BU Athletics

Team B

Robert Hannon: rhannon@bu.edu, Bargav Jagatha: bargav25@bu.edu Jeff Tedi: jtedi@bu.edu, Mounika Chowdary Akuraju: mounika5@bu.edu, Shangzhou Yin: syin10@bu.edu

Background and motivation

The main problem to examine for BU Athletics has been the relationship between admission statistics, such as test scores and high school GPA, and the ongoing academic performance of student athletes at BU, which is determined by their collegiate GPA. The athletics department is looking to gain deeper insights into the students that are members of their program to see if they can better support their athletes. Athletes face a unique challenge of balancing academics with athletic duties. With this added responsibility, we need to better support athletes in their academic pursuits to further streamline their college pathway. Furthermore, we want to ensure that athletes are able to fully integrate into their teams to ensure the success of teams across all athletic programs. By comprehending and deriving the correlations in the data we will learn more insights about the students the Athletics Department serves and supports.

To address the project's goals and answer the outlined questions access to essential student data is required. This data includes admission scores as mentioned before (SAT/ACT scores, high school scores), academic performance (Cumulative GPA), first language, current sports involvement, whether or not they switched majors, demographic details (gender, race) and Academic Index (a score from 60-240 assigned to students who are applying to colleges based on academic performance) for each student. The dataset provided by Spark! in CSV format comprises 994 records. Utilizing this dataset, we can proceed with analysis, identify correlations, and generate relevant visualizations.

In addition to this dataset we were also provided a table of student athlete survey data that we intended to use for our extension project. This polling data is given to each team each year, and is the average score 1-4 that student athletes vote on (4 being the highest) for questions related to their involvement in their sport, which includes questions about how they believe their academics are valued by the athletic department/their coaches while being a

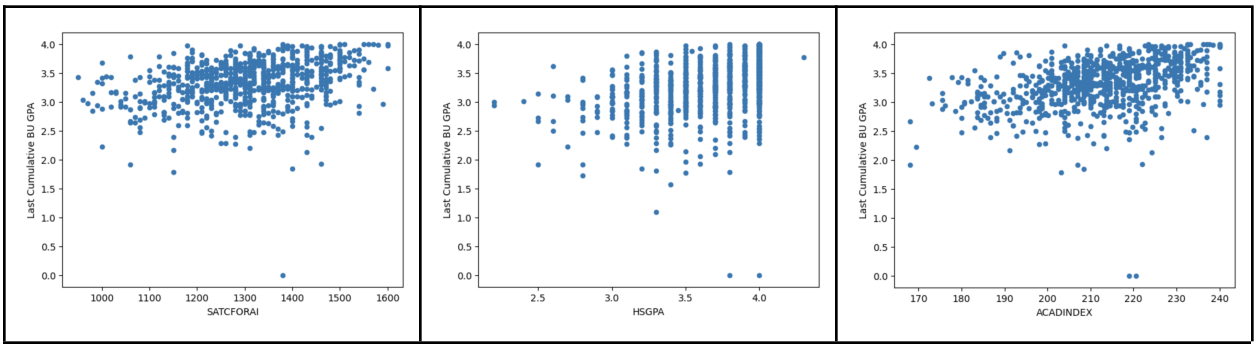
student athlete. A value of 1 indicates that the student athlete or team believes that they are being slighted in some way or another, while a greater value indicates that they believe the athletics department is catering to their unique needs and helping them fulfill their obligations as both a student and athlete.

Data Collection and Preprocessing

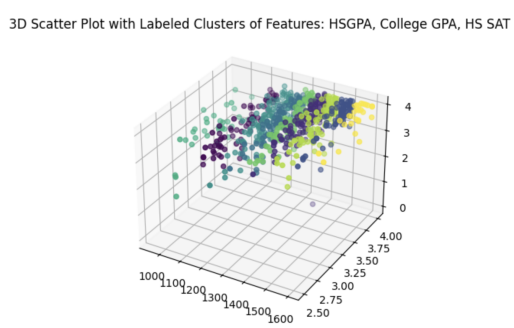
We did our data cleaning with the original dataframe we were given with the individual stats of each student athlete, not with the polling/survey data. We processed the initial Excel table we were given as a CSV then cleaned up the table with pandas. For example, in the dataframe we deleted the column ‘Student’, which was a unique student ID number. We also deleted a few other columns that were all null values, which for example included the 12th-22nd semester GPA values which were 0.0 in the dataframe for all students; we discovered this with the pandas nunique() function. We also renamed some columns to easier names such as ‘English Language Primary Language Of’ to ‘Language’, ‘Athletic Code Description 1’ to ‘Team Name’, ‘First Major (HEGIS) Description’ to ‘First Major Name’ and ‘Last Major (HEGIS) Description’ to ‘Last Major Name’. We also changed some values such as ‘Team Gender’ to a numerical 0 or 1 so that we could potentially use these features in developing future models. We lastly filled in some null values with more valuable assumed values such as in the ‘Country’ column where all foreign-born students had their home country listed; we filled in the other remaining values with ‘USA’, which was the only valid reason someone would not have their country listed. We did not have to collect any sort of data at the onset for the baseline questions, but had to request further data from Spark! And BU Athletics for our extension projects ideas and had to even go to the BU Athletics Office to get questions answered.

Base Analysis

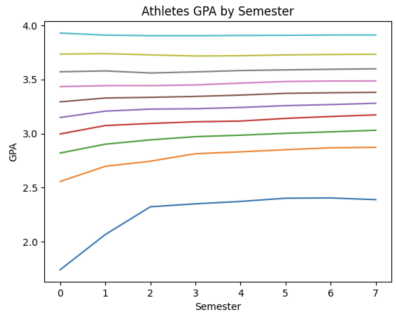
Our first baseline question was to answer how SAT/ACT and high school GPA translate to collegiate academic success. We learned that high school stats are not great indicators of how students perform academically at BU. We tried to make a model that would predict how the student would perform in college. The Decision Tree classifier that we tried to use did not do a great job predicting if athletes will fall into the A, B, C, or lower range of average GPAs. This, along with our scatter plots of Academic Index vs. HS GPA, HS GPA vs. College GPA and SAT vs. College GPA might be an indication that high school stats are not good indicators of how a student will perform in college. These scatter plots did not have any extremely well defined trends between these measurements.



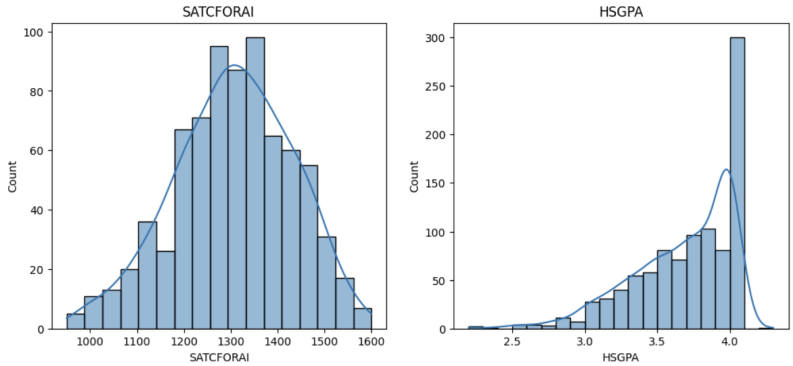
Our next question to answer was whether student athletes improve their GPA throughout their college years. After computing clusters using KMeans based on SAT, College GPA, and High school GPA to find if certain features were indicative of how a student athlete performed over their academic career, we did not find any surprising results, and rather solidified our thoughts from our scatter plots that these measurements do not have any extremely well defined correlations.

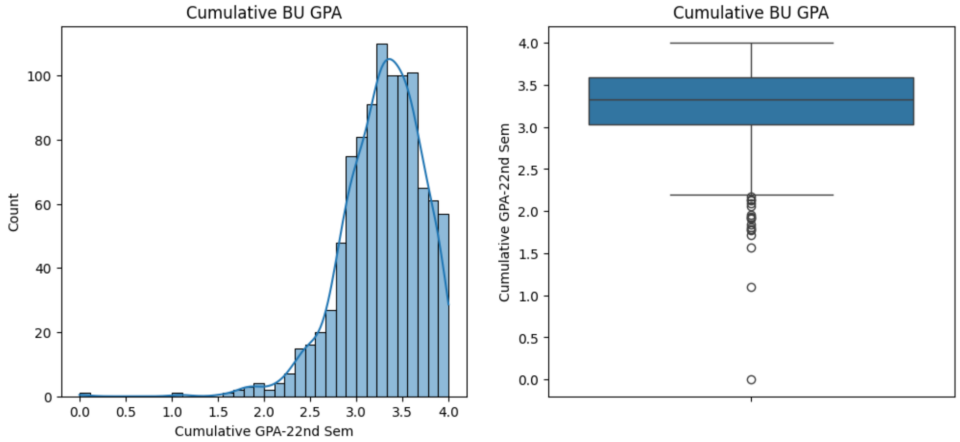


As High school GPA and SAT scores go up, the student usually performs better in college, but this assumption is not linear. Although we could not find any features that are indicative of how a student athlete will perform and change in performance, we did gain insights into how the “average student athlete” performs over their time at BU. We saw in our plots students that start off with lower GPAs usually improve at a greater rate than their peers who start off with higher GPAs as reflected by our graphs. This is not surprising as these students have more room to improve to begin with, but is a good indication that the athletic department works to improve their student athletes and provide support to their student athletes that need it. The following graph is 10 groupings of students just separated by how they performed academically by their starting GPA at BU and how they then improved.



Our next question was to find the range of accepted SAT/ACT scores and high school GPA for student-athletes, along with finding the range of BU GPA for student-athletes. The SAT range was 950 - 1600 and HSGPA range was 2.2 - 4.3, with averages respectively around 1300 and 4.0.





Our next question was to find the most common majors for student athletes. We grouped all students by first majors and last majors and plotted the figures. We can see that Business Administration & Management, Economics, and Health Science have the most students and more students transfer to these majors by comparing these figures. Most students stay in the same major and the distribution between the graphs looks similar.

