PUBG Death Data Analysis
Data-Driven Player Modelling

Data:
PUBG dataset from kaggle.

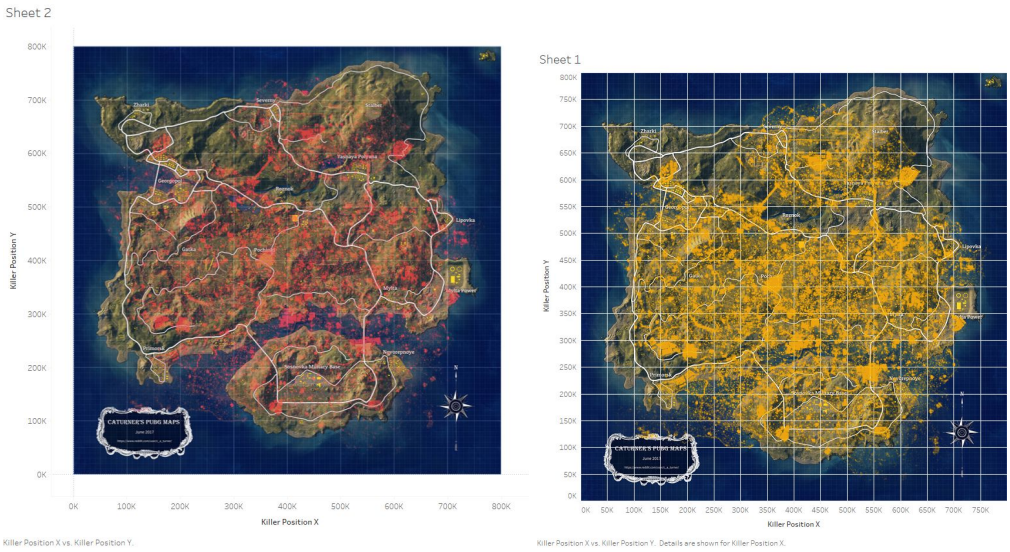
Aim:

1. Feature Engineering - We derived features and added new columns to the data to make to analyse it further.
2. We aim to identify the areas where players have a higher chance of killing other players.
3. Develop a model which predicts a gun based on the distance and elevation of the victim, player and game time.

- Original Data set contains:
    - 12 columns
    - Approx 60,000,000 rows

- There is so much redundant data like player name, victim name etc which won't contribute towards the analysis process. So we remove those columns
- Redundant Data:
    - Killer_name
    - Match_id
    - Map
    - Victim_name
    - killer_Placement
    - Victim_placement

