PUBG Data Visualization & Analysis

By Yuwen Zhang

Background

For this project, I decided to do something different: I am an avid lovers of video games and are especially passionate about the game PlayerUnknown's Battleground, so I decide to analyze data from that game.

Graph from here

Game Intro

- PlayerUnknown's Battleground, or PUBG in the following, is a first/third-person shooter battle royale style game that pits 100 players on a large island where players or teams fight to the last one remains.
- Players are airdropped onto the island where they can scavenge buildings to collect weapons, ammo, armor, first-aid, and other resources.
- Players strategize their game plan by either fighting or hiding until they become the last person or team surviving.
- A blue zone will appear randomly later into the game and players must remain outside the blue zone to avoid extra damage from the zone. The blue zone will zoom in every few minutes to corral players closer and closer together and the damage of staying in the blue zone will increase more and more.

Graph from here

Data Set Intro

I have found a dataset [1] of over 720,000 competitive matches on Kaggle, which is scraped from a game tracker website [2] .

I used the last three datasets of both kill and match stats, and analyzed games on the Erangel map in single player mode.

Graph from the Kaggle data set

Note

- ➔ There are several maps in pubg including Erangel, Miramar, Sanhok, and Vikendi. Erangel is the oldest and more popular map.
- ➔ There are three player modes including single (1 player/team), duo (up to 2 players/team), and squad (up to 4 players/team)

Some questions that I intend to explore:

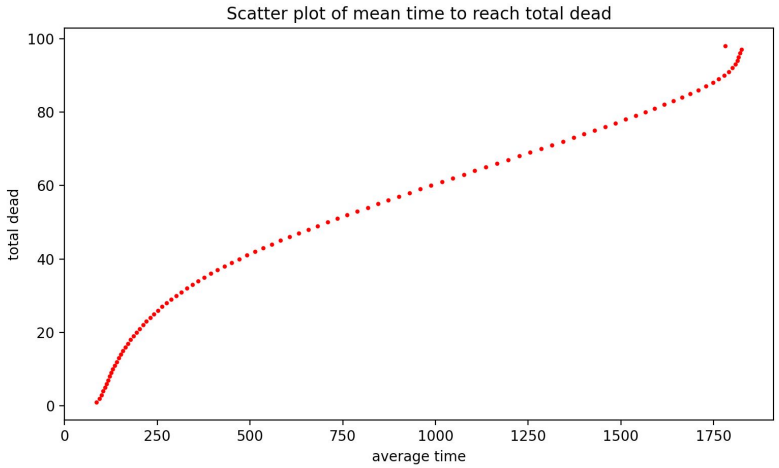
- ➔ Does time vs the amount of people dying in the games follow a linear relationship? Quadratic? Exponential?
- ➔ What are the most popular weapons during the start vs the end of the games?
- ➔ Where does most of the fighting take place?
- ➔ Is it better to fight or hide to gain a higher rank?

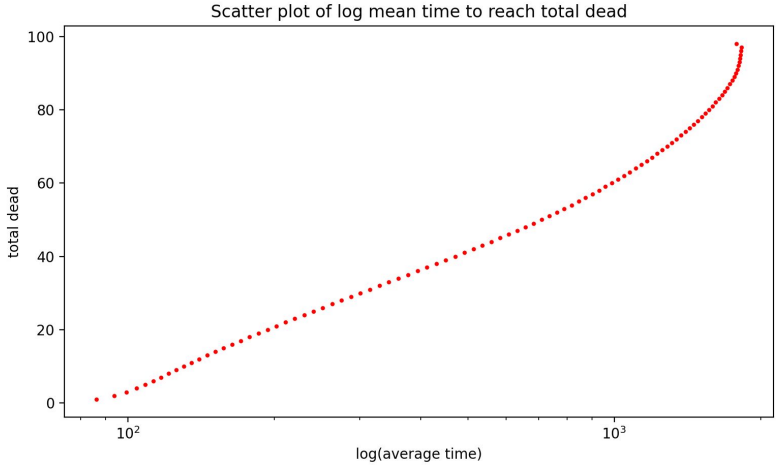
DEAD VS TIME

First we take a look at the relationship between average time and total players dead in each game.