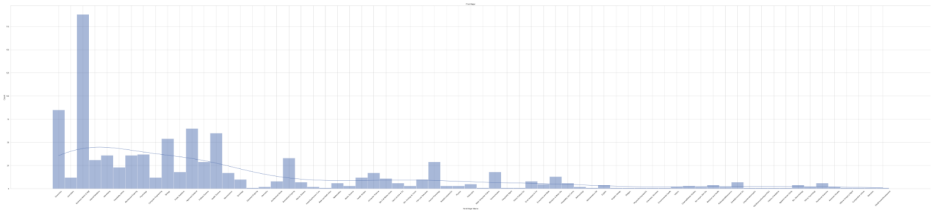


Figure 1: The distribution of first major over all students

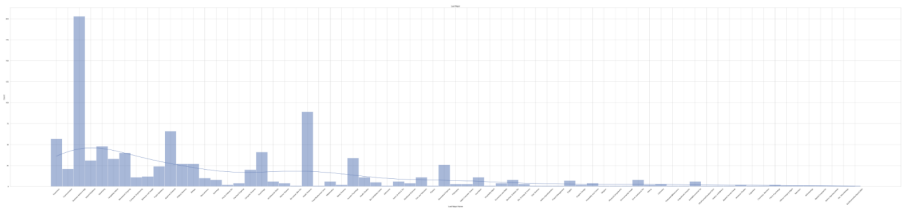
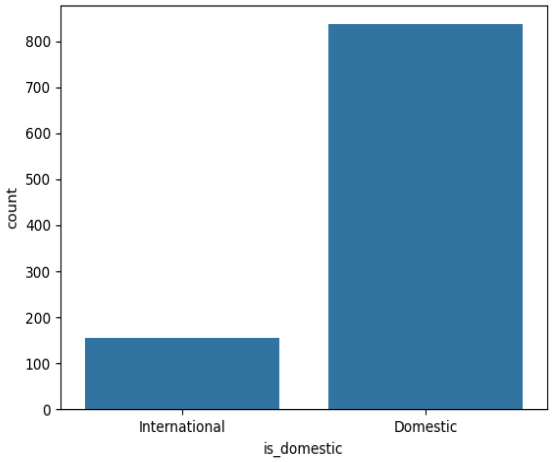
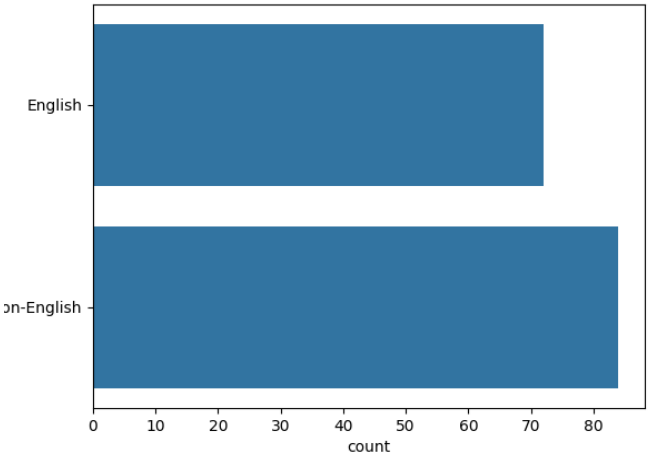


Figure 2: The distribution of last major over all students

The percentage of student-athletes staying in the same major was calculated by the number of students who had the same major over total students. Based on the calculation, it is easy to find that 76.6% of students stay in the same major.



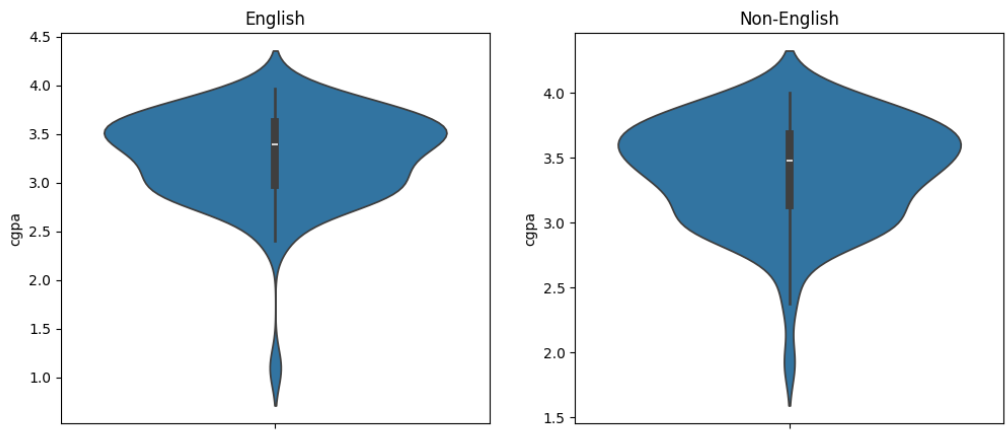
Continuing in answering baseline questions we found that the percentage of international students is approximately 15 percent of the student-athlete body and the percentage of domestic students is approximately 85 percent. There is a higher percentage of domestic students than international students. The bar plot to the left depicts the same.



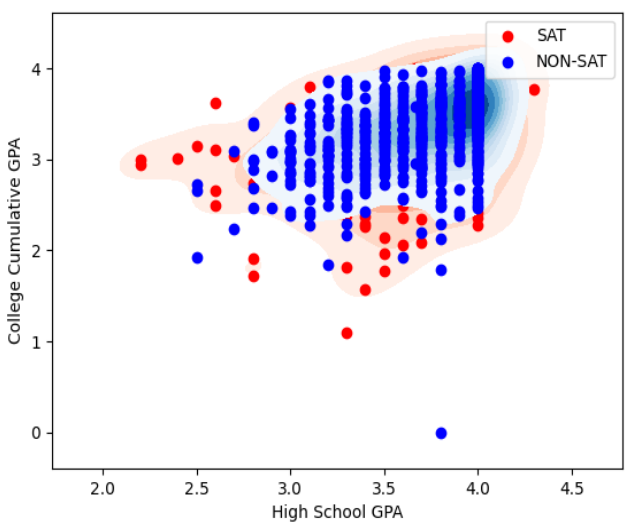
This next bar plot is the count comparing the number of student athletes whose primary language is English with those whose primary language is not.

We also answered the baseline question of whether English being the primary language of the country this student was native to impacted the students' performance. If you ignore the outliers,

the distribution appears to be similar,



suggesting to us that English being the primary language of a student athlete is not making any significant difference in their grades. The lower quartile range is slightly better for non-English students.

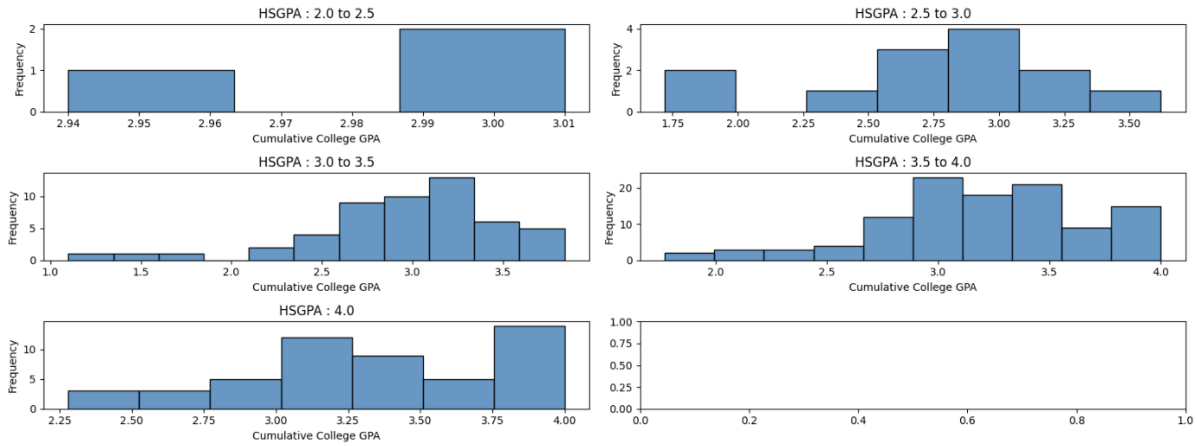


To continue answering baseline questions, we found that when there is a comparison between students who took the SAT and students who did not take the SAT with similar HS GPA, the students who took the SAT have a slight decrease in their GPA in college whereas students who did not take SAT are very consistent with their GPA.

We split the data into students who took the SAT exam and students who did not take the SAT exam. Later we split the high school GPA into buckets with a bucket size of 0.5. We plotted the college GPA of the students on the divided buckets. A clear visualization of the above steps can be seen below:

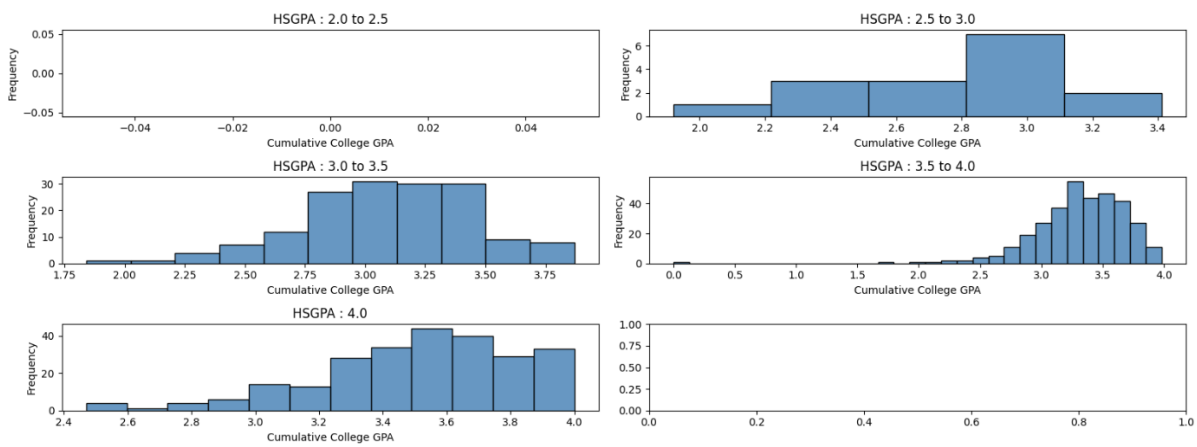
Students taken SAT

```
visualize_gpa_dist(sat_df)
```



NON SAT Students

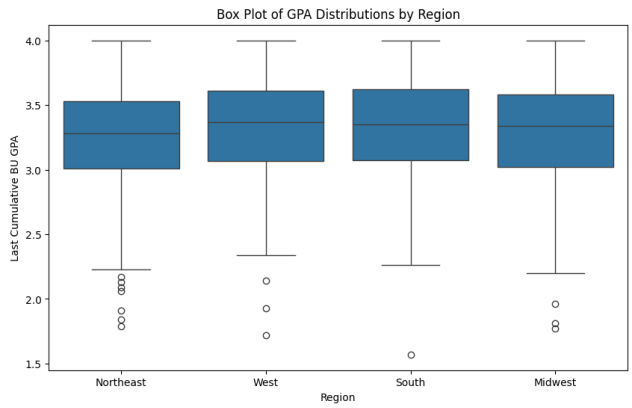
```
visualize_gpa_dist(nonsat_df)
```

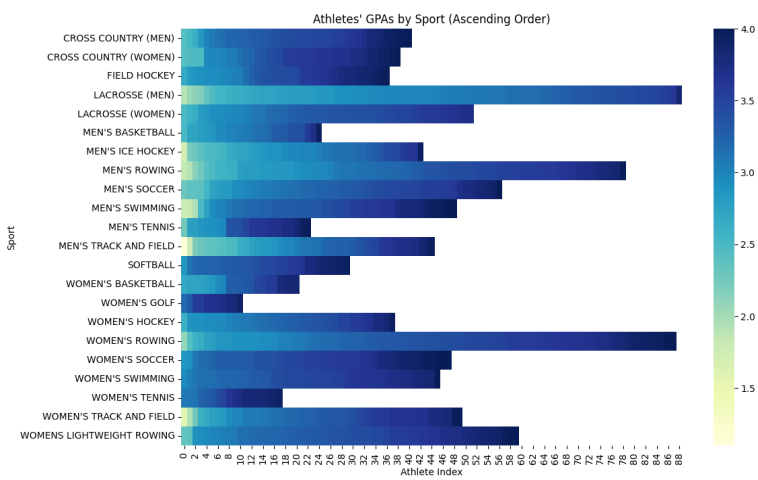


Ideally, for example, for a bucket of HSGPA 3-3.5, the college GPA should also lie in the same region or can go right in case of improvement. We observed (as seen in the above

section) students who did not submit SAT scores turned out to have a little more consistency in their performance than those who submitted the SAT.

Following in answering baseline questions for comparing US regions, we can also see that the average GPA in the Northeast is slightly lower than the other three major regions, but it is generally more consistent.





