NBA Analysis

Nestor Torrech, Sergio Zahira

5/9/2022

Abstracto

El propósito de este estudio es comparar el rendimiento entre los jugadores de la NBA, la liga profesional de baloncesto estadounidense, que fueron a la universidad y los que no fueron. Usando una base de datos de varios jugadores y sus métricas desde 1996 hasta 2020, esta investigación busca descubrir la existencia de alguna diferencia significativa entre los jugadores en términos de las métricas usadas. Además, también se busca encontrar que factores son significativos al predecir la experiencia universitaria de los jugadores. Los resultados de investigación van a ser útiles para los investigadores que quieran añadir mas detalles y los administradores de equipos profesionales de baloncesto que quieren optimizar sus decisiones sobre cuales jugadores comprar. Se concluye que los jugadores sin experiencia universitaria solían tener un mejor rendimiento.

Introducción

La proliferación de las bases de datos masivas ha afectado todos los aspectos de nuestra vida (Sagiroglu & Sinanc, 2013). El almacenamiento, la visualización, el análisis y la minería de datos son técnicas imperativas para la toma de decisiones en el mundo del siglo 21. Sin sorpresa alguna, este fenómeno también ha penetrado la industria lucrativa de los deportes (Azar et al., 2018). Desde los deportes más populares como el fútbol o "soccer" (Asif et al., 2016) hasta los menos reconocidos como el patinaje de velocidad (Cachucho et al., 2017), el análisis basado en datos se ha aplicado a los atletas, los entrenadores y lo árbitros. Siguiendo en línea con el entorno de la industria deportiva, esta investigación busca ejecutar un análisis comparativo de los jugadores de la NBA, también conocida como "National Basketball Association" y "Asociación Nacional de Baloncesto". En específico, usando técnicas de análisis, minería y visualización de datos, se busca encontrar factores y fenómenos que afectan el rendimiento de los jugadores. En este estudio se le da un énfasis particular en la experiencia educativa de los jugadores.

La NBA, fundada en el 1976 luego de la fusión entre la Asociación de Baloncesto Americana (BAA) y la Liga Nacional de Baloncesto (NBL), es la liga de baloncesto más popular en el mundo (Staffo, 1998). Además, curiosamente, el baloncesto es el único deporte puramente originario en Estados Unidos. Como en todo deporte, las estadísticas siempre han jugado una parte importante. Al principio, se colectaba la data a mano y se usaban métricas rudimentarias para analizar el deporte. Sin embargo, con el incremento del poder computacional, las estadísticas se volvieron mucho más complejas. Contrario a la percepción común, no fueron los científicos de datos ni los estadísticos quienes empezaron a crear y usar estas métricas. Sino que fueron los aficionados del deporte; este fenómeno también lo vimos en el Moneyball de beisbol (Lewis, 2004). Oliver (2004) y Hollinger (2003, 2004 & 2005) fueron los primeros estudios que segmentaron la base para las estadísticas del baloncesto. Sin embargo, estos escritos solo eran usados por áreas no académicas. A pesar de ser fuentes informales, añadieron mucho valor y crearon la base para el área académica. Particularmente, una de las grandes contribuciones fue la gran importancia que le dieron a la posesión como fundación estadística (Kubatko et al., 2007)

Aunque los primeros jugadores del espacio no eran académicos, la cantidad de datos disponibles y su posible aplicación a la administración, sociología, estadística, sicología, medicina y economía eran factores muy atractivos para los académicos en sus torres de marfil. Por lo tanto, se empezaron a usar técnicas aún más complejas para tratar de resolver todas las preguntas que crea el baloncesto. Cao (2012) y Belkham et al. (2018) usan técnicas de minería de datos para predecir los resultados de los juegos y analizar las estrategias óptimas para crear equipos.

Otra de las cuestiones más discutidas fue la del juego ofensivo contra el defensivo; cuando el equipo tiene el balón contra cuando no lo tiene. Anguera et al. (2009) analizaron las estrategias ofensivas para predecir su éxito. Franks et al. (2015) usan una combinación de procesos espaciales y espacio-temporales, técnicas de factorización de matrices y modelos de regresión jerárquicos para analizar el rol del juego defensivo en el baloncesto. Estos dos estudios demuestran que todos los aspectos que pasaban en la canchar eran meticulosamente analizados hasta mas no poder. Entonces, gracias a la llegada de los académicos se exploraron áreas del deporte que nadie jamás hubiera esperado que se exploraran. En vez del enfoque en las dinámicas del

juego, también se enfatizó el rendimiento de los jugadores individualmente. Esto último es el enfoque de este estudio. Moxley & Towne (2014) y Evans (2018) estudian cuales son los factores que aportan al rendimiento de los jugadores. Sin embargo, los investigadores se dieron cuenta que las características físicas no eran significativas para el éxito de los jugadores. Si no que era la experiencia del jugador que lo determinaba. Esto lleva a una de las preguntas principales de este escrito. ¿Habrá una diferencia entre el rendimiento de los jugadores de la NBA que fueron a la universidad y de los que no fueron?

Spurr (2002) hizo un estudio enfocado en el beisbol y encontró que la experiencia universitaria de los jugadores era infravalorada por los evaluadores del talento. Berri et al. (2010) hizo un estudio similar pero aplicado al baloncesto y tuvo hallazgos bastantes similares; las habilidades físicas y ofensivas eran sobrevaloradas. A pesar de que estos estudios predican sobre la importancia de los factores cualitativos para determinar el éxito de los jugadores, nunca se fueron en detalle sobre esto. Este espacio en blanco en la literatura del análisis del baloncesto es el que buscamos llenar con este estudio.

Ademas de añadir a la literatura, esta investigación es de gran valor para la toma de decisiones dentro de los equipos de la NBA, y posiblemente los equipos de otras ligas profesionales de baloncesto, al momento de escoger que jugadores deben incorporar a su equipo. Poder añadir los mejores jugadores es de suma importancia para la tesorería y éxito del equipo; tomar la decisión correcta puede generar millones y mejorar la posición del equipo (Hausman & Leonard, 1997; Walter & Williams, 2012). También esperamos que en el futuro se hagan mas investigaciones de los planteamientos de este estudio.

Metodologia

Aqui estaremos preparando los datos de estudio del NBA.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## Warning: package 'stringr' was built under R version 4.1.3

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(latexpdf)
all_season <- read_csv("C:\\Users\\nesto\\OneDrive\\Documents\\R Projects\\ESTA5504_2021S2\\Datasets\\N
```

```
## New names:
## * `` -> ...1

## Rows: 11700 Columns: 22

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (8): player_name, team_abbreviation, college, country, draft_year, draf...
## dbl (14): ...1, age, player_height, player_weight, gp, pts, reb, ast, net_ra...

##
## 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.
```

```
head(all_season)
```

```
## # A tibble: 6 x 22
##    ...1 player_name  team_abbreviati~   age player_height player_weight college
##   <dbl> <chr>        <chr>            <dbl>         <dbl>         <dbl> <chr>
## 1     0 Travis Knig~ LAL                 22          213.          107. Connect~
## 2     1 Matt Fish    MIA                 27          211.          107. North C~
## 3     2 Matt Bullard HOU                 30          208.          107. Iowa
## 4     3 Marty Conlon BOS                 29          211.          111. Provide~
## 5     4 Martin Muur~ DAL                 22          206.          107. None
## 6     5 Martin Lewis TOR                 22          198.          102. Seward ~
## # ... with 15 more variables: country <chr>, draft_year <chr>,
## #   draft_round <chr>, draft_number <chr>, gp <dbl>, pts <dbl>, reb <dbl>,
## #   ast <dbl>, net_rating <dbl>, oreb_pct <dbl>, dreb_pct <dbl>, usg_pct <dbl>,
## #   ts_pct <dbl>, ast_pct <dbl>, season <chr>
```

Analisis de Experiencia Educativa de Los Jugadores

Durante una exploracion de datos preliminar que hicimos antes de empezar este documento, encontramos algo interesante. Hay un porcentaje bastante significativo de jugadores en nuestros datos quienes *no* cuentan con experiencia universitaria. Aqui incluiremos los resultados de dicha exploracion

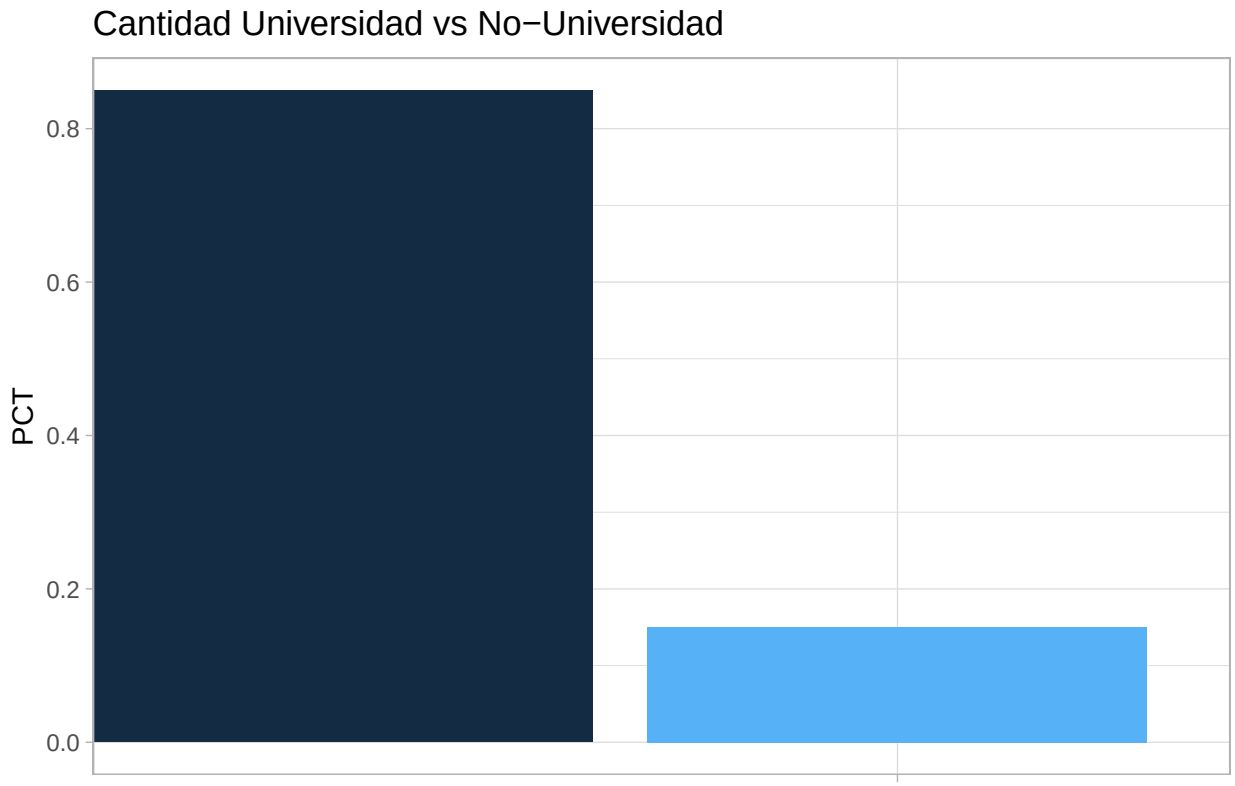
```
all_season %>%
    group_by(college) %>%
    count() %>%
    arrange(desc(n))
```

