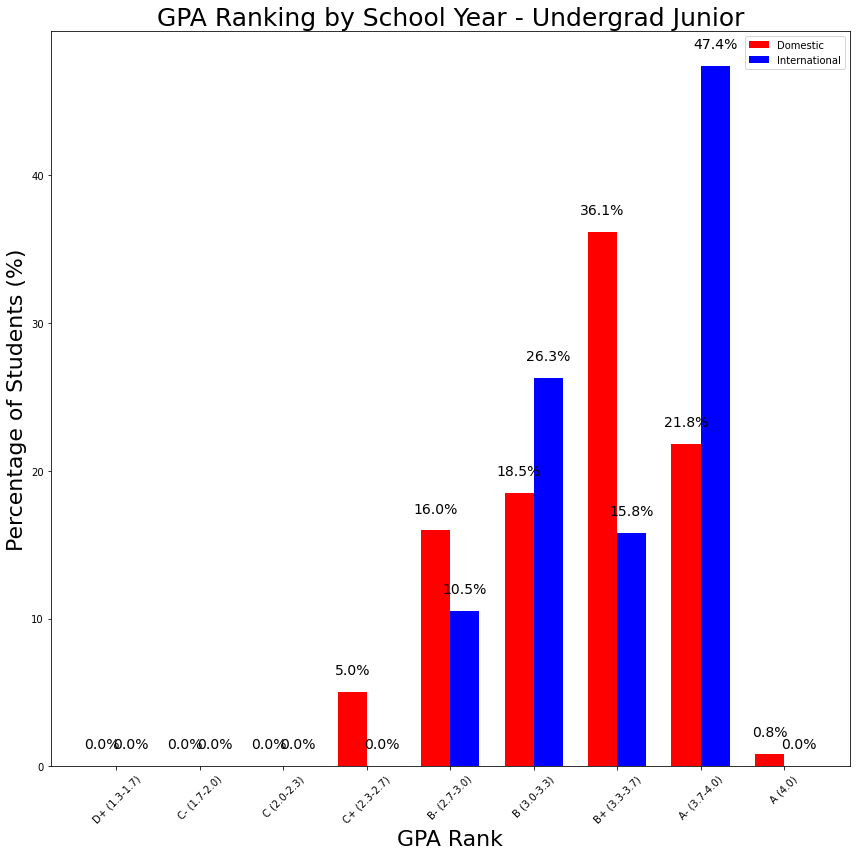
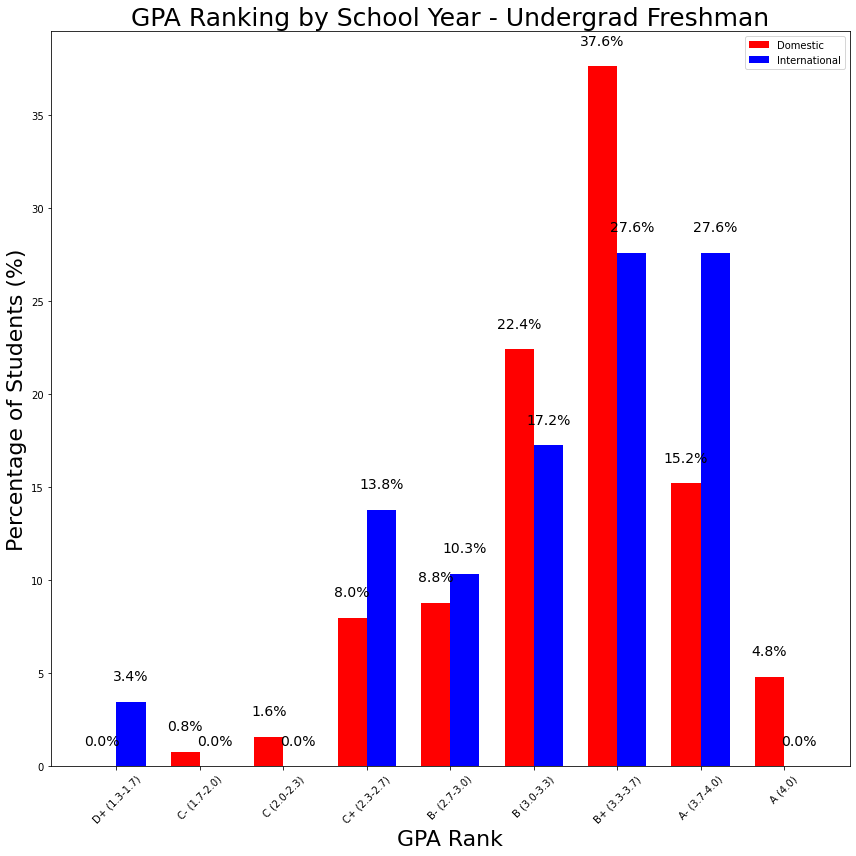
