

emory2 EDA

2025-07-02

```
library("readr")
library("dplyr")
library("ggplot2")
library("readr")
library("stringr")
library("glue")
```

```
g <- params$category
singular_game <- readr::read_csv(glue("Desktop/SURA project code/extended_cmu_data/extended_cmu_data_",
```

```
## New names:
## Rows: 20 Columns: 22
## -- Column specification
## ----- Delimiter: "," chr
## (1): LINEUP (NAMES) dbl (20): ...1, NUMBER OF GUARDS, OPPONENT POSSESSIONS, CMU
## POSSESSIONS, OP... time (1): LINEUP MINUTES
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
# if negatives in any columns (specifically had problem in possession column)
for (colName in colnames(singular_game)){
  singular_game[[colName]][singular_game[[colName]] < 0] <- 0
}
```

```
singular_game$`LINEUP MINUTES` <- sapply(singular_game$`LINEUP MINUTES`, function(t){
  parts <- as.integer(strsplit(as.character(t),":")[[1]])
  parts[1]*60 + parts[2]
})
```

```
singular_game <- singular_game %>% rename('LINEUP SECONDS' = `LINEUP MINUTES`) %>% mutate(LINEUP_SORTED =
  if (is.na(1)) return(NA)
  paste(sort(strsplit(1, ", ")[1]), collapse = " "))
}))
```

```
game <- singular_game %>% group_by(`LINEUP_SORTED`) %>% summarise(
  `NUMBER OF GUARDS` = mean(`NUMBER OF GUARDS`),
  `OPPONENT POSSESSIONS` = sum(`OPPONENT POSSESSIONS`, na.rm = TRUE),
  `CMU POSSESSIONS` = sum(`CMU POSSESSIONS`, na.rm = TRUE),
  `LINEUP SECONDS` = sum(`LINEUP SECONDS`, na.rm = TRUE),
  `OPPONENT PTS` = sum(`OPPONENT PTS`, na.rm = TRUE),
  `CMU PTS` = sum(`CMU PTS`, na.rm = TRUE),
