**NCAA APR Database**

**Exploratory Analysis**

Jack Huff, jhuff@bellarmine.edu

Brett Tully, btully@bellarmine.edu

**I.** **INTRODUCTION**

Our data set includes a list of all NCAA schools and their sports teams. It includes data on what conference and division they are in, as well as the sports teams respective APR score, retention rate, eligibility, and the amount of 4 year players they have. Our data set was found on Kaggle. We chose this dataset because we were interested in academic performance and its relationship to specific conferences, sports, and teams.

**II.** **DATA SET DESCRIPTION**

Our data set has a total of 6511 rows and 57 columns. The 57 columns have various data types including int64, float64, and object.

**Table 1: Data Types and Missing Data**

| *Variable Name* | *Data Type* | *Missing Data (%)* |
| --- | --- | --- |
| V1 SCHOOL\_ID | nominal/int64 | 0% |
| V2 SCHOOL\_NAME | nominal/object | 0% |
| V3 SCHOOL\_TYPE | nominal/int64 | 0% |
| V4 ACADEMIC\_YEAR | ordinal/int64 | 0% |
| V5 SPORT\_CODE | nominal/int64 | 0% |
| V6 SPORT\_NAME | nominal./object | 0% |
| V7 NCAA\_DIVISION | nominal/int64 | 0% |
| V8 NCAA\_SUBDIVISION | nominal/int64 | 0% |
| V9 NCAA\_CONFERENCE | nominal/object | 0% |
| V10 FOURYEAR\_ATHLETES | interval/int64 | 0% |
| V11 FOURYEAR\_SCORE | interval/int64 | 9.6% |
|  |  |  |
| V12 FOURYEAR\_ELIGIBILITY | interval/float64 | 9.6% |
| V13 FOURYEAR\_RETENTION | interval/float64 | 9.6% |
| V14 2014\_ATHLETES | interval/int64 | 9.6% |
| V15 2014\_SCORE | interval/int64 | 9.6% |
| V16 2014\_ELIGIBILITY | interval/float64 | 9.6% |
| V17 2014\_RETENTION | interval/float64 | 9.6% |
| V18 2013\_ATHLETES | interval/int64 | 9.6% |
| V19 2013\_SCORE | interval/int64 | 9.6% |
| V20 2013\_ELIGIBILITY | interval/float64 | 9.6% |
| V21 2013\_RETENTION | interval/float64 | 9.6% |
| V22 2012\_ATHLETES | interval/int64 | 9.6% |
| V23 2012\_SCORE | interval/int64 | 9.6% |
| V24 2012\_ELIGIBILITY | interval/float64 | 9.6% |
| V25 2012\_RETENTION | interval/float64 | 9.6% |
| V26 2011\_ATHLETES | interval/int64 | 9.6% |
| V27 2011\_SCORE | interval/int64 | 9.6% |
| V28 2011\_ELIGIBILITY | interval/float64 | 9.6% |
| V29 2011\_RETENTION | interval/float64 | 9.6% |
| V30 2010\_ATHLETES | interval/int64 | 9.6% |
| V31 2010\_SCORE | interval/int64 | 9.6% |
| V32 2010\_ELIGIBILITY | interval/float64 | 9.6% |
| V33 2010\_RETENTION | interval/float64 | 9.6% |
| V34 2009\_ATHLETES | interval/int64 | 9.6% |
| V35 2009\_SCORE | interval/int64 | 9.6% |
| V36 2009\_ELIGIBILITY | interval/float64 | 9.6% |
| V37 2009\_RETENTION | interval/float64 | 9.6% |
| V38 2008\_ATHLETES | interval/int64 | 9.6% |
| V39 2008\_SCORE | interval/int64 | 9.6% |
| V40 2008\_ELIGIBILITY | interval/float64 | 9.6% |
| V41 2008\_RETENTION | interval/float64 | 9.6% |
| V42 2007\_ATHLETES | interval/int64 | 9.6% |
| V43 2007\_SCORE | interval/int64 | 9.6% |
| V44 2007\_ELIGIBILITY | interval/float64 | 9.6% |
| V45 2007\_RETENTION | interval/float64 | 9.6% |
| V46 2006\_ATHLETES | interval/int64 | 9.6% |
| V47 2006\_SCORE | interval/int64 | 9.6% |
| V48 2006\_ELIGIBILITY | interval/float64 | 9.6% |
| V49 2006\_RETENTION | interval/float64 | 9.6% |
| V50 2005\_ATHLETES | interval/int64 | 9.6% |
| V51 2005\_SCORE | interval/int64 | 9.6% |
| V52 2005\_ELIGIBILITY | interval/float64 | 9.6% |
| V53 2005\_RETENTION | interval/float64 | 9.6% |
| V54 2004\_ATHLETES | interval/int64 | 9.6% |
| V55 2004\_SCORE | interval/int64 | 9.6% |
| V56 2004\_ELIGIBILITY | interval/float64 | 9.6% |
| V57 2004\_RETENTION | interval/float64 | 9.6% |

