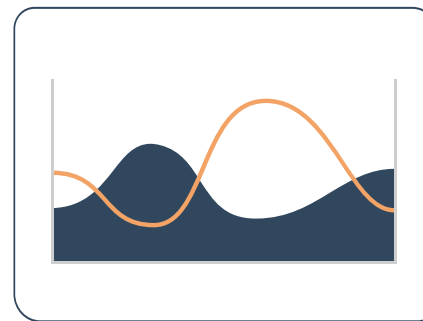
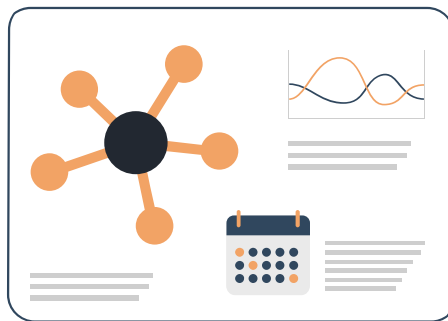
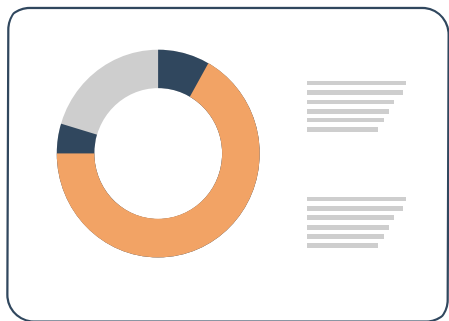


# Báo cáo đồ án

## Lập trình cho Khoa học dữ liệu



## Quy trình thực hiện



01

Thu thập dữ liệu

02

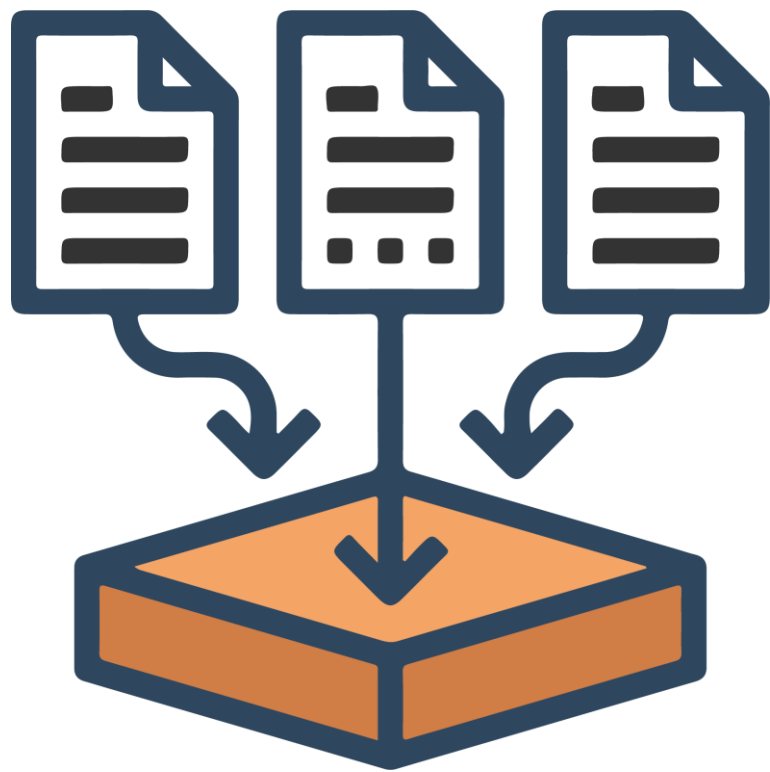
Tiền xử lí dữ liệu

03

Khám phá dữ liệu

04

Trả lời các câu hỏi



Thu thập  
dữ liệu

**Battle Rotal  
TPP/FPP**

100 Players

**PUBG**

**Người cuối  
cùng trên đảo**

**Bluezone**





Tập dữ liệu  
**Deaths**

- 5 file csv, mỗi file **13 triệu dòng**
- Mỗi dòng là một sự kiện **người chơi tử trận** trong 1 trận đấu

Tập dữ liệu  
**Aggregate**

- Thông tin tổng hợp về người chơi và các trận đấu
- Số lượng kill, sát thương,...
- Ngày diễn ra trận đấu, chế độ,...

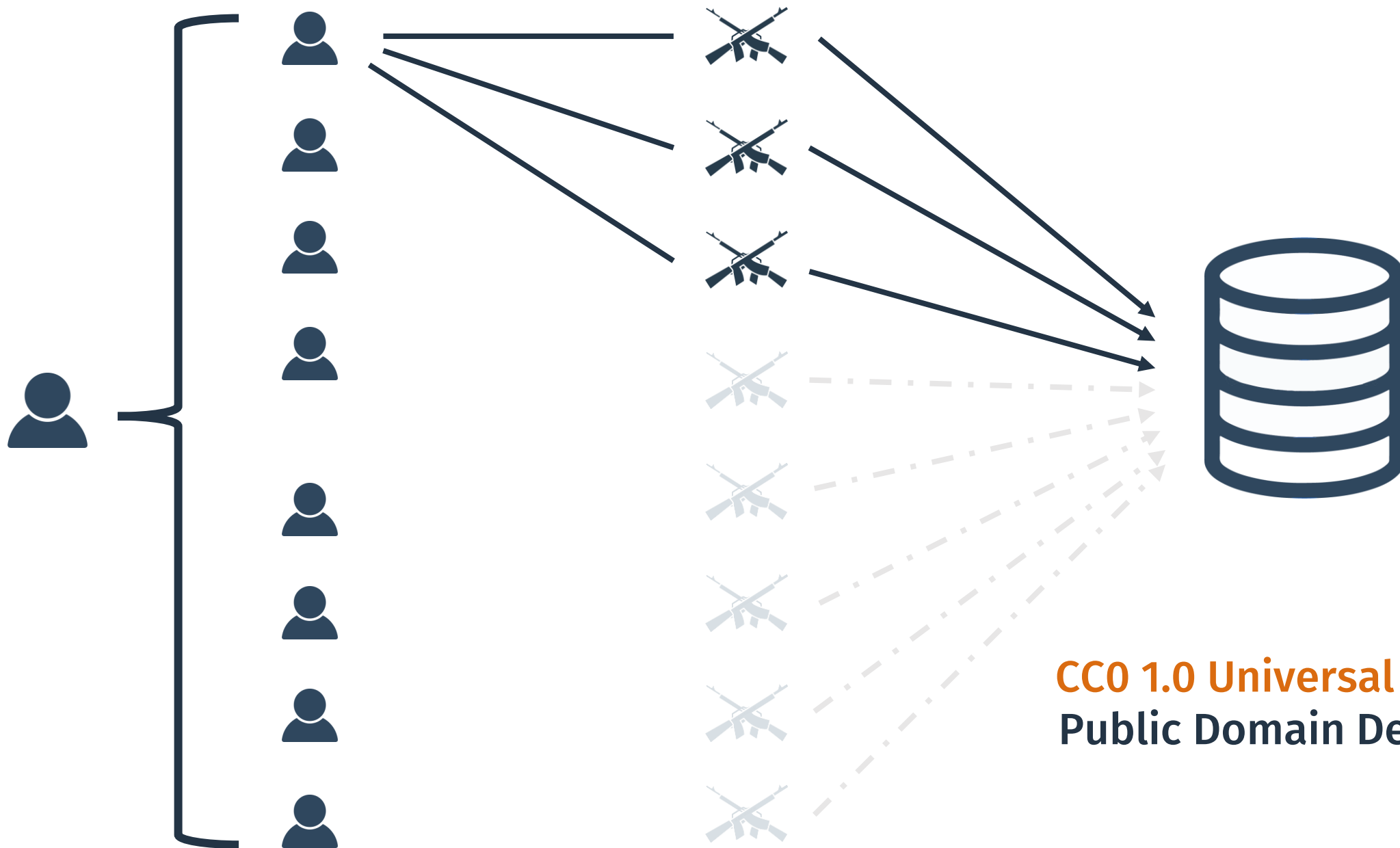


Tập dữ liệu  
**Deaths**

- 1 file csv **13 triệu dòng**
- Mỗi dòng là một sự kiện **người chơi tử trận** trong 1 trận đấu

Tập dữ liệu  
Aggregate

- Thông tin tổng hợp về người chơi và các trận đấu
- Số lượng kill, sát thương,...
- Ngày diễn ra trận đấu, chế độ,...



**CC0 1.0 Universal (CC0 1.0)**  
**Public Domain Dedication**



# Tiền xử lí dữ liệu



# Mô tả dữ liệu

- **killed\_by**: Nguyên nhân gây ra kill.
- **killer\_name**: tên player gây ra kill
- **killer\_placement**: thứ hạng của killer
- **killer\_position\_x**: tọa độ x của killer tại thời điểm gây ra kill.
- **killer\_position\_y**: tọa độ y của killer tại thời điểm gây ra kill.
- **map**: tên map của trận đấu.
- **match\_id**: id của trận đấu.
- **time**: thời điểm kill xảy ra tính từ lúc bắt đầu trận đấu.
- **victim\_name**: tên player tử trận
- **victim\_placement**: thứ hạng của victim.
- **victim\_position\_x**: tọa độ x của victim tại thời điểm kill xảy ra.
- **victim\_position\_y**: tọa độ y của victim tại thời điểm kill xảy ra.

killed\_by  
killer\_name  
killer\_placement  
killer\_position\_x  
killer\_position\_y  
map  
match\_id  
time  
victim\_name  
victim\_placement  
victim\_position\_x  
victim\_position\_y



killed\_by  
killer\_name  
killer\_placement  
killer\_position\_x  
killer\_position\_y  
map  
match\_id  
time  
victim\_name  
victim\_placement  
victim\_position\_x  
victim\_position\_y



killed\_by  
killer\_name  
killer\_placement  
killer\_position\_x  
killer\_position\_y  
map  
match\_id  
time  
victim\_name  
victim\_placement  
victim\_position\_x  
victim\_position\_y



killed\_by  
kx  
ky  
time  
vx  
vy

# Làm sạch dữ liệu

## Dữ liệu khuyết

kx và ky: 741,597 dòng

7% dữ liệu



Gán kx, ky bằng vx, vy

Nguyên nhân	Số lượng
Bluezone	468994
Down and Out	141913
Falling	61571
Drown	41150
RedZone	13683
Uaz	4451
Dacia	3002
Buggy	2091
Hit by Car	1719

# Làm sạch dữ liệu

Giá trị lặp

kx, ky, vx, vy: 257,932 dòng

2.5% dữ liệu



Xóa bỏ

kx	ky	vx	vy	Số lượng
0.0	0.0	0.0	0.0	257127
399045.9	300804.1	0.0	0.0	2
482652.2	446492.6	0.0	0.0	2
446353.9	629764.4	446300.3	629816.6	1
444775.8	623105.9	444581.1	623158.9	1
				...

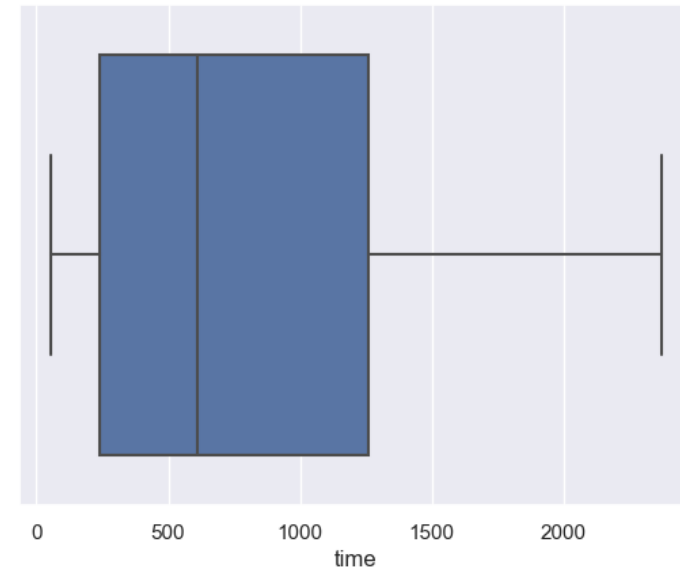
# Làm sạch dữ liệu - Giá trị ngoại lai

## Cột `killed_by`

<code>death.PlayerMale_A_C</code>	1
<code>Aquarail</code>	21
<code>death.RedZoneBomb_C</code>	90
<code>death.ProjMolotov_C</code>	251
<code>Boat</code>	1742
<code>death.Buff_FireDOT_C</code>	2711
<code>Sickle</code>	4338
<code>Crowbar</code>	4704

