PLAYERUNKNOWN'S
PUBG ANALYSIS

THE UNIVERSITY OF CHICAGO

**MScA Big Data Platforms**
Sheena Huang, Kai Li, Jessie Hung
March 2020

# Business Problem

Quick Overview–

Battlegrounds is a player versus player shooter game in which up to one hundred players fight in a battle royale, a type of large-scale last man standing deathmatch where players fight to remain the last alive. Players can choose to enter the match solo, duo, or with a small team of up to four people.

Comprehensive statistical analysis and models for  company and players:

- **Give company a better insight of player behaviors to make future improvement on game design.**

- **Cluster and identify outliers in order to ban cheaters.**

- **Provide key insights for players to strategize to win the game.**

# Data and Platform

# Source

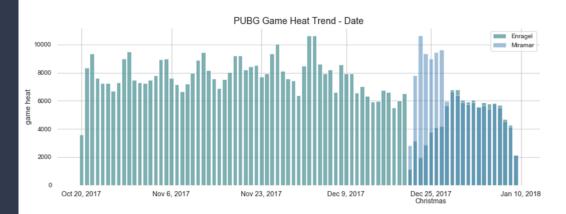- 19GB PUBG Match dataset & Death dataset (Oct.2017-Jan.2018)

- 10 CSV files in total

- 65 million+ rows in each dataset

- Match: 15 features

  Date, game size, match id, match type, party size, damage, assist, walk/ride distance, kills, survival time, player name, team id and team placement

- Death: 12 features

  Match id, map, time, killer name, killer placement, killer position, victim name, victim placement, victim position and weapon
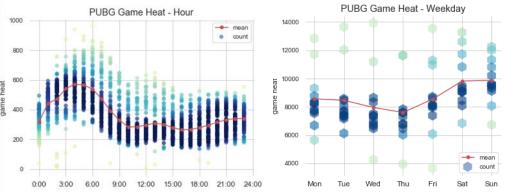
- Use Google Cloud Platform for storage and computing

# Data Visualization & Insights

# Game Time!

- A huge portion of players are more active at midnight and weekends

- New map launched on 2017-12-21. Great enthusiasm for new map but quickly gone
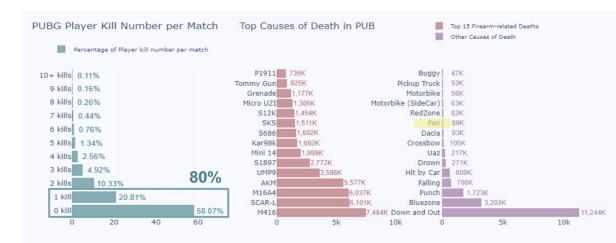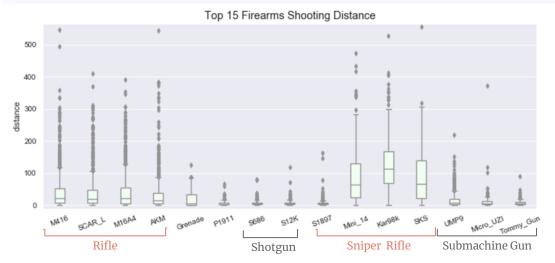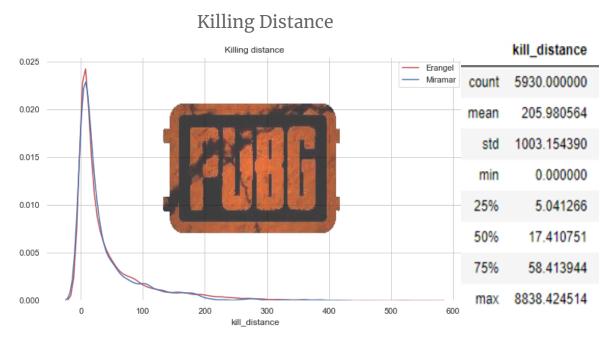
# Kill Stats

- 80% of players have no more than 1 kill

- 80% deaths are caused by firearm attacks. A huge proportion (50%) of players are killed by listed rifles and sniper rifles

- Blue zone/ red zone only eliminate 5% of players, the best tactic early on is to avoid everyone and find the right gears

# Killing Distance

### Killing Distance



| | kill_distance |
|---|---|
| count | 5930.000000 |
| mean | 205.980564 |
| std | 1003.154390 |
| min | 0.000000 |
| 25% | 5.041266 |
| 50% | 17.410751 |
| 75% | 58.413944 |
| max | 8838.424514 |

- 1/10000 sampling to draw distribution

- Units in meter

- Median is only ~17 meters

- No long distance kills

  (sidenote: longest shoot range 650m)

