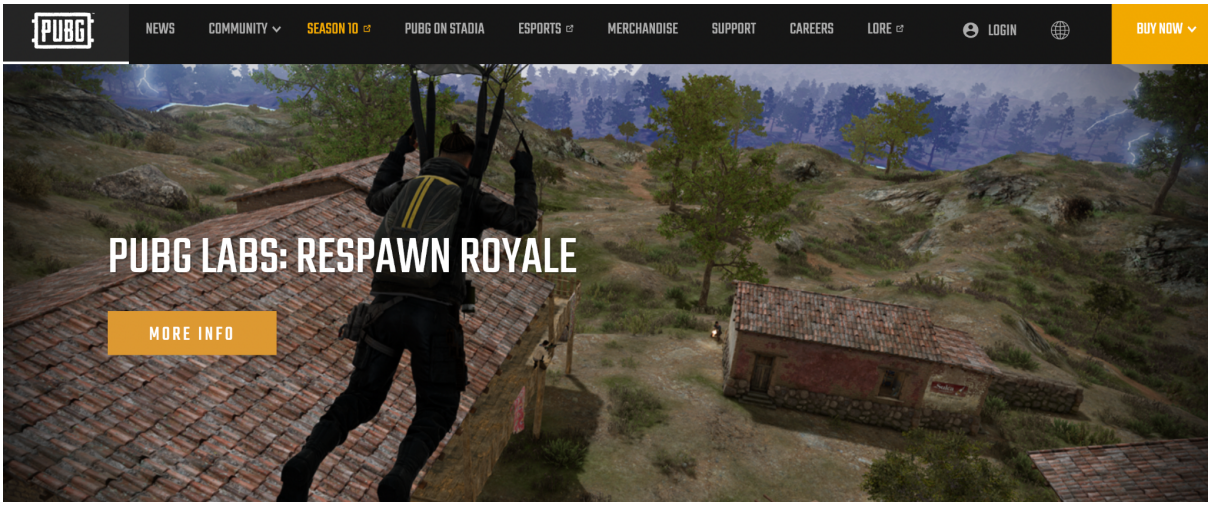
Big Data Project - Statistical Analysis on PUBG Data

494929 Sophia Yang, 489567 Yunfei Zhang,
492009 Yuxuan Deng, 490851 James Liao

Background introduction

Player Unknown's Battlegrounds is an online multiplayer shooter battle game developed by a South Korean video game company Bluehole. It was first published in 2017 and spread among young people swiftly. At each round of the game, more than 90 players are matched in a plane above a deserted island. Players can choose when they would like to be airdropped, and their task is to search through the towns and buildings for weapons, armor and first-aid kits for their combat. They can choose to search for people and kill, or to hide from being found by others. A bluezone will appear periodically a few minutes after the game so that players are forced to congregate into a smaller zone gradually and confront each other till the last survivor stands out.

One of the most distinct features of this game is that players form groups to help each other and the winner of this game is the group where the last survivor belongs. Because of this team-based feature, PUBG has become a must-have on parties and networking for young people worldwide. In the game, what matters is not just your own survival but also how you collaborate with teammates to search and exchange the weapons you collected, fight against other groups, and save each other's lives.



Description of the data:

Our dataset is collected from an online open source dataset called 'PUBG Match Deaths and Statistics'. As the name indicates, this dataset contains information from the popular online game Player Unknown's Battlegrounds. It was uploaded to Kaggle by KP, who claimed that a total of over 720,000 competitive matches data were extracted by using pubg.op.gg, a game tracker website. The dataset can be downloaded from https://www.kaggle.com/skihikingkevin/pubg-match-deaths

Our team used the 4th part of the death data, ‘kill_match_stats_final_4.csv’. The dataset is about 1.62 GB in size. It recorded meta information about deaths in each match.

Our dataset has 11,640,856 rows and 12 columns. Each row is a death record of one player and the 12 columns include detailed information about this ‘death’.

Detailed explanation of each column is shown in the table below:

Column Name	Description
killed_by	Reason why the player is killed, can be because of game settings such as bluehole, or be the name of a specific weapon the killer uses. There are in total 56 unique killed_by classes.
killer_name	name of the killer (alias)
killer_placement	the final ranking of the killer, ranging from 1 to 100
killer_position_x	the X position of the killer when the death happens
killer_position_y	the Y position of the killer when the death happens
map	the map on which this death happens. In our dataset there are only two maps, ERANGEL and MIRAMAR
match_id	An assigned unique code for each game
time	# of seconds since the game starts when the death happens
victim_name	name of the victim (alias)
victim_placement	the final ranking of the victim, ranging from 1 to 100
victim_position_x	the X position of the victim when the death happens
victim_position_y	the Y position of the victim when the death happens

Problem Statement

It is the idea of analyzing game data to produce interesting insights about how to play PUBG well that brought our team together. In this report, we aim to explore how to play games scientifically.

