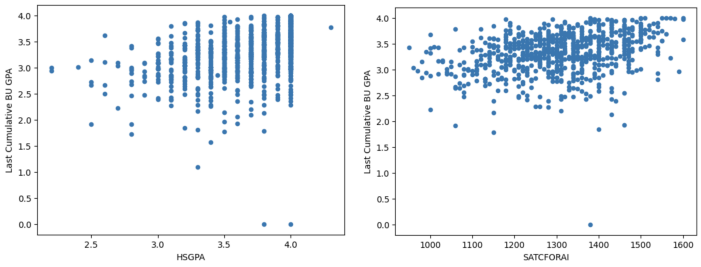
The main problem to examine for BU Athletics has been the relationship between admission statistics, such as test scores (SAT and ACT), Academic Index, and high school GPA, and the ongoing academic performance of student-athletes at BU. The athletics department is looking to gain deeper insights into the students that are members of their program to see if they can better support their athletes. To achieve the project's objectives and respond to the key questions, we needed the essential student data including some mentioned earlier. We were given SAT/ACT, high school GPA, and Academic Index (a score from 60-240 assigned to students who are applying to colleges based on academic performance) for each student. We were also given some other demographic information including first language, where they are from, their race, and their team gender. In terms of collegiate statistics we were given each student's sports participation, major, whether or not they switched majors, their cumulative GPA, and their GPA for each semester at BU. With 994 records in the provided BU dataset in CSV format, we analyzed for correlations and created pertinent visualizations. In addition to this data set we were also provided a table of student athlete polling data that we intend to use for our extension project. This polling data is given for 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.

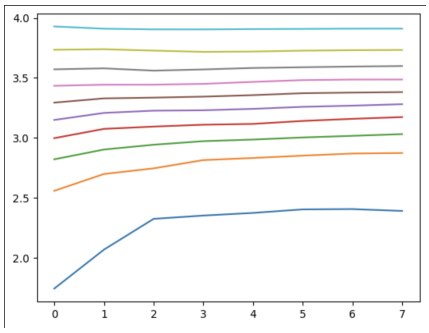
We did our data cleaning with the original data frame we were given with the individual stats of each student athlete, not with the polling data. We processed the initial Excel table we were given as a CSV then cleaned up the table with pandas. For example in the CSV we deleted the column ‘Student’ which was a unique student ID number along with a few other columns that were all null values including the 12th-22nd semester GPA values for students which were 0.0 in the dataframe for all students. We also renamed some columns to easier names such as ‘English Language Primary Language Of’ to ‘Language’, along with a few other

values to make easier use of. 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 potential 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’. The polling data did not have any missing values, however we may have to normalize the data in these values that are currently numbers 1-4. This may be necessary for our extension project for model performance, however we have not changed these values yet since the polling data was not necessary for answering the baseline questions. We did not have to collect any sort of data, but currently have requested further data from Spark! And BU Athletics that may be helpful for potential extension projects ideas.

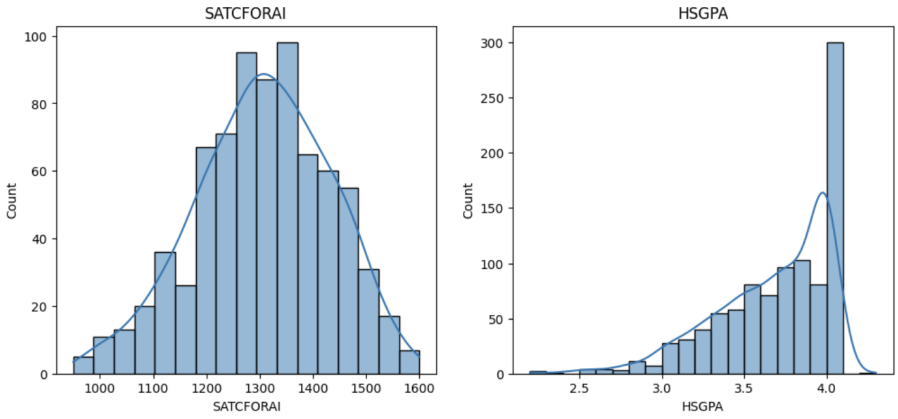
Our first baseline question was to answer how SAT/ACT and high school GPA translate to collegiate academic success. We learned that high school stats are not great indicators of how students perform academically at BU. We tried to make a model that would predict how the student would perform in college. The Decision Tree classifier that we tried to use does not end up being a great model for predicting if athletes will fall into the A, B, C, or lower range of average GPAs. This, along with our scatter plots of HS GPA v. 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 well defined trends between these measurements.

