Since the beginning of Rocket League Esports, analysis of players, teams, and their playstyles has contained a lot of speculation and generalization with no definitive numbers to back up any claims. Sometimes at the end of a game or a series, the casters and analysts will mention some of the scoreboard statistics to try to put some of the complicated puzzle together, but those numbers on the scoreboard don’t tell much of a story on their own, especially if you’re trying to calculate their playstyles, strengths, and weaknesses throughout the history of their team. In an attempt to uncover the mystery of Rocket League playstyles and statistics, I clustered select attributes of data from seasons two and three of RLCS gameplay and broke each cluster down into a position played on the field.

##Player Analysis##

I’ll start off by giving my analysis of the numbers I produced. The methods I used to get these numbers are at the end of the article. After running the clustering algorithm 100 times and choosing the best set of clusters, I was presented with five fairly distinct clusters of players, where all but one cluster had a roughly equal chunk of players associated with it. Since the algorithm only gives me numbers for the clusters, I had to come up with the names and implications of each one. This is where my analysis may turn subjective, but I think that I have my arguments pretty air tight. If you disagree, please present your argument to me; I want this to be a useful tool for the community and analysts to use, and any constructive criticism is helpful to achieving that goal.

First, I’ll show you the [scatterplot matrix](<http://i.imgur.com/Qi2KMBv.png>). This plot shows every variable compared to each other with the algorithmically assigned clusters color coded. Some plots have higher correlation than others while some more clearly show the separation between the clusters. It’s important to look at every possible combination of attributes in high-dimensional analysis like this to be able to see which factors play more of a role in determining playstyles and which ones don’t. Another useful tool the algorithm gives me is the [mean of each attribute]( http://i.imgur.com/aEFgMCo.png) in its cluster. With these two tools, I created some [tables](https://drive.google.com/open?id=0B3RqrLHaoDx-dS1DQmVlNWZxYUk) to help me understand the role each group of players has in a game of Rocket League. As you can see when comparing the tables to the analysis, I ignored any data that was too spread out, and I ignored comparisons between cluster attributes where the number was the same down to the thousandths digit. For example, the only notably higher group in the percent of team score was the carry, and everybody else was basically even. Here are the breakdowns:

Cluster 1: Striker

This group falls below the carry group in shots and goals, but they sit comfortably above everybody else. They’re below every other group in their saving percentage, meaning that this group likes to sit on the offensive side of the field, and while they still get some saves, they are usually the last in the rotation to do so. Interestingly, this group is the team MVP less of the time than any other group except defensive midfield.

Cluster 2: Passer

This group sits well above everybody else in their assist percentage and assists per goal (1.85), but they fall below everybody else in shots and goals. This group consists of the five players in RLCS seasons 2 and 3 who are much more comfortable passing the ball than they are shooting it, resulting in a specific playstyle that not many players share. They sit in the middle of the pack when it comes to saves and MVPs.

Cluster 3: Carry

The players in this group by far have the most goals, saves, shots, MVPs, and goal participation percentage, along with the highest score. They fall below in assists only to the passing group, while every other group has a higher assist per goal value. This player is all over the field doing every job more than their peers, and they’re able to score more unassisted goals than any other group of players.

Cluster4: Offensive Midfield

This group has the edge over the defensive midfield in goals and MVPs, but not in shots and assists. This can be attributed to the fact that they tend to play closer to the opponent’s goal, so they can get more shots on the net. Basically, they get passed to by a passer or a defensive midfield player, but they play more towards the defensive side than a striker.

Cluster5: Defensive Midfield

This group has the edge over the offensive midfield in shots and assists, but not goals and MVPs. They play more on the defensive side of the midfield, meaning that more of their shots will be saved while some will be scored by another teammate. This is proven by the significantly higher assists per goal ratio (1.22) that this group has than everybody else but the outlying passers. They get more saves than their offensive counterpart, meaning they will usually be the first back to net if the team doesn’t have a carry.

[Here](<https://drive.google.com/open?id=0B3RqrLHaoDx-c1dudTRIbjExaTQ>) is a table showing every player with their team, season, and cluster number, sorted by the group from 1-5. And [here](<https://drive.google.com/open?id=0B3RqrLHaoDx-VWhzcWlmYjJTTnc>) is the same data sorted by team. With this data, you can definitively tell what types of players each team is playing with and then try to see from there which combinations make a team successful and which don’t.

By clustering players based on their statistics, I was able to determine which playstyle any given player from seasons two and three of RLCS abides by most of the time. Of course, like in most fast-paced sports, players normally don’t confine themselves to one specific duty even though they will usually specialize in one thing or another. That means that, while this initial analysis of the players is useful for determining what a player will do on the field most of the time, it doesn’t tell the whole story. In the next article, I’ll talk about and show how we can use this initial analysis to take an even closer look at exactly what kind of playstyle a player has and which role each player has on their team.

Again, if you’d like to read up on how I calculated all of these numbers, I provided that information below. Otherwise, thank you for reading, and I hope I gave you a deeper insight into the statistics behind professional Rocket League.

##How I got the numbers##

The algorithm that I chose for this job is K-means clustering. If you aren’t familiar with the algorithm, here is a quick explanation. First, you specify the number of clusters that you would like to generate, and then you feed it an n-dimensional amount of data. It generates the number of centroids that you specified and then calculates the distance to each point. Every point will be assigned to its closest cluster, then the centroid will be moved to the center of all points associated with it. The last step will repeat until the position of the centroids stop changing. If you want a more in depth explanation, [here](https://en.wikipedia.org/wiki/K-means\_clustering) is the Wikipedia page. The statistics I used were compiled by /u/Slokh and published [here]( <http://rl.parallel.gg/stats/players/season-three-league-play#>).

When doing any kind of statistical analysis, you first need to normalize the data and make sure that the numbers are all within the same range. For this dataset, all I had to do was divide a few of the values by 100 to bring them down to the 0-3 range. I also took out all players who played less than 15 games in their respective seasons because, most of the time, those turned out to be outliers. [Here]( <https://drive.google.com/open?id=0B3RqrLHaoDx-dkNTXzJscDRUSnc>) is the final data that I ended up with.

The next step is choosing the variables and clusters that you’d like to use in the analysis, otherwise known as feature selection. You can try to use every single variable provided, but a lot of attributes are calculated using other attributes, so there is some redundancy to watch out for. Using every single attribute in the clustering algorithm returned very poor results, so I had to weed out unnecessary attributes and keep trying combinations that make sense together to see which ones return the best clusters. The “correlation” of a group of clusters is determined by a value calculated by dividing the between sum of squares by the total sum of squares. You can read about that [here](<https://stats.stackexchange.com/questions/82776/what-does-total-ss-and-between-ss-mean-in-k-means-clustering>). I had initially decided that I’d use three to five clusters for this analysis. A minimum of three because that’s a good general minimum to start off; a maximum of five because I feel that any more than that would be cherry picking niche midfield positions, and it wouldn’t be very meaningful. To test my theory of the maximum, I ran the algorithm with six clusters and ended up with a slight variation of midfield player, proving my hypothesis true.

After testing a [few groups](<http://i.imgur.com/5DcFlRo.png>), one really stood out as having the best correlation: Percent team score, percent team goals, percent team assists, percent team saves, percent team shots, percent team MVP, assists per goal, and goal participation percentage. This returned a BSS/TSS value of 75.3%, which is very convincing. Looking at the scatterplot matrix, the algorithm certainly did a good job of finding the clusters. If you have further questions, please don’t hesitate to ask.