Data Science and Analytics

Final Project Report

PLAYERUNKNOWN'S BATTLEGROUNDS

Team F

Jesse Perez □ Chi-Fang Li □ Nikita Nerkar □ Ian Huntley □ Junyi Wang

Introduction

The Data Science and Analytics course has given us the framework and tools needed to approach data-driven problem. In this final exercise, we will apply these tactics to find, assess and solve an issue encountered in the real world.

About PUBG

Specifically, we'll be reviewing the popular survival video game PUBG, and which factors make for a winning game for a player. This could include things such as the number of weapons picked up by the player, the distance the player travels in a game and more.

PUBG is a battle-royale survival video game. As a first person shooter, the main objective is to survive for the duration of the game. The last player in solo mode, or team in multiplayer mode left standing is declared the winner of the game. To provide context, a player begins a game sitting in an airplane flying over a remote nondescript island. This island serves as the game's *arena* of sorts. As the plane passes over the island, you can select where you'd like to jump out of the plane to parachute and land on the island. At this point, a player scrounges around finding supplies, weapons, power-ups and more to better their capabilities in playing the game.

One should note that as a player is jumping out of a plane, so are 99 other players. They may land on the opposite end of the island, or they may land right next to you. Depending on this, a player could encounter battle right at the start, or they could play most of the game without any action whatsoever. The variability of the game is what draws players into the gameplay. An interesting factor to this game is the *ring of death*, which slowly closes in on randomized areas of the island. Moving and staying within the ring encourages players to move closer and closer together as the game progresses. Should a player find themselves outside of the ring, they will start losing health, and may eventually die from damage.

In the end, out of the 100 total players, the last person alive is considered the winner. It’s evident from this introduction that there are many factors at play in the game, which makes the game so difficult. Finding the right equipment and guns for near- and long-distance combat can make the difference. Players may find themselves running all game, while others don’t move around the island as much.

Unfortunately, in this game, there is a notable number of people who are utilizing software to help them hack the game to give them a competitive edge. Overriding the system to be constantly using boosts, or even having a computer do the aiming on the head for a headshot are just a few ways these cheaters are improving their odds of winning games. Through our assessment, we wanted to explore this side of the PUBG game to help players recognize when another player might be cheating.

In this paper and associated presentation, we’ll follow the six-step process to set up the context of the problem, identify potential solutions or enhancements to gameplay, and present our findings in a way that communicate and visualize our insights and recommendations.

The Six-Step Framework for Collecting Data

Framing the Problem

Problem recognition - PUBG is a complex and competitive game, with global tournaments for cash prizes driving a lot of players to take part. Notably, in June 2018, it was estimated that 400 million players around the globe play the battle royale-style game. With so much emphasis on being the last person in the game alive, we hypothesized that there could be contributing common factors that are seen among winning players, and as such, a novice player adopting these behaviors could increase the likelihood of a player winning a game.

Because each game starts with 100 players, and the last person surviving is deemed the winner, the chance of winning is slim. Definitely those chances increase as one plays the game and becomes more proficient and experienced. Better accuracy, better reaction times and more can contribute to a winning combination for players.

As such, the problem many players face is that their chance of winning is very low, especially if they're a new player. It's understandable that every day thousands of searches online for winning tactics and tips for the game are conducted. So we set out to gather data that correlates with winning players to understand why certain actions in the game could lead to more kills and a longer lifespan in the game, which could result in the player winning the game.

Review previous findings - Our next step included a review of existing data, and if other data scientists have tried to analyze this problem. While people have explored the raw data, we have not found any deeper analysis of this data to draw conclusions. Especially because we want to identify cheaters' behaviors, we are confident that a consumer-available analysis has not been done. For the duration of this process, data sets were found via online repositories, mainly the site known as Kaggle.

Solving the Problem

Modeling and variable selection - In assessing the data provided to us, we were able to identify key pieces of information and variables that would be best leveraged to come to an answer on our hypothesis. Our identified independent and dependent variables are:

Dependent variable: Percentile Winning Placement

Independent variables: walkDistance, killPlace, boosts, weaponsAcquired, damageDealt, kills. Walk Distance consists of the distance traversed in the game on foot without the use of vehicles. Swim Distance consists of travel via swimming. Boosts are in-game perks that give players heightened abilities such as increased health and speed. The heal variable is perk that can be picked up in game that allows an injured player to regain lost health. Weapons Acquired details the amount of weapons picked up by a player as soon as they land and over the course of a game. Damage dealt is the sum of all damage one player has inflicted upon all other players regardless of weapon or whether or not the other player was killed or healed. Winning percentile placement is the distribution of players along where they placed in the winning percentile.

Collection - In the collection stage of our project, we met and discussed extensively which data points to pull together to help tell our story. Because PUBG has several gametypes, mainly playing solo or in teams, we wanted to be sure we were focusing solely on solo games and performance, so we made sure to filter to only this data. Data resource is from Kaggle competition: https://www.kaggle.com/c/pubg-finish-placement-prediction

This is a dataset with a large number of anonymized PUBG game stats, formatted so that each row contains one player's post-game stats.

Analysis - After digging through the data and visualizing it through R and Tableau, we were able to create a more accurate and visualized story to tell. We'll provide both the qualitative analysis of this information, as well as the quantitative reasoning and visualization of this data in this paper.

Communicating / Acting on Results

Communicating the Results - PUBG is an interesting game and topic on which we can tell an engaging and intriguing story. In reviewing our data, visualizations and analysis, we took care in showing how we could present the information in a way that would be engaging for the audience, but also encourage future action.

Hypothesis- It was hypothesized that certain variables would increase the likelihood of winning. High number of Headshots for example were hypothesized to be associated with winning, due to being quite effective at eliminating opponents. On that same note a high amount of kills was also hypothesized to increase chances of winning as it decreased the amount of opposition. Swimming was thought to be detrimental to the chances of winning due to being one of the slowest modes of transportation. High walk distance was thought to be negatively correlated with success due to the increased exposure while traversing the map at a normal speed. The use of vehicles was thought to be positively correlated with winning due increased speed of travel and cover from other players.

Our Hypothesis

Players who want to win should

- Quickly kill other players to reduce competition
- Use more boosts and heals
- Utilize more weapons in a game
- Aim for head-shots
- Use vehicles due to greater protection and faster travel
- Avoid the water since swimming as it gives highest exposure

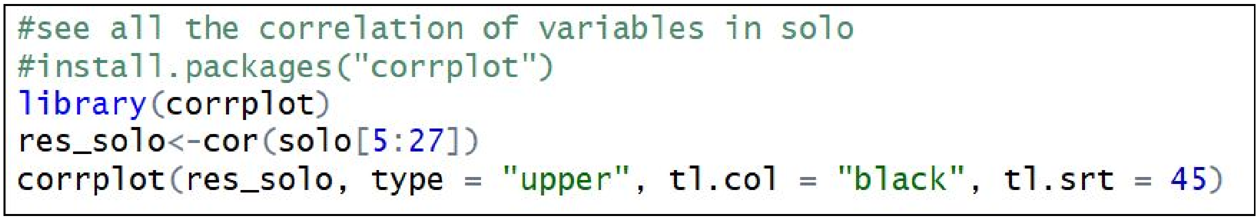
Quantitative Analysis and Creativity

Correlation

We divide the data into two categories according to the game mode.

