More Than Just Kills:
Meta-clustering player behaviors and playing styles in PLAYER UNKNOWN'S BATTLEGROUNDS (PUBG)

Thesis submitted in partial fulfillment of
the requirements for the degree
Masters of Science in Data Science: Business and Governance
at the Faculty of Humanities and Digital Sciences
of Tilburg University

Muhammad Daud
ANR: 275452, **SNR**: 2008311
Supervisor: Prof. dr. ing. S.C.J. Bakkes
Second Reader: Prof. dr. ing. P. H. M. Spronck

Date: June 2018
Faculty: Tilburg School of Humanities and Digital Sciences
Department: Data Science
Masters: Data Science Business & Governance

Preface

I would like to like to acknowledge Professor Bakkes for his support and advice throughout this year. He has been an incredible help in helping me narrow down the focus of my research, providing feedback and guidance on how to tackle different technical issues. I would also like to thank Maud Menard for her constant help, as she kept me on track to accomplish my goals and SonnySunTV for his expertise on PUBG. Without their tremendous help, this project would not been possible. Lastly, a special thanks to Justin Moore for compiling and publishing the PUBG dataset.

Abstract

Player behavior and playing styles, over the past few years, has become an integral part of game analytics. With the advent of new game genres, such as the Battle Royale, it has become imperative to explore in-game player behavior in detail and extract actionable insights. In this research thesis, different player behavior and playing styles are explored in the popular game PLAYER UNKNOWN'S BATTLEGROUNDS (PUBG). The study puts forward and implements an improved version of the K-Means algorithm to find meta-clusters of player behaviors and playing styles. The K-Means-o algorithm helps identify different clusters of player behavior and playing styles, detects and removes outliers relative to the clusters. Upon conclusion, we find several clusters within the following three categories: Combat, Support and Movement that identify different traits and strategies that players undertake in the PUBG squad mode.

Table of Contents

1 Introduction

The conception of simple games and simulations, as a part of academic research, in the 1950's can be traced back to as the birth of digital gaming we know today. From 8-bit arcade games to virtual worlds full of characters, the rise in gaming skyrocketed with the advent of personal computers and gaming consoles in the 1990's. The gaming industry is now one of the biggest and most profitable entertainment industries in the world (Fast Company, 2013). In 2015, video games pulled in more revenue than the music and movie industries combined (Nasdaq, 2018) and have established themselves as a dominant part of the entertainment industry. Globally, the gaming market is estimated to be worth around 100 billion US dollars as of 2016 (Venture Beat, 2016).

With such a meteoric rise in video games, game developers have worked extensively to bring forward new ideas, genres and styles. For example, in the recent years, we have seen games being developed for new platforms (mobile, VR, consoles) as well as innovative genres (MMORPG, FPS, Simulations, Battle Royale) that have taken the world with storm. Games have become more complex, support more players and have richer graphics and mechanics. It would not be wise to think that gaming is exclusively a leisure activity; competition has always been at the heart of gaming. From what started as getting high-scores in arcade games, competitive video gaming, has now become a global phenomenon (Li, 2017). eSports amass a community of millions who watch competitions and follow teams of professional players, who train full-time to compete for titles and cash prizes. The estimated number of people who watch eSports was reported to be 70 million in 2014 (Warr, 2014) and the latest estimates put the number around 380 million in 2018 (Statista, 2018). This astronomical rise has shaken the sports landscape to such an extent that the Paris 2024 Olympics bid team put forward the initiative to include eSports as a medaling event (BBC, 2018).

Despite the rise in leisure and competitive gaming, there has been relatively little research in the area, as compared to its impact and economic significance. With the increase in the number of online games, new business models and innovations, there has been a surge in the data collected (Bauckhage et al., 2014). Behavioral and player data is critical to understanding dynamics of the game and how game developers can move forward with improvements. This is an emerging area of game analytics where the aim is to understand player behavior in order to continually improve the game's design, ensure good user experience, identify players who do well, those who do not (and could be at risk of leaving the game) and revamp matchmaking (Bauckhage et al., 2014). Similarly, researchers have also looked to cluster player behavior

in online games (such as League of Legends, World of Warcraft) to explore how teams are composed, which ones do well and how this can be used to predict winners (Ong et al., 2015). Other research has been focused on finding different techniques to deal with behavioral data that is usually highly dimensional. Ramirez-Cano, Colton & Baumgarten (2010) propose a new meta-clustering approach to deal with high dimensional data and gain insights about player behavior in open world games. Drachen, Sifa, Bauckhage & Thurau (2012) have looked at telemetry data to cluster player behavior in major commercial games (Tera, Battlefield 2). Game analytics is a new exciting field that is rapidly developing different techniques and methodologies to analyze and interpret data. However, this doesn't come without challenges. Not only is this a new growing field, but the data varies from game to game, and so does the gameplay, style and design. Game developers look for different insights to achieve their own goals and similarly, researchers concentrate on the technical challenges of providing actionable insights from game data.

Such challenges have provided the impetus for my research. My goal is to explore and gain player behavior insights for one of the latest and popular game genres: Battle Royale. Battle Royale is a new format for online multiplayer first person shooter (FPS) games, where 100 players parachute on an island, scavenge for equipment and weapons to kill other players. The last person surviving wins. PLAYER UNKNOWN'S BATTLEGROUNDS (PUBG) is one of the game that features the Battle Royale mode. In addition to this, PUBG is of great significance as it popularized this genre. The game has been in such high demand that it has been released on multiple platforms (PC, Consoles, and Mobile) and have seen other companies try to capitalize on the popularity by releasing more games that feature Battle Royale.

It is interesting and important to explore player behavior in PLAYER UNKOWN'S BATTLEGROUNDS (PUBG) as all players start as equal, with no predefined classes, attributes or weapons. In first person shooter games, this is very uncommon. Most popular first person shooter games like Counter-Strike GO, Overwatch, Halo, Battlefield etc., all have predefined classes that can be chosen, and give the player access to certain weapons, traits and attributes. Here one can expect that different classes of players will behave differently. But since there are no predefined classes, weapons or attributes in PLAYER UNKOWN'S BATTLEGROUNDS (PUBG), extracting player behavior and playing style becomes very important. This leads me to the question: How would players behave and play when they all start on an equal footing in every game?

This research is also motivated by the limitations of past research. First, there hasn't been any recent research on player behavior in Battle Royale games. Secondly, the methodologies and frameworks used cannot be directly applied to the domain of Battle Royale genre. The purpose of thesis is to explore player

behavior in PLAYER UNKNOWN'S BATTLEGROUNDS (PUBG) and identify key strategies that players adopt in the game. Looking at the in-game data, the following research questions will be explored:

- What are the different strategies players adopt based on different skills and metrics?
- How do the players in the general population differ (in terms of strategies and play style) from those that are outliers?

The PLAYER UNKOWN'S BATTLEGROUNDS (PUBG) dataset that will be explored is a new dataset comprising of different attributes and metrics of players for the squad mode. To effectively find insights from this data, I will be proposing an adapted version of K-Means algorithm that allows us to look at the clusters as well as the outliers. Since, every player starts the game at the same point and there are no predefined classes, I will follow a meta-clustering approach that pairs different features together that are representative of different behaviors and playing styles. This will be explained further in detail. In the future, such analysis can be used to improve various aspects of the game, particularly matchmaking where player behavior and style of play are extremely important to now consider. With advancements in gameplay and game design, traditional metrics, matchmaking models and methodologies are not relevant enough. Player behavior and playing styles are a glimpse into the minds of the gamers, and can offer immense insight on how a game could be improved in every aspect.

This thesis is comprised of two main parts. The first half of the research focuses on the past research on player behavior, the techniques and methodologies used, and the in detail exploration of the Battle Royale genre and PUBG. The remaining half presents the empirical analysis to gain actionable insights about player behavior and in-game strategies. Section 2 reviews the theoretical framework and the literature that motivates this study. I will also present detailed information about Battle Royale genre and the PUBG, which will provide us with domain specific knowledge needed for analysis. Section 3 covers the exploratory analysis of the dataset and the development of the new K-Means algorithm needed to cluster player behavior. Section 4 provides the results and lastly, Sections 5 and 6 and 7 tie everything together to discuss the results and their implications, contributions and limitations of the study and conclude the thesis.

2 Theoretical Framework

As mentioned previously in the introduction, the goal of this study is to explore and extract different strategies and playing styles of the players in PLAYER UNKOWN'S BATTLEGROUNDS (PUBG). My research builds upon the work done previously in different research projects and academic articles. The modelling framework developed in the later sections of this thesis is the result of research published in different scholarly articles.

With the rise in popularity of casual and competitive gaming, there has been a parallel increase in interest on research related to this industry, especially in game analytics. This thesis aims to put forward new methodologies to explore and analyze game behavior and playing styles in an unconventional genre. This section presents the review of past literature, in-depth explanation of the Battle Royale genre and PUBG. The review of past literature has two goals. The first is to analyze how different researchers have explored player behavior in different games. The second is to analyze the framework and models, to understand how they have executed their research. The review presents the purpose of the scholarly articles, and a quick discussion of the models used. This is followed by the critique of the empirical analysis and results obtained. Furthermore, I discuss the assumptions and shortcomings in relation to my research questions. Lastly, I conclude this section with the historical background and the rise of the Battle Royale genre and the introduce readers to the world of PLAYER UNKOWN'S BATTLEGROUNDS (PUBG).

Before moving on to discuss different research and methodologies from past literature, it is important to note that all studies that I will mention employ similar techniques to different kinds of data of various games. Unsupervised learning techniques are highly dependent on the data and domain knowledge. That is one of the reasons that I will not explore the models of the studies in great detail. I believe that this would distract the reader and derail the focus of my research study. It is apparent that there has been no recent research on Battle Royale games, which, thus, would require a completely new approach. However, I do discuss different models and techniques that are highly relevant and have inspired my approach.

