

# washu2 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: 27 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%
```

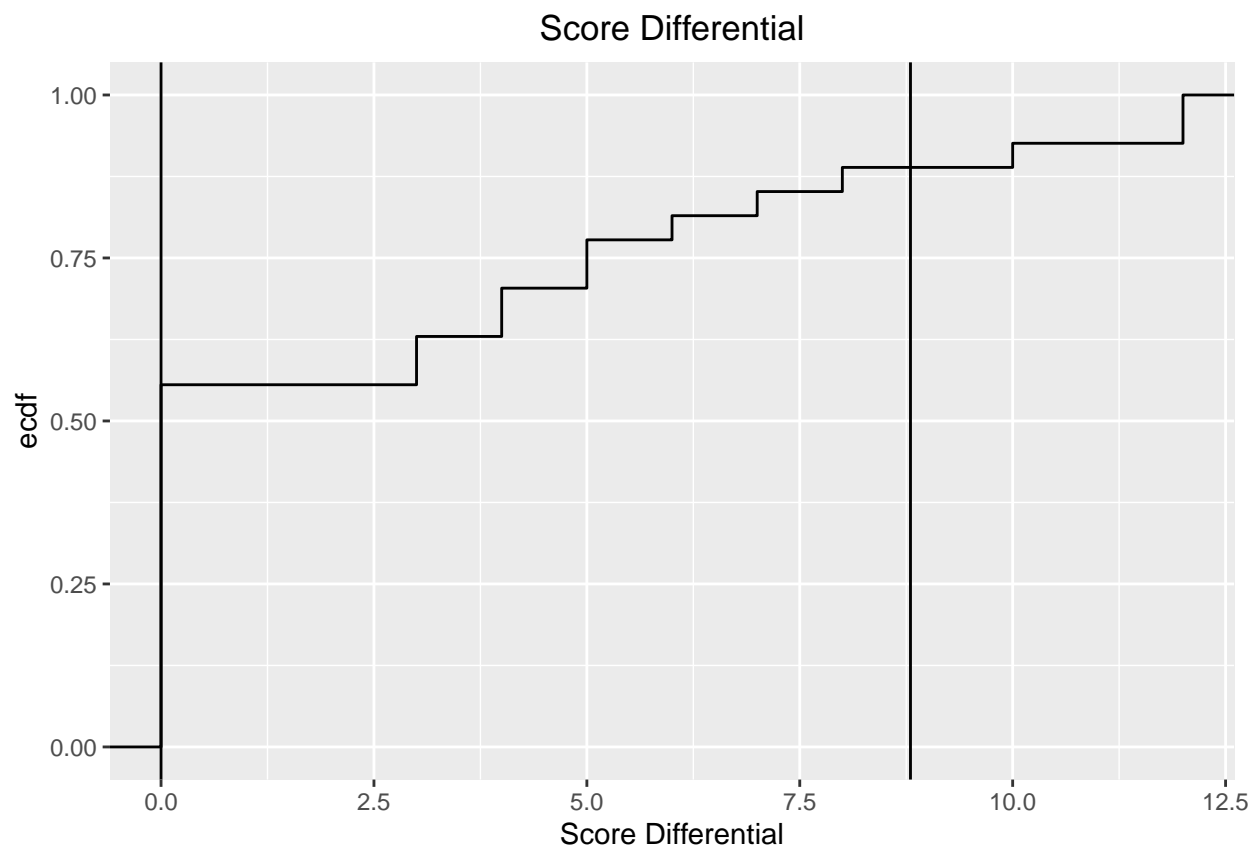
```
## 0
```