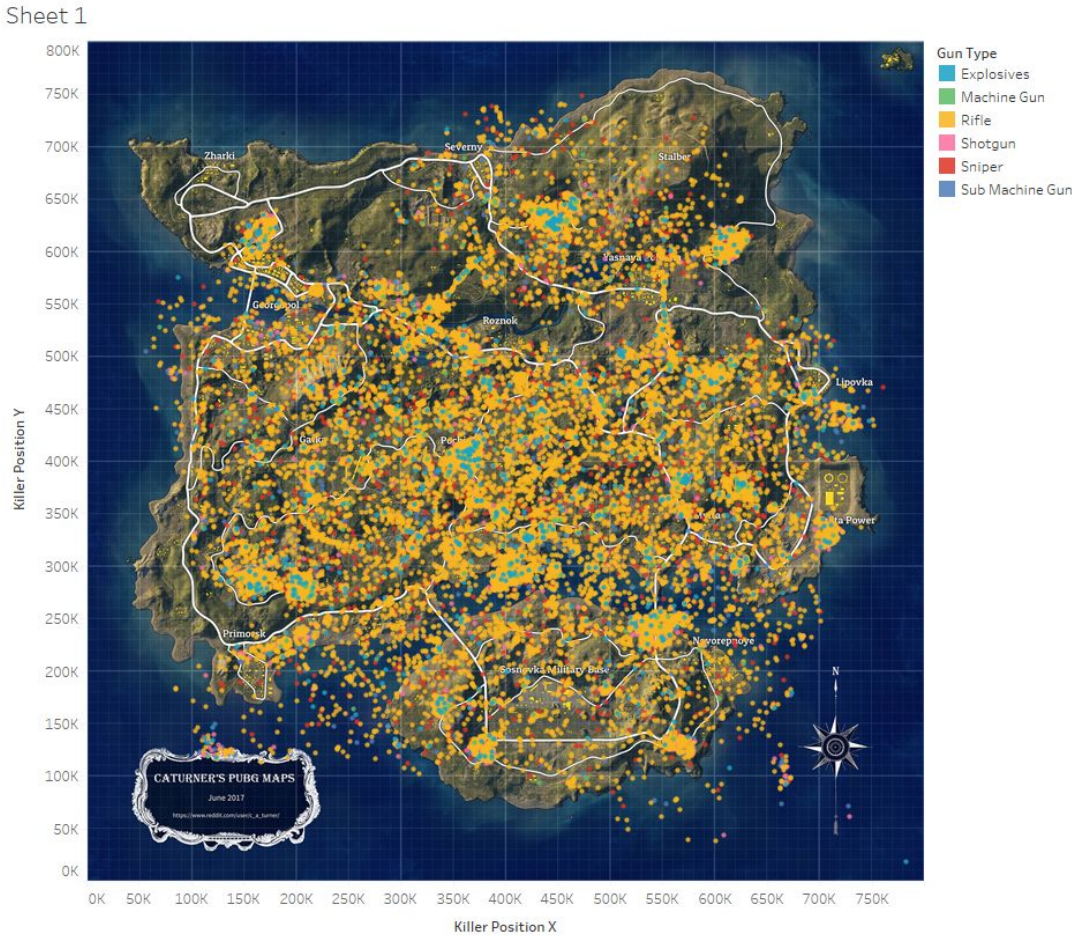
- Outlier Detection:
    - Outliers could only be present in killer and victim positional coordinates(x,y)



- Upon visualization, there were no points that were going out of map
- N/A values present in values falling under columns whose corresponding rows were removed:
  - Killed_by (Gun)
  - victim x, y position
  - Killer x, y position
- Many positional values have 0.0 and they were removed as well.
- “Killed_by” columns had some unidentified values and they were removed.
  - death.None
  - death.PlayerMale_A_C
  - death.Buff_FireDOT_C

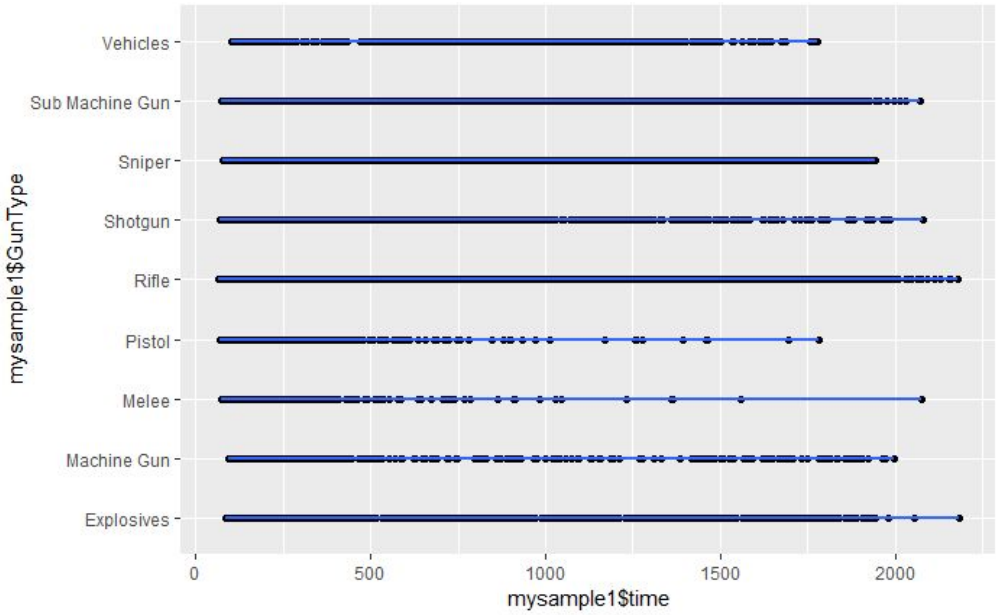
- Feature Engineering:
  - Gun Type: We classified the guns according to their class
  - Rifle = "Groza", "M16A4", "Mk14", "SCAR-L", "M416", "Mini 14", "AKM", "AUG"
  - MG = "M249", "DP-28"
- Killer and Victim Elevation:
  - We identified the hills and mountains from the map
  - We tagged killer/victim positions as “Elevated” if they were on a hill/mountain
- Kill Distance: Used distance formula to measure the trajectory of kill from killer to the victim