Xóa bỏ `death.PlayerMale_A_C` và  
`death.Buff_FireDOT_C`

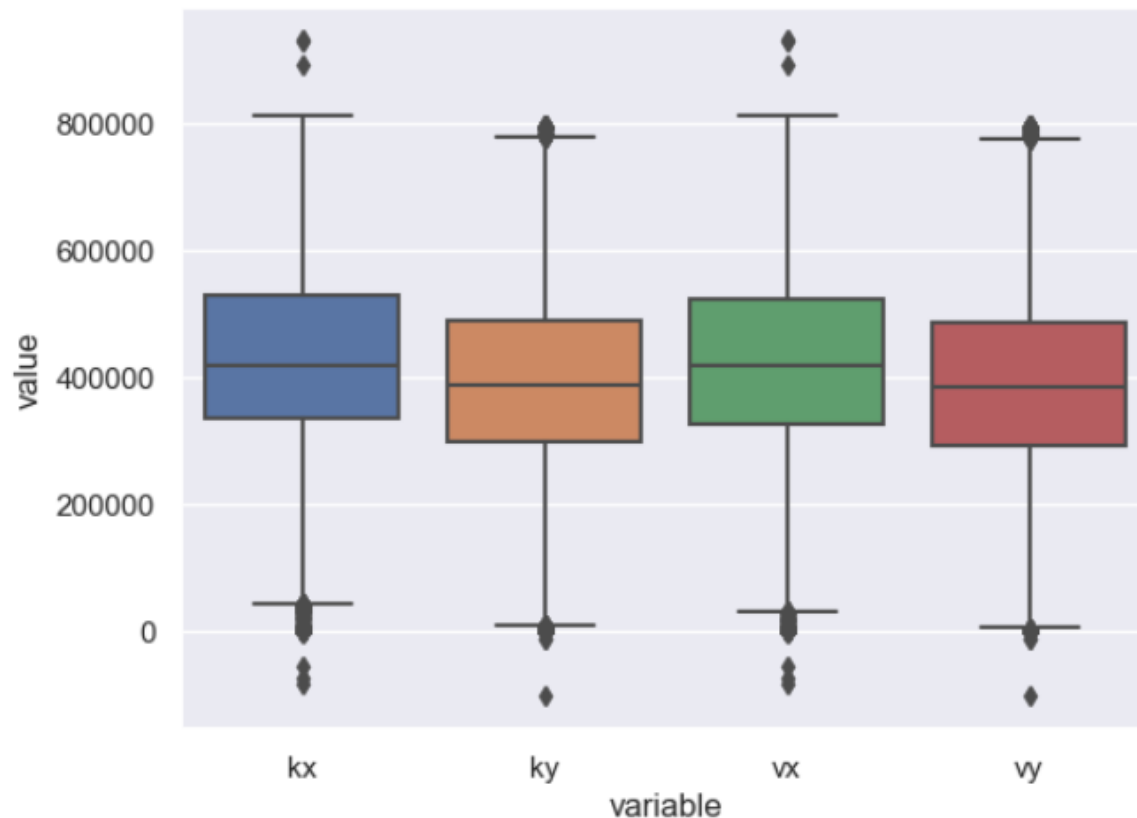
## Cột `time`



Giữ nguyên

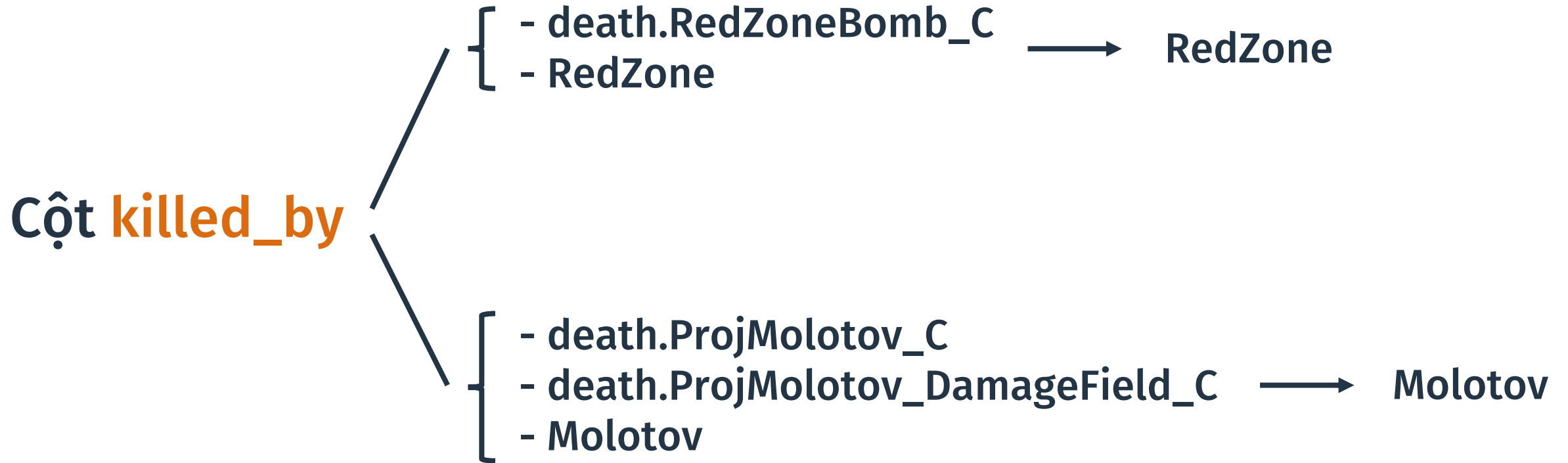
# Làm sạch dữ liệu - Giá trị ngoại lai

Các cột **kx, ky, vx, vy**



Xóa bỏ các dòng lớn hơn  
800,000 hoặc nhỏ hơn 0

# Đánh giá chất lượng dữ liệu



Các cột  
**kx, ky, vx, vy**

[0 -> 800,000] ~ [0 -> 8,000]

# Thêm một số cột

Cột **dis**

Khoảng cách từ  
killer tới victim

$$\text{dis} = \sqrt{(\text{kx} - \text{vx})^2 + (\text{ky} - \text{vy})^2}$$

Cột **phase**

time →

Phase	Bắt đầu	Kết thúc
1	121	720
2	721	1060
3	1061	1300
4	1301	1480
5	1481	1650
6	1651	1760
7	1761	1880
8	1881	1970
9	1971	2150

→ phase



# Thêm một số cột

**Cột type**

**killed\_by**

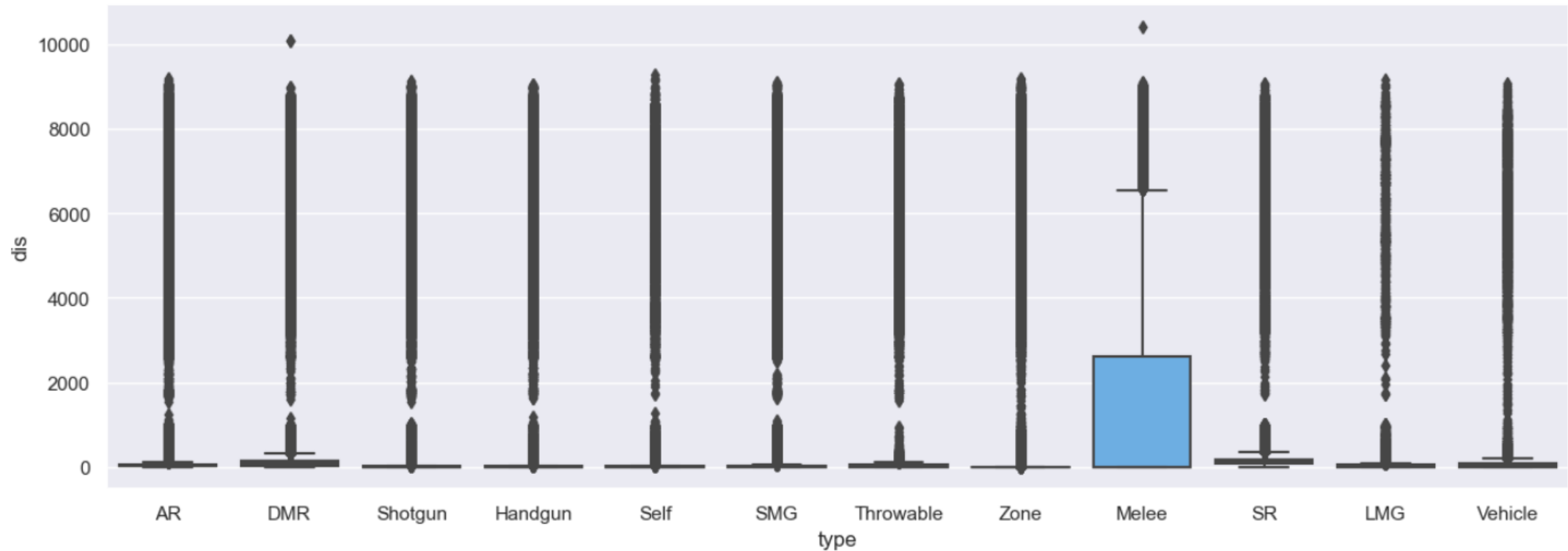
**Phân loại  
nguyên nhân  
theo nhóm**

**type**

AR	SMG	DMR
SR	LMG	Shotgun
Zone	Vehicle	Self
Throwable	Melee	Handgun

# Tiếp tục làm sạch dữ liệu

## Cột **dis**



# Tiếp tục làm sạch dữ liệu

Cột **dis**

dis  
type

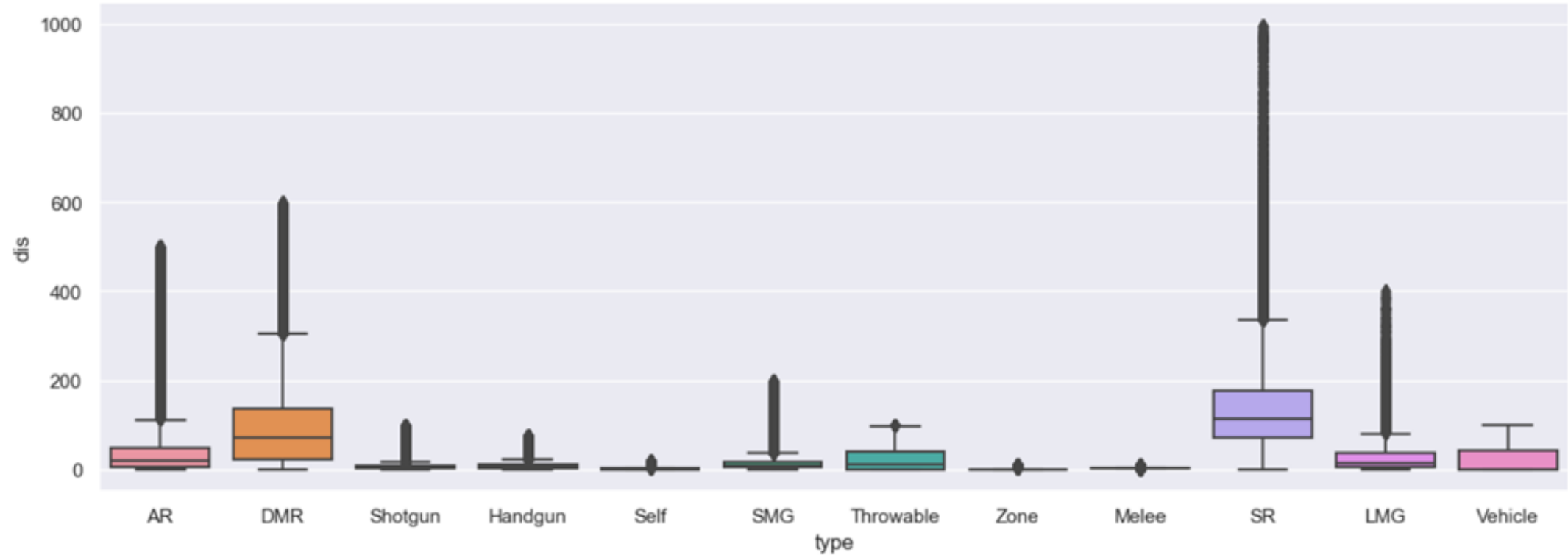


type	Giá trị dis tối đa
Self	20
AR	500
DMR	600
SR	1000
Shotgun	100
Handgun	75
SMG	200
LMG	400
Throwable	100
Zone	10
Melee	10
Vehicle	100



**dis**

# Tiếp tục làm sạch dữ liệu



# Tiếp tục làm sạch dữ liệu

Cột **time**  
và **phase**

Xóa các giá trị **nan** trong cột phase

Phase 9 kết thúc ở giây **2150**



Xóa các giá trị **lớn hơn 2160** trong cột time

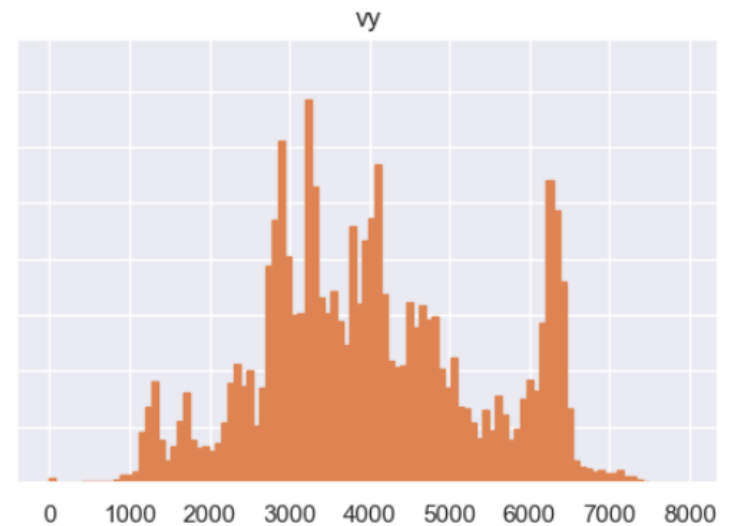
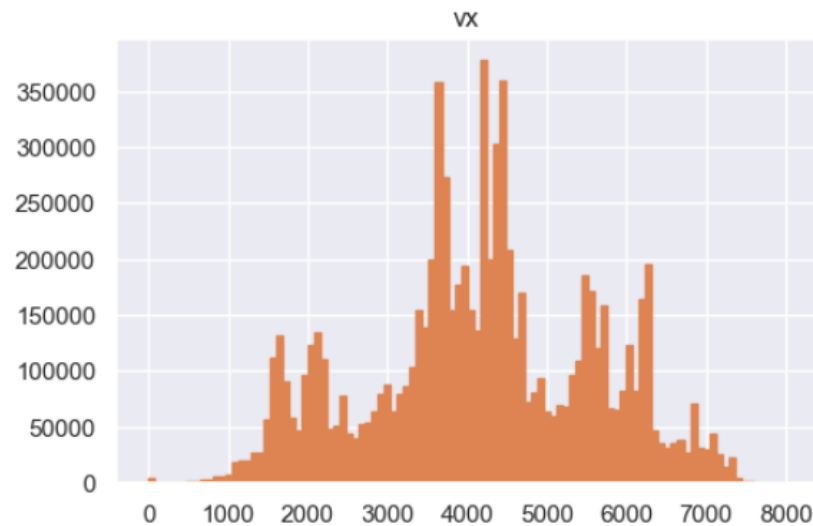
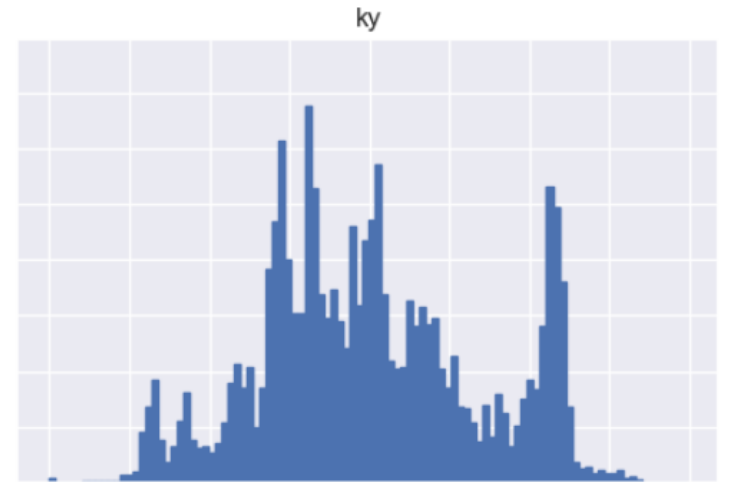
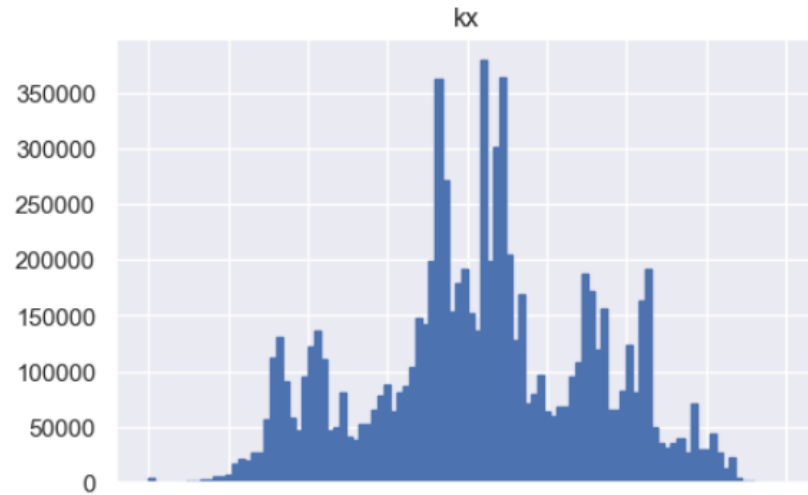