**III.** **Data Set Summary Statistics**

Here, we took a look at the major college sports conferences and their APR scores. There are 25 teams in each conference, and we did a simple statistical analysis on each conference and their scores.

**Table 2: Summary Statistics for Pivot (name of dataset)**

| *Variable Name* | *Count* | *Mean* | *Standard Deviation* | *Min* | *25th* | *50th* | *75th* | *Max* |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Atlantic Coast Conference | 25 | 980.269 | 8.320 | 960.945 | 975.745 | 981.151 | 985.162 | 992.948 |
| Atlantic Sun Conference | 25 | 966.136 | 12.370 | 932.636 | 966.136 | 966.136 | 975.575 | 983.909 |
| Big 12 Conference | 25 | 967.764 | 10.716 | 945.418 | 961.681 | 967.763 | 976.181 | 985.747 |
| Big East Conference | 25 | 983.851 | 9.106 | 960.072 | 978.363 | 986.584 | 989.893 | 996.090 |
| Big Ten Conference | 25 | 978.441 | 8.462 | 958.441 | 974.141 | 979.745 | 985.212 | 991.654 |
| Pac-12 Conference | 25 | 973.336 | 9.937 | 949.939 | 968.803 | 974.128 | 981.132 | 991.136 |
| Southeastern Conference | 25 | 971.081 | 11.650 | 947.344 | 965.836 | 971.081 | 978.370 | 997.363 |
| The Ivy League | 25 | 993.721 | 2.779 | 986.590 | 992.170 | 993.954 | 995.784 | 998.077 |