```
## # A tibble: 334 x 2
## # Groups:   college [334]
##    college            n
##    <chr>          <int>
##  1 None            1715
##  2 Kentucky         391
##  3 Duke             361
##  4 North Carolina   332
##  5 UCLA             295
##  6 Arizona          268
##  7 Kansas           263
##  8 Connecticut      225
##  9 Georgia Tech     185
## 10 Florida          181
## # ... with 324 more rows
```

Aqui vemos que, singularmente, el grupo mas grande en terminos de nivel de educacion universitaria es **"None"** . Pero para entenderlo mejor, vizualizemoslo

```
all_season_factor<- all_season %>%
              mutate(No_College = ifelse(college == "None",1,0) ) %>%
              group_by(No_College) %>%
              count() %>%
              mutate(PCT = round(n/length(all_season$college),2))

ggplot(all_season_factor, aes(No_College,PCT, fill=No_College)) +
  geom_col() + xlab(label = "") + xlim("") + theme_light() +
  theme(legend.position = "None") + labs(title = "Cantidad Universidad vs No-Universidad ")
```

```
## Warning: Removed 1 rows containing missing values (position_stack).
```




```
# Porcentaje

all_season_factor %>% mutate(PCT = round(n/length(all_season$college),2))
```

```
## # A tibble: 3 x 3
## # Groups:   No_College [3]
##   No_College     n   PCT
##        <dbl> <int> <dbl>
## 1          0  9980  0.85
## 2          1  1715  0.15
## 3         NA     5  0
```

Como podemos ver en el analisis porcentual, el grupo sin experiencia universitaria compone **15%** de la poblacion.

Aunque claro esta que los grupos **"None"** son lejos de la mayoria, si es claro que son un componente significativo de los datos. Es por esto que entendemos que son de interes para estudiar como ellos comparan en varias metricas de Basketball a jugadores que si tienen experiencia en la universidad.

Muestreo a Base de Experiencia Universitaria

Tomando en cuenta la diferencia en terminos de observaciones entre el grupo universitario y el no-universitario, decidimos que para hacer un analisis valido se tendria que sacar una **muestra aleatoria** del grupo universitario. Esta tendria que ser de un tamano igual al grupo no-universitario.

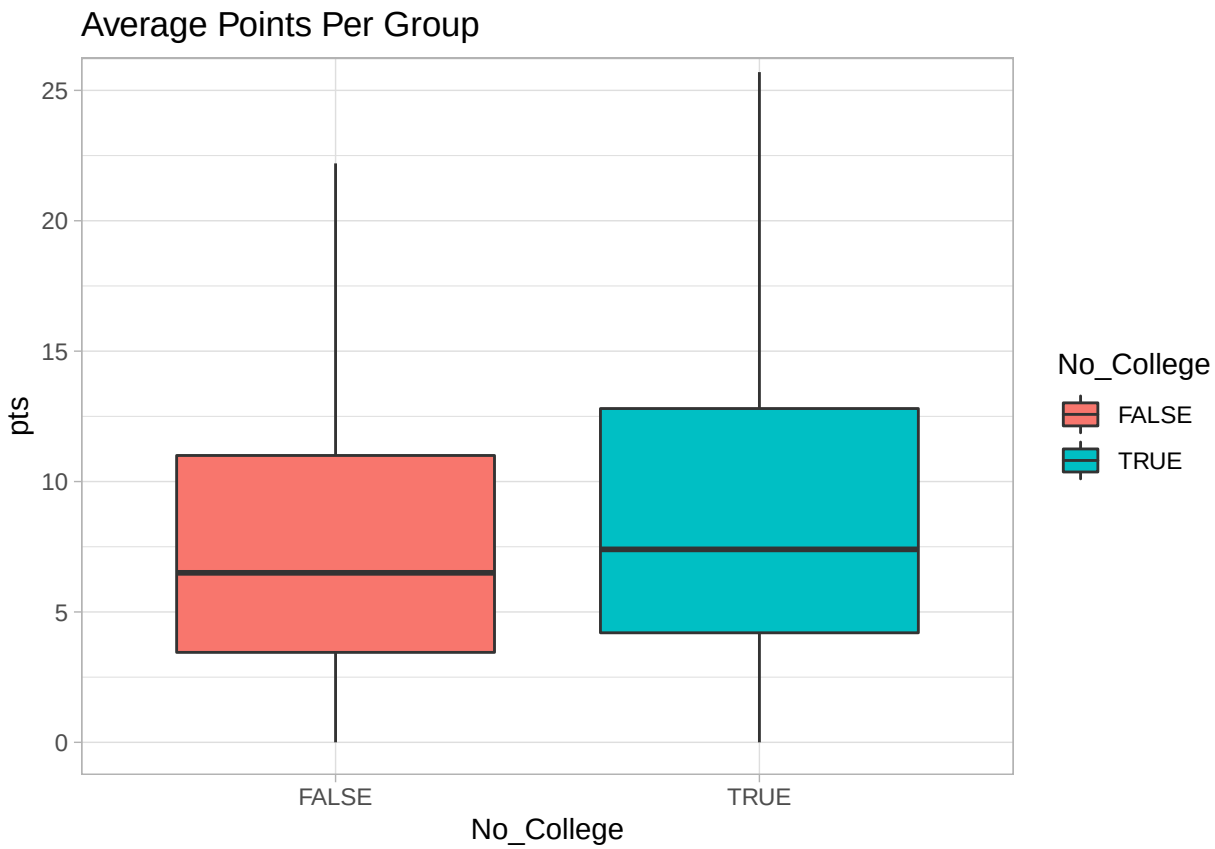
Para esto, utilizaremos la funcion **slice_sample**.

```r
set.seed(4120)
all_season_none <- all_season %>% filter(college == "None") %>%
  mutate(No_College = as.logical(ifelse(college == "None",1,0)))

all_season_college <- all_season %>% filter(college != "None")

college_sample <- all_season_college %>% slice_sample(n = 1715) %>%
  mutate(No_College = as.logical(ifelse(college == "None",1,0)))

all_season_sampled <- rbind(all_season_none, college_sample)
```

Ahora que tenemos nuestra muestra, podemos empezar con los analisis de las varias metricas de Basketball.

Analisis de Puntos Promedios

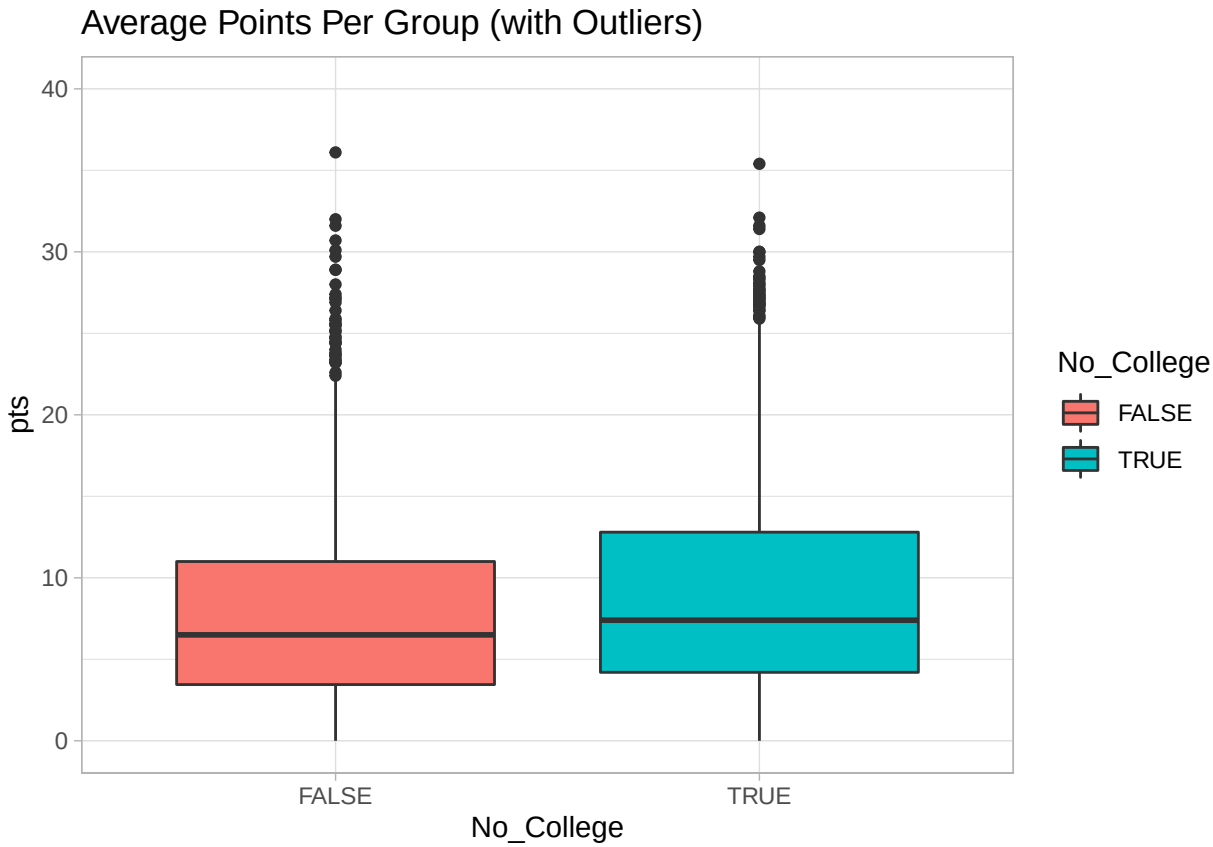
```r
ggplot(all_season_sampled, aes(No_College, pts, fill = No_College)) +
  geom_boxplot(outlier.shape = NA) + coord_cartesian(ylim = c(0,25)) +
  theme_light() + labs(title = "Average Points Per Group")
```




```r
# Diferencia Numerica

list("Mediana No-College" = median(all_season_none$pts),
  "Mediana College" = median(college_sample$pts))
```