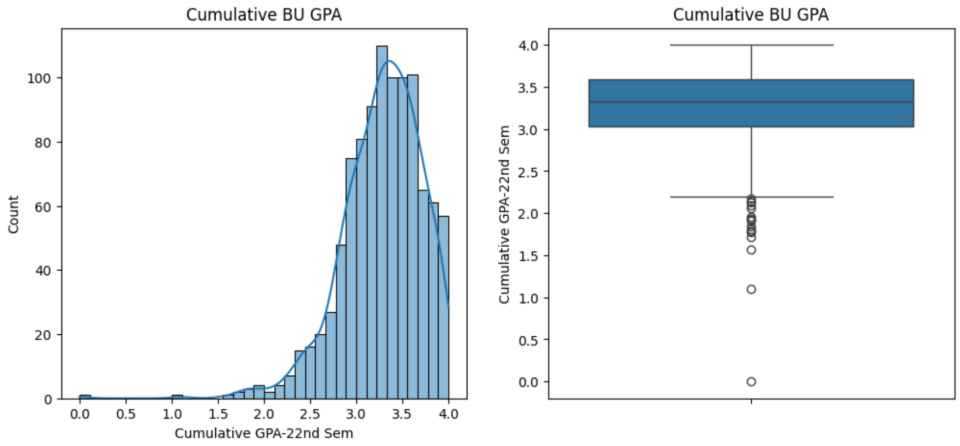


Our next question to answer was whether student athletes improve their GPA throughout their college years. After computing clusters using KMeans based on SAT, College GPA, and High school GPA, to find if certain features were indicative of how a student athlete performed over their academic career, we did not find any surprising results. 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.



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.





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 since the distribution looks similar.

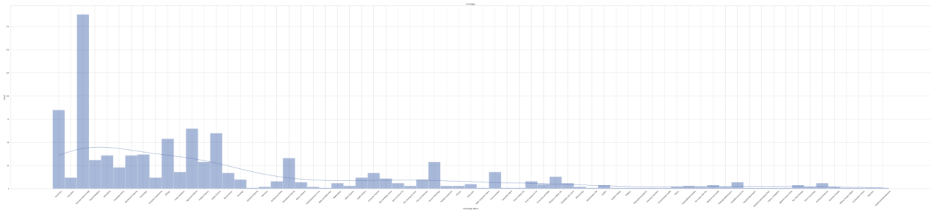


Figure1: the distribution of first major over all students

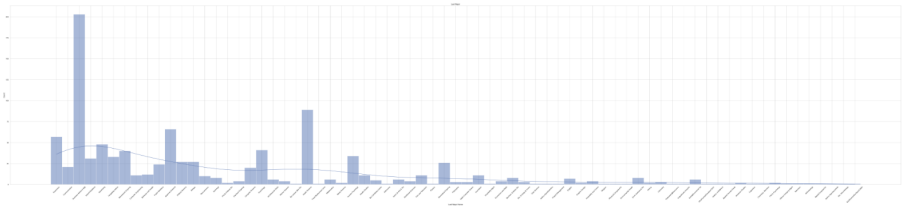
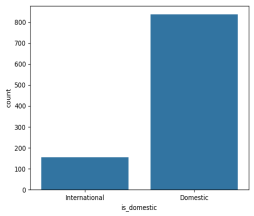


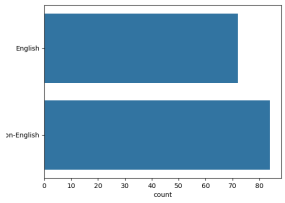
Figure 2: the distribution of last major over all students

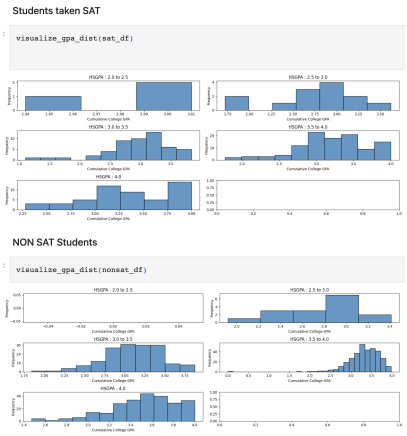
The percentage of student athletes staying in the same major was calculated by the number of students who had the same major over total students. Based on the calculation, it is easy to find that 76.6% of students stay in the same major.