We also had to compare differences between the average academic performance of different sports. We made a heat map to illustrate the difference (or the lack of it). We can see here that there are pretty big academic differences between sports although the p-value we get from a Kruskal-Wallis test shows that we don't have evidence to reject the claim that they are equal (p-value: 0.45).

Problems Encountered

The success of this project is heavily dependent on the quality and availability of the dataset. The data collected is only from several states in the US and several foreign countries. In addition, some relevant features are incomplete or inaccurate. There is also minimal data surrounding some features, such as from certain teams that are much smaller than others. This project analyzes the correlation between admissions data and the performance of student-athletes through their careers at BU, so the result of this analysis cannot be generalized to other institutions with different admission criteria.

Extension Motivation

As part of our extension proposal we were asked to discuss a variety of different mini-project ideas with our groups. Most of our ideas came from questions that we believed were worth asking as extensions to the baseline questions. We had seen various trends in the data we visualized and collected, and we have also seen that this dataset has a unique lack of trends as well, leading us to question why this might be. We had to frame all of our extensions around correlations to college GPA, since that is the only metric of academic performance we have for college.

Extension Mini-Project: Athlete Classification

The extension project that came to mind at first is one that is directly related to our problem statement. We were given many features of high school and personal/demographic data concerning each student, and wanted to see if there is a way we can develop a model to best predict the GPA of our student athletes. We had seen a lack of trends directly concerning high school GPA and testing scores versus college GPA, but only tried to use unsupervised learning techniques (clustering) to better classify these students. We had not been left with many informative insights by these techniques, but with experimentation of features and of

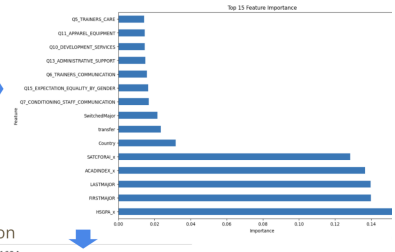
model type and model parameters believed we could discover if we can create some sort of regression model to predict the GPA of students.

The survey data did not have any missing values, however we did have to normalize the data as they were just numbers 1-4. This was necessary for our extension project for model performance, so we made the decision to normalize all our other data for model creation as well. We also turned all the categorical data into numeric data by assigning unique numbers to unique categories.

We had to predict the academic performance based on the feature “Last Cumulative BU GPA” of student athletes by constructing the binary classification model, since trying to actually calculate loss for an exact regression model that predicted GPA was extremely difficult, especially while focusing on logistic regression. We then made it our goal to predict whether we could construct models that would predict, or classify, if the student is likely going to be underperforming (GPA < 2.5) or overachieving (GPA > 3.5) in their time at BU and we tested them separately. We chose 2.5 as our academic threshold since this is the GPA to remain academically eligible, and we chose 3.5 as our other threshold for overachieving since this is roughly the threshold most BU colleges have for receiving Latin Honors distinctions upon graduation. The “Last Cumulative BU GPA” was first categorized into two bins (0 - 2.5) and (2.5 - 4.0) to test underperforming and then categorized into two bins (0 - 3.5) and (3.5 - 4.0). The GPA related features were excluded for training. The accuracy of predicting GPA for students who are underperforming is 0.96 and the accuracy of predicting GPA for students who are overachieving is 0.72. The detailed implementation of our logistic regression models and results can be found in our repository, along with seen in some of the below visualizations.

	Last Cumulative BU GPA
0	3.00
1	3.46
2	3.53
3	3.86
4	3.66
5	2.89
6	2.39
7	3.20
8	2.38
9	3.46
10	3.62
11	3.80
12	3.12
13	3.09
14	2.91
15	3.03
16	3.51

class 1: [0 - 2.5]
class 2: (2.5 - 4.0]



Logistic Regression

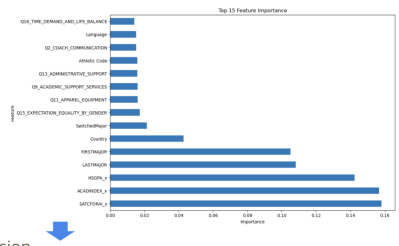
```
Accuracy: 0.9593908629441624
Classification Report:
              precision    recall  f1-score   support

           1       0.00      0.00      0.00         7
           2       0.96      0.99      0.98       190

    accuracy                           0.96       197
   macro avg       0.48      0.50      0.49       197
weighted avg       0.93      0.96      0.94       197
```

	Last Cumulative BU GPA
0	3.00
1	3.46
2	3.53
3	3.86
4	3.66
5	2.89
6	2.39
7	3.20
8	2.38
9	3.46
10	3.62
11	3.80
12	3.12
13	3.09
14	2.91
15	3.03
16	3.51

class 1: [0 - 3.5]
class 2: (3.5 - 4.0]



Logistic Regression

```
Accuracy: 0.7208121827411168
Classification Report:
              precision    recall  f1-score   support

           1       0.75      0.88      0.81       130
           2       0.64      0.42      0.50        67

    accuracy                           0.72       197
   macro avg       0.69      0.65      0.66       197
weighted avg       0.71      0.72      0.70       197
```

This infers that the logistic regression model can accurately predict GPA for students who are underperforming but cannot predict overachieving well because of the unbalanced dataset. This occurred for many reasons. First of all our feature space was small so we only had a few metrics for our model to work with. We also had to generalize for the model to train and perform the way we thought would be the best. We tried to first expand the feature space by including features from the survey dataset, merging the survey data and athletics data on team name. However the survey data is only an average for each team, while the athletics dataset is for individual athletes, which leads to generalizations of how each of the athletes voted in the survey. Also, there were large numbers of athletes who did not have HS GPAs or HS SAT scores. Since our dataset was small and we did not want to completely get rid of what would have been more than 200 entrants, we gave these missing values the average HS GPA and average SAT score of their teammates for their respective sport. We had to make this tradeoff to

hope for optimal performance of the model in training, since we were already making the testing set for model performance .1 to leave as many entrants for training.

Extension Mini-Project: Extending Baseline Questions

We have also formulated more questions for our extension as responses to our baseline questions. For example many of the baseline questions ask about the “average student athlete”. We then separated some of our baseline questions and asked about them on a team-by-team and major-by-major basis. This we believed would potentially give us insights into if certain teams are performing better than others and why that might be the case. Some of these baseline questions we looked into further for these specific team-by-team groupings also included if these athletes had significant fluctuations in their GPA during their career at BU and how their high school statistics compare to their academic performance at BU. We also expanded on one of our baseline questions specifically, which asked us to compare high school and college GPA. We expanded the correlation plots between these values across various groups based on different categories such as ‘Entering Term’, ‘Team Gender’, ‘FIRSTMAJOR’, etc. For the baseline questions this analysis was only asked to be conducted for students who submitted SAT scores versus those who did not.

To start answering these new questions we represented the data we had been given in more useful ways such as by the following groupings which give us better insights into the specific teams and major we are going to be asking questions about in the future. We made a new CSV that groups every team by the average GPA of the majors that make up the team and we also have another new CSV that we created that is of the average GPA of the team.

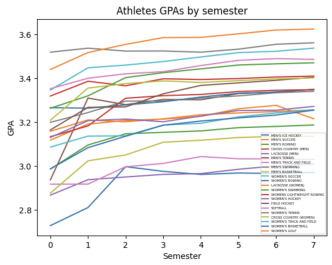
GPAgrouped

Team Name	Last Major Name	Last Cumulative BU GPA
CROSS COUNTRY (MEN)	Advertising	3.53
CROSS COUNTRY (MEN)	Business Admin & Mgt	3.51
CROSS COUNTRY (MEN)	Computer Science	3.65
CROSS COUNTRY (MEN)	Health Science	3.05666666666666700
CROSS COUNTRY (MEN)	Phil & Political Sci	3.6
CROSS COUNTRY (WOMEN)	Advertising	3.45
CROSS COUNTRY (WOMEN)	Biochem & Molec Bio	3.68
CROSS COUNTRY (WOMEN)	Computer Science	3.3
CROSS COUNTRY (WOMEN)	English	3.98
CROSS COUNTRY (WOMEN)	Health Science	3.59
CROSS COUNTRY (WOMEN)	Phil & Psychology	3.58

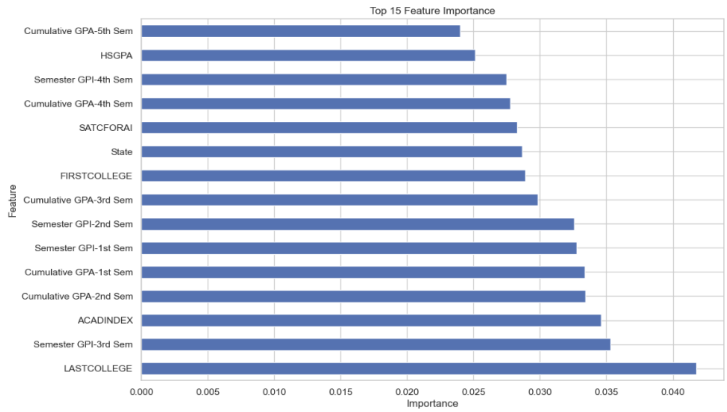
teamAVG

Team Name	Last Cumulative BU GPA
CROSS COUNTRY (MEN)	3.409756097560980
CROSS COUNTRY (WOMEN)	3.401794871794870 0
FIELD HOCKEY	3.404594594594590 0
LACROSSE (MEN)	3.008988764044940 0
LACROSSE (WOMEN)	3.214905660377360
MEN'S BASKETBALL	3.134800000000000
MEN'S ICE HOCKEY	2.970465116279070 7
MEN'S ROWING	3.185188873417720 0
MEN'S SOCCER	3.255438596491230 0
MEN'S SWIMMING	3.340204081632650 0
MEN'S TENNIS	3.348695652173910
MEN'S TRACK AND FIELD	3.029333333333330
SOFTBALL	3.486000000000000 0
WOMEN'S BASKETBALL	3.254285714285710 0
WOMEN'S GOLF	3.624545454545460
WOMEN'S HOCKEY	3.271578947368420
WOMEN'S ROWING	3.349204545454550
WOMEN'S SOCCER	3.537291666666670 0
WOMEN'S SWIMMING	3.470434782608700
WOMEN'S TENNIS	3.561666666666700
WOMEN'S TRACK AND FIELD	3.2558
WOMENS LIGHTWEIGHT ROWING	3.347704918032790