# Khám phá dữ liệu

Cột	Loại dữ liệu	Kiểu dữ liệu trong file
killed_by	nominal	string
type	nominal	string
kx	numerical	int
ky	numerical	int
vx	numerical	int
vy	numerical	int
dis	numerical	int
time	thời điểm	int
phase	thời điểm	int

# Phân bố trong từng cột

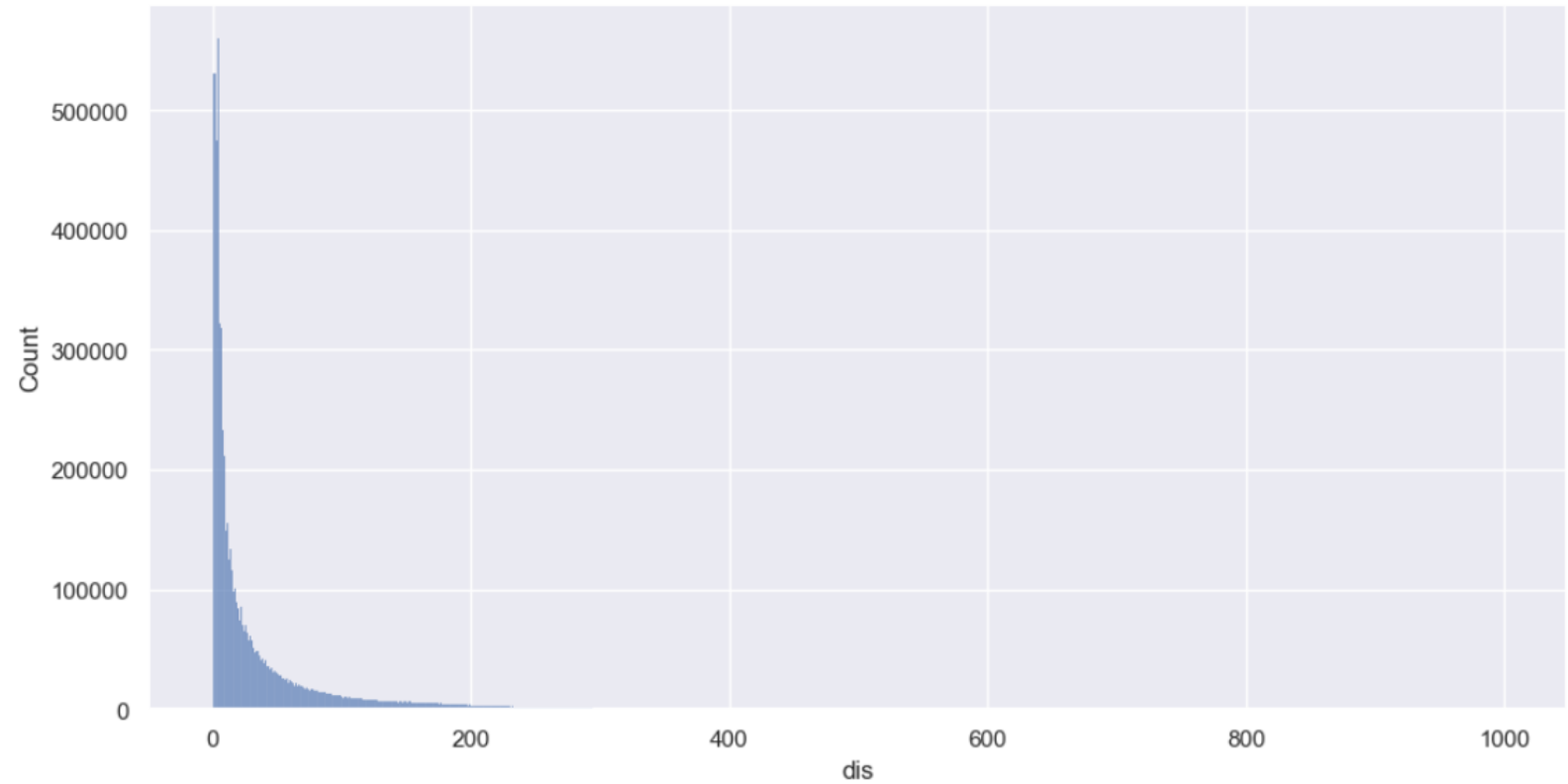
Các cột  
 $kx$ ,  $ky$ ,  $vx$ ,  $vy$





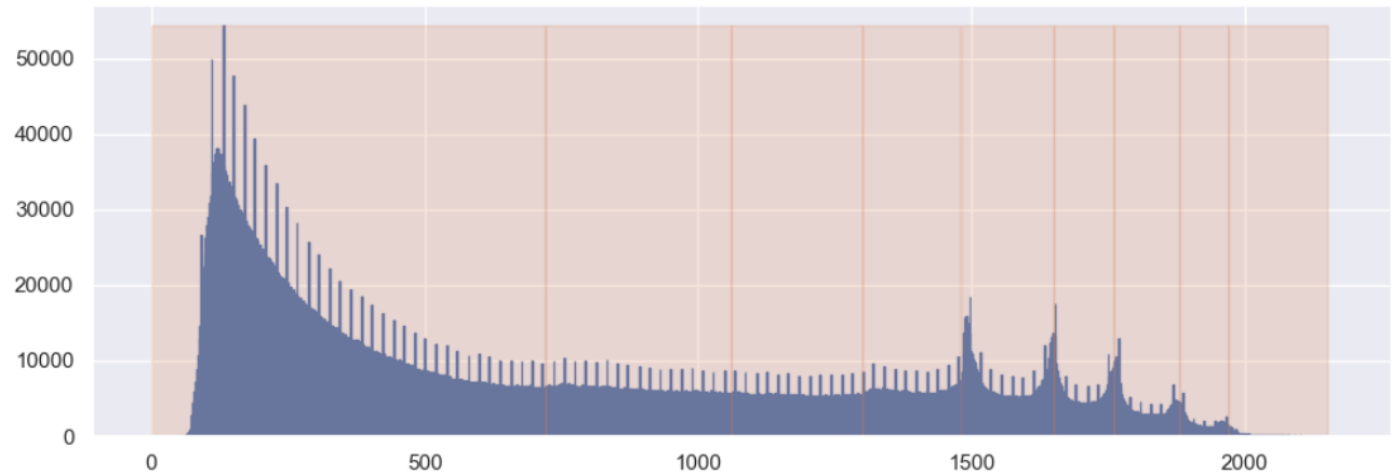
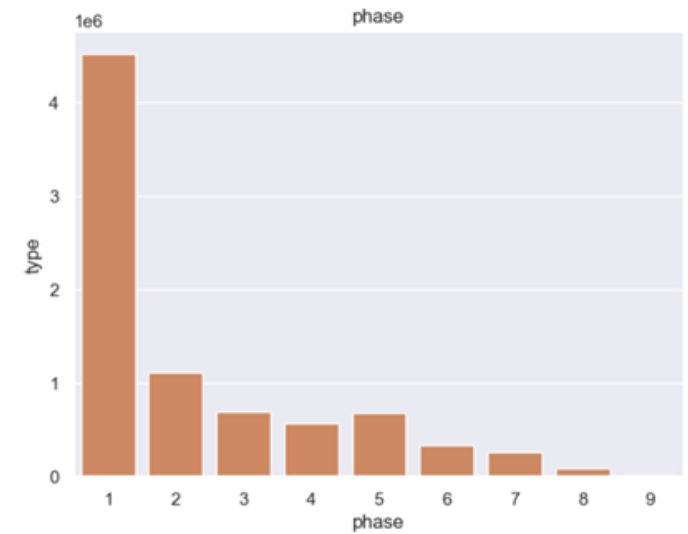
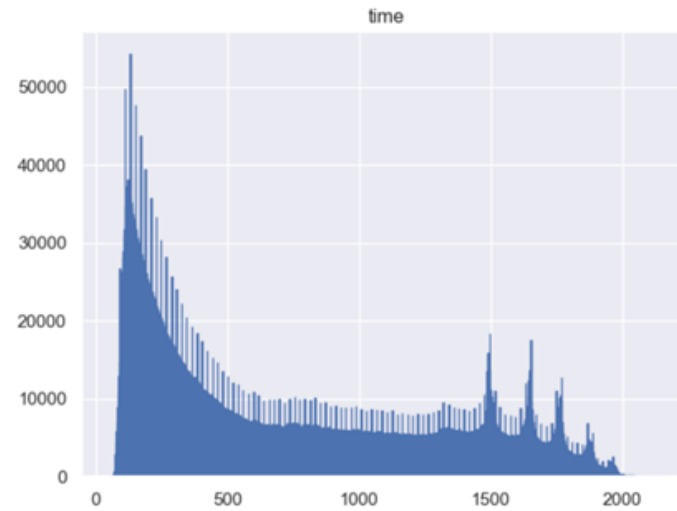
# Phân bố trong từng cột