  `CMU 3PA` = sum(`CMU 3PA`, na.rm = TRUE),
  `CMU FGA` = sum(`CMU FGA`, na.rm = TRUE),
  `CMU FTA` = sum(`CMU FTA`, na.rm = TRUE),
  `CMU REBOUNDS` = sum(`CMU REBOUNDS`, na.rm = TRUE),
  `TOTAL REBOUNDS` = sum(`TOTAL REBOUNDS`, na.rm = TRUE),
```

```

`SCORE DIFFERENTIAL WHEN ENTER` = paste(`SCORE DIFFERENTIAL WHEN ENTER`, collapse = ", "),
`QUARTER` = paste(`QUARTER`, collapse = ", ")
) %>%mutate(`PACE` = 40 * ((`CMU POSSESSIONS` + `OPPONENT POSSESSIONS`) / (2 * `LINEUP SECONDS`/60)),
`OFFENSIVE RATING` = 100 * (`CMU PTS` / `CMU POSSESSIONS`),
`DEFENSIVE RATING` = 100 * (`OPPONENT PTS` / `OPPONENT POSSESSIONS`),
`NET RATING` = `OFFENSIVE RATING` - `DEFENSIVE RATING`,
`3PA/FGA` = `CMU 3PA` / `CMU FGA`,
`TRUE SHOOTING %` = 100 * (`CMU PTS` / ( 2 * (`CMU FGA` + (0.44* `CMU FTA`)))),
`TRB%` = 100 * (`CMU REBOUNDS` / `TOTAL REBOUNDS`)

```

```

# see where to score differential cut off time -> SHOULD DO THIS AFTER OR BEFORE CUT SCRAP MINUTES?
l <- quantile(singular_game$`SCORE DIFFERENTIAL WHEN ENTER`,probs=c(0.1))
u <- quantile(singular_game$`SCORE DIFFERENTIAL WHEN ENTER`,probs=c(0.9))

```

```
l
```

```
## 10%
```

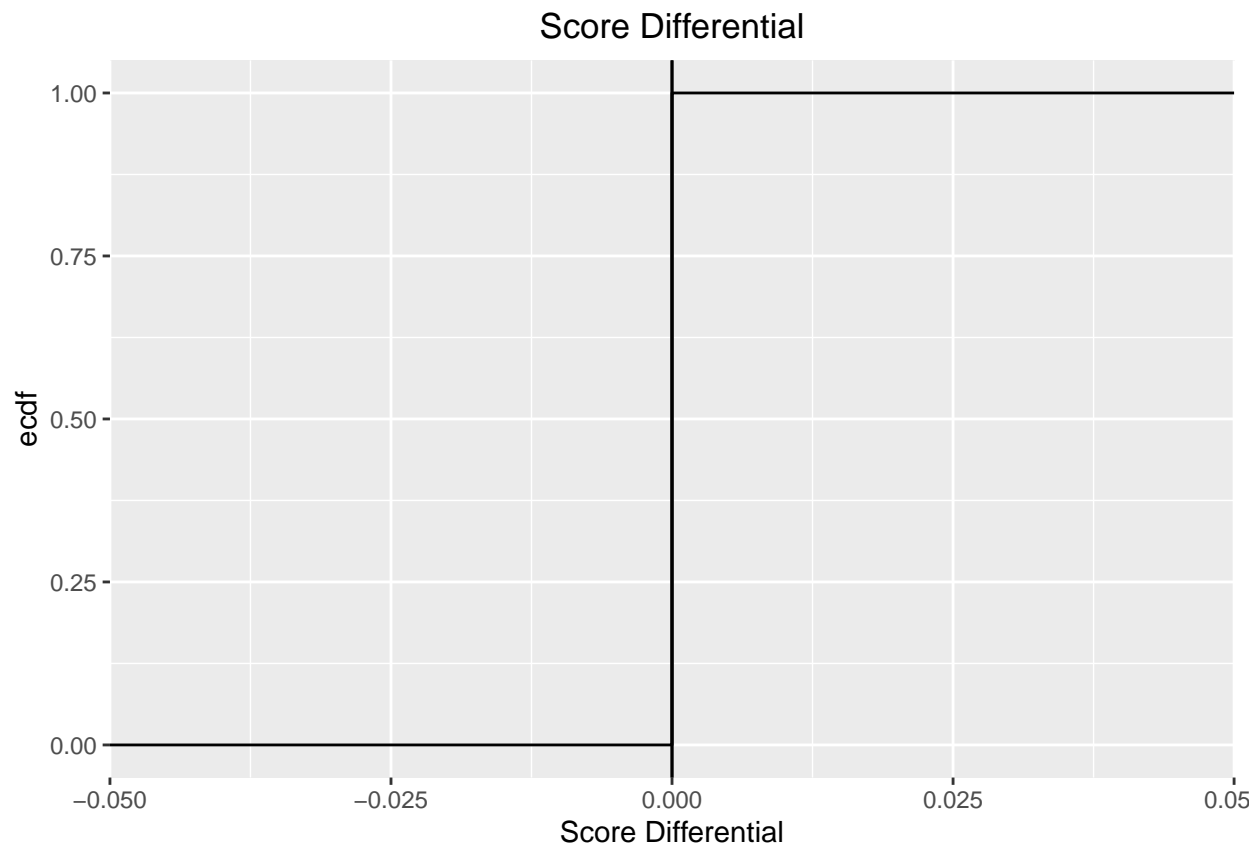
```
## 0
```

```
u
```

```
## 90%
```

```
## 0
```

```
ggplot(singular_game, aes(x = `SCORE DIFFERENTIAL WHEN ENTER`)) + stat_ecdf() + geom_vline(xintercept =
```



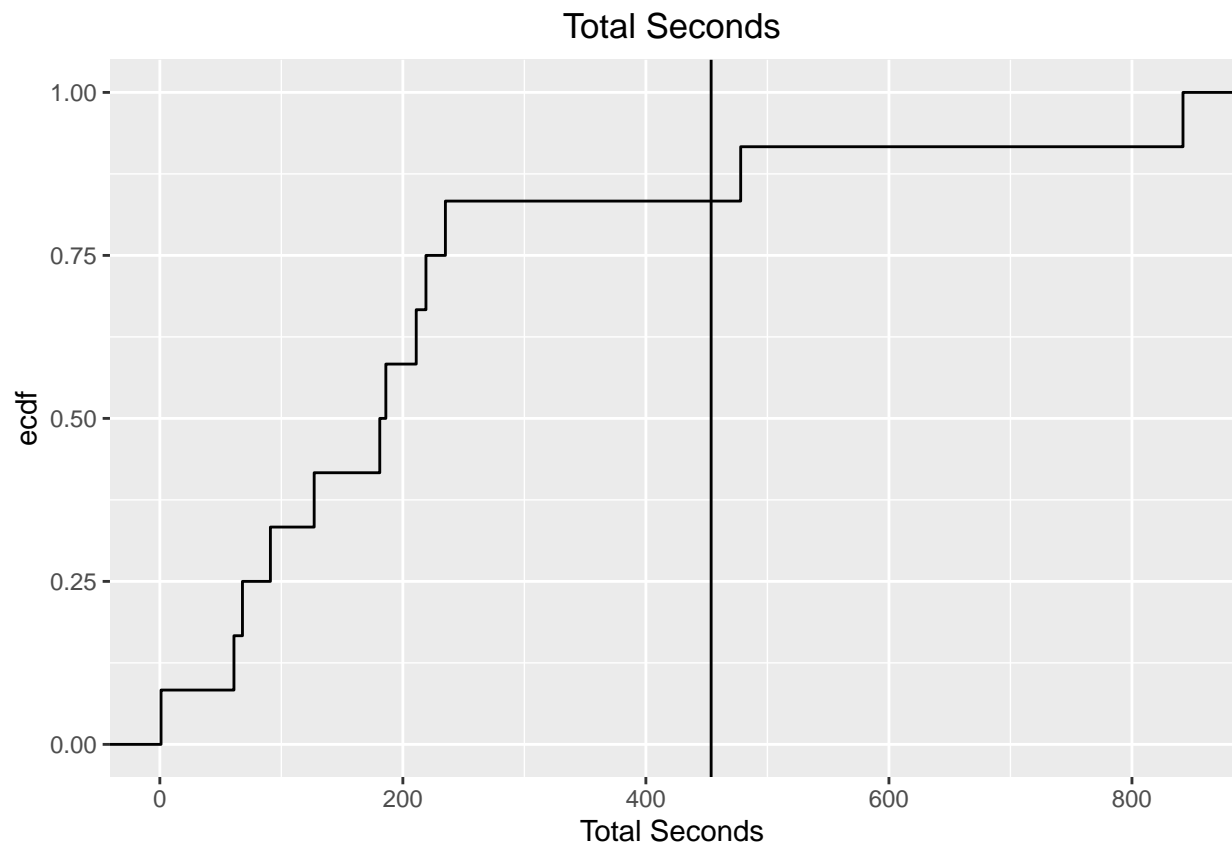
```
game <- subset(game, !((`SCORE DIFFERENTIAL WHEN ENTER` <= -11 | `SCORE DIFFERENTIAL WHEN ENTER` >= 15))
```

```