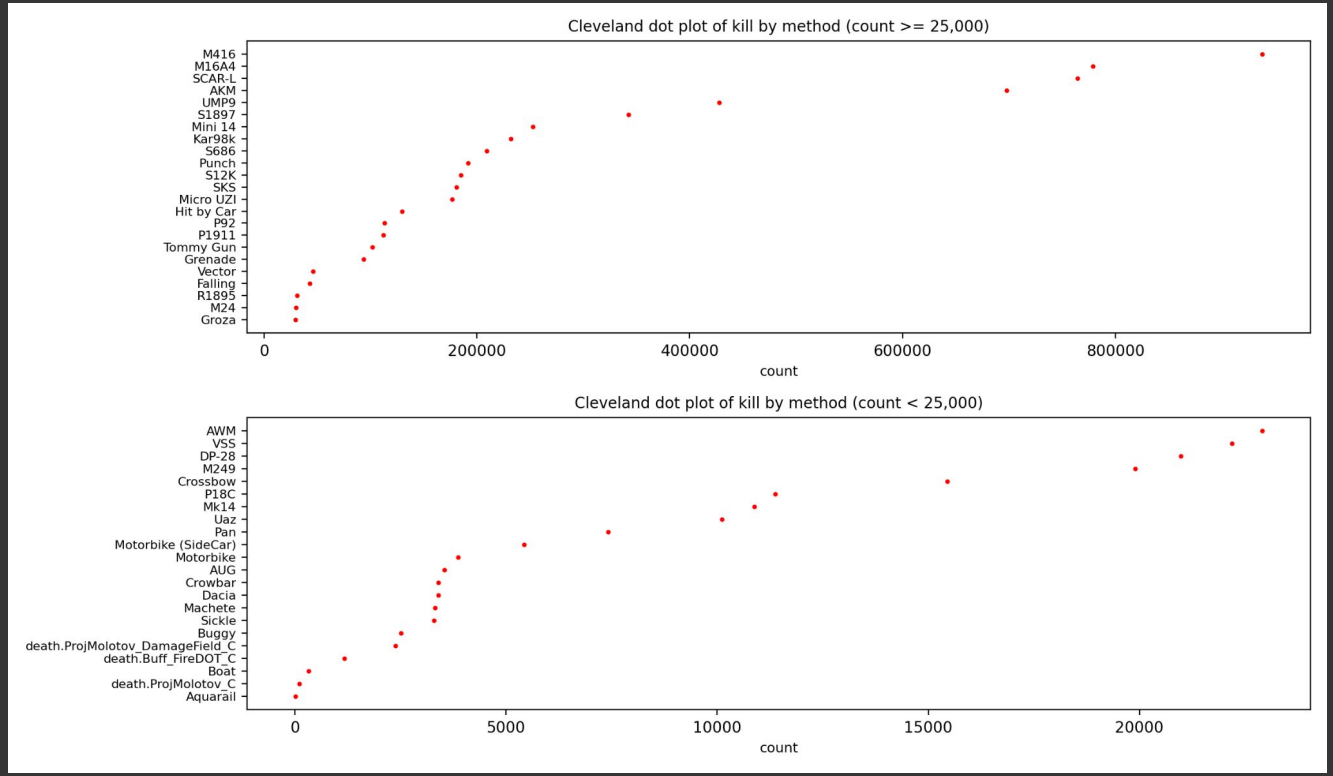
Clearly, there are more people dying at the beginning of the game, and as the game proceed, the amount of people dying slows down.

After transforming the x axis to a log frame, it's clear to see the the total players dead vs average time follows roughly a logarithmic relationship





KILL BY METHOD

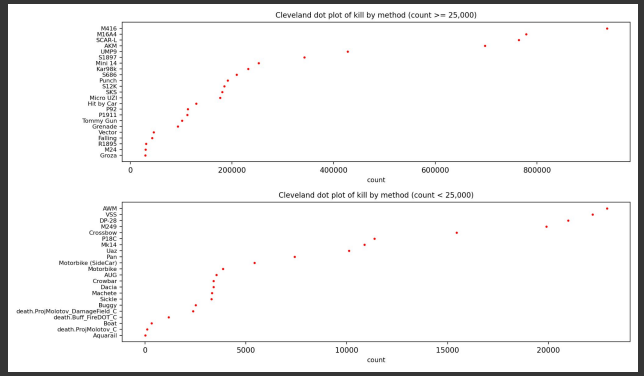


Now we move on to the popular weapon or kill method of the games.

To analyze that, we create a cleveland dot plot of players killed by each method.

From the graph we can see that M416 is the dominate weapon of the game.

KILL BY METHOD CONT



Moreover, weapons from the 5.56mm ammo family (M416, M16A4, SCAR-L etc), are in general much more popular than weapons from the other ammo family.

Assault Rifles, such as M416, M16A4, and SCAR-L from the 5.56mm ammo family, and AKM from the 7.62mm ammo family, are much more popular than other weapons kinds, such as Sniper Rifles (Kar 98k, M24 etc) or Designated Marksman Rifles (Mini 14, SKS etc).

Other non-ammo weapon, such as pan, crowbar, or molokov are in general less popular weapon than guns.

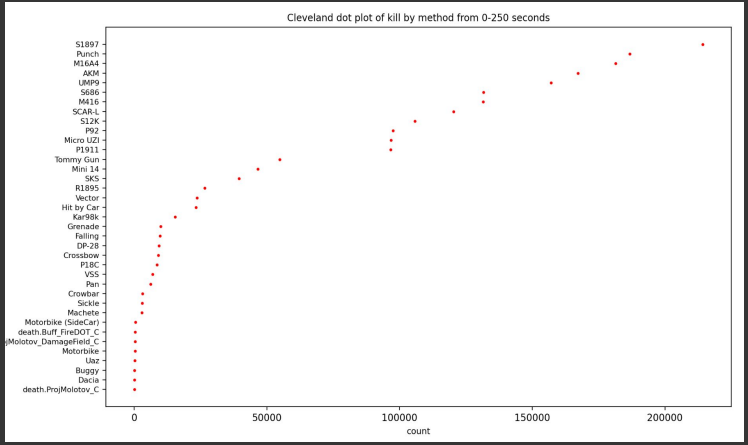
KILL BY TIME RANGE

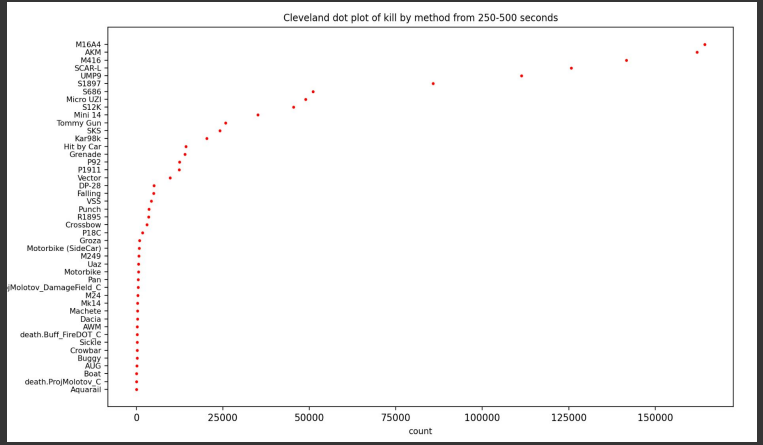
We now will analyze the dynamics of weapon use by different time frame.

