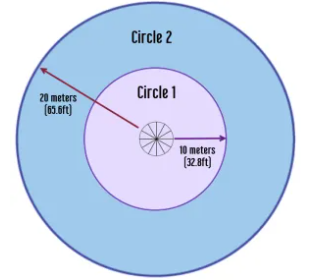
Disc Golf Data Mining Project

The data mining task I am attempting to solve is related to the sport of disc golf. Now disc golf has been around for a long time but it didn’t become popular until 2020 due to the Covid-19 outbreak which caused many people to start playing as it’s an outdoor activity. Due to the recency of the popularity, there are few resources focused on analyzing data within the sport unlike other mainstream sports such as basketball, football, or even baseball. I am attempting to solve this issue by analyzing pro player’s statistics from the 2023 Disc Golf Pro Tour (DGPT) season in hope’s of giving amateur players like myself the ability to see what aspects of our game we need to improve on to play at a higher level. This data mining task can be formulated into a clustering analysis problem. I had originally planned on doing anomaly detection to see if the winner of each tournament last year was considered to be anomaly when compared to the rest of the field playing but I decided against this as I thought clustering would provide a clearer picture to the data being analyzed. I believe looking at the top players from each tournament will clearly display what categories they’re exceeding in and also have enough variance to show the difference between a first and tenth place finish allowing amateur players to understand what they really need to hone their skills on.

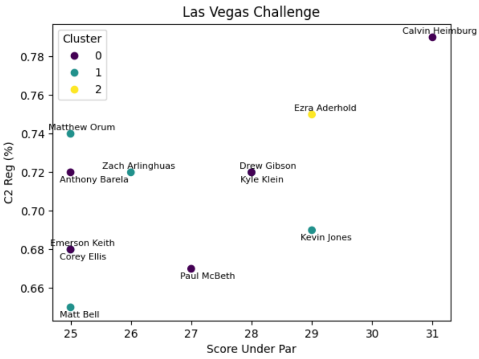
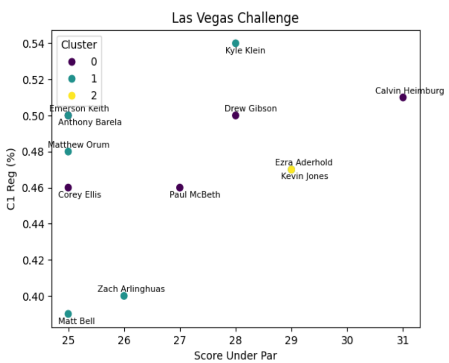
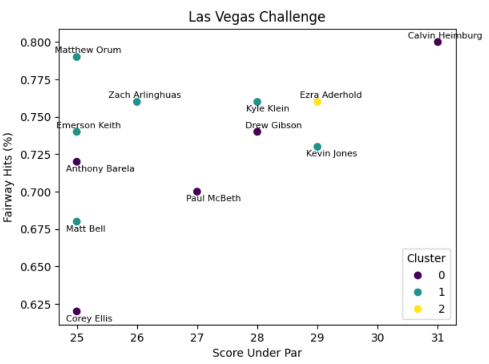
The input of my data mining task consists of ten tournaments from last year’s DGPT season with the top ten players’ statistics from each of those tournaments being analyzed. In some instances there were ties for 9th or 10th place so I’ve included any players who tied as well so some datasets have 11-13 players instead of ten. There are many categories of data relating to each player but the categories I decided to analyze were Total Score, Fairways Hit, Circle 1 in Regulation, Circle 2 in Regulation, Circle 1 Putting, Circle 2 Putting, and throws that landed OB. Now I know the average person probably isn’t familiar with some of the terminology here so let me explain. Total Score is an integer representing a player’s score under par. Fairways Hit is a percentage of how often a player threw a disc off the tee and it landed in the fairway. The category of “throws OB” is an integer representing how many times a player throws a disc during the tournament which landed out of bounds. For the remaining categories I must define what Circle 1 and Circle 2 are first. Circle 1 is an imaginary Circle ~33ft (10m) surrounding the basket which is designated with white/red flags during tournaments. Circle 2 is the same concept except it’s ~66ft (20m) until ~33ft (Circle 1) all the way around the basket and is marked with blue flags. An image is shown below for clarification.