# see where to cut time -> SHOULD DO THIS AFTER OR BEFORE CUT SCRAP MINUTES?

```

```
p <- quantile(game$`LINEUP SECONDS`, probs=c(0.9))
ggplot(game, aes(x = `LINEUP SECONDS`)) + stat_ecdf() + geom_vline(xintercept = p) + labs(title = "Total
```



```
#game <- subset(game, `LINEUP SECONDS` >= p)
```

```
p
```

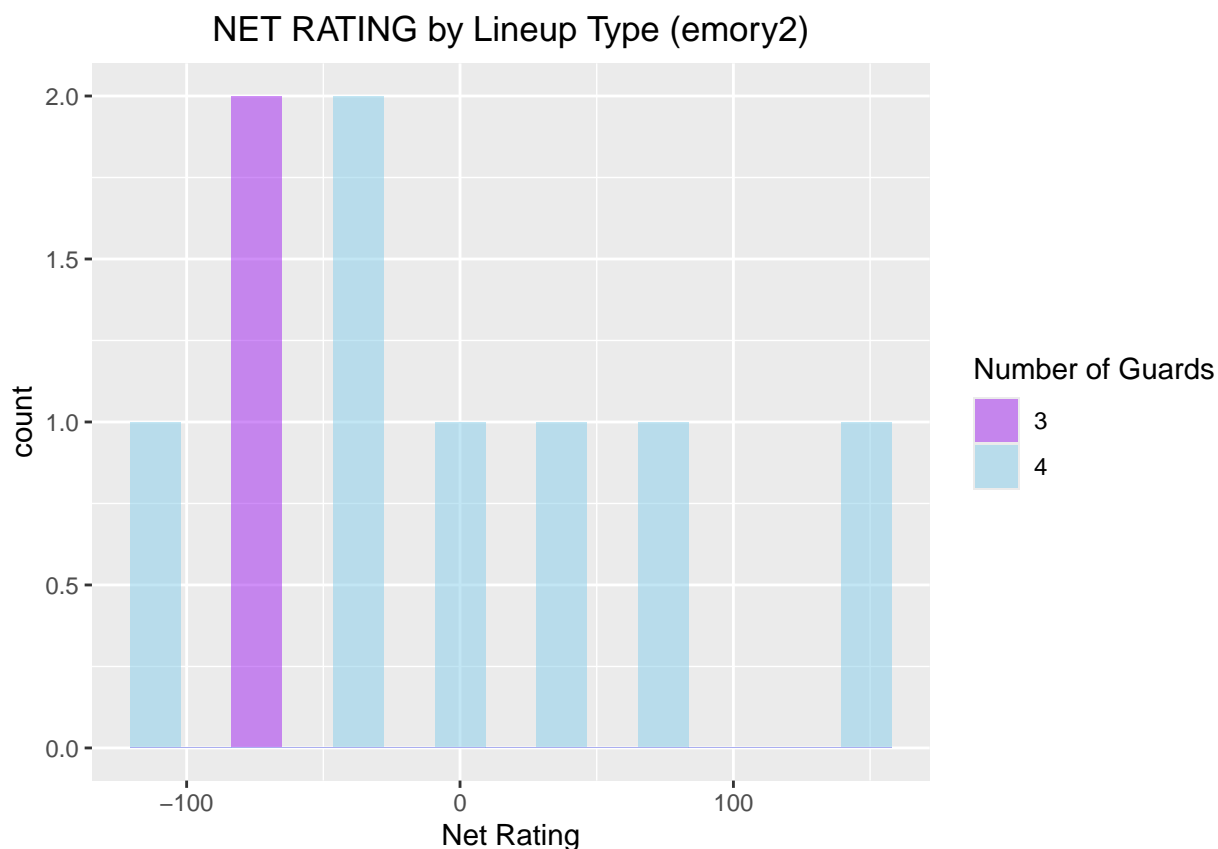
```
## 90%
```

```
## 453.7
```

```
#pdf(file = glue("Desktop/SURA project code/sing_game_EDA/{g}_plot.pdf"), width = 6, height = 5)
```

```
t_f <- c("3", "4")
```

```
ggplot(data = subset(game, subset = `NUMBER OF GUARDS` %in% t_f), aes(x = `NET RATING`, fill = factor(`
```



```

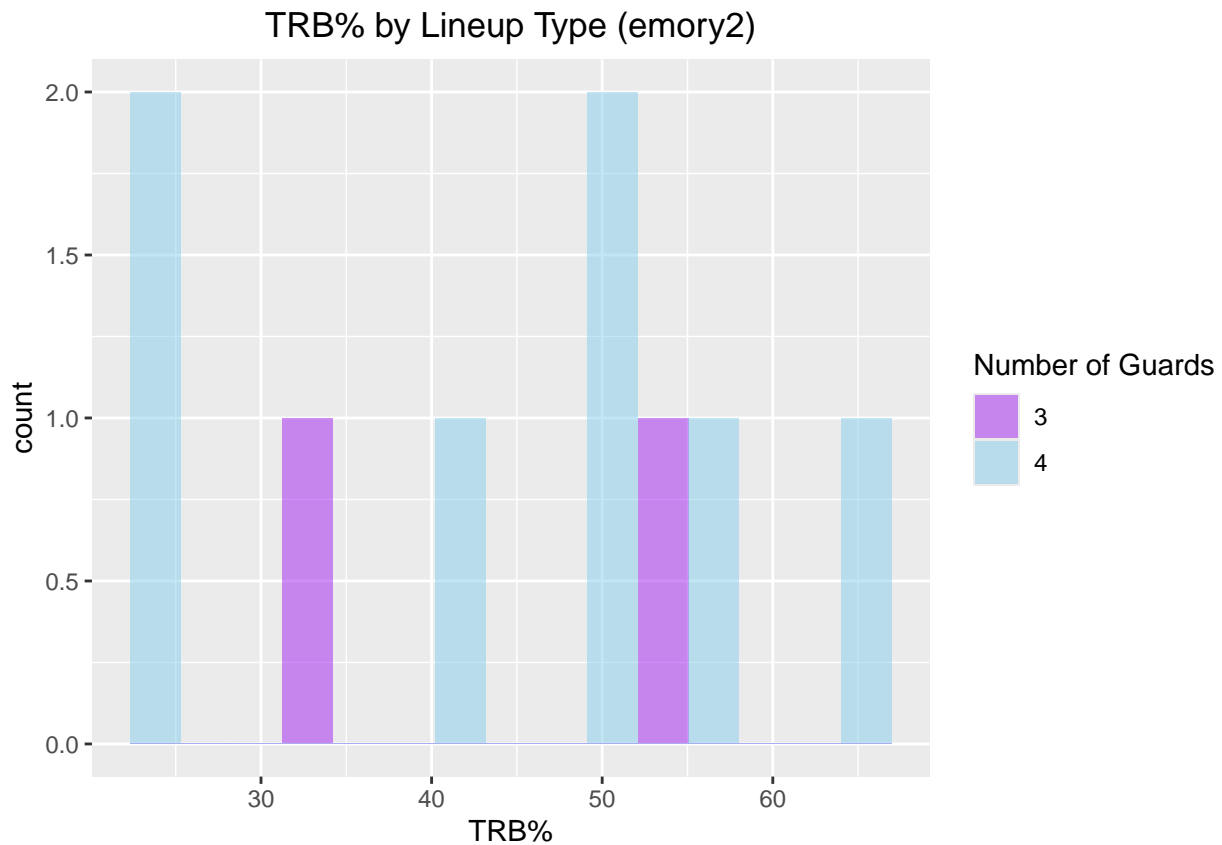
tapply(game$`NET RATING`[game$`NUMBER OF GUARDS` %in% t_f], game$`NUMBER OF GUARDS`[game$`NUMBER OF GUARDS` %in% t_f],
  FUN = function(x) {
    ## $`3`
    ##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
    ## -78.75 -75.73  -72.71  -72.71  -69.69  -66.67
    ##
    ## $`4`
    ##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
    ## -110.00 -36.67    0.00   12.07   58.92   150.00