```
## $`Mediana No-College`
## [1] 7.4
##
## $`Mediana College`
## [1] 6.5
```

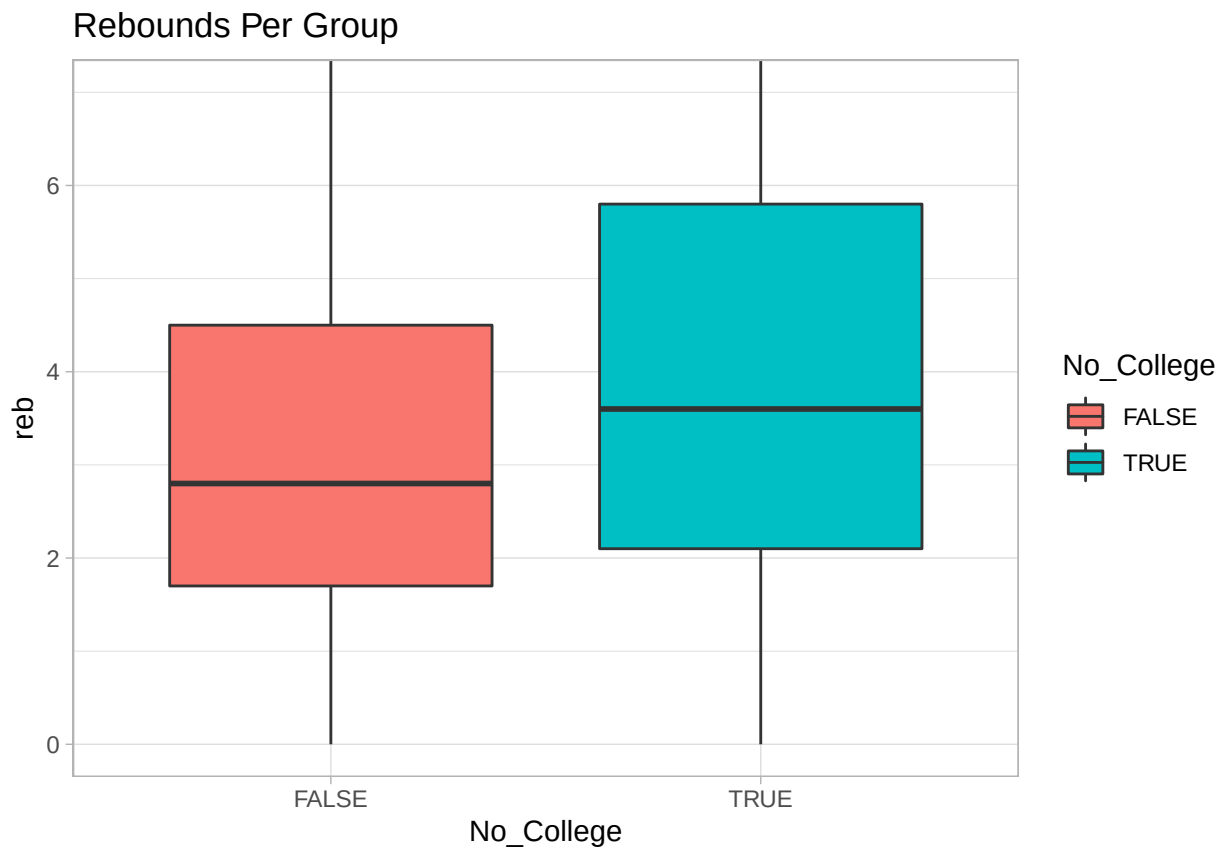
```
ggplot(all_season_sampled, aes(No_College, pts, fill = No_College)) +
  geom_boxplot() + coord_cartesian(ylim = c(0,40)) +
  theme_light() + labs(title = "Average Points Per Group (with Outliers)")
```



Como podemos ver, segun esta grafica parece que jugadores sin experiencia universitaria tienden a sacar puntos promedios un poco mas alto que jugadores que si cuentan con dicha experiencia.

Analisis de Rebounds

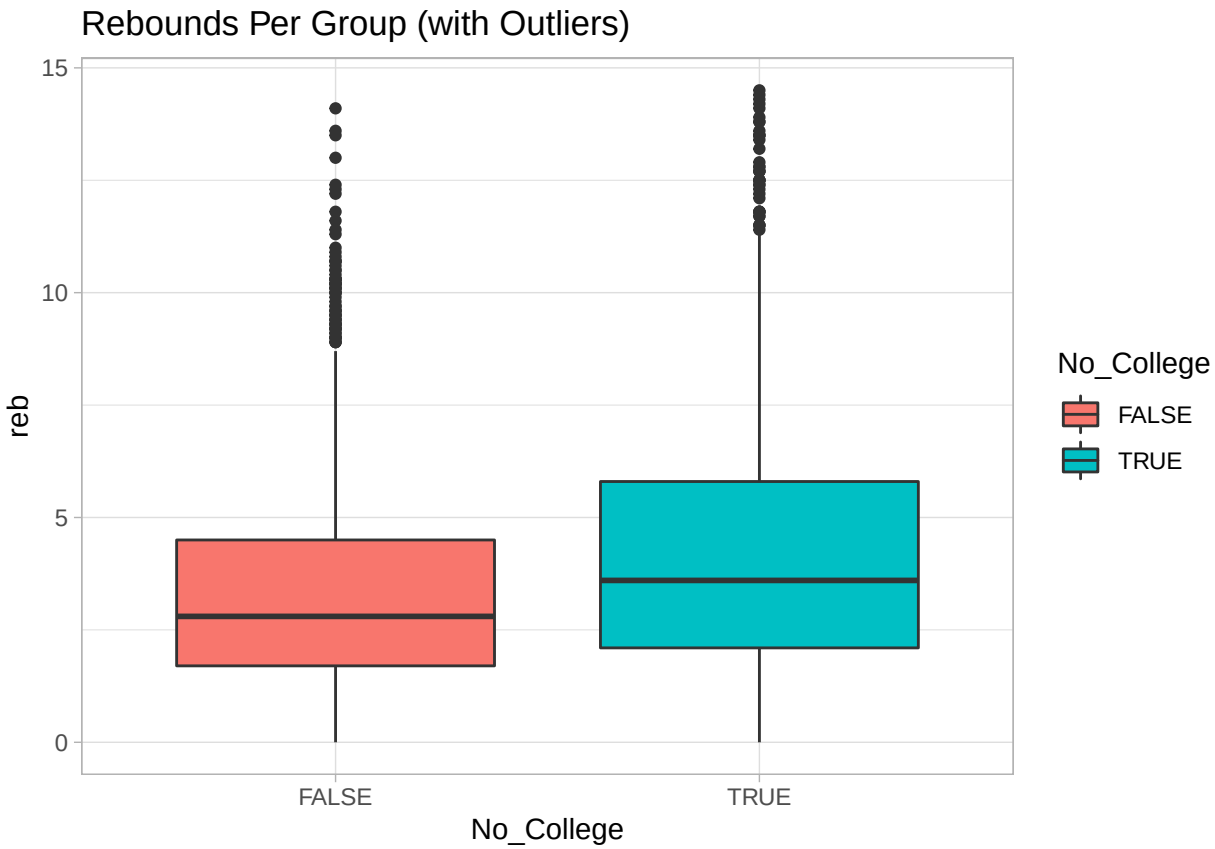
```
ggplot(all_season_sampled, aes(No_College, reb, fill = No_College)) +
  geom_boxplot(outlier.shape = NA) + coord_cartesian(ylim = c(0,7)) +
  theme_light() + labs(title = "Rebounds Per Group")
```




```
# Diferencia Numerica

list("Mediana No-College" = median(all_season_none$reb),
  "Mediana College" = median(college_sample$reb))
```

```
## $`Mediana No-College`
## [1] 3.6
##
## $`Mediana College`
## [1] 2.8
```

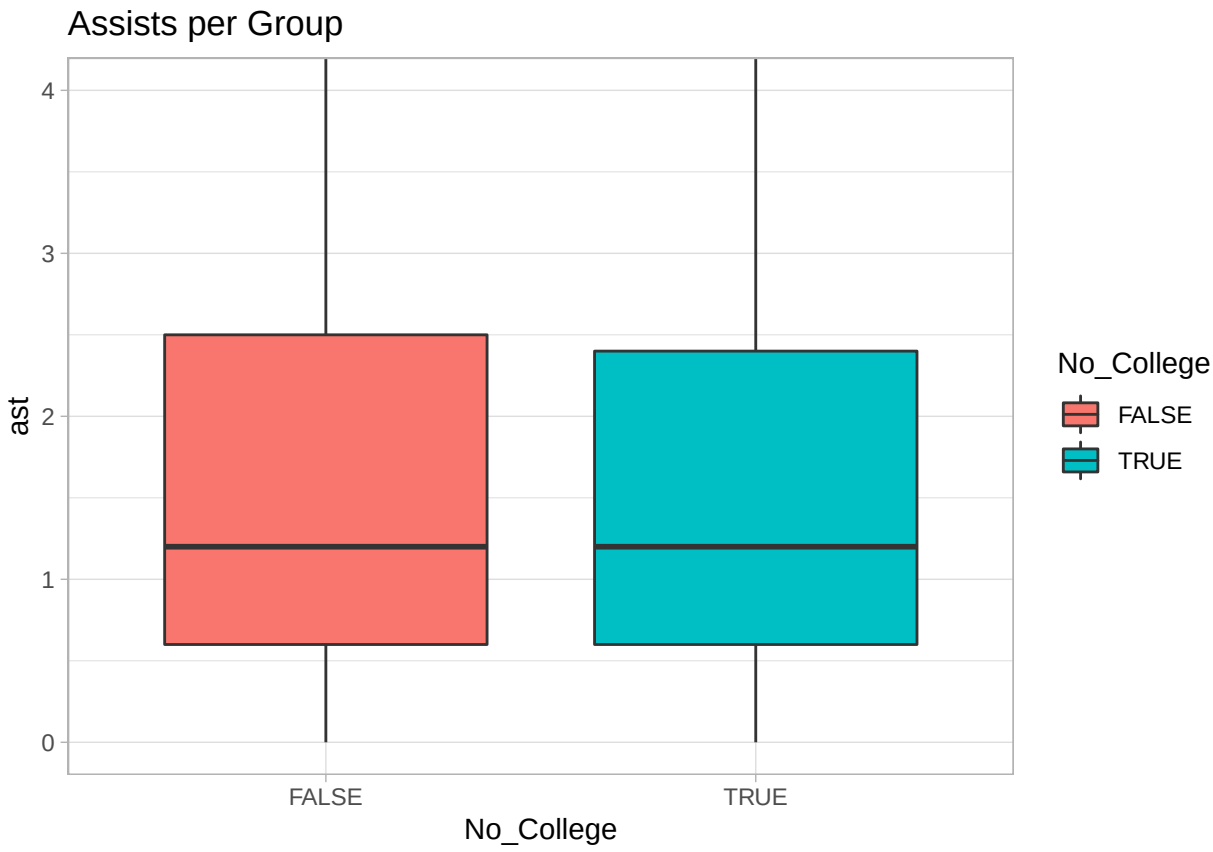
```
ggplot(all_season_sampled, aes(No_College, reb, fill = No_College)) +
  geom_boxplot() + theme_light() + labs(title = "Rebounds Per Group (with Outliers)")
```



Aqui vemos, vizualizado otra vez, que jugadores sin experiencia universitaria tienden a conseguir los rebounds mas consistentemente. Sin embargo, vale notar que cuando se habla de valores extremos, estos tienden a residir con el grupo universitario. Es decir que aunque **por lo general** el grupo de "No College" tiende a tener puntos un poco mas alto, los **mas** altos tienden a ser del grupo universitario

Analisis de Asistencias

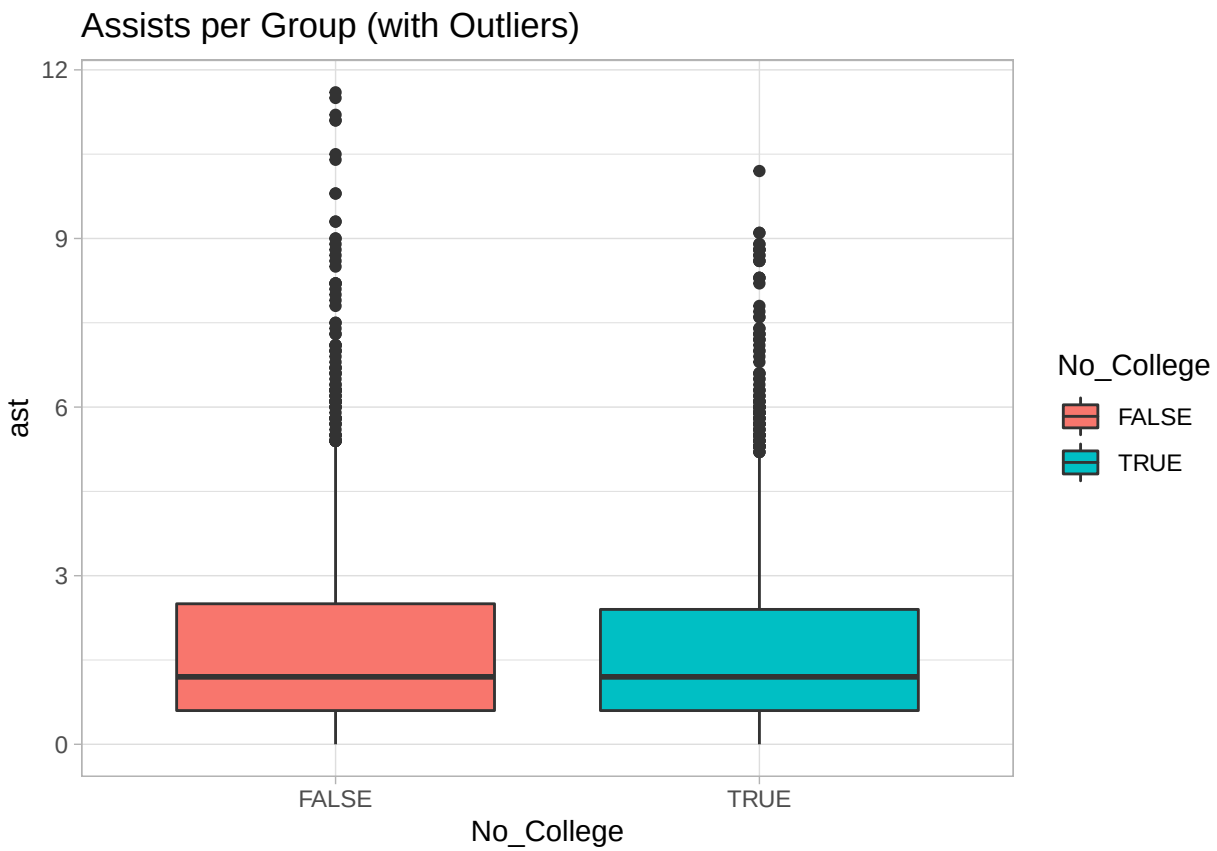
```
ggplot(all_season_sampled, aes(No_College, ast, fill = No_College)) +
  geom_boxplot(outlier.shape = NA) + coord_cartesian(ylim = c(0,4)) +
  theme_light() + labs(title = "Assists per Group")
```