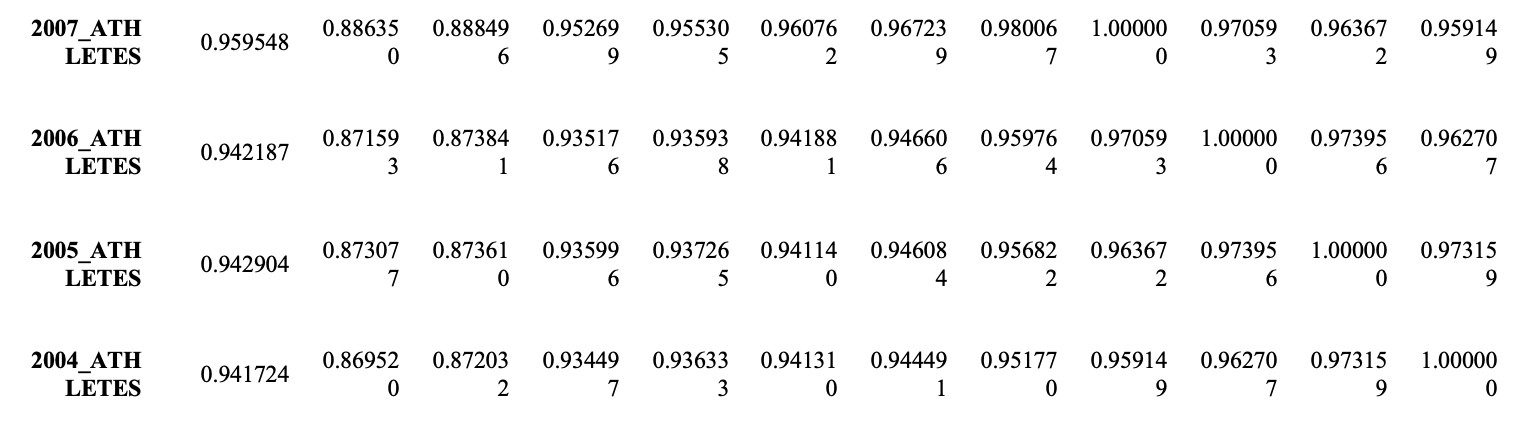
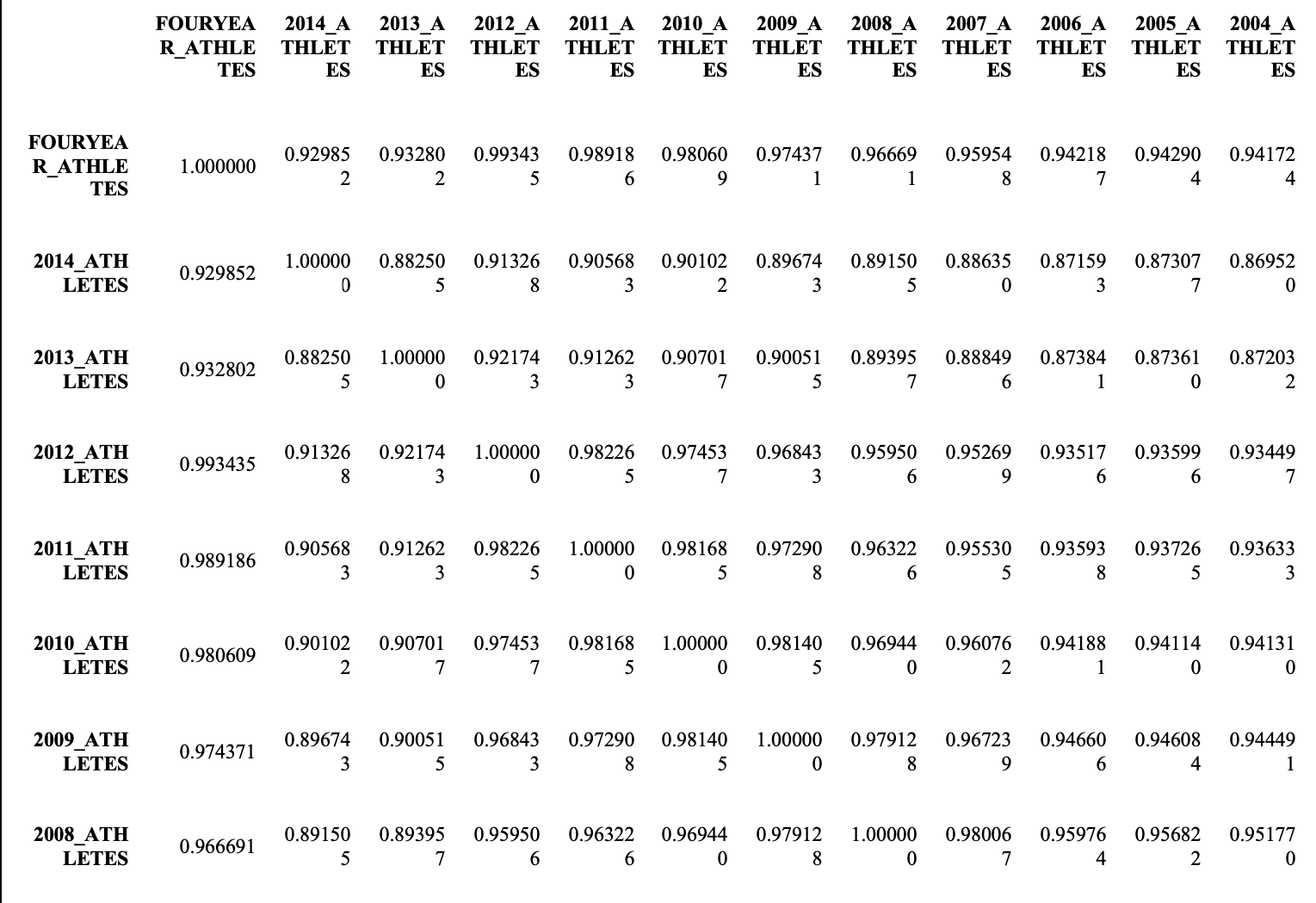
*We wanted to narrow down the Conferences to the main conferences that people generally pay more attention to.*