  },
  exact = TRUE)

wilcox.test(`NET RATING` ~ `NUMBER OF GUARDS`, data = subset(game, `NUMBER OF GUARDS` %in% t_f), exact = TRUE)

##
## Wilcoxon rank sum test with continuity correction
##
## data:  NET RATING by NUMBER OF GUARDS
## W = 2, p-value = 0.1877
## alternative hypothesis: true location shift is not equal to 0

ggplot(data = subset(game, subset = `NUMBER OF GUARDS` %in% t_f), aes(x = `TRB%`, fill = factor(`NUMBER OF GUARDS`)))

```



```
tapply(game$`TRB%`, [game$`NUMBER OF GUARDS` %in% t_f], game$`NUMBER OF GUARDS` [game$`NUMBER OF GUARDS` %in% t_f])
```

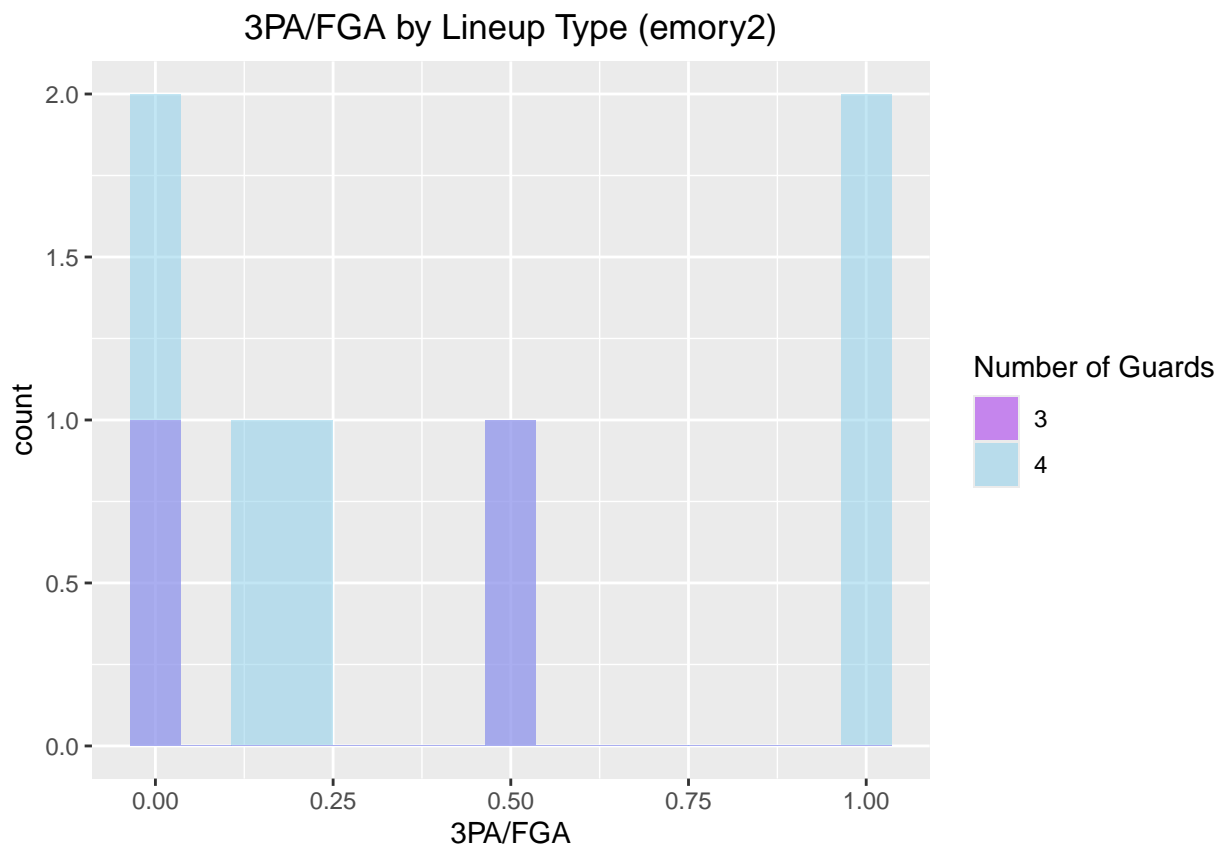
```
## $`3`
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  33.33  38.33   43.33   43.33  48.33   53.33
##
```

```
## $`4`
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  25.00  33.93   50.00   45.01  52.78   66.67
```

```
wilcox.test(`TRB%` ~ `NUMBER OF GUARDS`, data = subset(game, `NUMBER OF GUARDS` %in% t_f), exact = FALSE)
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data:  TRB% by NUMBER OF GUARDS