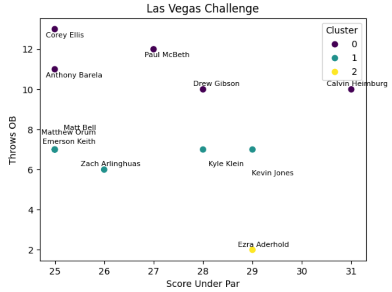
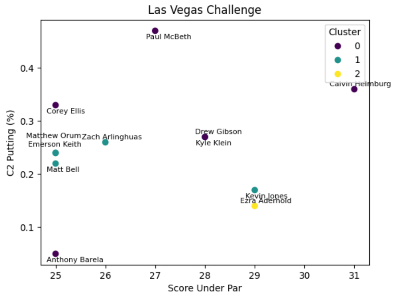


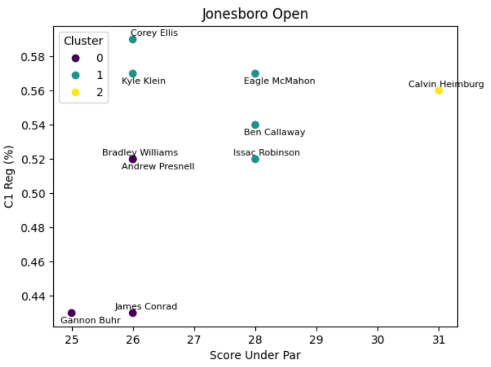
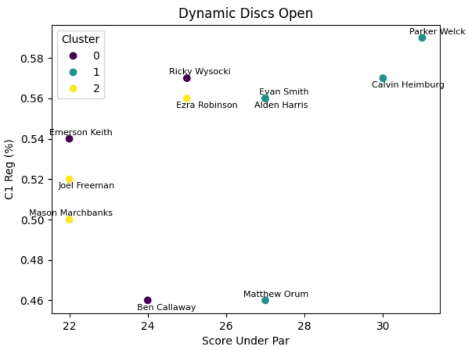
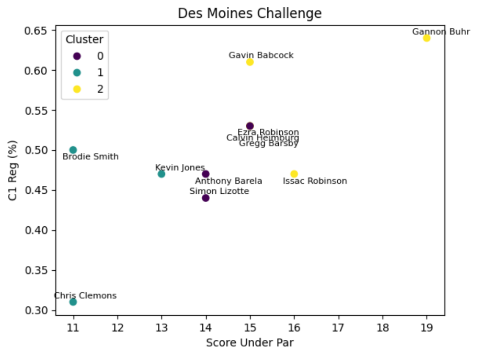
Hopefully you can better grasp the remaining categories now. Circle 1 in Regulation is a percentage of which a player throws a disc and it lands within 33ft of the basket and that player still has two throws remaining until they reach par giving them an opportunity for birdie. Circle 2 in Regulation has the exact same meaning except the disc the player has thrown has landed anywhere from ~66ft to ~33ft from the basket. Circle 1 Putting is the percentage of which a player makes their putt while they’re in Circle 1 (less than 33ft from the basket). Circle 2 Putting is the percentage of which a player makes their putt while they’re in Circle 2 (less than 66ft but more than 33ft from the basket). From these categories of data, the expected output is three clusters of data from each tournament displaying the average means of each category hopefully showing a correlation between player’s scores and statistics.