2.1 Game analytics and player behavior

Bauckhage et al. (2014) discuss that analyzing player behavior in games can be challenging due to the curse of dimensionality. There is a lot of data that is available, that players have control over and different variables that can impact results. The curse of dimensionality means that the data is highly dimensional and tends to behave counter-intuitively and thus, finding patterns can be hard (Bauckhage et al., 2014). In recent years, various methods of player modeling have been introduced. Henderson & Bhatti (2001) put forward neural networks implementations, and Kirman & Lawson (2009) (among others) looked at how artificial intelligence techniques can help identify and model player behavior.

One way of dealing with such data is to apply unsupervised learning methods, and in particular, clustering. Clustering allows for exploration of the data and can identify groups of players that exhibit similar behavior and detect features that influence such behavior (Bauckhage et al., 2014). Within clustering, there are a few different techniques that are routinely used by researchers: K-Means, K-Medoids, Matrix Factorization, and Hierarchical clustering. Bauckhage et al. (2014) emphasize on the fact that these clustering algorithms do not come as one size fits all; they have to be adapted to the data available and in-depth understanding of the foundations is needed to apply them correctly. Moreover, domain knowledge is crucial in identifying different clusters of player behavior.

There have been various research studies to explore and understand player behavior in games. These studies have been motivated by different reasons; from creating new business models, to adapting game designs, to improving game persistence (Bauckhage et al., 2014). Behavioral data from computer games tends to be very highly dimensional. And as mentioned before, clustering player behavior can offer researchers and game developers actionable insights into the behavioral patterns of the players (Drachen et al., 2012).

2.2 Player behavior in strategy games

Gagne et al. (2011) describe, in their research, the analysis of a free to play real time strategy game called Pixel Legions. The researchers, working together with the developers of the game, aim to use telemetry data to analyze player behavior in a match by outlining different factors such as how players play the game, and what strategies they used to win. Their use of telemetry data allows them to study if the players are actually playing the game the way it has been designed by the developers. This also allows them to see the strategies that are predominantly adopted by the players who win most, and what players are

learning due to the mechanics introduced in the different levels of the games. Drachen et al. (2012) put forward different techniques to deal with high dimensional telemetry data to cluster player behavior. They employ K-means and SVM algorithms to cluster players into different sets according to their playing styles. In their analysis for the game 'Tera', a small group of players that have high ranked scores in nearly all features are labelled as 'Elite', players that have low ranked scores in the majority of the features are identified as 'Stragglers'. Cluster of players that perform relatively better than 'Stragglers' are 'Average Joes' & 'Dependables'. These labels show the different spectrums the players lie on where the majority of players are 'Stragglers'. According to their analysis, it is important to further explore the class of 'Stragglers' as they are more likely to leave the game. Similarly, the 'Elites' warrant a closer inspection as well, so game developers can understand why these players do well, and how can they ensure a good mix of players in the game. This can used to study the journey of a player to see how they ascend from the bottom to the top ranked players. One of the limitations of their use of the standard K-Means algorithm is that the algorithm is less useful for finding players that exhibit extreme behavior (outliers) and that's where they use SVM. For my research, my goal is to adapt K-Means algorithm in such a way that it allows me to find a general cluster of player population but also, to identify and look at outliers for each clusters.

Another similar area of research is to explore player behavior to learn about team composition in online multiplayer games. Ong et al. (2015) look at player data from League of Legends, a popular team-based role-playing game to better understand the team compositions and predict game outcomes. The researchers grouped the 120 in-game features to different classes that dictate the character's play style. Ong et al. (2015) normalize the different features, and use two different clustering models: K-Means, and DP-Means. The results of their clusters were extremely accurate, as they were corroborated on by highly-ranked League of Legends players brought on as consultants. The researchers identified 12 different clusters that captured different types of player behavior. For example, 'Ranged Physical Attacker' prefer long range attacks as compared to 'Ambushers' who prefer to use stealth and agility to their advantage in close-ranged combat. It is also interesting to see that they found 2 clusters of players that have no defined way of playing. There could be two reasons for that; these players are either novice or prefer an all-round gameplay style. Accurate clustering player behavior in games with predefined classes can be deemed easier to accomplish. This is because characters with different classes have different attributes, strengths and weaknesses that can highly influence the playing style. To predict win or loss game outcome, Ong et al. (2015) used the following three methods: logistic regression (LR), Gaussian discriminant analysis (GDA) and support vector machines (SVM). The researchers predict win/loss outcomes based on the clustering player behavior and team composition and achieve an accuracy of over 70%. This approach can be used to see how team compositions can impact performance.

2.3 Player behavior in games with no pre-defined classes

Classifying player behavior in a game where no predefined classes influence play styles is a harder problem. Ramirez-Cano et al. (2010) propose new techniques to tackle such problems. In their research, they look at the game data of 'Hunter', a free-roaming hunting game where the objective is to track, spot and kill animals according to the specified rules of ethical hunting. Some of the challenges they faced due to the nature of the data and the game were: first, same techniques cannot be used every time as games measure and provide different kinds of metrics, second, datasets change as the expertise of the players evolves (especially in games where there no predefined classes), and lastly, this results in overlapping classes where players can fall into one or more classes. Ramirez-Cano et al.'s (2010) approach involves meta-classification that breaks the clustering into three different levels of data analysis: action/skill based clustering, similarity-based clustering and social based clustering. The researchers propose a framework of variable classification that consists of 4 classes: action, skills, exploration and social. Action class includes variables that define the player's behavior in basic activities like distance travelled, shots fired etc. The skills class contains features that can be used to analyze the level of player's expertise e.g. shooting accuracy, number of trophies received. Exploration class includes variable based on geographical features which describe the trails explored. And finally, the social class incorporates features that measure the level of social interaction of a player.

For action/skill based clustering, there were 4 distinct groups: shooters (players with high accuracy and kills), photographers/explorers (players who like to explore the world), hikers (players with lengthy trails but no shooting) and all-rounders (players with good tracking, spotting and shooting skills). To accurately capture these clusters, the researchers applied the K-Means algorithm after transforming the relevant variables with PCA analysis. This is done to reduce the high dimensionality of the data. Similarity and social based clustering are not relevant to my research question, thus, I will not be discussing them. One of the key takeaways from the research of Ramirez-Cano et al. (2010) is their approach in classifying player behavior where no predefined classes exist. Their approach of meta-clustering is looking at different variables (instead of all), has informed and inspired my meta-clustering approach, which I will detail in the later sections.

Another interesting research is regarding player behavior in one of the most famous game series; Tomb Raider. Mahlmann et al. (2010) present an explorative study on predicting aspects of player behavior in the game, Tomb Raider: Underworld. Their study is focused on understanding player behavior to

complement business strategies. Mahlmann et al. (2010) use supervised learning algorithms on in-game data to predict when a player will stop playing the game, and if he continues to do, how long it will take him to finish the game. Despite using a lot of features that inform the playing style and behavior, the researchers report the results of the classification techniques are average. This is because of high level noise in the dataset, missing information and aspects of behavior that cannot be captured by in-game features. In a similar project on the same game title, Drachen et al. (2009) focus on modeling player behavior using an unsupervised learning approach. They employ the K-Means clustering algorithm to normalized data and later, use a hierarchical clustering method. Their technique reveals four different kinds of players: Veterans, Solvers, Pacifists and Runners. Veterans are defined as players that die a few times, and their death is mainly due to environment. Solvers are players that are adept at solving puzzles but die often due to environmental reasons such as falling. Pacifist players die more than the players in the other groups and is mainly due to active enemies. Lastly, Runners complete missions very fast but die often because of environmental factors and enemies. One of the things to note is that the researchers normalize the data and look at finding clusters within the majority of the population. Their results do not include players who demonstrate extreme behavior. They conclude that such a study of player behavior can allow developers to adapt the game play according to different players to ensure low churn rate and variation in gameplay.

2.4 Clustering Techniques: incorporating outliers in analysis

Lastly, I want to discuss the research study by Barai & Dey (2017) on detecting and removing outliers using K-Means algorithm. Outlier detection is an important aspect of data mining that is used to detect and remove anomalous data that can occur because of various reasons (system behavior, human error, mechanical faults). Barai & Dey (2017) outline the methodology to detect and remove outliers using K-Means and Hierarchical clustering. I will only discuss their methodology relating to K-Means as it is relevant to my thesis. Though, it is important to note that this study has nothing to do with player behavior in computer games, the approach that Barai & Dey (2017) outline has inspired and informed my methodology. One of the reasons that outlier detection and removal is important is because having outliers in the data reduces the accuracy of the clusters. The researchers propose two ways of detecting and removing outliers with the K-Means algorithm: Distance based approach and Cluster based approach.

Distance based approach consists of three phases. First, the researchers run the K-Means algorithm on the data and find k clusters. They calculate accuracy and the silhouette index of the clusters. The first phase consists of finding a threshold value. This threshold value is calculated by finding the pairwise

distance between all paired observations in the dataset. The distance metric they use is Euclidean distance. Using the maximum and minimum pairwise distance values, they calculate the threshold value as follows:

Threshold value = (max. Distance - min. Distance) / 2

The second phase consists of finding the Euclidean distance of all observations in the dataset. If distance of an observation is greater than the threshold value, it is considered an outlier. And lastly, in the third phase, all outliers are removed from the dataset and the K-Means algorithm is used again to find the clusters, and the accuracy is recalculated. This time, the researchers expect the accuracy to be improved.