## W = 7, p-value = 1
## alternative hypothesis: true location shift is not equal to 0
```

```
ggplot(data = subset(game, subset = `NUMBER OF GUARDS` %in% t_f), aes(x = `3PA/FGA`, fill = factor(`NUMBER OF GUARDS`)))
```



```
tapply(game$`3PA/FGA`[game$`NUMBER OF GUARDS` %in% t_f], game$`NUMBER OF GUARDS`[game$`NUMBER OF GUARDS`
```

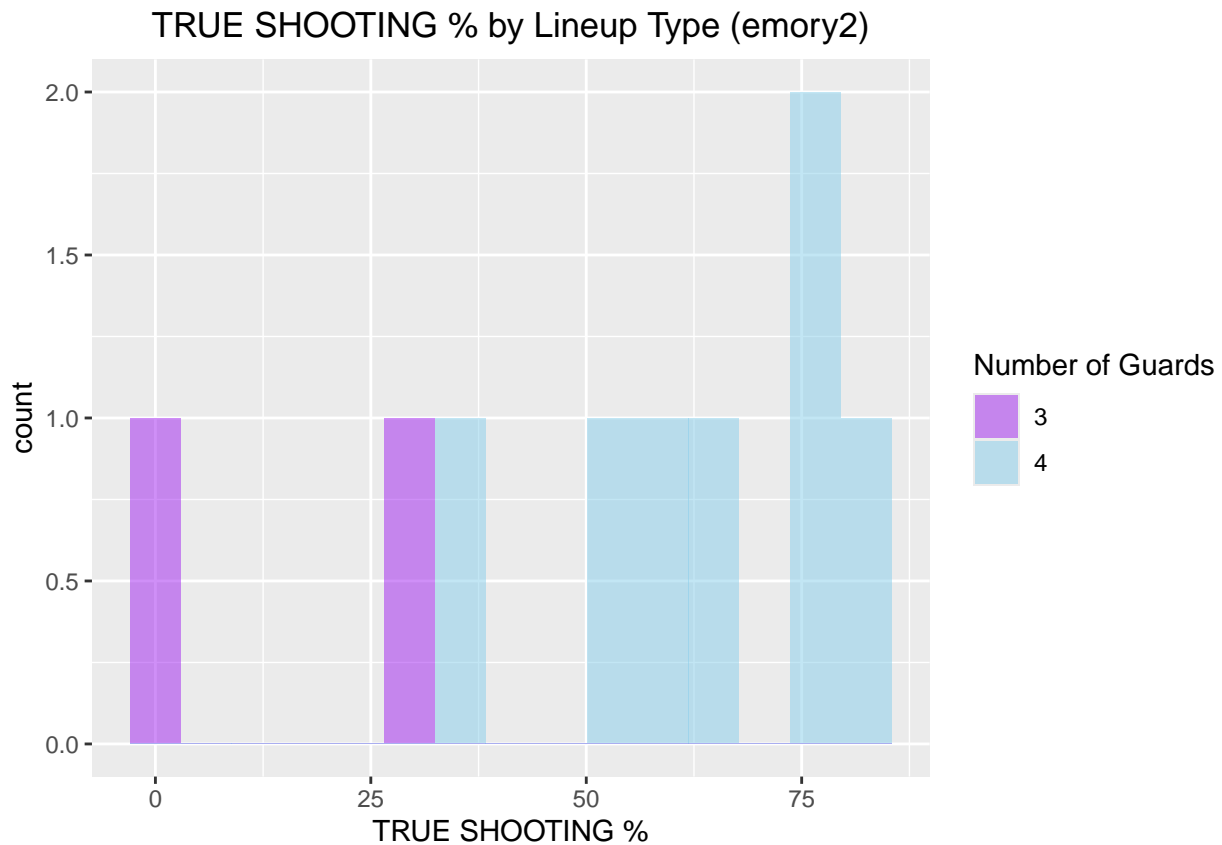
```
## $`3`
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.000  0.125   0.250   0.250  0.375   0.500
##
```

```
## $`4`
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.07143 0.20000 0.40612 0.75000 1.00000
```

```
wilcox.test(`3PA/FGA` ~ `NUMBER OF GUARDS`, data = subset(game, `NUMBER OF GUARDS` %in% t_f), exact = F
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data: 3PA/FGA by NUMBER OF GUARDS
## W = 5.5, p-value = 0.7639
## alternative hypothesis: true location shift is not equal to 0
```

```
ggplot(data = subset(game, subset = `NUMBER OF GUARDS` %in% t_f), aes(x = `TRUE SHOOTING %`, fill = fac
```



```