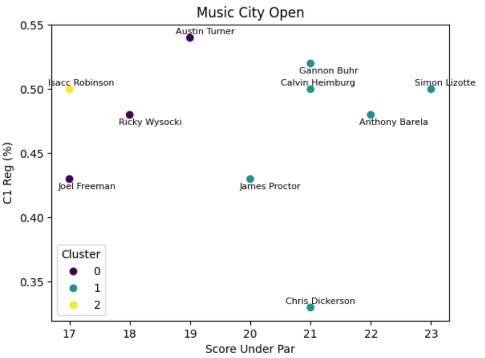
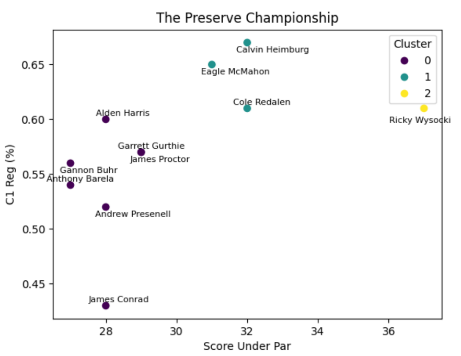
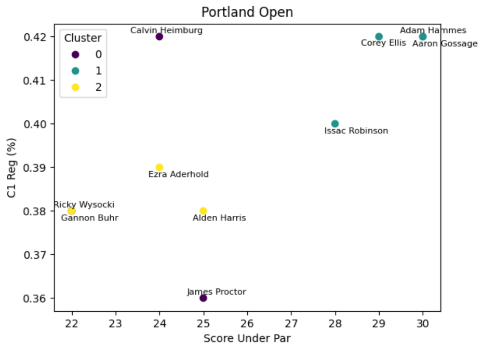
The datasets I used to obtain this data can be found at udisclive.com/stats. I had originally contacted statmando.com for the data as I was under the impression they were the ones collecting the data but once one of their representatives got back to me, he pointed me towards the udisc website. This was a huge relief as for a while I didn’t know if I was going to be able to get access to the data I wanted. The data I have pulled from the website is from 10 tournaments that took place all over the United States, from the top 10 players of each tournament including ties which comes out to 112 players statistics. This is not 112 unique players as many players place in the top 10 in multiple tournaments and sometimes even win more than one tournament. All the data on the website was in a huge table but there was no way of importing it over into a csv file. This meant I had to manually enter all the data into an array so that I could convert it into a csv file as I was using the K-means algorithm in a very similar fashion to what we did for homework 2. For preprocessing the data, there were no missing entries but I converted all percentages to decimals for ease of use. I treated each tournament as a separate dataset so I created ten csv files after I had gone through entering all ~10 players statistics for each tournament.

The output I obtained is six scatter plots for each tournament which display each player’s score against one of the other data categories. I will not be including all of the plots as that would be 60 scatter plots and this report isn’t supposed to be larger than 4 pages. However I have included all six plots for the first tournament along with the cluster attribute means. For the other nine tournaments I have the plot of what I believe to be the most important metric amateur disc golfers need to focus on and the cluster attribute means.

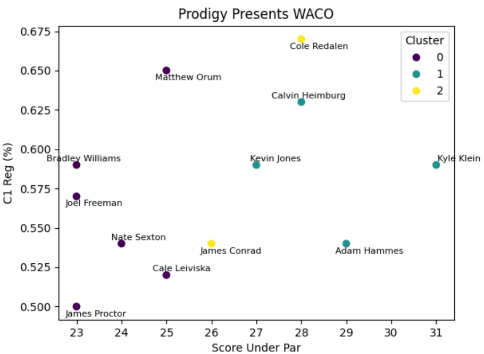
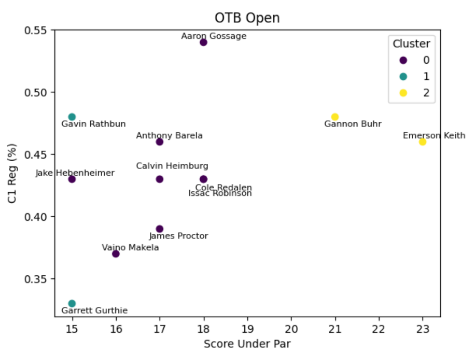
A graph with numbers and a number of players