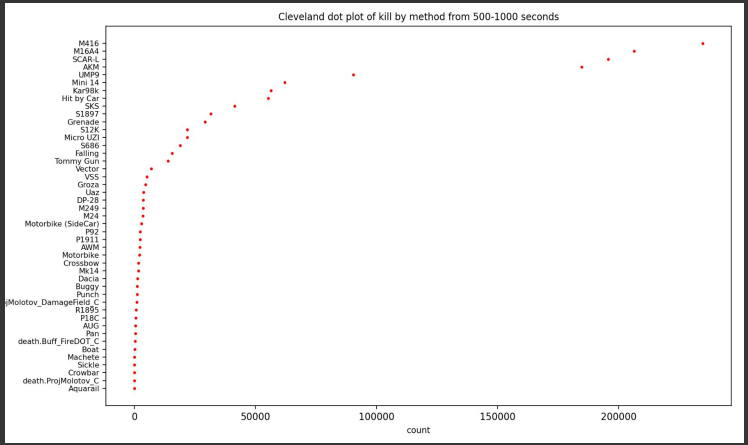
To do that, I split the data set based on kills happened in five time range.

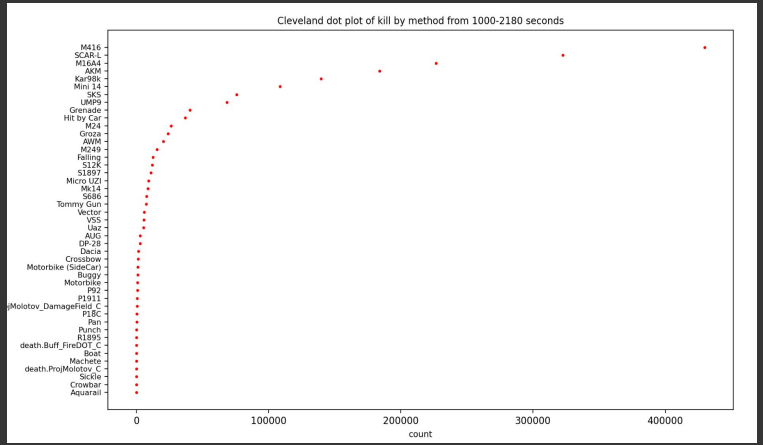
- **Early game**
  0-249 second
- **Mid-Early game**
  250-499 second
- **Mid-Late game**
  500-999 second
- **Late game**
  1000 second and above

Then I fit cleveland dot plots of weapon kills during different time range.

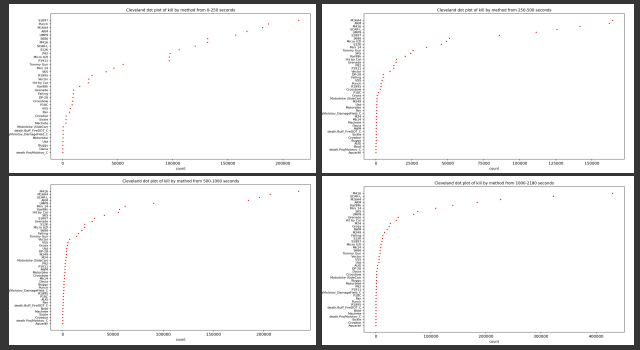








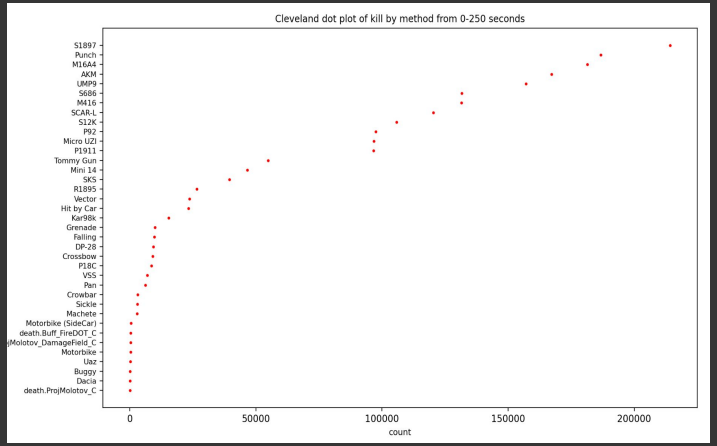
KILL BY TIME RANGE CONT

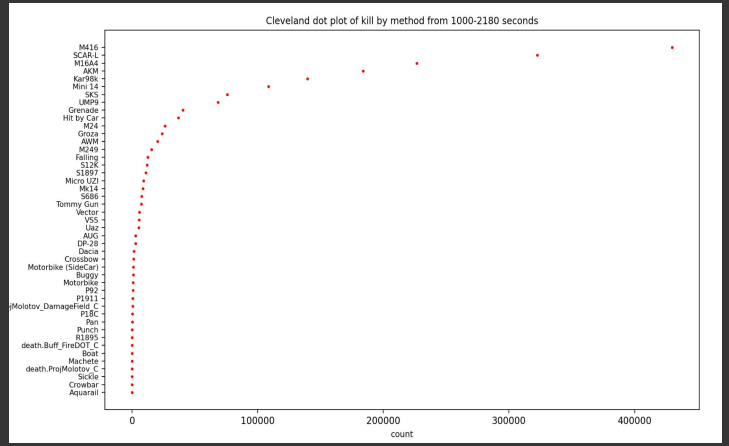


KILL BY TIME RANGE CONT

From the four plots, we can clearly see that at the beginning of each game, the most popular weapon is S1894, which is a shotgun well known for close distance combat. Since resources are limited at the beginning of each game, players will use shotgun to knock out close enemies as fast as possible.