```
# Diferencia Numerica

list("Mediana No-College" = median(all_season_none$ast),
  "Mediana College" = median(college_sample$ast))
```

```
## $`Mediana No-College`
## [1] 1.2
##
## $`Mediana College`
## [1] 1.2
```

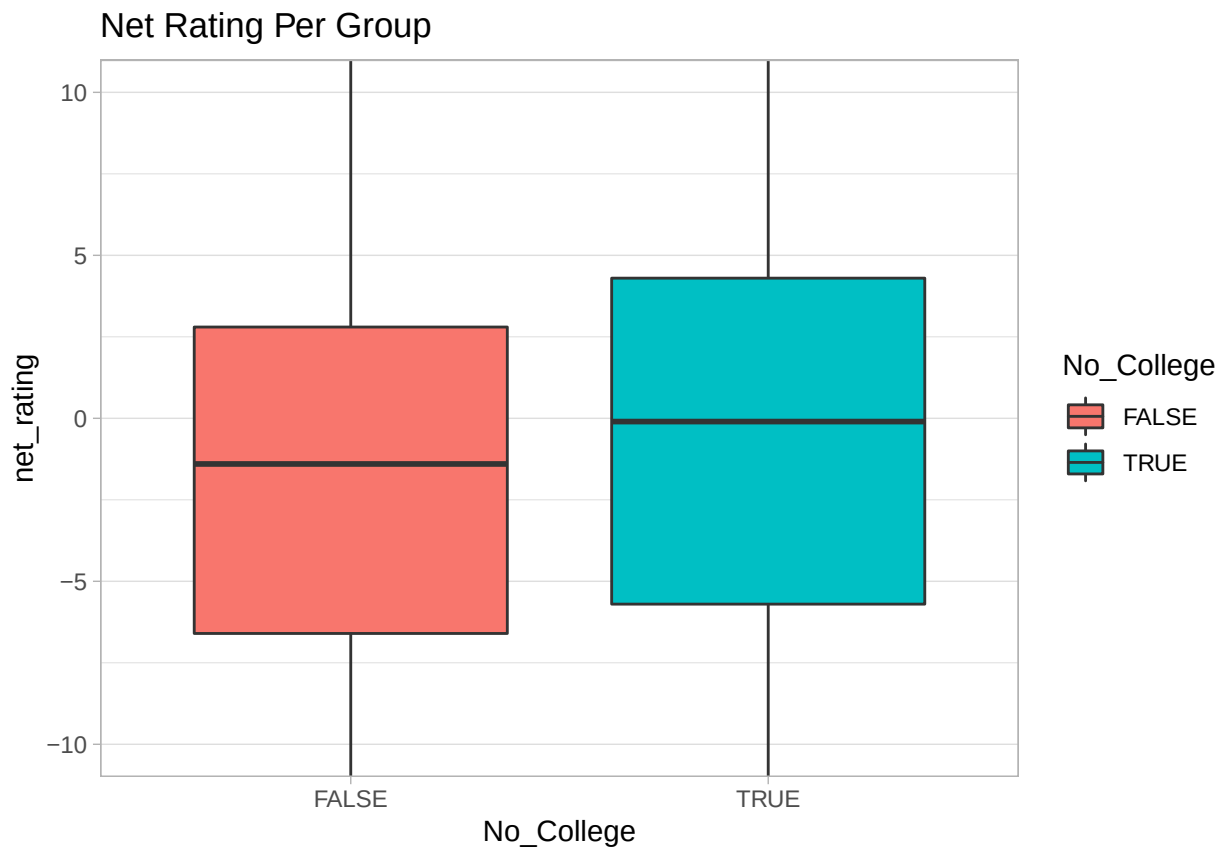
```
ggplot(all_season_sampled, aes(No_College, ast, fill = No_College)) +
  geom_boxplot() + theme_light() + labs(title = "Assists per Group (with Outliers)")
```



En el caso de las asistencias, parece que el "performance" en este caso favorece el grupo Universitaria. Ya que vemos, tanto an un nivel promedio como al nivel de outliers, los **universitarios** tienen una ventaja.

Analisis de Net Rating

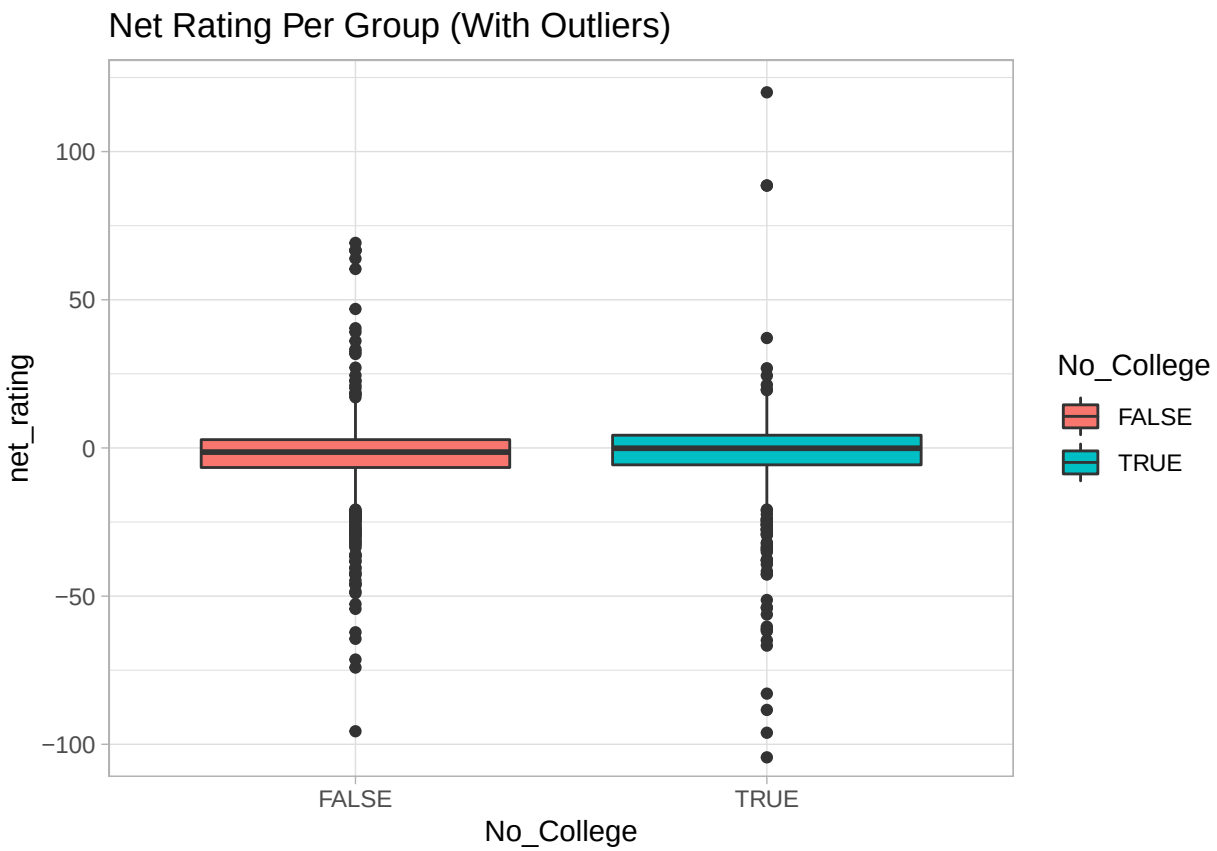
```
ggplot(all_season_sampled, aes(No_College, net_rating, fill = No_College)) + geom_boxplot() +
coord_cartesian(ylim = c(-10,10)) + theme_light() + labs(title = "Net Rating Per Group")
```




```
list("Mediana No-College" = median(all_season_none$net_rating),
  "Mediana College" = median(college_sample$net_rating))
```