```
u
```

```
## 90%
```

```
## 8.8
```

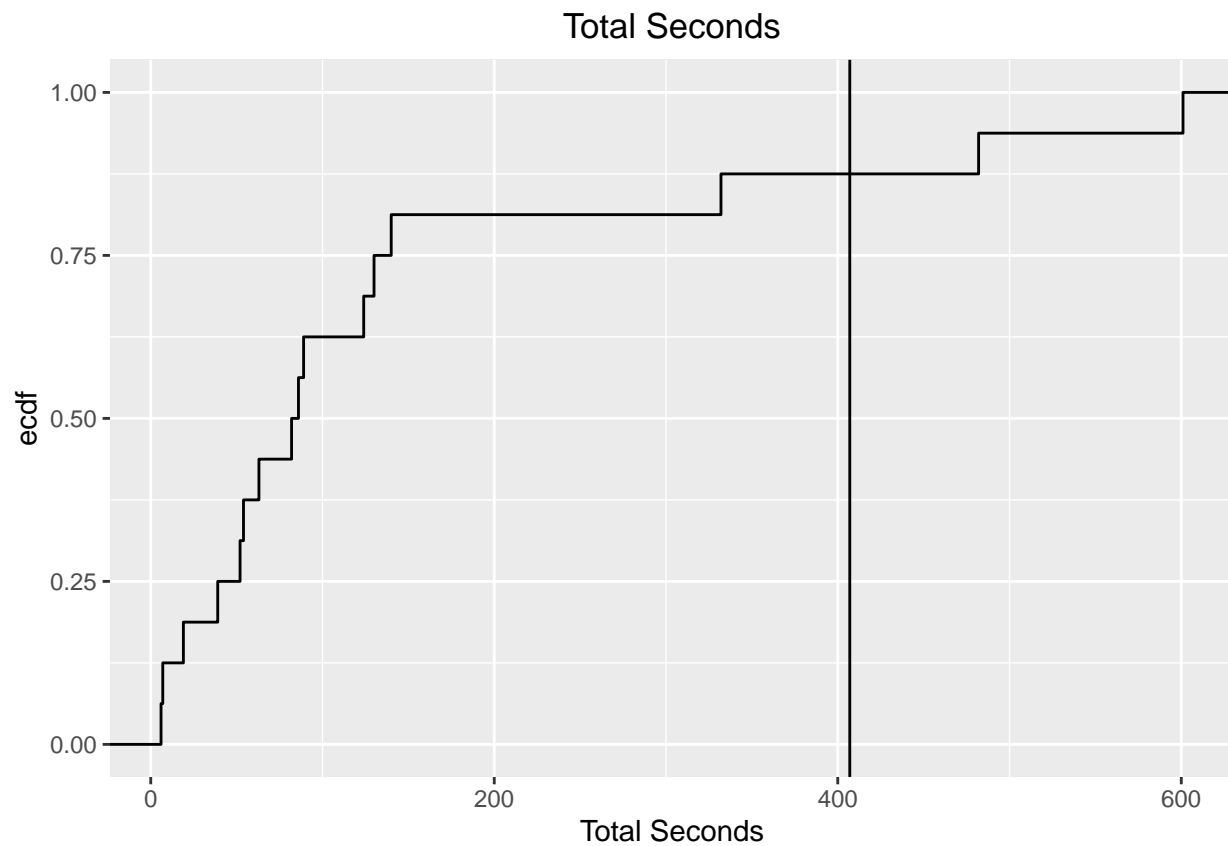
```
ggplot(singular_game, aes(x = `SCORE DIFFERENTIAL WHEN ENTER`)) + stat_ecdf() + geom_vline(xintercept =
```