The cluster based approach entails running the K-Means algorithm, finding the k clusters and calculating accuracy. And then it comprises of two phases. In the first phase, the researchers find clusters that are small, or find small clusters that are a part of a big cluster, and label them as outliers. The reasoning behind this is that since the cluster is small in number, they are assumed to be a cluster or forced to be a part of a cluster. During the second phase, all observations outlined in the first phase are considered as outliers and removed from the data, K-Means algorithm is run again on the new data and accuracy is recalculated. According to the researchers, it is expected that the accuracy of the clusters would improve. For my methodology, I will be using a modified approach based on the study of Barai & Dey (2017). I will be building up upon their study and presenting my methodology in the later sections.

2.5 Battle Royale

Battle Royale is a popular game genre these days. PLAYER UNKNOWN'S BATTLEGROUNDS (PUBG) was one of the first games to capitalize on the Hunger Games styled mode. Recently, there have been other games that have joined the Battle Royale genre, mostly notably Fortnite. While PLAYER UNKNOWN'S BATTLEGROUNDS (PUBG) was not one of the first games to feature Battle Royale modes, it is definitely the one that has popularized it.

The Battle Royale genre takes its name from the Japanese novel of the same name. The novel follows the story of a group of high school students who are forced by the government to fight to death until there is one person remaining. This survival setting is inspired by Gladiators in Ancient Rome and also, is an integral part of the Hunger Games book series. The Battle Royale genre blends different elements into one: survival, scavenging, and exploration. The Battle Royale games feature nearly the same format

where players drop simultaneously into a region, search for weapons and armor and eliminate opponents, while avoiding being trapped by the shrinking playable area.

As of February 2018, 30.1% of core PC gamers all over the world played Battle Royale games (Newzoo, 2018). These games have been extremely successful that game developers have released Battle Royale versions for consoles (Playstation, XBox, and Nintendo) and smartphones.

2.6 PLAYER UNKNOWN'S BATTLEGROUNDS

PLAYER UNKNOWN'S BATTLEGROUNDS (or called PUBG for short) is an online multiplayer Battle Royale game. It is developed and published by PUBG Corporation, which is a subsidiary of the game developer, Bluehole. PUBG has an interesting story of birth. This game is based on mods (modification to a video game) that were created by Brenden "PlayerUnknown" Greene, who used the Battle Royale genre for inspiration. With the success of the mods, Bluehole and Brenden began working on an independent game title that would feature the Battle Royale mode. Thus, PLAYER UNKNOWN'S BATTLEGROUNDS (PUBG) was born in March 2017.

PLAYER UNKNOWN'S BATTLEGROUNDS (PUBG) is a player versus player, last man standing FPS game where players can compete in solo, duo (a team of two) or squad (a team of four) mode. The goal, however, stays the same. Last person or team wins. Each match starts with a plane that flies over the island, where players can choose to jump whenever they want and parachute to different areas of the island. The flight of the plane changes each game, so the players have to adapt quickly to find the best spot to land. The island itself is 8 x 8 kilometers in size, with various locations featuring forests, towns, rivers etc.

All players start with nothing; no armor, no weapons, no gear. Once the players land, the goal is to quickly scavenge the buildings and find different weapons, armor, health kits and gear. These items are distributed all over the map. They are heavily concentrated inside buildings, and in areas where there are a lot of buildings. Thus, to find the best loot, players often go to such areas. However, presence of good loot attracts a lot of players which means that gun fights start immediately. Some players choose to go to crowded areas like towns and city centers while other resort to villages and outposts.

Every few minutes, the playable area of the map decreases in a circle towards a random location. The goal is to create the map smaller and smaller as less players are left alive, and to force players towards each other and into combat. During the course of the game, random areas of the map are highlighted in red and experience aerial bombing, presenting a new threat to players. On top of that, at random, a plane will drop a supply package that contains rare and unobtainable items. Players can navigate around the island on foot, or by using different vehicles (cars, motorbikes, quad bikes). On average, a full round lasts around 30 minutes.

Figure 1: PUBG Gameplay

Figure 2: PUBG Map Erangel

This game is quite complicated and it is very interesting to see the different behaviors that players exhibit in this game. For my research study, I will be looking at squad mode where up to 100 players compete against each other to be the last man standing. The goal is to explore and analyze player behavior to see what kind of strategies to players adopt, and for which strategies to players win often. I will present my hypothesis in detail in the next sections.

3 Experimental Setup

As I have mentioned regularly throughout the previous sections, player behavior and play style are of great importance when it comes to game analytics. Besides profit-making, player analysis provides us with an understanding of how the users are playing the game. Such insights, in the world of gaming, are essential for game developers. With new game genres such as Battle Royale, on which this study focuses, it is imperative to understand how players are playing the game, whether such styles of play expected, and how can we use insights from such analyses to bring forward a better experience for gamers. The purpose of thesis is to explore player behavior and player styles in PLAYER UNKNOWN'S BATTLEGROUNDS (PUBG). Looking at the in-game data, the following research questions will be explored:

1. What are the different strategies that players adopt in PUBG squad mode?
2. How do outliers differ from the general population?

The studies and results of past literature have informed my methodology and approach. This section is divided into three parts: Data collection and description, variable selection and experimental procedure. In the data collection and description section, I will go over the details of the data; collection, description and pre-processing. For variable selection, I will explore in detail the different variables in the data and how they will be used to detect different strategies that players adopt. And finally, in the experimental procedure section, I will develop my framework, as inspired by past literature, on how to accurately capture and identify clusters of player behavior and strategies. For ease of reading, from here onwards, I will refer to PLAYER UNKNOWN'S BATTLEGROUNDS (PUBG) as simply, PUBG.

3.1 Data Collection

The developer of PUBG, Bluehole, has made the game's player statistics and match histories freely available through a web-based programming application interface (API). However, their data policies have been changed over the last few months and now, not a lot of data is accessible. Fortunately, the dataset, collected in July 2017, is available on Kaggle (published by Justin Moore). This dataset includes 87989 random PUBG players and their game statistics. Overall, the dataset includes 152 features per player. These includes all features for solo, duo and squad mode and are aggregated across all regions. For the purpose of this study, I will be only looking at the squad mode features for all the players. This is because it is one of the most popular modes for Battle Royale games and PUBG. The motivation to look at squad data is to find behavior relating to combat as well as supporting your team within the game.

3.2 Data Description

The dataset includes 152 different features and a total of 87989. Since, there are a lot features, and I won't be using them all for my study, I will list the different categories of features[1]. These features are of 7 different categories, as follows:

- **Identifiers:** Player Name, Tracker ID (unique ID per player)
- **Performance:** Rating, Win Ratio, Top 10 Ratio, Number of Wins, Number of Top 10s etc.
- **Combat:** Kills, Assists, Headshots, Damage Dealt etc.
- **Health:** Heals, Revives, Boosts, DBNO (down but not out) etc.
- **Movement:** Ride Distance, Walk Distance
- **Time:** Time Survived
- **Achievements:** Daily Kills, Weekly Kills, Longest Kill, Max Kill Streaks etc.

3.3 Pre-processing

The first step in pre-processing the data is to filter out only the features that are related to squad mode. Next, only players that have played more than 100 games/rounds were filtered out. This was done to make sure that we don't include players that have just started playing the game, as the learning curve is steep. This leaves us with a total of 59911 observations. Since, a lot of the features are not directly relevant to finding out specific player behavior and playing styles, only key features were included in the dataset. So, as a result, we have statistics of 59911 unique players, with 52 features, comprising of different categories as outlined above. For the purpose of the study, here are the features that I will be using for my analysis:

- **Identifiers:** Tracker ID
- **Performance:** Win Ratio, Top 10 Ratio
- **Combat**: Kills per Game, Damage Dealt per Game
- **Support:** Assists per Game, Revives per Game
- **Survival:** Time Survived per Game, Heals per Game
- **Movement:** Walk Distance per Game, Ride Distance per Game

The feature Time Survived per Game was converted from seconds to minutes. Win Ratio and Top 10 Ratio were converted to percentages (instead of ratios). Walk distance and Ride distance were converted to

[1] The complete list of all the features are presented in Appendix I

meters. For all other in-game features, the averages were calculated so we have statistics per round. The averages were calculated by dividing the total statistic with the total number of rounds played. The dataset was inspected to see if we have any anomalous observations, but no such observations were found. The goal is to keep reasonable outliers within the data, so we can compare how these players do as compared to the majority of the players. This is essential to understand why some players are do well (or worse) as compared to others. All features were normalized over their range of values, to prevent clusters from forming due to the order of magnitude. For example, Damage Dealt per Game is several magnitudes higher than Revives per Game, which means that small variations in Damage Dealt per Game are considered much more important than Revives per Game.