Visualisations:

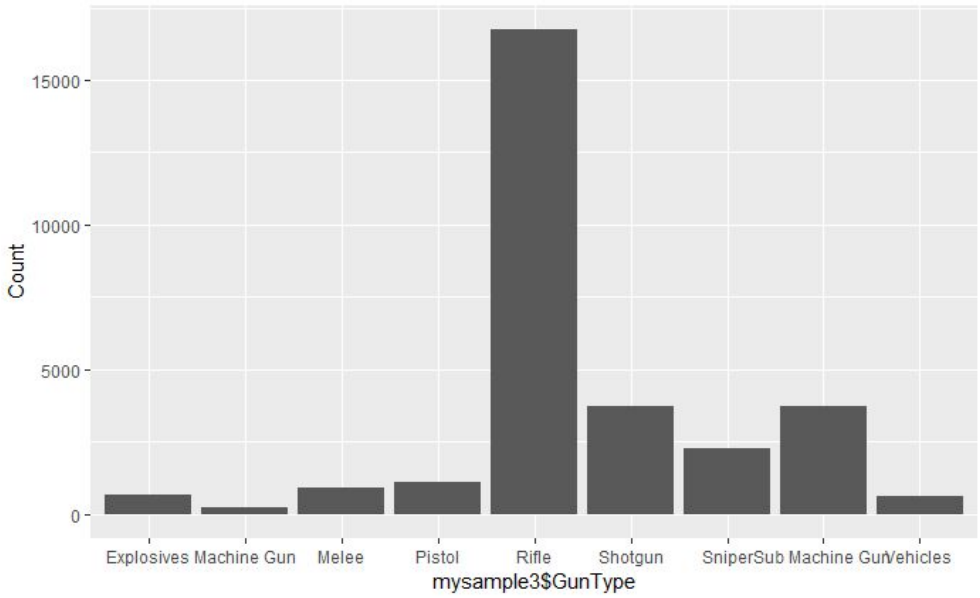


Killer Position X vs. Killer Position Y. Color shows details about Gun Type. The view is filtered on Gun Type, which excludes Melee, Pistol and Vehicles.

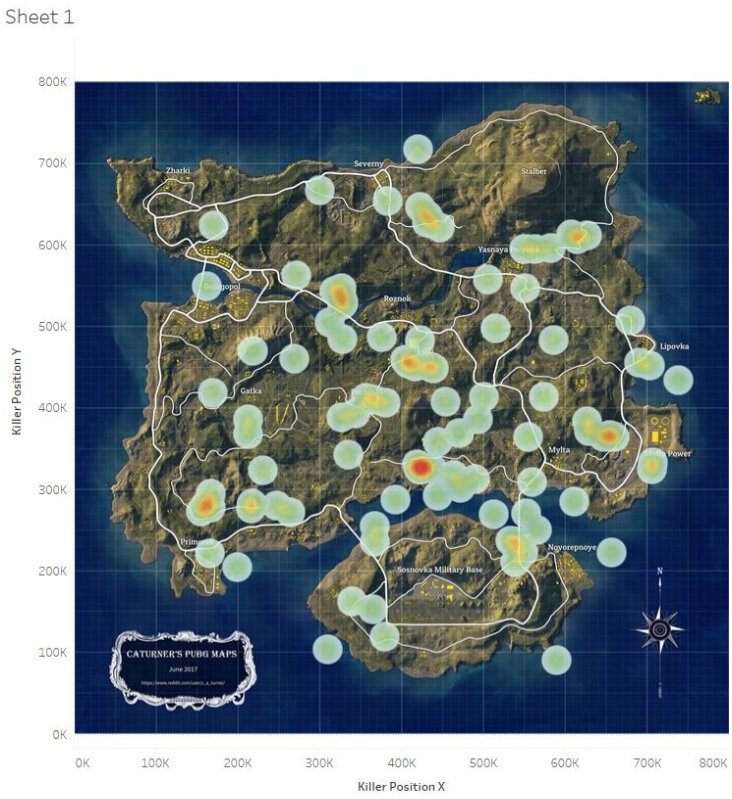
The above visualisation shows the distributions of weapons used by the killer players to kill the victim players. Ther different types of guns used are shown by the different colors. From this image, we notice that the explosives are used in a less scattered way than the other guns.