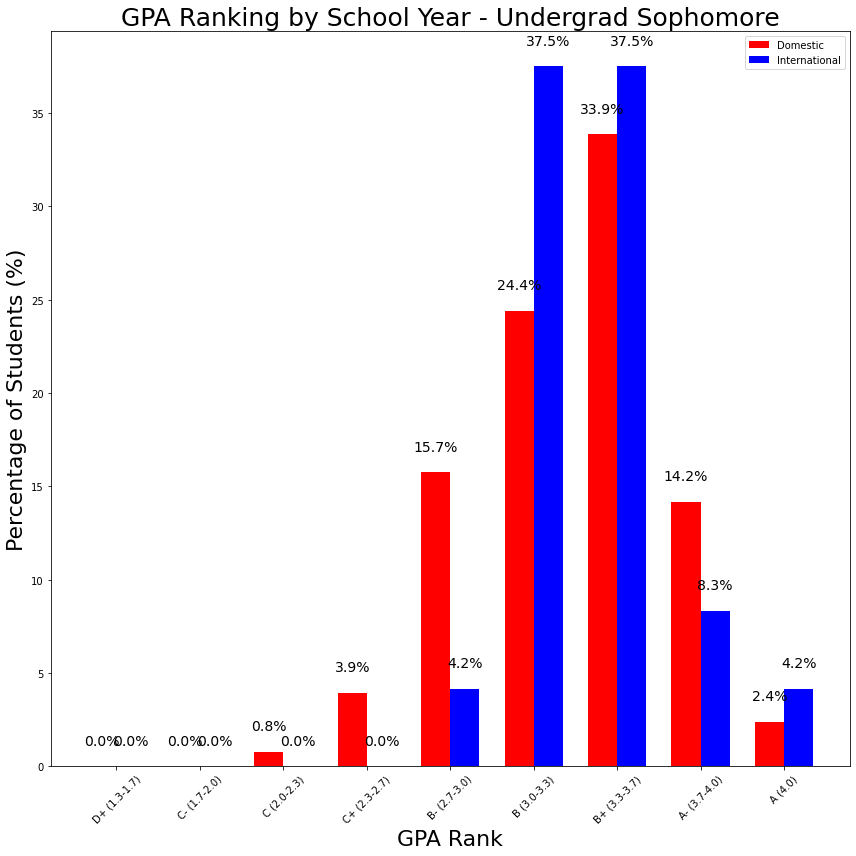
## Appendix