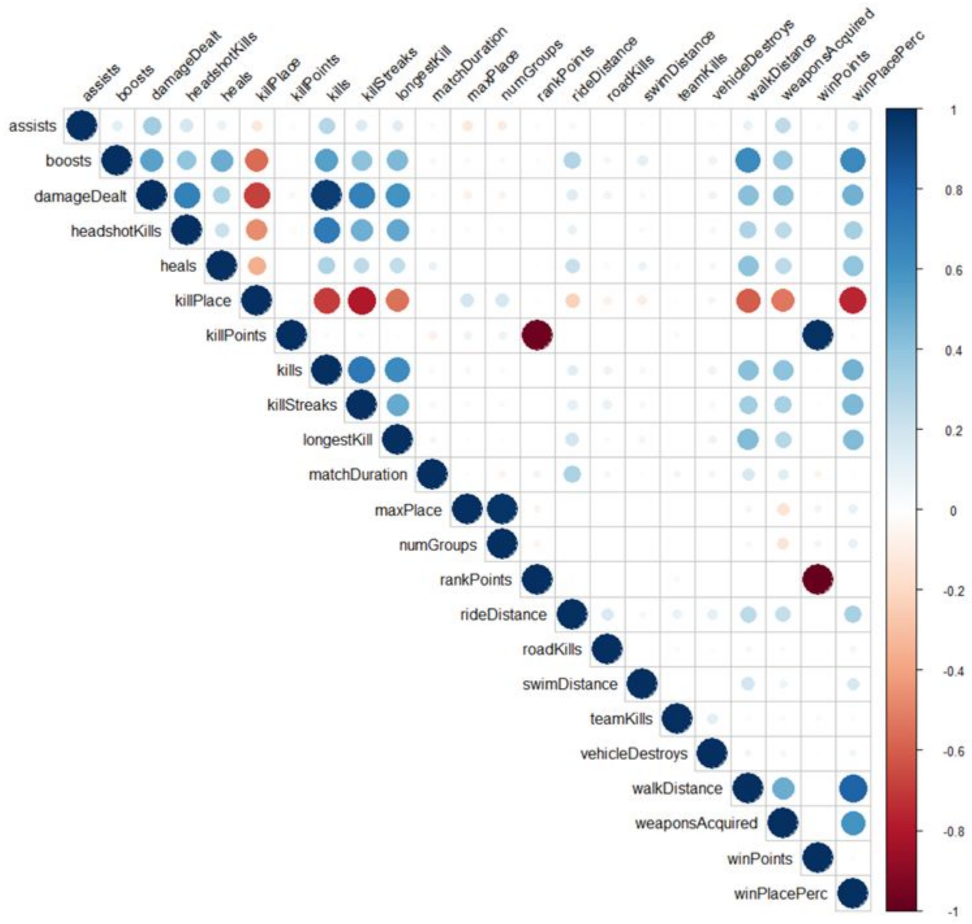
- Solo, a game mode where you spawn into the world alone and you just rely on your own tactics and skill to push you to the end and be the last player alive.
- Multi, for this game mode, players are organized into teams and will compete to be the last ones alive.

Plot the correlation in R.

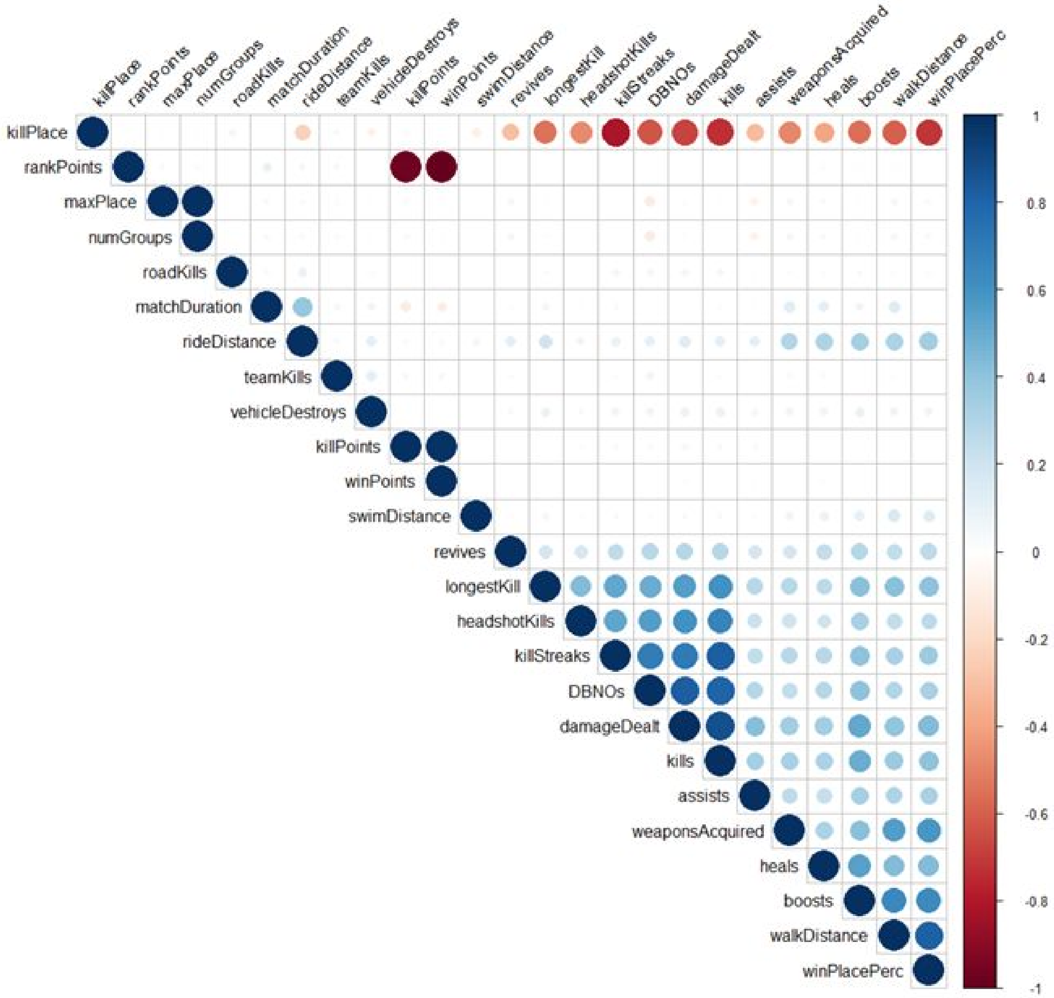


```
#see all the correlation of variables in solo
#install.packages("corrplot")
library(corrplot)
res_solo<-cor(solo[5:27])
corrplot(res_solo, type = "upper", tl.col = "black", tl.srt = 45)
```

This is all the correlation of variables under **solo mode**. The highest positive correlation is walkDistance and the highest negative the killPlace.



This is all the correlation of variables under **multiplayer mode**. The highest positive correlation is walkDistance and the highest negative the killPlace.

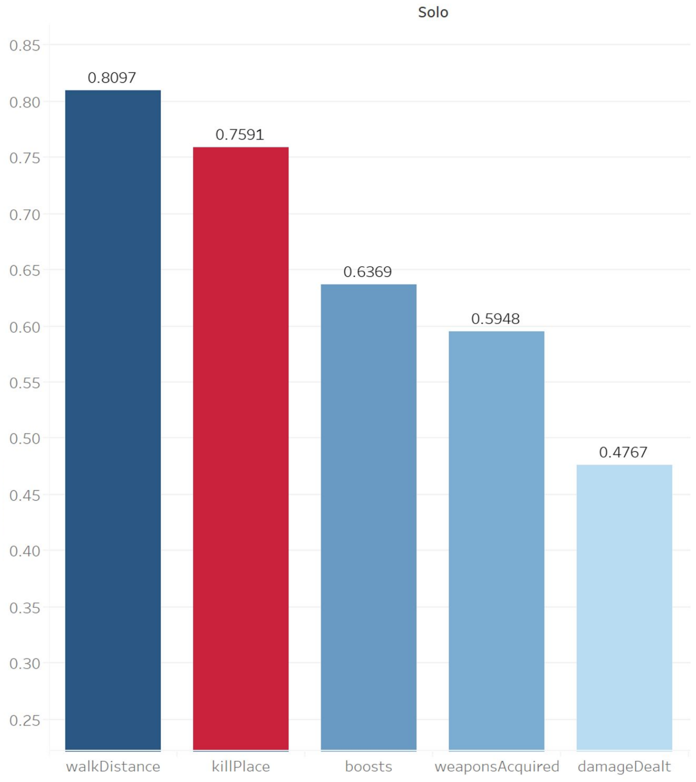


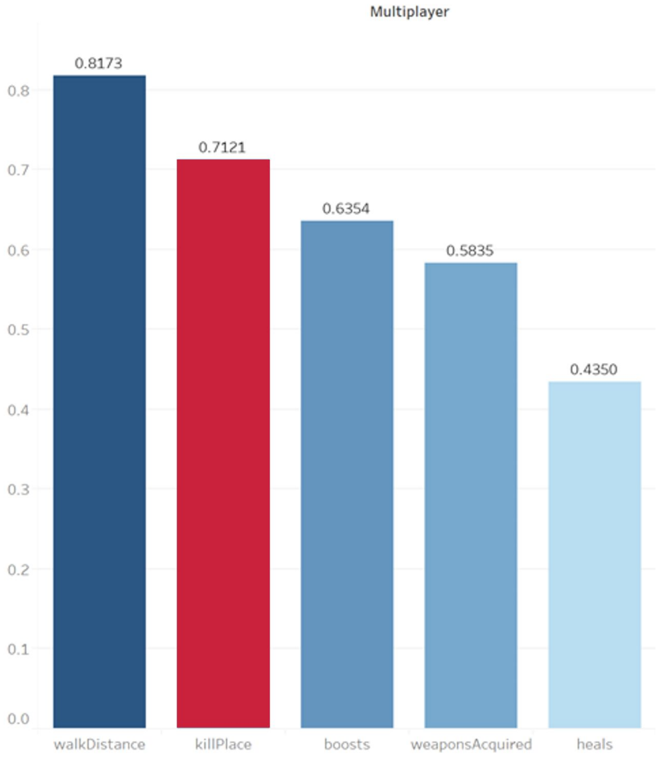
From the analysis of the correlation between winplaceperc and other variables, whether in the solo or multi mode, to survive for a long time, the most important first four factors are: walkDistance, killplace , boost, weaponsAquired. But the correlation of killplace and winplaceperc is negative.