```
## $`Mediana No-College`
## [1] -0.1
##
## $`Mediana College`
## [1] -1.4
```

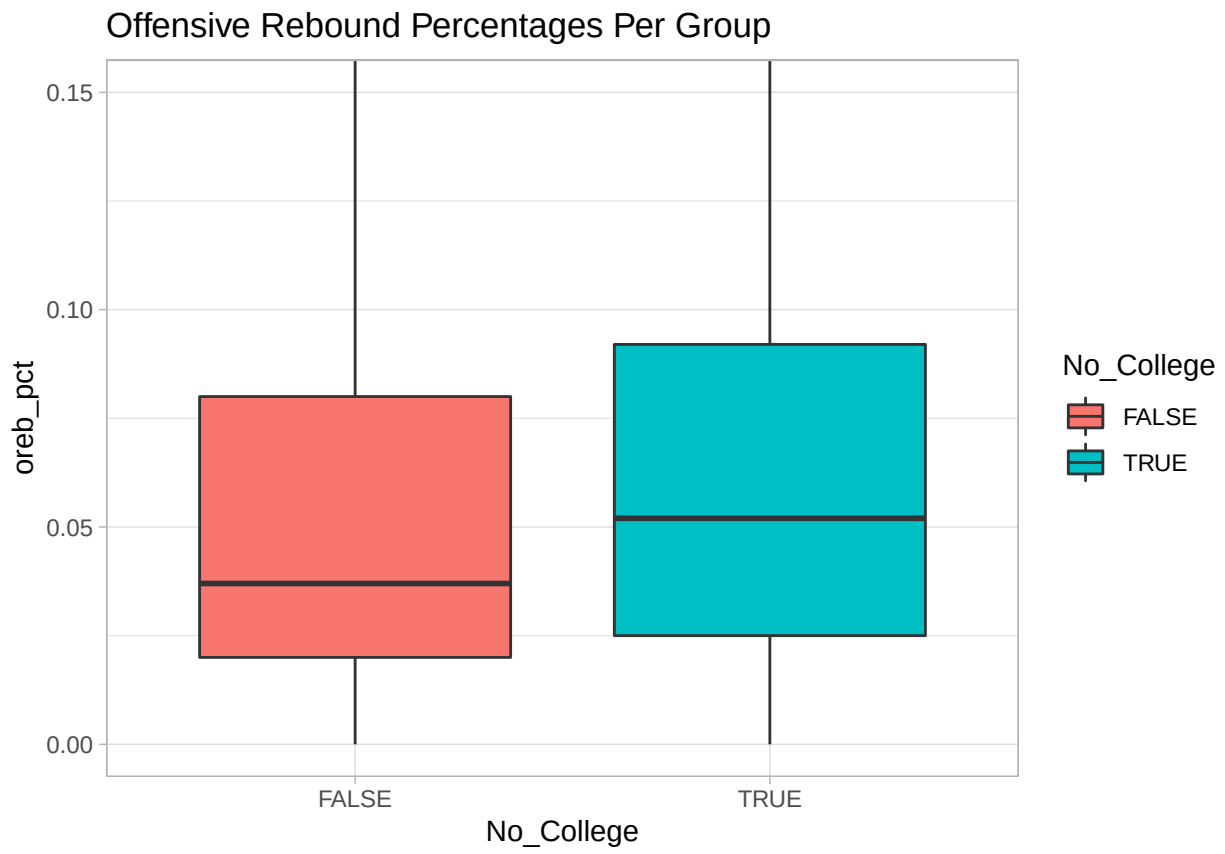
```
ggplot(all_season_sampled, aes(No_College, net_rating, fill = No_College)) + geom_boxplot() +
coord_cartesian(ylim = c(-100,120)) + theme_light() + labs(title = "Net Rating Per Group (With Outliers)
```



Con net rating, vemos que los dos grupos por lo normal estan a la par. Sin embargo, los outliers son peores en el grupo de No_College y mejores en el grupo Universitario

Analisis Rebounds Ofensivos

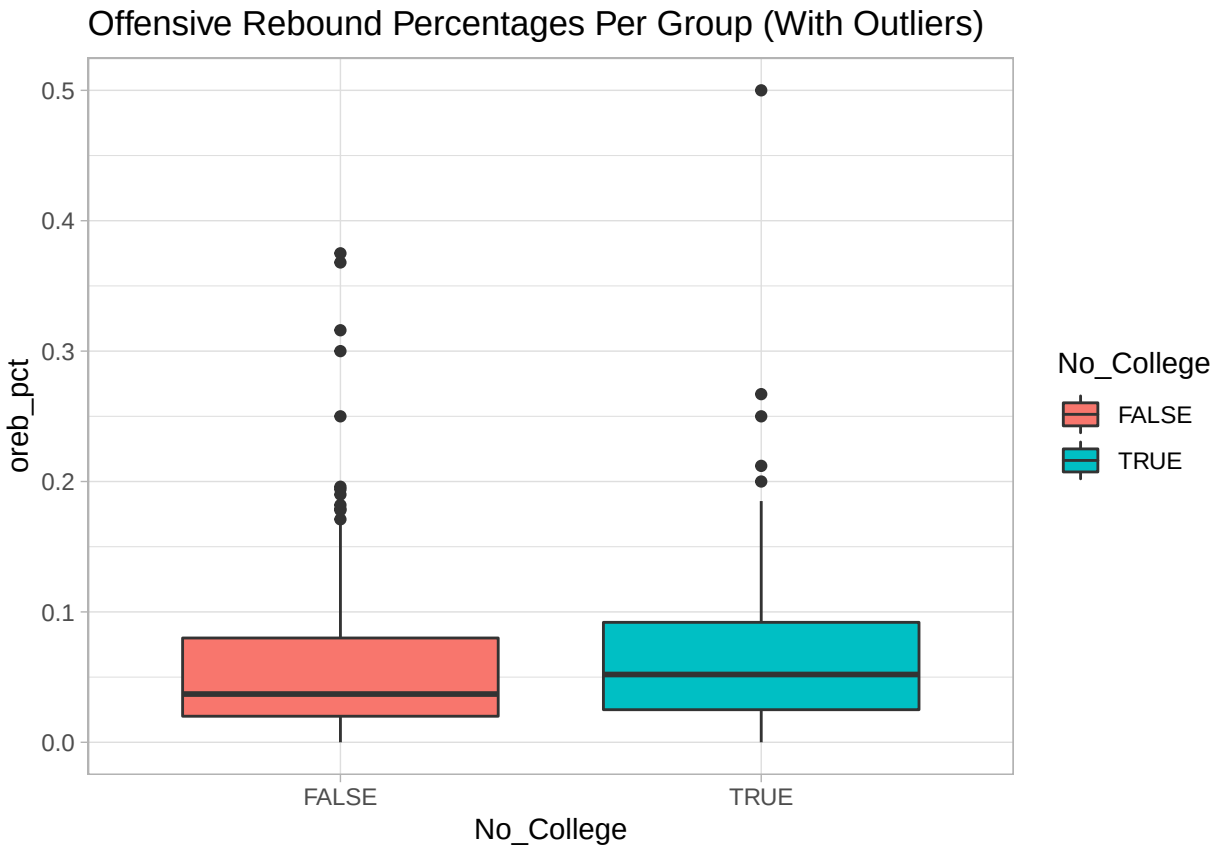
```
ggplot(all_season_sampled, aes(No_College, oreb_pct, fill = No_College)) +
geom_boxplot() + theme_light() + labs(title = "Offensive Rebound Percentages Per Group")+
coord_cartesian(ylim = c(0,.15))
```




```
list("Mediana No-College" = median(all_season_none$oreb_pct),
  "Mediana College" = median(college_sample$oreb_pct))
```

```
## $`Mediana No-College`
## [1] 0.052
##
## $`Mediana College`
## [1] 0.037
```

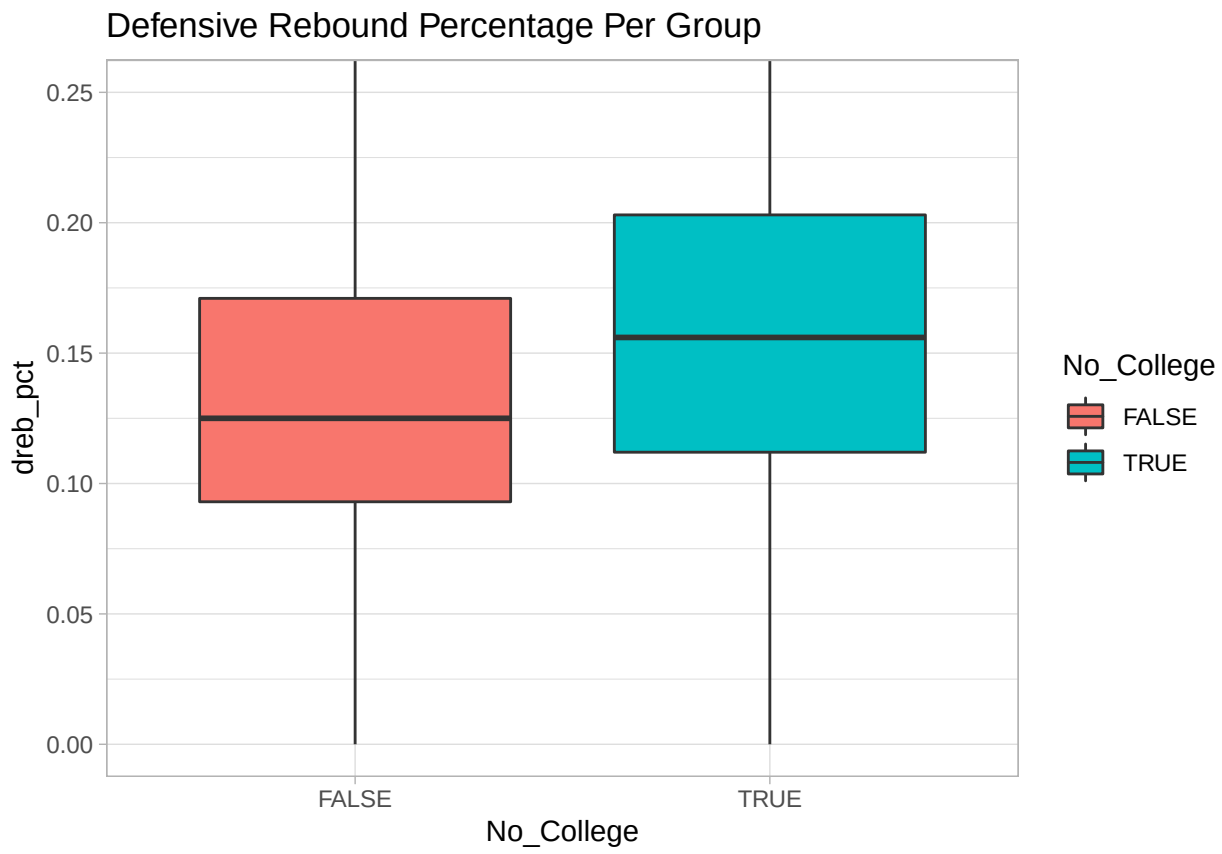
```
ggplot(all_season_sampled, aes(No_College, oreb_pct, fill = No_College)) +
geom_boxplot() + theme_light() +
labs(title = "Offensive Rebound Percentages Per Group (With Outliers)")+
coord_cartesian(ylim = c(0,.50))
```



En el caso de los rebounds ofensivos, el grupo **No Universitario** cuenta con la ventaja. Tanto al promedio como en los outliers

Analisis de Rebounds Defensivos

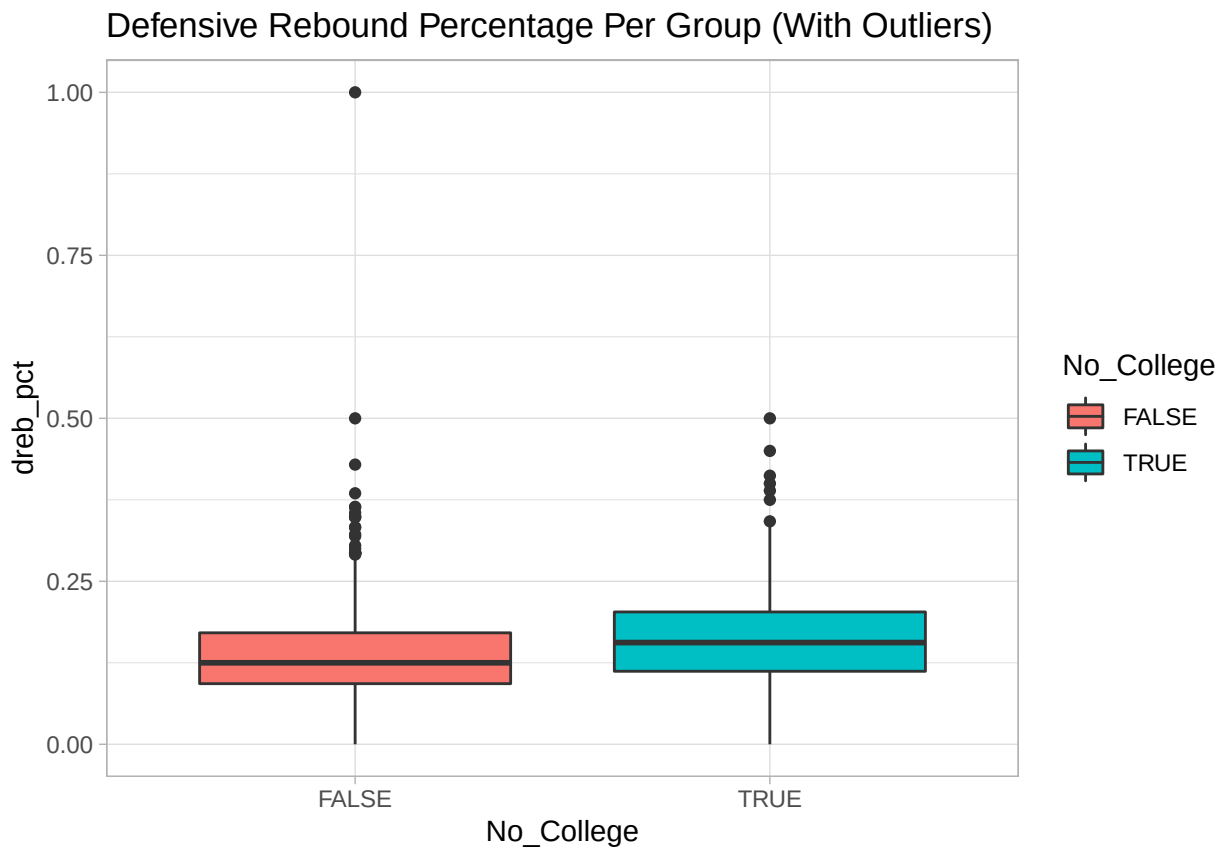
```
ggplot(all_season_sampled, aes(No_College, dreb_pct, fill = No_College)) +
geom_boxplot() + coord_cartesian(ylim = c(0,.25)) + theme_light() +
labs(title = "Defensive Rebound Percentage Per Group")
```