Specifically, we would like to know what weapons are most destructive or most frequently used, and also we’re curious if there are any map-specific features (i.e. which locations are the most dangerous) that can guide us to win in these two maps.

Why is this big data?

As introduced in Module 1 of this class, there is no clear definition of big data and as long as it cannot be processed using traditional tools it can be categorized as big data. Big data has three V's as its main features: Volume, Variety and Velocity. This dataset definitely fits into these features.

Volume: With more than 11 million rows, this data is impossible to be processed using traditional analytical tools like excel or in any local environment (personal computer) and we have to rely on web-based cloud computing platforms for data cleaning and analysis.

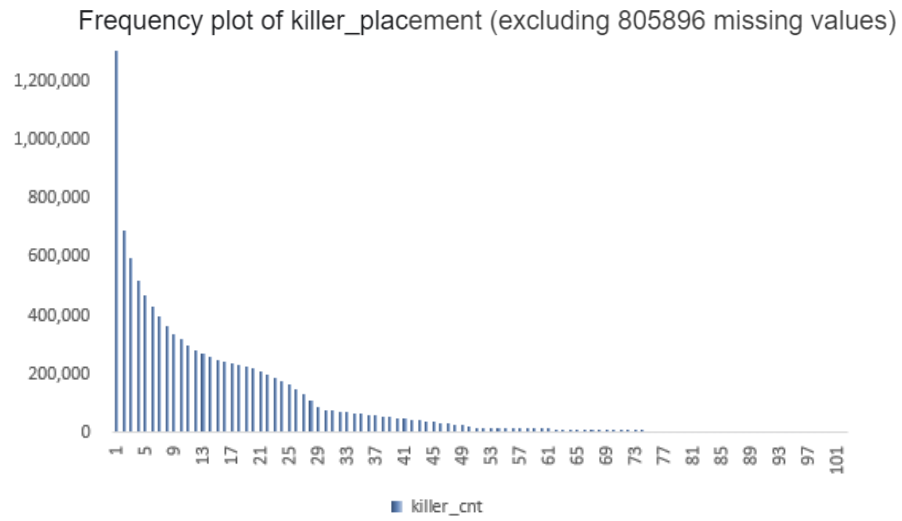
Variety: This dataset includes multi-dimensional data like time related information, user information in string format and geographic information that can be linked to a map.

Velocity: This game is very popular among young people. Every second, there are millions of players online and in the game, therefore such death data is generated at a very fast speed.

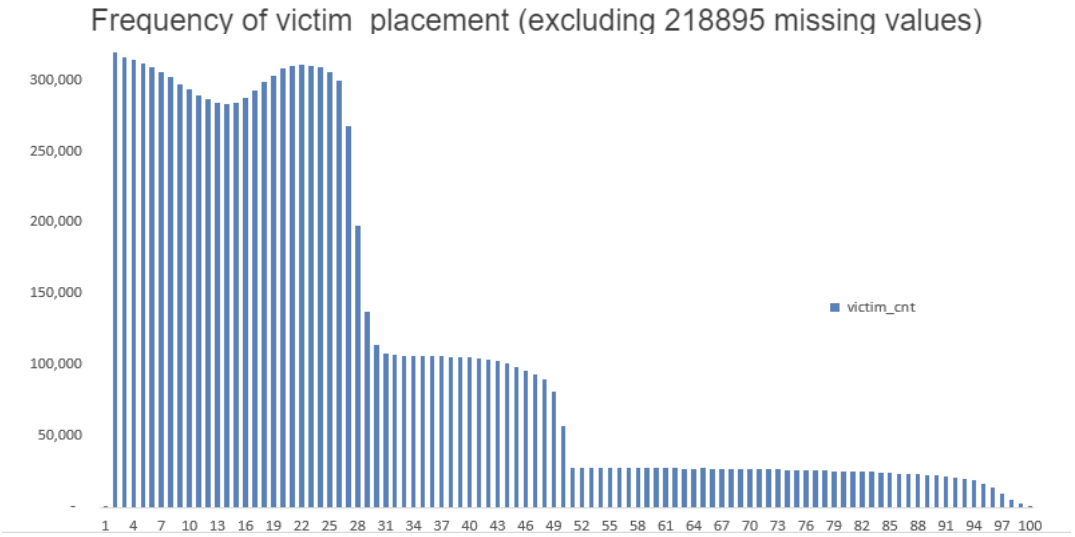
Descriptive Analysis

The data we use is fairly clean as it is well-structured and consistent in its format. Although there is a decent amount of missing values in each column, it is not affecting our analysis. For some records where players are killed by the game environment setting such as 'down and out' or 'bluehole', it is understandable that there will not be any information about the killer. In other cases where we observe missing values in placement, it may be because of multiple reasons, for instance, connection issues and system outage.