Cột **dis**



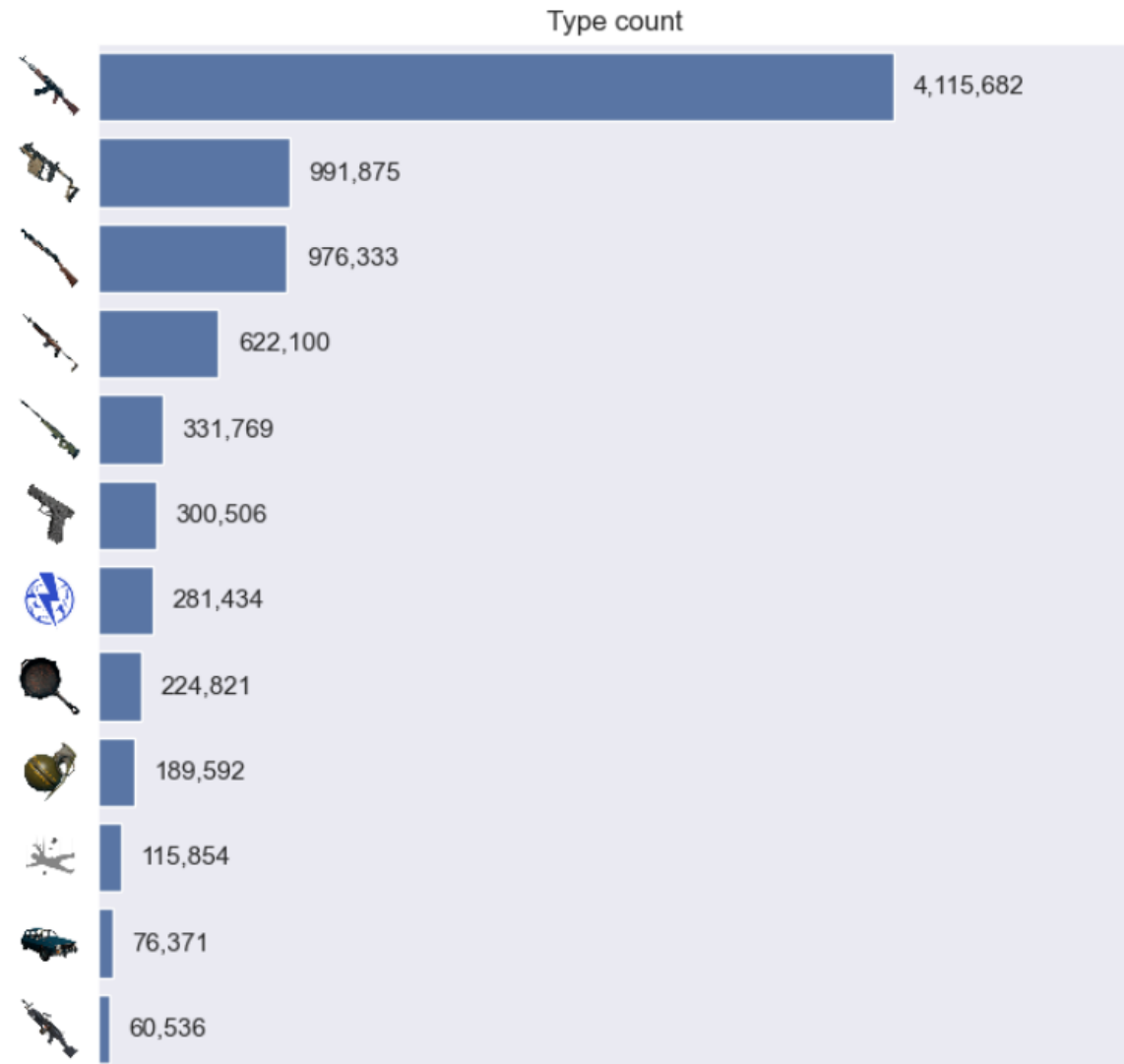
# Phân bố trong từng cột

Cột **time**  
và **phase**



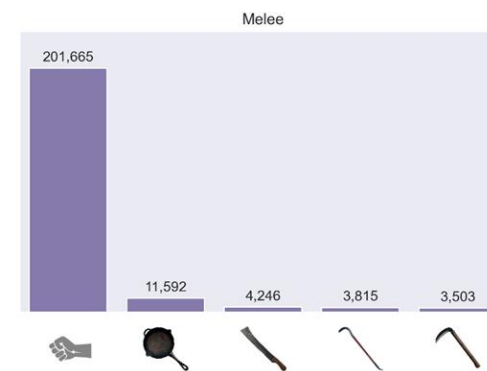
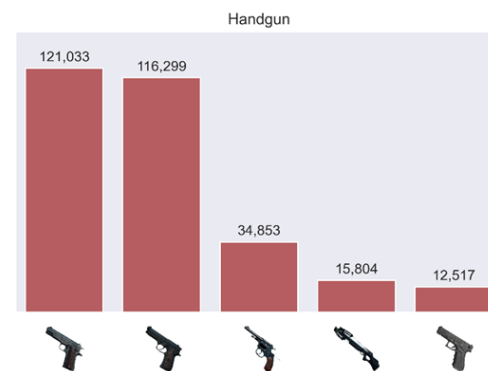
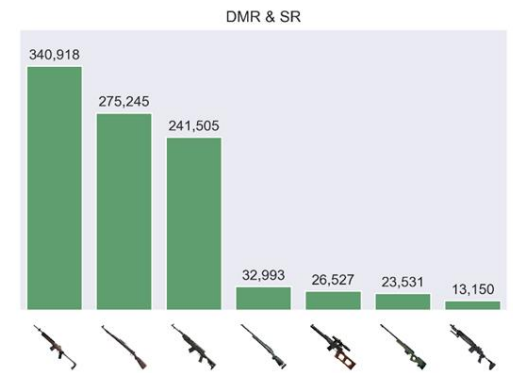
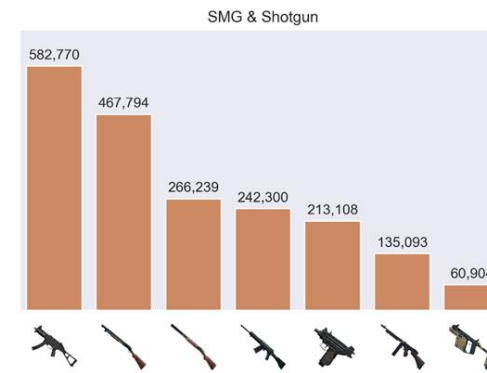
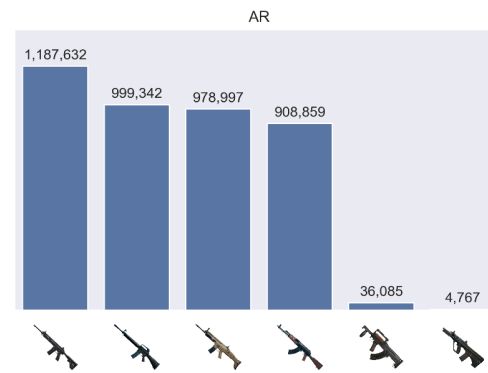
# Phân bố trong từng cột

Cột type

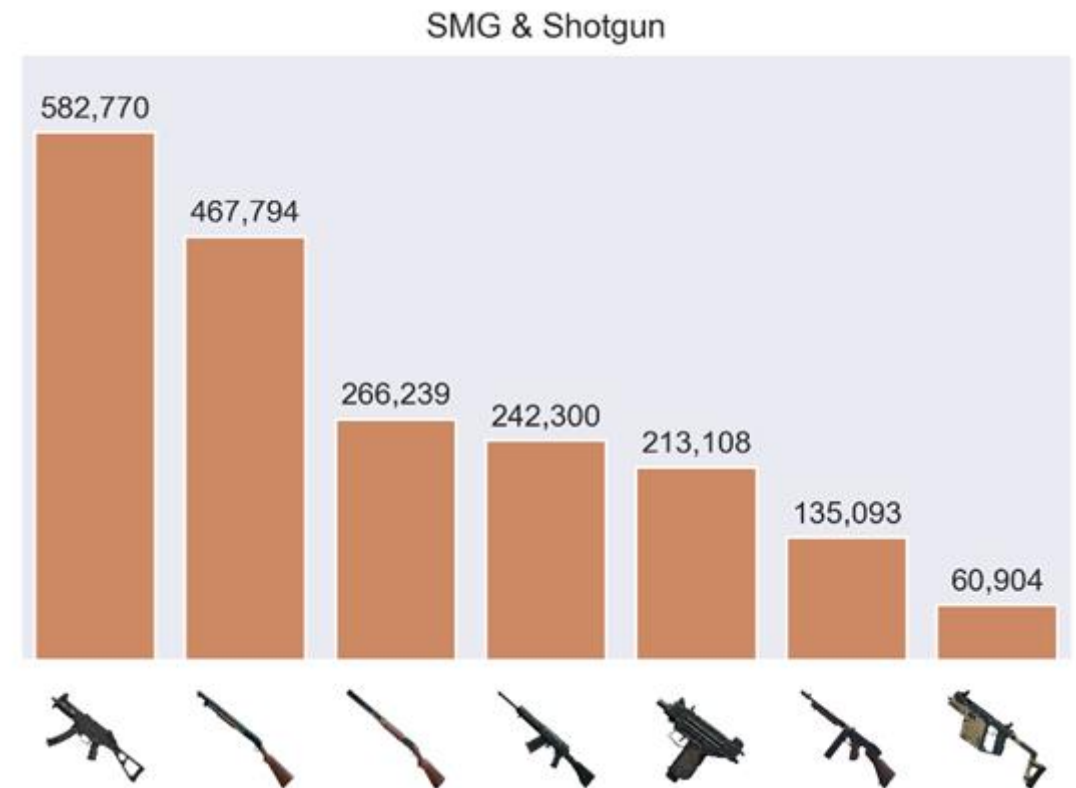
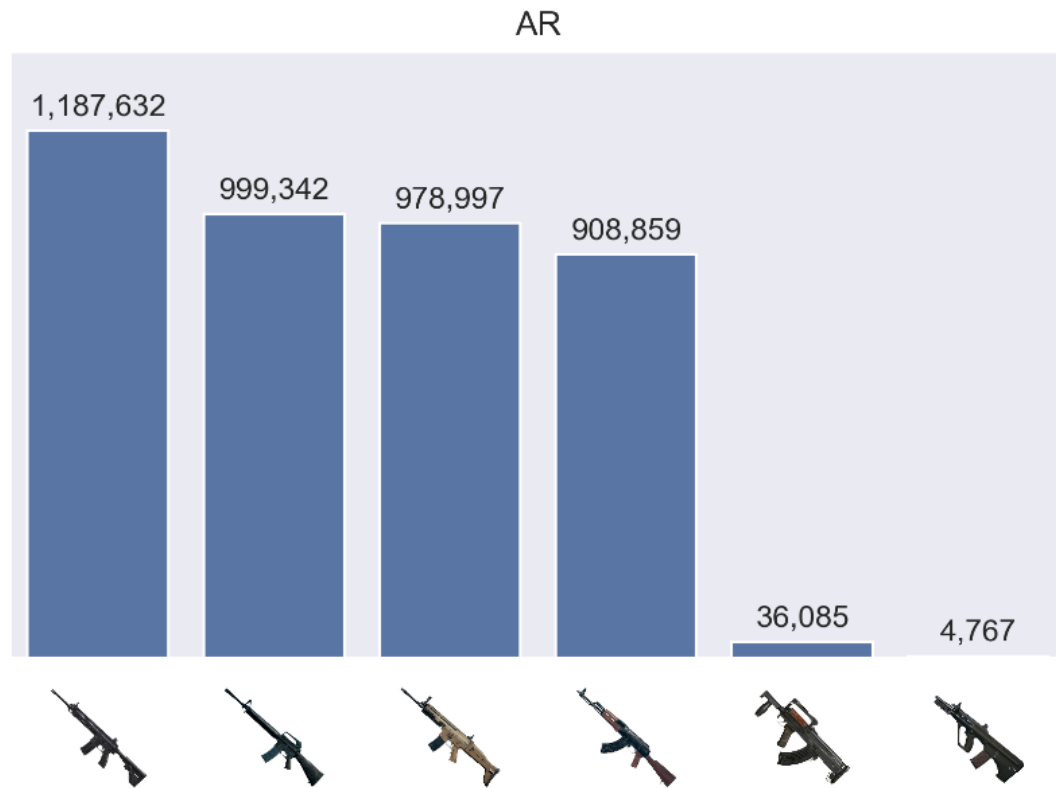


# Phân bố giữa các cột

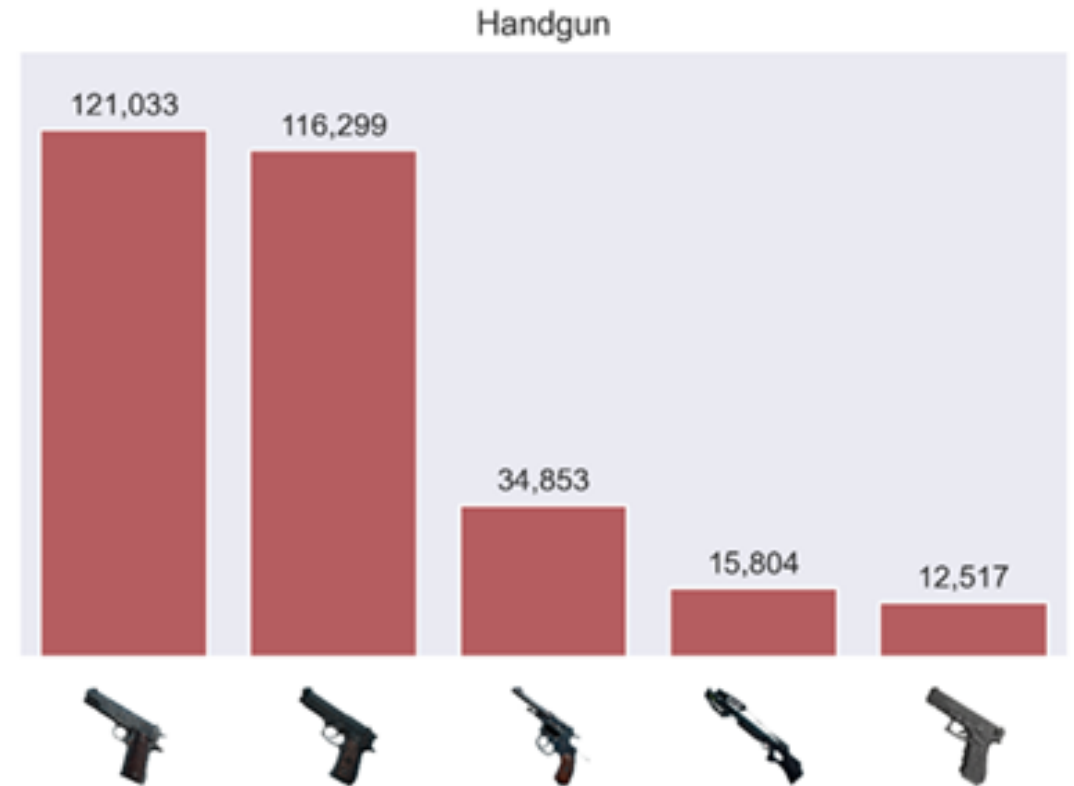
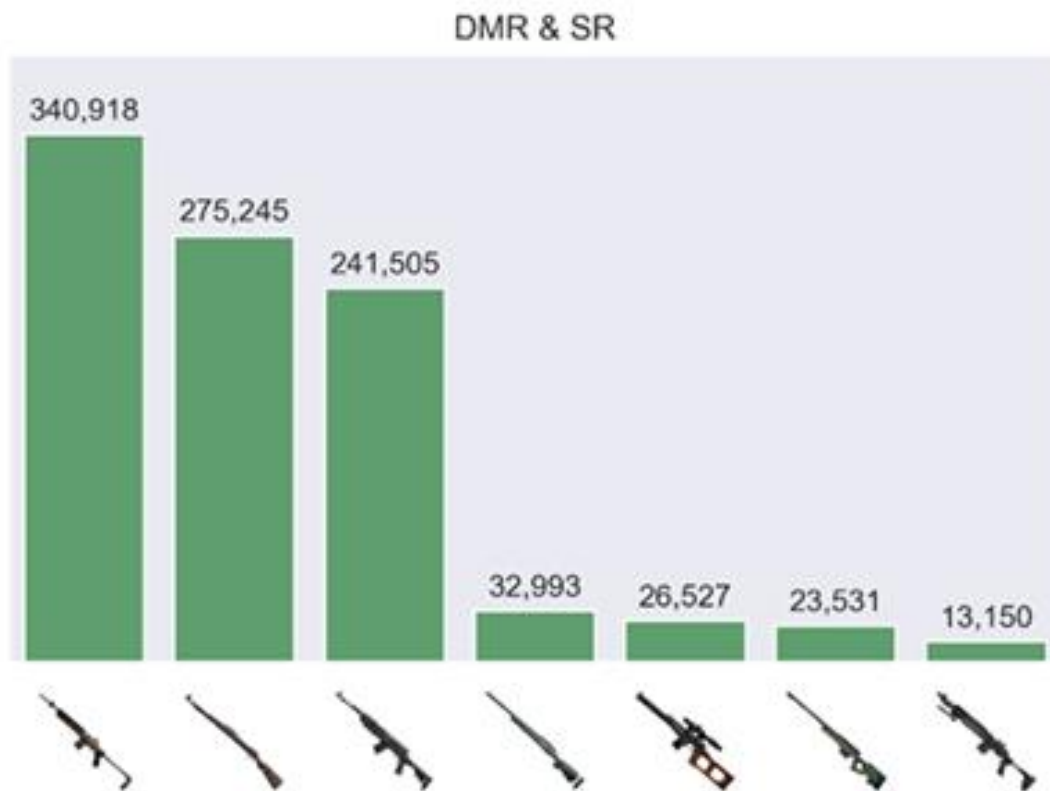
Đếm killed\_by  
theo từng type



