Report on
Battlegrounds Analytics: Unearthing Victory

Akash Verma *Nidhi Verma* *Sravani Reddy*

1 Problem Statement

The problem is to identify, address, and mitigate technical, social, and gameplay-related aspects effectively, ensuring a more enjoyable, fair, and enriching PUBG gaming experience for all players and stakeholders.This project seeks to understand and tackle these issues comprehensively, aiming to enhance various features of PUBG gameplay and its surrounding ecosystem.

2 Introduction

In PUBG (PlayerUnknown's Battlegrounds), there can be up to 100 players in a single game. Each game is identified by a unique match ID (matchId). Players in PUBG can either play alone, in pairs (duos), or in groups (squads).

The goal of the game is to survive and be the last player or team standing. At the end of the game, players or teams are ranked based on how well they did compared to the other players or teams.
The project focuses on conducting Exploratory Data Analysis (EDA) to thoroughly understand these issues and their underlying causes. By leveraging data-driven insights, the goal is to enhance various aspects of PUBG's gameplay and its broader ecosystem.
Through rigorous data analysis and the implementation of targeted strategies, this project seeks to contribute to the ongoing evolution of PUBG, ensuring that it remains enjoyable, fair, and engaging for both current and future players. The findings and recommendations generated from this EDA will serve as a valuable resource for PUBG developers and the gaming community at large, ultimately striving to elevate the gaming experience within the PUBG universe.

3 Motivation

a) **Research and Analysis**: Some individuals or groups may be interested in studying player behavior, in-game dynamics, or trends within PUBG. This can be motivated by a desire to contribute to academic research or gain insights into gaming culture.
b) **Competitive Gaming**: PUBG offers competitive gaming opportunities, including tournaments with prizes. A project could be motivated by a desire to compete at a high level, win tournaments, and gain recognition in the gaming community.
c) **Mental Challenges**: PUBG can be mentally challenging, requiring players to think on their feet and adapt to rapidly changing situations. Some may be motivated by the intellectual challenge the game offers.

4 Dataset Description

The dataset is available online on Kaggle here which consists of large number of anonymized PUBG game stats, formatted so that each row contains one player's post-game stats. The dataset have 1111742 rows and 30 columns with a mix of 20 columns having integer 6 columns having decimal and 4 columns having string

values.
The description of the features are as follows:-

- DBNOs- Number of enemy players knocked.
- assists- Number of enemy players this player damaged that were killed by teammates.
- boosts - Number of boost items used.
- damageDealt- Total damage dealt. Note: Self inflicted damage is subtracted.
- headshotKills- Number of enemy players killed with headshots.
- heals- Number of healing items used.
- Id- Player's Id
- killPlace- Ranking in match of number of enemy players killed.
- killPoints- Kills-based external ranking of players. (Think of this as an Elo ranking where only kills matter.) If there is a value other than -1 in rankPoints, then any 0 in killPoints should be treated as a "None".
- killStreaks- Max number of enemy players killed in a short amount of time.
- kills - Number of enemy players killed.
- longestKill - Longest distance between player and player killed at time of death. This may be misleading, as downing a player and driving away may lead to a large longestKill stat.
- matchDuration - Duration of match in seconds.
- matchId - ID to identify matches. There are no matches that are in both the training and testing set.
- matchType - String identifying the game mode that the data comes from. The standard modes are "solo", "duo", "squad", "solo-fpp", "duo-fpp", and "squad-fpp"; other modes are from events or custom matches.
- rankPoints - Elo-like ranking of players. This ranking is inconsistent and is being deprecated in the API's next version, so use with caution. Value of -1 takes the place of "None".
- revives - Number of times this player revived teammates.
- rideDistance - Total distance traveled in vehicles measured in meters.
- roadKills - Number of kills while in a vehicle.
- swimDistance - Total distance traveled by swimming measured in meters.
- teamKills - Number of times this player killed a teammate.
- vehicleDestroys - Number of vehicles destroyed.
- walkDistance - Total distance traveled on foot measured in meters.
- weaponsAcquired - Number of weapons picked up.
- winPoints - Win-based external ranking of players. (Think of this as an Elo ranking where only winning matters.) If there is a value other than -1 in rankPoints, then any 0 in winPoints should be treated as a "None".
- groupId - ID to identify a group within a match. If the same group of players plays in different matches, they will have a different groupId each time.