Other than S1894, Punch is surprising also a popular method to kill enemies at early stage games: when resources are scarce, players have to utilize everything to kill enemies.





KILL BY TIME RANGE CONT

As game proceeds, we can see that the plot moves gradually to favor popular weapons like M416 and SCAR-L. The graph also gets increasingly polarized, as popular weapons are used repeatedly and unpopular weapons are being left out more and more.

WINNER WEAPON ANALYSIS

We now will analyze the winning player of each game and their choice of weapon.

For this analysis I will look at the winning player of each game and the weapon they used for the **last 10 kills** that they achieved in each game, if exist. Then I will create an **ALLUVIAL DIAGRAM** to track the movement of weapons of these last 10 kills.

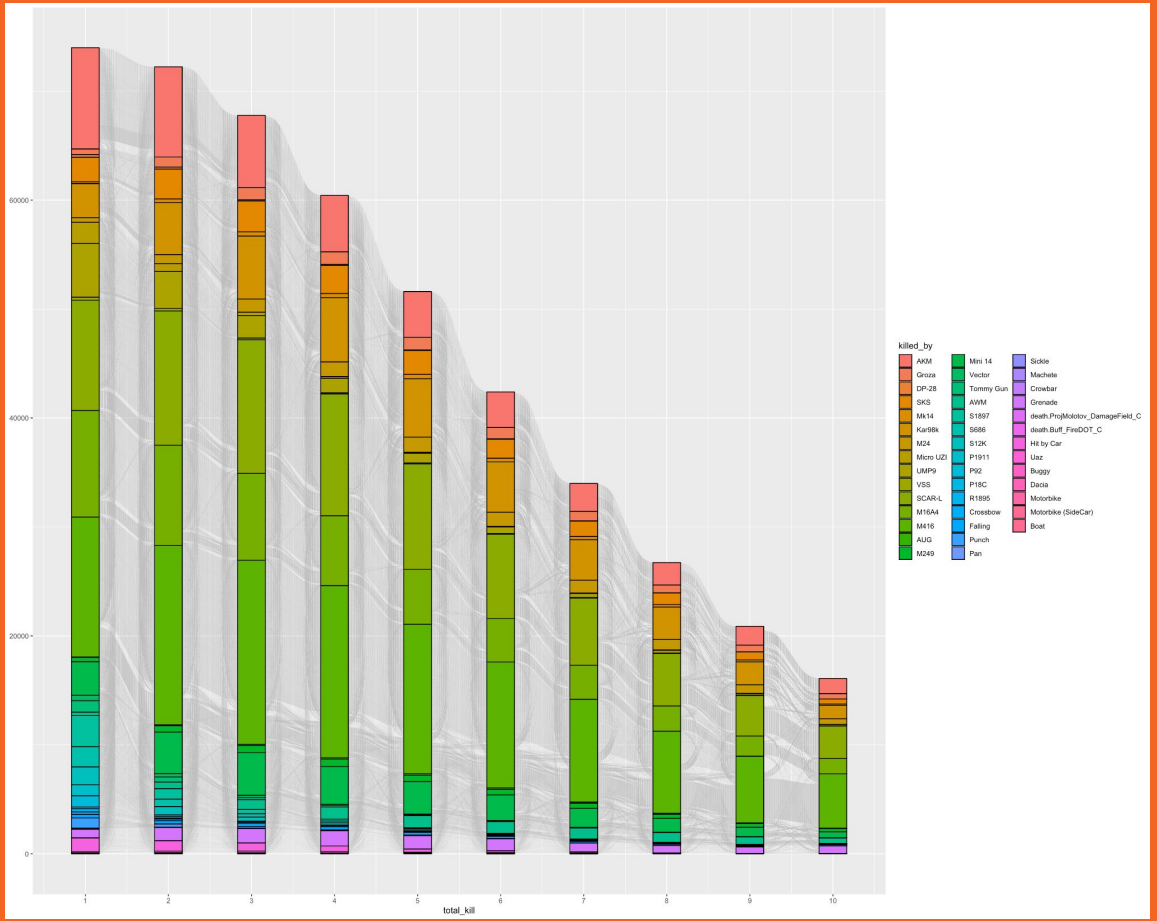
(Note that this is the only part where I ran my analysis in R, the rest are all in Python)

