The Best Way to Win a Match of PUBG

JOSHUA RAMOS, University of Colorado at Boulder
TYLER MOORE, University of Colorado at Boulder

1 PROBLEM STATEMENT

Player Unknown Battlegrounds, better known as “PUBG”, is an online multiplayer battle royal video game. The game starts with one hundred players and the only way to win is to be the last man standing. There are many variables that may better your chances of winning. Those variables are dropping location, team size, walking distance and accuracy. With millions of data point and using the skills we have learned from this data mining class we hope to find what combinations will assist a player in ranking in a higher percentile.

We hope to find some interesting stats on combining unique datasets. For example, it would be interesting to find the correlation of walking certain distances and winning or things like if there is a sweet spot for team size that would increase your chances of winning. It's always interesting finding stats that many players don’t even think about, and with a passion in data mining and video games we believe we’ll find some useful information that is new and creative to improve community knowledge on PUBG.

2 LITERATURE SURVEY

Player Unknown Battlegrounds is a relatively new game with an initial release in 2017, so there is some research done but it hasn’t been mined as much as we would’ve liked. From the Kaggle page and the dataset we are mining on we found two unique projects.

A user without his actual name listed but a screen name called Skihikingkevin. Skihikingkevin took the data set and created heatmaps of the final circle based on where the final people died and their coordinates. Skihikingkevin used python for his data processing using the libraries as listed matplotlib, numpy, pandas, seaborn and used my jupyter notebook to show exactly what he did. In conclusion he found, “Erangel: Although a lot more spread out, there are a few pockets of land that is more contested for final circle, primarily near very open areas between pochinki and mylta. It is worth noting that a lot of close river areas north are avoided so players based off of seeing the first circle, should be able to make an educated guess where the next circle will be.
Miramar: A lot different and more clustered. Notable clusters are between bendita and Leones, south of San Martin and west of Pecado. It is interesting to see that a large portion on the of the map is often never in the final circle” With this information it can assist a player in jumping in a location where they won't have to move as much to stay in the circle.

Another user named “Christoph Egerland” Mined the same dataset we will be using and his focus was on making maps of where players jumped the most. The tools he used were as follows python, matplotlib, numpy, pandas, seaborn and used my jupyter notebook to show exactly what he was doing. He focused on finding were the first kills were to find out where exactly the players were landing and using their

coordinates to make a heat map. The heat map is useful because it can tell players where the best place to be in the final circle.

3 PROPOSED WORK

Using these datasets, we can hopefully find multiple answers to the questions we want to figure out. Before we actually begin mining, we need to make some alterations to our data so that we can use it.

The first problem we need to solve involves **cleaning our data** so that there are no NULL values. If we do not fix this problem, then we will either get errors or our answers will be skewed. The reason there are NULL values is because sometimes a player will enter the match and then leave early. When this happens, the game will enter NULL values into the data. To fix this problem, we will just need to replace all the NULL values with 0's to signify they left the match early.

Another thing we need to do before actually mining is **normalizing** our data. For example, the rules of the game involve having a team of up to 4. However, some teams might only have 3 people in their group. If we decide to look at the amount of kills a certain team has, it would not be fair to compare a team of 4 to a team of 3. This is why we need to normalize these numbers in a way that makes it fair. A solution would be to look at the kills per player on a certain team and see if a team of 4 has a higher ratio than a team of 3, or vice versa. Also, if we are going to look at team stats, one thing we will have to do is make a whole new dataset with just team information on it. Right now, our current datasets only have individual player information, including their team ID. So we can make a new dataset that finds all the unique team ID's and put their data in it.

Once we have all the datasets we need and all the data normalized, it will be time to start mining it to hopefully find answers to our questions. One of the first things we will be doing involves clustering the data. A lot of our questions involve finding patterns and one of the best ways to do that is from clustering. We plan to put all the players together who have the same type of stats, such as: same distance traveled, same amount of kills, same amount of players on a team. Once this data is all sorted out, we will sort through it to see who has the highest win percentage to hopefully figure out the best strategy to win the match.

How it is Different Than Previous Work: Our project is much different than other people's previous work because there's mostly identifies heat maps of where people started out and where they died. Ours goes more in depth because we want to know the strategy behind this game, like if a bigger team increases the chance to win or if hiding the whole game will make you less likely to die. We will have to integrate the datasets way more than just finding out where people died.

4 DATA SET

When we downloaded the datasets, we realized there are multiple attributes to look at and evaluate. In one of the datasets, it has all the information on deaths. This involves who the killer was, their position, how much time has gone by, the victims name, the victims position, etc. This dataset has over 65 million entries. There is also another dataset based purely on a specific person for an individual match, including how far they traveled, how many kills they had, the amount of time survived, their individual and team placement. Every single attribute in both tables involves integers, except for the match ID which is just a random string value.

URL : https://www.kaggle.com/skihikingkevin/pubg-match-deaths/

---

5 EVALUATION METHODS

The two main evaluation methods we will be using in the beginning will be clustering and association. One of the first things we need to do is put similar players together, such as number of kills or distance traveled. We currently have 65 million death entries to look at, so by grouping similar players together, we can break up our dataset into different categories and test certain ones specifically. This can significantly speed up our algorithms because we will be testing groups of players, not each one individually.

We will also be using association to figure out certain patterns. For example, do players who travel more during the match also get more kills? This will be very helpful in figuring out the best way to win a match because hopefully we can find a correlation between certain attributes that help a player increase their chances to win.

After we have found certain patterns among the dataset, the only thing left to do is to apply this knowledge and figure out if it works. One way to do that is to actually play the game and follow the patterns to see if this increases our winning percentage. Another way is to use conditional probability. We can delete all the players who do not meet the requirements we found, and then see if the remaining players have a higher winning percentage than the general group.

---

6 TOOLS

-Python: A general purpose programming language that is great with manipulating big datasets
-MatPlotLib: A plotting library that can produce powerful graphs for displaying dataset information
-Numpy: A library for complex matrices and array operations.
-Pandas: A library for reshaping and viewing datasets in a variety of ways.
-Seaborn: A library based on MatPlotLib that can create heat maps to better visualize our dataset

---

7 MILESTONES

-Data Preprocessing (getting rid of NULL values and normalizing) : March 12th, 2018
-Data Integration (creating new team dataset) : March 19th, 2018
-Simple Techniques : April 2nd, 2018
    -Clustering (break up dataset into different categories)
    -Association (find correlations between two or more items)
    -Other techniques we will learn later
-Back testing (once patterns are formed, look back at the data to see if we are right) : April 16th, 2018
-Get everything formalized for final presentation : May 1st, 2018

---

8 SUMMARY OF PEER REVIEW SESSION

In summary, we had positive feedback at last weeks presentations. Professor Boese had nothing to say after we presented. We believe that some additions that we could have made was a little more focus on exactly what we wanted to mine and the tools we would be using. With a little more research I believe we

could have found the exact tools we would be using continuing the project. I think it would have been useful if we did a little more research before presenting.

At this time our feedback hasn't been given to us, but we can discuss feedback we would give to other groups. Many of the other groups used very similar or the same datasets and didn't improve presentations after watching the previous groups. We also think it will be useful for groups doing similar projects to talk to each other to get an idea how and what other groups are doing. Aside from that the groups presentations went really well and almost every group spoke really well and had great presentation skills.

Egerland, Christoph. "PUBG Data Analysis." PUBG Data Analysis | Kaggle, www.kaggle.com/chegerland/pubg-data-analysis.

KP. *Final Circle Heat Map*. www.kaggle.com/skihikingkevin/final-circle-heatmap.