We see from the above image that the melee ways of killing other players is rarely used. Using Pistols in the late game is also very rare as pistols are not very efficient weapons.

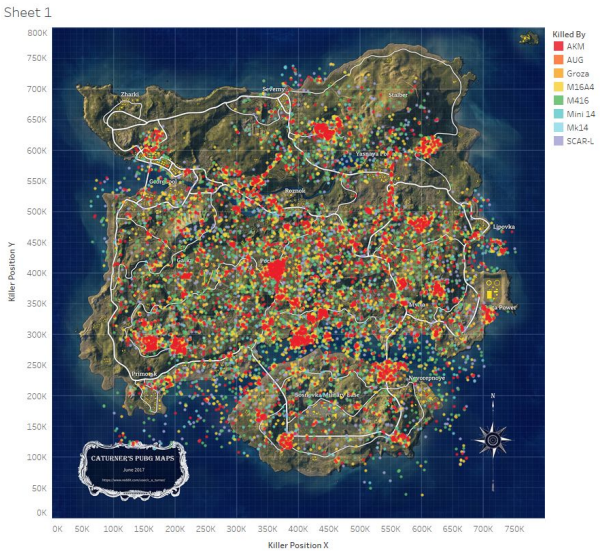


The bar chart shows the gun types used. The top 4 kinds of guns used were the Rifles, Shotguns, Sniper and Submachine guns. We will explore them in more detail.



Killer Position X vs. Killer Position Y.

The heatmaps show the areas on the may where the killer players have killed the victims using Rifles. The image below shows the different kinds of rifles used.



Killer Position X vs. Killer Position Y. Color shows details about Killed By. The data is filtered on Gun Type, which keeps Rifle, Sniper and Sub Machine Gun. The view is filtered on Killed By, which excludes 10 members.

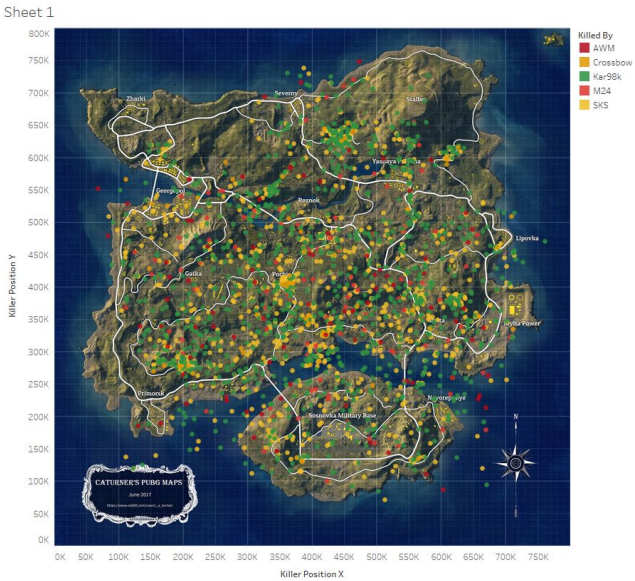
The rifle AKM is used is the areas that have urban structures with some elevation like warehouses, buildings etc. The rifles are long to medium range weapons and

hence can be used from the top of the buildings.