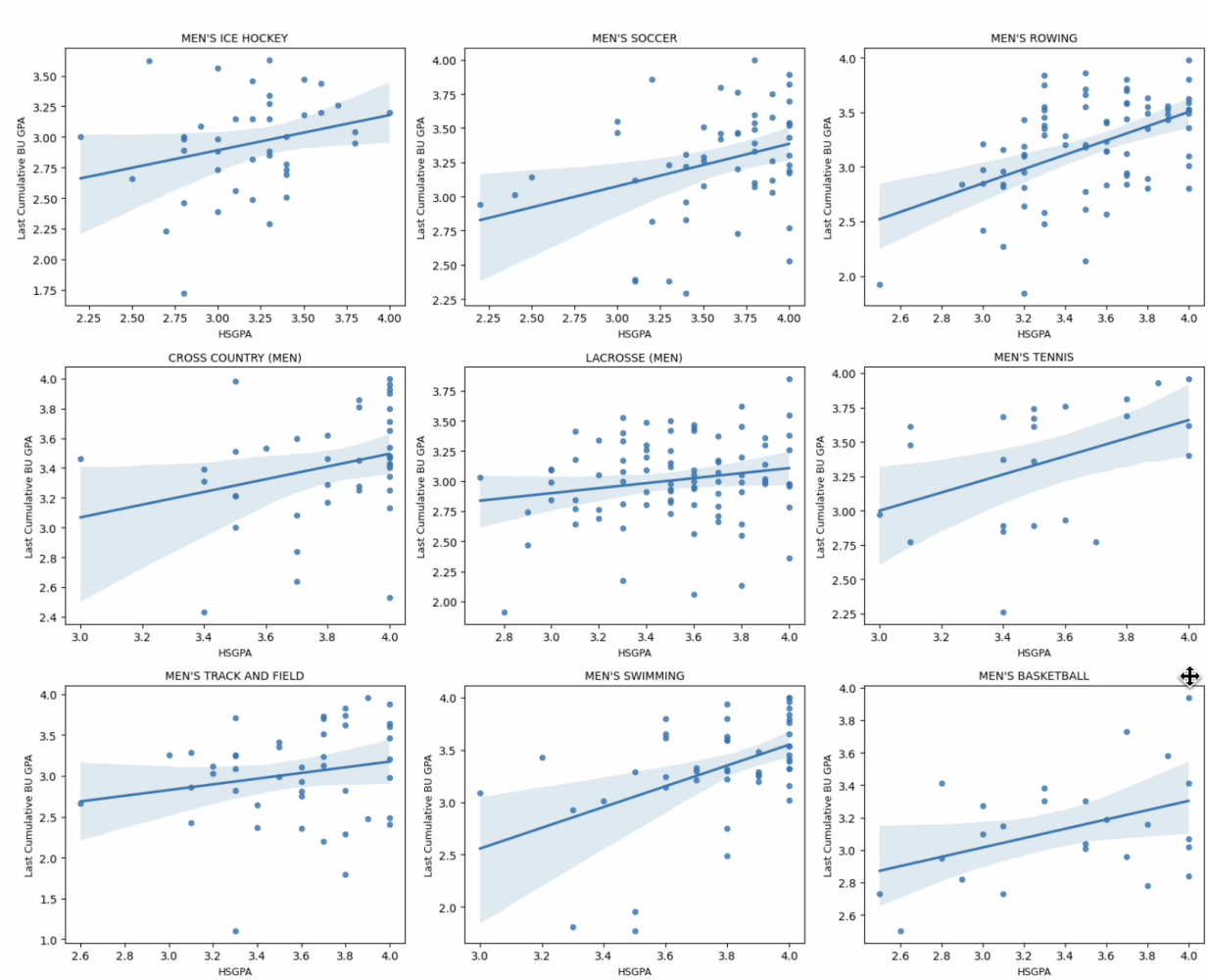
We have also done more advanced exploratory data analysis into these baseline questions, such as by examining how GPA changed over time for student athletes every semester, except this time we broke up this data team by team.



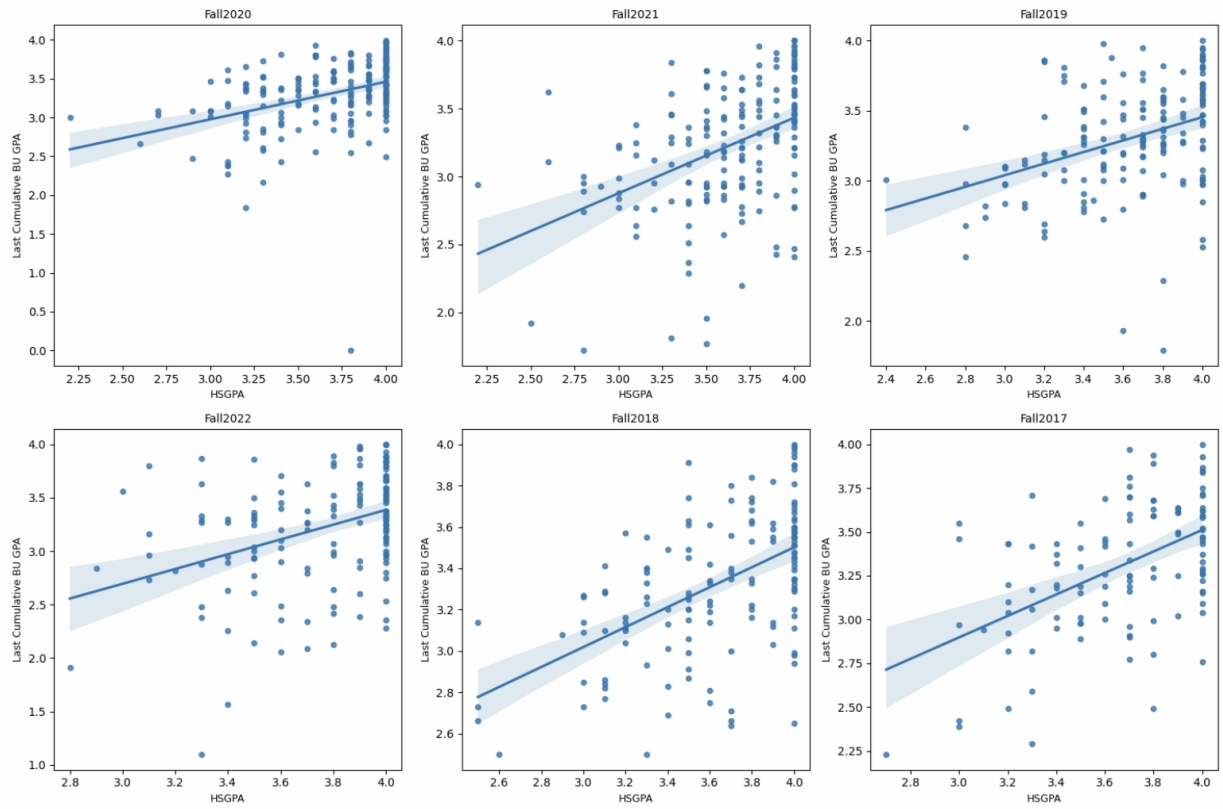
To also help start answering these questions a new binary feature “MajorChange” was added to the dataset to denote a change in a student's major, which can be used to find out the correlation between major related features and other features and predict the reasons for changing majors of student athletes. Value 1 represents students changing their majors and 0 otherwise. We also added additional similar binary features such as if said student took any standardized test or not at all, which we denoted as ‘tested’, and if a student started at BU or transferred into school, which we denoted as ‘transfer’ We used Random Forest Classifier to find the feature importance of each of these features with college GPA to see if any feature stood out or if any semester was more deterministic in determining cumulative GPA than another. The top 15 features were found and shown below.



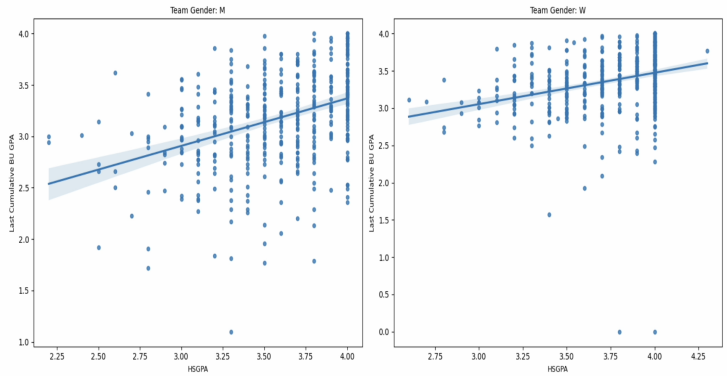
To start analyzing these extended baseline questions a graph is plotted between high school GPA and college GPA, but with respect to sports such as men's ice hockey, men's soccer, men's rowing, cross country, lacrosse, tennis, basketball and swimming.



A graph is also plotted between high school GPA and college GPA, but with respect to year enrolled such as fall 2020, fall 2021, fall 2019, fall 2022, fall 2018 and fall 2017.



Finally a graph is also plotted between female team sports and their GPAs versus male team sports. The GPA looks better for women. Also the best fit line indicates their consistency with high school GPA which is better than men.



Limitations in Extension Mini-Projects

We were also interested in how athletic performance relates to academic performance to see if these values have any relation. Our largest issue in answering this extension question is in our data collection. Our first thought was to answer this question for a subset of student athletes whose individual athletic performance we may be able to measure. This might be possible for track and field athletes. The site https://tf.tfrrs.org has all results for every track and field athlete in collegiate sports. We can compare the results of an athlete for their specific event (personal best for example) and assign a data value to it from this World Athletics Organization for the score of the athlete in that event. This score is currently the most tested way to compare track events, since it is difficult just based on raw times, distance, and results to compare unrelated events such as the high jump to the mile run. For a better example of what we are referring to here is an entry into the World Athletics Organization scoring table:

Points	100m	200m	300m	400m	500m	110mH	400mH	4x100m	4x200m	4x400m
400	12.97	26.62	42.41	59.20	1:17.85	18.57	1:08.43	51.51	1:47.34	4:05.00

This tells us that someone that runs 12.97 seconds in the 100 meters is close to equivalent from an athletic standpoint to someone that runs 18.57 seconds in the 110 meter hurdles. There are more track athletes than any other sport in the dataset, which is why it could potentially be useful to invest the time to find the score of each athlete for the event they compete in for track and field, as the more data the more useful the insights. It is also potentially useful since it is difficult to measure how an individual performs in a team sport such as soccer or hockey where an individual's athletic ability may not be as quantifiable. In the scenario where we are able to get this data we could attempt to make some sort of regression model that measures academic performance based on how an athlete performs in their sport (or potential just track and field). However, this would mean that the dataset would lack the ‘anonymous’ factor it holds since we would need to request names of athletes in order to find their athletic

results. We were awaiting a request for this information from BU Athletics and Spark, but were not able to get this information. We then decided to do some exploratory research and data analysis for each team based on whether or not the team that the student athlete is a member of won their conference championship for that season. While this is less tailored to the individuals that we are examining in the rows of the dataset we thought it could potentially yield some interesting results that are worth exploring. This data can be found on the following line of the polling dataset we were given.

League Champion		N	N	N	Y	Y	Y	Y	Y	N	N	N	N	N	N	Y	N	N	N	N	Y	N	N

However our fears over this being less tailored to the specific athletes rang true, and models we were trying to create to predict whether or not a student would be high performing academically given the other features along with their athletic performances were extremely low performing. We also tried to create models to predict if they would be high athletically performing based on the other features and GPAs, but this model performed almost equally as bad. We believe if we were able to get more specific individual performance data such as what we described earlier we could have done more with this question.