3.4 Variable Selection

In addition to this, I will be using other variables from the dataset to compare behaviors and playing styles between different clusters. I will be detailing these later in the Results section. But before we move forward towards data preparation and pre-processing, here are the descriptions of these features:

- **Tracker ID:** Unique ID per player
- **Win Ratio:** Number of wins as a ratio of the number of total rounds played
- **Top 10 Ratio:** Number of times the player was placed Top 10 as a ratio of the number of total rounds played
- **Kills per Game:** Average number of kills per game
- **Damage Dealt per Game:** Average damage dealt per game
- **Assists per Game:** Average assists per game. An assist is defined as when you deal damage to an opponent, but the opponent is killed by another player
- **Revives per Game:** Average revives per game. When a player's health goes down to zero, he enters the DBNO (down but not out) mode where they can only crawl. Teammate can revive the player, and he can continue the game without dying
- **Time Survived per Game:** Average time survived per game. It is important to note that on average the game lasts around 30 minutes
- **Heals per Game:** Average heals per game. When a player incurs damage, he can use bandages, health kits to restore their health. This is counted as Heals
- **Walk Distance per Game:** Average distance walked per game
- **Ride Distance per Game:** Average distance travelled in vehicles per game
- **Headshot Kills per Game:** Average number of headshot kills per game

- **Move Distance per game:** Average distance moved (walked and ride) per game

3.5 Experimental Procedure

Classification of player behavior consists of taking player's actions as inputs and finding insights about these actions grouped as playing styles. Simple classification methods of looking at complete data would not be able to offer us the insights we need. This is because all players converge towards the same point when you look at the data as a whole, as there are minor variations that separate them. Instead, I propose a meta-classification approach in which I look at clusters in different features that can provide us with useful insights. This allows me to take advantage of different characteristics available in the dataset. The meta-clustering approach consists of finding the right features that can tell us about specific playing styles and behavior. The categories are defined as follows:

1. **Combat:** Are players aggressive in approach? Do they pursue the enemy and go for the kill? Or engage when necessary and stay defensive?
   Features: Kills per Game, Assists per Game
2. **Support:** Do players do more damage or more inclined to help their teammates?
   Features: Damage Dealt Per Game, Revives per Game
3. **Movement:** Do players prefer to walk or ride in vehicles?
   Features: Average Ride Distance, Average Walk Distance

3.6 Modeling Framework

K-Means is a popular clustering algorithm used for clustering player behavior. The objective of the K-Means algorithm is to partition the data into k sets of clusters, so that the inter-cluster similarity is minimum and the intra-cluster similarity is maximum. The algorithm operates as follows:

Input:
K: number of clusters
DF: Dataset

Method:

1. Choose K objects in DF as the initial cluster center.
2. Calculate the distance between each point and the cluster centers.

3. Assign the data point to the cluster where the distance between the data point and the cluster center is the smallest.
4. When all data points are placed, recalculate the cluster centers.
5. Repeat 2 and 3 until the position of K doesn't change anymore.

Output:
K set of clusters

For the clustering, I employ the Lloyd's algorithm which consists of randomly choosing observation as cluster centroids and iteratively assigning other points to clusters, as defined in the process above. To find the k number of clusters for each category, I run a 10-fold cross validation to find the local optimizer. One of the drawbacks of the simple K-Means algorithm is that, even if the data is normalized, presence of outliers can heavily influence the results. This is because the K-Means algorithm is sensitive to outliers, as it tries to minimize the error. Presence of outliers in the data shift the centroids of the cluster, and thus, can influence what observations go in what cluster. However, I don't want to completely remove outliers from my data. This is because, in my study, it is desirable to see the behaviors of the outliers, and see why they do well (or worse) as compared to the rest of the players. Removing outliers is thus, not beneficial. For this reason, based on the study by Barai & Dey (2017), I have formulated an adapted version of the K-Means algorithm that allows for the detection and removal of outliers from the clusters, but keeps them in the dataset, so we can visualize where they lie as compared to all the other observations in the clusters.

The 'K-Means-o' algorithm operates as follows:

Input:
K: number of clusters, as calculated through the Elbow Method and domain knowledge.
DF: Dataset

Method:

1. Choose K objects in DF as the initial cluster center.
2. Calculate the distance between each point and the cluster centers.
3. Assign the data point to the cluster where the distance between the data point and the cluster center is the smallest.
4. When all data points are placed, recalculate the cluster centers.

5. Repeat 2 and 3 until the position of K doesn't change anymore.
6. Find distance threshold for each cluster. Distance threshold is calculated as the 0.9 quantile of the distances of all the points from the nearest cluster center.
7. If data point of each cluster is greater than 0.9 quantile of the distances from the cluster center, it is labelled as outlier and removed.
8. Repeat all steps until no more outliers and no changes in cluster centers and K.

Output:

K set of clusters

This model allows me to identify observations as outliers relative to their cluster center. The important takeaway is that these observations are not outliers in the complete dataset, but are considered outliers as compared to their clusters. In the analysis, this helps me see why these observations are labelled as outliers and how these players perform relative to the nearest cluster. Moreover, this approach increases the overall accuracy of the clusters. As I am formulating a new version of the K-means algorithm, by also looking at outliers, I have chosen to present only one algorithm for analysis. This is because first, this is a unique problem and there are no other significant algorithms to deal with such a problem, and second, comparing results between different algorithms (that do not incorporate outliers) would bring forward drastic differences, and that is not ideal, as the models and their respective approaches are not compatible. Even if I could incorporate algorithms, like DBSCAN, that do incorporate outliers in their framework, there is a fundamental difference between K-Means and DBSCAN. DBSCAN is a density-based clustering algorithm as it finds clusters based on the estimated density distribution of the nodes. Comparing results from K-Means and DBCAN would not make sense as they are so fundamentally different. Thus, for these reasons, in this thesis, I concentrate on developing K-Means-o and presenting the analysis based on this algorithm, to see what results we can produce.

All code was implemented in Python and computations were executed on a 2.7 GHz Intel Core i3 processor with 8 GB RAM. Due to the random initializations, I ran a 10-fold cross validation for the K-Means-o clustering algorithm for each category in order to obtain the best locally optimal centroids. The process of finding the optimal number of clusters, outlier observations relative to the clusters, and final clusters are outlined in the Results section.

4 Results

The purpose of this section is to implement the modeling framework, outlined in the previous section, and find clusters within the 3 following categories:

1. **Combat:** Are players aggressive in approach? Do they pursue the enemy and go for the kill? Or engage when necessary and stay defensive?
2. **Support:** Do players do more damage or more inclined to help their teammates?
3. **Movement:** Do players prefer to walk or ride in vehicles?

Before we begin the analysis, it is important to look at the relationship between the features that we are considering. To do that, I look at how these features might be correlated with each other. From the results, it can be seen that performance features are highly correlated with combat features (kills, assists) and time survived. This is expected as having more kills and assists would mean that the players survive till the end of the game, and hence achieve higher ranking within the rounds. Similarly, kills and assists are correlated with damage dealt, as expected. Not only these correlations give us an idea how the features are related, but more importantly, that my approach of meta-clustering won't be hindered by these features being correlated. The correlations between all the variables can be seen in the visualization below:

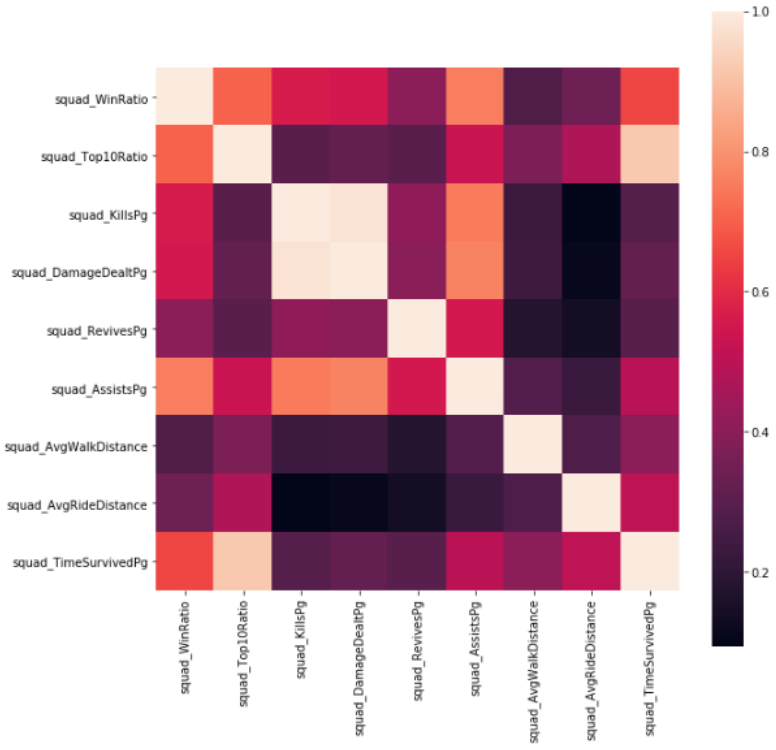


Figure 3: Correlations between all features used for clustering

For this research study, I also consulted with semi-professional PUBG players, and they corroborated on the different strategies that player's exhibit in the squad mode in PUBG. Moreover, I have extensively played PUBG and other Battle Royale games, and do have in-depth knowledge about the game from a player's perspective. All this suggests that the clusters are intuitively correct, which were selected using the Elbow Method and domain knowledge. For the purpose of the analysis, I will go over the categories one by one, present and interpret the different clusters within these categories on how they relate to different player strategies.

4.1 Combat

The number of K-Means-o clusters for Combat are a total of 3, and are explained and visually represented below:

1. **Cluster 1: Stragglers**

   Players that have overall low assists and kills per game
2. **Cluster 2: Mediocres**

   Players that are right in the middle, they get a good amount of assists and kills per game. There are three different groups here, players that have high assists as compared to kills, high kills as compared to assists, and high assists and kills (as compared to other players within the cluster)
3. **Cluster 3: Elites**

   Overall, the players have on average high kills and assists. As with Cluster 2, there are three different groups within the cluster, players that have high assists as compared to kills, high kills as compared to assists and high assists and kills.