# Phân bố giữa các cột

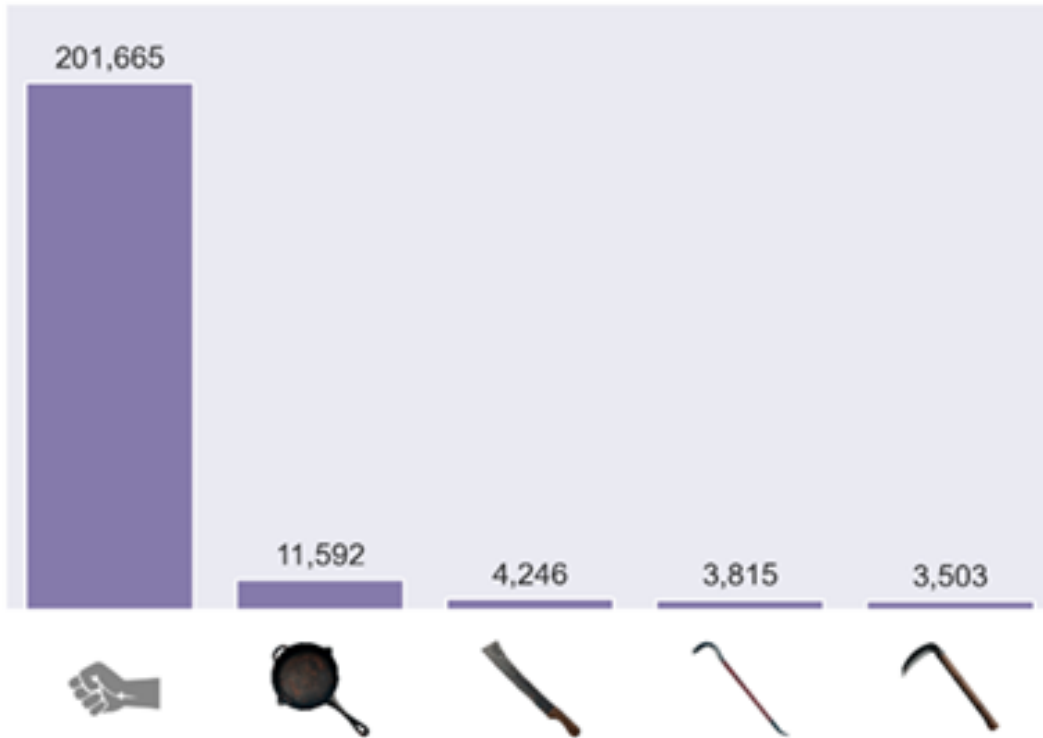


# Phân bố giữa các cột

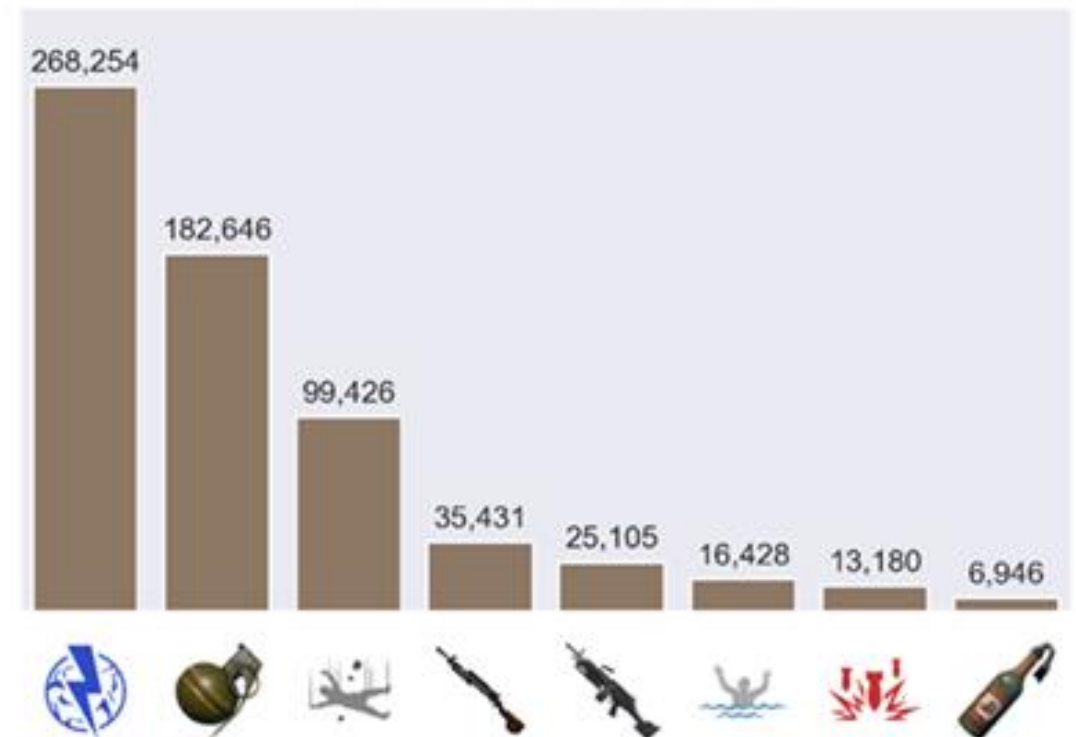


# Phân bố giữa các cột

Melee

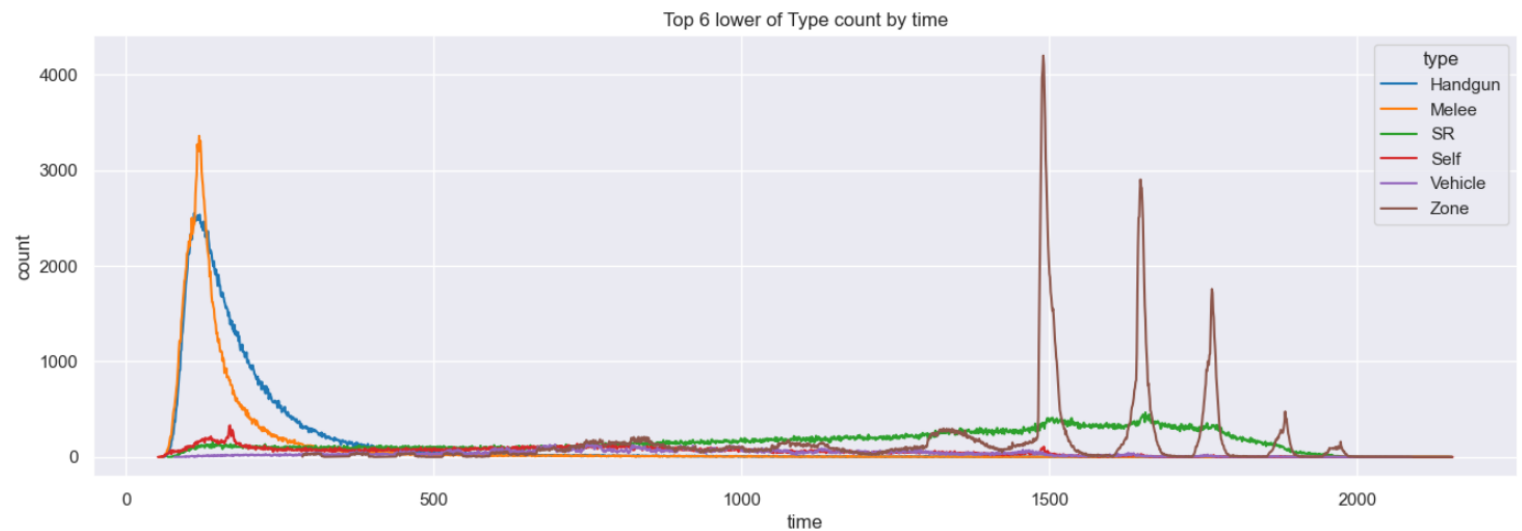
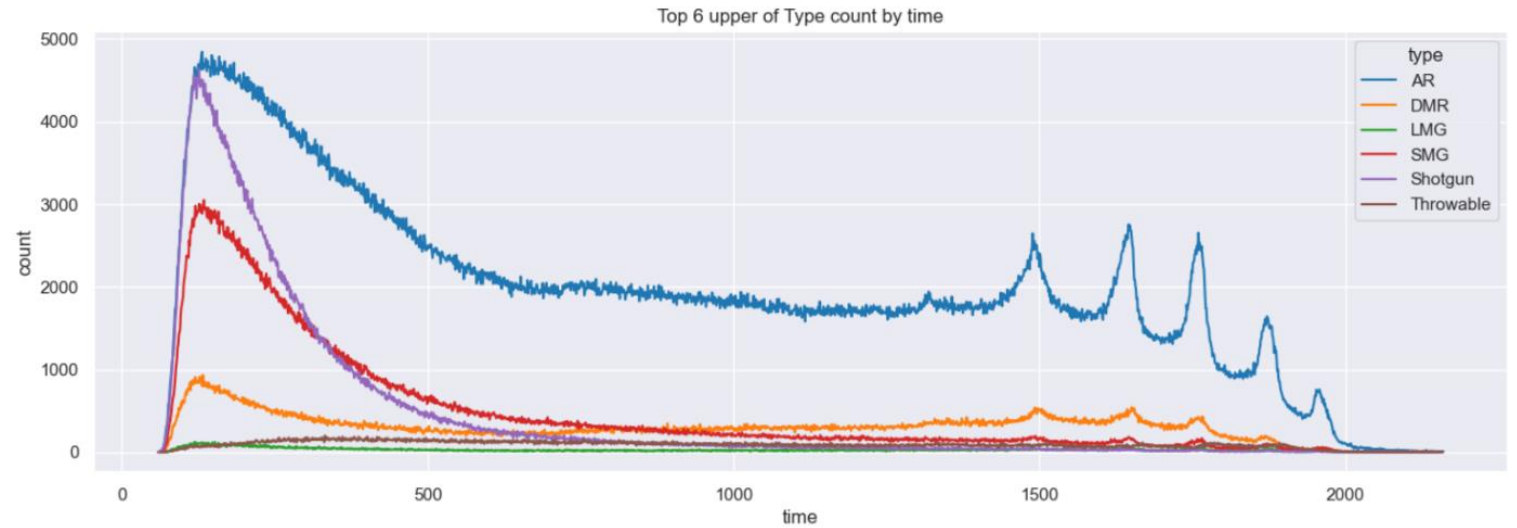