```
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

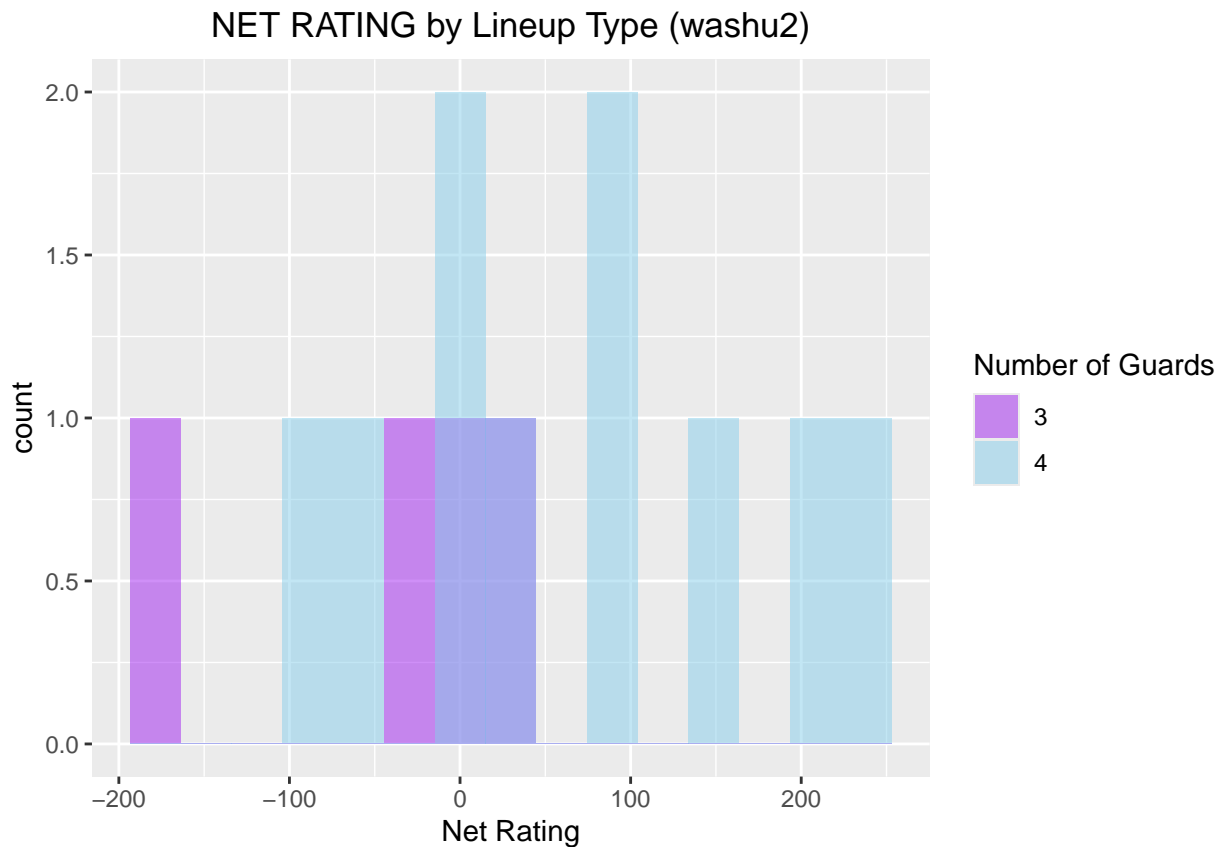
## 90%
## 407

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

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

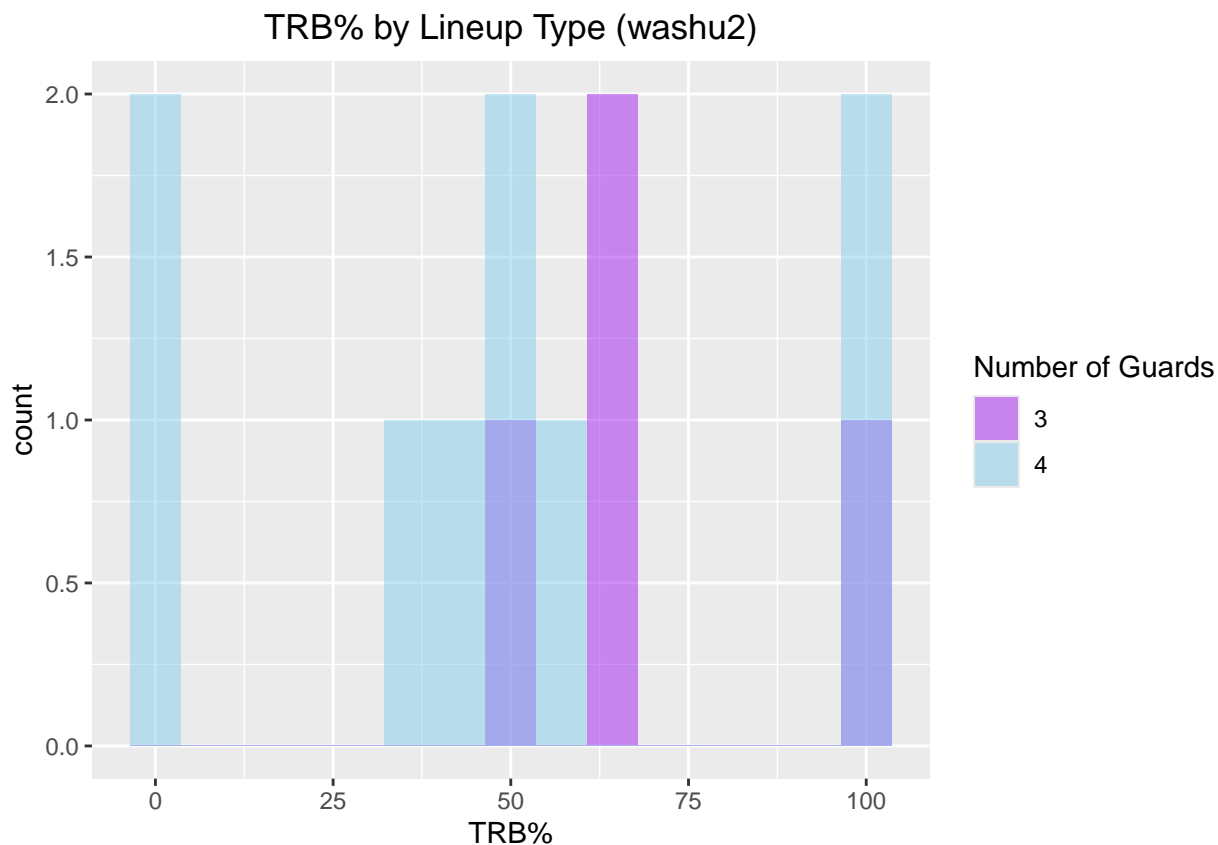
ggplot(data = subset(game, subset = `NUMBER OF GUARDS` %in% t_f), aes(x = `NET RATING`, fill = factor(`

## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



```
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.
    ## -166.667 -66.667 -16.667 -41.667   8.333  33.333
    ##
    ## $`4`
    ##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.      NA's
    ## -100.00  -6.25   69.16   66.33  137.50  250.00         2
    wilcox.test(`NET RATING` ~ `NUMBER OF GUARDS`, data = subset(game, `NUMBER OF GUARDS` %in% t_f), exact = FALSE)

    ##
    ## Wilcoxon rank sum test with continuity correction
    ##
    ## data:  NET RATING by NUMBER OF GUARDS
    ## W = 9.5, p-value = 0.1564
    ## 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`)))
    ## Warning: Removed 3 rows containing non-finite outside the scale range
    ## (`stat_bin()`).
```



```
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.
##  50.00   59.38   64.58   69.79   75.00   100.00
##
```

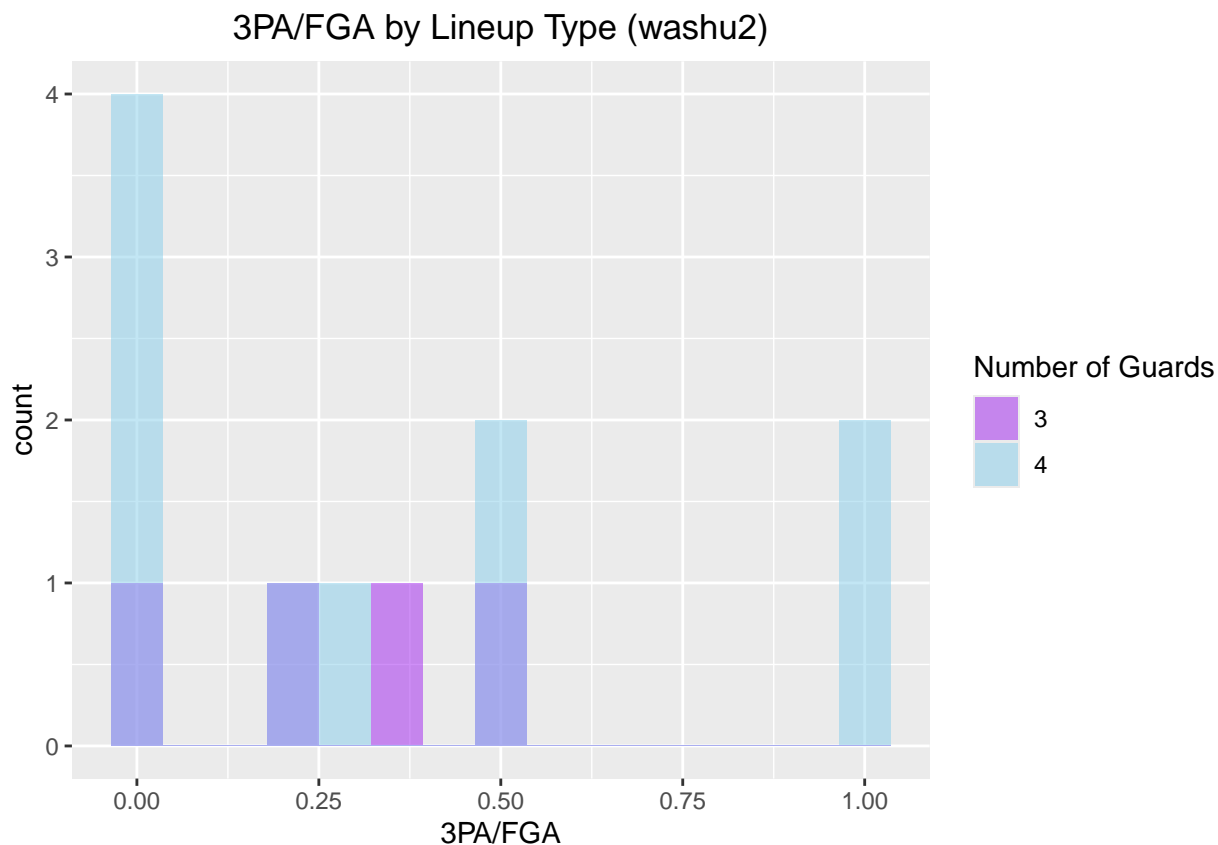
```
## $`4`
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   33.33   50.00   48.15   60.00   100.00        3
```

```
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 = 27, p-value = 0.1841
## 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`)))
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