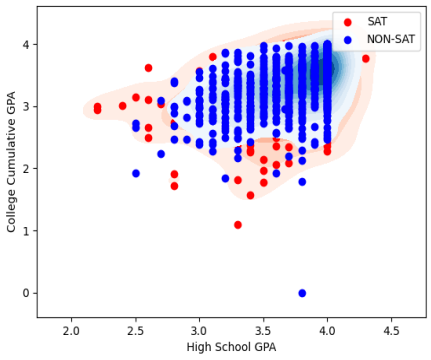
Continuing in answering questions we found that the percentage of International students is 15 percent and the percentage of domestic students is 84 percent. Clearly there is more percentage for domestic students then the international students

We also answered if English being the primary language of the country impacted the students' performance. If you ignore the outliers, the distribution is pretty much the same, suggesting to us that English as a primary language or not, is not making any necessary difference for athletes in grades. In Fact 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 SAT and non SAT students with similar HS GPA, the students who took SAT have a little decrease in their GPA in college whereas students who did not take SAT are very consistent with their score.

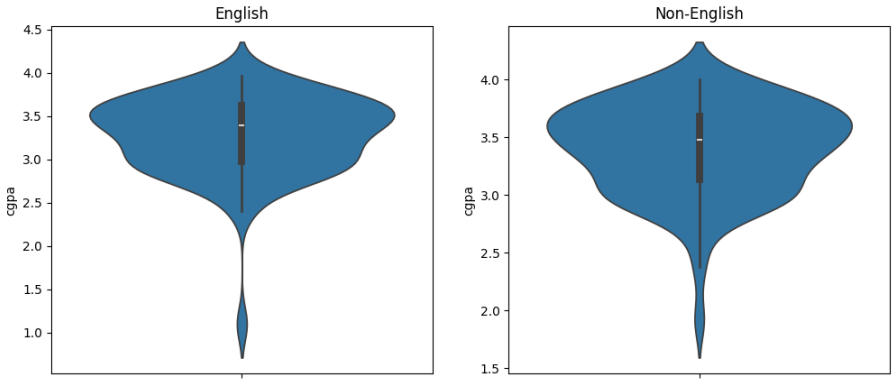




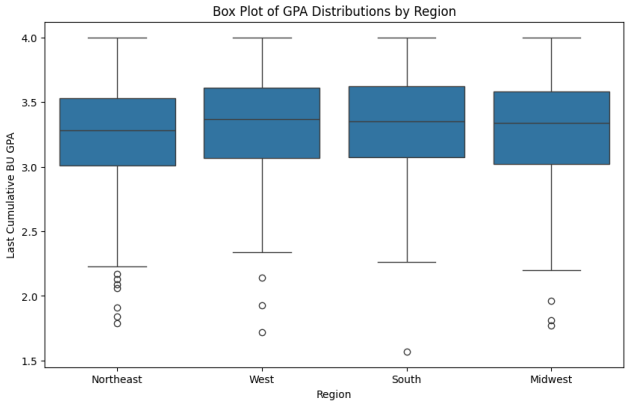


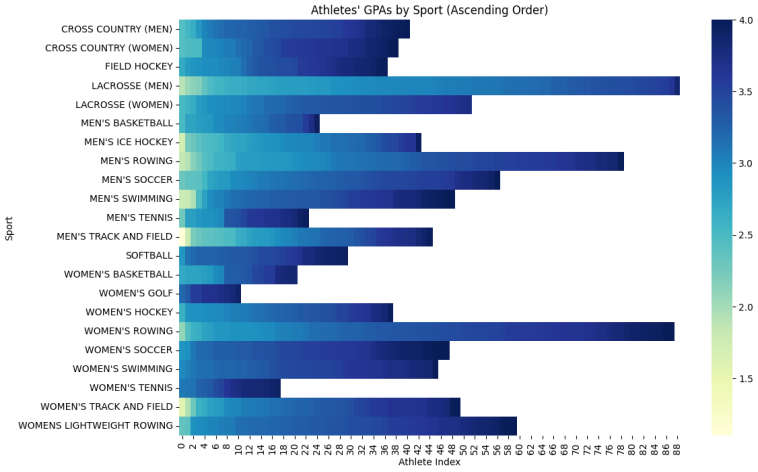


To find out whether English as a primary language or not impacts international students' grades, the CGPA distribution of both groups is observed, and the distribution is pretty much the same. This tells us that English as a second language is not affecting students's grades.



