PUGB-Survival Analysis

Hai Long, Le

August 20, 2019

```r
library(sparklyr)
library(dplyr)
library(data.table)
library(ggplot2)
library(tidyverse)
library(magrittr)
library(survival)
library(survminer)
```

```r
rm(list=ls())
```

Connecting to Spark and Data Manipulation.

```r
sc <- spark_connect(master = "local")
```

Import Data.

```r
agg_match_stats_0 <- fread("D:/Github/PUBG-Survival/agg_match_stats_0.csv")
```

Import data using fread() is flashing and convient for this big dataset. The dataset has ~1.4M rows.

```r
agg1_tbl <- copy_to(sc, agg_match_stats_0[1:500000,], "agg1")
```

Only copy 500,000 observations to Spark.

```r
src_tbls(sc)
```

```
## [1] "agg1"
```

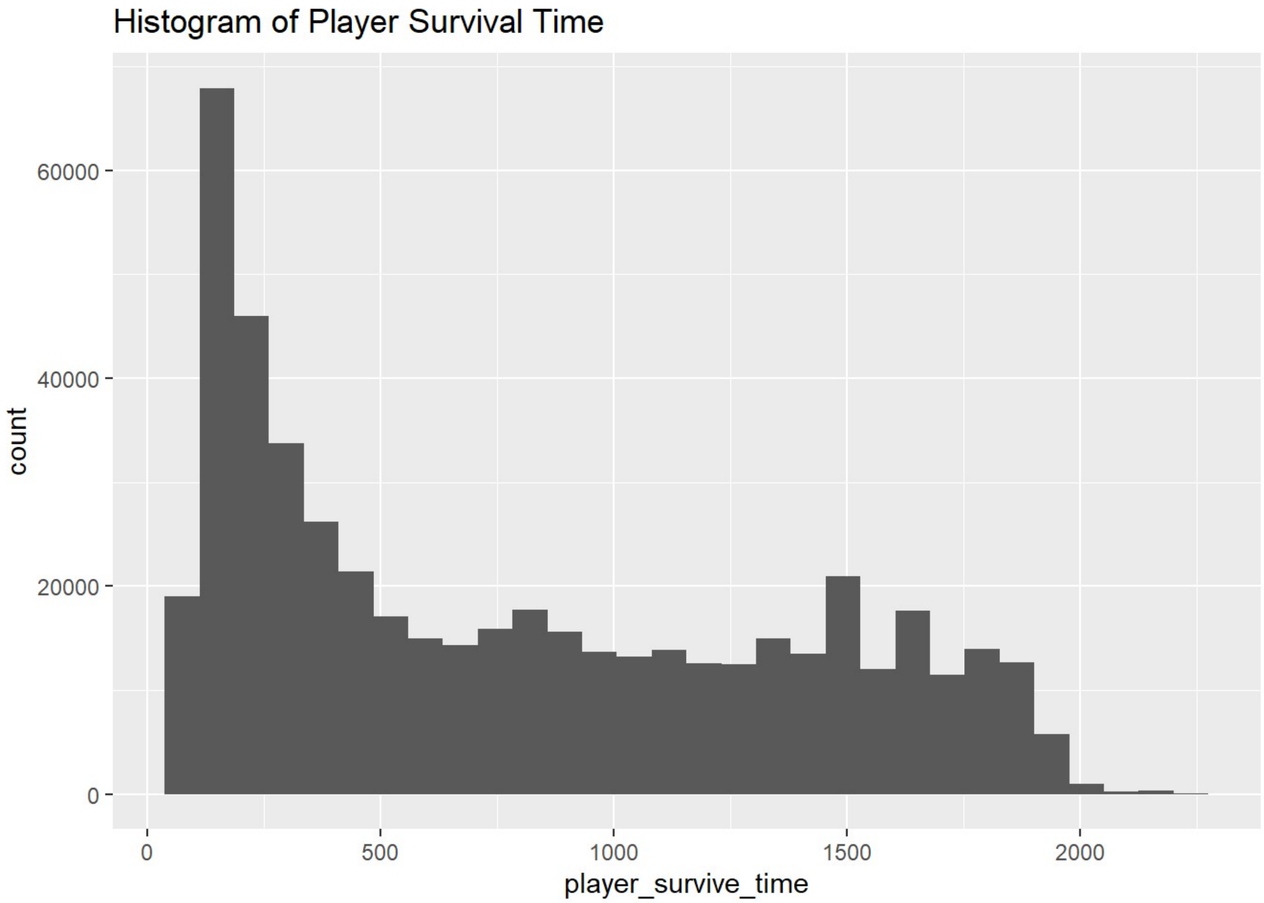
```r
glimpse(agg1_tbl)
```

```
## Observations: ??
## Variables: 15
## Database: spark_connection
## $ date               <chr> "2017-11-26T20:59:40+0000", "2017-11-26T20...
## $ game_size          <int> 37, 37, 37, 37, 37, 37, 37, 37, 37, 37, 37...
## $ match_id           <chr> "2U4GBNA0YmnNZYkzjkfgN4ev-hXSrak_BSey_YEG6...
## $ match_mode         <chr> "tpp", "tpp", "tpp", "tpp", "tpp", "tpp", ...
## $ party_size         <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...
## $ player_assists     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ player_dbno        <int> 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, ...
## $ player_dist_ride   <dbl> 2870.724, 2938.407, 0.000, 0.000, 2619.077...
## $ player_dist_walk   <dbl> 1784.84778, 1756.07971, 224.15756, 92.9351...
## $ player_dmg         <int> 117, 127, 67, 0, 175, 65, 0, 0, 79, 101, 5...
## $ player_kills       <int> 1, 1, 0, 0, 2, 0, 0, 0, 0, 1, 1, 0, 0, 2, ...
## $ player_name        <chr> "SnuffIes", "Ozon3r", "bovize", "sbahn87",...
## $ player_survive_time <dbl> 1106.320, 1106.315, 235.558, 197.553, 1537...
## $ team_id            <int> 4, 4, 5, 5, 14, 14, 15, 15, 17, 17, 22, 22...
## $ team_placement     <int> 18, 18, 33, 33, 11, 11, 17, 17, 24, 24, 4,...
```