```
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.
## 0.0000 0.1500 0.2667 0.2583 0.3750 0.5000
##
```

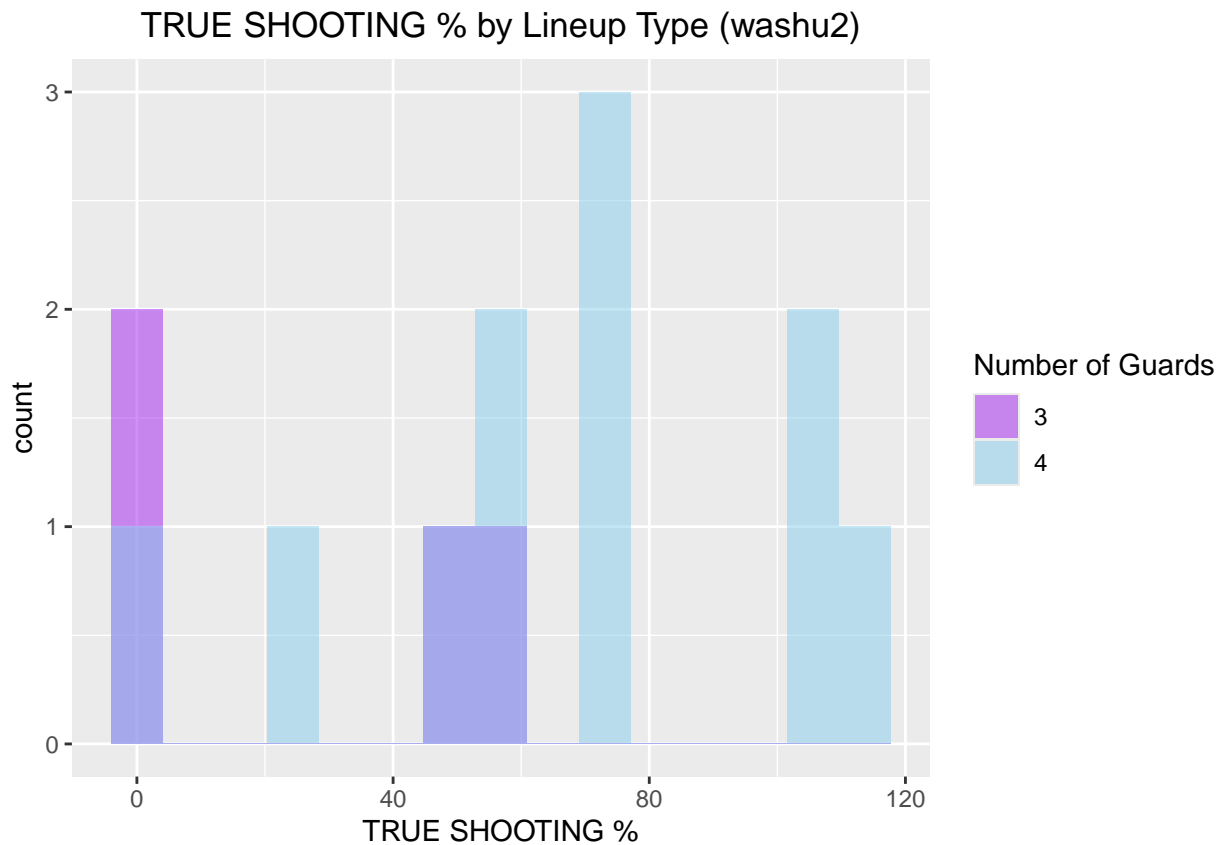
```
## $`4`
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
## 0.0000 0.0000 0.2409 0.3482 0.5000 1.0000     2
```