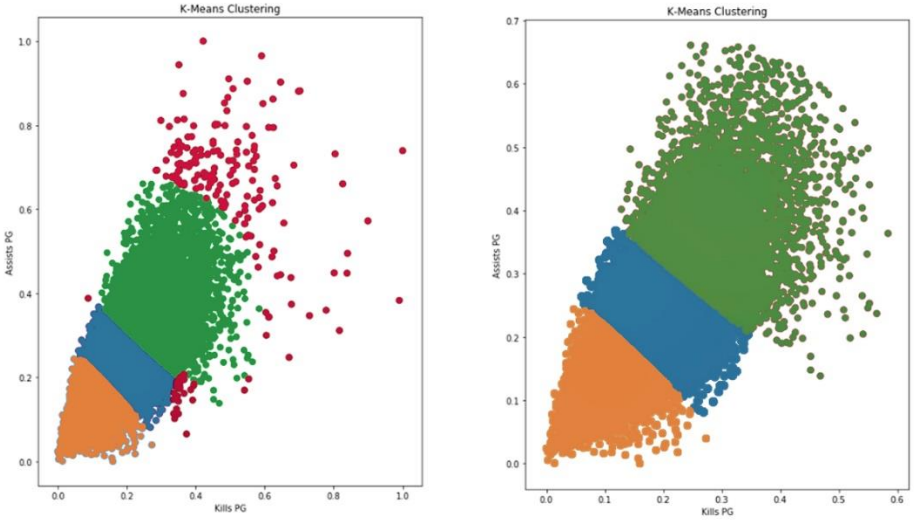


Figure 4: Combat Clusters. Red = Outliers, Green = Elites, Blue = Mediocres and Orange = Stragglers

With the box plots[2], we can see that the Elites have on average higher Win Ratio, followed by the Mediocres and then the Stragglers. Also, Elites are also dominant in time survived per game, even though the number of rounds played are on average the same between all these classes. This definitely shows that there is clear difference in level of skill. Next, Elites have a higher headshot kill ratio and tend to move a lot in the game. Not only is the accuracy of the players higher, but they seek out enemies, pursue them and kill them.

Outliers:

For Cluster 2, the Mediocres, there are two groups of outliers, relative to the player population in the cluster. First, there is one player that has relatively high assists per game, but very low kills per game. Second, players that have relatively high kills per game, but low assists per game. For Cluster 3, the Elites, there are two distinct groups of outliers as well. Some of these players have exceptionally high kills per game, assists per game or both. Overall, there is no major difference or something unexpected between the clusters and their outliers.

4.2 Support

The number of K-Means-o clusters for the Support are a total of 4, and are explained and visually represented below:

1. **Cluster 1: Recons**
   Players that low revives per game, and low damage dealt per game.
2. **Cluster 2: Medics**
   Players that have high revives per game, and relatively low damage dealt per game
3. **Cluster 3: Assault Team**
   Players that have high damage dealt per game and relatively low revives per game.
4. **Cluster 4: Squad Leaders**
   Players that are all-rounders. They have high damage dealt and revives per game, on average. There are three different types of Squad Leaders: those who have high revives but relatively low damage dealt, high damage dealt and relatively low revives and high revives and damage dealt.

[2] The box plots for the Combat clusters are presented in Appendix II

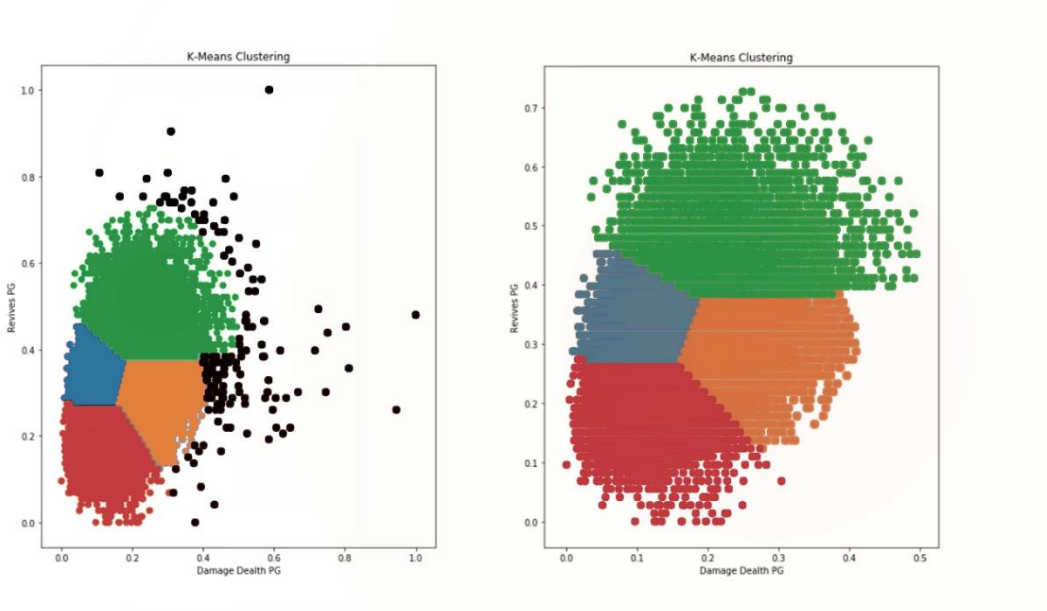


Figure 5: Support Clusters. Black = Outliers, Green = Squad Leaders, Blue = Medics, Orange = Assault Team and Red = Recons

Furthermore, looking at the box plots[3] of how these clusters are distributed when looking at other features. Recons have relatively low win rate, top 10 rate, distance moved and time survived. These players are on the lookout, staying stationary and are most often exposed in combat. These players often drift away from the rest of the squad members to provide information. However, they do have good shooting accuracy, but since they are shooting from longer distances, their kills are lower. The Medics do relatively better as compared to Recons. They have a relatively higher win rate and top 10 rate as compared to Recons. They have low kills and assists as compared to Assault and Squad Leaders. For Medics, the priority is to provide support in terms of medical assistance and firepower. The Assault team have relatively high win and top 10 Ratios. In terms of kills, they are on par with Squad Leaders but have the highest accuracy and headshot kills as compared to other players. Moreover, they are more likely to incur damage and die. Finally, the Squad Leaders dominate other players in nearly every aspect. They have a good sense of the game, and lead the team on what to accomplish. They have high win and top 10 ratios. Squad Leaders have high assists, which means they approach and engage enemies first. They have relatively lower accuracy but higher distance travelled.

[3] The box plots for the Support clusters are presented in Appendix III

Outliers:

As far as the outliers are concerned, there are two distinct groups. For the Squad Leaders, there are players that have unusually high revives per game, and/or damage dealt per game. For the Assault Team, there are players that have high damage dealt per game, as compared to their cluster average. Overall, there is no major difference or something unexpected between the clusters and their outliers.

4.3 Movement

The number of K-Means-o clusters for the Movement are a total of 2, and are explained and visually represented below:

1. **Cluster 1: Walkers**

   These players prefer to walk.

2. **Cluster 2: Drivers**

   These players prefer to drive and move around in vehicles.

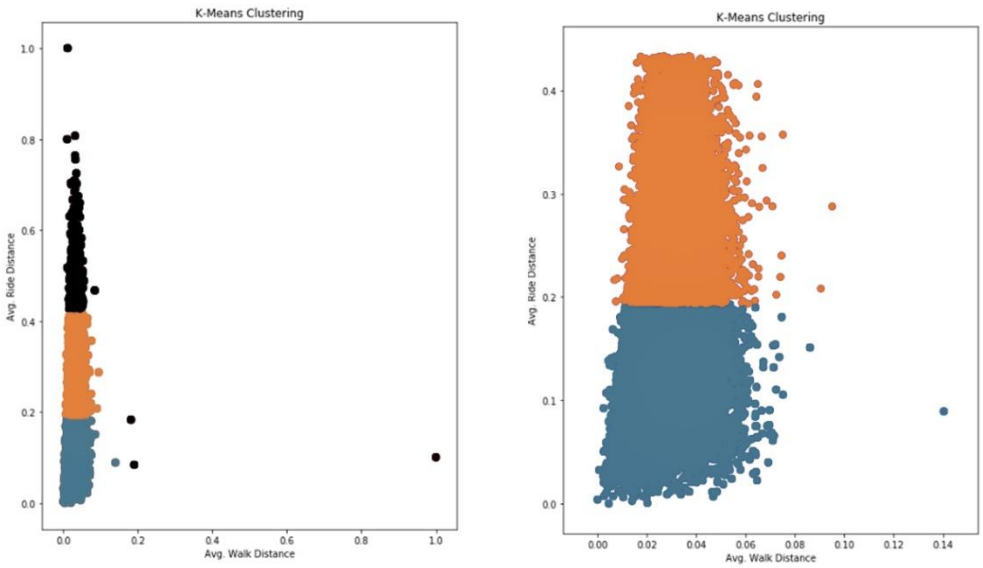


Figure 6: Movement Clusters. Black = Outliers, Blue = Walkers and Orange = Drivers

From the box plot distributions[4], we can clearly see that Drivers tend to survive longer in the game. This is because they can evade enemies faster, explore areas quickly and stay within the playable area of the map more. This results in the fact that they finish higher than walkers. In all other aspects of combat

[4] The box plots for the Movement clusters are presented in Appendix IV

engagement, they are nearly the same. This means that the mode of movement, really impacts how well players do in the game.

Outliers:

There are two distinct groups of outliers. Players who tend to walk way more than other players, and players who prefer vehicles more than anything else. These players are not so different from the walkers and the drivers, they are just considered outliers as the extent to which they walk or ride is way higher as compared to the means of the rest of the players. Overall, there is no major difference or something unexpected between the clusters and their outliers.

5 Discussion

Game analytics, and in-game player behavior specifically, has risen in demand in the last few years. The applicability of such analysis is key to improving different areas in the gaming industry; from innovative business models, introducing new genres, to improving the gameplay for the gaming community. This research study tries to explore different player behavior and playing styles in the PUBG squad mode. In particular, my goal was to find player behavior through meta-clustering in three different categories: Combat, Support and Movement, considering the dataset that was available.