Using glimpse() to have a brief overview at the dataframe.

```
ggplot(agg1_tbl, aes(x=player_survive_time)) +
  geom_histogram() + labs(title = "Histogram of Player Survival Time")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Using Histogram to learn more about distribution of *Player Survival Time*.

Focus on meaningful variables only.

```
agg1_tbl <- agg1_tbl %>%
                select("game_size", "party_size", "player_assists", "player_dbno", "player_dist_ride", "player_dist_walk",
                       "player_dmg", "player_kills", "player_survive_time", "team_placement") %>%
                collect
```

Feature Engineering

```
agg1_tbl <- agg1_tbl %>%
                mutate(drive.or.not = ifelse(player_dist_ride>0, 1, 0),
                       party.type = ifelse(party_size==1,"Solo",ifelse(party_size==2,"Dual","Team")),
                       gamesize =  ifelse(game_size<=30,"Small",ifelse(game_size>30&game_size<=60,"Medium","Big")))
```

Create 3 new Variable which are *drive.or.not*, *party.type*, *gamesize* from player_dist_ride, party_size. Basically, converting numerical variable to categorical variable.

drive.or.not tells if the player driving or not. *party.type* tells if the player playing Solo, Dual, or in Team. *gamesize* tells about the size of the match.

```
agg1_tbl$dist_walk <- arules::discretize(agg1_tbl$player_dist_walk, method = "interva
l", breaks = 5)
```

```
table(agg1_tbl$dist_walk)
```

```
##
##        [0,1.83e+05) [1.83e+05,3.65e+05) [3.65e+05,5.48e+05)
##              499980                   9                   5
##  [5.48e+05,7.3e+05)  [7.3e+05,9.13e+05]
##                   4                   2
```

Convert *player_dist_walk* from Continuous to Categorical Variable using **interval** - (equal interval width) with 5 levels. These 3 levels equivalent to Walking Distance Range. Majority of the players walk in distance range [0,1.83e+05).

```
agg1_tbl <- agg1_tbl %>% select(-"party_size", -"player_dist_walk", -"player_dist_rid
e",-"game_size")
```

```
agg1_tbl$drive.or.not %<>% as.factor
agg1_tbl$party.type %<>% as.factor
agg1_tbl$gamesize %<>% as.factor
```

Removing numerical variable because they are already converted to categorical variables.

```
agg1_tbl$status <- 1
```

Add the Status varible. The recorded data measures how long the players can survive in the game so the status of all observation should be equal to 1.