As previously introduced in the background, the ranking of a player is not the ranking of this person among the 100 individuals but that of the team this player belongs to, so we see many records where the killer_placement is even larger(meaning worse) than the victim_placement.



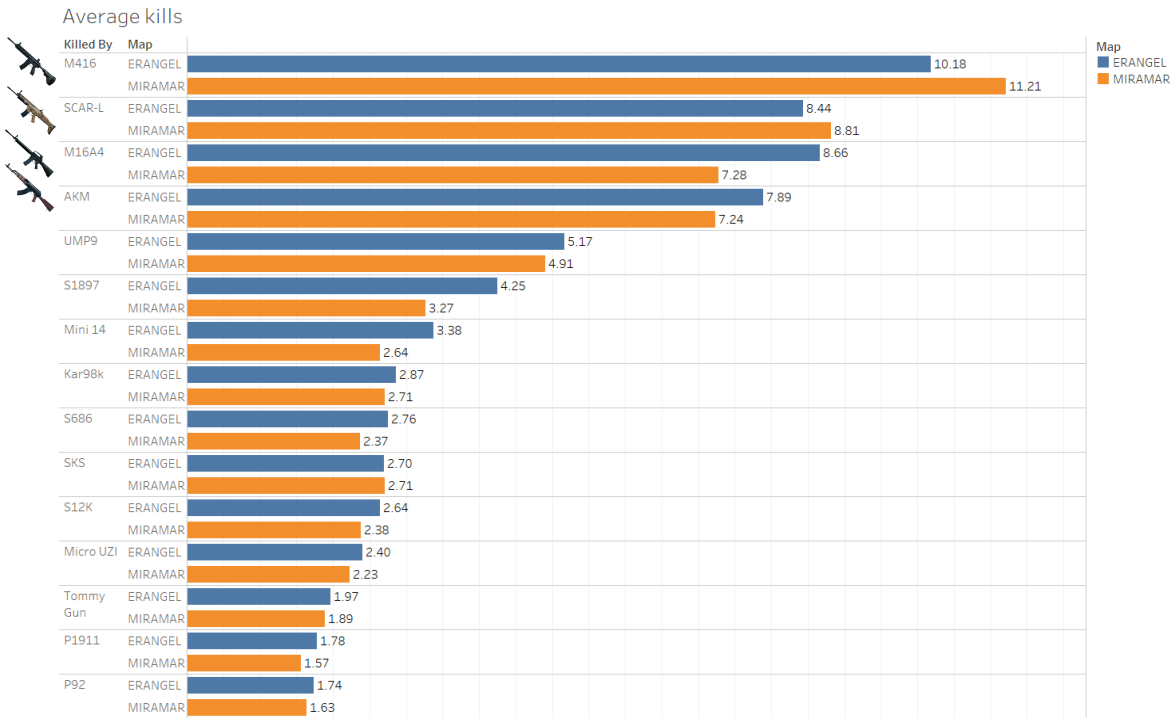
There are three ways for teaming up , “”Solo”, “Duo”, and “Squad”. Under the limitation of 100 players per game, the database should show flat stepping structure through the critical point- placement of 100, 50, 25 (the least placement of three teaming up types). However, the placement between 7~19 shows a convex outline, representing the missing team member phenomenon. The missing team member may result from wireless disconnection or be banned from the PUBG.



Method & Results

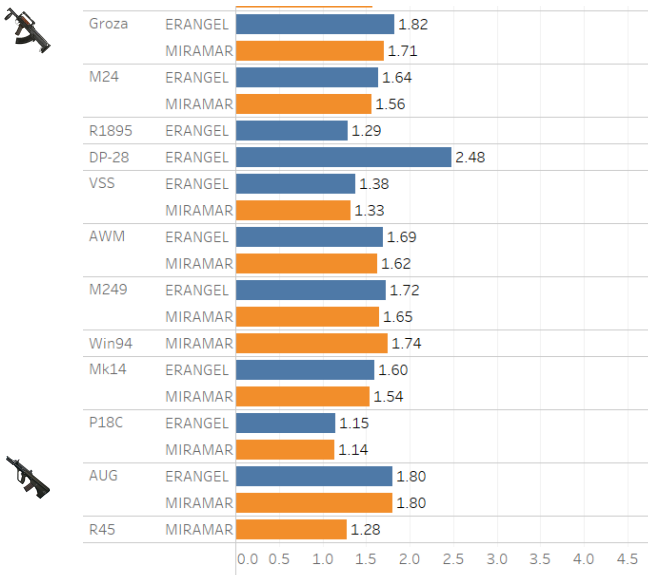
In this project, we mainly apply statistical analysis to demonstrate a high - level overview of interesting facts in the game. We apply MapReduce for simple calculation, and Impala, Tableau for visualization.We will start from statistics of weapon performance, then calculate the main threatening weapon in each period of a match, main threatening vehicle in each period of a match, and finally plot the top 5 most popular weapons used in the place where victim density is overall the highest.