- numGroups - Number of groups we have data for in the match.
- maxPlace - Worst placement we have data for in the match. This may not match with numGroups, as sometimes the data skips over placements.
- winPlacePerc - The target of prediction. This is a percentile winning placement, where 1 corresponds to 1st place, and 0 corresponds to last place in the - - match. It is calculated off of maxPlace, not numGroups, so it is possible to have missing chunks in a match.

5 Existing Analysis

1. One of the existing work did visualization of the dataset and finds out diverse features types including numerical as well as categorical features. It also finds out that there are various outliers in multiple column, average kills per person, maximum number of players killed in short time, match count per match type and finds that most played matchtype is squad-fpp and least played matchtype is normal-duo.

They also mentioned in their study that the duration of a PUBG match seems to have little impact on the ultimate winPlacePerc but some players have managed to secure victory in just a little over 2 minutes, while more typical winning durations fall in the range of around 1400 to 1850 seconds. Therefore duration of match is not useful feature which can be used for predicting the winPlacePer.

It also describes that what are the effects of using heal items and boost items are used compared to each other and observed that using few healing items increases the chance of winning. Few of the graphs related to above analysis are shown below.

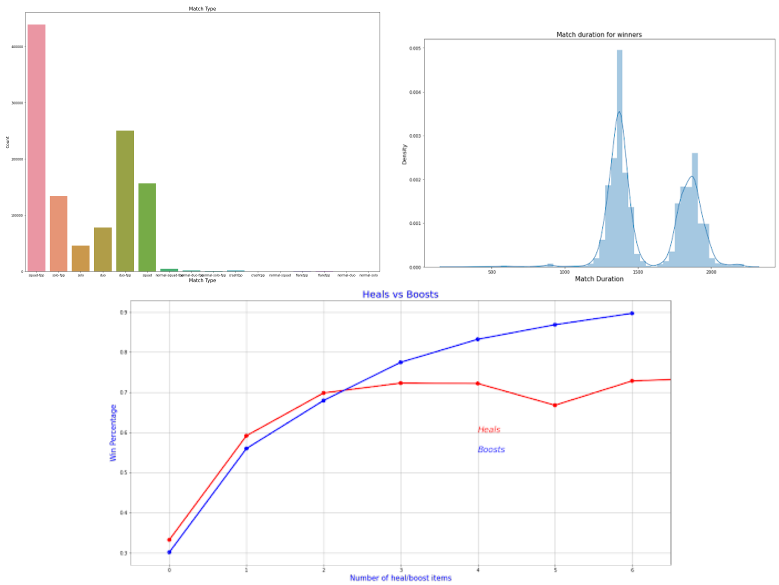

Figure 1: Analysis of Dataset, a)count vs match type, b) match duration vs density, c) win percentage vs no of heal items.

2. This existing work is on pubg dataset which is most relevant to our dataset . It derived key insights regarding player behavior and strategies for increasing the chances of winning.
The EDA implied the following results:
i) killer players-The 99th quantile of kills was computed, which revealed that 99% of values fell below 7 kills. The plot indicated that most players do not eliminate opponents.Additionally, for non-killer players (those with 0 kills), a distribution plot of damage dealt was created. The majority of non-killer players were found to inflict minimal damage to opponents.

ii) Team size Analysis-The distribution of match types (Solo, Duo, Squad, etc.) was visualized. It was observed that squads, especially in FPP mode, were the most common match type.The data was then transformed to combine similar match types for ease of analysis.
iii) Movement Analysis-The relationship between winning and walking distance was examined. A scatter plot indicated a positive correlation, suggesting that players who walk more tend to win.A point plot was generated to demonstrate the impact of destroying vehicles on the chance of winning. Destroying atleast one vehicle was associated with an approximately 35% increase in the chance of winning.
iv) Boosts and Healing Elements-The correlation between winning and healing elements, as well as health boosts, was explored. Healing elements showed a correlation of approximately 0.43 with winning, while health boosts exhibited a correlation of 0.634.