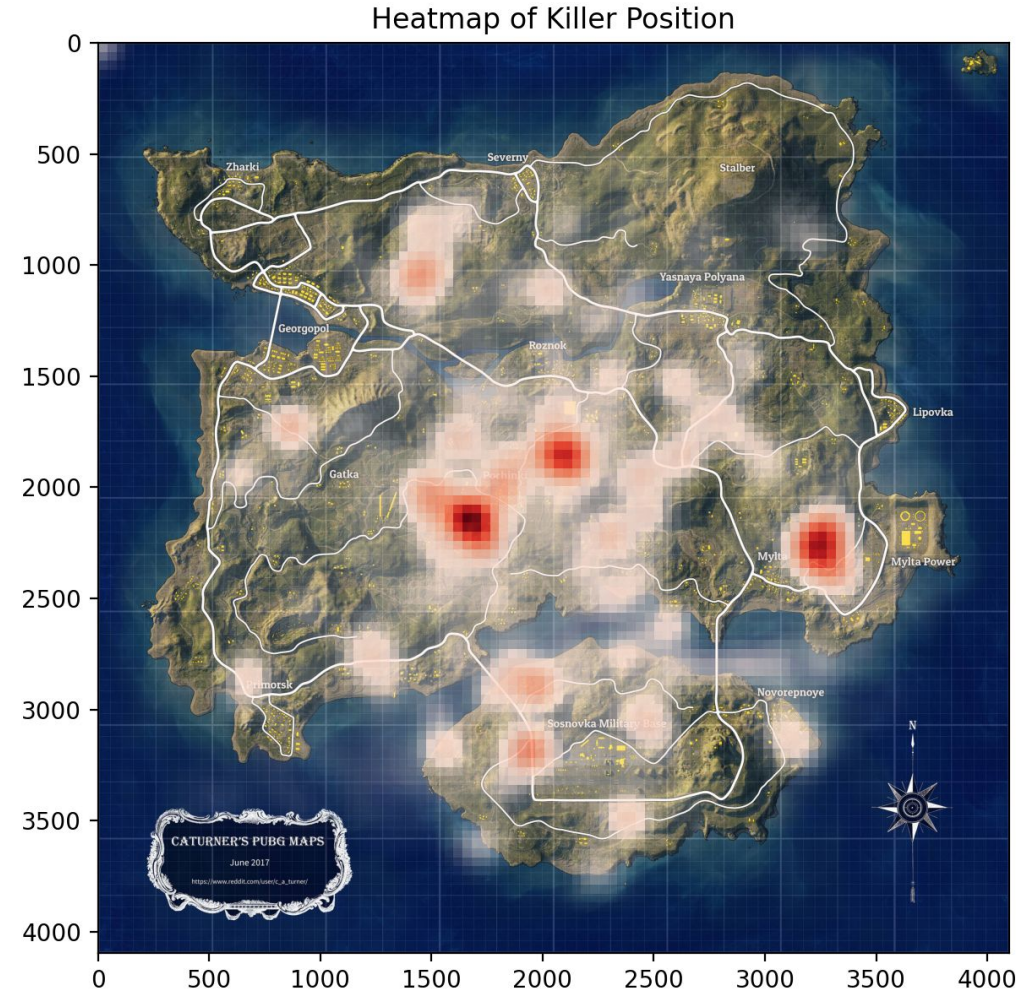
WINNER WEAPON ANALYSIS CONT

From the diagram, the first thing we can tell is that in reality, most winning players are not high-kill players, as we can see a gradual decrease of the amount of total kills of each player.

we can see that 5.56mm ammo weapons are again favored by most winning players.

Many transactions of weapons happened between switching a non-ammo weapon to ammo weapon, or an assault rifle to sniper rifle





KILLER POSITION HEATMAP

Now we move on to analyze the popular places where most fights take place. To do that we will create a heatmap of the killer position of each game.

From the heat map we can clearly see that the most popular locations for fights are Pochinki, Mylta, Sosnovka Military Base and their surrounding areas.

A good advice for beginner is to avoid these areas and choose places that were less popular in fighting.

K/D ANALYSIS

Moving forward, we will analyze the performance of players.

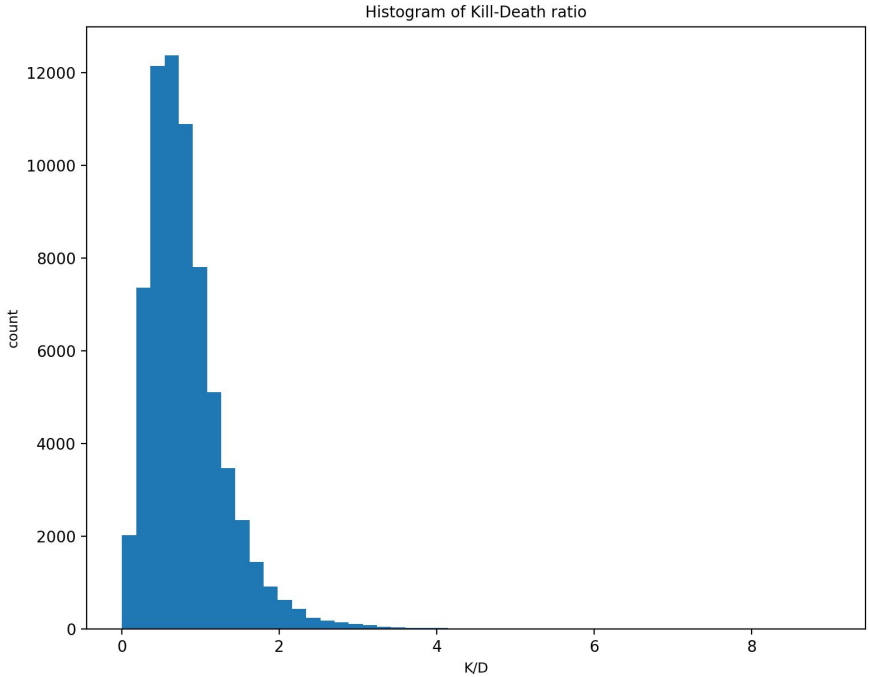
For that, I will introduce a concept called K/D, which is short for Kill-Death ratio. K/D is the total amount of people a player killed divided by the total amount of times a player got killed.

Typically, the higher the K/D is, the better the player is at the game.

K/D ANALYSIS CONT

From the histogram of K/D value, we can see that the distribution of K/Ds is extremely skewed to the right.

To gain more insight, we split the data set based on high-low K/D values.

