PUBG Data Royale

CSCI 4502 Final Project

Matthew Davenport, Federico Aragon, Thomas Alder, Haotian Dong

University of Colorado
Boulder, Colorado, USA

PROBLEM STATEMENT/MOTIVATION

In this project we are looking to analyze data from the results of a player's matches they played in a Battle Royale video game named PlayerUnknown's Battlegrounds. Using this dataset, we hope to answer questions regarding the game's mechanisms through an analysis of of player damage and deaths and how this statistic could be improved. We would also like to see what the likelihood of a player being a hacker is through understanding the inflicted damage they cause and whether or not there are significant outliers which may suggest an unfair advantage. Lastly we would like to find the best strategy to use when playing the game in terms of determining the most dangerous places to avoid, places which may be beneficial, and the general strategy of players that tend to win more.

LITERATURE SURVEY

1 Final Circle Heatmap

Kevin Pei. 2018. Final Circle Heatmap: PUBG Match Deaths and Statistics. https://www.kaggle.com/skihikingkevin/final-circle-heatmap

This previous work explored the "final circle" of games played in PlayerUnknown Battleground's to visualize patterns in the ultimate positions of the ever decreasing circle of survival. Throughout a match in PUBG, players are restricted to a circled playing area which continuously has periods in which it decreases in circumference and times where it stagnantes. Using python and jupyter notebooks, Kevin Pei uses the deaths of players who placed first and second place who did not die outside of the circle to determine the location of each final circle. He then plots the results on a heatmap which shows the distribution of final circle locations demonstrating areas of congestion in which the final circle ended up in a concentrated area. Visualizating both maps shows that the edges of the map never see the final circle at their location while numerous pockets of high density are dotted in areas of more diverse end locations.

2 Survival Analysis Using Kaplan-Meier Estimator

Robert Kwiatowski. 2018. PUBG - Survival Analysis (Kaplan-Meier): PUBG Match Deaths and Statistics. https://www.kaggle.com/datark1/pubg-survival-analysis-kaplan-meier

This previous work shows the damage and die rate data for each players outside the playzone before PC 1.0 Patch #15 of

PlayerUnknown Battleground's to visualize the survival situations. The data set also has the data of the party situation including SOLO, DUO, and SQUAD, the use of vehicle for each players, and the numbers of players for each game. He plots the results for different groups with those conditions to illustrate the survival rate for players. Then, Kwiatowski uses a Kaplan-Meier estimator to use non-parametric analysis using the number of events that occur during a time frame and the survived units from the previous time frame. Each time frame represents the amount of time the circle is a stagnant playzone. Using this, he creates three plots which shows the survival probability based on the team members a player has which are solo (no team members), duo (one team member), and a squad (2-3 team members). Here it is seen that the highest chance of survival is for squads. The same process is done for two different scenarios in which a player uses a vehicle versus when they do not which yields results that show using a vehicle has a greater probability of survival. A final stratification is done based on the number of players in a match between small (< 33 players), medium (33-66 players), and large (> 66 players). Here the results show that the smaller number of players creates the larger survival chances.

3 **Data Visualizations**

Peter Ott. 2018. PUBG - You're not as bad as you think: PUBG Match Deaths and Statistics. Various data visualizations - https://www.kaggle.com/ottpeterr/pubg-you-re-not-as-bad-as-you-think

In this previous work, Peter Ott creates a variation of data visualizations mainly histograms which plots different relationships which can be found in the data sets. First, he plots the relationship between the rank of team placement and player kills before splitting the dataset into quartiles based on the final ranking of teams following a match. Here he finds that making it into the top five isn't a task that requires a lot of kills which many might think with a "King of the Hill" style of game. Through his next visualizations, the same theme occurs. He finds that a lot of kills is necessary to be a winner otherwise known as being part of the "Chicken Dinner Club" with the majority of winners falling into the three to six kill range.

PROPOSED WORK

1 **Data Cleaning**

- Shorter games in the dataset which have missing data must be removed.
- Longer games that have few missing values will have those values filled with placement values.
- If the number of players or teams in a game is too small then remove this data
- as well as it might skew the information we want to gain.

2 **Data Preprocessing**

- Reducing the number of attributes through correlation analysis.
- Removing columns which have no association with match statistics like player_name and team_id.
- Transforming data when necessary, for example if percentages are stored as integers or floats greater than 1, then convert them to their corresponding decimal values.
- Normalization of attributes.

- Removal of games that are not played on the map of Erangel seen in figure 1.

3 Data Integration

- Merging the first few of the separate csv files for the deaths and aggregates before a integration of the final deaths and aggregates all of which are provided in the PUBG Match Deaths and Statistics Kaggle page.
- The remaining csvs will be saved for testing our classification model

4 Data Warehousing

- Storing the historical data in a PostgresSQL Data warehouse which then can be queried for the data mining process.

5 Data Mining

- Analyze the data to gain useful knowledge and answer our initial questions.
- Classification of aggregates that will yield a win and those that will not.

6 Data Visualization

- Presenting the information found in a manner that is understandable by players of the game and non-players alike.
- Heat maps of the “dangerous” areas of the game, most likely drop areas which results in a win.