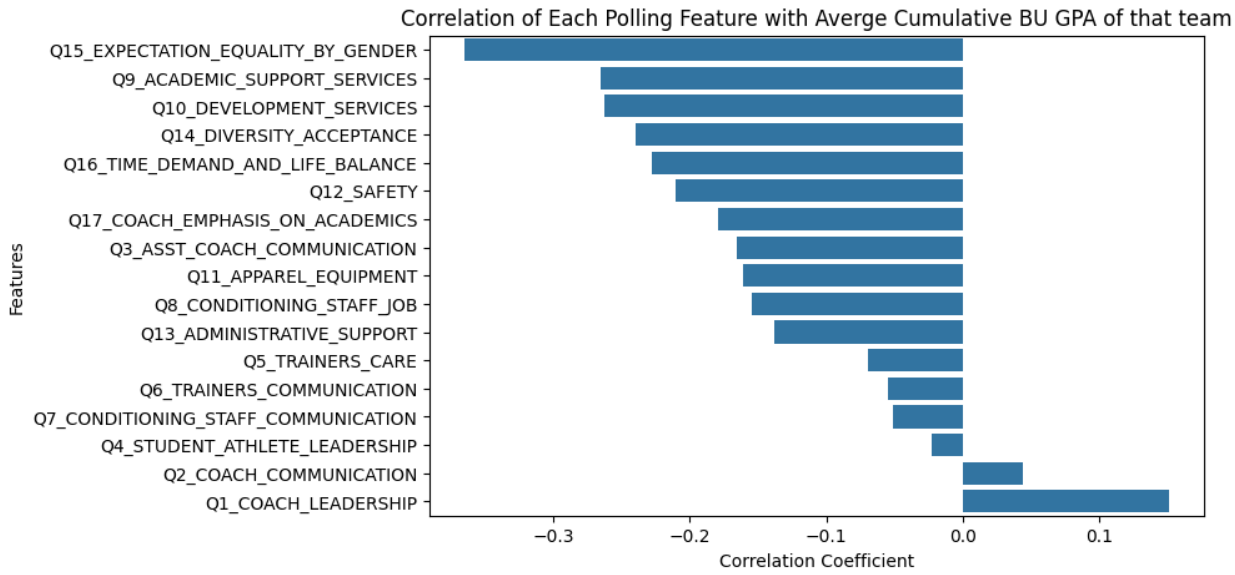
Main Extension Project: Analysis of Gender Disparity

Our final extension project that we were most focused on is examining how the survey dataset relates to academic performance. Our basic idea is to find if the data in ‘Athletics Annual Surveys Data’ correlates team-by-team to academic performance, since the polling data is only given to us as a team average for each question, but was the only additional data we were able to receive from BU Athletics and Spark outside of the data needed to answer our baselines questions We tested features from polling data, or the average responses of each team to the polling data, such as how each team voted on ‘Ability to successfully manage time demands and achieve appropriate balance between academics, athletics, and community and social

experiences' and 'Accessibility of resources for academic or student-athlete development assistance (tutoring, sports psychology, health and wellness, career services, community service)' as just two examples. We wanted to see if polls for certain questions reflect on student athletes academic performance, and if certain questions are indicative of why a team is performing well academically versus another team. We also added these polling questions as features to our model extension project that is predicting GPA.

We started at a standstill in trying to explore this question further as we are awaiting further information from BU Athletics and Spark! As seen above were given all of these teams as only abbreviations in the survey dataset. As it might seem like a simple task to link abbreviations to team names, these abbreviations are only two letters long and not the most intuitive. We tried our best figuring out which team name has which abbreviation, but since we could not waste time, we went to the BU Athletic Department to get more information on these abbreviations. Once we got these cleared up we did a mapping between the team names in this polling dataset and merge it on the team names from the athletics dataset, assigning each person from each team the average score that they voted on for that particular question, meaning we added 17 new features (the number of questions in the survey) to the dataset we have been working with.

To understand the importance of these polling attributes, we grouped the teams and took Cumulative GPA's mean for each team. We then retrieved the correlation between these mean GPAs and each of the polling scores to understand the polling features' importance.

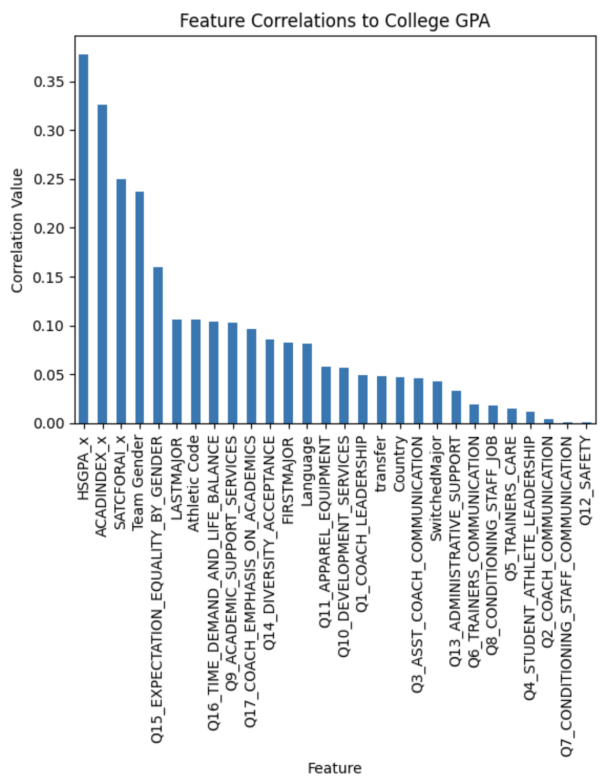


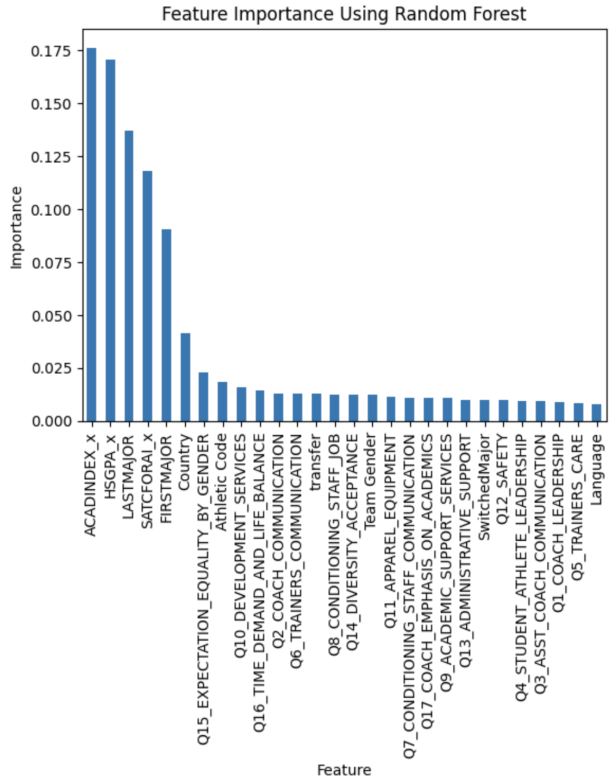
Most features appear to have a slight negative correlation with the average GPA, as indicated by the bars extending to the left. The feature with the highest positive correlation seems to be "Q1_COACH_LEADERSHIP", suggesting that higher perceptions or ratings of coach leadership are associated with a slightly higher average GPA among the teams. The feature with the highest negative correlation is “Q15_EXPECTATION_EQUALITY_BY_GENDER”. This means that higher perceptions or ratings of expectation equality by gender are associated with a lower average GPA among the teams. The exact questions that refer to these two questions respectively are “Effectiveness of head coach's leadership and direction for the team” and “Similar performance expectations, behavioral standards and support provided compared to your male/female counterparts”.

We then decided to do OLS regression using HS GPA and the important polling attributes from above as predictors, and with BU GPA as the response variable. From the p-values under, except "Q15_EXPECTATION_EQUALITY_BY_GENDER", no other attribute from polling data is a significant predictor in determining college GPA, which tells us Gender Equality is an important factor in determining academic success.

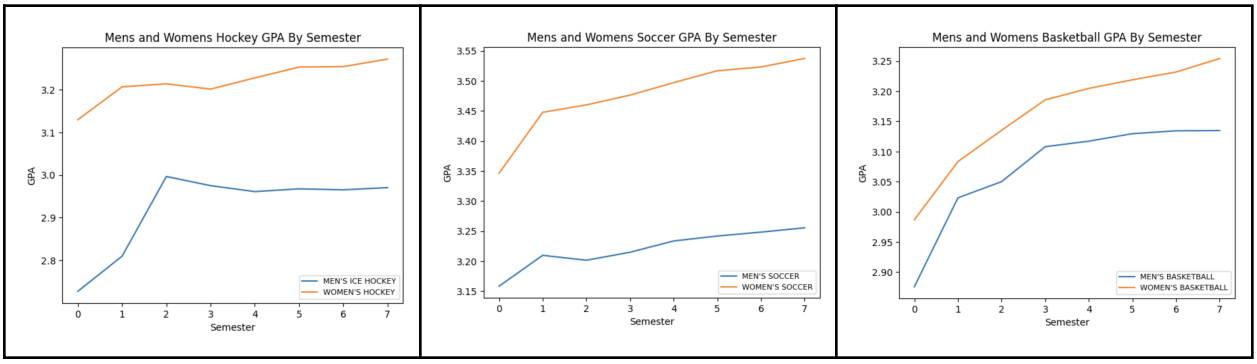
```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Last Cumulative BU GPA   R-squared:                       0.215
Model:                                OLS   Adj. R-squared:                  0.208
Method:                     Least Squares   F-statistic:                     34.33
Date:                    Tue, 05 Dec 2023   Prob (F-statistic):           9.59e-37
Time:                            18:54:08   Log-Likelihood:                -295.61
No. Observations:                     761   AIC:                             605.2
Df Residuals:                         754   BIC:                             637.7
Df Model:                               6
Covariance Type:                nonrobust
==========================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                                1.4312      0.333      4.297      0.000       0.777       2.085
Q15_EXPECTATION_EQUALITY_BY_GENDER  -0.1049      0.035     -2.966      0.003      -0.174      -0.035
Q9_ACADEMIC_SUPPORT_SERVICES        -0.0237      0.054     -0.441      0.659      -0.129       0.082
Q10_DEVELOPMENT_SERVICES             0.0044      0.090      0.049      0.961      -0.173       0.182
Q1_COACH_LEADERSHIP                  0.0173      0.032      0.541      0.588      -0.045       0.080
HSGPA                                0.4075      0.046      8.882      0.000       0.317       0.498
SATCFORAI                            0.0006      0.000      4.931      0.000       0.000       0.001
==============================================================================
Omnibus:                      375.576   Durbin-Watson:                   1.911
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5285.690
Skew:                          -1.863   Prob(JB):                         0.00
Kurtosis:                      15.362   Cond. No.                     3.42e+04
==============================================================================
```