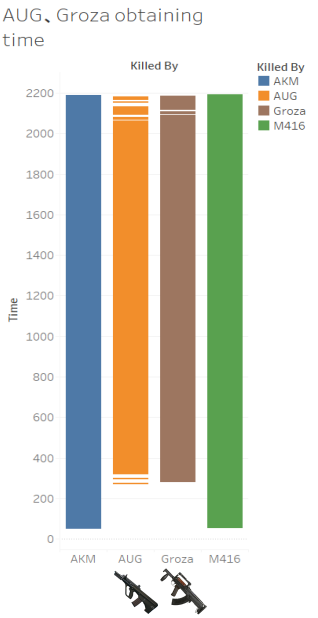
Chart(1) Average kills by weapon



From Chart(1) we can find that the assault rifle is overall the most popular category in both two maps.In a single match, M416 performs the best, with 10.18 people killed in the map ERANGEL, and 11.21 in MIRAMAR. However, among all the assault rifles, Groza and AUG perform the least best compared to others, resulting from their scarcity and difficulty to obtain.

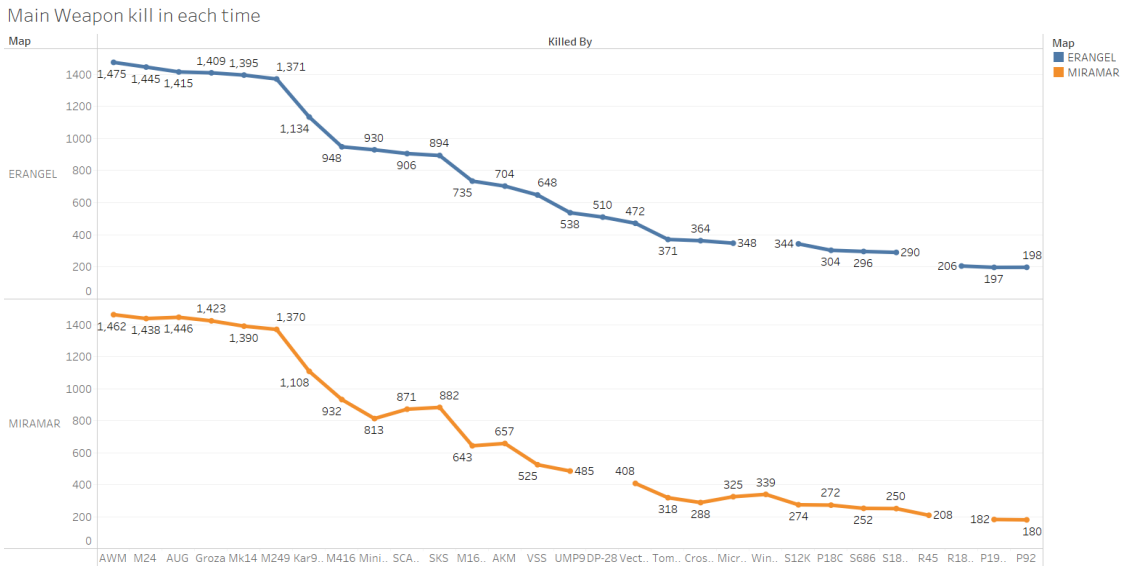
Chart(2) Average kills by weapon - cont'd / AUG, Groza obtaining time





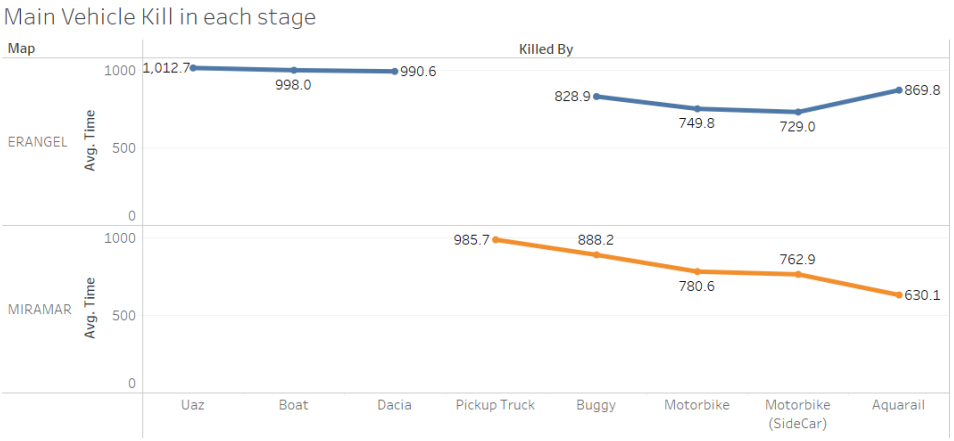
From Chart(2), we can find that AUG and GROZA are only available after 300 seconds from the game start. Because they are the airdrop guns (only very few people can obtain them from the limited airdrops in a single match), there are multiple separations both in the hologram of AUG and GROZA, meaning no player uses these two weapons in a certain period of time.