```
glimpse(agg1_tbl)
```

```
## Observations: 500,000
## Variables: 11
## $ player_assists      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ player_dbno         <int> 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, ...
## $ player_dmg          <int> 117, 127, 67, 0, 175, 65, 0, 0, 79, 101, 5...
## $ player_kills        <int> 1, 1, 0, 0, 2, 0, 0, 0, 0, 1, 1, 0, 0, 2, ...
## $ player_survive_time <dbl> 1106.320, 1106.315, 235.558, 197.553, 1537...
## $ team_placement      <int> 18, 18, 33, 33, 11, 11, 17, 17, 24, 24, 4,...
## $ drive.or.not        <fct> 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, ...
## $ party.type          <fct> Dual, Dual, Dual, Dual, Dual, Dual, Dual, ...
## $ gamesize            <fct> Medium, Medium, Medium, Medium, Medium, Me...
## $ dist_walk           <fct> "[0,1.83e+05)", "[0,1.83e+05)", "[0,1.83e+...
## $ status              <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
```

Take a brief look at the dataframe now and we are ready to perform Survival Analysis. There are 11 Variables with 500,000 observations. We want to learn how each Covariate impacts the player_survive_time.

Survival Analysis.

Survival Analysis is statistical techniques to analyze a “time to event outcome variable”. A Survival Object represents the time until a participant has an event of interest. Survival Analysis has wide range of application in Customer Analysis (Account Length), Reliability Engineering (Machine Failure), and Medical Research.

There are 3 types of analysis for the Survival Analysis will be shown. *Kaplan-Meier Estimate* will focus on how categorical varialbes impact the Survival Time. *Cox-proportional hazards (Cox-PH)* and *Accelerated Failure Time (AFT)* will show the Hazard Ratios of all Covariates.

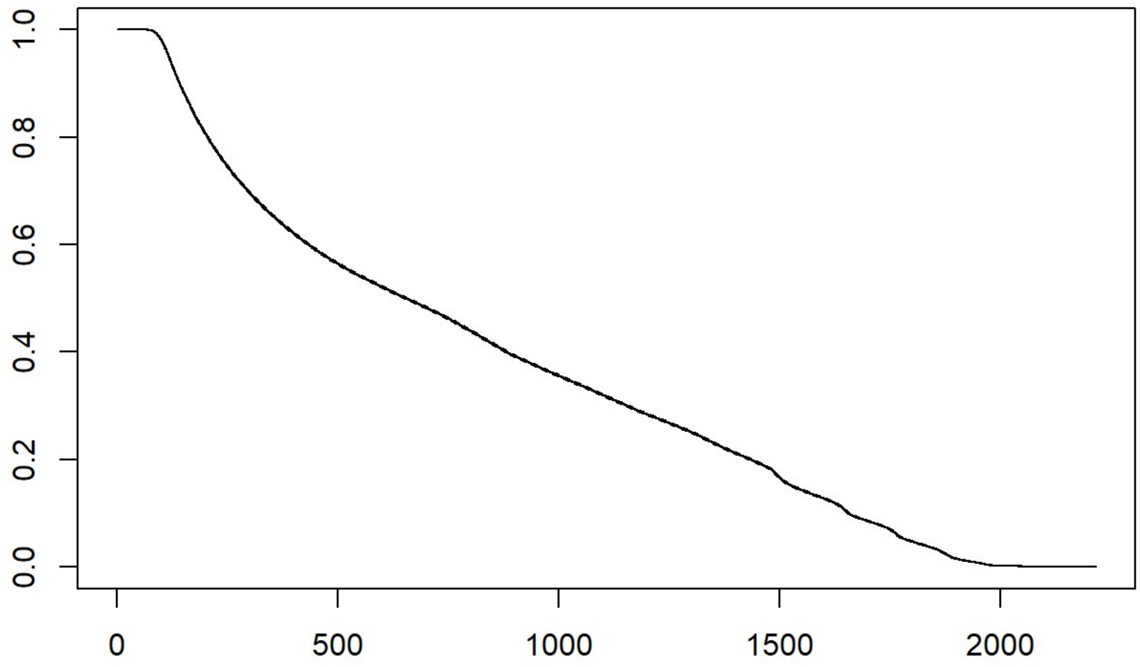
2.1 Kaplan-Meier Estimate.

2.1.1 KM Estimate.

```
surv_object <- Surv(time = agg1_tbl$player_survive_time, event = agg1_tbl$status)
```