```
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 = 20, 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 = `TRUE SHOOTING %`, fill = fac
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_bin()`).
```



```

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) {

```

```

## $`3`
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.00   0.00   26.04   28.02   54.06   60.00
##
## $`4`
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      0.00   51.60   69.44   65.64   88.73   113.64      1

```

```

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

```

```

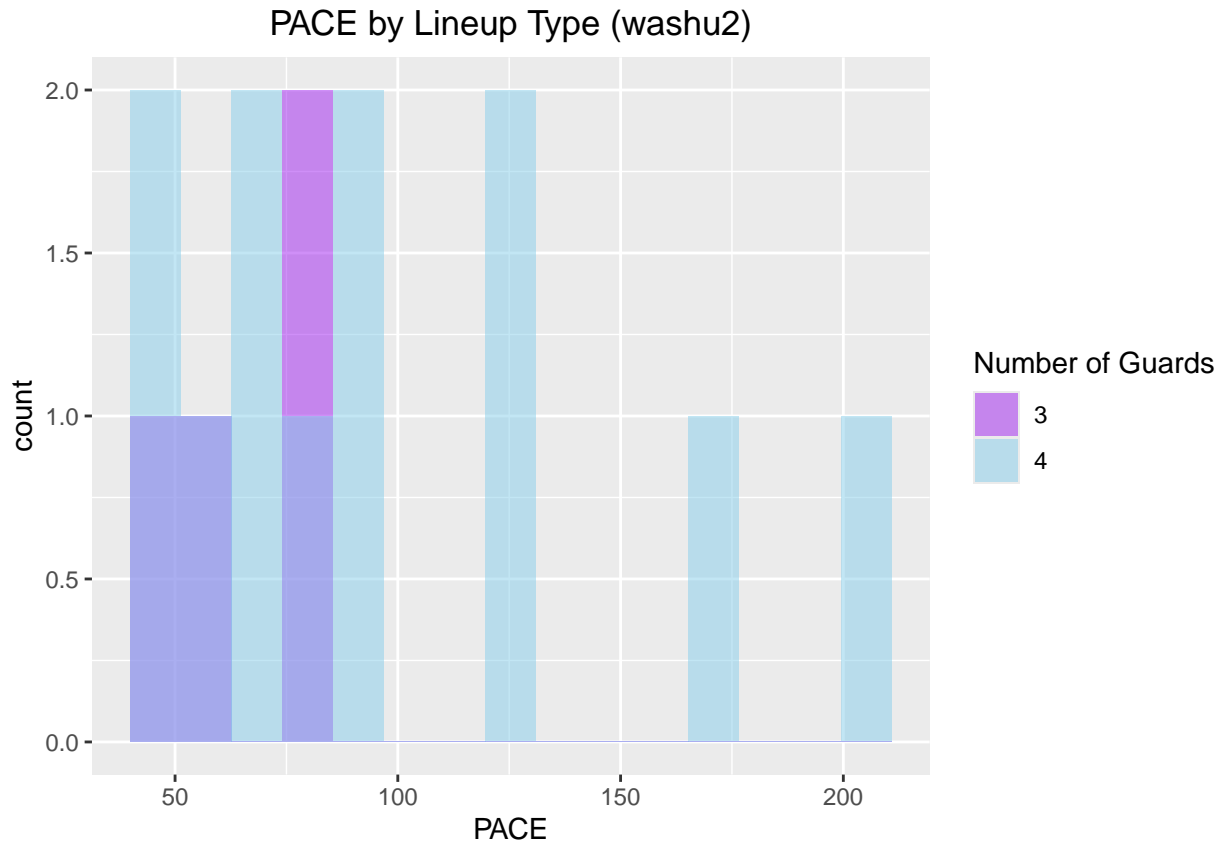
##
## Wilcoxon rank sum test with continuity correction
##
## data: TRUE SHOOTING % by NUMBER OF GUARDS
## W = 9, p-value = 0.1015
## alternative hypothesis: true location shift is not equal to 0

```