Chart(3) Main threatening weapon in each period



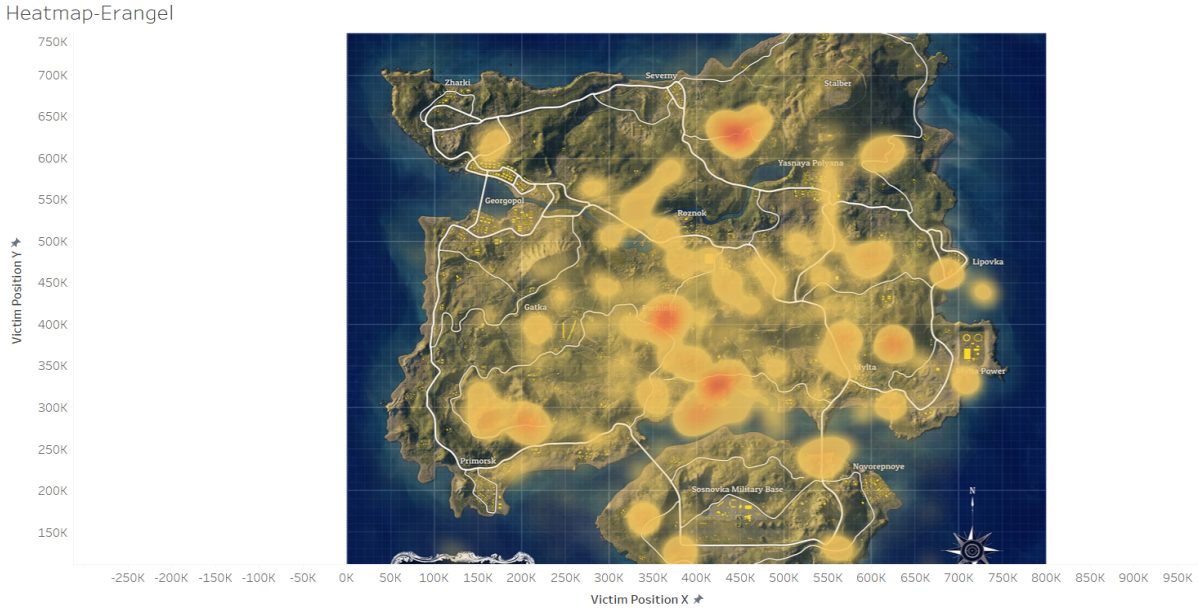
From Chart(3), It is obvious that the main threatening weapons in each period are identical in two maps. Undoubtedly, the top 6 main dominating weapons in the final stage of the game are all from the airdrops (see Appendix(1)), and the dominating weapon in the final stage is AWM. We can also find that pistols and shotguns are mainly used during the early period of the game, but as the match keeps continuing, players tend to switch their weapons to assault rifles and sniper rifles, and finally the airdrops weapons. The main difference between the two maps is the placement of SKS. SKS is the powerful sniper rifle which has the ability in killing enemies at long range, especially in the broad map as MIRAMAR.

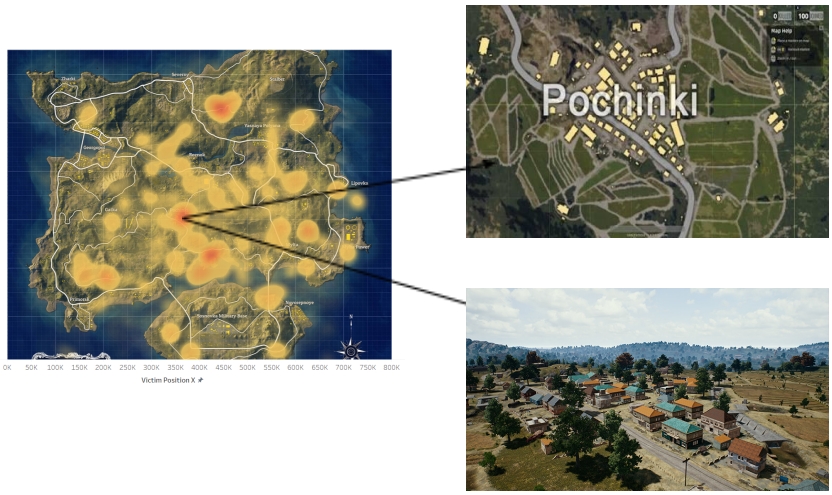
Chart (4) Main threatening vehicle in each period