Throwable & Zone & Self & LMG



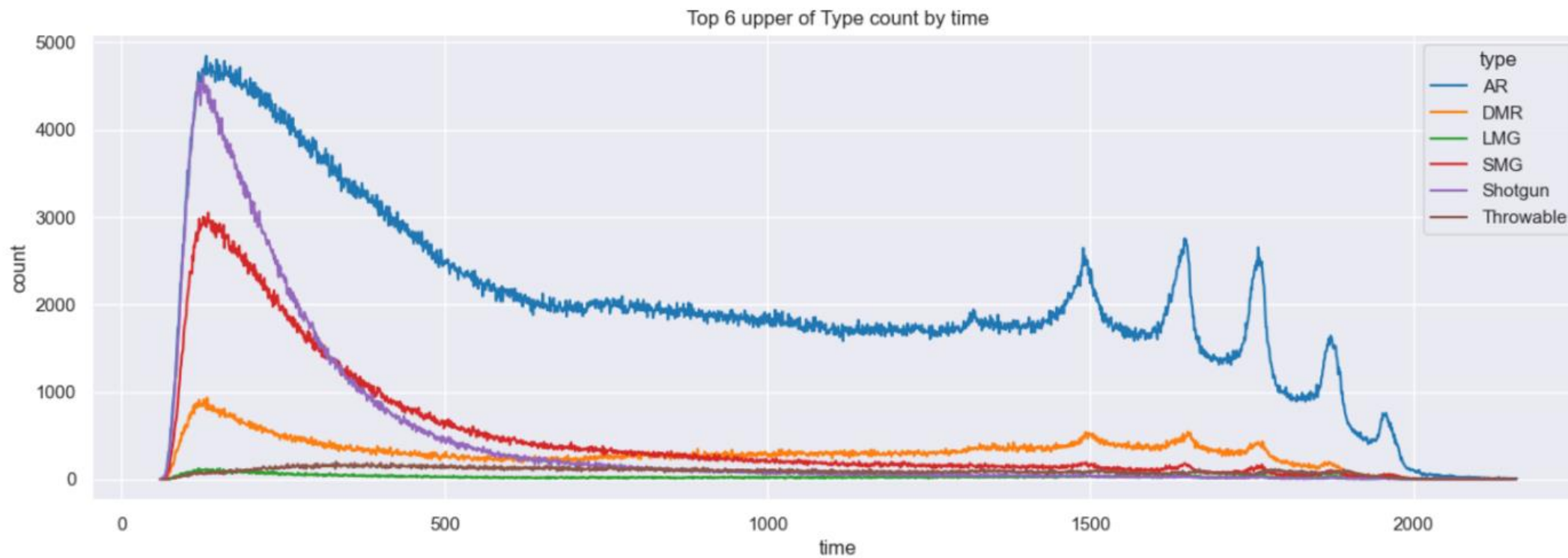
# Phân bố giữa các cột

Đếm **type** theo  
từng **time**

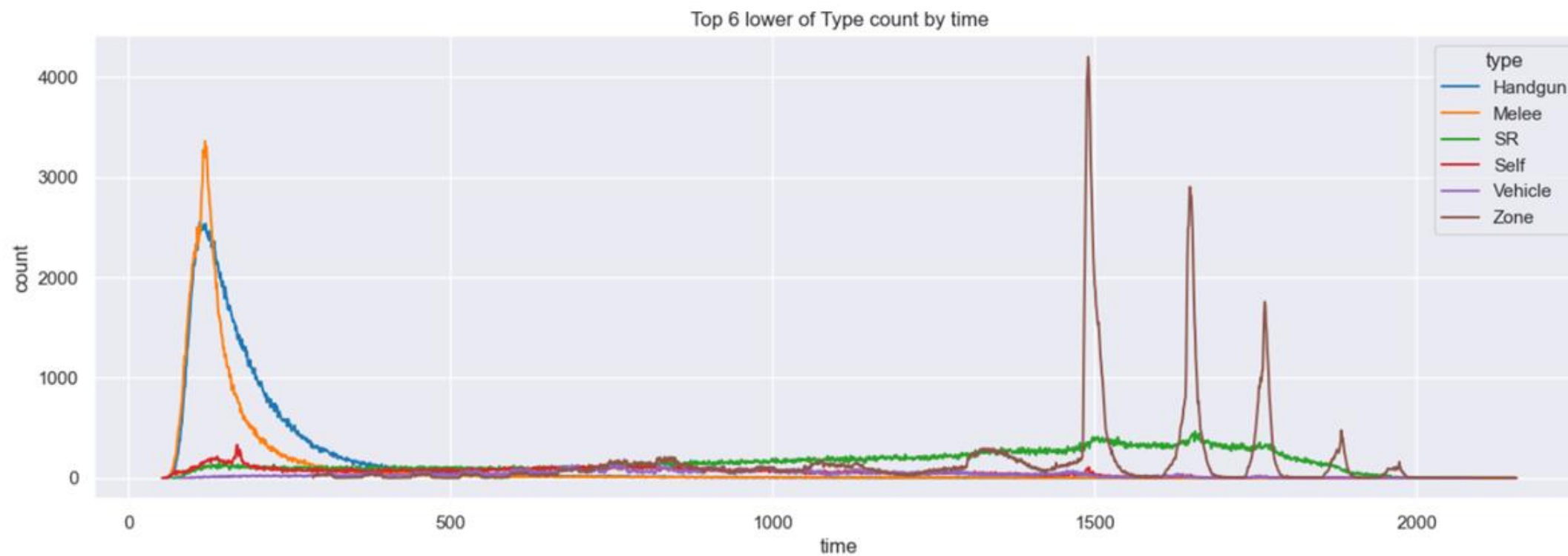




# Phân bố giữa các cột

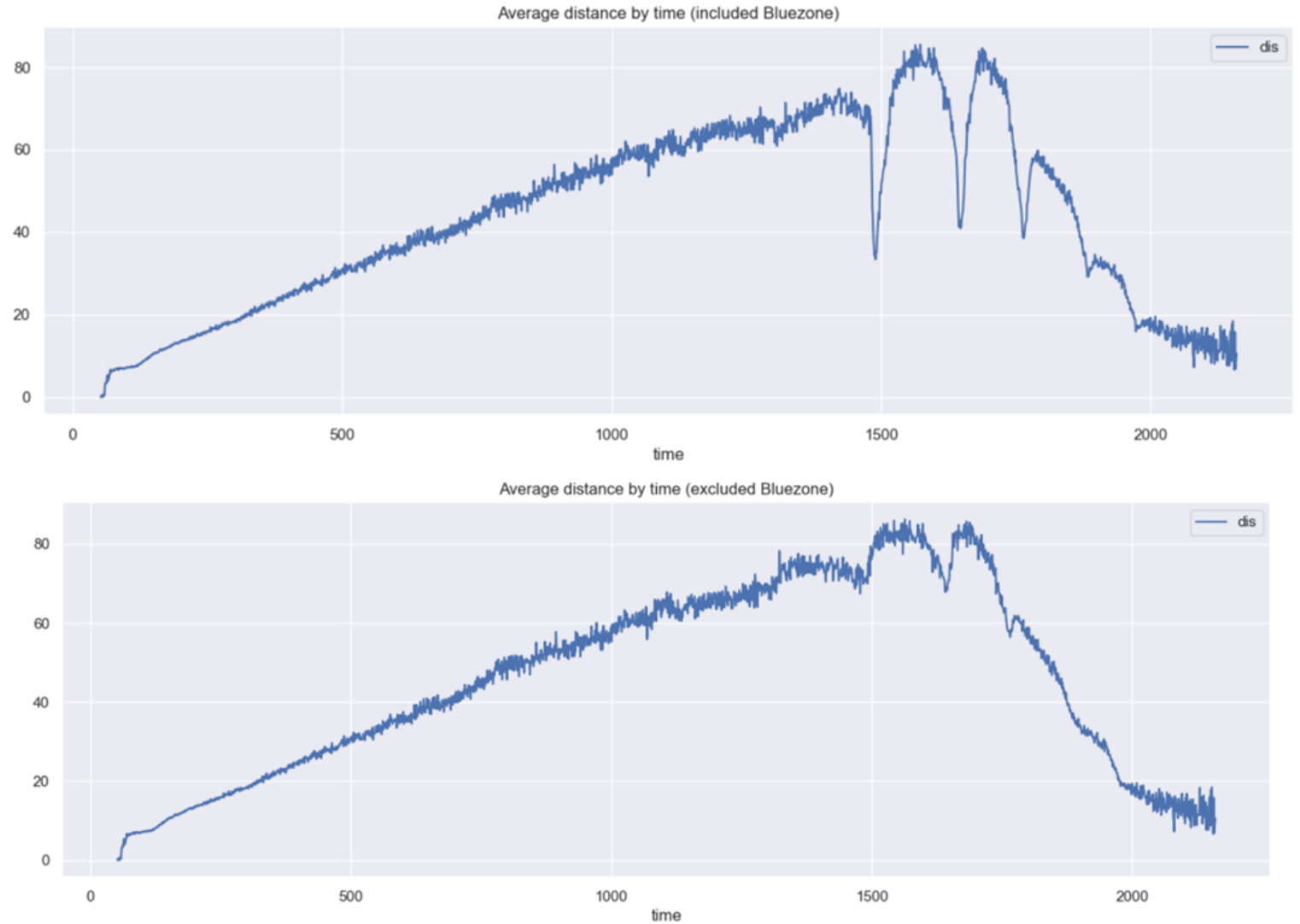


# Phân bố giữa các cột



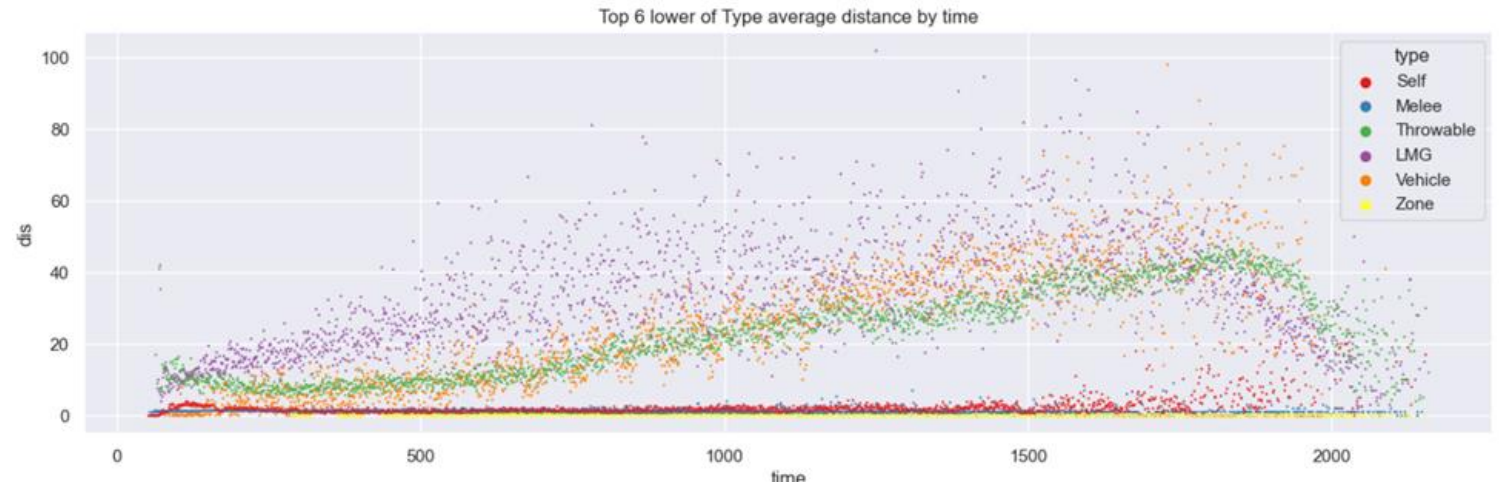
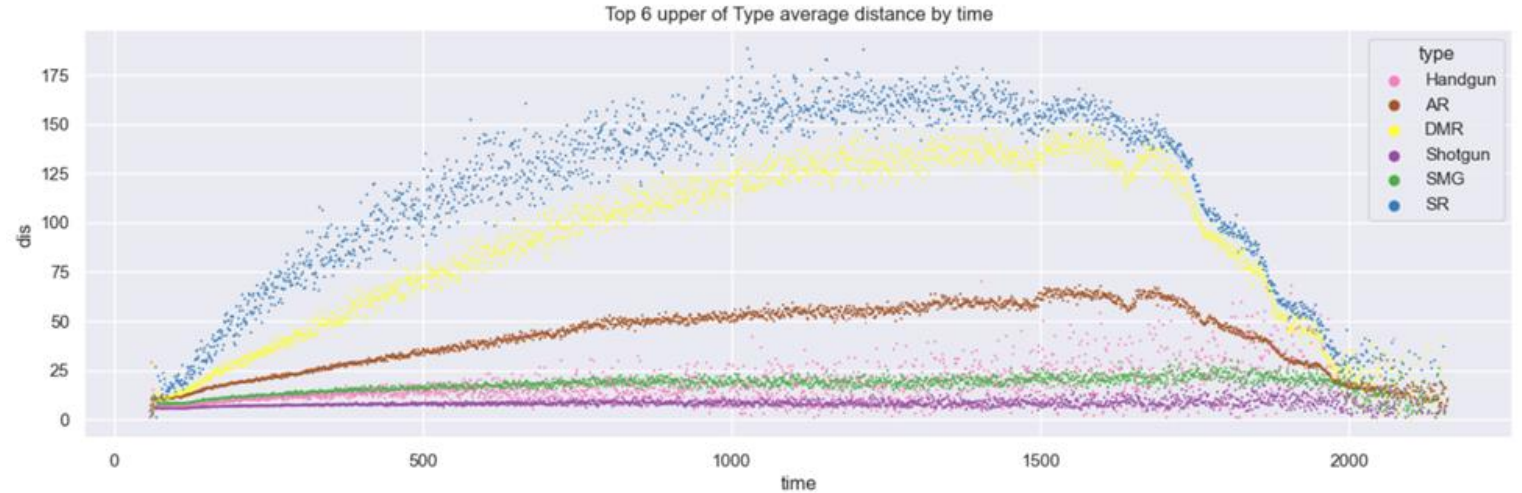
# Phân bố giữa các cột

Trung bình **dis**  
theo **time**



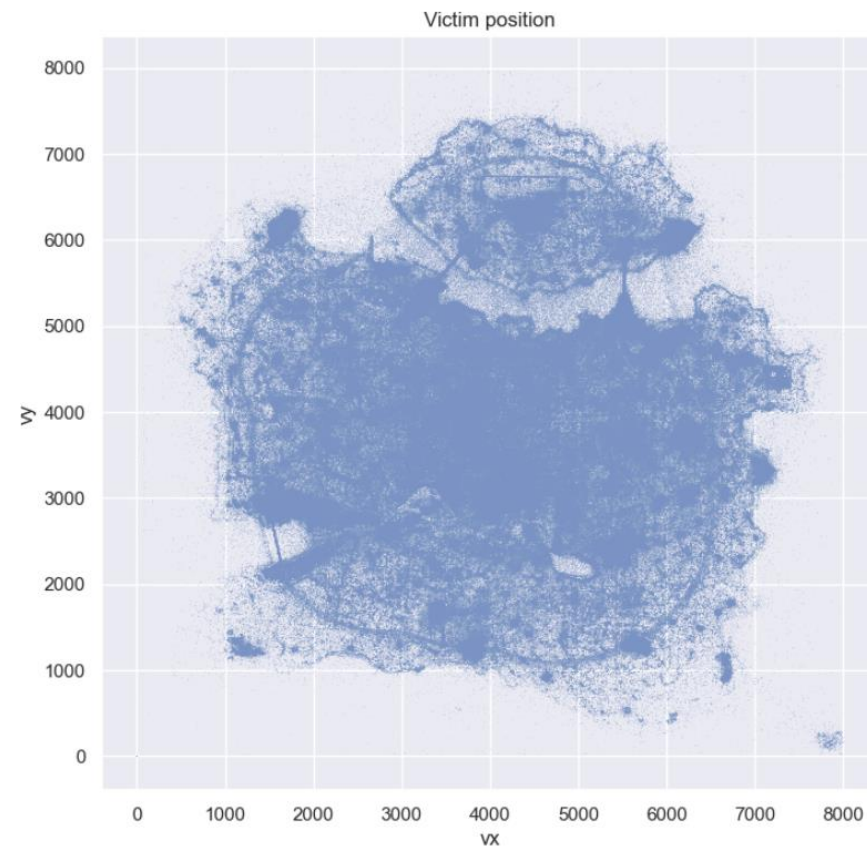
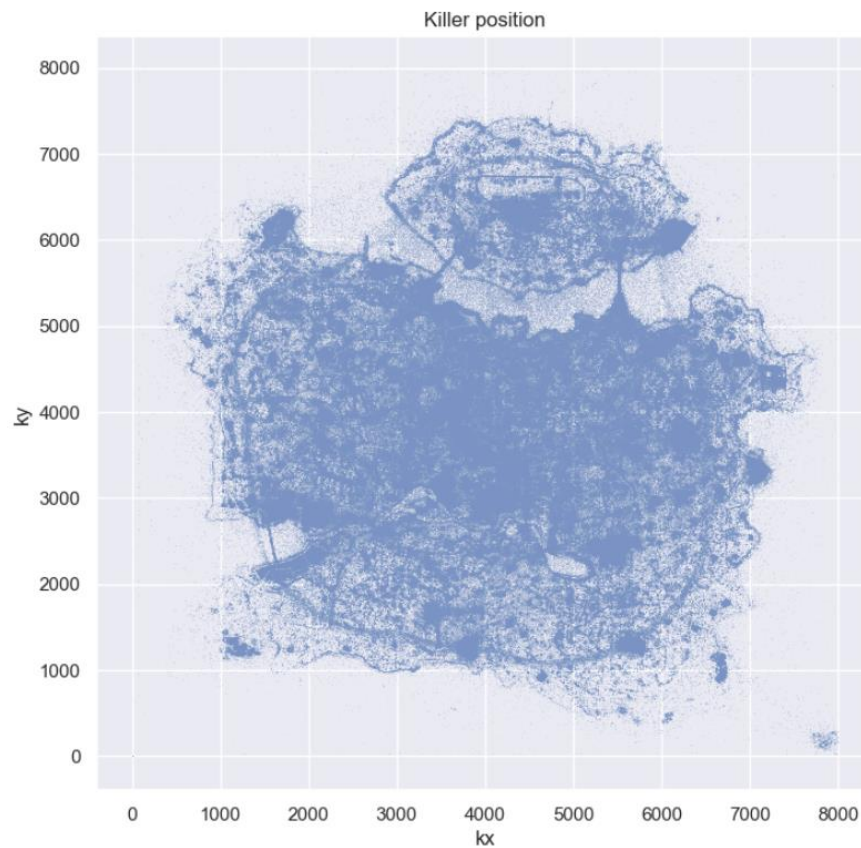
# Phân bố giữa các cột