```
list("Mediana No-College" = median(all_season_none$dreb_pct),
  "Mediana College" = median(college_sample$dreb_pct))
```

```
## $`Mediana No-College`
## [1] 0.156
##
## $`Mediana College`
## [1] 0.125
```

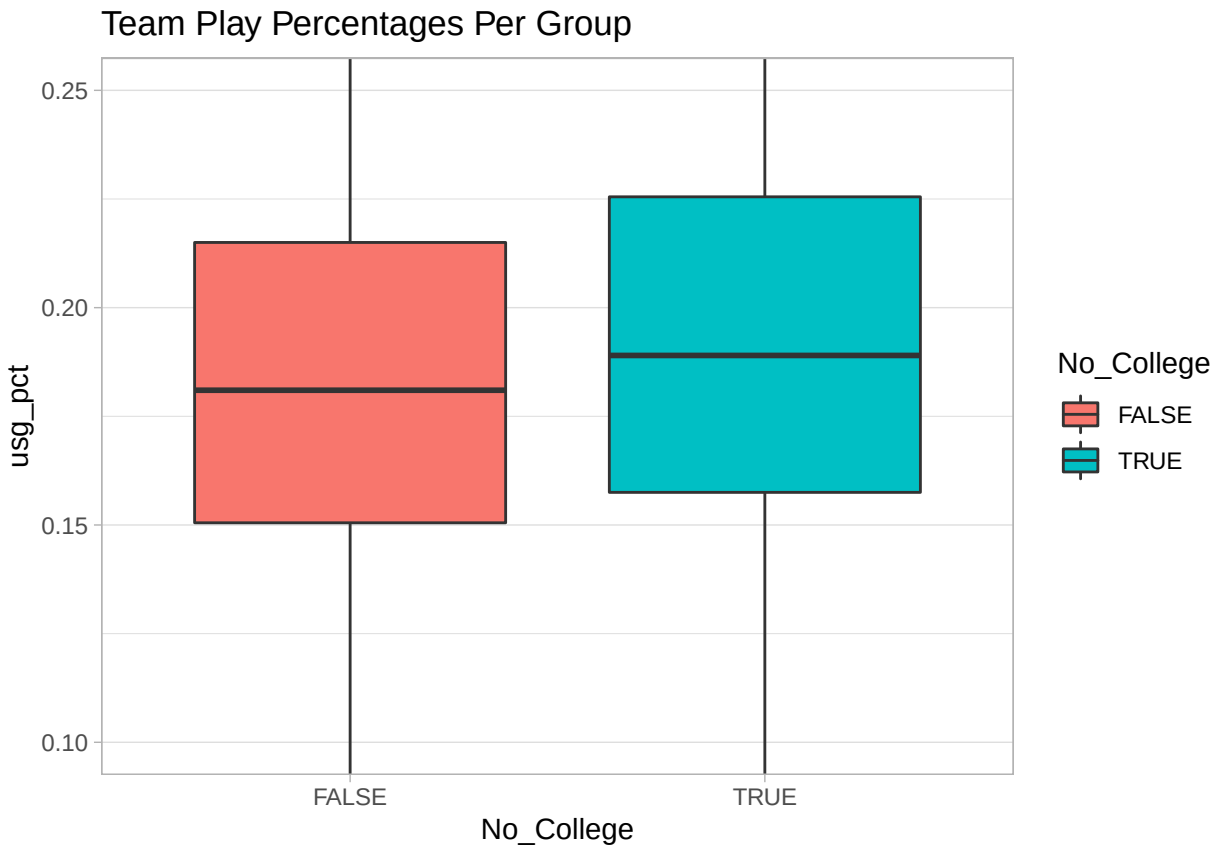
```
ggplot(all_season_sampled, aes(No_College, dreb_pct, fill = No_College)) +
geom_boxplot() + theme_light() +
labs(title = "Defensive Rebound Percentage Per Group (With Outliers)")
```



En el caso de los rebounds ofensivos, el grupo **No Universitario** cuenta con la ventaja al promedio, y los outliers siendo generalmente ciertos

Analisis de Team Plays

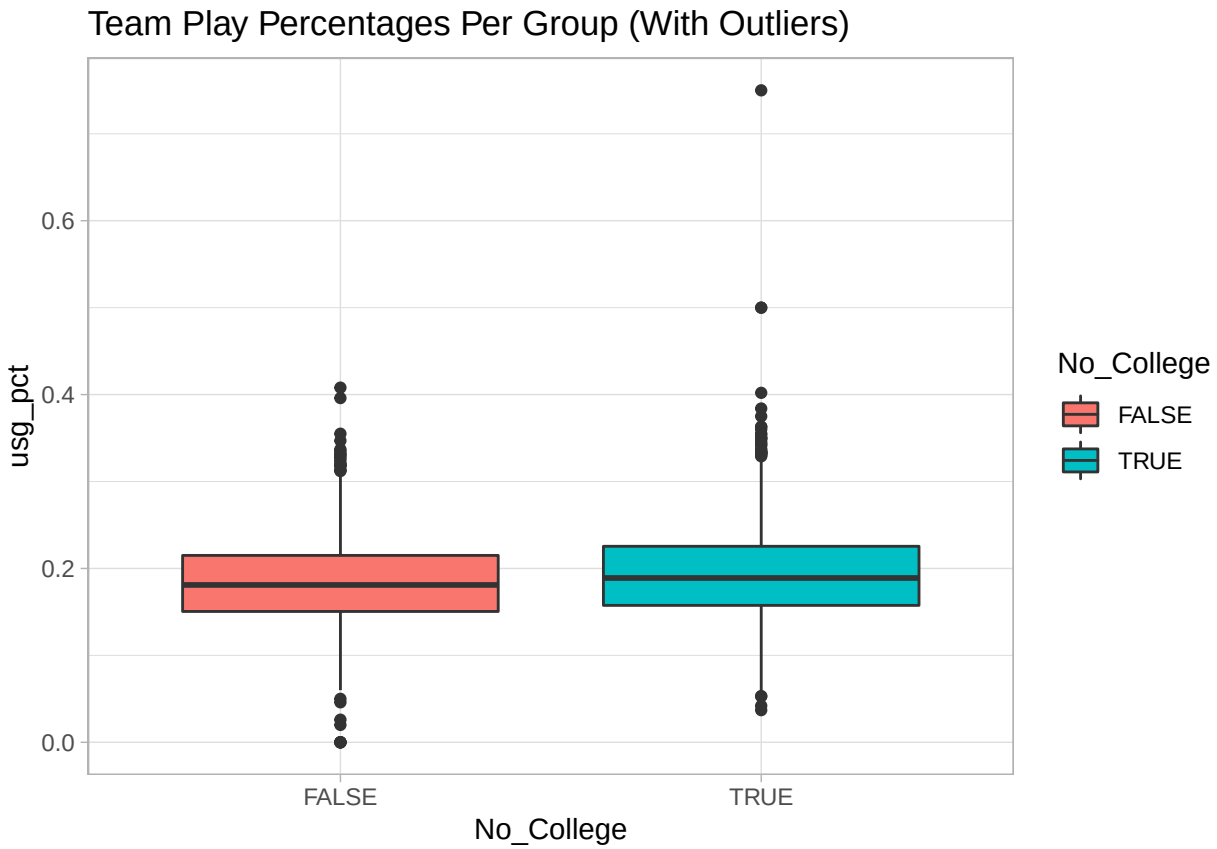
```
ggplot(all_season_sampled, aes(No_College, usg_pct, fill = No_College)) + geom_boxplot() +
coord_cartesian(ylim = c(0.1,0.25)) +theme_light() + labs(title = "Team Play Percentages Per Group")
```