From the Chart(4), we can clearly see that all vehicle kills start in the mid period to the final period in a match, the period when players must drive to the safe zone. The UAZ(a type of truck) is dominating the final round in the map ERANGEL, while the Pickup Truck is representative in the map MIRAMAR.

Chart(5-1-1) Victim Heatmap-ERANGEL

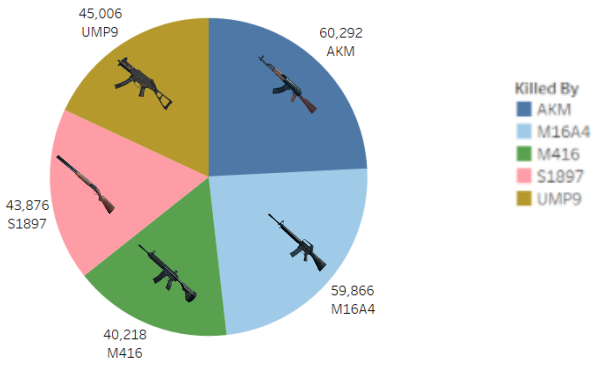




X axis:350928 ~ 379328
Y axis:386400 ~ 417020
(Total map area:80k x 80k)

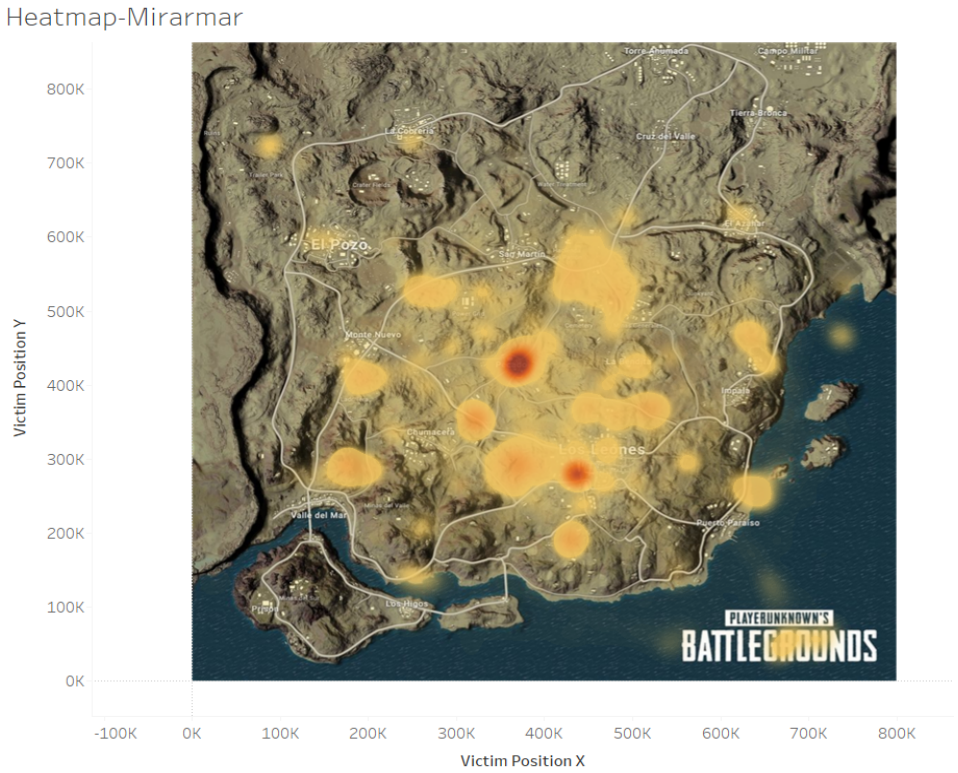
The heatmap shows the density of players who get killed in the match. From the Chart(5-1-1), we can find the location “Pochinki” is the most dangerous place on this map.

Chart(5-1-2) Top5 weapons in hot zone-ERANGEL



Chart(5-1-2) shows the weapons mainly used in the hot zone “Pochinki”. From the charts, the AKM has the largest percentage of top 5 weapons. Interesting fact is that other than assault rifles(AKM, M16A4, M416), the shotgun(S1897) and the submachine gun(UMP9) are both popular in this location as well. As the scenery displayed above, “Pochinki is the house congested area, where players may choose to commit kills with short range weapons like shotgun or submachine gun.