The results in this thesis show that there are indeed different types of player behavior or playing styles within PUBG. One of the most important aspects of this thesis is that it tries to look at different combinations of features to find clusters, instead of looking at all features at the same time. For some games, this technique works well because, either there are predefined classes or characters have different attributes (even if there are no predefined class structure). The interesting thing about PUBG is that all players in the start of every game are the same, they can have no advantage over the others through weapons, attributes or items. Every decision they make once the game starts; where to land, what weapons and items to take and when to engage, is vital in the final placement.

With the results provided in the previous section, we can see that there are indeed different classes of players when we cluster their behavior according to different combinations of features. These features were chosen because they inform us of different decision making and choices of the players. For Combat, we have three different classes: Stragglers, Mediocres and Elites. For Support, we have four different classes: Recons, Medics, Assault Team, and Squad Leaders. And finally, for Movement, we have two different clusters: Walkers and Drivers. All these classes/clusters show us different behavioral traits of the players and give us an understanding how they behave within the game. The semi-professional PUBG players, that I consulted, corroborated on the different clusters. However, one criticism they put forward was that, considering the nature of the game, players do not stick to these classes all the time, they adapt based on the game. This may be true when looking at individual games, but my goal was to explore overall data and see what behaviors are most predominant within players.

Furthermore, in the cluster analysis, we see that there is some overlap between clusters and player behavior. Some players exhibiting the same behavior (but with different magnitudes) are put in different clusters. This is a problem that we could look in the future with the help of hierarchical clustering. It would

be interesting to further explore these classes, in the future, if we have more data available from Bluehole. For example, match and telemetry data would allow us to see what players are doing in individual matches and at given times during these matches. This would allow us to explore if these strategies of behavior change as the game progresses. In the future, partnership with Bluehole should be considered so we can avail the vast amount of data that they possess. Lastly, I did explore players that exhibit outlier behavior, but found no major differences. Only the magnitude of the features is different, and they show extreme behavior (as expected) but we do not see differences in playing styles as compared to their nearest cluster.

Research on player behavior and playing style is vital in the gaming industry. Such an analysis, as presented in this thesis, can inform game developers, competitive players, and other stakeholders on the dynamics of the game. Not only can such analysis help game developers see how the players are actually playing the game, they can also use this information to create a better gameplay, introduce new features in the game and improve the overall gaming experience. For example, proper matchmaking is vital for any online multiplayer game. Game developers can use player behavior clusters to identify different types of players and pair them with other players so that there is an even distribution of players (with different behaviors and playing styles) to create a good balance. No one likes a game where every player behaves the same way, or there is a dominant playing style. A good balance is key to keep the interest of the players and grow the community. Another example of such an analysis is to explore the players that exhibit anomalous behavior. This can be used to study the progression of player's performance over time, and examine why some players do well and the others do not, and what can the game developers do to stop players from leaving the game. All in all, the potential for such a research study in the game industry is huge, and actionable insights can create a huge impact in the gaming industry.

6 Contribution and Limitations

This study contributes to the current landscape of game analytics and player behavior research on various fronts. As reiterated several times in this thesis, the purpose of this study is to explore different player behavior and playing styles in PUBG using a meta-clustering approach. The purpose of this section is to look at different areas of research where this study contributes to and discuss, in detail, the limitations of this study.

6.1 Contributions

First, as I have mentioned in the introduction of this study, Battle Royale is a new genre that has taken the world by storm. Despite its popularity, there has not been a significant amount of academic research on games that comprise this genre. The unique aspect of the Battle Royale games, such as PUBG, is that these games rely solely on the skills and strategies of the players. Unlike other famous game titles, in PUBG, all players start at the same point; with no weapons, health packs, items etc. To win, a player must not only have a good level of skill, but also a good understanding of the game and the underlying strategies. Moreover, PUBG is unique as there are no pre-defined classes. Choosing a player, attire, or any other attributes (aesthetic or not) has no impact on the player. Thus, there are no pre-defined classes, and players differ from each other solely on the basis of how they play, and the decisions they make in the game. Thus, this thesis primarily contributes by laying the groundwork for research on player behaviors and playing styles in PUBG and in Battle Royale.

Second, because of the unique nature of the game, it is imperative to find an approach that can provide us with suitable results, to identify different clusters of player behavior and playing styles. This study, on this aspect, contributes on two different fronts. First, this study proposes a meta-clustering approach where I look at combinations of different features that can offer us insight into the decision-making and playing style of the players. Other research studies have used clustering techniques that look at all the feature together (often, pre-processed with techniques like Principle Component Analysis), my study contributes to the literature by presenting a slightly different approach. The choice of meta-clustering is based on the fact that, as all players start equally, one would expect that (when looking at all features) the differences would be too marginal to find clusters. As part of the initial analysis, I did try this technique and found that this approach was invalid and did not really work (and thus, I left it out of the research). Second, in this study, I propose a new algorithm ‘K-Means-o’ that is born out of the need to look at all players. The common technique is to pre-process the data and remove outliers before clustering. This is

because algorithms like K-Means are sensitive to outliers. However, in my case, I believe that this doesn't make sense as studying players that exhibit outlying behavior is essential. Moreover, we can look at these anomalous players in relation to their nearest clusters, and see how they differ. For such an analysis, the K-Means-o algorithm is quite unique and provides accurate results.

Upon conclusion, this study contributes significantly to the current landscape of literature on game analytics by providing research on a new genre in gaming, proposing a meta-clustering approach to identify player behaviors and playing styles and finally, by implementing a new algorithm to deal with particular problems related with this study.

6.2 Limitations

For the analysis, this study primarily focuses on meta-clustering player behaviors and playing styles using the K-Means-o algorithm. In the Experimental Setup and Contribution sections, I have explained the K-Means-o algorithm in detail and how it helps us counter the problems posed by the research in question. Considering this, there are certain limitations of this study.

First, due to the lack of data available from Bluehole, it is hard to find match data which we could use to predict the win outcomes based on the player behavior clusters. This would help us see if certain playing styles can help us predict if a player or team will win the round. It would also have been nice to have data that could give us demographic information about the players, where we could explore if external factors have an impact on the player behavior and playing styles.

Second, for the purpose of the analysis, I chose to only look at meta-clustering by employing the K-Means-o algorithm. Comparisons with other algorithm would have made the significance of the results more clear. My choice to only use K-Means-o algorithm were motivated by a few reasons; unique nature of data given the Battle Royale genre and not removing outlier values from the dataset. Due to these limitations of the study, I, myself, was limited in what algorithms I could employ. There are other algorithms that do not exclude outliers (such as DBSCAN) but that there fundamentally too different from the base K-Means algorithm. Such a difference would have produced drastically different results that, in the end, would be rendered incomparable. Furthermore, as highlighted in the Discussion section, hierarchical clustering would also be another option to explore. Given more time and resources, this is something that I would like to pursue in the future. All in all, this is a promising start to the research question at hand.

7 Conclusion

The purpose of this thesis is to explore different player behavior and playing styles in the PUBG Battle Royale game. For the purpose of this analysis, I presented an adapted version of the K-Means algorithm called, K-Means-o. K-Means-o algorithm aggregates data into various clusters based on similarity and iteratively, removes data points that are considered outliers. This allows me to look at the overall population of the players relative to the outliers. Such an approach was necessary for this analysis, as we cannot simply remove players that exhibit extreme behavior as it might give us certain insights into the game.

For the analysis, this thesis attempts to answer the following research questions:

- What are the different strategies players adopt based on different skills and metrics?
- How do the players in the general population differ (in terms of strategies and play style) from those that are outliers?

As we can see in the Results section, using the K-Means-o algorithm, we can cluster different player behaviors and detect different classes of players. Moreover, we can see that even though these clusters are based on specific combination of feature values, the impact of behavioral clusters can be seen in other features. For example, for Support, the Assault Team not only does well in doing damage to other players and are at par with the Squad Leaders, their shooting accuracy and headshot kill ratio is higher. As far as the players that exhibit extreme behaviors, we can see from the analysis that most of these players do unusually well as compared to their clusters (though there are outliers that do worse). The strategies adopted by such players are the same as those in the clusters, but it is reasonable to consider that these players are just better in terms of their skill and judgement (or worse, if we are looking at the outliers that don't do well).

On conclusion, this thesis presents a different analytical approach to tackle player behavior clustering in a new game genre. Not only is this genre popular, but on a technical and gaming standpoint, it is different and harder to analyze. Thus, the analysis presented is a good first step given the limitations described in the previous section. Since, there has been no significant research on player behavior in Battle Royale games, this analysis might lay the groundwork for future studies. Future work could include look at differences in results calculated by employing different algorithms as well as incorporating match and demographic data to make predictions about game outcomes.

References

1. Barai, A., Dey, L., 2017. Outlier Detection and Removal Algorithm in K-Means and Hierarchical Clustering. *World Journal of Computer Application and Technology*, [Online]. 5/2, 24-29. Available at: [Accessed 4 June 2018].

2. Bauckhage, C., Drachen, A. & Sifa, R., 2014. Clustering Game Behavior Data. *IEEE Transactions on Computational Intelligence and AI in Games*. Available at: https://andersdrachen.files.wordpress.com/2014/07/clustering_game_behavior_data_tciaig2014.pdf [Accessed 2 June 2018].

3. BBC Sport. 2018. *Paris 2024 Olympics: Esports 'in talks' to be included as demonstration sport*. Available at: https://www.bbc.com/sport/olympics/43893891. [Accessed 3 June 2018].

4. Drachen, A., Canossa, A. & Yannakakis, G.N., 2009. Player modeling using self-organization in Tomb Raider: Underworld. *22nd IEEE International Symposium on Computer-Based Medical Systems*. Available at: https://ieeexplore.ieee.org/document/5286500/references [Accessed 7 June 2018].