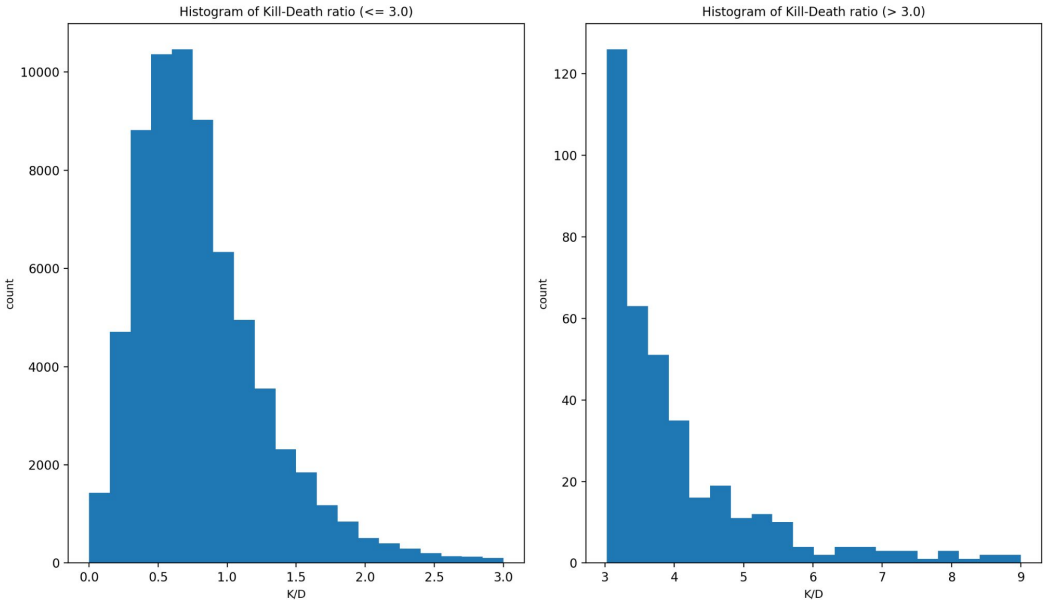


K/D ANALYSIS CONT

I split the data set based on low K/D value (K/D <= 3.0) and high K/D value (if K/D > 3.0).

From the graph, we can see that in reality, most of the K/Ds ranged from 0.5 to 1.5.

In reality, is extremely rare to kill more that 3 people per death on average.



INDIVIDUAL ANALYSIS

Now we will look at the individual behavior good vs bad player.

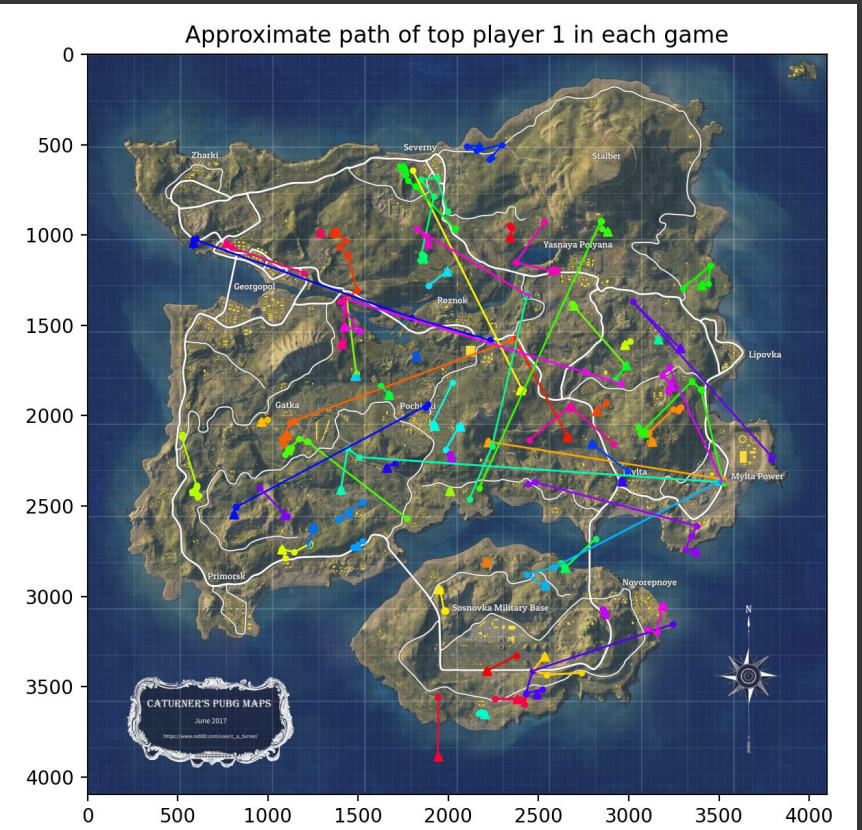
For this, I shrunk the dataset to players whose total death counts range from 46 to 50 and have killed at least 40 people in total.

We will look into 3 players with the highest K/D and 3 players with the lowest K/D, and their approximate path travelled in each game.

Note

For each path, the triangle indicates the end point of this player, either they were killed or won the game at the spot.