**Plot 1-1**

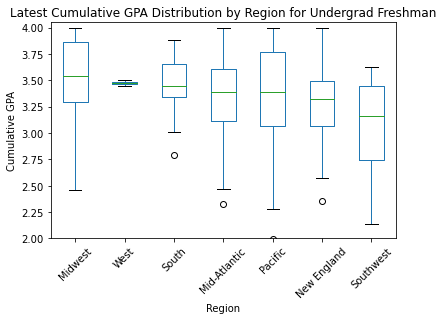
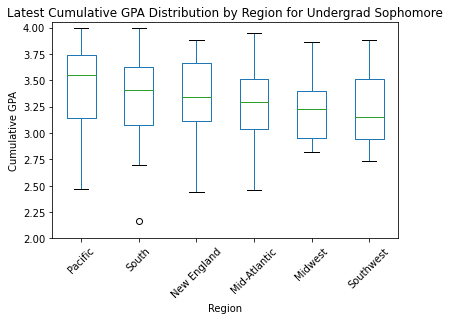


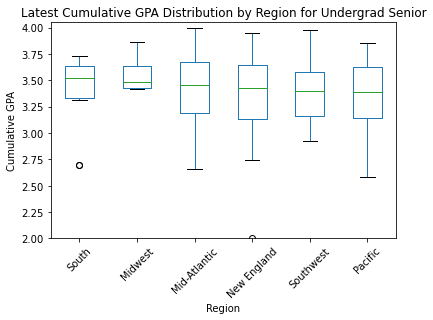
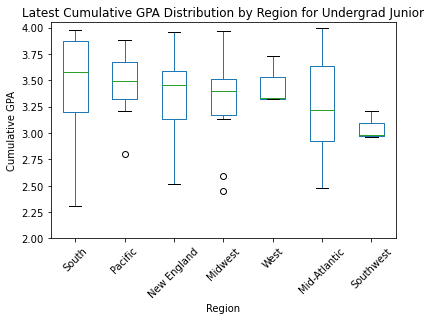


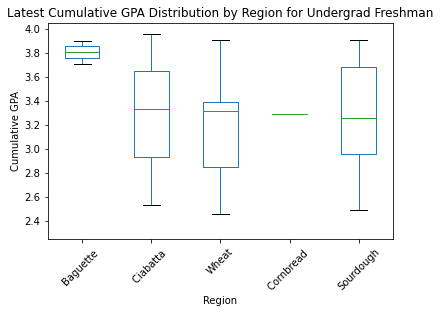
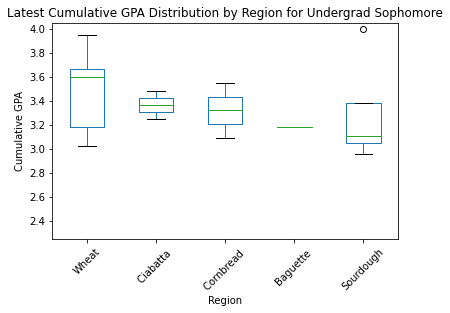
**Plot 1-2**

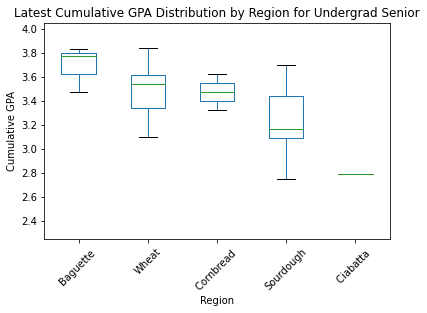
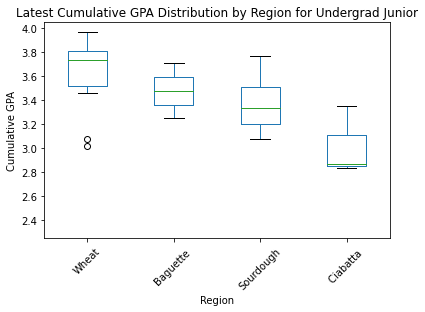


**Plot 5**









1. Quad Education. “Average ACT Score in 2024 | Stats by State, Schools + Guide.” *Quadeducationgroup.com*, Quad Education, 16 Nov. 2024, www.quadeducationgroup.com/blog/understanding-your-act-score-average-good-highest-act-scores. Accessed 8 Dec. 2024. [↑](#footnote-ref-0)
2. “What Is a Good SAT Score in 2024?” Appily.com, 2024, www.appily.com/guidance/articles/college-admission-tests/what-is-a-good-SAT-score. [↑](#footnote-ref-1)