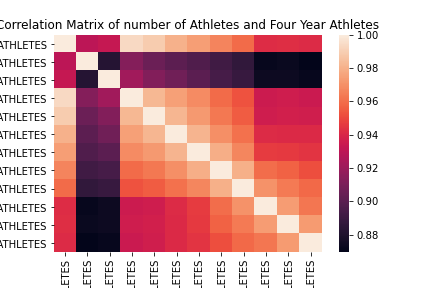
There should be a table for **EACH** categorical variable.

Table 3: Proportions for Sports (n=yyy)

| *Category: Sports Teams* | *Frequency* | *Proportion (%)* |
| --- | --- | --- |
| *Men's Basketball* | *322* | *5.91%* |
| *Women's Basketball* | *321* | *5.89%* |
| *Women's Cross Country* | *273* | *5.01%* |
| *Women's Volleyball* | *303* | *5.56%* |
| *Women's Track, Outdoor* | *282* | *5.17%* |
| *Women's Soccer* | *290* | *5.32%* |
| *Women's Tennis* | *260* | *4.77%* |
| *Women's Track, Indoor* | *274* | *5.03%* |
| *Men's Cross Country* | *209* | *3.83%* |
| *Men's Golf* | *189* | *3.47%* |
| *Baseball* | *270* | *4.96%* |
| *Women's Softball* | *257* | *4.72%* |
| *Men's Track, Outdoor* | *232* | *4.26%* |
| *Women's Golf* | *189* | *3.47%* |
| *Men's Tennis* | *211* | *3.87%* |
| *Men's Track, Indoor* | *212* | *3.89%* |
| *Football* | *228* | *4.19%* |
| *Men's Soccer* | *183* | *3.36%* |
| *Women's Swimming* | *165* | *3.03%* |
| *Men's Swimming* | *114* | *2.09%* |
| *Women's Lacrosse* | *71* | *1.30%* |
| *Women's Rowing* | *65* | *1.19%* |
| *Women's Field Hockey* | *71* | *1.30%* |
| *Men's Wrestling* | *67* | *1.23%* |
| *Men's Lacrosse* | *49* | *.89%* |
| *Women's Gymnastics* | *60* | *1.10%* |
| *Men's Ice Hockey* | *56* | *1.02%* |
| *Women's Ice Hockey* | *25* | *.46%* |
| *Women's Water Polo* | *25* | *.46%* |
| *Women's Bowling* | *16* | *.29%* |
| *Women's Fencing* | *11* | *.20%* |
| *Men's Water Polo* | *17* | *.31%* |
| *Men's Volleyball* | *17* | *.31%* |
| *Mixed Rifle* | *15* | *.27%* |
| *Men's Fencing* | *12* | *.22%* |
| *Men's Gymnastics* | *14* | *.26%* |
| *Women's Skiing* | *10* | *.18%* |
| *Men's Skiing* | *9* | *.16%* |