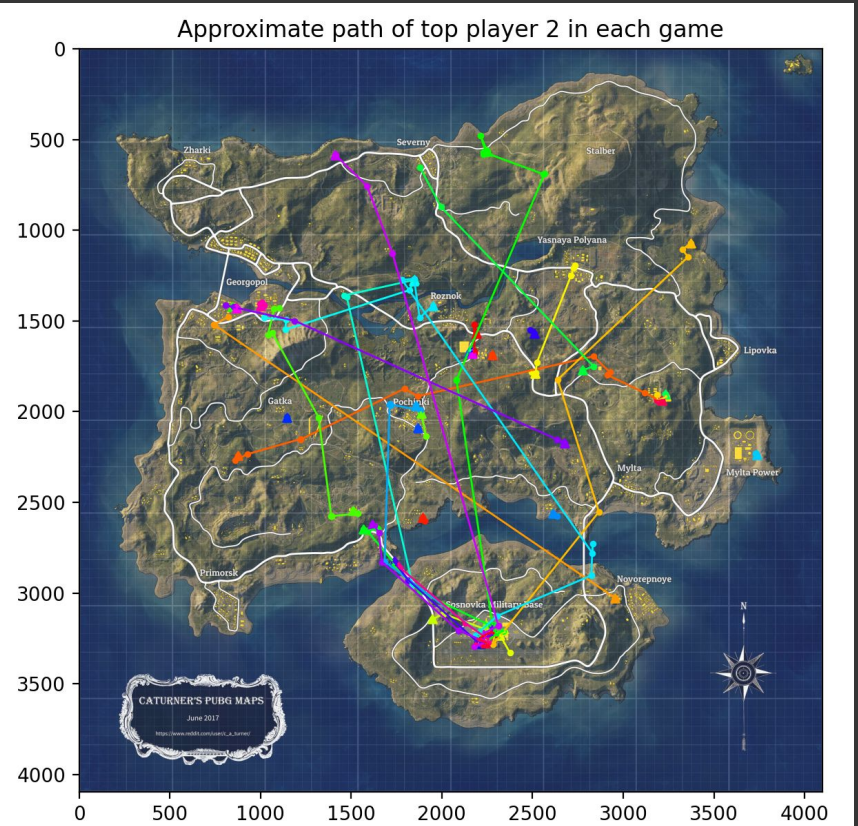
INDIVIDUAL ANALYSIS CONT (TOP PLAYER 1)



The behavior of the first top player seems to be pretty random.

They will land at random places and won't have games with particularly high kills, but they can still have a consistent kills so maintain a high K/D

INDIVIDUAL ANALYSIS CONT (TOP PLAYER 2)

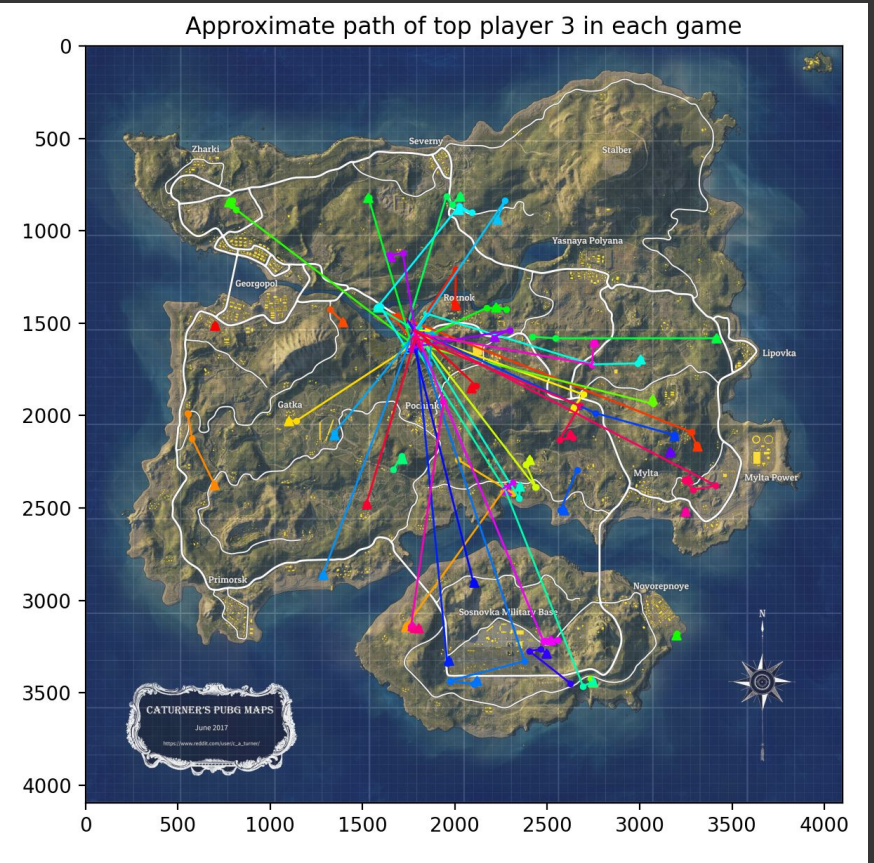


The behavior of the second top player is more predictable.

They love to start at military base, which is well known for heavy fighting and abundant resources.

Then they will travel upwards following the circle.

INDIVIDUAL ANALYSIS CONT (TOP PLAYER 3)