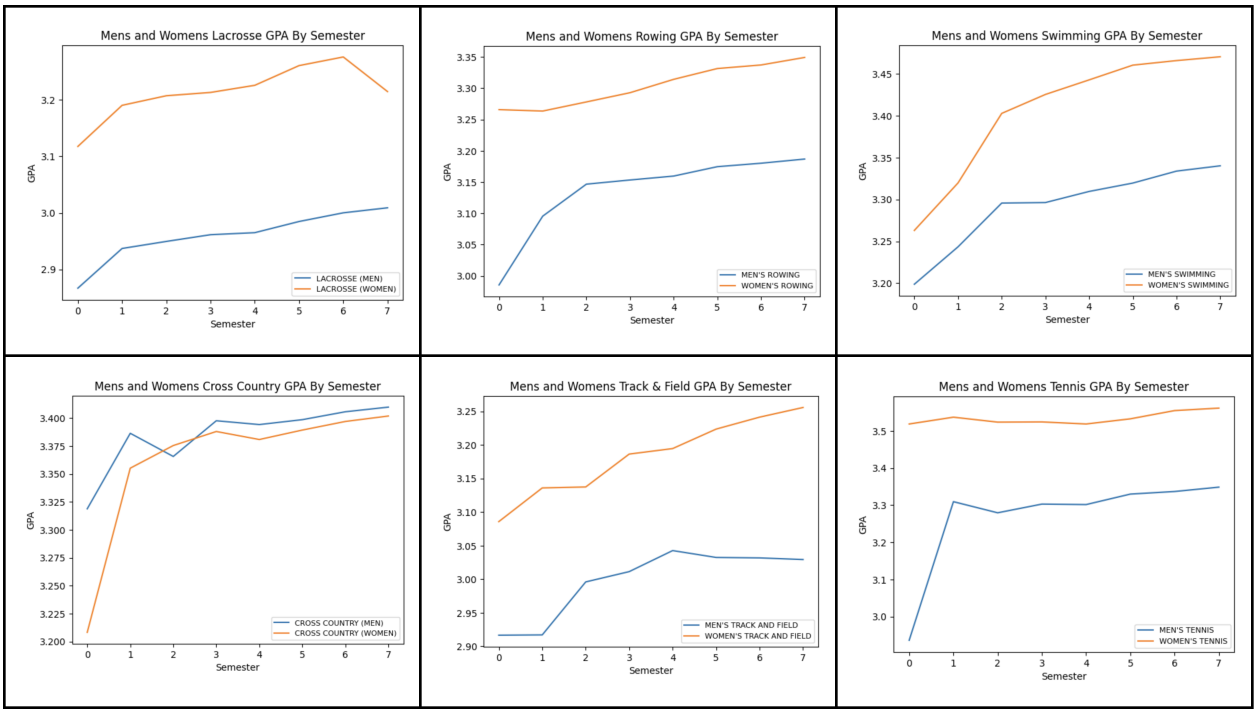
We remembered from the BU athletics extension proposal that “mini-projects are meant to let you explore interesting questions related to your core project. These might be sociological questions (are there systemic inequalities ingrained in the athletics system?)” Following this idea, we decided to do a deeper dive into an analysis into gender and why this survey question is the most correlated to academic performance. We started with trying to graph the correlation between all features and college GPA to see how any of the survey questions related, though our focus was the one related to gender related to collegiate GPA. We also tried to weight the feature importance by using a Random Forest in order to see what features would weigh the most heavily in a model when using all combined athletics data and survey features to predict GPA.



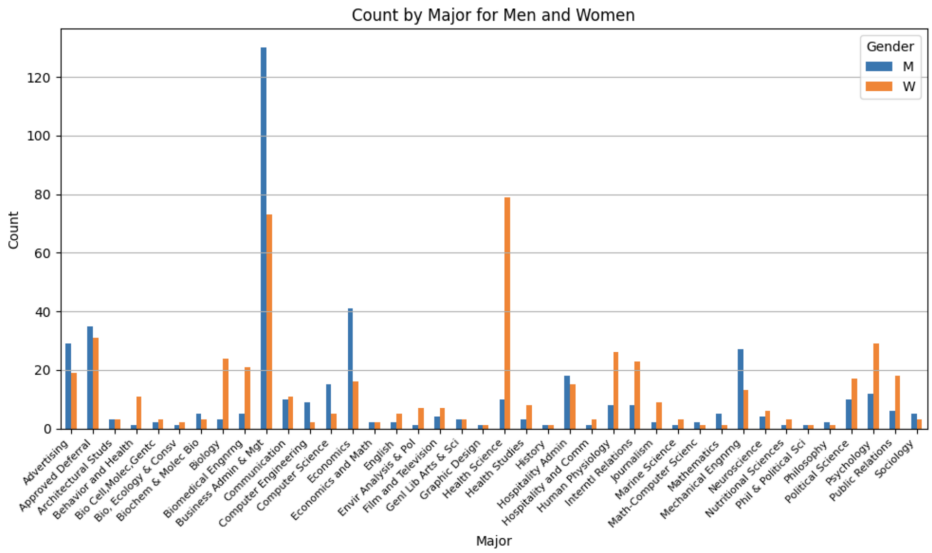


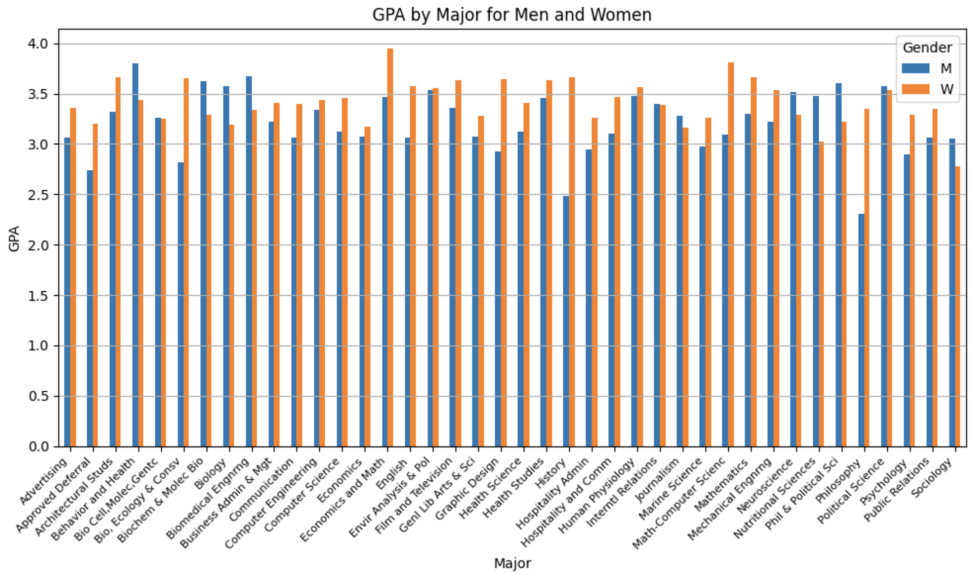
We then did some more exploratory research into men's team sports versus women's team sports to see if there was any correlation in the data with this survey feature. We first graphed all of the teams GPAs against each other.



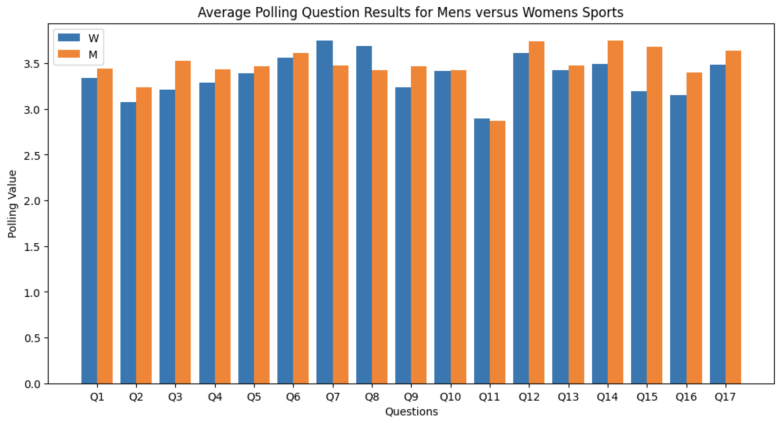


From this we can clearly see that women outperform their respective men's teams for almost every single sport. We were curious what majors men and women are choosing to take the most of so we plotted bar charts of the counts of each major. We can see from this chart below that the most popular male majors for athletes are Business, Economics, Advertising, Mechanical Engineering, and Hospitality while the most popular majors for females are Health Science, Business, Psychology, Human Physiology and Biology. The majors that males take much more than females are Math, Computer Engineering, Computer Science, Economics, and Mechanical Engineering, while the majors that females take much more than males are Behavior and Health, Biology, Health Science, Environmental Analysis and Politics, and Journalism. Many of the male dominated majors seem more STEM related than the female dominated majors, which might lead people to believe that this is the reason for GPA discrepancies, since these majors are usually considered “more difficult”. However, when comparing this to majors' average GPA across genders we can see that women (the blue columns) outperform men (the orange column) in 28/40 majors.

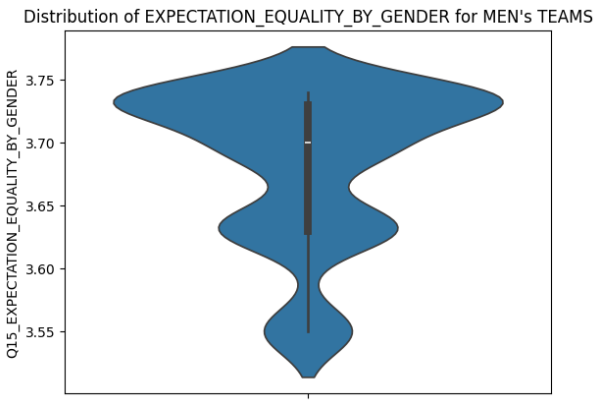


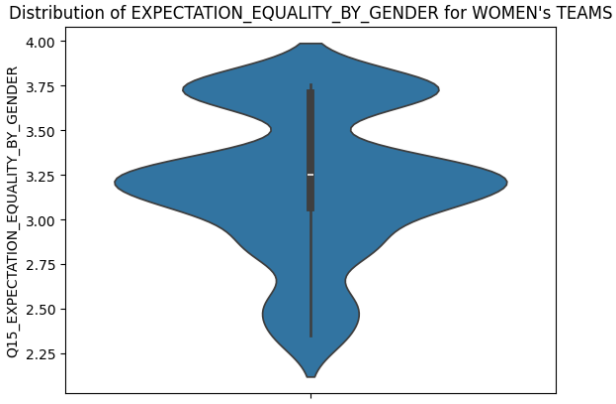


After seeing the obvious differences in academic performance between mens and womens athletes, we were curious to see if these athletes are mindful of this gap, which we thought we could visualize by seeing which survey questions mattered the most to female and male athletes.



It seems from this graph that most of the athletes across genders share the same mentalities, but unsurprisingly the survey question that has the most different results is the question related to gender equality. To dive deeper in this question we visualized the distribution of "EXPECTATION_EQUALITY_BY_GENDER" for men's and women's sports teams