```

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` %
```

```
## $`3`
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  42.86  52.64   66.66   64.98   78.99   83.72
##
```

```
## $`4`
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  40.45  62.39   83.28   96.49  123.89  200.00
```

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

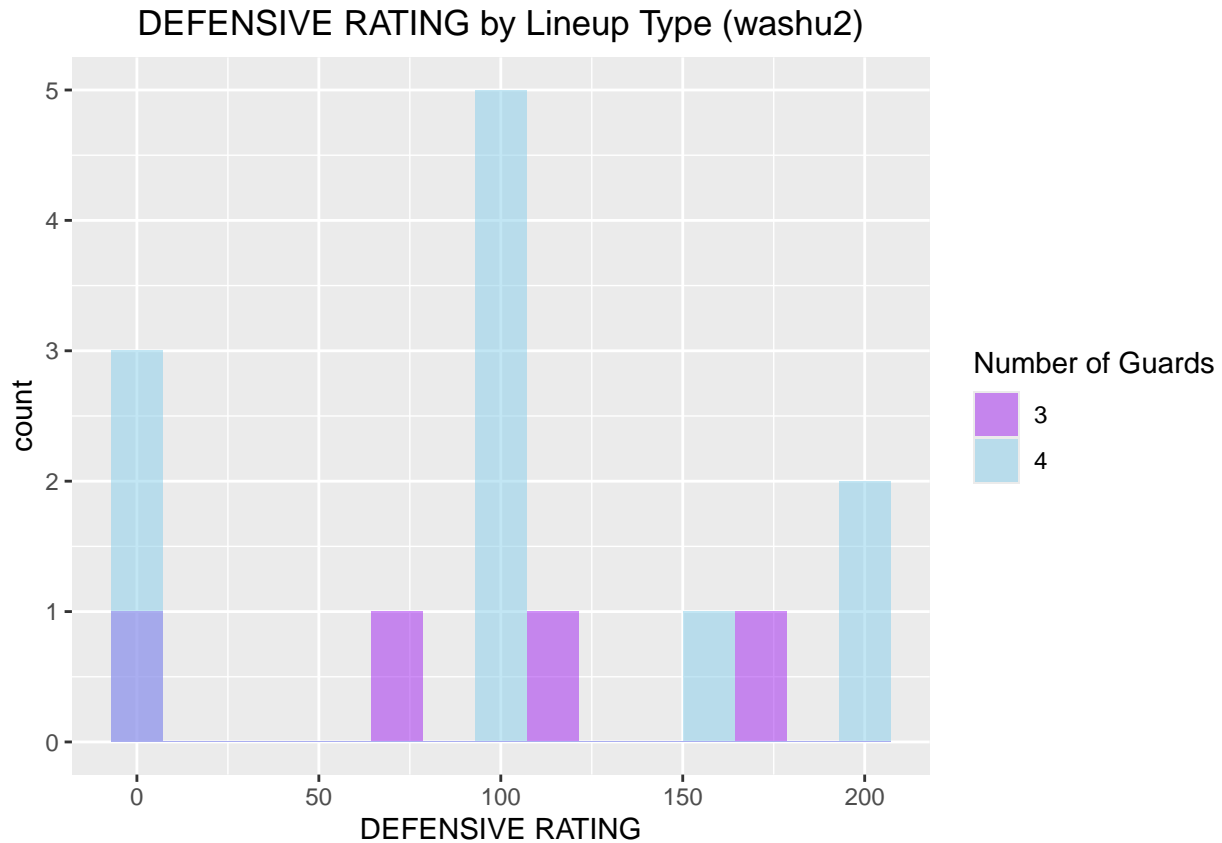
```
##
## Wilcoxon rank sum test with continuity correction
##
```