# Team Statistics

**party size=1**

| summary | avg_survive_time | total_kill | total_assists | total_dbno | avg_ride_dist | avg_walk_dist |
|---|---|---|---|---|---|---|
| count | 14039649 | 14039649 | 14039649 | 14039649 | 14039649 | 14039649 |
| mean | 145792 | 0 | 0 | 0 | 882 | 1164 |
| stddev | 96089538 | 1 | 0 | 0 | 1854 | 5308 |
| min | 0 | 0 | 0 | 0 | 0 | 0 |
| 25% | 207 | 0 | 0 | 0 | 0 | 151 |
| 50% | 578 | 0 | 0 | 0 | 0 | 688 |
| 75% | 1178 | 1 | 0 | 0 | 1105 | 1819 |
| max | 63647875100 | 94 | 6 | 0 | 442536 | 1188375 |

**party size=4**

| summary | avg_survive_time | total_kill | total_assists | total_dbno | avg_ride_dist | avg_walk_dist |
|---|---|---|---|---|---|---|
| count | 9192801 | 9192801 | 9192801 | 9192801 | 9192801 | 9192801 |
| mean | 19280 | 3 | 1 | 2 | 1254 | 1340 |
| stddev | 26530186 | 3 | 1 | 2 | 1811 | 2234 |
| min | 0 | 0 | 0 | 0 | 0 | 0 |
| 25% | 330 | 0 | 0 | 1 | 0 | 326 |
| 50% | 724 | 2 | 0 | 2 | 216 | 1051 |
| 75% | 1259 | 4 | 2 | 4 | 2112 | 2102 |
| max | 63648215000 | 71 | 28 | 152 | 163364 | 1006273 |

**party size=2**

| summary | avg_survive_time | total_kill | total_assists | total_dbno | avg_ride_dist | avg_walk_dist |
|---|---|---|---|---|---|---|
| count | 10981317 | 10981317 | 10981317 | 10981317 | 10981317 | 10981317 |
| mean | 12354 | 1 | 0 | 1 | 1131 | 1225 |
| stddev | 19206974 | 2 | 0 | 1 | 1869 | 3054 |
| min | 0 | 0 | 0 | 0 | 0 | 0 |
| 25% | 255 | 0 | 0 | 0 | 0 | 222 |
| 50% | 628 | 1 | 0 | 1 | 0 | 831 |
| 75% | 1212 | 2 | 1 | 2 | 1833 | 1940 |
| max | 31824108320 | 78 | 18 | 42 | 254516 | 897151 |

- Max: hackers (detect hackers!), but more likely bad data

- Squad usually survives longer

- Not many kills

- No assists 😢

# Survival Time
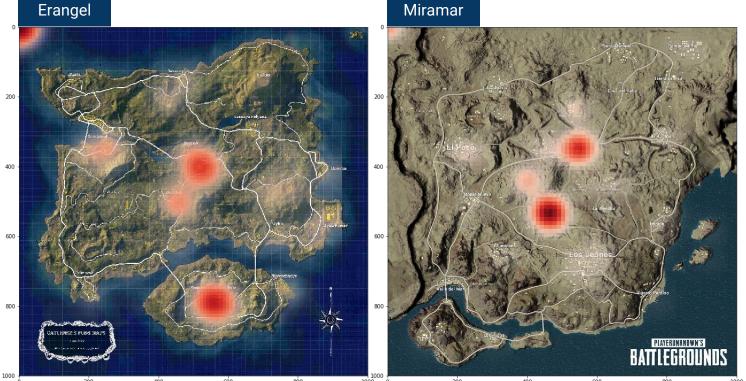
- Solo game is fiercer at the early game than duo and squad

- Peaks corresponding to the start of 2nd, 4th, and final Blue Zone (14 minutes, 25 minutes and 30 minutes)

- Wanna better game experience?
  Don't jump to those places....



Survival Time Distribution by Party Size

# Death HeatMap(<4min)

Death Heat Map within 4 min after skydiving

Erangel

Miramar
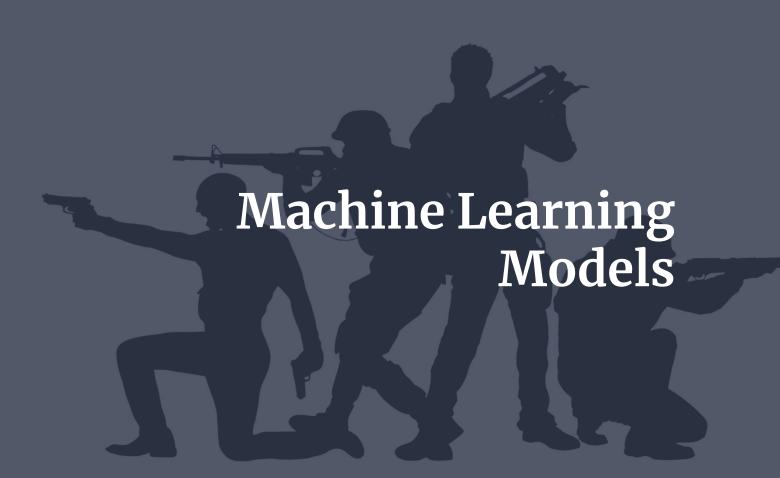
# The Final Circle – Safe Zone

- Survival heatmap after the last blue zone starts shrinking

- A large portion on the map is never in the final circle

- A few land is more contested for final circle, primarily near very open areas and road intersections