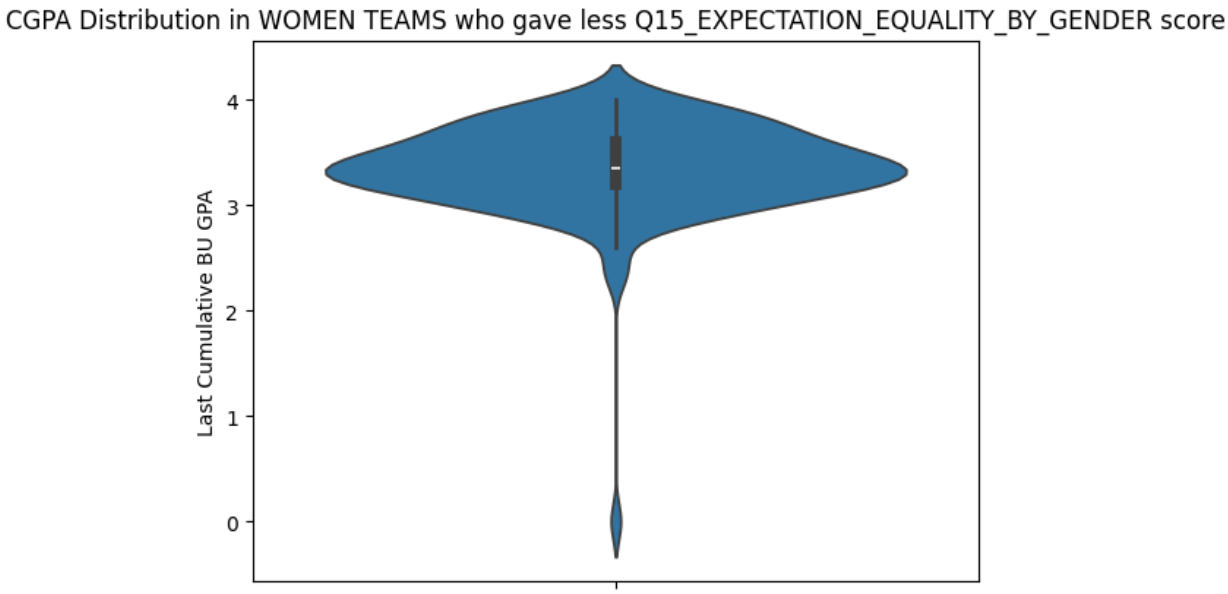


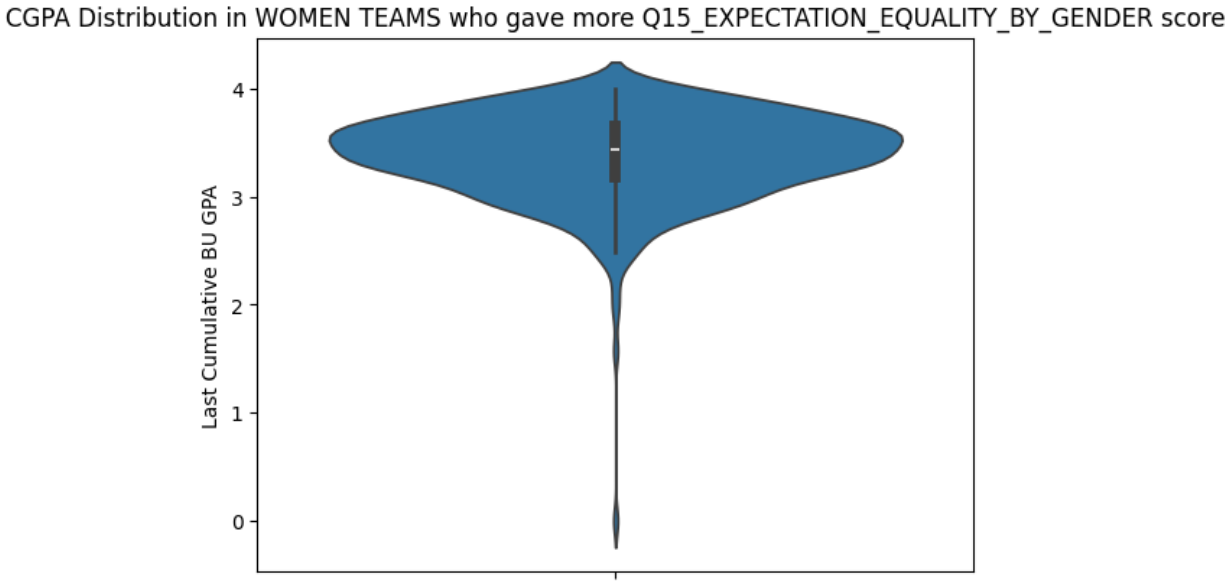


For the men's teams, the higher median score of approximately 3.70 indicates that they generally perceive a higher level of equality in treatment compared to their opposite-gender counterparts within the athletic department. The narrow distribution suggests that this perception is fairly consistent across the men's teams. On the other hand, the women's teams have a median score of around 3.24, which is significantly lower than that of the men's teams. This suggests that women's teams, on average, perceive less equality in treatment compared to their male counterparts. The wider distribution of scores for the women's teams points to a more

varied perception of equality; some teams feel they are treated with a level of equality close to that of the men's teams, while others perceive much less equality.

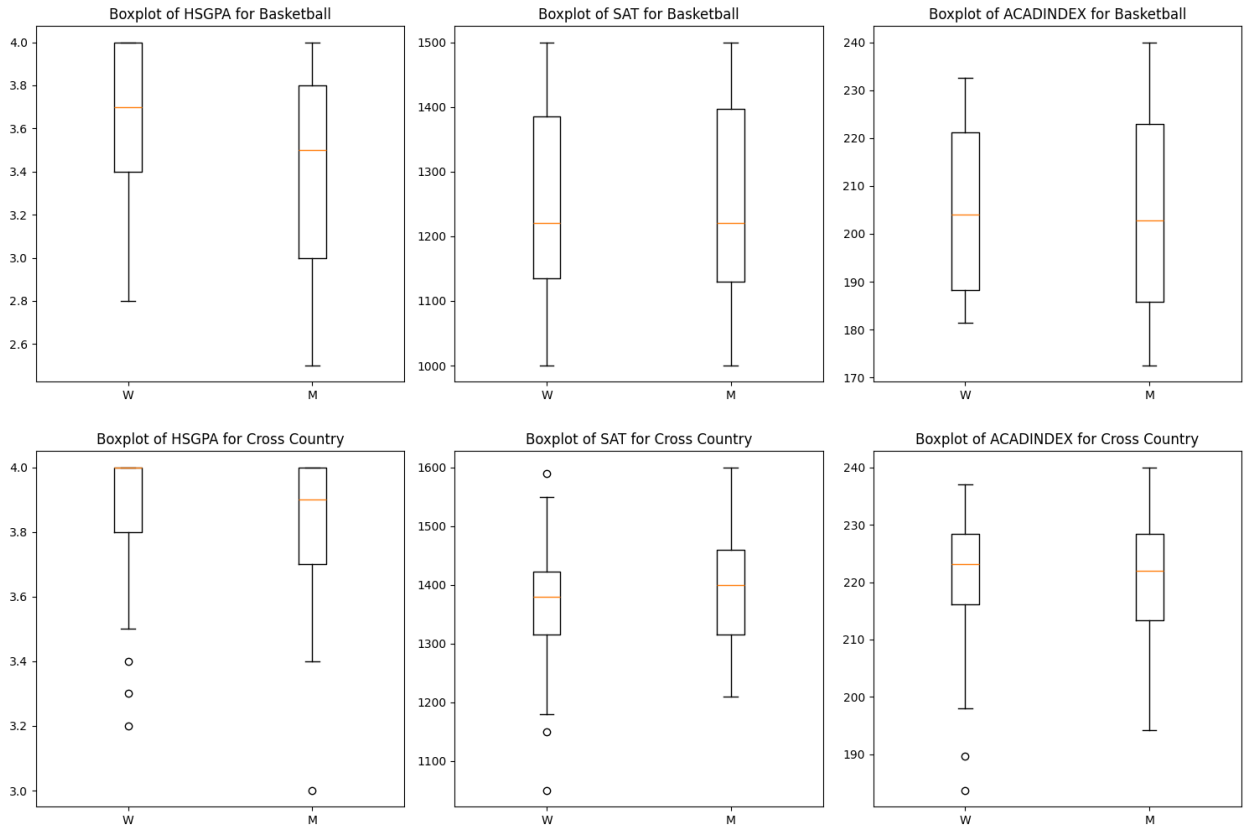
In summary, the data infers that there is a noticeable difference in the perception of gender equality in treatment within the athletic department, with men feeling more equally treated than women. The variation in the women's scores also highlights that this perception of inequality is not uniform and varies more widely among different women's teams.