Table 4: Correlation Table/Tables





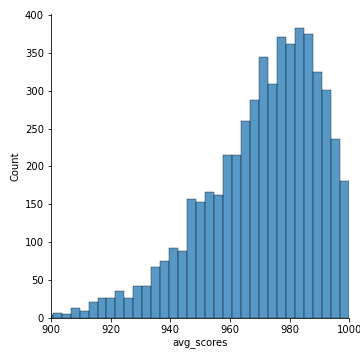
**IV.** **DATA SET GRAPHICAL EXPLORATION**

After we narrowed down our data set for what we were interested in looking at, we decided to visualize the data. We graphed distributions, scatterplots, barcharts, and heatmaps. The visualization of the data helped us see some interesting patterns.

**Titles of some of the charts are cut-off/not showing in the document**

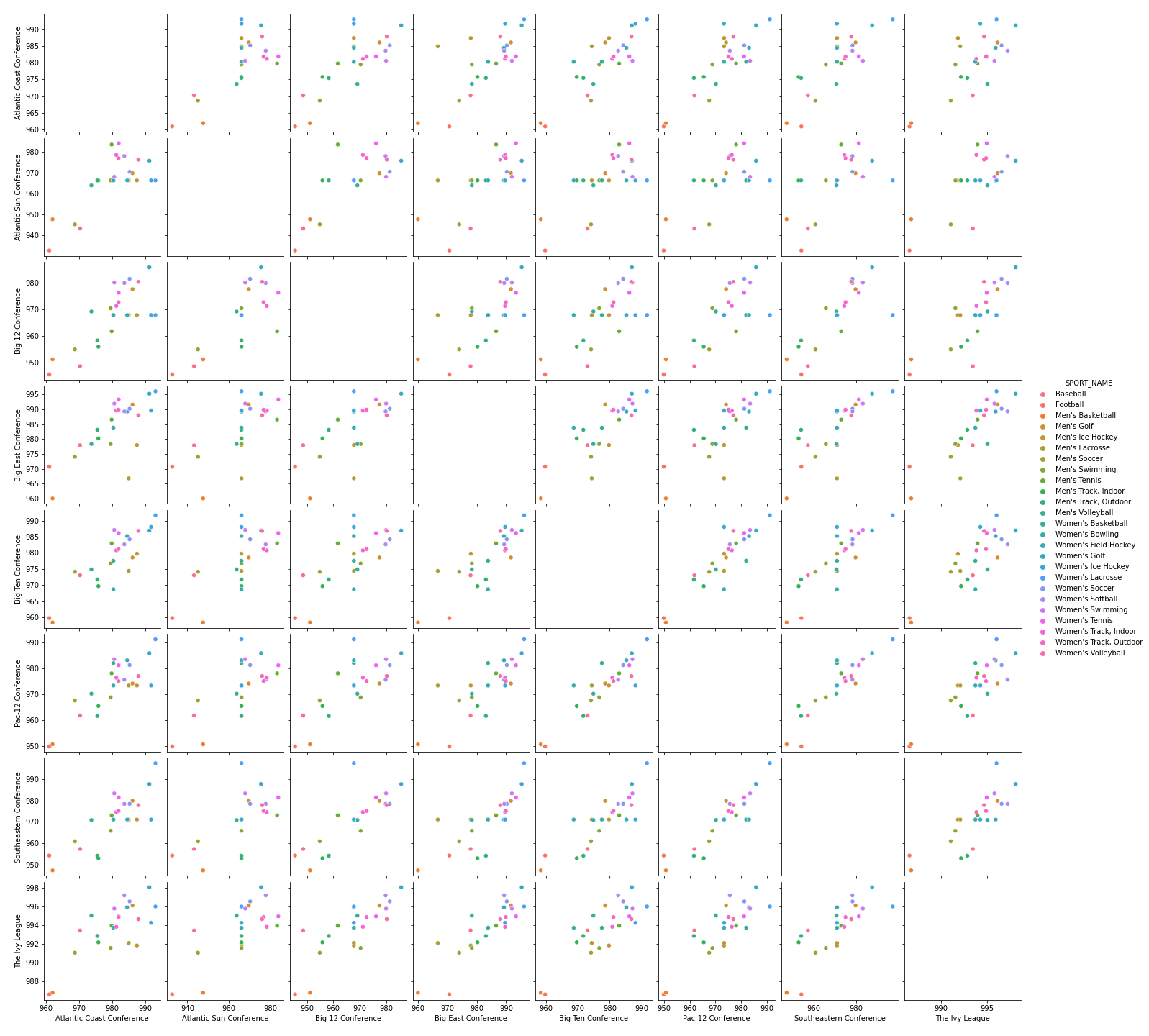
*A.* *Distributions*

**Figure 1: Histogram of Distribution from Dataset**

**

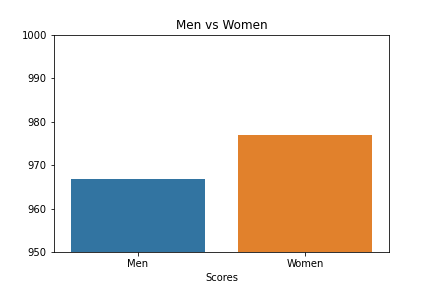
*B.* *ScatterPlots / Pairwise Plots (continuous variables)*

**Figure 2: Pairwise Plots for Sports and comparing each Conference to each other**

**

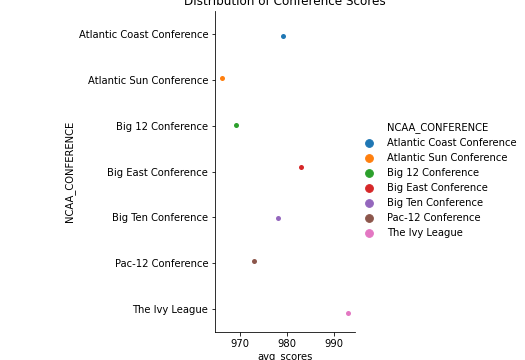
*C.* *Barcharts (categorical variables)*

