a place of mind
THE UNIVERSITY OF BRITISH COLUMBIA

PlayerUnknown's Battlegrounds (PUBG) Exploratory and Descriptive Data Analysis Integrated With Data Visualizations

Name: Wells Wang

a place of mind
THE UNIVERSITY OF BRITISH COLUMBIA

Student Number: 48750665

2018-12-01

1. Introduction

As one of the most popular online shooting games, PlayerUnknown's Battlegrounds (PUBG) stands out from the other games for a reason. The rules and contents of the game has shaped the style of playing into a new dimension. Players can't focus only on boosting their shooting techniques anymore in order to win this game. They have to work on their strategies to avoid getting killed at the same time.

"In the game, up to one hundred players parachute onto an island and scavenge for weapons and equipment to kill others while avoiding getting killed themselves. The available safe area of the game's map decreases in size over time, directing surviving players into tighter areas to force encounters. The last player or team standing wins the round (PUBG Introduction). "

PUBG's massive number of players created a large amount of data as they play their game. One of the data set can be found on the Kaggle website, which is uploaded two months ago for a prediction competition. Kaggle is a website for data scientist to communicate, collaborate, and compete with each other.

1.1 Motivation and background

As more people are communicating on the social media or blogs about specific techniques in PUBG such as taking shots or a better place on the map to hide and avoid getting killed, not many of them talked about the factors that is essential in winning this game (i.e. Focus on killing or hiding to avoid getting killed).

With the data set on Kaggle, the above problem might be solved if we do analysis into it. Furthermore, Kaggle users help each other out by posting their analysis results to kernel. As a result, if we do this research, we might help out two communities at once.

1.2 Research Questions and Objectives

For the purpose of helping PUBG users, the objective of this research is to find out the correlation between "*player controllable factors"* (some variables in the data set that indicates what actions did the players made during the match) and the "*winPlacePerc*" (the final placement of the player's performance, will be further discussed in the data section), find out what players have to do to stand a higher chance in winning the game, and visualize the analysis results to make them understand.

For the purpose of helping Kaggle users, the objective of this research is to upload the analysis results as well as some explanations on the Kaggle website to give some insights about the data for their further predictions and model constructions.

2. Data Summary

The definition of the 29 PUBG variables are provided on the Kaggle website, they can be categorized as groups below:

Groups	Variables	Level of Measurement
Ids	Id(players)/groupId/matchId	nominal
Match Stats	matchDuration/matchType/numGroups	ratio/nominal/ordinal
Kill Counts	DBNOs(knocks)/kills/assists/killStreaks/ headshotKills/roadKills/teamKills	ratio
In Game Stats	boosts/heals/revives/vehicleDestroys/ weaponsAcquired	ratio
Distances	longestKill/swimDistance/rideDistance/ walkDistance	interval
Points	damageDealt/killPoints/winPoints/rankPoints	ordinal
Performance	killPlace/maxPlace/winPlacePerc	ordinal/ordinal/ratio

The Ids are a series of numbers that represents the players, the match, and the group they are in for the match. Apart from *matchId*, the duration of the players in a match, the total groups in the match, and the type (solo, duo, squad) of the match are also important as Match Stats. During the game, Kill Counts, In Game Stats, and Distance represent the actions that the players did in the game. These are the variables we would want to look into for they imply the strategies of which

the player has acquired. Finally, Points and Performance indicates how the player do at the end of the game.

Supplementary Definitions:

- DBNOs - Number of enemy players knocked. If you knocked someone, they can't make actions until there teammates revive them.
- assists - Number of enemy players this player damaged that were killed by teammates.
- boosts - Number of boost items used. (boost items are things like adrenaline that keep players alive in the game)
- killStreaks - Max number of enemy players killed in a short amount of time.
- damageDealt - Total damage the players dealt to their enemies, in a cumulative point scale.
- 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". The "-fpp" stands for "first person perspective", which is another type of playing the game.
- longestKill - Longest distance between player and player killed at time of death.
- roadKills - Number of kills while in a vehicle.
- teamKills - Number of times this player killed a teammate.
- swimDistance - Total distance traveled by swimming measured in meters.
- rideDistance - Total distance traveled in vehicles measured in meters.
- 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.