```
KM.estimate <- survfit(surv_object~1,data = agg1_tbl)
```

```
plot(KM.estimate)
```

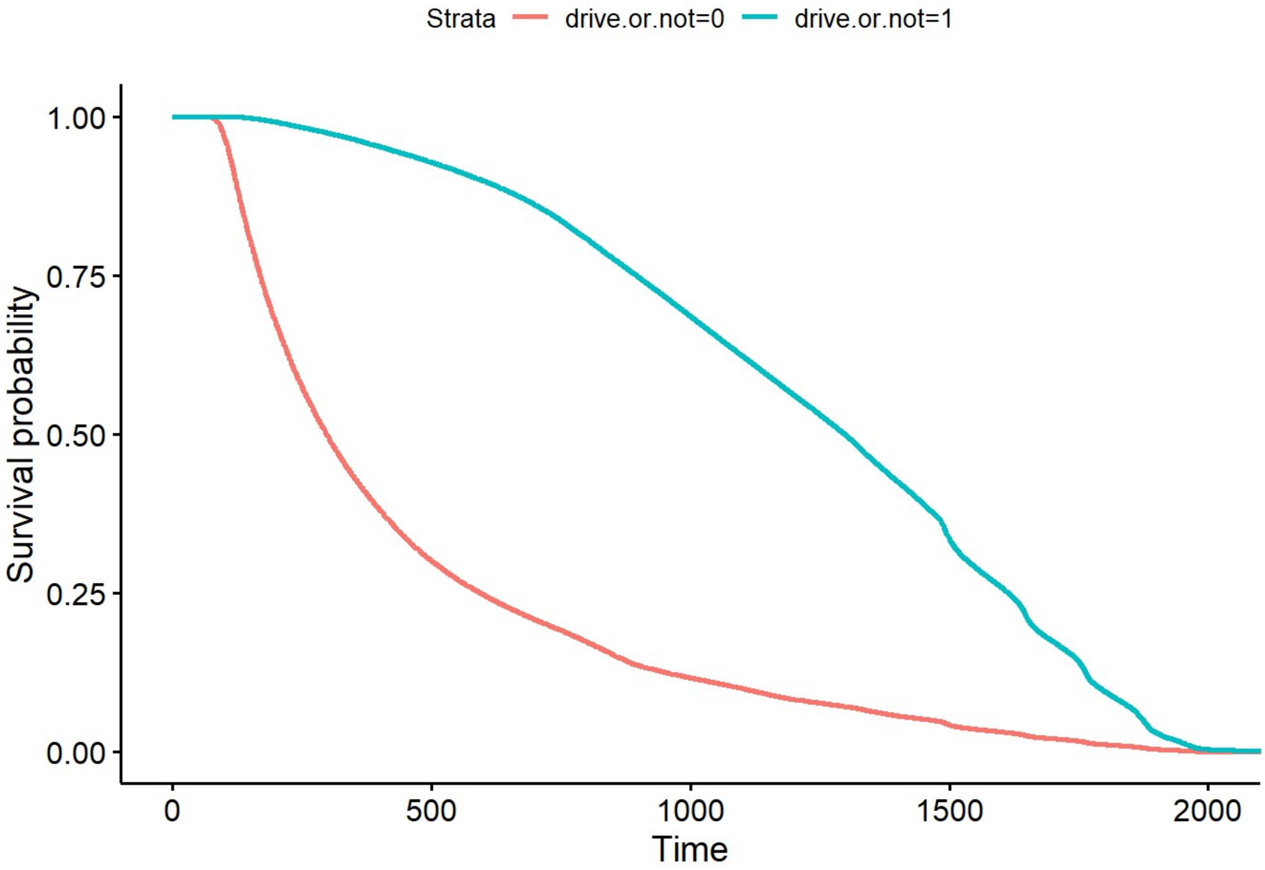


2.1.1 KM-Estimate Curve if Driving impacts Survival.

```
surv_object <- Surv(time = agg1_tbl$player_survive_time, event = agg1_tbl$status)
```