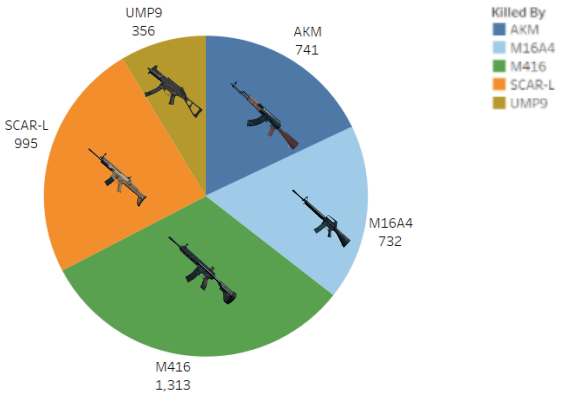
Chart(5-2-1) Victim Heatmap-MIRAMAR



X axis:357607 ~ 387289
Y axis:414248 ~ 442564
(Total map area:80k x 80k)

From the Chart(5-2-1), we can see that the location “Pecado” is the most dangerous place on this map.

Chart(5-2-2) Top5 weapons in hot zone-MIRAMAR



From Chart(5-2-2), we can find that almost all the most popular weapons in “Pecado” are assault rifles. Slightly different from “Pochinki” in the map ERANGEL, “Pecado” is the area with broader road and wider interval between houses, as the scenery image shown above. Players tend to choose mid range weapons like assault rifles to knock down their enemies, and both M416 and SCAR-L stand obviously the top 2 largest percentage of all top 5 weapons.

Summary

In this project, we demonstrate interesting facts through statistical analysis, aiming at teaching you how to play games scientifically. According to all the results above, we are able to generate the following conclusion:

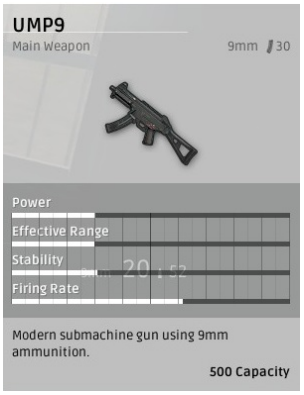
1. Assault rifle is the most popular weapon among all categories.
2. M416 shows the most balanced ability in both two maps, and it is also the ideal weapon in the most dangerous area where victim density is the highest.
3. Compared to the players in the map ERANGEL, those in MIRAMAR have a higher chance to kill their enemies with sniper rifles in the later stage of the game.
4. If players want to live longer, stay away from “Pochinki” and “Pecado”!
5. Be aware of the vehicle kills, especially after 600 seconds from the game start.

Statistics of top guns from three main categories:

King of the Assault Rifles
- M416

Merciless Assassin in the Final
- AWM

Close Range Specialist
-UMP9



The common characteristic: **Stable, Stable, and Stable!**

Appendix

Appendix(1)

Airdrops Guns:

AWM
M24
AUG
GROZA
MK14
M249

Appendix(2)

Code for data extraction and cleaning (In Linux environment):

Get data dimensions:

```
[sophia.y@ip-172-31-68-14 ~]$ wc -l kill_match_stats_final_4.csv
11640856 kill_match_stats_final_4.csv
```

```
[sophia.y@ip-172-31-68-14 ~]$ head -1 kill_match_stats_final_4.csv|tr "," "\n"|wc -l
12
```