3. Methodology and Applications

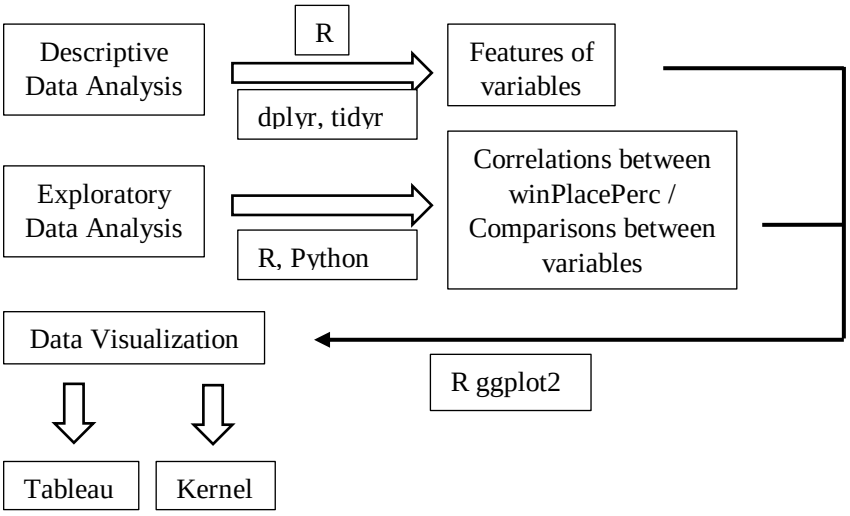

Fig. 1. Methodology Flow Chart

3.1 Methodology

As shown in Fig. 1, The methodology of this research is as follow:

First, the descriptive data analysis was executed for looking into the features of the variables. In this part, the R *dplyr* and *tidyr* package was used to examine the data set. Then the exploratory data analysis was executed with the help of R and python to find out variables' correlation with the *winPlacePerc*, and to do comparisons to see how each variable are related to the *winPlacePerc* differently. Last, the above analysis was integrated with data visualizations to generate understandable graphs with the help of r *ggplot2* package, which was uploaded to Kaggle kernel for giving the Kaggle users insights. Finally, a Tableau dash board contains only the winner's data was constructed to help the PUBG players get a hint of how they perform comparing to the winners.

3.2 Applications

The applications are split into three parts, the first part is individual features visualizations based on descriptive data analysis, where we look into the insights of each variables. The second part is multiple features visualizations based on exploratory data analysis, where we investigate how each variable is related to *winPlacePerc* differently. The third part is Tableau visualization dash board based on the correlation analysis with python.

3.2.1 Descriptive Data Analysis

The descriptive data analysis will focus on *Match Stats*, *Kill Counts*, *In Game Stats*, and *Distances* group mentioned in the Data Summary section. These groups consist of variables that is controlled by the players as they imply the actions they made in each match. First, we look into the *Match Stats* group.