1. **walkDistance** - Total distance traveled on foot measured in meters.
2. **killPlace** - Ranking in match of number of enemy players killed.
3. **boosts** - Number of boost items used.
4. **weaponsAcquired** - Number of weapons picked up.

Our interpretation of this data is that the player who can survive to the end is because the game time is longer and in order to collect more weapons and materials, the walking distance must be higher. If players want to survive for a long time, they must collect more weapons and boost materials in order to have enough ability to fight against other players. What interesting is, since killplace and winplaceperc are negatively related, the players who can survive for long time, they didn’t try to kill other players, which is contrary to our initial assumption. Therefore, players should try to avoid fighting with other players. Players should choose a place where people are less likely to land, and avoid moving to more places at the beginning of the game. These are more favorable strategies for survival.

Let’s compare and zoom to **the top-5 most correlated variables** with the target.





In the multiplayer mode, we found that the correlation of heal to winplaceperc is very high (0.43). But under the solo model, the influence of heal only reached the ninth place.

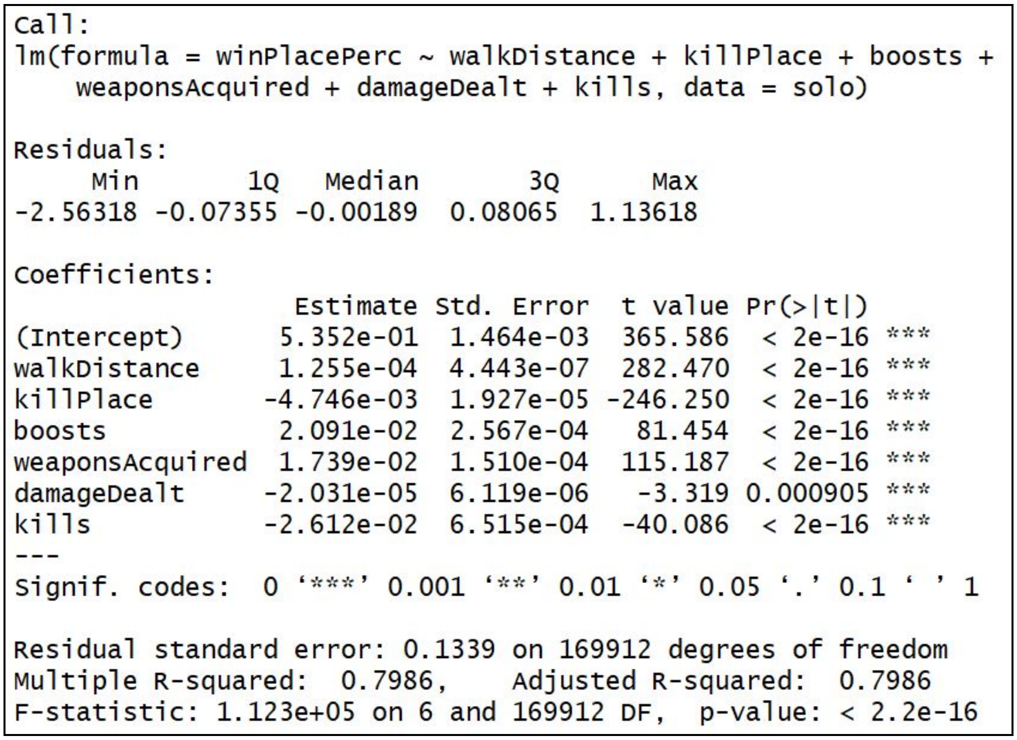
Consequently, if the player wants to get a higher ranking under the multiplayer mode, besides killing other players, he must focus on collecting more healing items and try to heal the injured teammates.

Data Driven Winning Formula

We built two models, one is for Solo Mode, another is for Multiple Mode.

The regression model for **Solo Mode**:

- Dependent variable: winPlacePerc
- Independent variable: walkDistance, killPlace, boosts, weaponsAcquired, damageDealt, kills

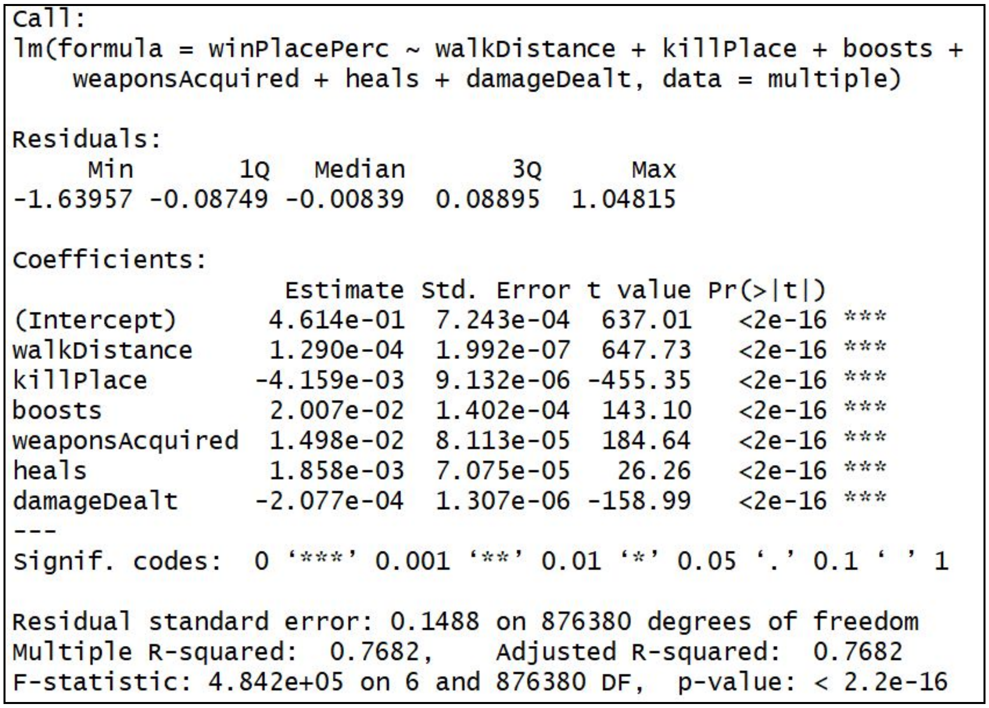


```
Call:
lm(formula = winPlacePerc ~ walkDistance + killPlace + boosts + 
    weaponsAcquired + damageDealt + kills, data = solo)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.56318 -0.07355 -0.00189  0.08065  1.13618 

Coefficients:
                  Estimate Std. Error  t value Pr(>|t|)    
(Intercept)      5.352e-01  1.464e-03  365.586  < 2e-16 ***
walkDistance     1.255e-04  4.443e-07  282.470  < 2e-16 ***
killPlace       -4.746e-03  1.927e-05 -246.250  < 2e-16 ***
boosts           2.091e-02  2.567e-04   81.454  < 2e-16 ***
weaponsAcquired  1.739e-02  1.510e-04  115.187  < 2e-16 ***
damageDealt     -2.031e-05  6.119e-06   -3.319 0.000905 ***
kills           -2.612e-02  6.515e-04  -40.086  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1339 on 169912 degrees of freedom
Multiple R-squared:  0.7986,    Adjusted R-squared:  0.7986 
F-statistic: 1.123e+05 on 6 and 169912 DF,  p-value: < 2.2e-16
```

All the independent variables are significant, and the Multiple R-squared is 0.7986, means the model fits the data well.