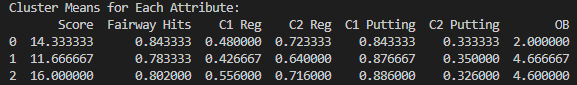
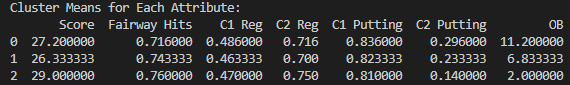
Description automatically generated with medium confidence



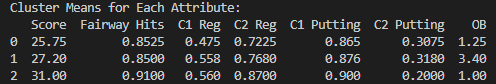
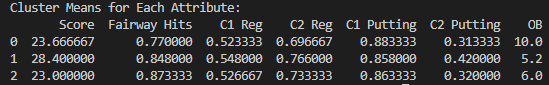


A graph with numbers and names

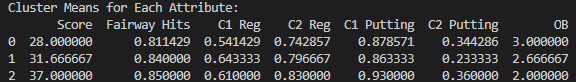
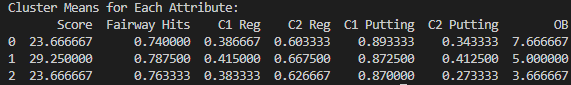
Description automatically generated with medium confidence



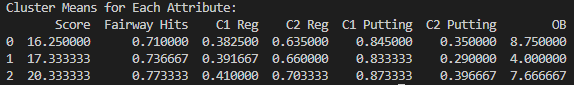
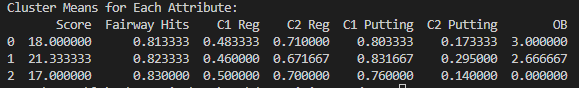
Las Vegas Challenge Attribute Means Des Moines Challenge Attribute Means



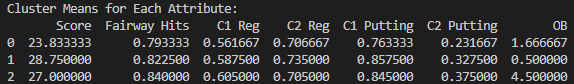
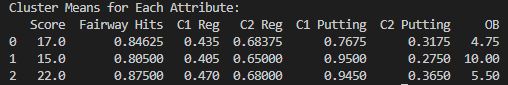
Dynamic Discs Open Attribute Means Jonesboro Open Attribute Means



Portland Open Attribute Means The Preserve Championship Attribute Means



Music City Open Attribute Means Open at Austin Attribute Means



OTB Open Attribute Means Prodigy Presents WACO Attribute Means

The first six scatter plots are from the Las Vegas Challenge player statistics so you can get an overall view of the data being displayed. The next nine plots are the Circle 1 in Regulation plots from the other tournaments. I believe the Circle 1 in Regulation statistic to be the most important to focus on as 7/10 tournaments the cluster with the highest average score also has the highest C1 in Reg. This can be seen in the cluster means for each attribute images shown above. Also 3/10 tournament winners had the highest C1 in Reg percentage with two other winners being a close second. I believe this is definitive enough to make it the number one priority for amateur players to improve upon as no other attribute has a stronger correlation. If I were to suggest the 2nd most important attribute of disc golf, it would be C1 Putting. In 6/10 tournaments the cluster with the highest average score also had the highest C1 putting percentage. This makes sense as it’s absolutely crucial that if you’re inside C1 that you make your putt whether it be for par or birdie. The last thing I’ll say is regarding the mental game of disc golf. Looking at the winning scores of each tournament, there’s quite a bit of variability in the winning score which can be attributed to the difficulty of the course, the number of rounds played during that tournament, and the weather involved. This being said, as a player you never know what the winning score may be before the tournament so don’t ever give up on yourself during the tournament no matter how well you feel you’re playing as you may do better than you think. After doing this project, I’ve realized there’s even more potential than what I had originally imagined for data utilization within the disc golf community. I hope to continue using my gained data mining skills to further dive into the data within disc golf in hopes of advancing the sport in the years to come.