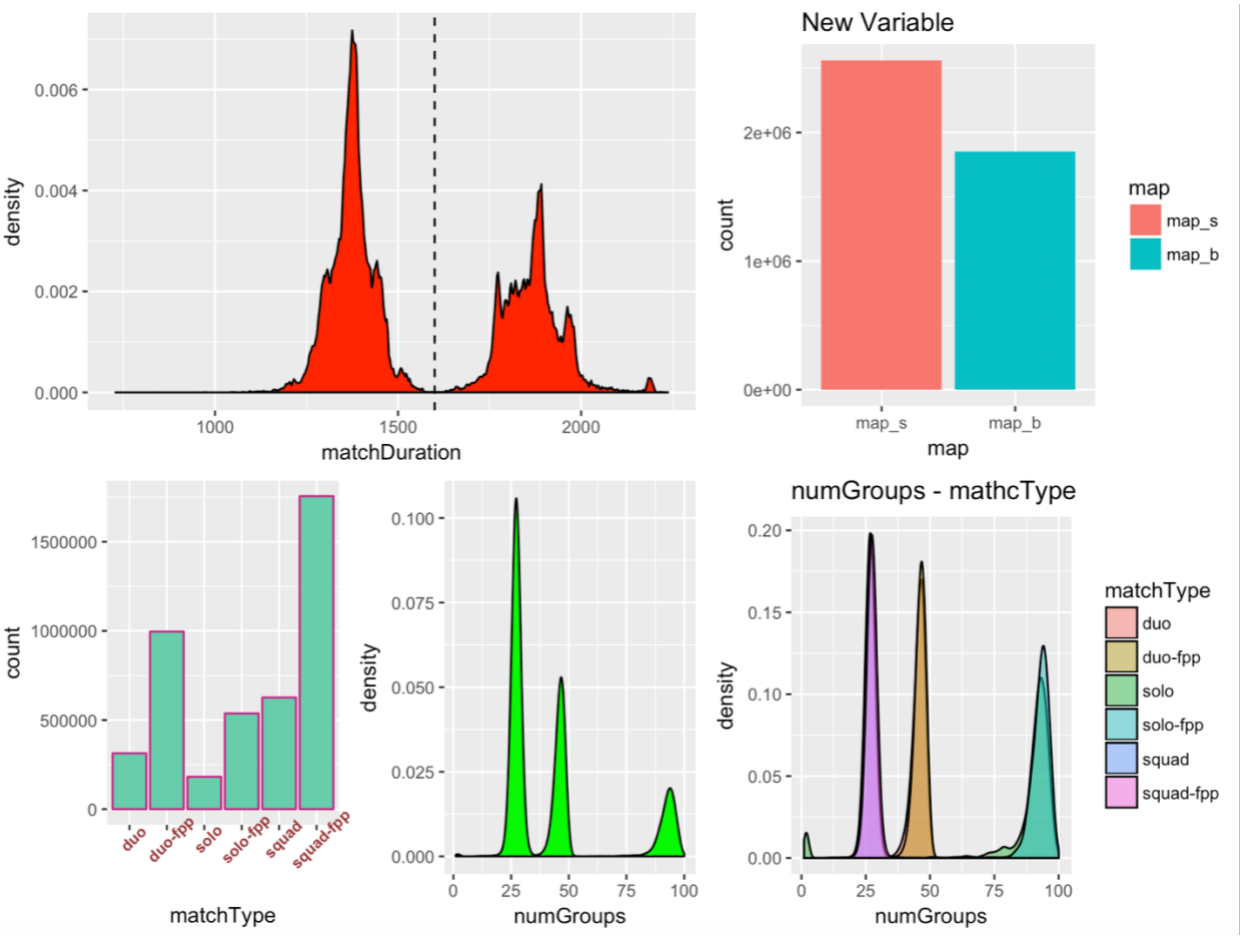

Fig. 2. Match Stats Group

As shown in Fig. 2, For the *matchDuration*, we find that there are two chunks among all players which can be separated by x intercept 1600. It might be influenced by the map the players have chosen. Unfortunately, we don't have such information in our data. However, this could be an interesting feature if we generate a new variable called map base on *matchDuration*. If the assumptions are correct, we can see the distribution of the *New Variables* graph to see how people are choosing the map. The *squad-fpp* is by far the most popular one, follow by *duo-fpp*, leaving 16% of the players choose *solo* and *solo-fpp*. This implies that the majority of the players likes to team up to play this game. The *numGroups* peak at around 25, 50 and 100, there is an obvious trend for this situation as it is the default group number for *solo*, *duo*, *squad*, and the first-person perspective mode of each of them. We can see the trend from *numGroups* – *matchType* plot.