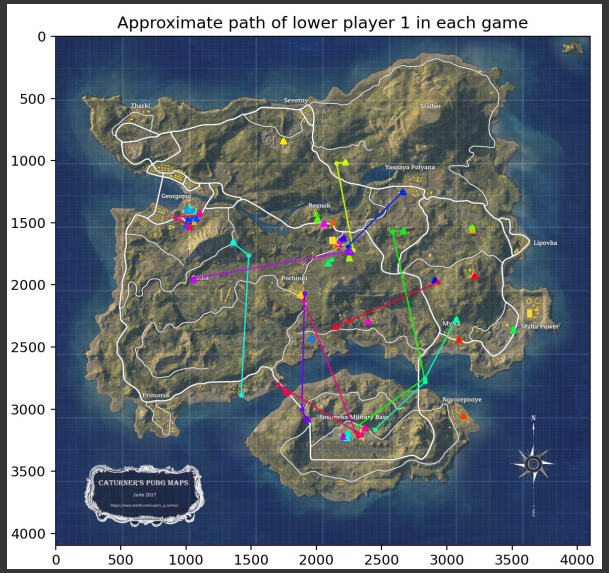


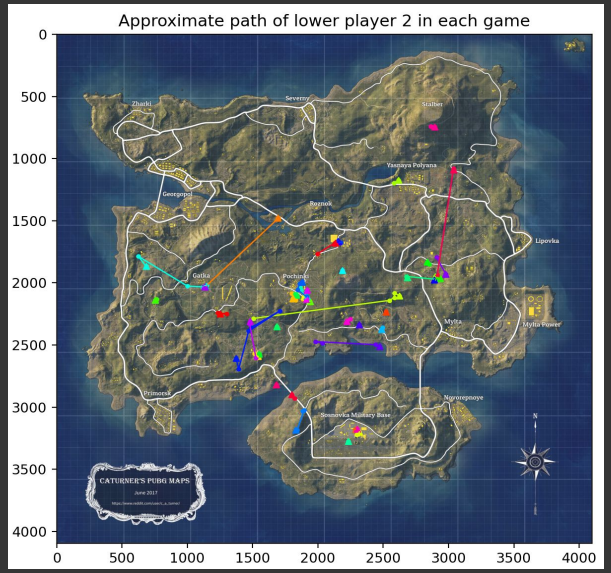
The behavior of the third top player is the most interesting one.

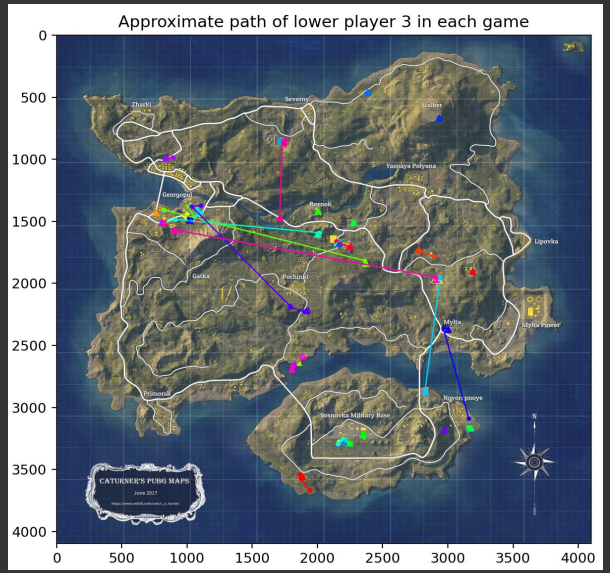
Their favorite drop spot is around the Pochinki area, which is probably where the most heavy fighting takes place.

This shows that this player is really confident about their skills and love to take risks to fight against other players.

INDIVIDUAL ANALYSIS CONT (BOTTOM 3)







From the three graphs, we can see that these three players with lower K/A also have their preferred drop spot. Player 1 prefers **Georgopol and Military Base** area, Player 2 prefers **Pochinki area and Military Base** area, and Player 3 also prefers **Georgopol and Michitary Base** area. However, perhaps because of skills, they never are able to out-compete other players in those heavy fighting region, or they hide until the last minute to win the game.

KILL RATIO VS RANKING

Finally, I will analyze whether it is better to fight or hide to gain a better placement in the game.

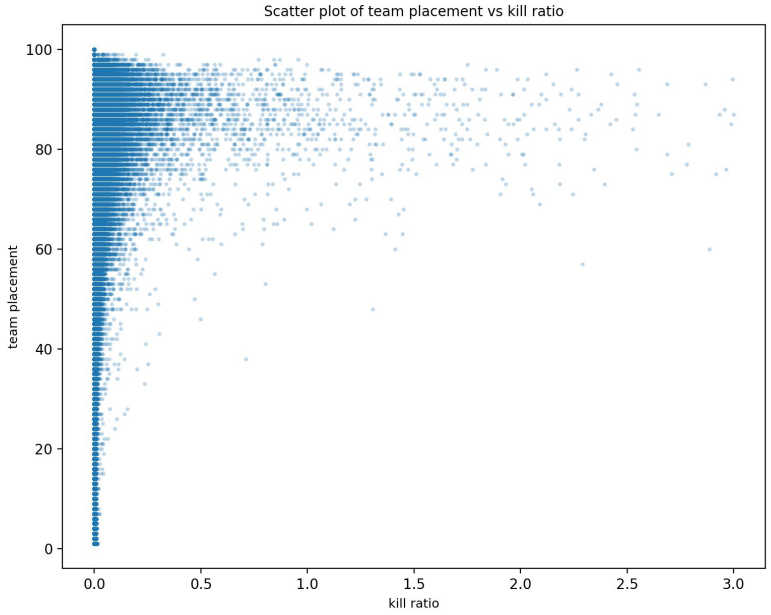
To do that, I will introduce a new variable called the **kill ratio**.

Kill ratio is measured by the amount of people that a player killed for each game divided by the sum of player distance ride and player distance killed.

Since the bigger amount of player killed indicates heavier level of fighting and more distance of traveling means higher possibility of hiding, the bigger the kill ratio, the more fighting and less hiding of the player.

From the graph of team placement vs kill ratio, it is clear that the there is a positive correlation between higher kill ratio and lower team placement.

Therefore, combined with the behavior analysis of top vs low K/D players, for regular users, to get a better rank, it is wiser to **HIDE** than **FIGHT**



THANK YOU!