```
KM.estimate.driving <- survfit(surv_object~drive.or.not,data = agg1_tbl)
```

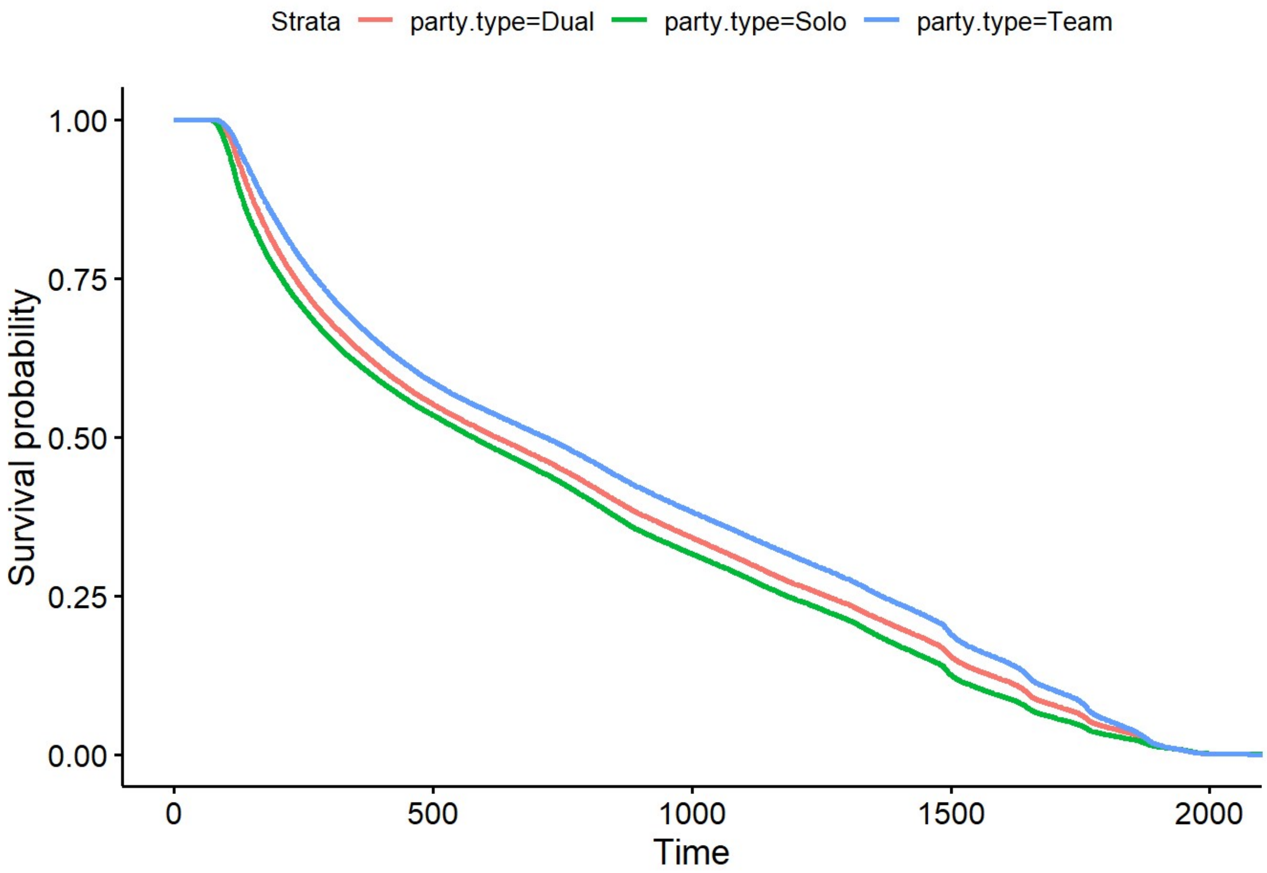
```
ggsurvplot(KM.estimate.driving, data = agg1_tbl)
```



2.1.2 How Party Size impacts the Survival in the game.

