Chicken for Dinner

A data mining project on discovering how to win in PUBG

Zijun Xu
University of Colorado Boulder
zixu6769@colorado.edu

Jonathan Young
University of Colorado Boulder
joyo2566@colorado.edu

Jianyi Chen
University of Colorado Boulder
jich2029@colorado.edu

Problem statement:

Playerunknown's Battleground (PUBG for short) is a popular multiplayer online battle royale game developed and published by Bluehole, a Korean publisher. Typically, there will be maximum 100 players in a game. The game starts by putting all of the players on an airplane, which goes through the map in a random direction. The players are free to choose when to jump out of the plane after the back door opens. Various resources are provided across the map, including weapons (with a large number of possible combinations of parts), armors, medicine, vehicles, etc. Additionally, after a short while when all players land on the ground, the game forces all the players to move to the “safe zone”. If a player remains outside the safe zone, they will continually lose vitality. The size of this safe zone decreases with time, picking a new circle of same relative size inside the new circle. This is to force the player’s closer together to instigate struggle (as shown in figure 1). The game ends when there is only one team left. A team may consists of 1-4 players at the start of the game, but only one must survive till the end for the team to win.

Statistically speaking, at the start of a game a single player has only a 1 percent chance of winning. By mining the death statistics data, we aim to create a survival guide for any player who wishes to have a better performance in the game. Our survival guide should be able to answer some interesting questions like: Does skill or luck win more games? Do players that get more kills win more games? Where do the most player deaths occur?

Figure 1: A example of the safe zone reduction in playerunknown’s battleground

Literature survey:

We have found some interesting prior work that is related to our project. Christoph Egerland has done an exploration on where most of the people jump out of the plane in PUBG.[1] By using Jupyter notebook, Christoph scaled both maps' coordinates to fit the map image in order to draw density points later on the image file (shown in figure 2). Since the dataset doesn't contain the positions that players jumped out but only the positions of players' deaths, Christoph restricted his program to only take death positions in the first two minutes of the game - because most of the players that die early do not usually move far away from their jump locations.

Another interesting project was done by Justin Moore and published on August 8, 2017.[2] In his project *To drive...or not to drive*, he analysed vehicle usage in depth. First of all, he showed that the mean of average ride distance grows as there are more people in a team. Justin then drew a scatter plot showing the correlation between average ride distance and wins, as shown in figure 3. The result indicates that players who have the average ride distance between 2000 and 4000 are the players with the most wins. In other words, players die more often if they don't drive at all, or drive too much.

Figure 2: Density graph of where the players jump at the most

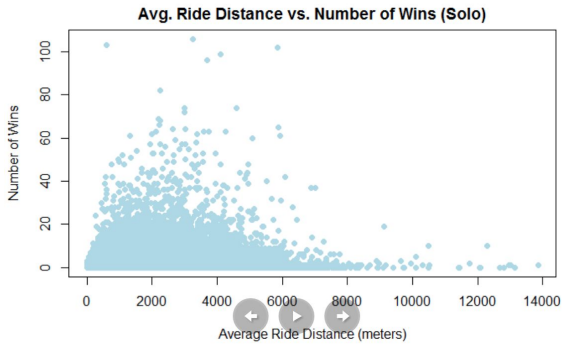

Figure 3: Average ride distance vs # of wins

Proposed Work:

Before we process our dataset, we will have to detect and clean any corrupt or inaccurate data to ensure the completeness and accuracy of our dataset. Specifically, we will locate those data that have missing or inaccurate attributes then remove them from our dataset. For example, there are some rows in our dataset that are missing the player's coordinates, which is inaccurate and needs to be cleaned. We are likely to remove attributes like victim name and killer name, which are also useless to us. In addition, to create the density graph showing players' coordinates, we will have to normalize the player position attributes in our dataset. The coordinate in our dataset is represented by big numbers in range from 0 to 800000, while our pictures of the maps are of the size 1000 * 1000. We will have to scale the player position by multiplying them with 1000/800000 in order to keep the player position and the picture size in the same page. Another step we will take in the data cleaning is the removal of all deaths that occured before the first safe zone reduction. This is to help with our proposed analysis.

After preprocessing, we will start analysing the data. Our analysis will be whether or not player skill or player luck has a higher precedence for determining a

player's final ranking. Luck and skill will be defined by the relations between travel distance and number of kills. The players with lower travel distances had luck on their side because they were closer to the center of the safe zones during the size reductions. And those who have more kills (over some particular threshold) often have a higher level of skill. We will analyze a players luck vs their skill. To do this we will be looking at four different relations: low distance-low kills, low distance-high kills, high distance-low kills, and high distance-high kills.

We also would like to analyse the relationship between the time index of a safe zone reduction and the locations of the most player deaths in this time in order to create some mapping of where to avoid during particular instances.

Data set:

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

There are two main parts of our dataset. The first part provides detailed information on the match, such as weapon type, player position, player placement (ranking), round time. The second part gives more general information on the match, including distance traveled by foot and vehicles, player placement, player damage dealt, player kill numbers.

Evaluation methods:

We will be looking at the four relations described above, blocking all players into these four bins. We hope to be able to use these bins to analyze a players relative luck or skill and which is better for survival.

For the deaths during the reductions we plan to group the players who died during particular time indexes. Since the size reductions of the safe zone always occur at the same time intervals during any given game, we can create bins such that the size of the bin is the time between circle reductions, i.e. (0-first),(first-second) and so on. After the bins have been made we will look at the player locations and create a heat map for the given game.

Tools:

We will be using MySQL and Jupyter Notebooks for cleaning and analysis. As well as Weka for visualization.

Milestones:

- Preprocessing completed
- Data integration
    - Combining our two datasets into one
- Creating the evaluation methods for luck and skill
    - Bin creation/bucketing of players
- Making bins for time indexes
- Mine the data/extract desired information
- Creating visuals

Summary of peer review session:

Finding prior works is important for our project because they can provide us with information on how to approach our problem and act as guideline for us to follow.

Reference:

1. Egerland, Christoph. "PUBG Data Analysis." *PUBG Data Analysis | Kaggle*, Jan. 2018,

www.kaggle.com/chegerland/pubg-data-analysis.
2.Moore, Justin. “To Drive... or Not to Drive...” *PUBG Data Analysis*, 8 Aug. 2017, pubganalysis.wordpress.com/2017/08/08/to-drive-or-not-to-drive/

Feedback:

We were given the feedback that we had not found any prior work to help us with our mining idea. To fix this we went online and found some prior work which helped us to narrow down our idea.