The regression model for **Multiple Mode**:

- Dependent variable: winPlacePerc
- Independent variable: walkDistance, killPlace, boosts, weaponsAcquired, heals, damageDealt

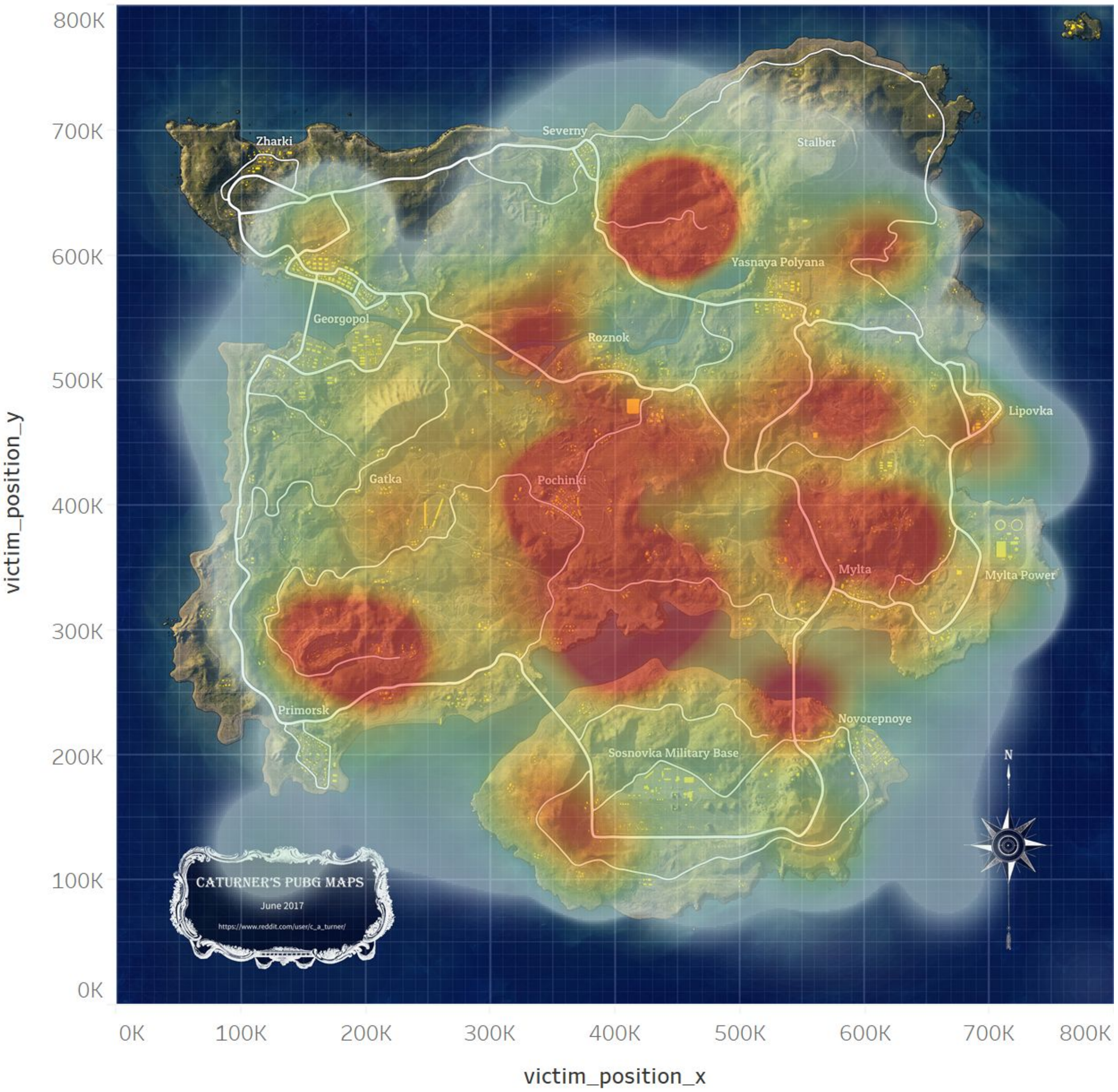


```
Call:
lm(formula = winPlacePerc ~ walkDistance + killPlace + boosts + 
    weaponsAcquired + heals + damageDealt, data = multiple)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.63957 -0.08749 -0.00839  0.08895  1.04815 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      4.614e-01  7.243e-04  637.01   <2e-16 ***
walkDistance     1.290e-04  1.992e-07  647.73   <2e-16 ***
killPlace       -4.159e-03  9.132e-06 -455.35   <2e-16 ***
boosts           2.007e-02  1.402e-04  143.10   <2e-16 ***
weaponsAcquired  1.498e-02  8.113e-05  184.64   <2e-16 ***
heals            1.858e-03  7.075e-05   26.26   <2e-16 ***
damageDealt     -2.077e-04  1.307e-06 -158.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1488 on 876380 degrees of freedom
Multiple R-squared:  0.7682,    Adjusted R-squared:  0.7682 
F-statistic: 4.842e+05 on 6 and 876380 DF,  p-value: < 2.2e-16
```

All the independent variables are significant, and the Multiple R-squared is 0.7682, means the model fits the data well.

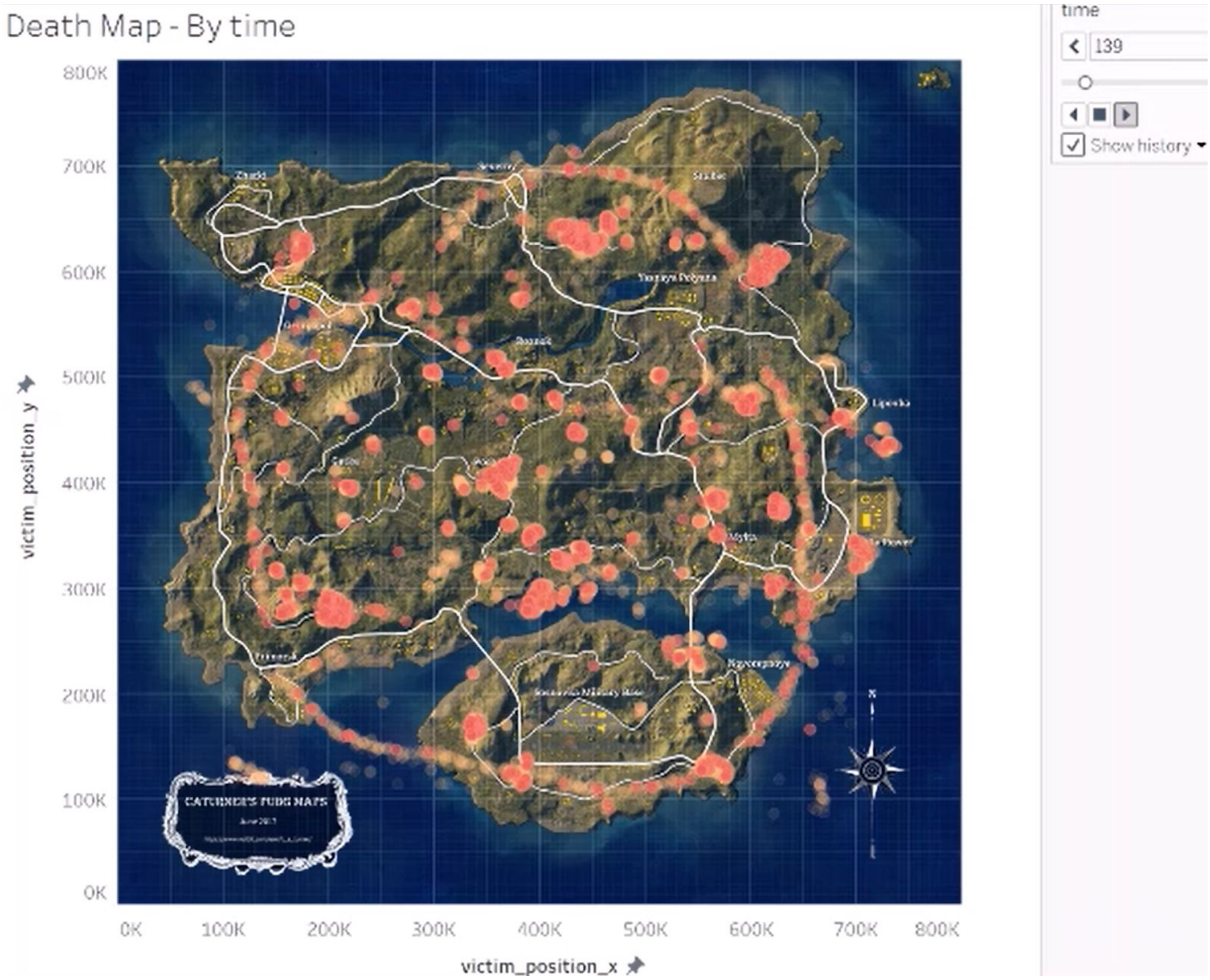
Death Heat Map:

This is the Erangel map in PUBG. We draw the location of each player's death on the map to form a density heat map. You can see that the death rate in the center of the map is very high, so players can avoid these high death rates area, such as the Pochinki city.



Death Map by Time

When we analyze the player's death position by game time, we find that in the second to third minutes, the player's death position will present a circle on the map. We think it’s because the first blue circle is closing, so the player will die if he stays outside the circle for too long.