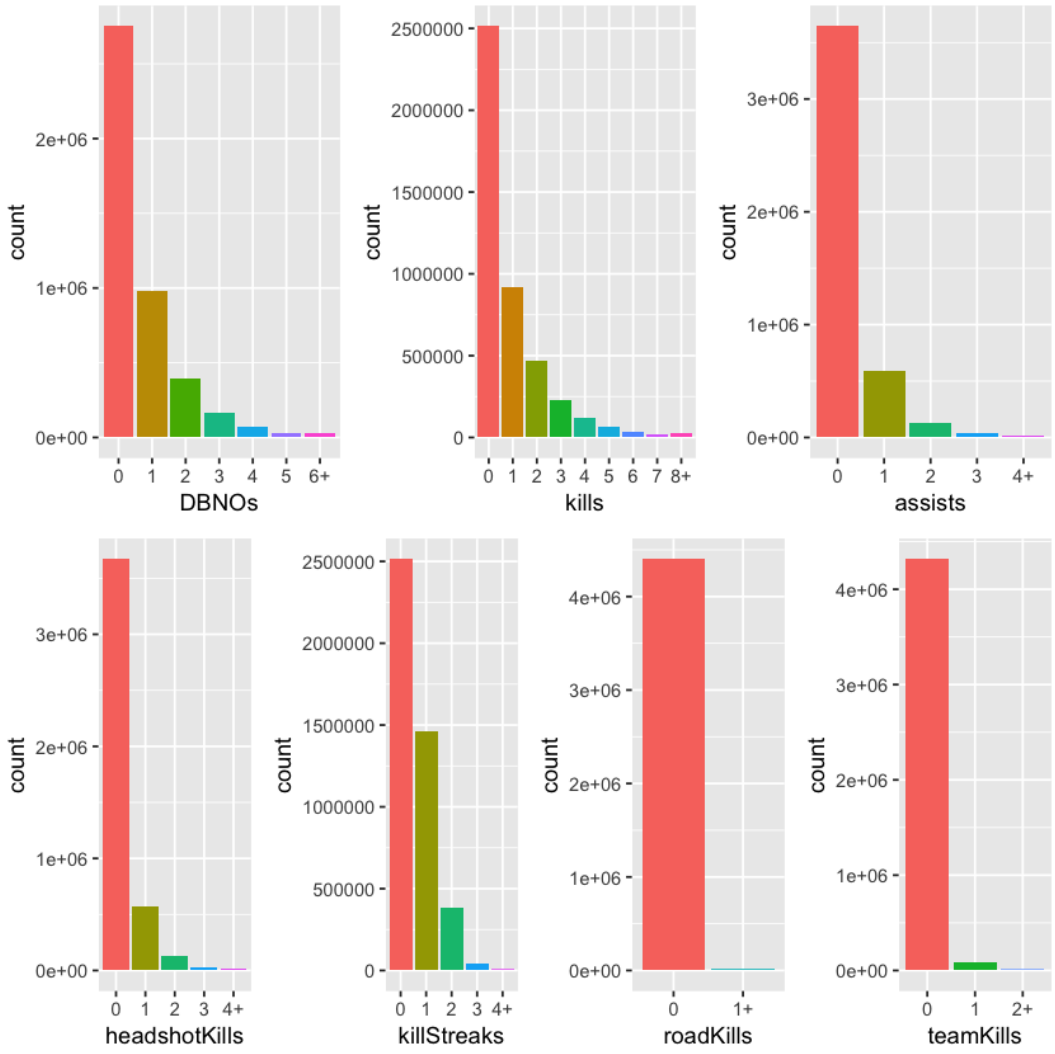

Fig. 3. Kill Counts Group

Next, Fig.3 shows the features of variables inside the *Kill Counts* group. Among the *Kill Counts* group, a clear majority of players didn't even succeed in knock out or kill an enemy. However, for *DBNOs*, *kills*, the count from 1 to 6 still contribute a notable fraction compared to those bigger numbers. A high count of *killStreaks* at 1 indicates that most people who did make kills can't kill more than 1 enemy within a short period of time. Still, there are a notable amount of people who did kill 2 enemies in a roll. As for *roadKills* and *teamKills*, such incident rarely happens in a match.