To compare students' consistency with their high school grades, HSGPA is divided into buckets of 0.5 size, and the distribution of college cumulative GPA is plotted over them. 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. When observed (as seen in the above section), Students who didn't submit SAT scores turned out to have a little more consistency in their performance than those who submitted the SAT. We see that the average GPA in the Northeast is slightly lower than the other three major regions, but it is generally more consistent.





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).

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 believe are worth asking as extensions to the baseline questions. We have seen various trends in the data we have visualized and collected, and we have also seen that this data set has a unique lack of trends as well, leading us to question why this might be.

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

experimentation of features and of model type and model parameters are going to discover if we can create some sort of regression model to predict the GPA of students.

We have also formulated more questions as responses to our baseline questions. For example many of the baseline questions ask about the “average student athlete”. We are now going to separate some of our baseline questions and ask about them on a team-by-team and major-by-major basis. This may potentially give us insights into what majors best fit certain teams, if teams are frequently switching majors and why that might be the case, and if certain teams are performing better than others and why that might be the case. Some of these baseline questions we will be looking further into for these groupings also include if these athletes have significant fluctuations in their GPA during their career at BU and how their high school statistics compare to their academic performance at BU. We will also be expanding on one of our baseline questions specifically, which asked us to compare high school and college GPA. We will expand the correlation plots between these values across various groups based on different categories such as ‘Entering Term’, ‘Team Gender’, ‘FIRSTMAJOR’, etc. So far, this analysis has only been asked to be conducted for students who submitted SAT scores versus those who did not.

To start answering these baseline questions we have represented the data we have 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 have a new CSV that we created 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.

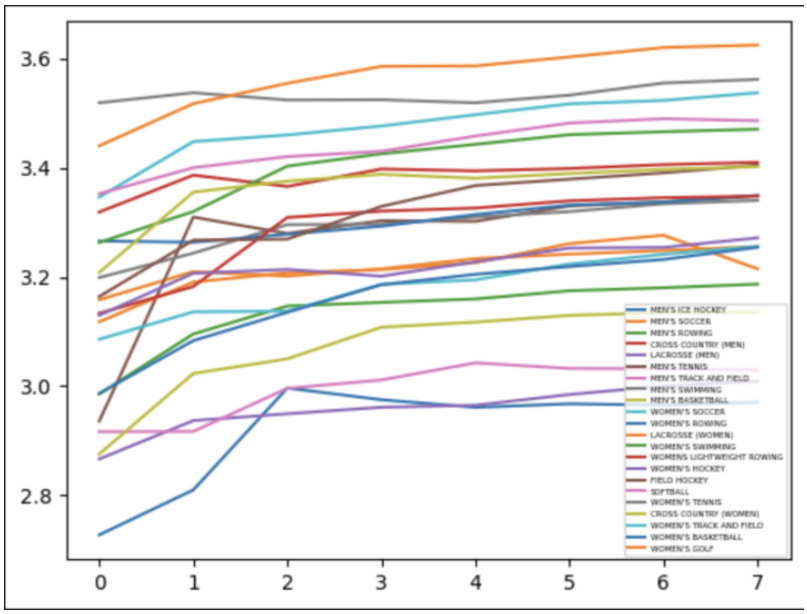
GPAgrouped

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	3.0566666666666700
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