**Figure 3: Comparison between Men and Women scores**

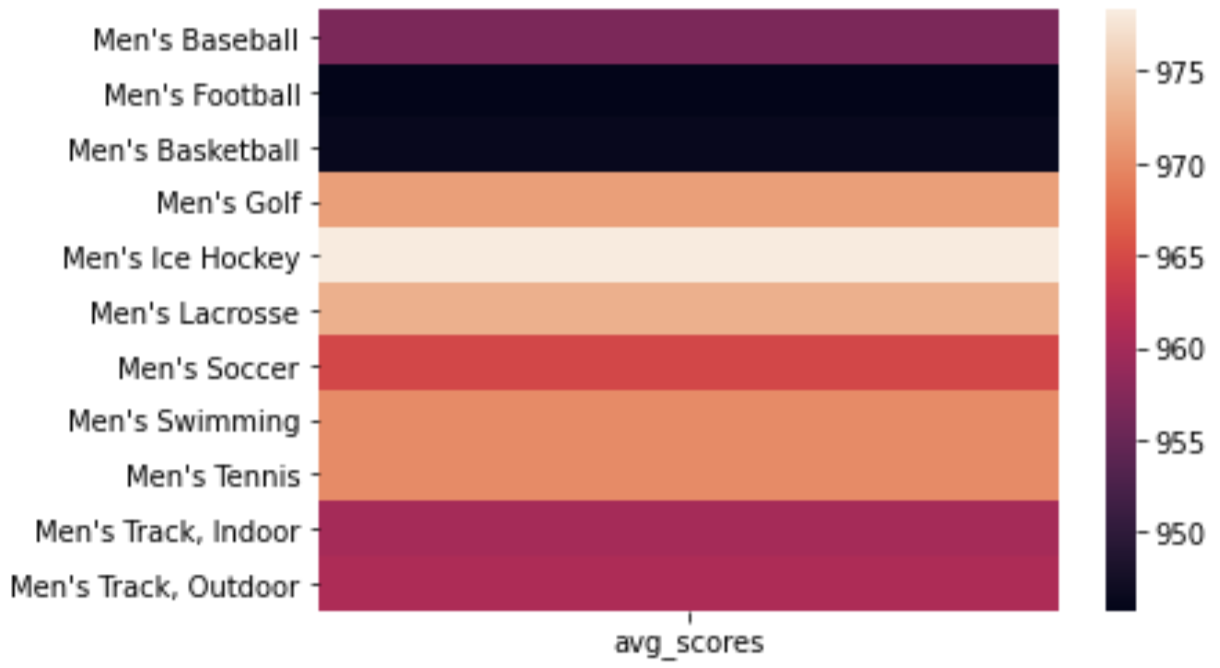
**

*D.* *Other Plots*

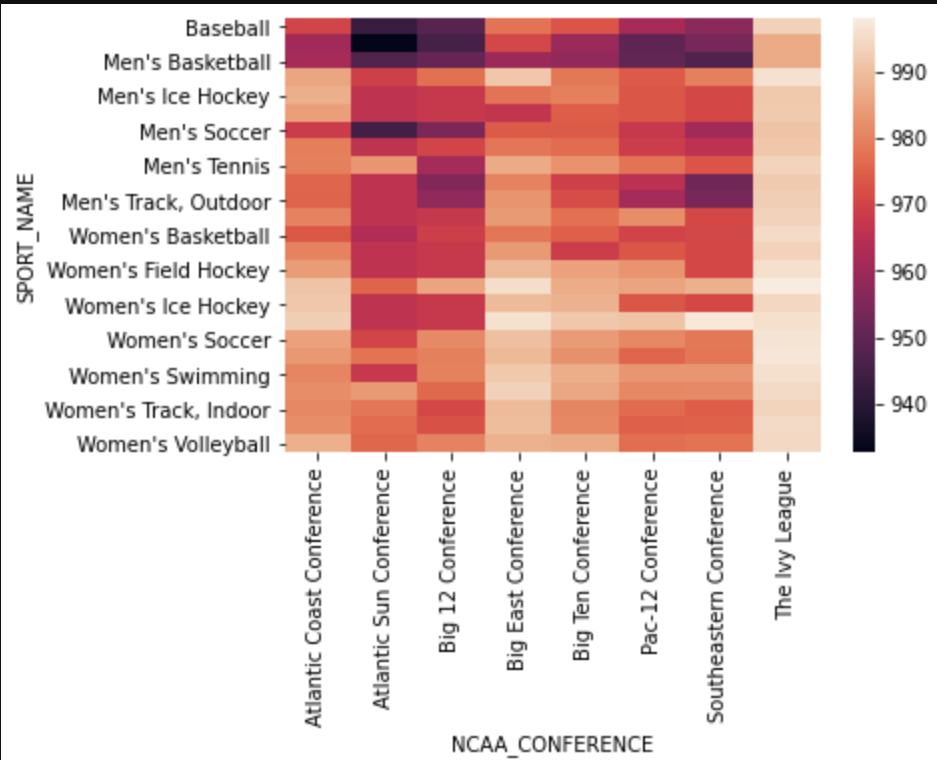
**Figure 4: Scatter Plot of 7 specific Conferences**