Feature Engineering is also done here and new features were created, including the total number of players in a match, a combined feature for heals and boosts,and a total distance traveled feature. The team size was transformed into a categorical variable indicating 1-player teams, 2-player teams, or 4-player teams. Skewed columns (damage dealt and total distance) were transformed using a cube root transformation.The plots of the above implications are as follows:

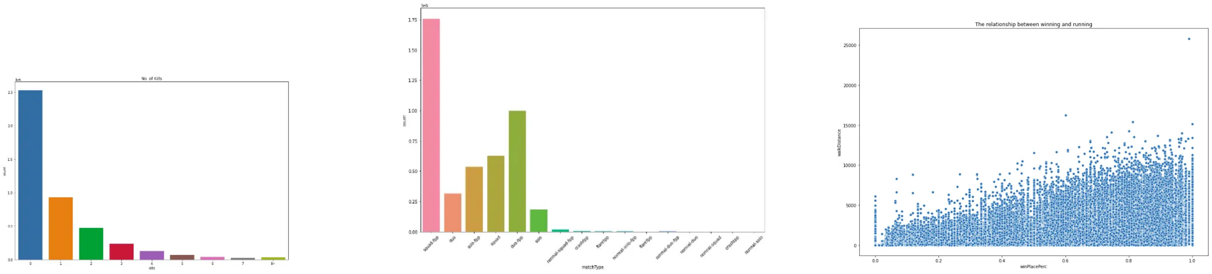

Figure 2: Analysis of Dataset, a) kill vs count, b) match type vs count, c) winning vs running.

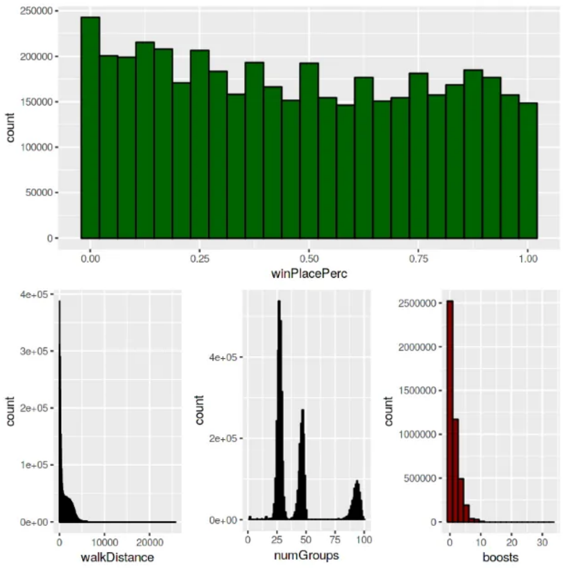

Figure 3: count vs win place percentage.(wrf analysis-3)

3. This work deals with pubg dataset which is mostly relevant to our dataset with some additional rows and features and tries to help in increasing the winning chances of the players.The EDA and feature engineering includes the following implications:

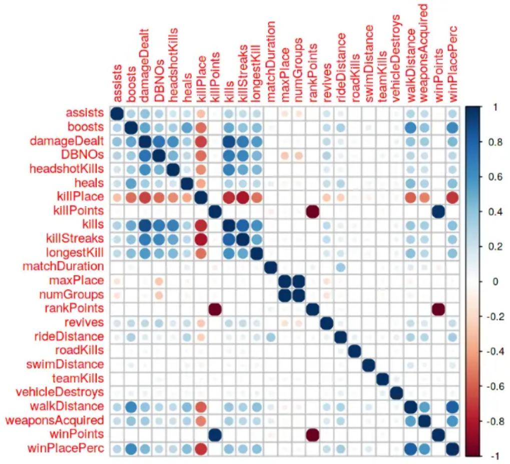

Figure 4: Feature Mapping.

i) walk-distance-It is identified that walk distance is a critical factor in winning. Since the running speed is relatively constant, players have to keep moving to increase their chances of winning.
ii) Boosts-can help players stay alive longer, which is crucial for survival in the game.
iii)Number of Groups: It is noticed that the number of groups in the dataset corresponds to different game modes, such as Squad, Duo, and Solo modes. Values less than 10 may indicate custom games or disconnect errors.
iv) Correlation Analysis: They have performed correlation analysis to identify relationship between variables. Some key findings include: Positive correlation with Walk Distance, Weapons Acquired, and Boosts.Negative correlation with Kill Place, indicating that a lower rank (higher number) is associated with a higher chance of winning.v)Kills: They have created a flag for players with more than 40 kills, possibly identifying exceptional players or cheaters.The following are the plots for the above implications:

6 Hypothesis

There can be many hypothesis which can be inferred from the dataset. Some are:-

1. **Hypothesis 1**: The average kill count per match is higher for players with more playtime (total hours played) compared to those with less playtime.
   Null Hypothesis (H0): The average kill count per match is the same for players with more playtime and those with less playtime.
   Alternative Hypothesis (Ha): The average kill count per match is higher for players with more playtime than those with less playtime.