Death Map - By time

time

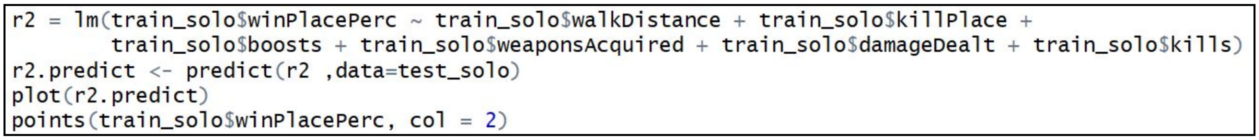
< 139

☑ Show history ▾

Prediction

Solo mode prediction:

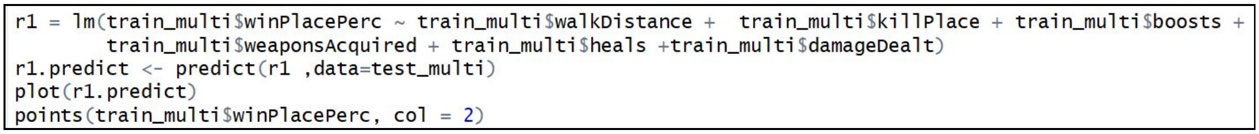
r2.predict is the prediction value generated by regression model r2.



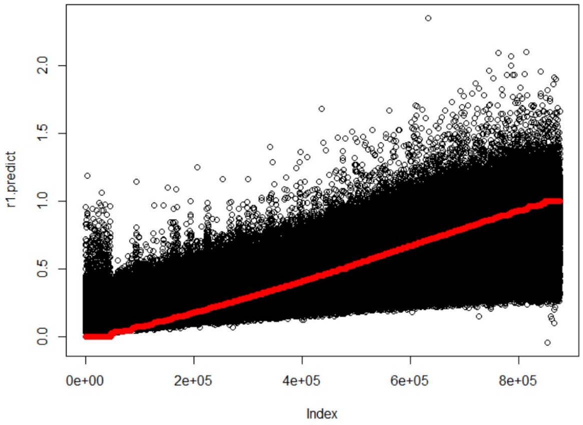
```
r2 = lm(train_solo$winPlacePerc ~ train_solo$walkDistance + train_solo$killPlace +
        train_solo$boosts + train_solo$weaponsAcquired + train_solo$damageDealt + train_solo$kills)
r2.predict <- predict(r2 ,data=test_solo)
plot(r2.predict)
points(train_solo$winPlacePerc, col = 2)
```

Multi mode prediction:

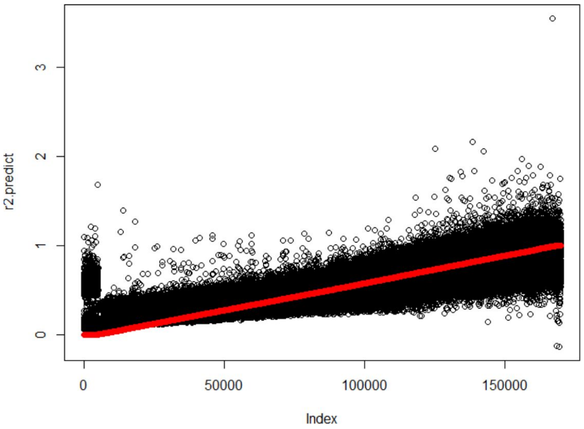
r1.predict is the prediction value generated by regression model r1.



```
r1 = lm(train_multi$winPlacePerc ~ train_multi$walkDistance +  train_multi$killPlace + train_multi$boosts +
        train_multi$weaponsAcquired + train_multi$heals +train_multi$damageDealt)
r1.predict <- predict(r1 ,data=test_multi)
plot(r1.predict)
points(train_multi$winPlacePerc, col = 2)
```



Multi mode



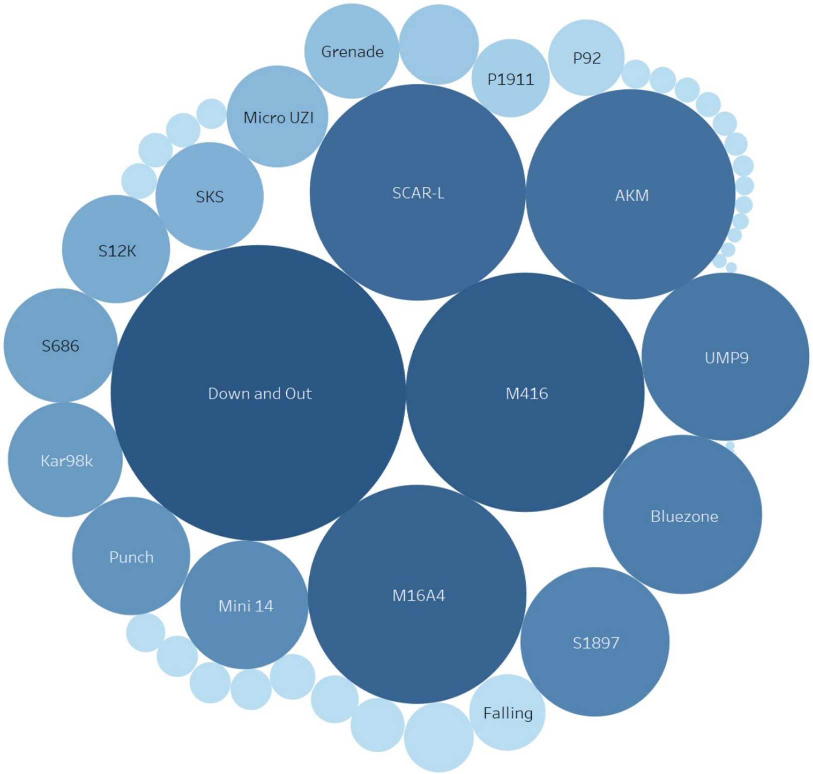
Solo Mode

We used the regression model of solo and multiplayer mode to predict the percentage of winning placement, where 1 corresponds to 1st place, and 0 corresponds to last place in the match. Black dots are the prediction value generated by regression model. Red dots are the actual value of winning placement.

Data Visualizations

Causes of death in PUGB:

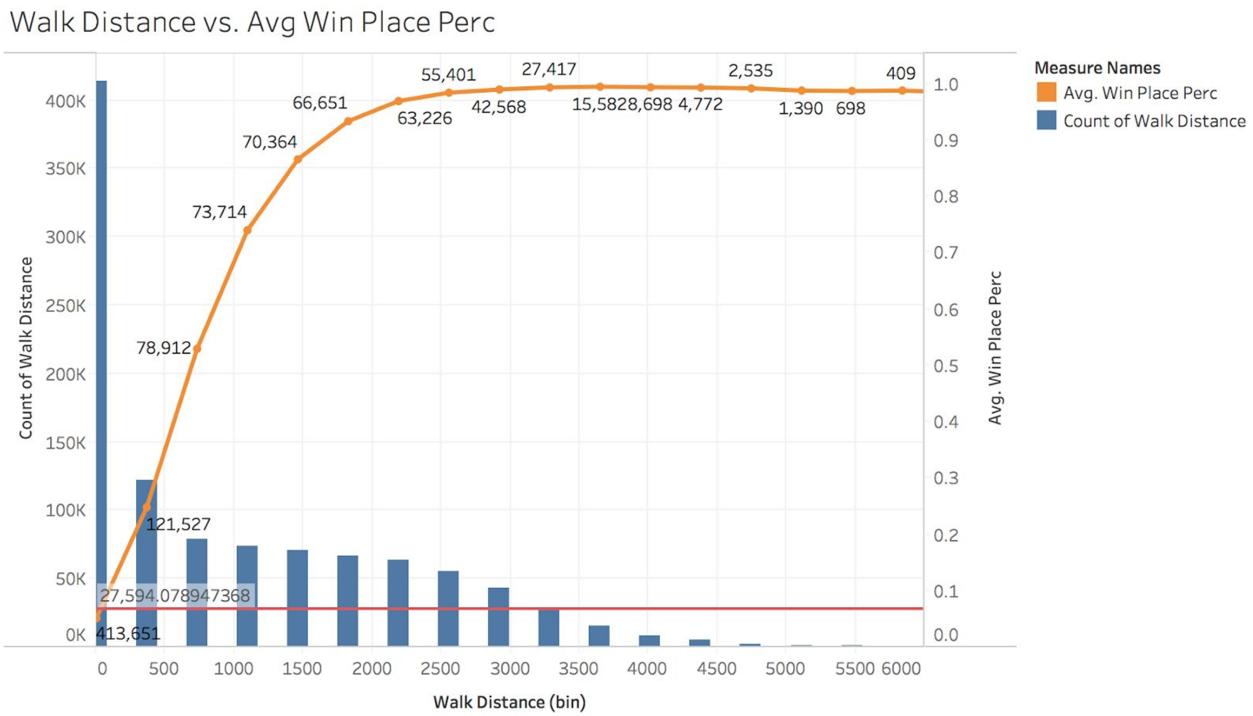
There are three biggest causes of death in PUGB. The first one is down and out, which has 17% players. Down and out is a player who got a kill from others, but they did not die immediately, they bleed slowly to death. There are 11% of people died due to the gun of M416. Moreover, there are 5% people died due to Bluezone, it is the district in which player will keep losing blood to die.