**

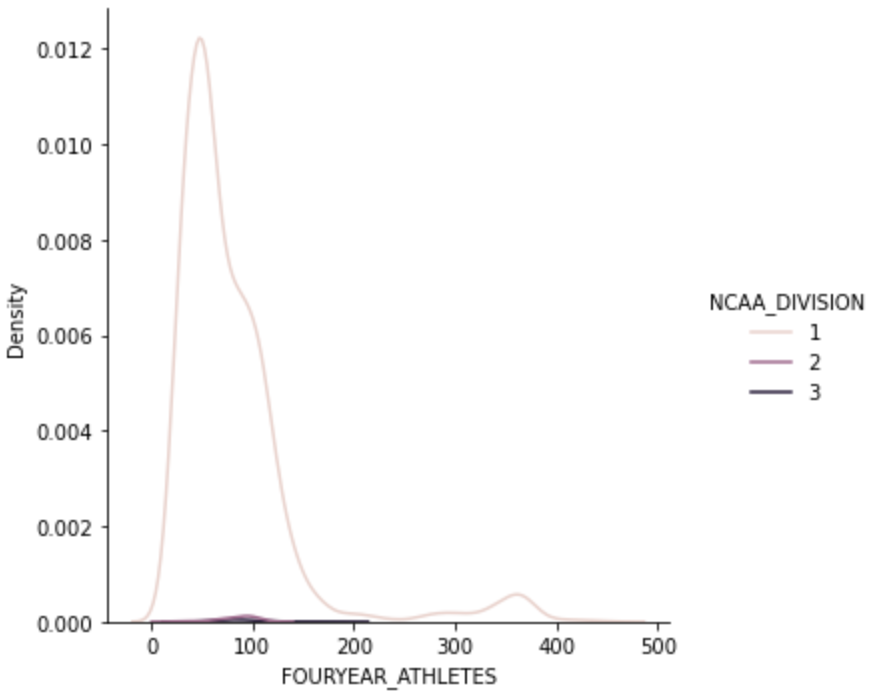
**Figure 5: Heatmap of Men’s sports from every Conference in NCAA**



**Figure 6: Heatmap for the main Men/Women sports in our specific Conferences**



**Figure 7: KDE of the amount of Four Year Athletes in all of NCAA**



**V.** **SUMMARY OF FINDINGS**

In **Figure 1**, we graphed the distribution of the average APR score for each sports team. It seems to have a pretty normal distribution. A perfect APR score is 1000. **Figure 2,** the pairwise plot, shows how the APR score in each sport in each Conference pairs up to the scores of the same sports in other Conferences. **Figure 3**, the bar chart shows the average APR scores for men's versus women's sports. It appears that on average, women’s sports teams tend to outperform men’s sports teams when it comes to academic performance. **Figure 4** shows the average APR score for the most talked about Conferences, with Ivy League added in to view its academic superiority and Atlantic Sun Conference for obvious reasons. Unsurprisingly, the Ivy League has the highest average APR score, and unfortunately, the A-Sun conference has the lowest. **Figure 5;** This heat map shows the average APR scores for the main mens sports throughout the entire NCAA. Basketball and Football seem to have the lowest average, and Golf, Hockey, and Lacrosse appear to have the highest scores. This is a good way to quickly compare sports averages against each other. **Figure 6;** This heat map is similar to the last one, but this one shows each sports team's average APR score within its respective conference. This would be useful if one wanted to compare how a particular sport in a particular conference compares to another sport in a different conference. **Figure 7;** This kernel density chart shows the density of four year athletes in division 1, 2, and 3. It is clear that division one sports have a much higher density of four year athletes when compared to division 2 and 3 sports. This most likely is caused by scholarship options that Division 1 players recieve.

Working with this data set was very interesting. Coming into it, we both had a few ideas in mind about what sports would have good/bad academic performance, and it was cool to see if we were correct. We found that men's basketball tended to have the worst academic performance across all conferences, while men's ice hockey had the best academic performance. We also found that, generally speaking, women’s sports teams outperform men’s sports teams academically. To no surprise, the Ivy league sports teams substantially outperformed all other conferences, and the Atlantic sun conference had the worst academic performance. When it comes to 4 year athletes, Division 1 schools had a much larger amount than Division 2 and 3 schools. We assume this has to do with the scholarship opportunities Division 1 players have that Division 2 and 3 players don’t. Overall, working with this data set was interesting because it did reveal that there are some relationships between specific sports, conferences, and gender when it comes to academic performance within the NCAA.