```
KM.estimate.party <- survfit(surv_object~party.type,data = agg1_tbl)
```

```
ggsurvplot(KM.estimate.party, data = agg1_tbl)
```

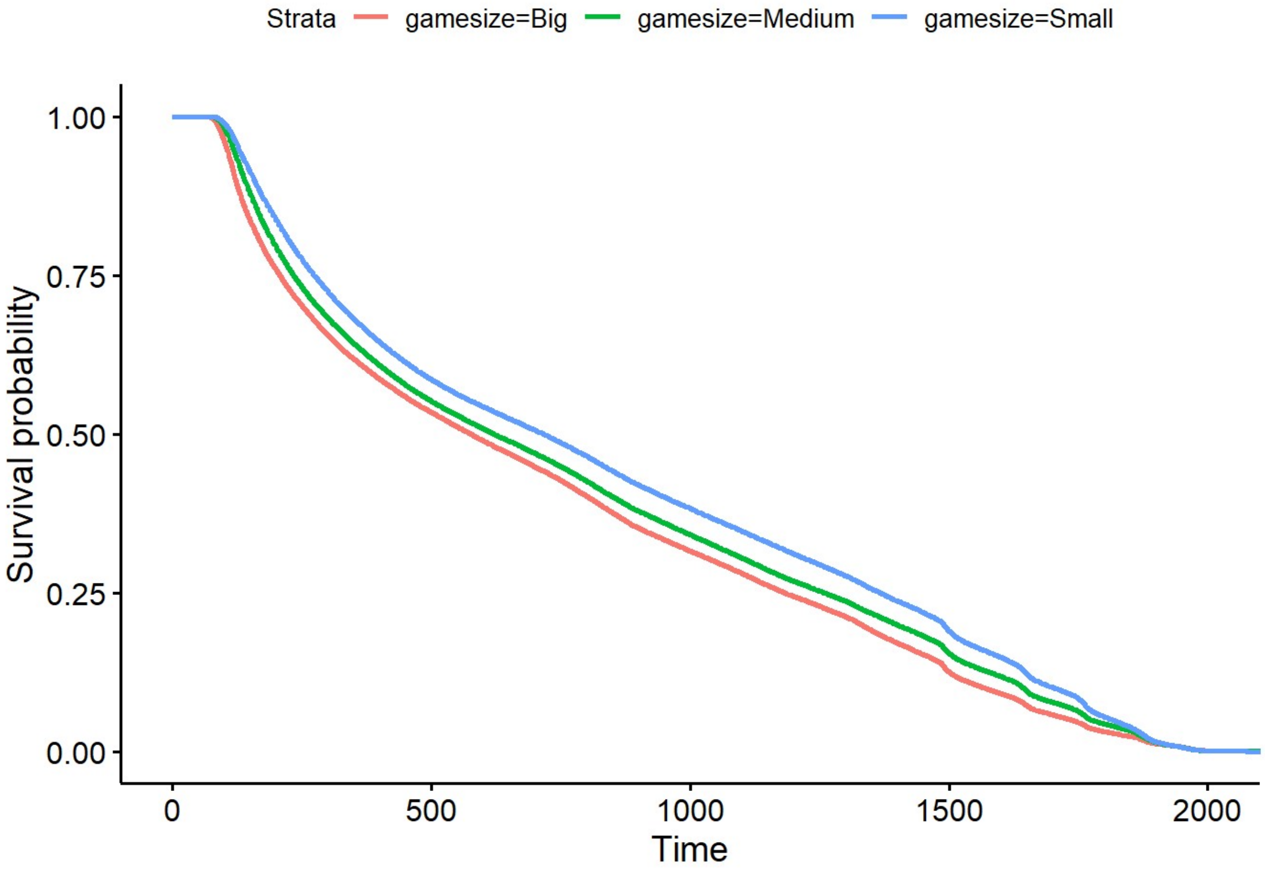


The Survival Curve based on KM-Estimate of all 3 types of Party Size are quite similar.

2.1.3 How Game Size impacts the Survival in the game.

```
KM.estimate.game <- survfit(surv_object~gamesize,data = agg1_tbl)
```

```
ggsurvplot(KM.estimate.game, data = agg1_tbl)
```



The Survival Curve based on KM-Estimate of all 3 types of Game Size are quite similar.

- The sameple size of Distance Riding is too small so we should not perform KM-Estimate Curve.

2.2 Cox-PH Model.

```
fit.cox <- coxph(formula = Surv(time = agg1_tbl$player_survive_time, event = agg1_tbl
$status) ~ ., data = agg1_tbl)
```