tapply(game$TRUE SHOOTING % [game$NUMBER OF GUARDS %in% t_f], game$NUMBER OF GUARDS [game$NUMBER OF GUARDS %in% t_f], FUN = function(x) {
  summary(x)
})
```

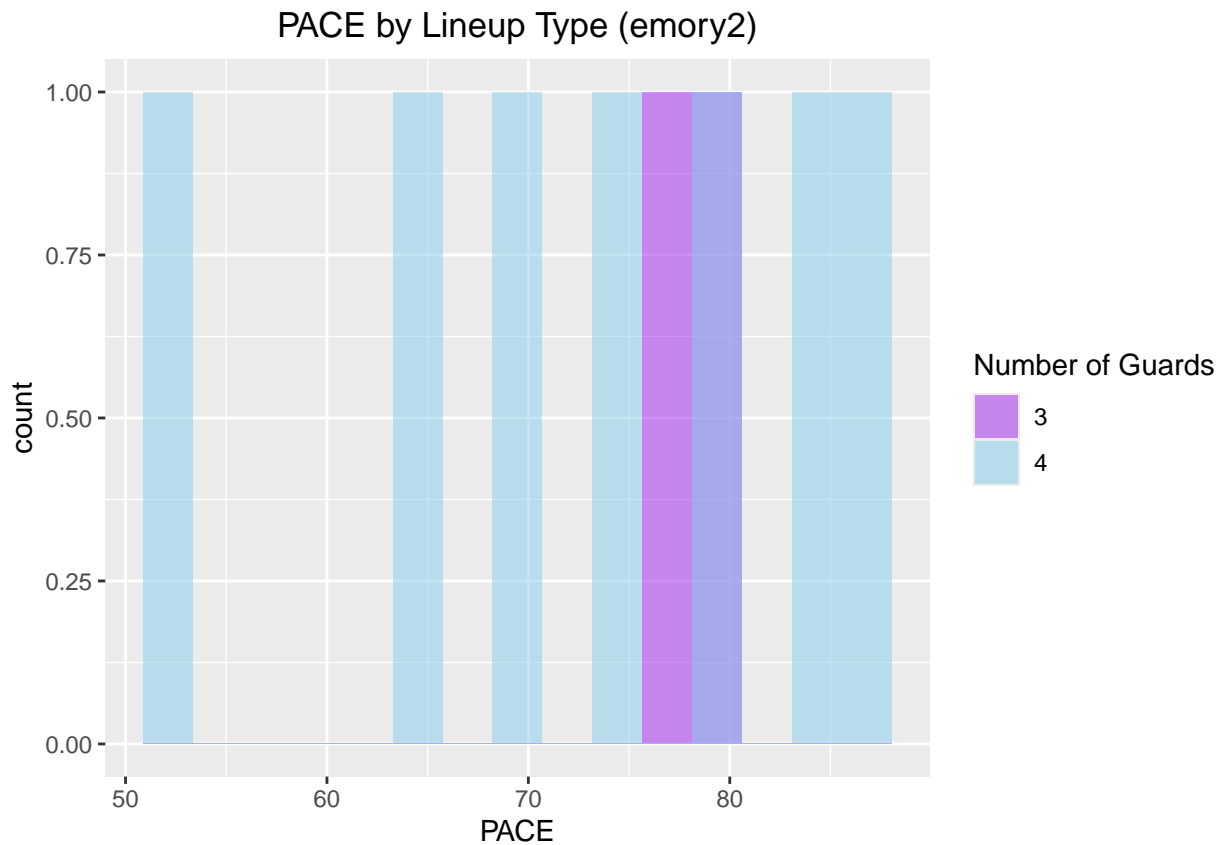
```
## $`3`
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   7.684   15.369   15.369   23.053   30.738
##
```

```
## $`4`
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   33.33   54.11   66.67   62.96   75.00   82.49
```

```
wilcox.test(`TRUE SHOOTING %` ~ `NUMBER OF GUARDS`, data = subset(game, `NUMBER OF GUARDS` %in% t_f), exact = FALSE)
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data: TRUE SHOOTING % by NUMBER OF GUARDS
## W = 0, p-value = 0.05601
## alternative hypothesis: true location shift is not equal to 0
```

```
ggplot(data = subset(game, subset = `NUMBER OF GUARDS` %in% t_f), aes(x = `PACE`, fill = factor(`NUMBER OF GUARDS`)))
```



```
tapply(game$`PACE`[game$`NUMBER OF GUARDS` %in% t_f], game$`NUMBER OF GUARDS`[game$`NUMBER OF GUARDS` %in% t_f],
```

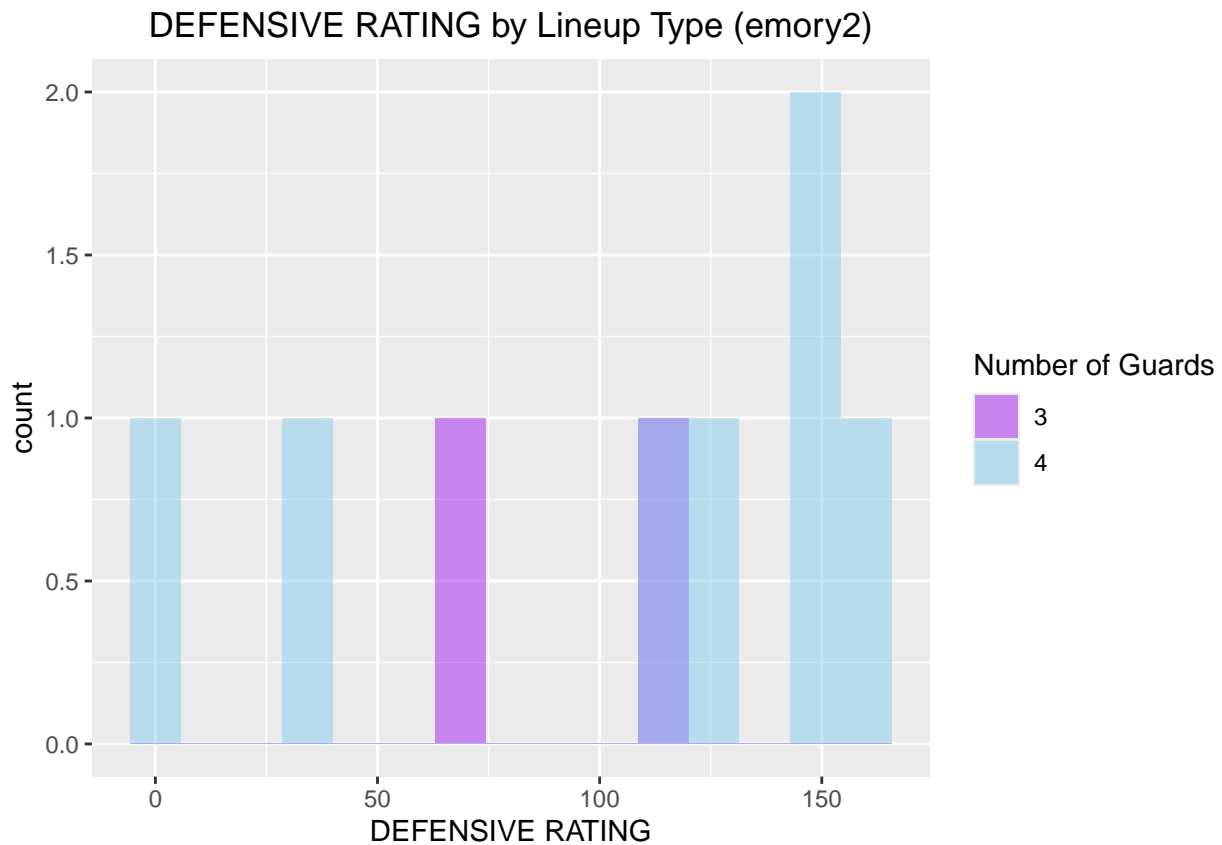
```
## $`3`
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  77.82  78.15   78.47   78.47  78.80   79.12