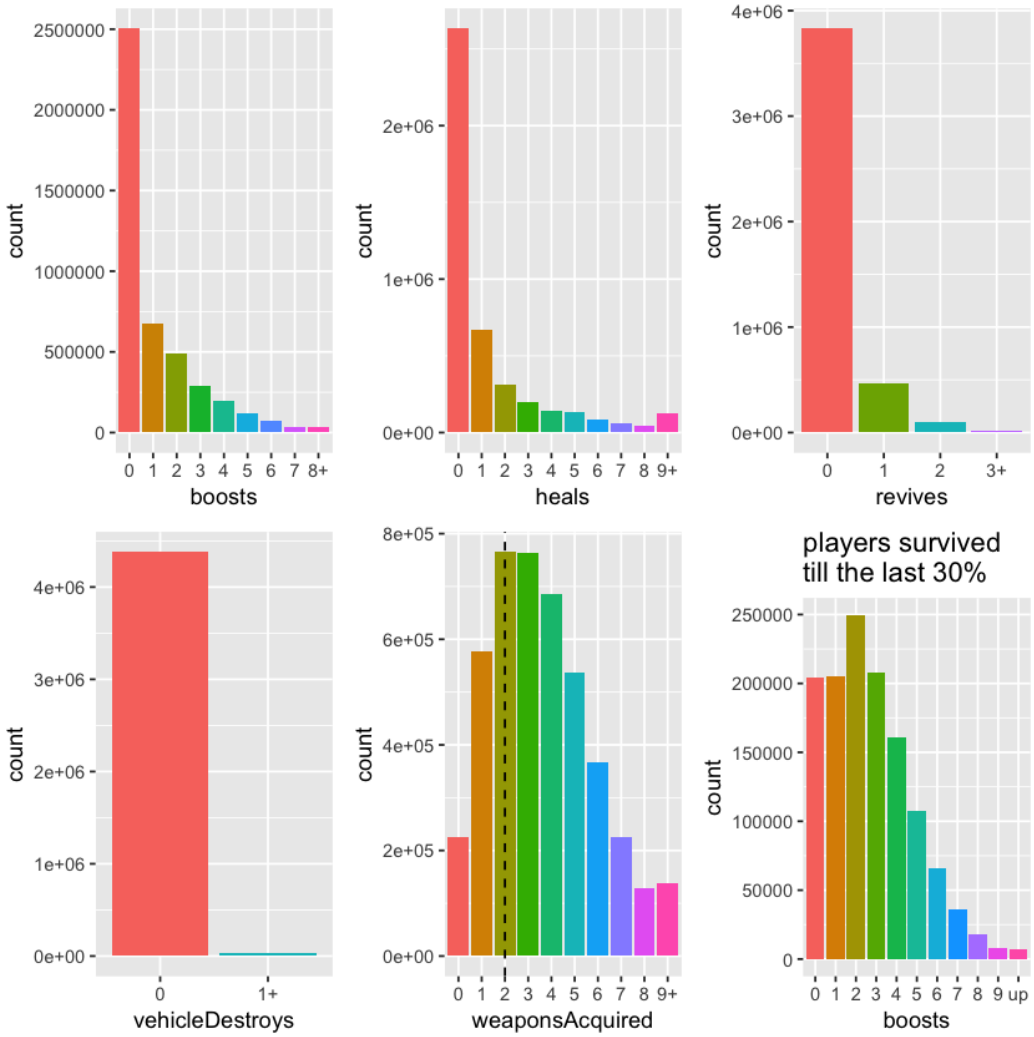

Fig. 4. In Game Stats Group

Fig.4. shows the features of the *In Game Stats* group. Around 60% of players didn't boosts and heals. We have reasons to believe that these players have died in an early stage of the game. As a reference, we plotted a graph of *players survived till the last 30%*. We can see that a lot of players with 0 boosts are eliminated. *WeaponsAcquired* seems to be roughly normal distributed with a median of 2 (vertical dashed line). As for *revives* and *vehicleDestroys*, such incident rarely happens in a match. Next, Fig.5 shows the features in the *Distances* group.