5. Drachen, A., Sifa, R., Bauckhage, C. & Thurau, C., 2014. Guns, Swords and Data: Clustering of Player Behavior in Computer Games in the Wild. *Conference: IEEE Computational Intelligence in Games*. Available at: https://pdfs.semanticscholar.org/3944/89e7f97fa5dd73c09704cdfb1233c0ba2946.pdf [Accessed 29 May 2018].

6. Fast Company. 2013. *Why video games succeed where the movie and music industries fail*. Available at: https://www.fastcompany.com/3021008/why-video-games-succeed-where-the-movie-and-music-industries-fail. [Accessed 30 May 2018].

7. Fernandez, M. 2018. *Professional competitive gaming on the rise, Overwatch shows olympic potential*. Variety. Available at: https://variety.com/2018/digital/news/esports-video-games-olympics-1202709110/. [Accessed 3 June 2018].

8. Gagné, A.R., El-Nasr, M.S. & Shaw, C.D., 2011. A deeper look at the use of telemetry for analysis of player behavior in RTS games. *ICEC '11 Proceedings of the 10th international conference on Entertainment Computing*. Available at: https://dl.acm.org/citation.cfm?id=2176617 [Accessed 9 June 2018].

9. Henderson, T. & Bhatti, S., 2001. Modelling user behaviour in networked games. *Proceedings of the ninth ACM international conference on Multimedia*. 1, 212-220. Available at: https://dl.acm.org/citation.cfm?doid=500141.500175 [Accessed 4 June 2018].

10. Kirman, B., & Lawson, S., 2009. Hardcore classification: Identifying play styles in social games using network analysis. *In Proceedings of the 8th International Conference on Entertainment Computing, ICEC*. 9, 246–251. Berlin, Heidelberg: Springer-Verlag.

11. Li, R., 2017. *Good Luck Have Fun: The Rise of eSports*. 1st ed. New York City: Skyhorse Publishing.

12. Mahlmann, T., Drachen, A., Togelius, J., Canossa, A., Yannakakis, G.N., 2010. Predicting player behavior in Tomb Raider: Underworld. *Computational Intelligence and Games (CIG), 23rd IEEE International Symposium on Computer-Based Medical Systems*. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.180.6405&rep=rep1&type=pdf [Acces sed 6 June 2018].

13. Minotti, M.. 2016. *Video games will become a $99.6B industry this year as mobile overtakes consoles and PCs*, Venture Beat. Available at: https://venturebeat.com/2016/04/21/video-games-will-become-a-99-6b-industry-this-year-as-mobile-overtakes-consoles-and-pcs/. [Accessed 1 June 2018].

14. Nasdaq. 2018. *Investing in video games*. Available at: https://www.nasdaq.com/g00/article/investing-in-video-games-this-industry-pulls-in-more-revenue-than-movies-music-cm634585?i10c.encReferrer=&i10c.ua=1&i10c.dv=14. [Accessed 2 June 2018].

15. Ramirez-Cano, D., Colton, S. & Baumgarten, R., 2010. Player Classification Using a Meta-Clustering Approach. *Imperial College London*. Available at: http://ccg.doc.gold.ac.uk/wp-content/uploads/2016/10/ramirez_cgat10.pdf [Accessed 30 May 2018].

16. Statista. 2016. *eSports audience size worldwide from 2012 to 2021, by type of viewers (in millions)*. Available at: https://www.statista.com/statistics/490480/global-esports-audience-size-viewer-type/. [Accessed 4 June 2018].

17. Warr, P. 2014. *eSports in numbers: Five mind-blowing stats*, Red Bull. Available at: http://www.redbull.com/en/esports/stories/1331644628389/esports-in-numbers-five-mindblowing-stats [Accessed 11 June 2018 2015].

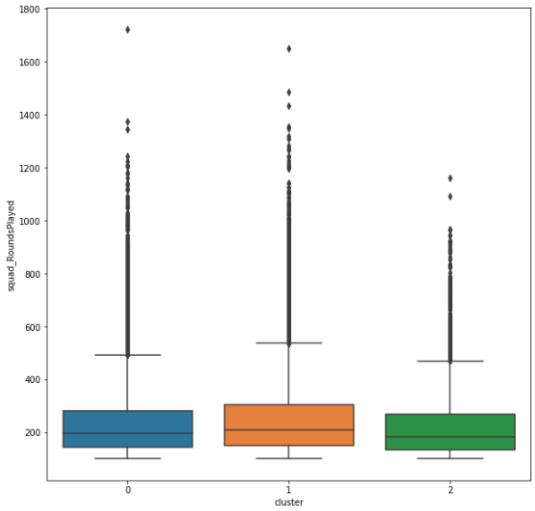
18. Yi Ong, H., Deolalikar, S. & Peng, M.V., 2015. Player Behavior and Optimal Team Composition in Online Multiplayer Games. *Stanford University*. Available at: http://cs229.stanford.edu/proj2014/Hao%20Yi%20Ong,%20Sunil%20Deolalikar,%20Mark%20Peng,%20Player%20Behavior%20and%20Optimal%20Team%20Compositions%20for%20Online%20Multiplayer%20Games.pdf [Accessed 2 June 2018].

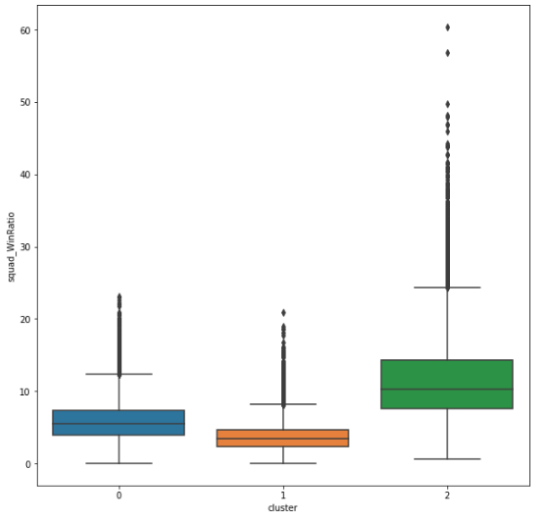
Appendix I: PUBG Features

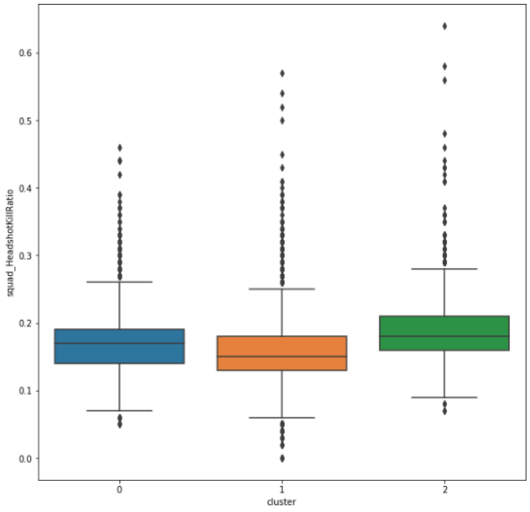
Here is the complete list of all the features available in the dataset:

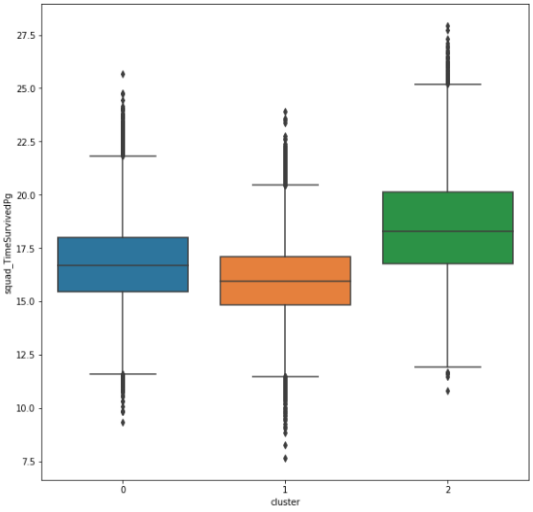
KillDeathRatio	WinRatio
TimeSurvived	RoundsPlayed
Wins	WinTop10Ratio
Top10s	Top10Ratio
Losses	Rating
BestRating	DamagePg
HeadshotKillsPg	HealsPg
KillsPg	MoveDistancePg
RevivesPg	RoadKillsPg
TeamKillsPg	TimeSurvivedPg
Top10sPg	Kills
Assists	Suicides
TeamKills	HeadshotKills
HeadshotKillRatio	VehicleDestroys
RoadKills	DailyKills
WeeklyKills	RoundMostKills
MaxKillStreaks	WeaponAcquired
Days	LongestTimeSurvived
MostSurvivalTime	AvgSurvivalTime
WinPoints	WalkDistance
RideDistance	MoveDistance
AvgWalkDistance	AvgRideDistance
LongestKill	Heals
Revives	Boosts
DamageDealt	DBNOs
player_name	tracker_id