Figure 1: **Map of Erangel which will be used with geographic player locations to produce a heatmap of player deaths. This map was produced by Reddit user /u/c_a_turner on the /r/PUBATTLEGROUNDS subreddit. Other maps can be found at https://www.reddit.com/r/PUBATTLEGROUNDS/comments/6g1u2v/alternative_pubg_maps_topographic_realistic_raw/.**

Our proposed work looks to be different that the previous literature survey. While these previous works have created data visualizations and used Kaplan-Meier to show trends and estimate survival respectively, we hope to make more fleshed out data visualizations which incorporates the Erangel map more as well as creating our own survival prediction score using the attributes provided in the dataset.

DATA SET

For this project, we are going to be using the dataset PUBG Match Deaths and Statistics which can be found on Kaggle[1]. This dataset was supplied by Kaggle user skihikingkevin or Kevin

[1] https://www.kaggle.com/skihikingkevin/pubg-match-deaths/home

Pei who compiled the results of over 720,000 matches extracting the data from pubg.op.gg which tracks games of PUBG. Currently, the dataset is on Thomas's and Matthew's machine. As the data set provides four csv files for all of the deaths that occur during the game as well as the aggregate data for each player during a match this means each csv file has over 13,000,000 rows of data. As this is the case, we are going to use the first one or two csvs as a training set for our classification on if a match was won or not while one of the remaining csv files can be used as a testing set for our classification model.

EVALUATION METHODS

Ideally with the correct results we would be able to create prediction scores that will be used to predict winners for a given match and players statistics. We would also like to predict match results given certain scenarios and environmental factors within a game such landing locations, total number of teams, equipment picked up, etc. These factors can be displayed on heat map data visualizations which portray "hot spots" around the two different maps of the game, scores can be given to different areas regarding their "dangerous" levels. Classification will provide a prediction model so that we can try to see if given a set of attributes that correspond to the ones provided in the data set how likely is a player to win. In addition, time permitting, a interactive data visualization can be created using NodeJS and d3js in which people can drop a location for a starting point during the game and see the likelihood of winning or the relevant statistics regarding that locations e.g., the average amount of damage that occurs here during a match hinting at how dangerous the area is or kills that occur in the area.

Figure 2: **Another, less realistic but monotone map of Erangel which can be used to produce heatmaps for data visualizations. This map was produced by Reddit user /u/c_a_turner on the /r/PUBATTLEGROUNDS subreddit. https://www.reddit.com/r/PUBATTLEGROUNDS/comments/6g1u2v/alternative_pubg_maps_topographic_realistic_raw/.**

TOOLS

Throughout this project we will be using a variety of tools much of which will revolve around Python as we will use it for our main programming language. With Python we can clean and integrate the data, create data visualizations using libraries like Numpy, Pandas, and SciPy for machine learning as well as using Python Cubes to create a data warehouse for our historical data with the potential to use it for further cleaning and integration. For project reports and documentation, we will be using Google Drive and Docs and for the centralized project repository and milestones we will be using GitHub.

MILESTONES

By the progress report, we plan to have our data set cleaned, integrated, and hosted on a Python Cubes OLAP server. Once we have accomplished that, we can mine the model and data warehouse created with Python cubes using drill-down and aggregation techniques. By the final report, we will have the final project completed with the data mined for the interesting questions, correlations, and analysis we are looking to find as well as creating data visualizations which we hope will provide a more statistical insight of the game, its mechanics and scenarios in which players can learn or understand how to better play PUBG. Time permitting, this would ideally be interactive so a user could experiment and try different areas to drop giving certain scenarios that would allow us to show relevant information.

1 Milestones Completed

Currently, we have achieved cleaning and preprocessing the data as well as uploading the data onto a local MySQL server which will provide to be a back end to the Data Warehouse we have set up with Python Cubes. Through Jupyter Notebooks, we were able to clean the data with Python's Pandas and NumPy libraries in which the results were stored into a .csv file. Using Python scripts with the same libraries, those .csv files were then uploaded into a MySQL local server which provides to be the backend for our cube models in our data warehouse. Panda's powerful dataframes and methods such as to_sql were used to format the data frames to a relational model. The process of uploading our data to a MySQL backend took a series of days with the first few .csv files containing over 13 million rows of data meaning the data had to be uploaded as chunks to the database otherwise the to_sql function caused extensive memory use and crashing. Now that this is completed, we are set up to do the data mining process and to discover correlations and begin the classification process to determine the outcome of certain scenarios with the attributes from the dataset.

2 Milestones Todo

Now we have our setup phase completed, we can begin to use drill down and other aggregation techniques to start truly exploring our data. Over the course of the next few weeks, we are going to start building our classification model which will help provide predictions for whether or not games are winning matches based on certain attributes for the classification, for example the drop location and the locations where fights between teams occur the most. After we have finished our classification model we will be able to produce visualizations and heatmaps of the best locations to drop according to our model's predictions.

I. Segregate our classification/training data that will be used for the model.

II. Train the model using this segregated data.

III. Test accuracy of model against other data to see how are predictions match up to the expected results.

IV. Create Visualizations and heatmaps from our accurate models predictions and the expected results.

V. Analyze the heatmaps and give some suggestions to players about how to play PUBG better.

In order to accomplish these milestones, Python scripting will be employed in addition to a mix of Python for static data visualizations, and JavaScript for potentially interactive visuals.

RESULTS SO FAR

As we have only completed the pre-processing and integrations steps so far, we have limited results that can be observed. However, we can now start the steps to determine correlations, create data visualizations, and begin the classification of winning matches. Currently, we have found that we have an extremely large amount of data to mine and explore, meaning that there are plenty of substantial correlations and predictions waiting to be found, we just haven't been able to find them without an appropriate data model/cube to work against.