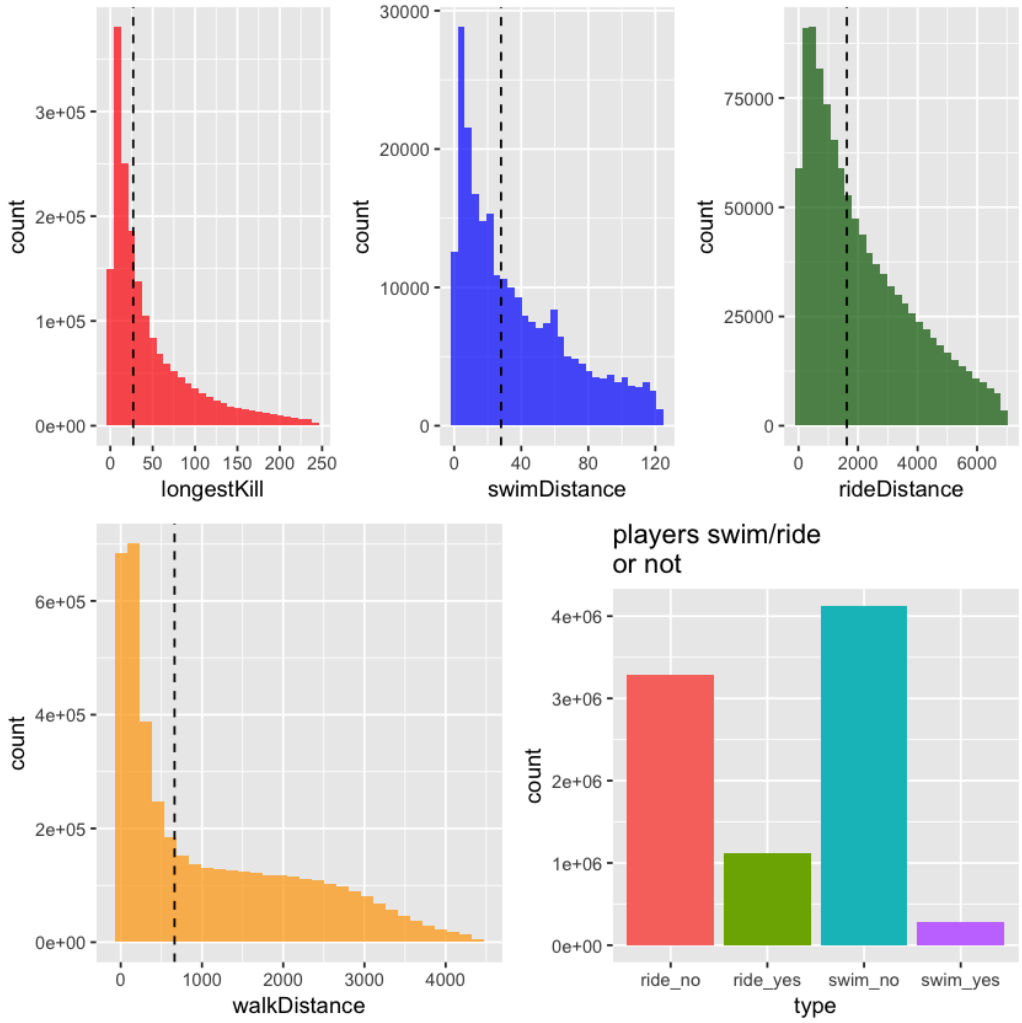

Fig. 5. Distances Group

Since over half of the players didn't even make a kill, resulting the *longestKill* for them to be 0, we filtered them out to see how the other players perform. We can see that these players have a median of 27 meters longest kill distance, and a significant amount of them only have kills shorter than 27 meters. We can imply that a majority of the combat happens in a smaller amount of range. Again, we filtered the 0s out for rideDistance and swimDistance because only around 1/4 of the players ride, most of them don't swim. The graph *players swim/ride or not* is reference to that. Interestingly, *swimDistance* for those who did swim have a median at around 28 meters, and also a significant amount of them have only swum shorter than 27 meters. As for *walkDistance* over half of the players can't walk over 700 meters, with a median of this variable at 662.2.

3.2.2 Exploratory Data Analysis

The exploratory data analysis will focus on how the features in *Kill Counts* group and the *In Game Stats* group are related to the *winPlacePerc*. First, three graphs of the group *In Game Stats* are shown to demonstrate how the data differ when the *winPlacePerc* is different (Comparisons 1). The graph of *winPlacePerc* at 0 and 1 are plotted and compared with the general players, which consist of the whole data set.

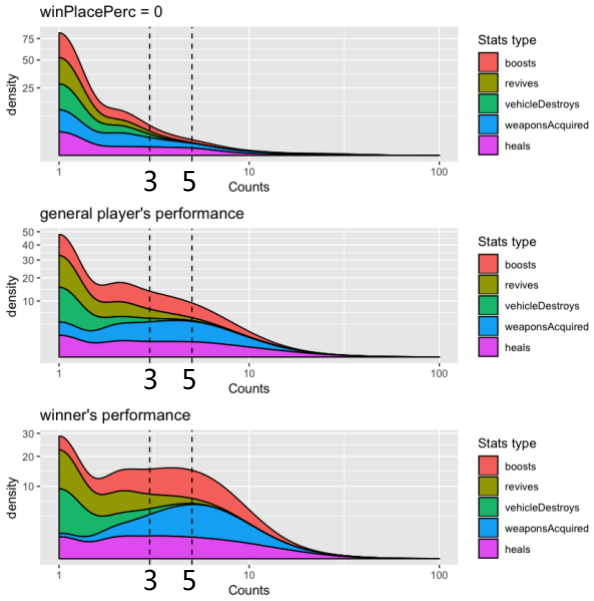

Fig. 6. Comparisons 1 – In Game Stats

winPlacePerc = 0

heals > 3	percentage
FALSE	99.91680112
TRUE	0.08319888

boosts > 3	percentage
FALSE	99.98758912
TRUE	0.01241088

General Players

heals > 3	percentage
FALSE	86.50678
TRUE	13.49322

boosts > 3	percentage
FALSE	89.72074
TRUE	10.27926

Winners

heals > 3	percentage
FALSE	64.21498
TRUE	35.78502

boosts > 3	percentage
FALSE	45.96025
TRUE	54.03975

Fig. 7. Percentage Difference

As shown in Fig. 6, most of the players with 0 in *winPlacePerc* didn't get the count of these actions over 3 times. Fig.7 shows how the percentage of heals and boosts shifts for the general players and the winners, while differences between revives and vehicle destroys are not very obvious. In the next comparisons (comparisons 2), the mean and the median of *winPlacePerc* at different counts of each variables are calculated and plotted. The *Kill Counts* group and the *In Game Stats* group are examined in this part.