When we made heatmaps for the snipers, we see that snipers are more scattered throughout the map. The major use of snipers are near elevated areas as snipers are long range weapons.

Killer Position X vs. Killer Position Y.

The below map shows the use of all kinds of snipers. The Sniper Kar98k was used the most and is used all over the map.



Killer Position X vs. Killer Position Y. Color shows details about Killed By. The data is filtered on Gun Type, which keeps Rifle, Sniper and Sub Machine Gun. The view is filtered on Killed By, which excludes 13 members.

The below map shows the use different kinds of submachine guns all over the map.

Submachine guns are medium to short range weapons and we see that the players use them near urban structures.

The heatmap below highlights all the urban structures where sub machine guns are used more.

Data Analysis:

- The Data Set:
    - The clean data set has 39,705,577 rows which my computer couldn't handle.v So we sampled 3 different data sets of 30000 rows each.
        - Each data set has random values
        - All the computations will be performed on the three data sets and compared for validation
- Correlation Matrix:
    - We used one hot encoding method to create dummy variables for rifles and elevation. After which we ran correlation analysis. We got some interesting results.
        - Explosives, Rifle and Sniper have a **positive correlation** with time, meaning that the usage of these classes of weapons depends on the time of the game.
        - Rifles,and Snipers have a positive correlation with killer elevation meaning that the killer will have an advantage over the victim if the killer is at an elevated position.
        - Sub machine guns have a **positive correlation** when both the **victim and killer are on normal elevations**
- Data Splitting:
    - The three data sets were split:
        - 0.75 Training and 0.25 Testing randomly using a seed to replicate the results. These sets are unscaled since using min max normalization is not appropriate for positional data and distance value. Min max normalization further reduces distance value
        - Instead we create another data set and zscale it and run analysis on it
- Columns used for classifications:
    - We are going to predict the gun type for a scenario which contains the following factors:
        - Killer position (x,y)
        - Victim position (x,y)
        - Time in the game, since the battlefield reduces in size as time progresses
        - Killer Elevation

        - Victim Elevation
- kNN:
    - Performed kNN classification with k fold cross validation
    - Moderate accuracy
    - Bad Kappa
    - Bottom line, the model is pretty inaccurate and even more unreliable

```
Overall Statistics

               Accuracy : 0.4811
                 95% CI : (0.4697, 0.4924)
    No Information Rate : 0.5595
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0912
 Mcnemar's Test P-Value : NA
```

    -

```
Overall Statistics

               Accuracy : 0.4849
                 95% CI : (0.4736, 0.4963)
    No Information Rate : 0.5564
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0913
 Mcnemar's Test P-Value : NA


Overall Statistics

               Accuracy : 0.4907
                 95% CI : (0.4793, 0.502)
    No Information Rate : 0.5586
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1066
 Mcnemar's Test P-Value : NA
```

    - kNN on scaled data:

```
Overall Statistics

               Accuracy : 0.4825
                 95% CI : (0.4712, 0.4939)
    No Information Rate : 0.5595
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0941
 Mcnemar's Test P-Value : NA
```

    -
    - There's no difference in kNN when scaled or unscaled data is used
- Rule Learner:

- Mediocre Accuracy
- Low Kappa so the Reliability score bit better than kNN

```
Overall Statistics

               Accuracy : 0.5814
                 95% CI : (0.5701, 0.5926)
    No Information Rate : 0.5586
    P-Value [Acc > NIR] : 3.545e-05

                  Kappa : 0.1035
 Mcnemar's Test P-Value : NA
```

```
Overall Statistics

               Accuracy : 0.584
                 95% CI : (0.5728, 0.5952)
    No Information Rate : 0.5595
    P-Value [Acc > NIR] : 9.453e-06

                  Kappa : 0.1214
 Mcnemar's Test P-Value : NA
```

```
Overall Statistics

               Accuracy : 0.5815
                 95% CI : (0.5702, 0.5927)
    No Information Rate : 0.5564
    P-Value [Acc > NIR] : 6.289e-06

                  Kappa : 0.1063
 Mcnemar's Test P-Value : NA
```