The Effect of Distance and Win Place Perc

There are some different distance will impact the win place percentile. We analyzed the relationship between the win place percentile and different factors of distance, which are walk distance, swim distance, and ride distance.

<u>Walk Distance vs. Avg. Win Place Perc:</u>

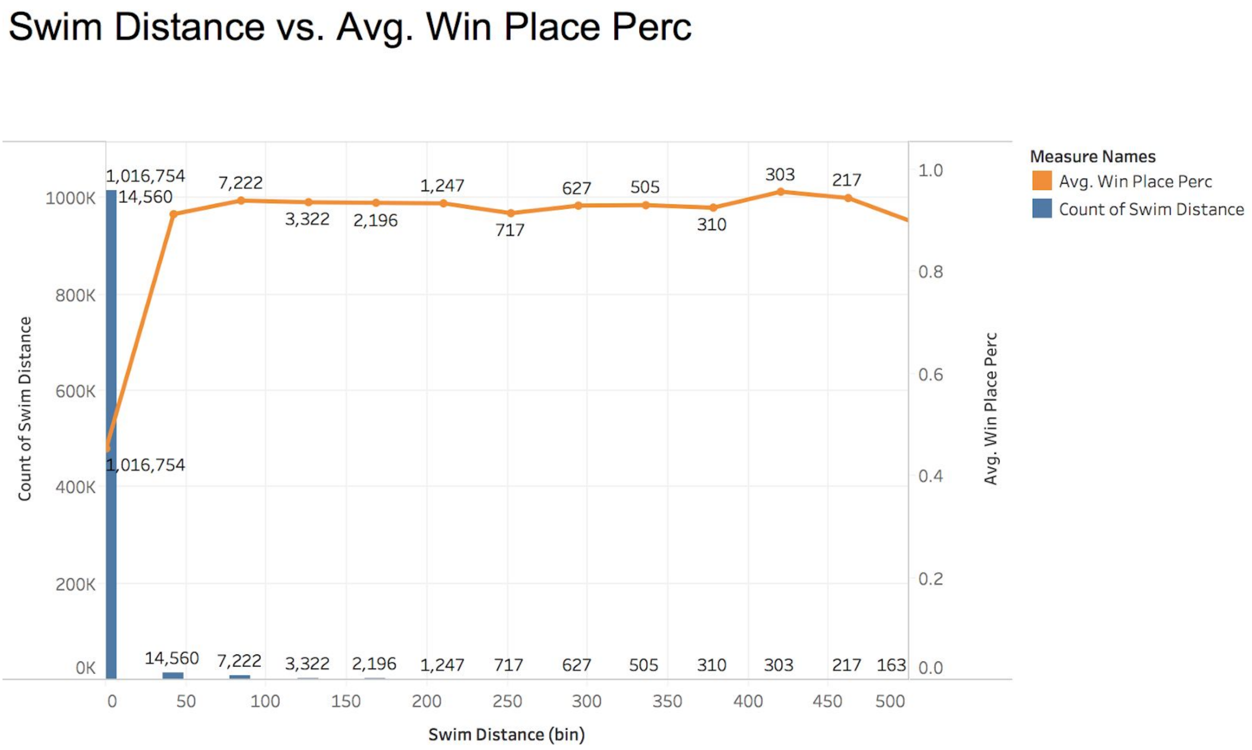


The trends of count of Walk Distance and Avg. Win Place Perc for Walk Distance (bin). Color shows details about count of Walk Distance and Avg. Win Place Perc. For pane Average of Win Place Perc: The marks are labeled by count of Walk Distance.

From the above graph, y-axes represent the walk distance and the x-axes represent the count of walk distance; the color of orange is indicating that the avg. win placement percentile. Walk distance and win place perc are strongly positively correlated. A player who has a longer distance can have the higher wins place perc. The average of walk distance by a player is

1150.557 meters (according to the R-studio). And there are 413,651 players have no walking distance, those players may be killed or offline when they land. As the graph shows, a player has 5840-meter walk distance approach to the winning place perc of 1.

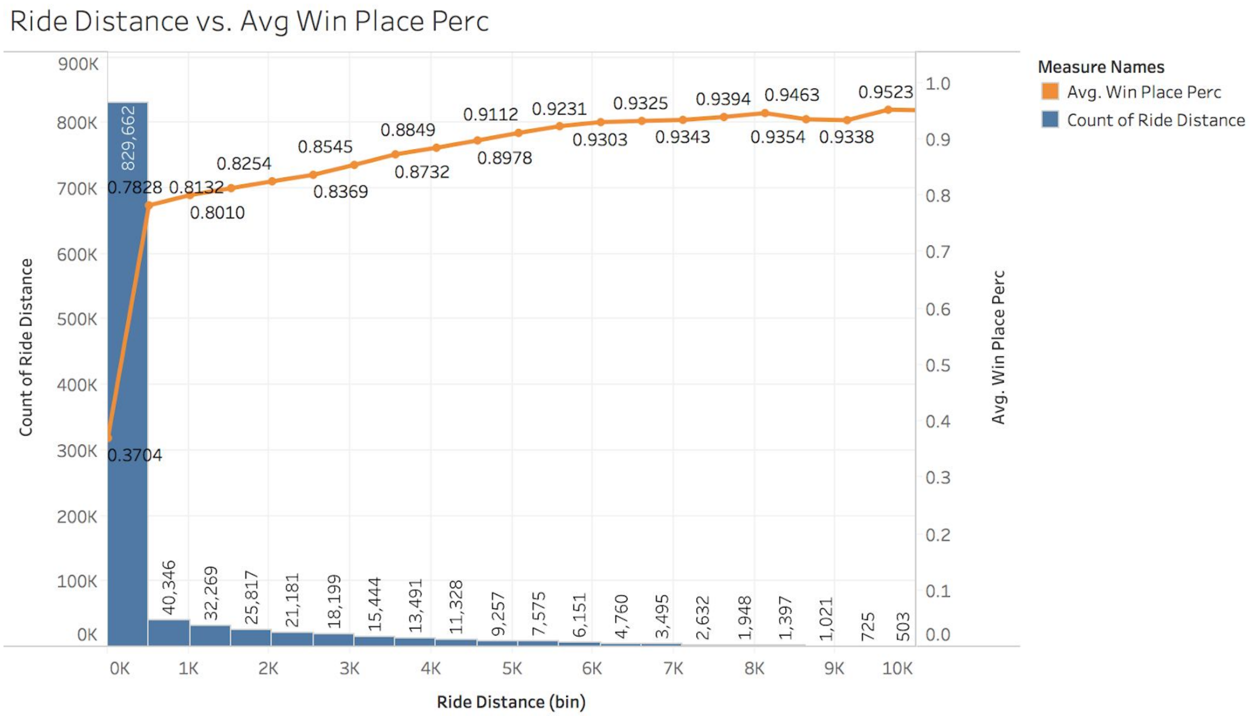
<u>Swim Distance vs. Avg. Win Place Perc:</u>



The trends of count of Swim Distance and Avg. Win Place Perc for Swim Distance (bin). Color shows details about count of Swim Distance and Avg. Win Place Perc. The marks are labeled by count of Swim Distance.

Swim distance and the average win place perc have a small correlation. The average of player's swim distance is only 4.5 meters (according to the R-studio). More than 1 million players do not swim during the game. Most players will not choose to swim unless the restriction of the bluezone.