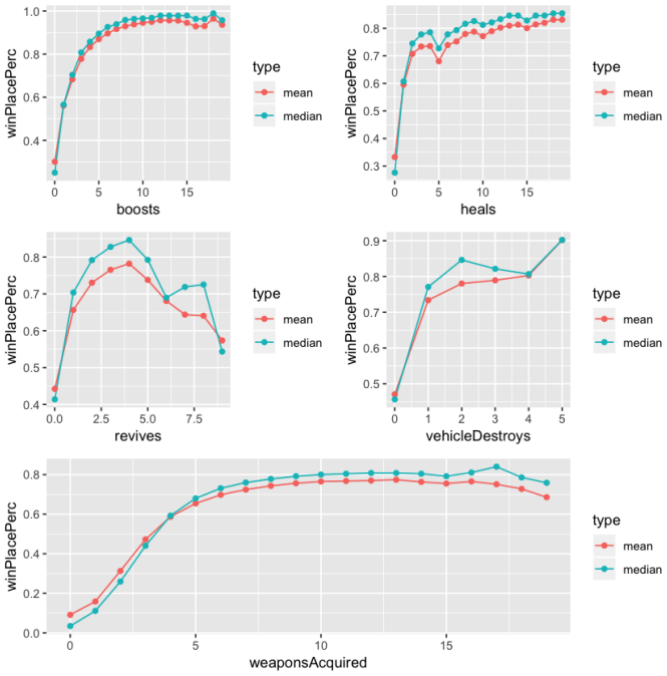

Fig. 8. Comparisons 2 – In Game Stats

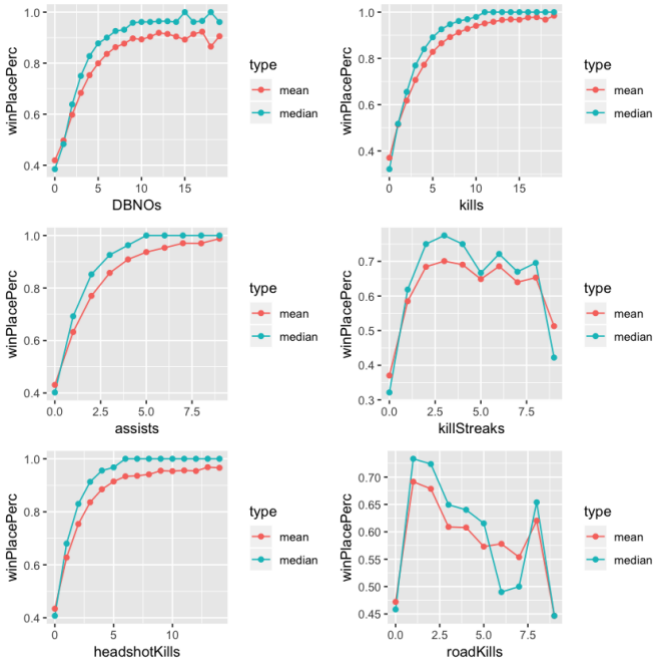

Fig. 9. Comparison 2 – Kill Counts

Fig. 8 shows the comparisons 2 of *In Game Stats* group. The *winPlacePerc* grows rapidly as each boost items were used. The influence gets not very obvious as *boosts* goes over 5. The situation for *heals* and weapons Acquired is quite similar with *boosts*. As for *revives*, the *winPlacePerc* grows as people started to revive their teammates. However, once it gets over 3 times of reviving teammates, the *winPlacePerc* would actually start dropping. For *vehicleDestroys*, those who destroyed above 1 vehicle have a way higher *winPlacePerc* than those who didn't. However, we must keep in mind that only 1% of the players did destroy vehicles.

Fig. 9 shows the comparisons 2 of *Kill Counts* group. The trend for *DBNOs*, *kills*, *headshotKills*, and a*ssists* is quite similar to the *boosts* in *In Game Stats* group. As for *killStreaks*, killing 2 in a short period of time did indicates growth in *winPlacePerc*. However, the *winPlacePerc* actually starts to drop when *killStreaks* gets over 2, the trend is even more obvious when it gets to 7. This might be indicating the players who kills too quick has a bigger chance in getting killed.

3.2.3 Tableau Visualization Dash Board

Through the python analysis, the correlation between *winPlacePerc* and each variable are calculated. The six variables that have higher correlations with *winPlacePerc* are *walkDistance, boosts*, *weaponsAcquired*, *damageDealt*, *heals*, and *kills* (shown in Fig. 10 and Fig. 11). As a result, the 6 variables are included within the Tableau dash board (link here) for data visualization process. In the Tableau dash board, only the winner's data (*winPlacePerc* = 1) are included. Players can type in their game statistics to see how they performed compared to the winners.