Players are most frequently killed by:

```
[sophia.y@ip-172-31-68-14 ~]$ nano killed_by_map.py
[sophia.y@ip-172-31-68-14 ~]$ cat killed_by_map.py
#!/usr/bin/env python
import sys
for line in sys.stdin:
        line = line.strip()
        words = line.split(",")
        killed_by = words[0]
        if killed_by != "killed_by":
                print '%s' %killed_by
```

```
[sophia.y@ip-172-31-68-14 ~]$ nano killed_by_reducer.py
[sophia.y@ip-172-31-68-14 ~]$ cat killed_by_reducer.py
#!/usr/bin/env python
import sys
by = {}
for line in sys.stdin:
        cause = line.strip()
        try:
                by[cause] = by[cause] + 1
        except:
                by[cause] = 1
for v in by.keys():
        print '%s\t%i' %(v,by[v])
```

```
[sophia.y@ip-172-31-68-14 ~]$ chmod +x killed_by_map.py
[sophia.y@ip-172-31-68-14 ~]$ chmod +x killed_by_reducer.py
```

```
[sophia.y@ip-172-31-68-14 ~]$ nano killed_by_bash.sh
[sophia.y@ip-172-31-68-14 ~]$ cat killed_by_bash.sh
#!/bin/bash
hadoop jar /opt/cloudera/parcels/CDH-5.15.2-1.cdh5.15.2.p0.3/jars/hadoop-streaming-2.6.0-cdh5.15.2.jar \
        -Dmapred.reduce.tasks=1 \
        -input /user/sophia.y/kill_match_stats_final_4.csv \
        -output /user/sophia.y/killed_by_output \
        -file killed_by_map.py \
        -file killed_by_reducer.py \
        -mapper "python killed_by_map.py" \
        -reducer "python killed_by_reducer.py"
```

```
[sophia.y@ip-172-31-68-14 ~]$ hdfs dfs -ls killed_by_output
Found 2 items
-rw-r--r--   3 sophia.y sophia.y          0 2021-04-26 15:18 killed_by_output/_SUCCESS
-rw-r--r--   3 sophia.y sophia.y        812 2021-04-26 15:18 killed_by_output/part-00000
[sophia.y@ip-172-31-68-14 ~]$ hdfs dfs -cat killed_by_output/part-00000
death.ProjMolotov_DamageField_C 7532
Buggy   8360
Van     687
Motorbike       10012
SKS     269537
Micro UZI       232421
S686    284897
Mk14    14697
AKM     994727
Falling 126018
M249    25272
death.Buff_FireDOT_C    2961
Pan     16155
M16A4   1077583
P1911   131165
P92     129467
Dacia   16529
Punch   306558
UMP9    641447
death.ProjMolotov_C     277
S12K    258959
AWM     25616
Drown   48194
death.RedZoneBomb_C     143
Bluezone        571403
Aquarail        68
Pickup Truck    9508
Machete 5478
Crossbow        18689
RedZone 14675
DP-28   32110
AUG     10007
```

```
[sophia.y@ip-172-31-68-14 ~]$ hdfs dfs -cat killed_by_output/part-00000|sort -rnk 2|head -5
M416    1331669
SCAR-L  1088986
M16A4   1077583
AKM     994727
UMP9    641447
```

Code for analysis (In Impala environment):

Create Table:

```
CREATE TABLE kill_match_2 ( killed_by STRING, killer_name STRING, killer_placement STRING, killer_position_x STRING, killer_position_y STRING,
map_name STRING, match_id STRING, time int,
victim_name string, victim_placement STRING, victim_position_x STRING, victim_position_y STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE TBLPROPERTIES("skip.header.line.count"="1")
```

```
Load DATA INPATH '/user/d.yuxuan/kill_match_stats_final_4.csv' INTO TABLE `kill_match_2`
```

Descriptive Analysis:

```
select killer_placement, count(killer_placement) as killer_cnt
from kill_match_2
group by killer_placement
```

```
select victim_placement, count(victim_placement) as victim_cnt
from kill_match_2
group by victim_placement
```

Count popular weapons in two maps:

```
select killed_by, count(killed_by) from kill_match_2 where killer_name is not null and map_name = "MIRAMAR" group by killed_by
order by count(killed_by) desc limit 20
```

```
select killed_by, count(killed_by) from kill_match_2 where killer_name is not null and map_name = "ERANGEL" group by killed_by
order by count(killed_by) desc limit 20
```

```
select killed_by, count(killed_by) from kill_match_2
where map_name = "MIRAMAR" and cast(killer_position_x as decimal) > 357607 and cast(killer_position_x as decimal) < 387289 and cast(killer_position_
as decimal) <414248 and cast(killer_position_y as decimal) > 332564
group by killed_by order by count(killed_by) desc limit 20
```

```
select killed_by, count(killed_by) from kill_match_2
where map_name = "ERANGEL" and cast(killer_position_x as decimal) > 350938 and cast(killer_position_x as decimal) < 379328 and cast(killer_position_y
as decimal) > 386400 and cast(killer_position_y as decimal) < 417020
group by killed_by order by count(killed_by) desc limit 20
```

Count main weapons in high density area:

```
select killed_by, count(killed_by) from kill_match_2
where map_name = "MIRAMAR" and cast(killer_position_x as decimal) > 357607 and cast(killer_position_x as decimal) < 387289 and cast(killer_position_
as decimal) <414248 and cast(killer_position_y as decimal) > 332564
group by killed_by order by count(killed_by) desc limit 20
```

```
select killed_by, count(killed_by) from kill_match_2
where map_name = "ERANGEL" and cast(killer_position_x as decimal) > 350938 and cast(killer_position_x as decimal) < 379328 and cast(killer_position_y
as decimal) > 386400 and cast(killer_position_y as decimal) < 417020
group by killed_by order by count(killed_by) desc limit 20
```

Data Visualization using Tableau:

Heatmap

Top 5 weapons in each map

All Weapons used in hot zone

Average Kills by weapons