Trung bình **dis**  
theo **time** và **type**



# Phân bố giữa các cột

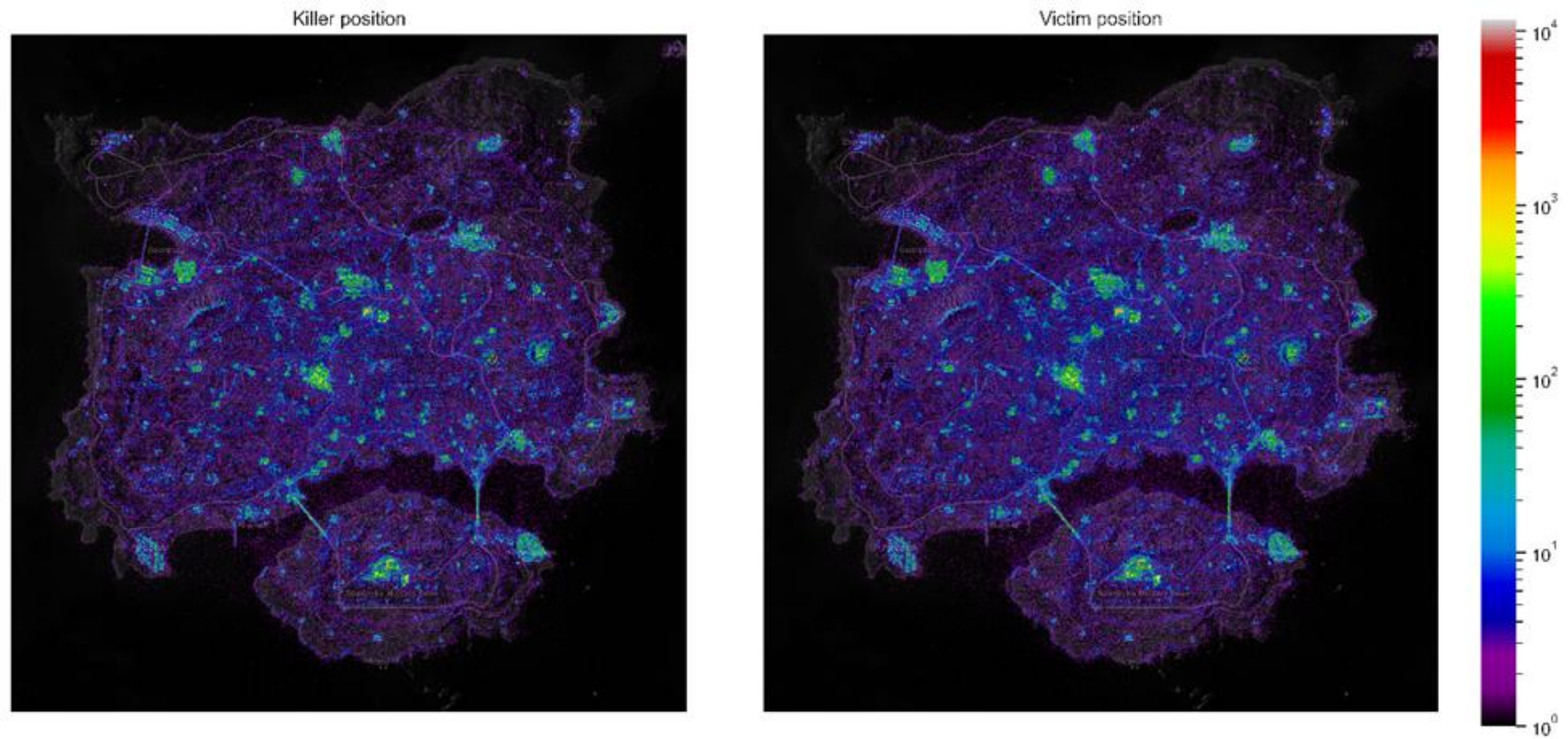
## Phân bố tọa độ trên mặt phẳng 2 chiều





# Phân bố giữa các cột

## Phân bố tọa độ trên mặt phẳng 2 chiều



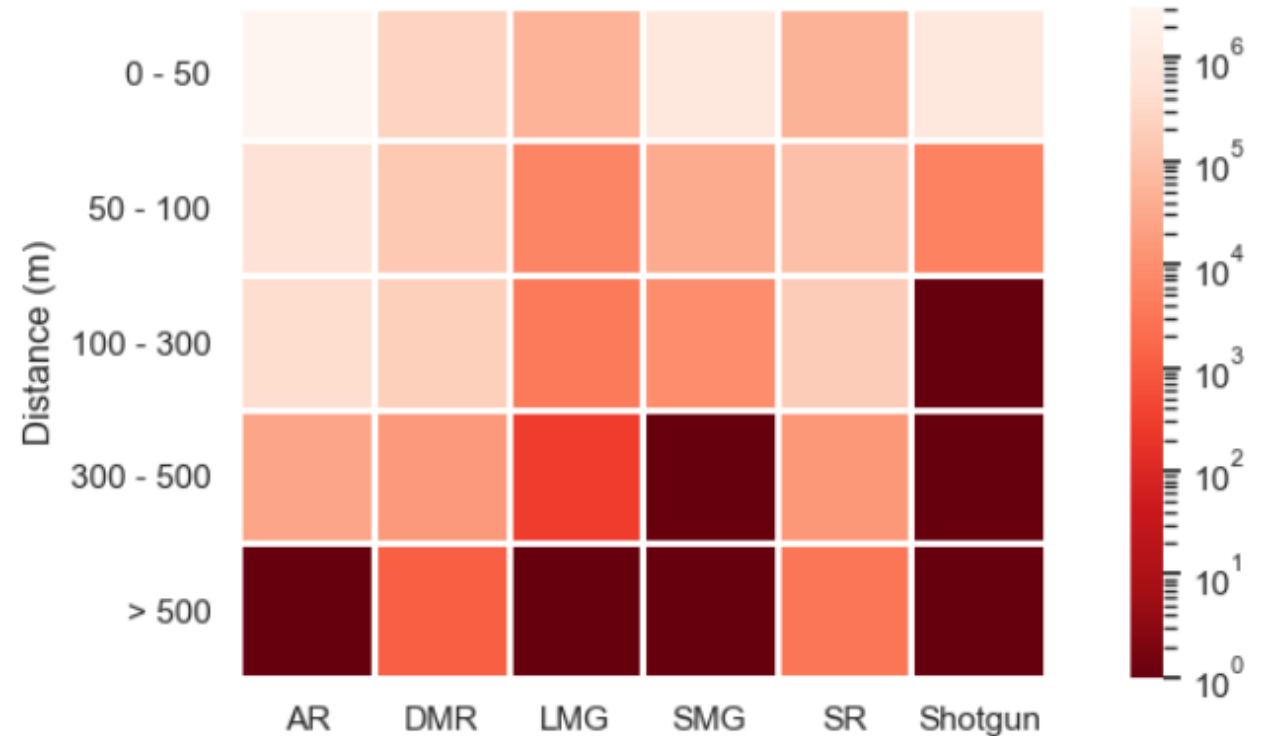


Các câu hỏi  
có ý nghĩa

# 1. Loại vũ khí hiệu quả nhất

Mỗi player chỉ được trang bị **2 vũ khí chính**, do đó cần chọn loại vũ khí hiệu quả cho các trường hợp giao tranh khác nhau.

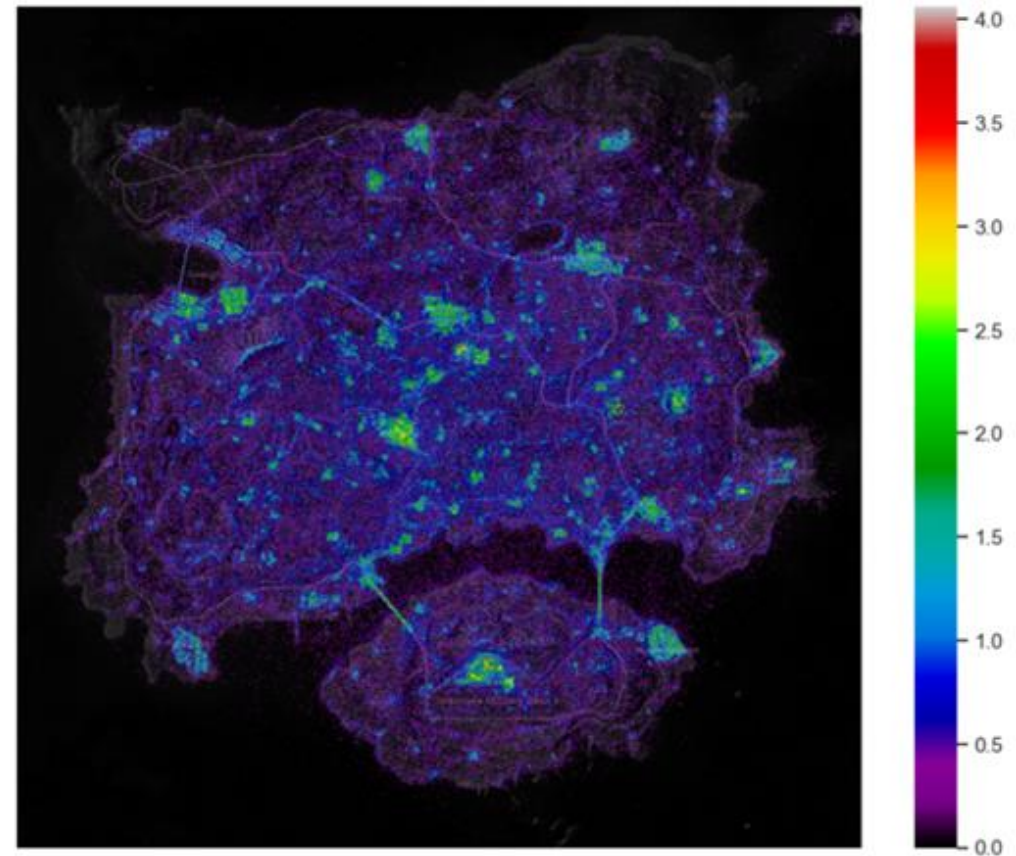
Đánh giá độ hiệu quả của vũ khí bằng **số lượng kill** gây ra ở từng **khoảng cách**

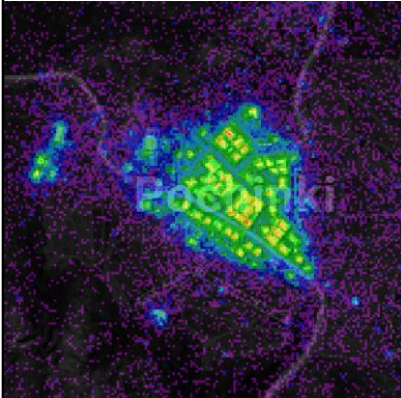
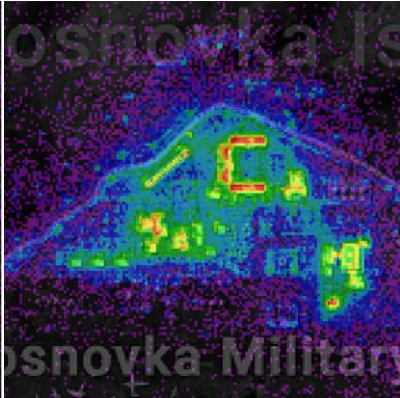

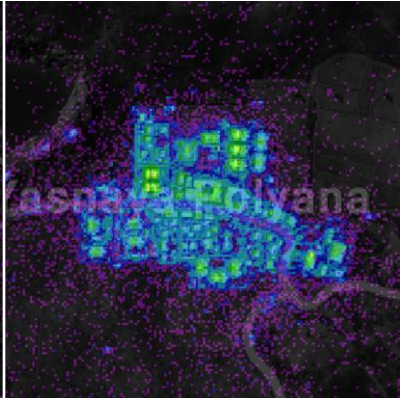
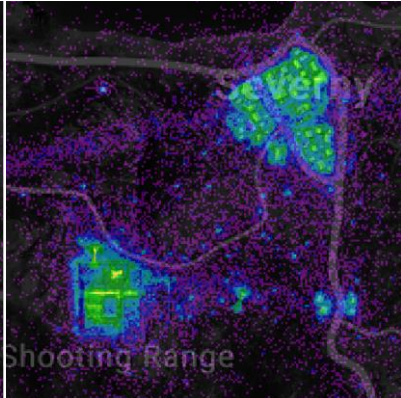
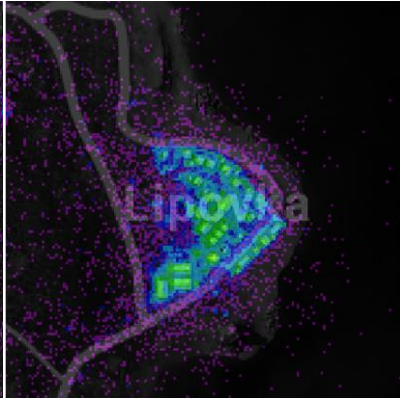
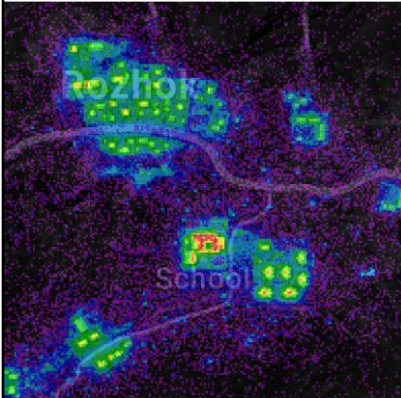
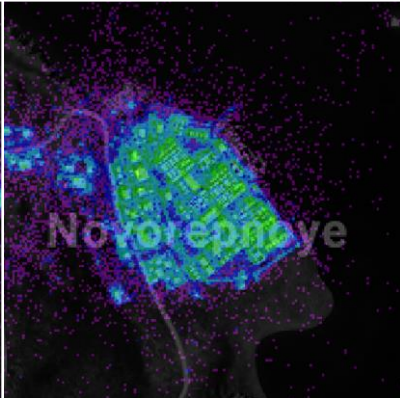
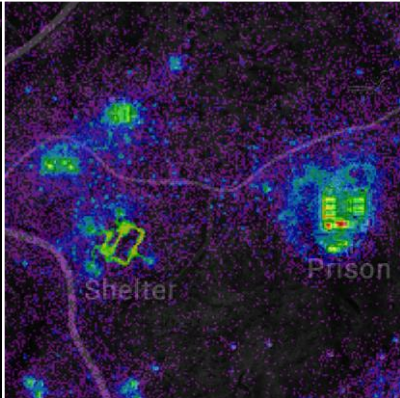
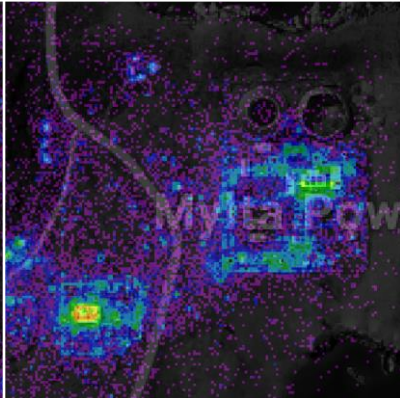
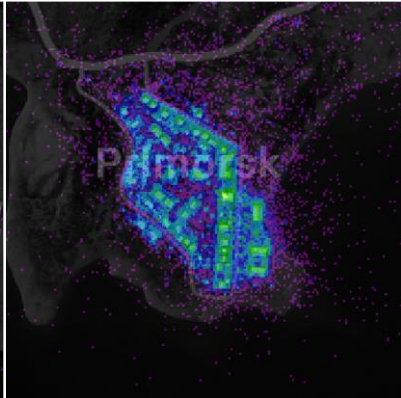
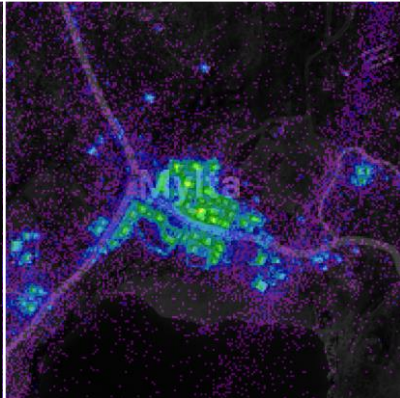