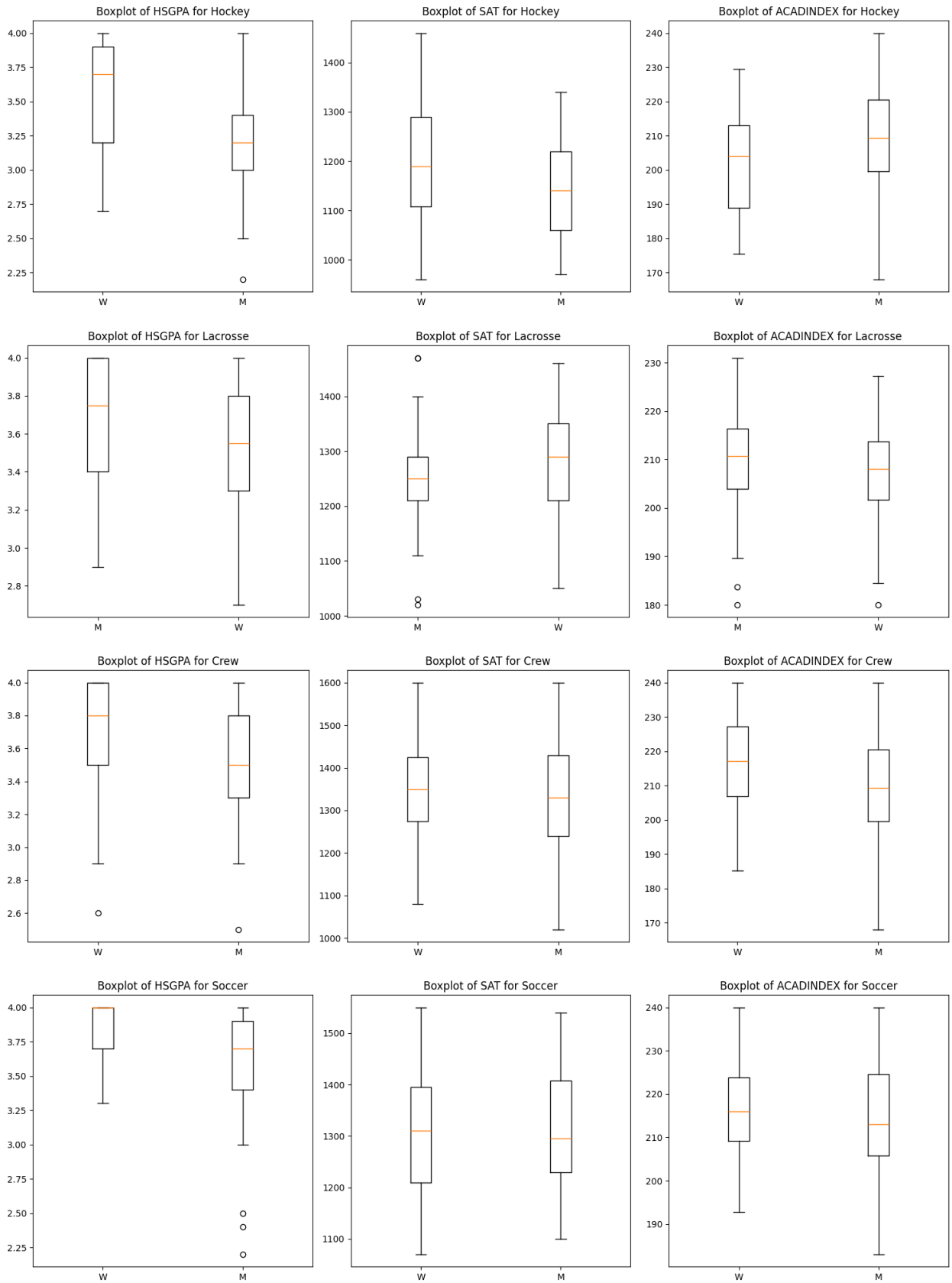


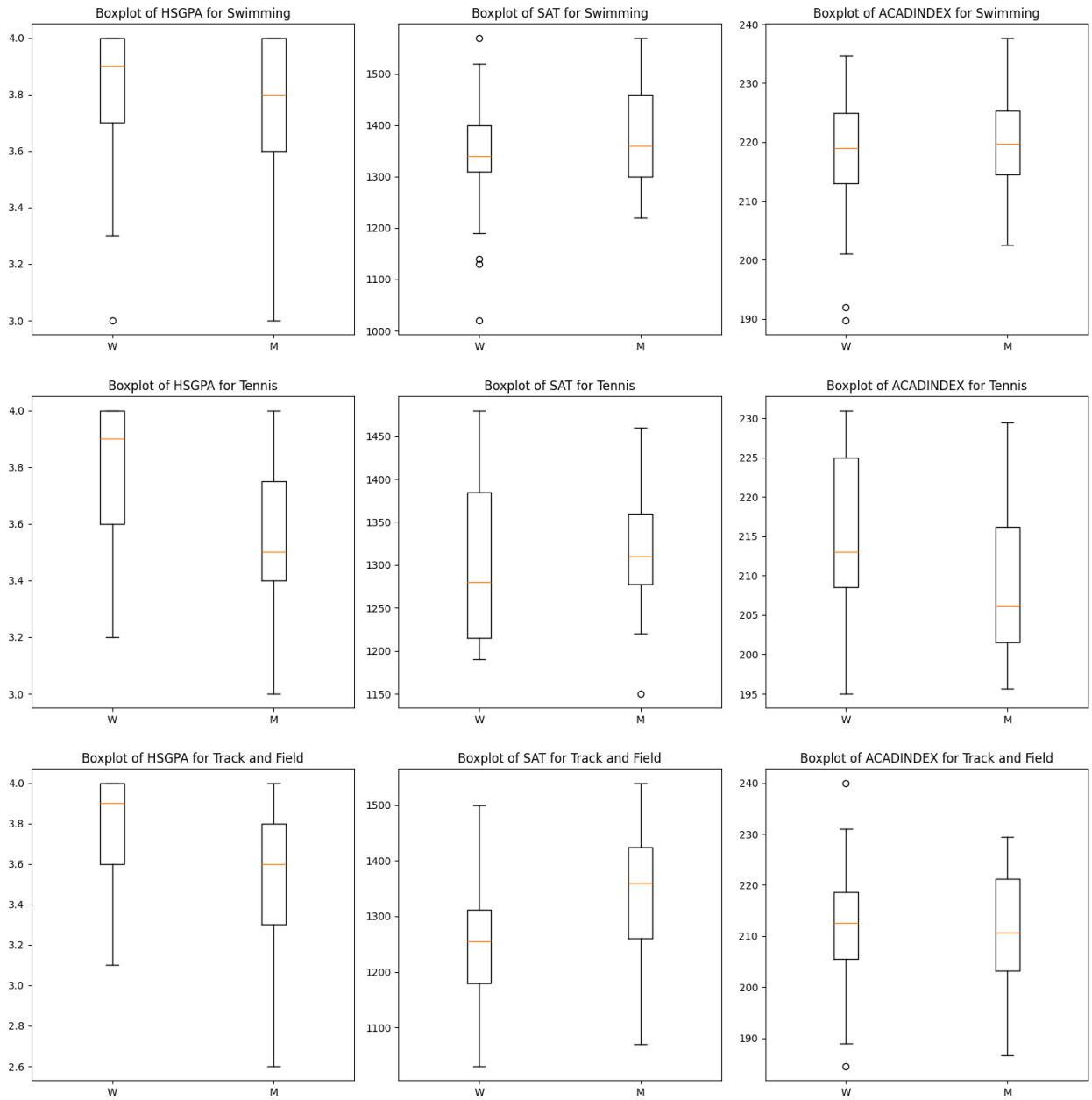


It appears that the CGPA distribution among women's teams is fairly consistent regardless of their perception of gender equality within the athletic department. Both groups—those who perceive less equality and those who perceive more—exhibit a range of GPAs with a similar shape and median value. This would suggest that there isn't a strong correlation between the grades of female athletes and their perception of gender equality in their respective sports teams, but rather that all women's sports teams see gender inequality.

From the above analyses we can see that women's sports teams perform better in college and also believe that they are treated unfairly. This might be correlated with the idea that they believe they might need to outperform their male counterparts to receive equal recognition, a common equal rights belief throughout history. We wanted to see if this outperformance was also prevalent in just admissions data as visualized through the box plots below.

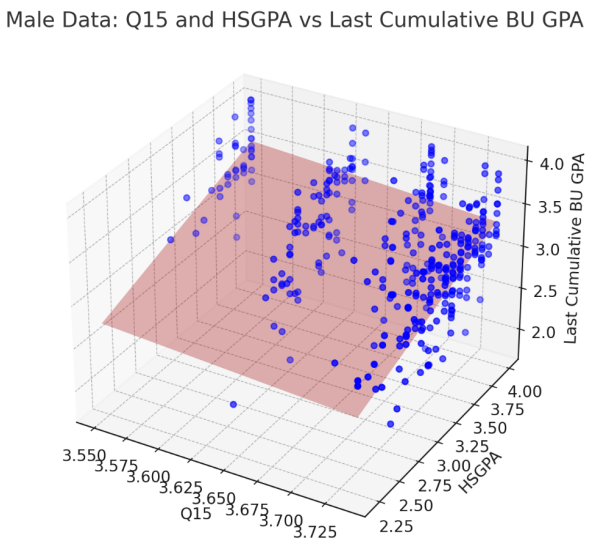


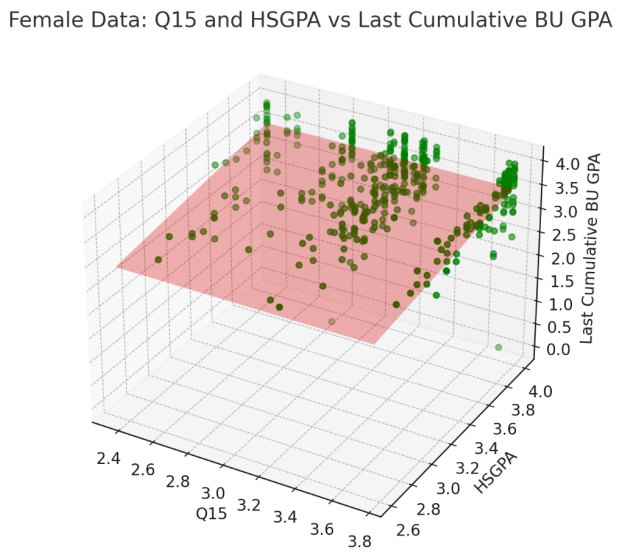

Boxplot of HSGPA for Hockey
Boxplot of SAT for Hockey
Boxplot of ACADINDEX for Hockey
Boxplot of HSGPA for Lacrosse
Boxplot of SAT for Lacrosse
Boxplot of ACADINDEX for Lacrosse
Boxplot of HSGPA for Crew
Boxplot of SAT for Crew
Boxplot of ACADINDEX for Crew
Boxplot of HSGPA for Soccer
Boxplot of SAT for Soccer
Boxplot of ACADINDEX for Soccer
W
M



The women's teams boxplots for Academic Index reflect a higher median than their male counterparts for 7/9 sports. Academic Index takes all high school admissions stats into account, meaning HS GPA, rank, and test scores, so this is the most complete metric of a student athlete's academic profile before college. While the numbers are not staggeringly different, the men's academic index lower quartile is less for 8/9 team sports. These numbers reflect potential inequalities in the recruiting standards for mens versus women's sports, and may hint at more leniency for mens sports over women's sports for recruiting standards

We also performed regression analysis with 'Q15_EXPECTATION_EQUALITY_BY_GENDER' and 'HSGPA' as the independent variables to predict 'Last Cumulative BU GPA' for each gender.





The constancy of the graph representing women in academics, despite their perception of unequal treatment, suggests a complex interplay of motivation, expectations, and cultural factors. It's plausible that female athletes, facing perceived or real inequalities, might be more driven to perform at a higher level as a form of empowerment or as a way to level the playing field to receive comparable recognition or opportunities. For male athletes, the perception of more fair treatment and perhaps a different set of cultural and motivational factors might lead to different academic and/or athletic outcomes. This doesn't mean less effort or capability, but possibly different priorities or pressures set forth by society systematically.

The model's findings alone neither conclusively support nor dispute our claim. However, they do highlight important dynamics regarding gender perceptions and academic/athletic performance especially among women and this warrants further investigation and could be indicative of underlying issues in how female athletes are treated or perceive their treatment.

Final Conclusions from Main Extension

Based on our findings from our last extension, we can conclude that there are gender-based disparities within BU Athletics. Key indicators include the dissimilarity between performance and the behavioral and support standards between women and their male counterparts. Just the fact alone that women so highly value the gender equality survey question as opposed to men should be an indication that there are widespread beliefs by more than half the student athlete body that must be examined within BU Athletics. Additionally, academic excellence skews in favor of women with 28 of their majors having a higher average GPA than men out of the 40 majors. In addition, there is also more leniency granted to men in admission standards for athletics. There is also shown to be more consistent academic outperformance of women's teams over men's in similar sports. The lack of variability in our regression model further supports the argument that there are societal externalities that are pressuring women into performing better and consistently, highlighting the higher standards they are held at. After further analysis and alternative model development, no other features outside of gender and admissions statistics seem to have much relevance towards the model's performance, which means this issue is extremely prevalent in the overall problem statement of the project, which is to examine how and why student athletes perform academically the way that they do. We conclude that we observe certain patterns within Boston University Athletics that suggest a need for further examination and discussion regarding gender equality.

Individual Contributions

Robert: Cleaned and imported the data for the initial project and the extension project from the excel files given; data exploration and graphing done with the initial datasets; cleaning and preprocessing the data; adding more feature variables than the dataset originally had to increase feature space then worked on merging the athletics dataframe and the survey dataframe; did my baseline questions assigned to me and came up with relevant plots/graphs to answer these questions; scoped out and came up with ideas and conclusions for the extension project; did work on the correlation between variables for the main extension project, used random forest for weighting variables, made all bar graphs for extension doing a deeper dive into mens versus women's sports across majors, did the exploratory analysis into mens versus women's sports with graphs above, did dive into admissions data with box-plot analysis; worked on normalizing and getting the data processed so that it could fit properly into the model for that mini extension project; worked with Spark! And BU Athletics and went to their offices directly for data; worked with our TPM Matthew and Spark! and was at lab every week besides 1; wrote up/put together all besides one of the Scrum Reports and all of the Deliverable deliverable reports and all of the presentation videos and final report

Bargav: Answered baseline questions assigned and helped in extension making GPA distributions for different sub-categories, made more pertinent graphs concerning how mens and womens sports differ in terms of GPA and how they vote in the survey dataset and discovered gender bias using model evaluation, did exploratory analysis with distribution plots

Mounika: Answered baseline questions assigned and helped in extension making GPA distributions for different sub-categories, helped Bargav with extension and exploratory analysis

Jeff: Answered baseline questions by plotting appropriate graphs assigned (i.e. box plot and a heatmap) and concluded that they are not good indicators for our extension project. Moved on to extension after baseline questions were answered which gave us preliminary visions on what our extension project should focus on (or ignore). After the baseline extension was answered, developed alternate models to possibly identify extenuating variables that explain gender disparity (no alternative models were found given limited data). Attended labs with team lead to provide updates to TPM. Created a “model” with a primary focus on understanding the relationship between perceptions of gender equality in athletic programs (as indicated by Q15), High School GPA, and Last Cumulative BU GPA. Conducted separate regression analyses for male and female teams using the given variables above. Created 3D plots to visualize the regression models and the variability between the variables for both male and female teams. Generated residual plots for each gender group to evaluate the fit of the regression models and to identify any patterns in the residuals.

Shangzhou: Answered baseline questions assigned and helped make models for extensions, evaluated models, and worked with the team in trying to find best feature selection and model evaluation methods.