2. **Hypothesis 2**: Players who land in highly populated areas (e.g., major cities) have a higher chance of surviving longer in a PUBG match compared to those who land in less populated areas (e.g., remote areas).
   Null Hypothesis (H0): There is no significant difference in the survival time between players who land in highly populated areas and those who land in less populated areas.
   Alternative Hypothesis (Ha): Players who land in highly populated areas have a higher survival time than those who land in less populated areas.

3. **Hypothesis 3**: There is a correlation between the number of kills a player achieves in a match and their final ranking in that match.
   Null Hypothesis (H0): There is no significant correlation between the number of kills and final ranking.
   Alternative Hypothesis (Ha): The number of kills is positively correlated with the final ranking.

4. **Hypothesis 4**: The choice of weapon (e.g., assault rifle, shotgun, sniper rifle) significantly affects a player's average kill count per match.
   Null Hypothesis (H0): The choice of weapon does not significantly affect a player's average kill count per match.
   Alternative Hypothesis (Ha): The choice of weapon has a significant effect on a player's average kill count per match.

5. **Hypothesis 5**: There is a difference in the average playtime (total hours played) between players who play solo matches and players who play squad matches.
   Null Hypothesis (H0): There is no significant difference in average playtime between solo players and squad players.
   Alternative Hypothesis (Ha): Solo players have a different average playtime compared to squad players.

6. **Hypothesis 6**: The region (e.g., North America, Europe, Asia) in which a player is located affects their preferred game mode (e.g., TPP or FPP).
   Null Hypothesis (H0): The region of the player's location does not significantly affect their preferred game mode.
   Alternative Hypothesis (Ha): The region of the player's location is associated with their preferred game mode.

7. **Hypothesis 7**: There is a difference in the average survival time between players who use voice chat for communication with their squad and those who do not use voice chat.
   Null Hypothesis (H0): There is no significant difference in average survival time between players who use voice chat and those who do not.
   Alternative Hypothesis (Ha): Players who use voice chat have a different average survival time compared to those who do not.

7 Pre EDA Observations

The link for Pre-EDA observation is here: Visit the Dataset profile report
We observed that many of the researchers have used this dataset to analyze winner in the game and important features which are leading towards finding the winner. Our aim excels towards identifying Hacks used in the game to ensure fair play in the game and popularity of favourite gameplay and startegy in the game, through which we can help build obstacles, NPC and more amusments which make the user experience better and smooth. We identified Null Values in the dataset and we found it clean except some rows, we dropped the rows. Also, the dataset is sparse,so we need to identify important columns which makes the data useful. We plotted correlation matrix to identify which features are highly correlated.

From the correlation matrix we know,

- If correlation is positive, one variable increases with other.
- If correlation is negative, as one variable increases, the other decreases.
- If correlation is 1, it means that either the variables are same or they are almost same.

From the Figure above we can see that many columns are having very high correlation between them (greater than 0.8). Some are kills-damageDealt, killStreaks-kills, winPoints-killPoints, etc. These features can play an important role during training a model.

Also, further Feature Engineering and many other techniques can be applied to create important features and extract other features and Information Gain can be extracted to identify edge of one feature over other. This project can be further extended to extract a lot of information from the dataset to understand gameplay.

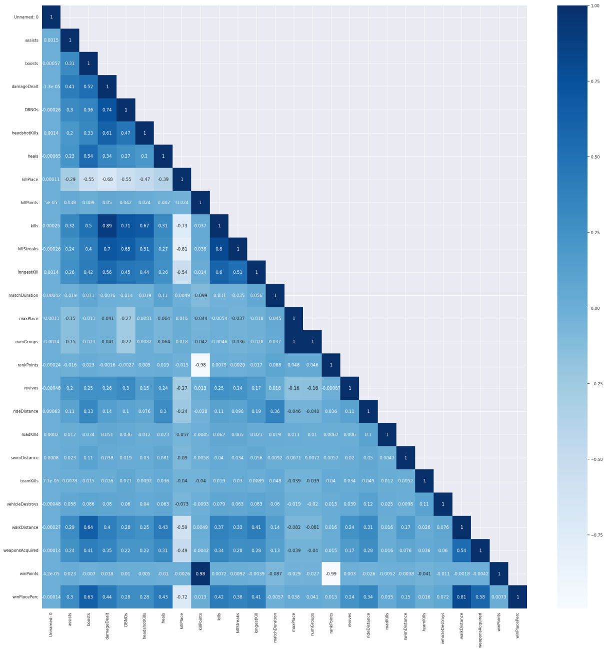

Figure 5: Correlaton of Different Features with Target Column.