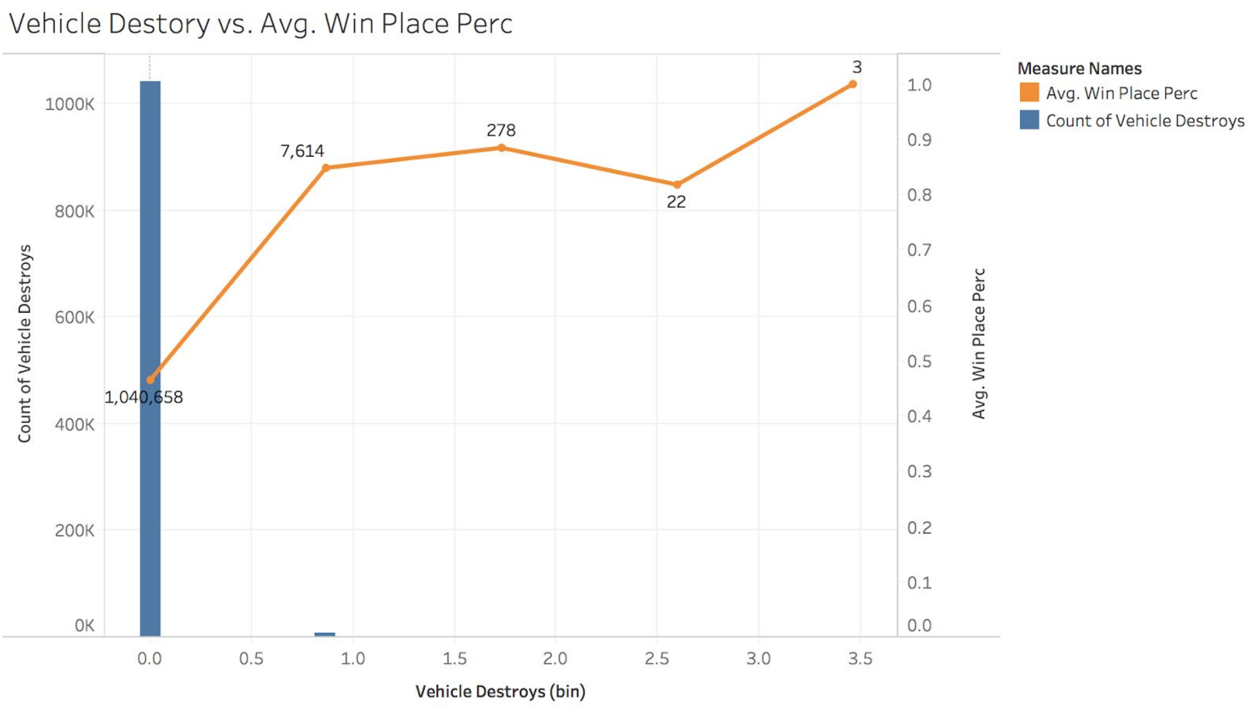
Ride Distance vs. Avg. Win Place Perc:



The trends of count of Ride Distance and Avg. Win Place Perc for Ride Distance (bin). Color shows details about count of Ride Distance and Avg. Win Place Perc. For pane Average of Win Place Perc: The marks are labeled by Avg. Win Place Perc. For pane Count of Ride Distance: The marks are labeled by count of Ride Distance.

The average of player drive a distance is 590 meters. There is a small correlation between ride distance and avg. win place perc. As the graph shows, the father of ride distance drives, a player has higher avg. win place perc.

Vehicle Destroy vs. Avg. Win Place Perc:



The trends of count of Vehicle Destroys and Avg. Win Place Perc for Vehicle Destroys (bin). Color shows details about count of Vehicle Destroys and Avg. Win Place Perc. For pane Average of Win Place Perc: The marks are labeled by count of Vehicle Destroys.

We find the interesting thing is that a player who can destroy a vehicle indicates that the player has skills and destroy a Vehicle may increasing a player's chances of winning. For example, a player destroyed three vehicles and his win place per is 1.0.

Identifying Cheaters

We can not only understand the different variables which correlate to the winning percentile but we can use the data to identify cheaters and tricks and techniques which are used to cheat during the game. From the data that we have collected, we could analyze the following types of cheaters:

1. **Aim Hacks**

   There are some cheating softwares which will take control of a players aim and automatically target it towards opponents. These cheaters can have high kills and headshot rates.

2. **Speed Hacks**

   The cheating software gives the player a massive speed increase, meaning they can go from one side of the map to the other in seconds.

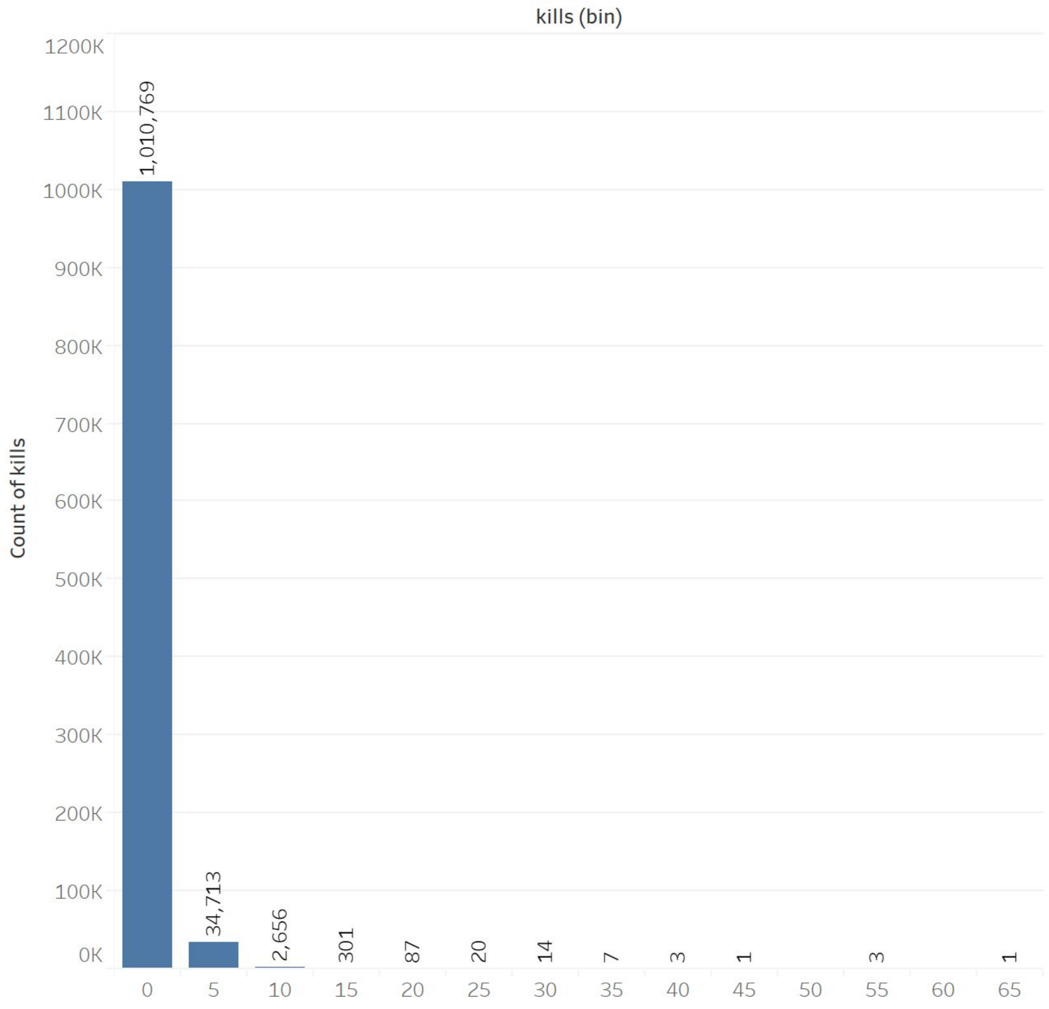
3. **Super Snipers**

   The weapon with the longest range in PUBG is AWM, which can reach up to 650 meters. So if the longest distance between player and player killed at time of death is higher than 650 meters, these players have a high chance of cheating.