```
print(fit.cox)
```

```
## Call:
## coxph(formula = Surv(time = agg1_tbl$player_survive_time, event = agg1_tbl$status)
~
##     ., data = agg1_tbl)
##
##                                    coef  exp(coef)   se(coef)        z
## player_assists               -2.445e-01  7.831e-01  2.762e-03  -88.531
## player_dbno                  -9.503e-02  9.093e-01  2.358e-03  -40.297
## player_dmg                   -6.465e-05  9.999e-01  2.070e-05   -3.123
## player_kills                 -4.188e-02  9.590e-01  2.115e-03  -19.801
## team_placement                6.128e-02  1.063e+00  1.166e-04  525.544
## drive.or.not1                -7.606e-01  4.674e-01  3.486e-03 -218.167
## party.typeSolo               -7.381e-01  4.780e-01  1.079e-01   -6.841
## party.typeTeam                3.079e-01  1.361e+00  2.285e-02   13.473
## gamesizeMedium                6.936e-01  2.001e+00  1.079e-01    6.425
## gamesizeSmall                 7.673e-01  2.154e+00  1.098e-01    6.985
## dist_walk[1.83e+05,3.65e+05) -9.114e-01  4.019e-01  3.333e-01   -2.734
## dist_walk[3.65e+05,5.48e+05) -3.733e-03  9.963e-01  4.472e-01   -0.008
## dist_walk[5.48e+05,7.3e+05)  -4.555e-01  6.341e-01  5.000e-01   -0.911
## dist_walk[7.3e+05,9.13e+05]  -6.402e-01  5.272e-01  7.071e-01   -0.905
##                                     p
## player_assists                < 2e-16
## player_dbno                   < 2e-16
## player_dmg                    0.00179
## player_kills                  < 2e-16
## team_placement                < 2e-16
## drive.or.not1                 < 2e-16
## party.typeSolo               7.87e-12
## party.typeTeam                < 2e-16
## gamesizeMedium               1.32e-10
## gamesizeSmall                2.84e-12
## dist_walk[1.83e+05,3.65e+05)  0.00625
## dist_walk[3.65e+05,5.48e+05)  0.99334
## dist_walk[5.48e+05,7.3e+05)   0.36231
## dist_walk[7.3e+05,9.13e+05]   0.36531
##
## Likelihood ratio test=521061  on 14 df, p=< 2.2e-16
## n= 500000, number of events= 5e+05
```

If a Covariate has Coef < 0, that Covariate will decrease the Hazard function and increase the Survival function. If a Covariate has Coef > 0, that Covariate will increase the Hazard function and decrease the Survival function.

The Covariates have multiplicative relationship. The Exp(Coef) shows how much how many times the Hazard Ratio. If Exp(Coef) < 1, the Hazard Ratio is decrease, and vice versa. Hazard Ratio Plot below provides better visualization.

One unit increase in *player_assists*,*player_dbno*, *player_kills*, the Hazard Ratio is decrease by factor of

0.78, 0.91, 0.96, respectively.

One unit increase in *player_dmg*, the Hazard Ratio does not change because exp(player_dmg) = 1.0.

One unit increase in *team_placement*, the Hazard Ratio is increased by factor of 1.06.

Moving from Not Driving to Driving, the Hazard Ratio is decreased by factor of 0.47.

Moving from Dual to Solo, the Hazard Ratio is decreased by factor of 0.48.

Moving from Dual to Team, the Hazard Ratio is increase by factor of 1.36.

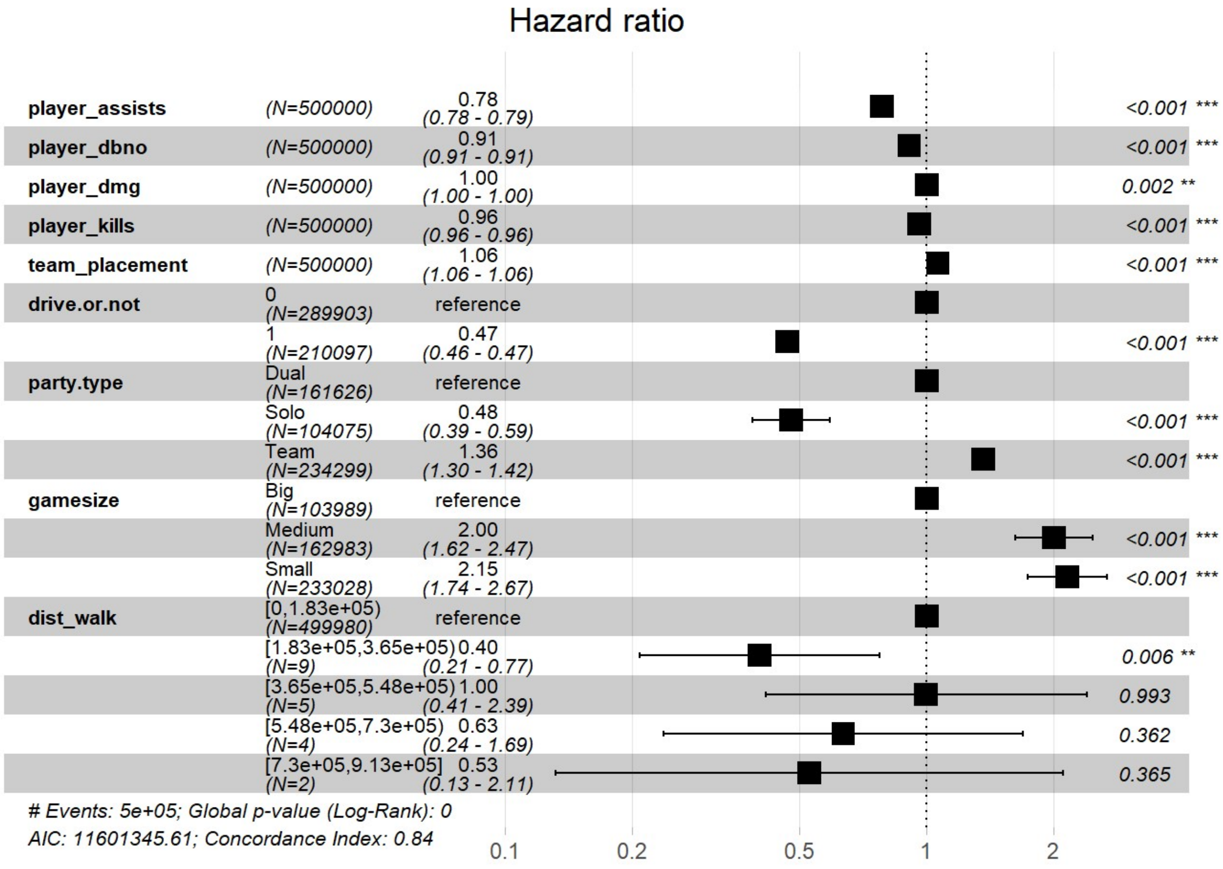
Moving from Big Game to Medium Game, the Hazard Ratio is increase by factor of 2.0.

Moving from Big Game to Small Game, the Hazard Ratio is increase by factor of 2.15.

The higher the Walking Distance, the lower the Hazard Ratio.