```
## data: PACE by NUMBER OF GUARDS
## W = 16, p-value = 0.3631
## 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
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_bin()`).
```

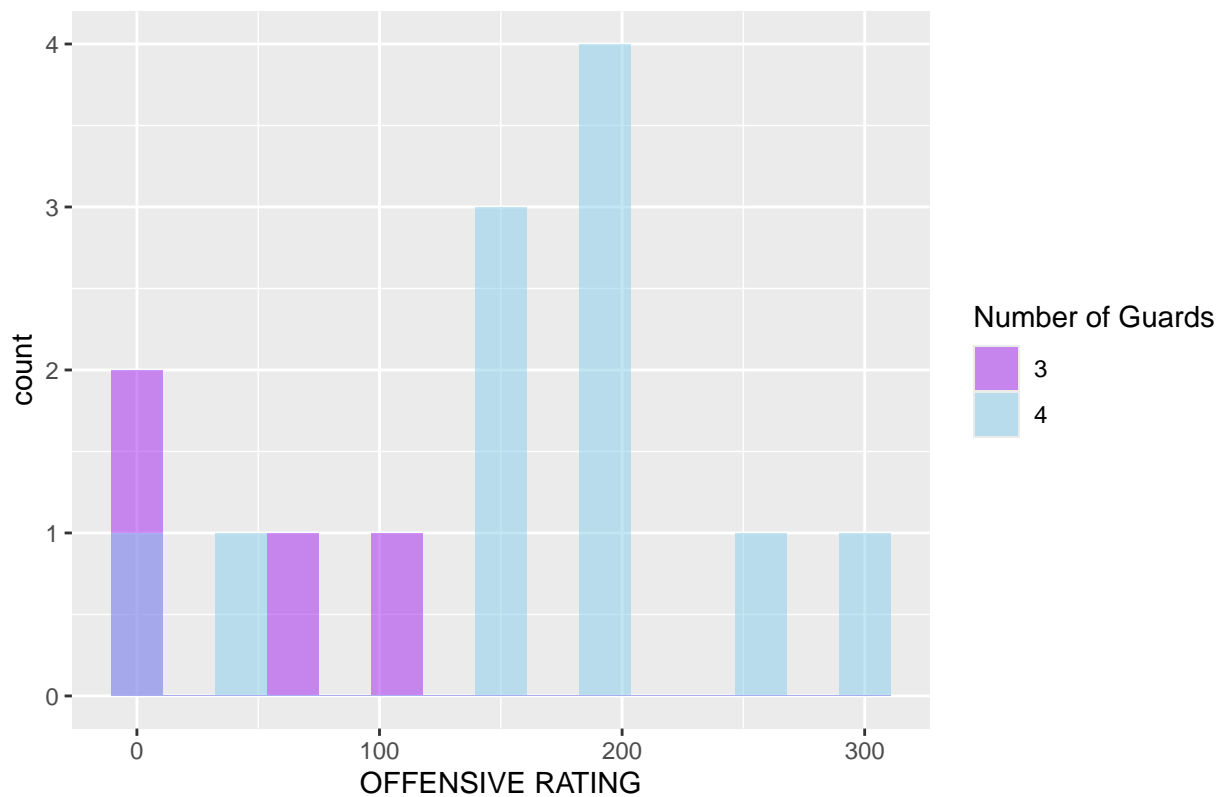




```

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) {
  ## $`3`
  ##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  ##   0.00  50.00   87.50   85.42 122.92   166.67
  ##
  ## $`4`
  ##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
  ##   0.00  50.00  100.00   96.86 132.74   200.00     1
  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 = 21.5, 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 = `OFFENSIVE RATING`, fill = fa
  ## Warning: Removed 1 row containing non-finite outside the scale range
  ## (`stat_bin()`).
  
```

## OFFENSIVE RATING by Lineup Type (washu2)



```

tapply(game$`OFFENSIVE RATING`[game$`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.
##   0.00   0.00   37.50   43.75   81.25  100.00
##

```

```

## $`4`
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
##   0.0   147.7   200.0   166.3   200.0   300.0     1

```

```

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

```

```

##
## Wilcoxon rank sum test with continuity correction
##
## data:  OFFENSIVE RATING by NUMBER OF GUARDS
## W = 5, p-value = 0.02899
## alternative hypothesis: true location shift is not equal to 0

```

```

#dev.off()

```