Aim Hacks

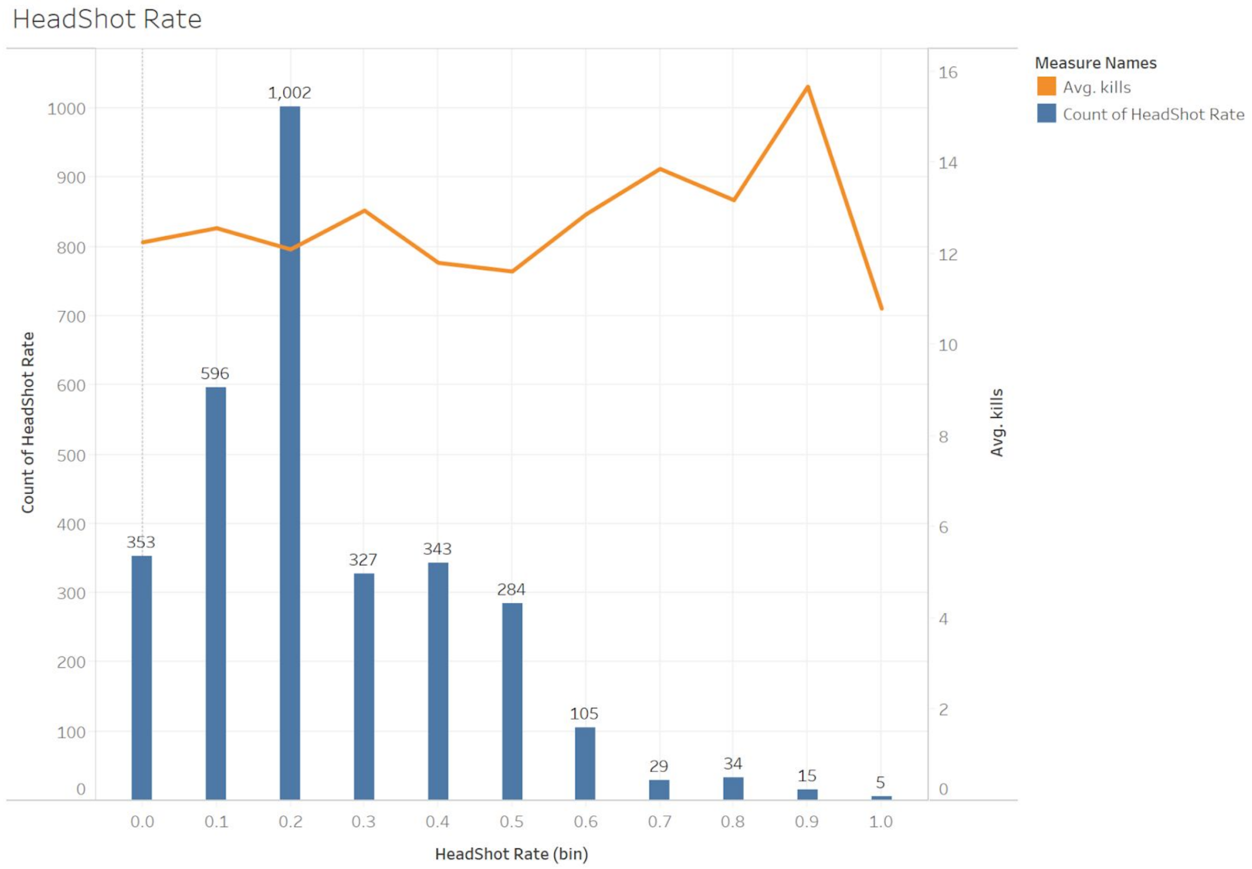
Unrealistic number of kills

In the total of 1048575 players in the dataset, the average number of kills is 0.92. Below is the histogram of kills. More than 1 million players ended the game without killing anyone. But there are total 8 players who had killed more than 40 players in a match. As we are aware of the fact that there are hackbots or trigger bots which can be used during the game to kill the players, unrealistic number of kills can help us flag such players and monitor them for further symptoms of cheating.



<u>A Closer Look at the Headshots</u>

We tried to analyze the relationship between their headshot rate and kill in players with kill>10. In more than 1 million players, a total of 3093 players killed more than 10 players in a single game. Among these players, the headshot rate of most players is 0.2. However, there are 5 players with a headshot rate of 1, which means that each kill is a headshot. That doesn't sound normal, but after we analyzed these five players, we found that their other variables are normal. Therefore, we can't assert that they have cheated, maybe they are just really good players.



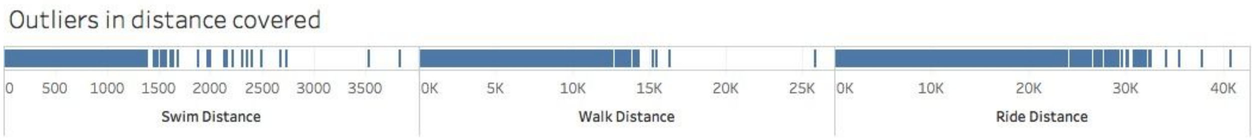
Speed Hacks

Outliers of swimDistance, walkDistance or rideDistance can be flagged to check if they are trying to cheat. After analyzing the data, we found the max and mean values. The standard deviation can be used to find the outliers.

- Walk distance maximum: 14910 meters
- Swim distance maximum: 2395 meters
- Driving distance maximum: 40700 meters

- Walk distance mean: 1150.557 meters

- Swim distance mean: 4.480585 meters
- Driving distance mean: 590.5788 meters


Outliers in distance covered

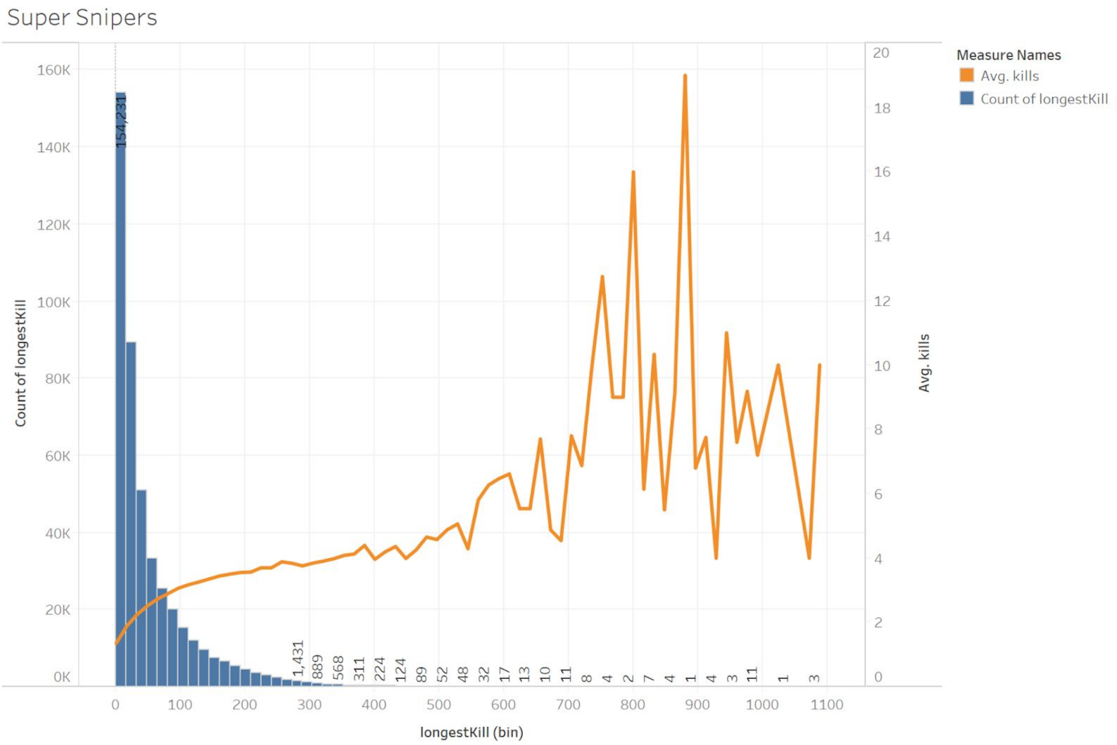
Swim Distance, Walk Distance and Ride Distance. The data is filtered on Walk Distance, which keeps 38,599 of 38,599 members.

Super Snippers

The average attack range for all weapons in pubg is 36.78 meters.

The weapon with the longest range in PUBG is AWM, which can reach up to 650 meters.

About 150,000 players have the longest killing distance of 0 meters, which means they had close quarters combat with other players. The longest range of weapons in PUBG is only 650 meters. But among all players, 134 have a longestKill greater than 650. So we believe that these players are likely to use cheating software.


Super Snipers

Final Thoughts

After verifying our hypotheses with the data, following are the findings from our Exploratory Data Analysis.

- To increase your chances of winning, play multiplayer mode. From analysis, we found that the chances of lasting in the game are higher as you can have team-mates helping you with heels to recover from injuries.
- Use our map to strategize your starting point.
- The SCAR and M16 may be the most useful weapons to you. Loot first and then shoot.
- Swim only if you need to hide , but don't enter the water if you are running from someone as swimming has a highest exposure with no deference.
- Wheels are better than legs as you are protected with the additional shield and from the analysis we saw that very few players target the vehicles to destroy
- Be defensive and not offensive - Number of kills do not help in improving your chances to win
- Reposition in advance and don't wait for circle to shrink as at the time of shrinking, many people are off guard and are at a higher risk of getting killed.
- Don't cheat using data analysis, we can flag and monitor players to target cheaters

References

Data Resources:

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

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

EDA:

https://www.kaggle.com/deffro/eda-is-fun/notebook

https://www.kaggle.com/rejasupotaro/cheaters-and-zombies

Weapons:

https://pubg.op.gg/weapons