- The rules tree:

```
JRIP rules:
===========

(killdistance <= 1.552417) and (time >= 537) and (killer_position_x <= 334228.4) => GunType=Vehicles (51.0/23.0)
(killdistance <= 0) and (time >= 704) and (killer_position_y >= 367525.4) and (time <= 847) => GunType=Vehicles (19.0/6.0)
(killdistance <= 1.389244) and (time >= 914) and (killer_position_y <= 210491.7) => GunType=Vehicles (7.0/1.0)
(killdistance <= 0) => GunType=Explosives (254.0/73.0)
(killdistance <= 98.595588) and (time <= 174) and (killer_position_x <= 354326.3) => GunType=Melee (115.0/21.0)
(killdistance <= 131.158263) and (time <= 128) => GunType=Melee (314.0/108.0)
(killdistance <= 93.407708) and (time <= 284) and (killdistance >= 78.751317) => GunType=Melee (190.0/69.0)
(killdistance <= 115.702031) and (killer_position_x >= 488458.7) and (killer_position_x <= 535550.8) => GunType=Melee (17.0/6.0)
(killdistance <= 128.232016) and (time <= 160) and (killer_position_x >= 432195.8) => GunType=Melee (37.0/18.0)
(killdistance >= 9867.813132) and (time >= 1678) and (killdistance <= 15483.686419) => GunType=Sniper (166.0/81.0)
(killdistance >= 10981.405343) and (killer_position_y >= 322453.1) and (killer_position_y <= 369829.1) and (time >= 1505) => GunType=Sniper (55.0/23.0)
(killdistance <= 1043.960708) and (killer_position_y >= 414874.8) and (victim_position_y <= 627478.2) and (victim_position_y >= 613295.3) and (killdistance >= 242.88699) and
(time <= 138) and (victim_position_x <= 451097.7) => GunType=Shotgun (59.0/18.0)
(time <= 268) and (killdistance <= 685.914936) and (killdistance >= 603.52299) and (killer_position_x >= 420193.6) and (victim_position_y >= 325850.6) => GunType=Shotgun
(111.0/51.0)
 => GunType=Rifle (21107.0/8738.0)

Number of Rules : 14
```

- In the rule tree, u can see that killer position, victim position and time are the main factors that help predict guns.
- Rule learner on scaled data

```
Overall Statistics

               Accuracy : 0.584
                 95% CI : (0.5728, 0.5952)
    No Information Rate : 0.5595
    P-Value [Acc > NIR] : 9.453e-06

                  Kappa : 0.1214
 Mcnemar's Test P-Value : NA
```