## 2. Những vị trí giao tranh **nhều ở khoảng đầu trận**

Trên bản đồ sẽ có những khu vực giao tranh nhộn nhịp hơn những khu vực khác, biết được các khu vực này giúp player chọn nơi nhảy dù, trú ẩn phù hợp với lối chơi của mình

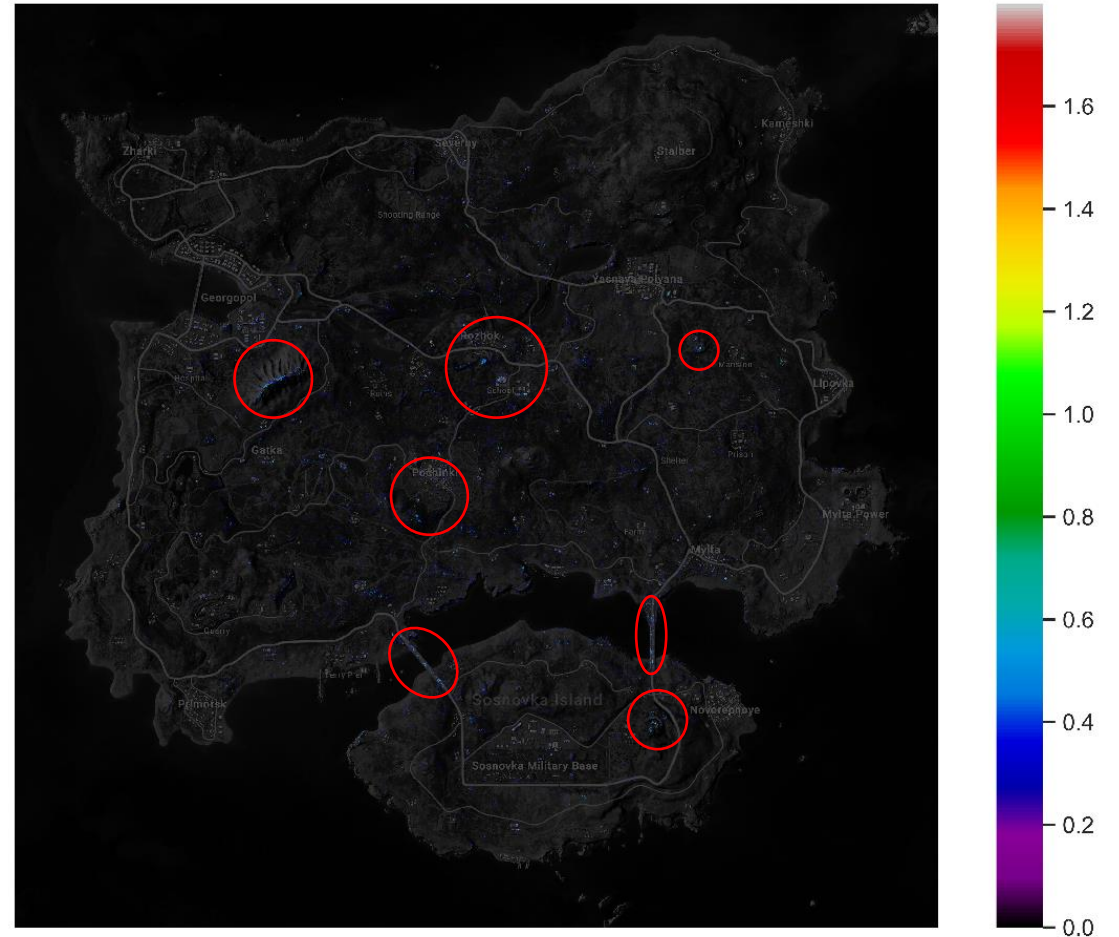


Pochinki	Military Base	Georgopol	Yasnaya Polyana	Serveny	Lipovka
					
Rozhok & School	Novorepnoye	Prison & Shelter	Myta	Primorsk	Myta Power
					



### 3. Những vị trí thuận lợi gây kill từ xa

Trong đa số giai đoạn của trận đấu thì **khả năng gây kill từ xa** sẽ mang lại lợi thế rất lớn cho các player. **SR và DMR** là các vũ khí ưa chuộng dùng để gây kill từ xa. Biết được những vị trí thuận lợi để gây kill từ xa sẽ giúp các player **có thêm những sự lựa chọn trong chiến lược** của mình.



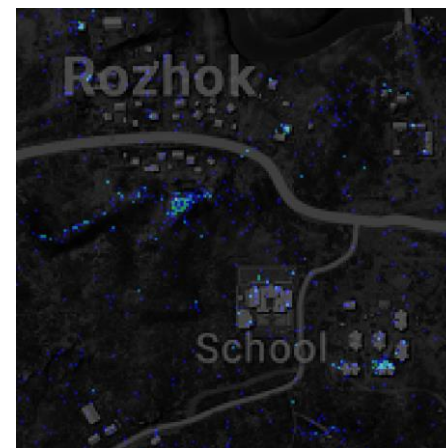
**Đồi xương cá**



**Mansion**



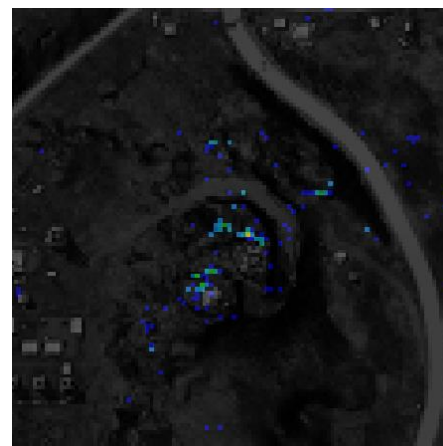
**Rozhok**



**Pochinki**

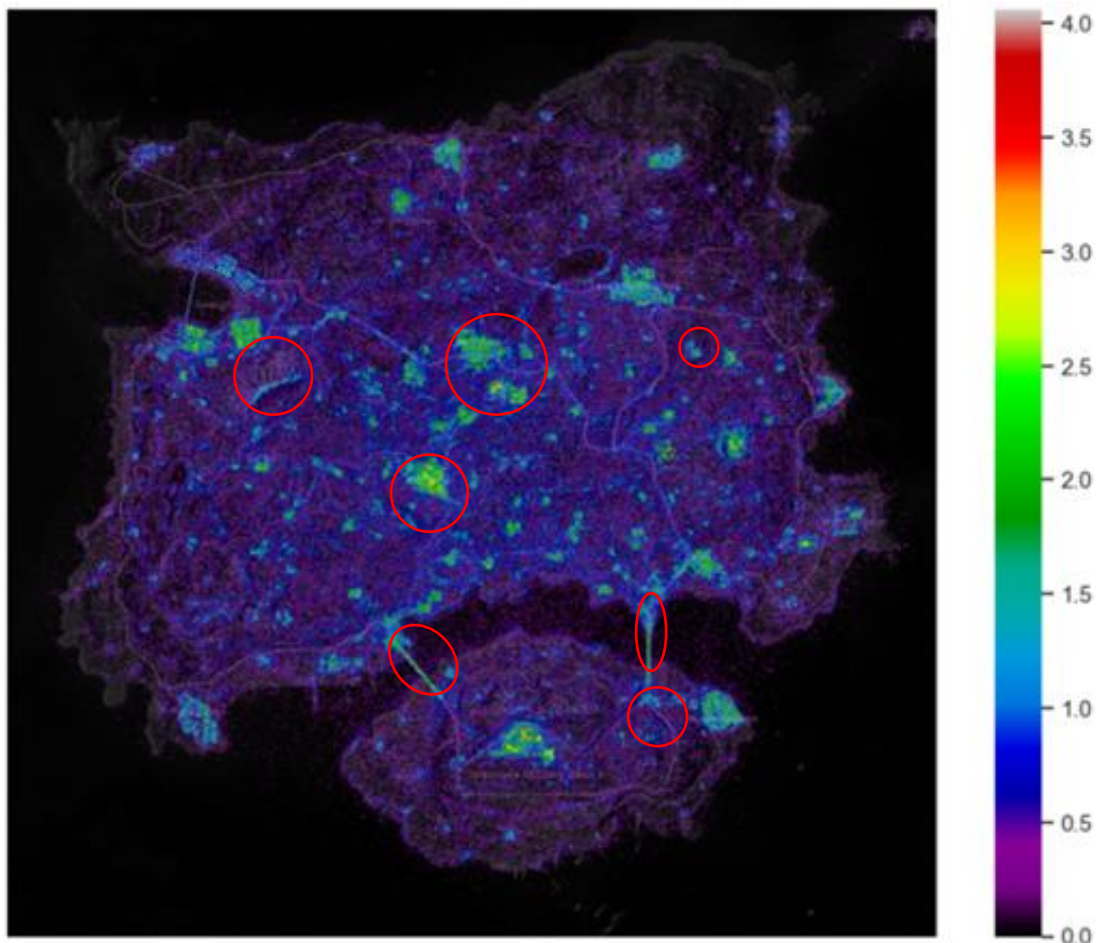


**Military base**



## 4. Những vị trí thuận lợi gây kill từ xa có thực sự an toàn?

Tuy việc biết được các vị trí thuận lợi gây kill từ xa sẽ mang lại lợi thế nhưng các player cũng cần biết **tình hình giao tranh tại các khu vực này.**



## 4. Những vị trí thuận lợi gây kill từ xa có thực sự an toàn?

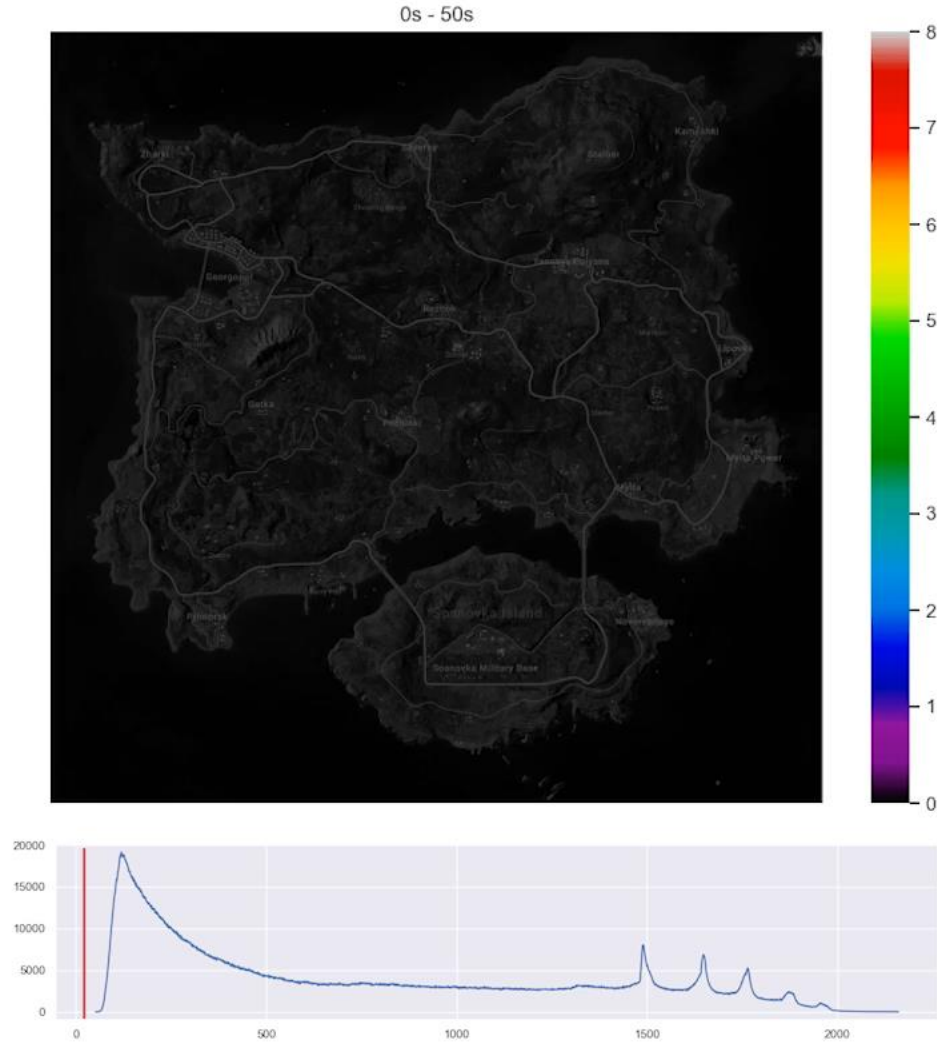
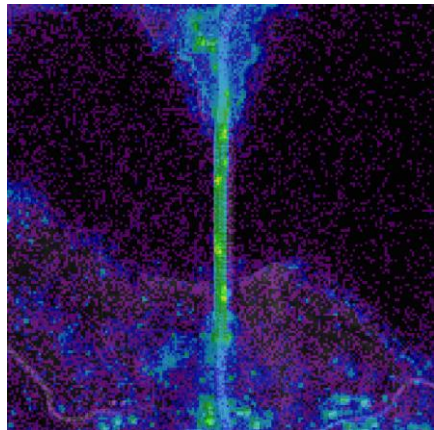
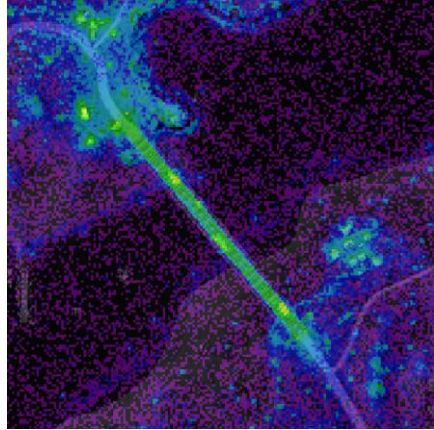
Tuy việc biết được các vị trí thuận lợi gây kill từ xa sẽ mang lại lợi thế nhưng các player cũng cần biết tình hình giao tranh tại các khu vực này.







## 5. Thời điểm thích hợp để qua cầu





**The end**

20120088 Lê Nguyễn Thanh Hoàng  
20120055 Nguyễn Thế Đạt  
20120085 Trần Xuân Hòa  
20120105 Lê Hoàng Huy