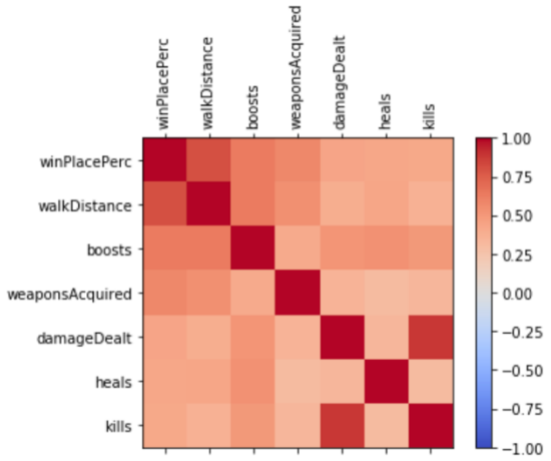

Fig. 10. Correlation Matrix

winPlacePerc	1.000000
walkDistance	0.810888
boosts	0.634234
weaponsAcquired	0.583806
damageDealt	0.440507
heals	0.427857
kills	0.419916

Fig. 11. Variables with Higher Correlations

4. Conclusions

In the conclusion section, some interesting results is pointed out for the players reference. On top of that, some advises is given base on the results.

4.1 Results

From the results of the applications above we can sum up with two features that is interesting and useful for the PUBG users. First, an important part of this game is killing or attacking enemies, which is the variables of actions inside the *Kill Counts* group. It can be concluded in few points:

- Half of the players didn't make a kill, 60% of them didn't Knocked.

- Killing more do help increase the chances of winning but killing above 5 won't help that much.
- Killing more than one people in a short period of time do increase chances of winning but killing over three in a roll will decrease the chances of winning. This might be indicating the players who kills too quick has a bigger chance in getting killed.

Second, as seen in the application section, there are many variables related to *winPlacePerc*. As a result, we can conclude that only focusing on killing won't help in winning the game. Other than killing, it can be concluded in few points:

- Using boost items helps increase the chances of winning but not much after 5 boosts, over 50% of the winners boosts more than 3 times.
- Half of the kills were made within 27 meters, so get too close to the enemy will increase the chances of dying.
- Reviving teammates do increase chances of winning but reviving over three times will decrease the chances of winning.

4.2 Discussions

From the conclusions above, some implications are provided to the PUBG users as advises. First, killing more than 5 people won't help too much in increasing *winPlacePerc*, PUBG players with too many kills might want to adjust their strategies and put more focus on the other factors in the game such as boosting or be more preserved about killing to avoid getting killed. On top of that, killing more than 2 people in a roll would put the player in higher risk of getting killed. Next, Using boost items between 3~5 times would be most effective for increasing the *winPlacePerc.* Finally, 27 meters is a range that players would want to be ware of because it increases the chances of getting killed but it also increases the chances of killing others.

Using the Tableau dash board can also give the players a hint about how they perform compared to the winners. For instance, if the players walk distance is more than most of the winners, they might want to stay in a place for longer time span instead of moving around in the map.

References

PUBG Introduction: https://en.wikipedia.org/wiki/PlayerUnknown%27s_Battlegrounds

Data Scource: https://www.kaggle.com/c/pubg-finish-placement-prediction/data

Tableau Dash Board: https://public.tableau.com/profile/wells5566#!/vizhome/BabyNameExplorer_2/PUBGDashboard

ggplot2, https://ggplot2.tidyverse.org/

dplyr, https://www.rdocumentation.org/packages/dplyr/versions/0.5.0

tidyr, https://tidyr.tidyverse.org/

ggridges, https://cran.r-project.org/web/packages/ggridges/vignettes/introduction.html

data.table, https://cran.r-project.org/web/packages/data.table/vignettes/

matplotlib.pyplot, https://matplotlib.org/api/_as_gen/matplotlib.pyplot.plot.html

pandas, https://pandas.pydata.org/

numpy, http://www.numpy.org/

sklearn, https://scikit-learn.org/stable/

seaborn, https://seaborn.pydata.org/

Calculations, https://www.tutorialspoint.com/tableau/tableau_functions.htm

Dashboard, https://onlinehelp.tableau.com/current/pro/desktop/en-us/dashboards_create.htm