```
list("Mediana No-College" = median(all_season_none$usg_pct),
  "Mediana College" = median(college_sample$usg_pct))
```

```
## $`Mediana No-College`
## [1] 0.189
##
## $`Mediana College`
## [1] 0.181
```

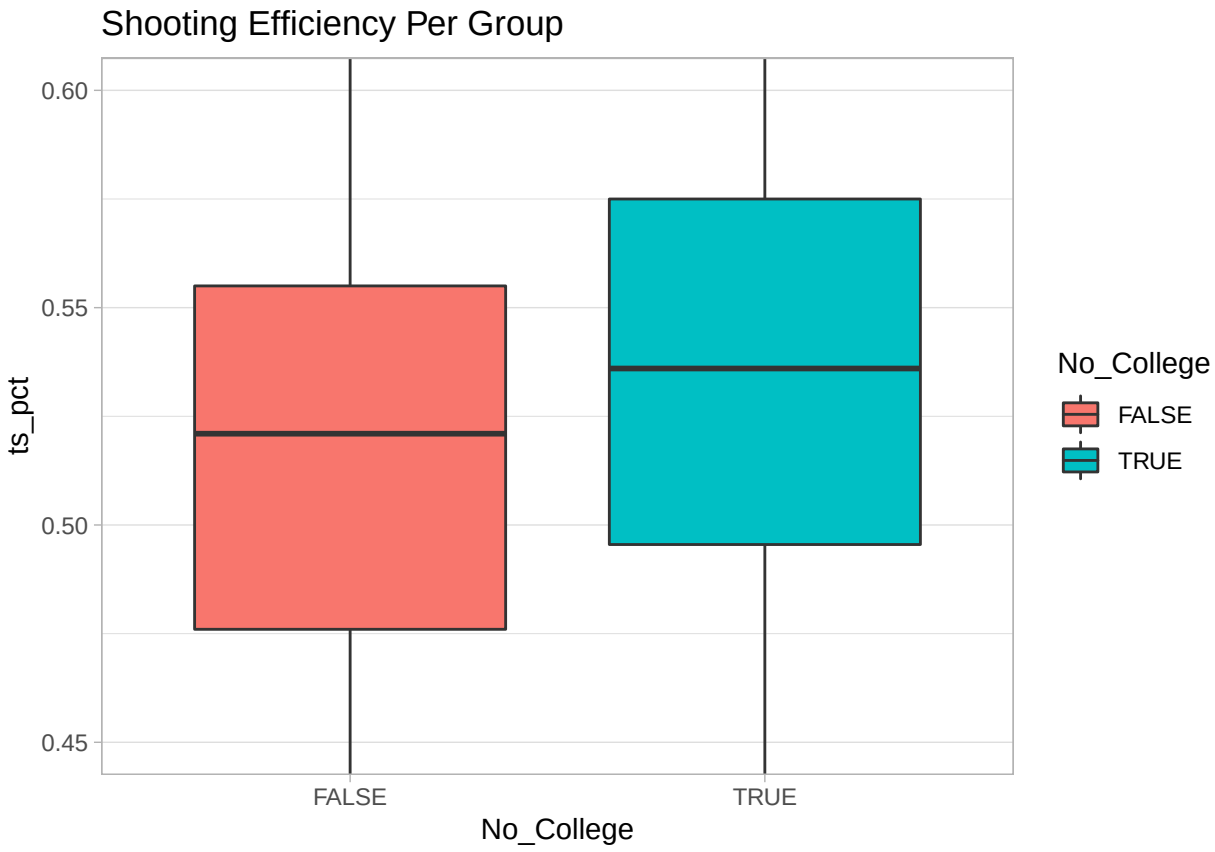
```
ggplot(all_season_sampled, aes(No_College, usg_pct, fill = No_College)) + geom_boxplot() +
coord_cartesian(ylim = c(0,0.75)) +theme_light() +
labs(title = "Team Play Percentages Per Group (With Outliers)")
```



En este caso, ambos grupos son mas o menos igual al promedio. Pero el **No Universitario** tiene ventaja en terminos de outliers.

Analisis de Shooting Efficiency

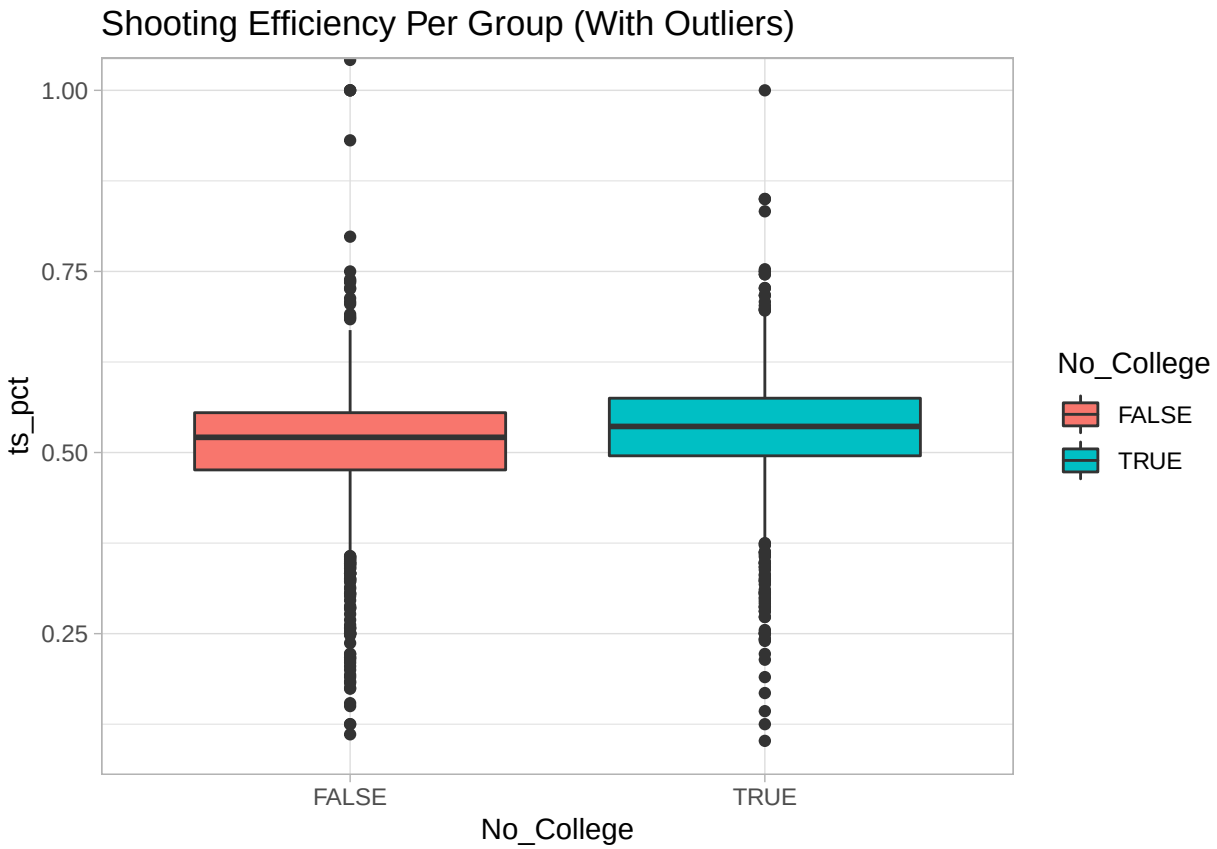
```
ggplot(all_season_sampled, aes(No_College, ts_pct, fill = No_College)) + geom_boxplot() +
coord_cartesian(ylim = c(0.45,0.6)) + theme_light() + labs(title = "Shooting Efficiency Per Group")
```




```
list("Mediana No-College" = median(all_season_none$ts_pct),
  "Mediana College" = median(college_sample$ts_pct))
```

```
## $`Mediana No-College`
## [1] 0.536
##
## $`Mediana College`
## [1] 0.521
```

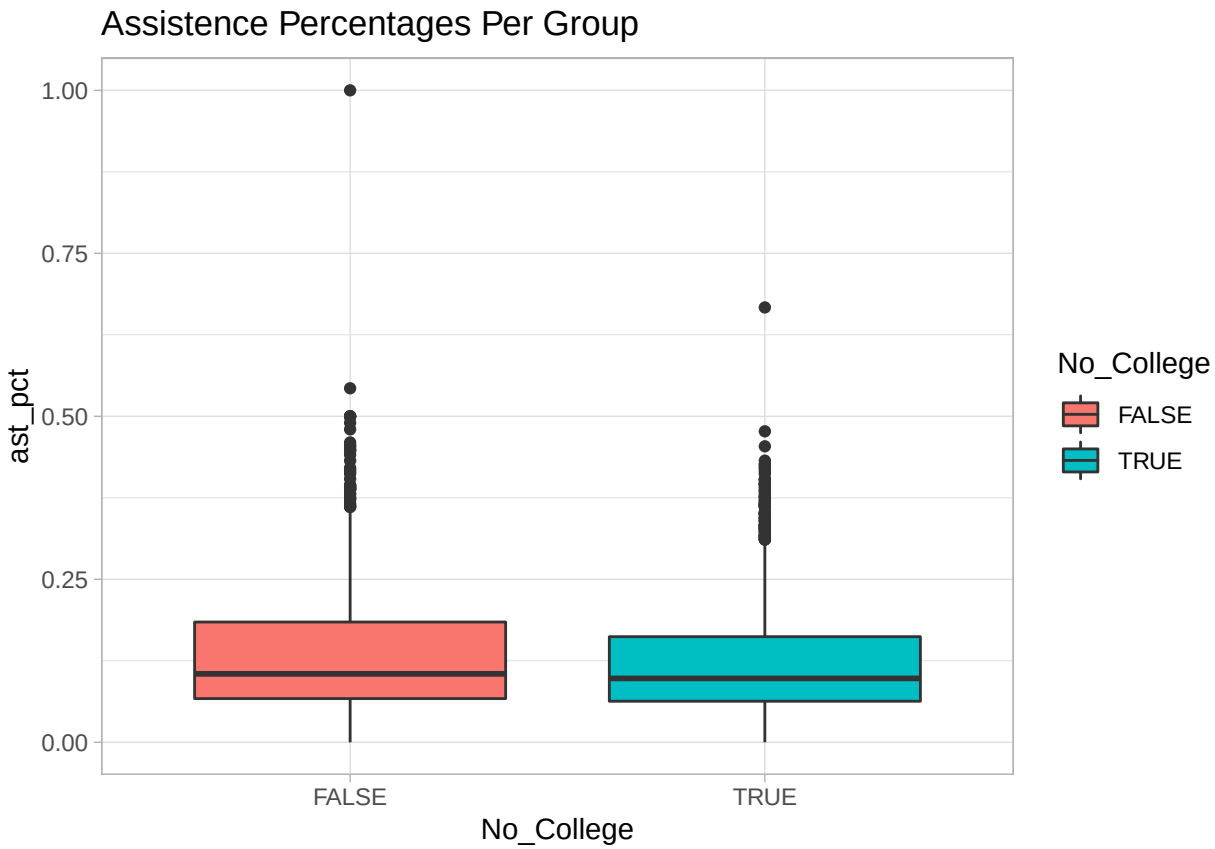
```
ggplot(all_season_sampled, aes(No_College, ts_pct, fill = No_College)) + geom_boxplot() +
coord_cartesian(ylim = c(0.10,1.0)) + theme_light() +
labs(title = "Shooting Efficiency Per Group (With Outliers)")
```



Aqui vemos un patron en donde por lo normal, los de No_College salen con un score mas alto, pero con muchos mas outliers. En esta ocasion los outliers parecen ser mas o menos a la par

Porcentaje de Asistencias

```
ggplot(all_season_sampled, aes(No_College, ast_pct, fill = No_College)) +
geom_boxplot()  + theme_light() + labs(title = "Assistence Percentages Per Group")
```



En terminos del porcentaje de asistencia, el grupo **Universitario** cuenta con ventajas en tanto el promedio como los outliers

Significancia de Variables

Considerando lo patrones qu hemos visto hasta ahora en los boxplot, hara sentido ver si se puede crear un modelo a base de estas. Deseamos ver cuales variables prueban ser mas significativas a las hora de predecir el nivel educativo de un jugador.

Para esto, crearemos un modelo de regresion logistica. En dicho modelo, se tomara la variable 'No_College' como la variable respuesta. Incialmente el modelo tendra al resto de las variables como predictores. Pero mediante que vallamos identificando cuales factores suelen ser mas significativos, ajustaremos el modelo.