##
```

```
## $`4`
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  52.94  66.38   75.53   73.23  81.86   87.67
```

```
wilcox.test(`PACE` ~ `NUMBER OF GUARDS`, data = subset(game, `NUMBER OF GUARDS` %in% t_f), exact = FALSE)
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data: PACE by NUMBER OF GUARDS
## W = 9, p-value = 0.6605
## alternative hypothesis: true location shift is not equal to 0
```

```
ggplot(data = subset(game, subset = `NUMBER OF GUARDS` %in% t_f), aes(x = `DEFENSIVE RATING`, fill = fa
```

```
tapply(game$`DEFENSIVE RATING`[game$`NUMBER OF GUARDS` %in% t_f], game$`NUMBER OF GUARDS`[game$`NUMBER OF GUARDS` %in% t_f], FUN = function(x) {
  summary(x)
})
```

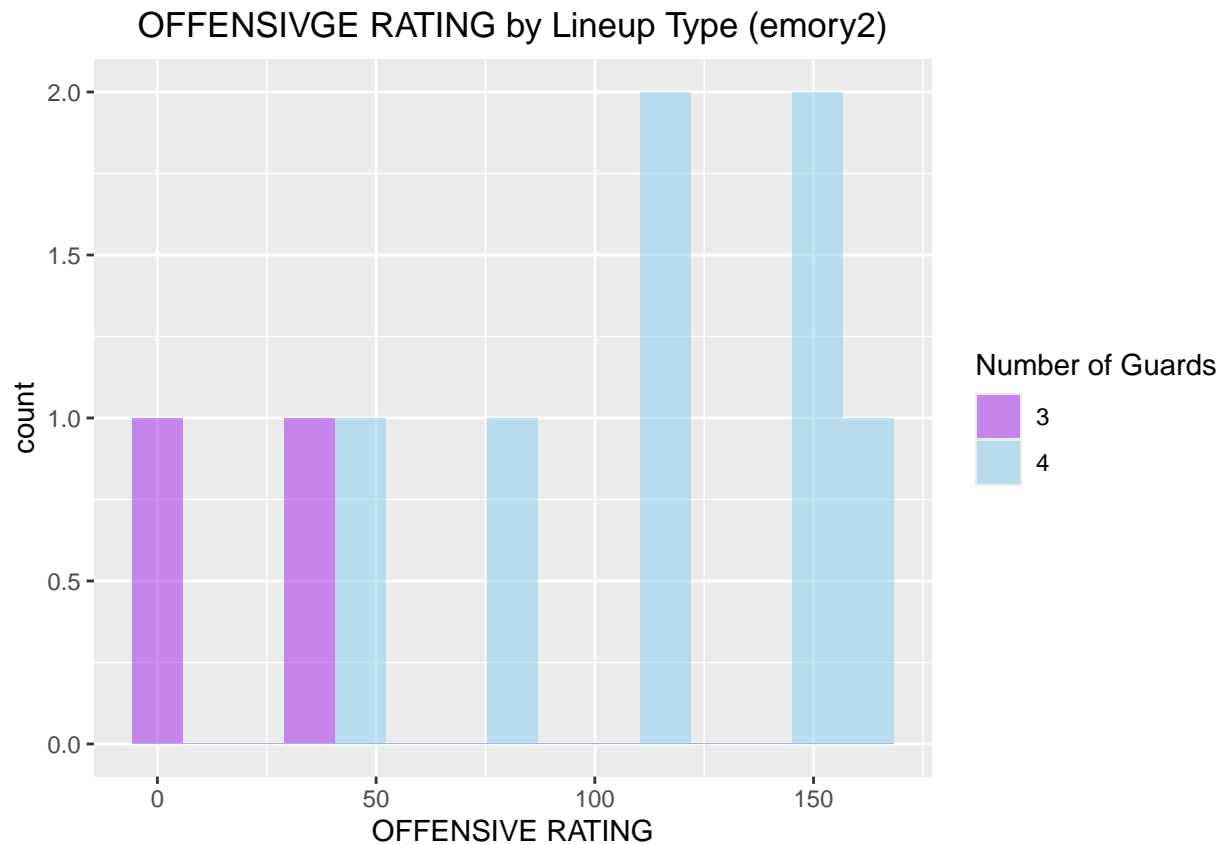
```
## $`3`
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  66.67   79.69   92.71   92.71  105.73  118.75
##
```

```
## $`4`
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   75.38  125.00  105.11  150.00  160.00
```

```
wilcox.test(`DEFENSIVE RATING` ~ `NUMBER OF GUARDS`, data = subset(game, `NUMBER OF GUARDS` %in% t_f), c
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data: DEFENSIVE RATING by NUMBER OF GUARDS
## W = 4, p-value = 0.4623
## alternative hypothesis: true location shift is not equal to 0
```

```
ggplot(data = subset(game, subset = `NUMBER OF GUARDS` %in% t_f), aes(x = `OFFENSIVE RATING`, fill = fa
```



```
tapply(game$`OFFENSIVE RATING`[game$`NUMBER OF GUARDS` %in% t_f], game$`NUMBER OF GUARDS`[game$`NUMBER OF GUARDS` %in% t_f], FUN = function(x) {
  summary(x)
})
```

```
## $`3`
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##         0      10      20      20      30      40
##
## $`4`
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    50.00  95.56  116.67  117.18  150.00  162.50
```

```
wilcox.test(`OFFENSIVE RATING` ~ `NUMBER OF GUARDS`, data = subset(game, `NUMBER OF GUARDS` %in% t_f), c("n", "p"))
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data:  OFFENSIVE RATING by NUMBER OF GUARDS
## W = 0, p-value = 0.05601
## alternative hypothesis: true location shift is not equal to 0
```

```
#dev.off()
```