Erangel

Miramar

# Machine Learning Models

# Team Ranking Prediction by Performance
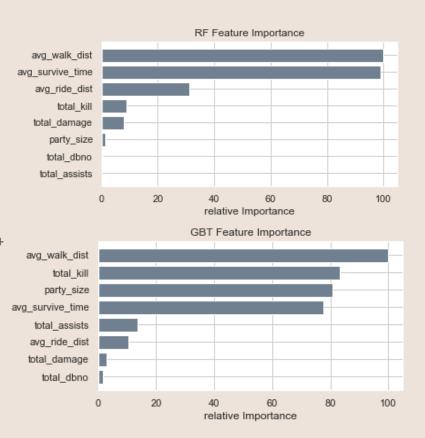
- Target Variable:
  - Relative Ranking : Team placement / Game size
- Features:
  - Party Size
  - Team average survival time
  - Total kills
  - Total assists
  - Total DBNO(down but not out)
  - Total Damage
  - Average walk distance
  - Average ride distance

- Models:
  - Linear Regression
  - Random Forest
  - Gradient Boosting Tree(GBT)
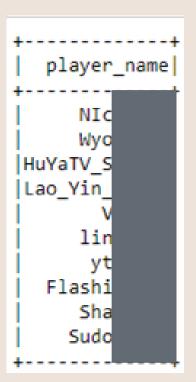
- Linear Regression:
  - Linear Regression R2 on training data: 0.454778
  - Linear Regression R2 on test data = 0.458131

- Random Forest Regressor:
  - 20 trees
  - Random Forest R2 on test data = 0.906374
  - Huge improve
  - Average survival time and average walk distance are the most important features
- Gradient Boosting Trees(GBT):
  - Maxiter – 5, num of trees – 5
  - GBT R2 on test data = 0.917593
  - Different feature importance



RF Feature Importance



GBT Feature Importance

# Hacker Detection – Clustering

- Use K-means to clustering players and try to identify hackers
- Find hackers have abnormal value in all three features: player_kills, player_dist_ride, and player_dist_walk
- Identified:
  - 574 players out of 7.6 million players
- Limitation: unable to distinguish with real talented players and more information required

# Network Graph

```
+-----------+------+
|        id|degree|
+-----------+------+
|feitengdedan|  5460|
|  hzxiaobin|  5346|
|  Slh_Bunny|  5314|
|     JZalan|  5132|
|  yangyl123|  4210|
|   Sterealz|  4206|
|  jimmy0572|  4124|
|   yirui887|  4106|
|BlueChick666|  3966|
|Matthew_wang|  3940|
+-----------+------+
```

- Identify Important Players:
  - Degree for each player
  - Send queries to those influencers in order to improve the game experience
- Friends / Teammates Recommendation:
  - Motif subgraph finding
  - Providing recommendation for players to improve game experience and game-player cohesion

```
+-----------------------+-----------------------+------------------------+
|a                      |b                      |c                       |
+-----------------------+-----------------------+------------------------+
|[Bizzy-Bin, LeLocle0512]|[LeLocle0512, LeLocle0202]|[LeLocle0202, Cannvaro]|
|[SMer_, babyt2]        |[babyt2, 333yinuo]     |[333yinuo, babyt2]      |
|[saoqi-NQR, DouYu_GuLang]|[DouYu_GuLang, CNJianChen]|[CNJianChen, fdgerggs]|
+-----------------------+-----------------------+------------------------+
```

# Recommendation

Tips for Players:

- If you enjoy late-game combat, try to avoid those areas while skydiving
- If you exciting with the early-game fight, those areas are top choices
- Try to enter the next safe zone earlier to avoid the high death chances at the end each safe zone shrinkage

Insights for PUBG Corp.:

- Better algorithms to detect hackers in different perspectives such as aim hack, speed hack, and super sniper
- Could recommend possible known players in the squad mode based on the network graph
- Use triangle counting find influencers and send questionnaires to improve the game experience

# Future Work

- Adding more features in the clustering model to detect hackers (higher dimensions)

- Get internal data such as players' report of hackers to improve the accuracy of hacker detection

- Larger memory for computing motif subgraph finding and providing recommendations for players while matching

- Develop a model to predict waiting time between two matches

WINNER WINNER
CHICKEN DINNER

Questions ?