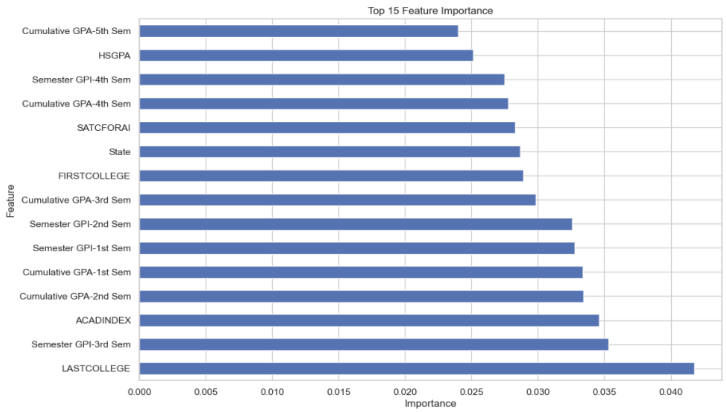
teamAVG

Team Name	Last Cumulative BU GPA
CROSS COUNTRY (MEN)	3.409756097560980
CROSS COUNTRY (WOMEN)	3.401794871794870
FIELD HOCKEY	3.404594594594900
LACROSSE (MEN)	3.008988764044940
LACROSSE (WOMEN)	3.214905660377360
MEN'S BASKETBALL	3.134800000000000
MEN'S ICE HOCKEY	2.970465116279070
MEN'S ROWING	3.185188987341772
MEN'S SOCCER	3.256438698491230
MEN'S SWIMMING	3.340204918032790
MEN'S TENNIS	3.348695652173910
MEN'S TRACK AND FIELD	3.028333333333330
SOFTBALL	3.486000000000000
WOMEN'S BASKETBALL	3.254285714285710
WOMEN'S GOLF	3.624545454545460
WOMEN'S HOCKEY	3.271578947368420
WOMEN'S ROWING	3.349204545454550
WOMEN'S SOCCER	3.537291666666700
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 started our new more advanced exploratory data analysis into these 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 used Random Forest Classifier to find the feature importance except those major related features. The top 15 features were found and can be used for future exploration.



We are 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 data set, 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 data set would lack the ‘anonymous’ factor it holds since we would need to request names of athletes in order to find their athletic results. We are currently awaiting a request for this information from BU Athletics. If we are unable to get this information we could potentially try 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 it could potentially yield some interesting results that are worth exploring. This data can be found on the following line of the polling data set we were given.

With the additional heat-map of GPAs across sports, we can try to determine which group of student-athletes might need the most academic assistance. With our proposal up to this point, we can try to identify at-risk sports. For starters, we can identify the sports with the lowest average GPAs, highest variability, or a significant number of athletes with GPAs below a certain threshold (e.g. below 2.0). Furthermore, we should also consider other factors that might affect academic performance. Based on the extension proposal we have, we came up with three factors

- Training Schedule: Some sports may have more demanding training schedules, leaving less time for academics.
- Season Timing: Sports with seasons that coincide with crucial academic periods might impact GPAs.

- Team Culture: Investigate if there's a correlation between team culture and academic performance (we can use the polling data for this but the layer of anonymity will make this particular factor a challenging one to decipher).

Our final extension project that we are most focused on is examining how the polling 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. We are going to test 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 are going 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 will also try to add these polling questions as features to our regression model extension project that is predicting GPA.

We are currently at a standstill in trying to explore this question further as we are awaiting further information from BU Athletics and Spark! As seen above we are given all of these teams as only abbreviations in the polling data set. 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. Once we have the information on which team name has which abbreviation, we are going to do a mapping between the team names in this polling data set and merge it on the

team names from the athletics data set, assigning each person from each team the average score that they voted on for that particular question, meaning we are going to be adding 28 new features (the number of questions in the survey) to the data set we have been working with.

Rob - I did my two questions and the original data cleaning/pre-processing. I have also done the work not specifically centered around the baseline questions as the team lead. I compiled the video for this deliverable and came up with the extension project and proposal. I also handled coming up with the challenges that our team had to report on at Checkpoint A. As part of this deliverable I also handled the sections concerning the Problem Statement, Data Cleaning and collection steps; the Exploratory Data Analysis; the extension proposal; and the visualization and insights for the extension proposal. I have started work on adding new features to our data set for our extension projects and have started uploading this code to our branch on GitHub. I am in communication with our TPM Matthew on getting the additional data we are hoping to collect for our extension project.

Bargav - I did the questions: Whether English as a primary language or not affects the international students' grades, and also how students are consistent in academic performance - HSGPA vs College Cumulative GPA, also comparing the same with students who submit SAT and students who didn't.

Shangzhou - I did two questions assigned to me and plotted figures to support my calculations. I also plotted other figures to see the percentage of students changing majors over semesters. and demographic distributions of majors. I explored more to find possible reasons which cause the students to change their majors. All the figures can be seen in the github repo.

Mounika - Questions related to all stats of High School GPA and BU GPA. Also the distribution of Domestic and International students, along with the English native speakers percentage in the international students.

Jeff - I illustrated the difference in academic performance with a heat map of GPAs across all sports. With this, I proposed an extension project to determine which group of student-athletes might need the most academic assistance. I also tried to get data that is otherwise not readily available from Spark/TPM/BU Athletics in order to answer my extension questions.