- There's no change with scaled and unscaled data too

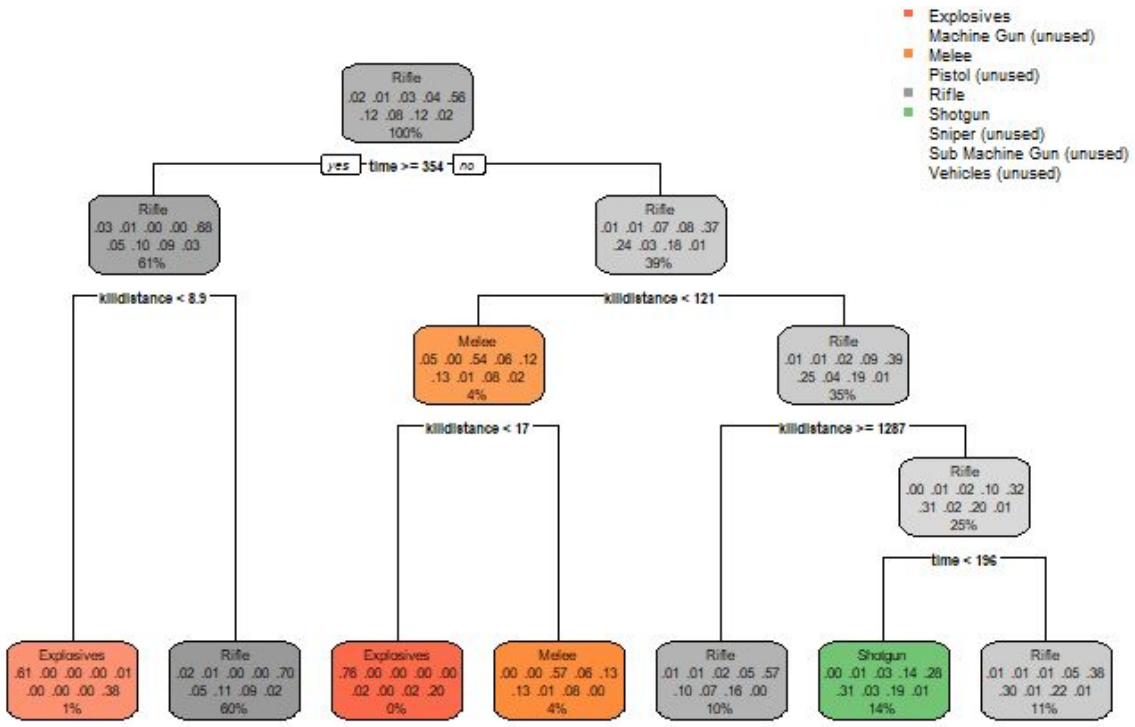
- Rparts
  - An accuracy of approx 0.60
  - With repeated cross validation
  - Moderate Kappa value of approx 0.22
  - Overall this model is pretty good model compared to other models

```
Overall Statistics

               Accuracy : 0.5902
                 95% CI : (0.5789, 0.6013)
    No Information Rate : 0.5595
    P-Value [Acc > NIR] : 4.323e-08

                  Kappa : 0.2286
 Mcnemar's Test P-Value : NA
```

```
Overall Statistics

               Accuracy : 0.5844
                 95% CI : (0.5732, 0.5956)
    No Information Rate : 0.5564
    P-Value [Acc > NIR] : 5.273e-07

                  Kappa : 0.2323
 Mcnemar's Test P-Value : NA
```

```
Overall Statistics

               Accuracy : 0.5875
                 95% CI : (0.5763, 0.5987)
    No Information Rate : 0.5586
    P-Value [Acc > NIR] : 2.228e-07

                  Kappa : 0.2017
 Mcnemar's Test P-Value : NA
```

  - Tree plot for rparts:



- In the tree, we can see that rifle, melee, explosives and shotguns are the classes of guns that make the tree
- Rparts for scaled data:

```
Overall Statistics

               Accuracy : 0.5902
                 95% CI : (0.5789, 0.6013)
    No Information Rate : 0.5595
    P-Value [Acc > NIR] : 4.323e-08

                  Kappa : 0.2286
 Mcnemar's Test P-Value : NA
```

-
- There's no change for scaled and unscaled data here too.

- SVMs:
  - Used the radial kernel with repeated cross validation
  - Other kernels never converged
  - Mediocre accuracy and 0 Kappa score
  - Not reliable at all

```
Overall Statistics

               Accuracy : 0.5593
                 95% CI : (0.548, 0.5706)
    No Information Rate : 0.5595
    P-Value [Acc > NIR] : 0.5141

                  Kappa : -2e-04
 Mcnemar's Test P-Value : NA
```

○

```
Overall Statistics

               Accuracy : 0.5564
                 95% CI : (0.5451, 0.5677)
    No Information Rate : 0.5564
    P-Value [Acc > NIR] : 0.5048

                  Kappa : 0
 Mcnemar's Test P-Value : NA
```

```
Overall Statistics

               Accuracy : 0.5586
                 95% CI : (0.5472, 0.5698)
    No Information Rate : 0.5586
    P-Value [Acc > NIR] : 0.5048

                  Kappa : 0
 Mcnemar's Test P-Value : NA
```