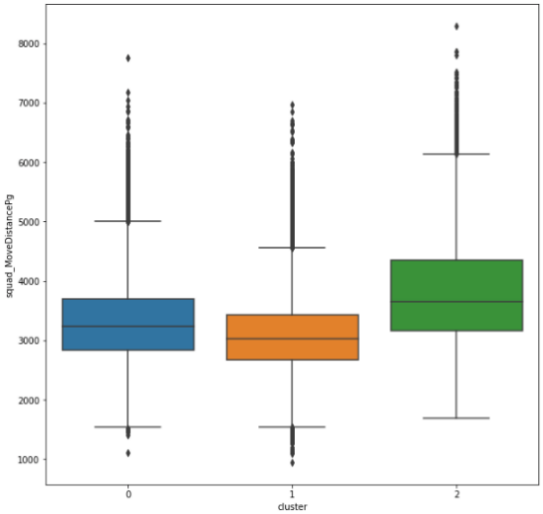
Appendix II: Combat



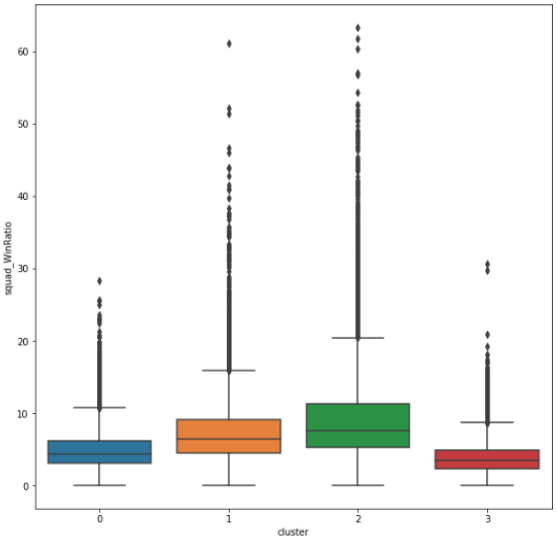


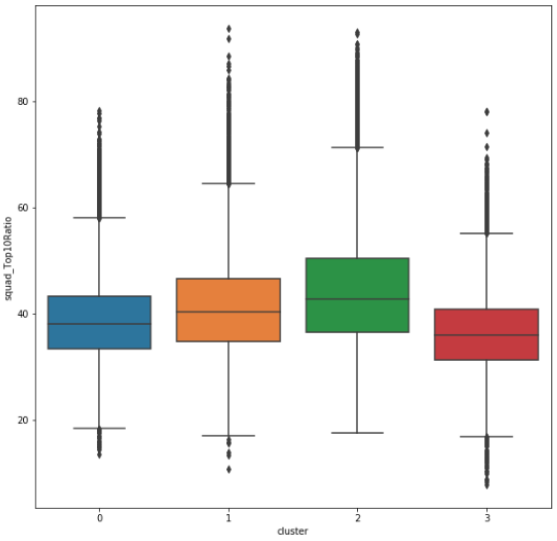


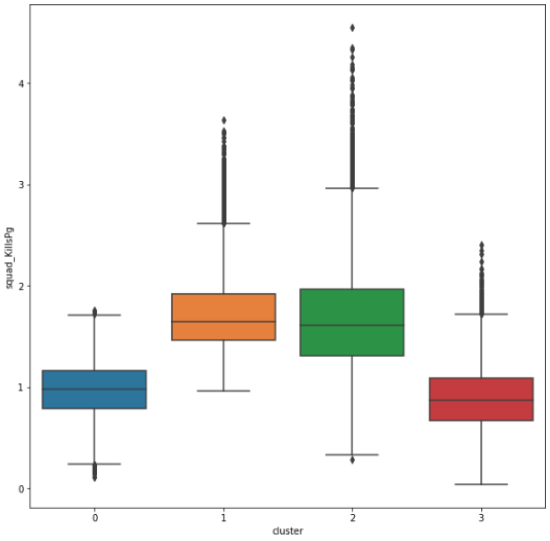


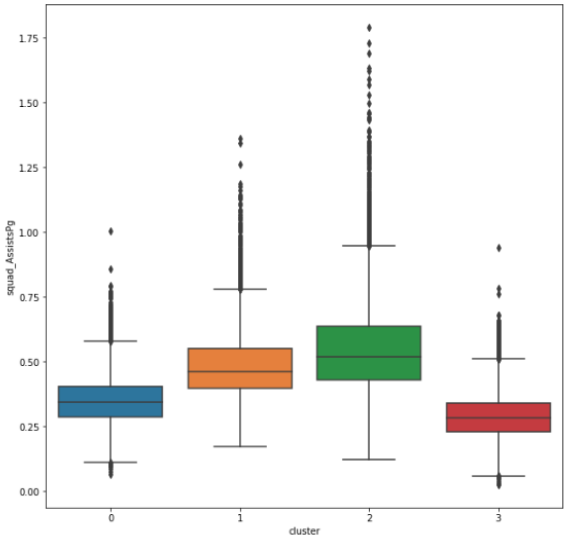


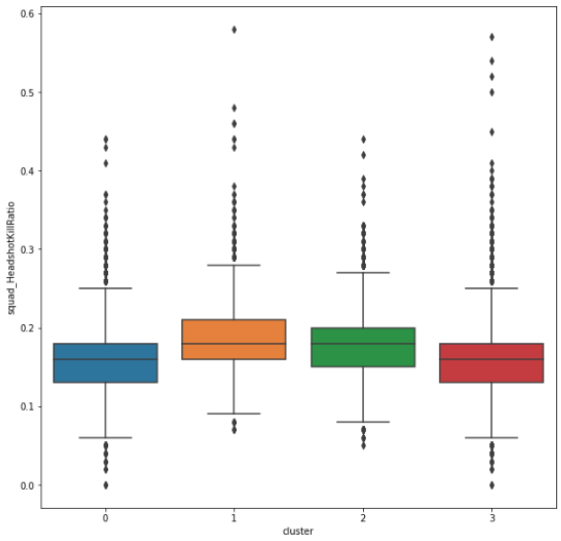
Appendix III: Support

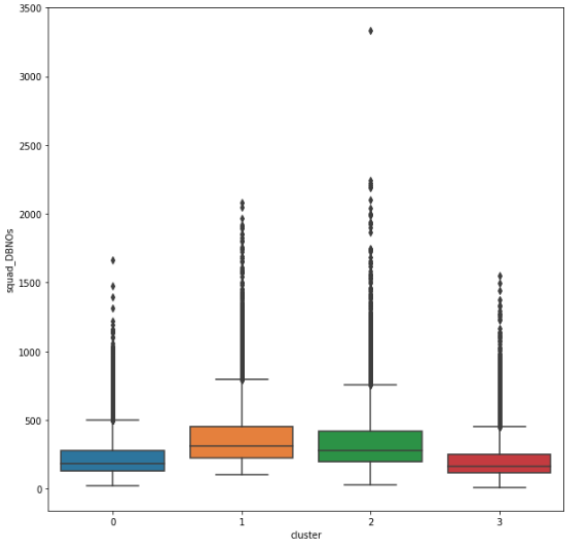


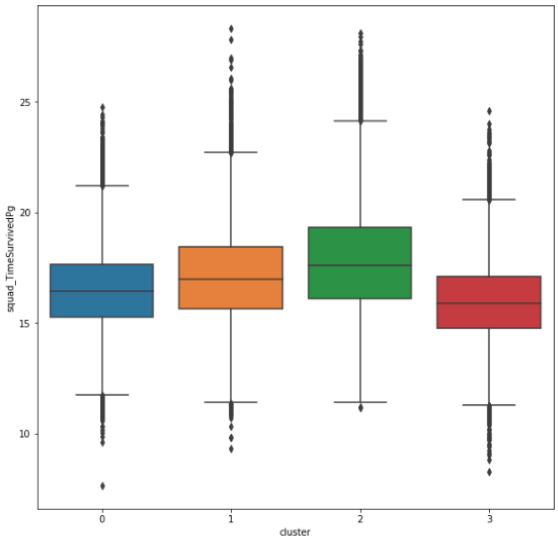


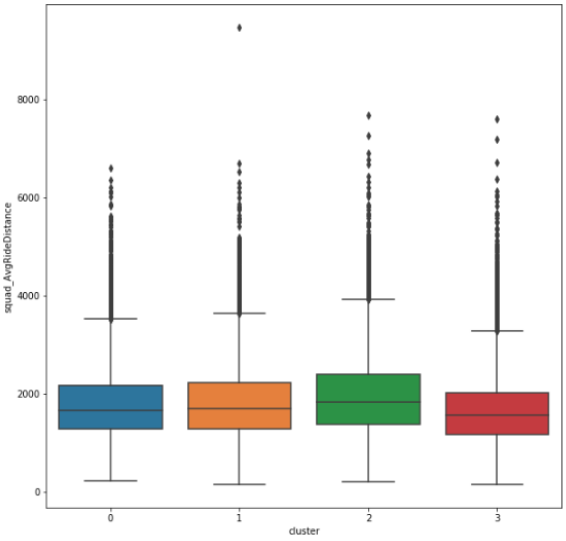


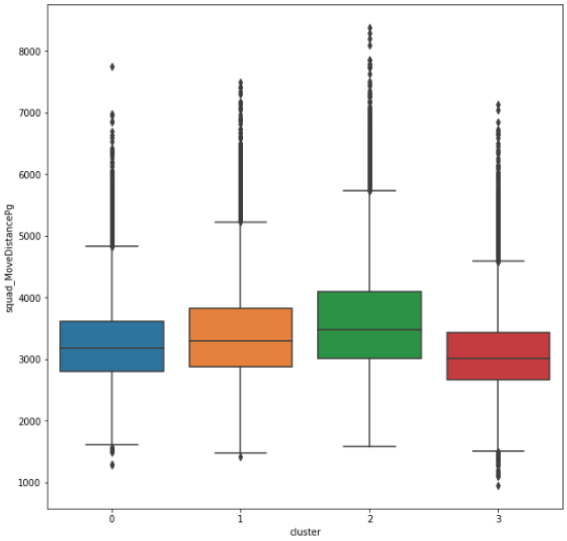





squad_DBNOs
cluster


squad_TimeSurvivedPg
cluster


squad_AvgRideDistance
cluster


squad_MoveDistancePg
cluster

Appendix IV: Movement