```
ggforest(fit.cox, data = agg1_tbl)
```

```
## Warning: Removed 4 rows containing missing values (geom_errorbar).
```



2.3 Accelerated Failure Time (AFT).

```
Fit.LogNormal <- survreg(formula = Surv(time = agg1_tbl$player_survive_time, event = a
gg1_tbl$status) ~ .,
                         data=agg1_tbl, dist = "lognormal")
```

```
Fit.LogNormal
```

```
## Call:
## survreg(formula = Surv(time = agg1_tbl$player_survive_time, event = agg1_tbl$statu
s) ~ 
##     ., data = agg1_tbl, dist = "lognormal")
## 
## Coefficients:
##                  (Intercept)               player_assists 
##                 7.3319765927                 0.1016749671 
##                  player_dbno                   player_dmg 
##                 0.0890300454                 0.0001059295 
##                 player_kills               team_placement 
##                -0.0203028776                -0.0341658459 
##                drive.or.not1               party.typeSolo 
##                 0.5928003433                 0.2932008047 
##               party.typeTeam               gamesizeMedium 
##                -0.2358030990                -0.5229604206 
##                gamesizeSmall dist_walk[1.83e+05,3.65e+05) 
##                -0.6081516124                 0.7454753248 
## dist_walk[3.65e+05,5.48e+05)  dist_walk[5.48e+05,7.3e+05) 
##                 0.3733941656                 0.3466073172 
##  dist_walk[7.3e+05,9.13e+05] 
##                 0.4455577296 
## 
## Scale= 0.4939679 
## 
## Loglik(model)= -3512903   Loglik(intercept only)= -3822767
##  Chisq= 619727.5 on 14 degrees of freedom, p= <2e-16 
## n= 500000
```

```
summary(Fit.LogNormal)
```

```
##
## Call:
## survreg(formula = Surv(time = agg1_tbl$player_survive_time, event = agg1_tbl$statu
s) ~
##     ., data = agg1_tbl, dist = "lognormal")
##                                  Value Std. Error       z      p
## (Intercept)                   7.33e+00   5.35e-02  137.04 <2e-16
## player_assists                1.02e-01   1.33e-03   76.53 <2e-16
## player_dbno                   8.90e-02   1.07e-03   82.87 <2e-16
## player_dmg                    1.06e-04   1.06e-05   10.03 <2e-16
## player_kills                 -2.03e-02   1.07e-03  -19.01 <2e-16
## team_placement               -3.42e-02   5.60e-05 -609.76 <2e-16
## drive.or.not1                 5.93e-01   1.70e-03  349.41 <2e-16
## party.typeSolo                2.93e-01   5.34e-02    5.49  4e-08
## party.typeTeam               -2.36e-01   1.07e-02  -21.98 <2e-16
## gamesizeMedium               -5.23e-01   5.34e-02   -9.79 <2e-16
## gamesizeSmall                -6.08e-01   5.38e-02  -11.31 <2e-16
## dist_walk[1.83e+05,3.65e+05)  7.45e-01   1.65e-01    4.53  6e-06
## dist_walk[3.65e+05,5.48e+05)  3.73e-01   2.21e-01    1.69  0.091
## dist_walk[5.48e+05,7.3e+05)   3.47e-01   2.47e-01    1.40  0.161
## dist_walk[7.3e+05,9.13e+05]   4.46e-01   3.49e-01    1.28  0.202
## Log(scale)                   -7.05e-01   1.00e-03 -705.28 <2e-16
##
## Scale= 0.494
##
## Log Normal distribution
## Loglik(model)= -3512903   Loglik(intercept only)= -3822767
##  Chisq= 619727.5 on 14 degrees of freedom, p= 0
## Number of Newton-Raphson Iterations: 5
## n= 500000
```

Disconnecting to Spark.

```
spark_disconnect_all()
```

```
## [1] 1
```

Reference:

https://www.kaggle.com/datark1/pubg-survival-analysis-kaplan-meier/data (https://www.kaggle.com/datark1/pubg-survival-analysis-kaplan-meier/data)

https://spark.rstudio.com/ (https://spark.rstudio.com/)