```
glm.inicial <- glm(
    No_College ~., data = all_season_sampled[-c(1:3,6:11,22)],
    family = "binomial")

summary(glm.inicial)
```

```
##
## Call:
## glm(formula = No_College ~ ., family = "binomial", data = all_season_sampled[-c(1:3,
##     6:11, 22)])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.3347  -1.0216   0.0459   1.0071   3.6937
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -25.245788   1.483787 -17.014  < 2e-16 ***
## age            -0.078853   0.008963  -8.797  < 2e-16 ***
## player_height   0.124513   0.007288  17.084  < 2e-16 ***
## gp              0.006364   0.002030   3.135 0.001719 **
## pts            -0.053207   0.019091  -2.787 0.005320 **
## reb             0.003395   0.035467   0.096 0.923750
## ast             0.056477   0.058633   0.963 0.335429
## net_rating      0.010398   0.003827   2.717 0.006589 **
## oreb_pct       -7.406184   1.415460  -5.232 1.67e-07 ***
## dreb_pct        0.538011   1.035605   0.520 0.603403
## usg_pct         5.317728   1.249793   4.255 2.09e-05 ***
## ts_pct          1.826898   0.486987   3.751 0.000176 ***
## ast_pct         4.192627   1.008607   4.157 3.23e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4755.0  on 3429  degrees of freedom
## Residual deviance: 4146.4  on 3417  degrees of freedom
## AIC: 4172.4
##
## Number of Fisher Scoring iterations: 4
```

Explicacion

Mirando los resultados de este analisis **preliminar**, podemos concluir que las siguientes variables son significativas a la hora de determinar si un jugador es universitario o no:

age player_height oreb_pct usg_pct ts_pct ast_pct

Para poder realizar un analisis mas preciso, crearemos otro modelo con solamente esta variables

```
glm.trimmed <- glm(
    No_College ~age + player_height + oreb_pct + usg_pct + ts_pct +
      ast_pct, data = all_season_sampled[-c(1:3,6:11,22)],
    family = "binomial")

summary(glm.trimmed)
```

```
##
## Call:
## glm(formula = No_College ~ age + player_height + oreb_pct + usg_pct +
##     ts_pct + ast_pct, family = "binomial", data = all_season_sampled[-c(1:3,
##     6:11, 22)])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4445  -1.0255   0.0605   1.0222   3.8068
##
```

```
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -24.337208   1.344957 -18.095  < 2e-16 ***
## age            -0.078172   0.008825  -8.858  < 2e-16 ***
## player_height   0.122618   0.006698  18.307  < 2e-16 ***
## oreb_pct       -6.361440   1.193232  -5.331 9.75e-08 ***
## usg_pct         2.155769   0.783523   2.751  0.00593 **
## ts_pct          1.777650   0.412712   4.307 1.65e-05 ***
## ast_pct         4.814332   0.580567   8.292  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4755.0  on 3429  degrees of freedom
## Residual deviance: 4171.5  on 3423  degrees of freedom
## AIC: 4185.5
##
## Number of Fisher Scoring iterations: 4
```

Nota

Se debe notar que **usg_pct**, lo cual era originalmente significativa, no es significativa bajo este modelo. Antes de decidir si nos quedamos con ella o no, pasaremos a otra fase de determinacion.

Backward Stepwise

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
backward_step <- stepAIC(glm.inicial, direction = "backward")
```

```
## Start:  AIC=4172.37
## No_College ~ age + player_height + gp + pts + reb + ast + net_rating +
##     oreb_pct + dreb_pct + usg_pct + ts_pct + ast_pct
##
##                 Df Deviance    AIC
## - reb            1   4146.4 4170.4
## - dreb_pct       1   4146.6 4170.6
## - ast            1   4147.3 4171.3
## <none>               4146.4 4172.4
## - net_rating     1   4153.9 4177.9
## - pts            1   4154.3 4178.3
## - gp             1   4156.3 4180.3
```

```
## - ts_pct         1   4160.6 4184.6
## - ast_pct        1   4164.5 4188.5
## - usg_pct        1   4165.9 4189.9
## - oreb_pct       1   4176.9 4200.9
## - age            1   4226.5 4250.5
## - player_height  1   4490.5 4514.5
##
## Step:  AIC=4170.38
## No_College ~ age + player_height + gp + pts + ast + net_rating +
##     oreb_pct + dreb_pct + usg_pct + ts_pct + ast_pct
##
##                 Df Deviance    AIC
## - dreb_pct       1   4146.9 4168.9
## - ast            1   4147.3 4169.3
## <none>               4146.4 4170.4
## - net_rating     1   4153.9 4175.9
## - gp             1   4156.8 4178.8
## - pts            1   4159.5 4181.5
## - ts_pct         1   4160.6 4182.6
## - ast_pct        1   4164.9 4186.9
## - usg_pct        1   4168.7 4190.7
## - oreb_pct       1   4181.4 4203.4
## - age            1   4226.5 4248.5
## - player_height  1   4496.1 4518.1
##
## Step:  AIC=4168.87
## No_College ~ age + player_height + gp + pts + ast + net_rating +
##     oreb_pct + usg_pct + ts_pct + ast_pct
##
##                 Df Deviance    AIC
## - ast            1   4147.8 4167.8
## <none>               4146.9 4168.9
## - net_rating     1   4154.7 4174.7
## - gp             1   4157.2 4177.2
## - pts            1   4159.7 4179.7
## - ts_pct         1   4161.1 4181.1
## - ast_pct        1   4165.5 4185.5
## - usg_pct        1   4169.2 4189.2
## - oreb_pct       1   4183.6 4203.6
## - age            1   4226.5 4246.5
## - player_height  1   4574.8 4594.8
##
## Step:  AIC=4167.79
## No_College ~ age + player_height + gp + pts + net_rating + oreb_pct +
##     usg_pct + ts_pct + ast_pct
##
##                 Df Deviance    AIC
## <none>               4147.8 4167.8
## - net_rating     1   4155.6 4173.6
## - gp             1   4158.5 4176.5
## - ts_pct         1   4161.5 4179.5
## - pts            1   4162.1 4180.1
## - usg_pct        1   4169.4 4187.4
## - oreb_pct       1   4184.4 4202.4
```

```
## - ast_pct        1   4220.9 4238.9
## - age            1   4227.1 4245.1
## - player_height  1   4575.4 4593.4
```

```
summary(backward_step)
```

```
##
## Call:
## glm(formula = No_College ~ age + player_height + gp + pts + net_rating +
##     oreb_pct + usg_pct + ts_pct + ast_pct, family = "binomial",
##     data = all_season_sampled[-c(1:3, 6:11, 22)])
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.429  -1.020   0.056   1.010   3.649
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -25.562406   1.415432 -18.060  < 2e-16 ***
## age            -0.078193   0.008926  -8.760  < 2e-16 ***
## player_height   0.126359   0.006828  18.505  < 2e-16 ***
## gp              0.006492   0.001990   3.262 0.001105 **
## pts            -0.042870   0.011448  -3.745 0.000181 ***
## net_rating      0.010462   0.003791   2.760 0.005785 **
## oreb_pct       -7.036461   1.216436  -5.784 7.27e-09 ***
## usg_pct         4.922984   1.087502   4.527 5.99e-06 ***
## ts_pct          1.781353   0.483595   3.684 0.000230 ***
## ast_pct         4.990369   0.598660   8.336  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4755.0  on 3429  degrees of freedom
## Residual deviance: 4147.8  on 3420  degrees of freedom
## AIC: 4167.8
##
## Number of Fisher Scoring iterations: 4
```

Ahora tomamos el Forward Stepwise a consideracion tambien

Forward Stepwise

```
forward_step<-stepAIC(glm.inicial, direction = "forward")
```

```
## Start:  AIC=4172.37
## No_College ~ age + player_height + gp + pts + reb + ast + net_rating +
##     oreb_pct + dreb_pct + usg_pct + ts_pct + ast_pct
```

```
summary(forward_step)
```

```
##
## Call:
## glm(formula = No_College ~ age + player_height + gp + pts + reb +
##     ast + net_rating + oreb_pct + dreb_pct + usg_pct + ts_pct +
##     ast_pct, family = "binomial", data = all_season_sampled[-c(1:3,
##     6:11, 22)])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3347  -1.0216   0.0459   1.0071   3.6937
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -25.245788   1.483787 -17.014  < 2e-16 ***
## age            -0.078853   0.008963  -8.797  < 2e-16 ***
## player_height   0.124513   0.007288  17.084  < 2e-16 ***
## gp              0.006364   0.002030   3.135 0.001719 **
## pts            -0.053207   0.019091  -2.787 0.005320 **
## reb             0.003395   0.035467   0.096 0.923750
## ast             0.056477   0.058633   0.963 0.335429
## net_rating      0.010398   0.003827   2.717 0.006589 **
## oreb_pct       -7.406184   1.415460  -5.232 1.67e-07 ***
## dreb_pct        0.538011   1.035605   0.520 0.603403
## usg_pct         5.317728   1.249793   4.255 2.09e-05 ***
## ts_pct          1.826898   0.486987   3.751 0.000176 ***
## ast_pct         4.192627   1.008607   4.157 3.23e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4755.0  on 3429  degrees of freedom
## Residual deviance: 4146.4  on 3417  degrees of freedom
## AIC: 4172.4
##
## Number of Fisher Scoring iterations: 4
```

Comparando los Tres Modelos

Antes de este paso, creamos tres modelos mediante metodos diferentes. El modelo trimmed se baso en manualmente mirar los pvalues de las variables, y quedandose con solamente esas que eran significativas segun el sistema. Los otros dos modelos se utilizaron mediante AIC tanto con el metodo **backwards** y **forwards**.

Ahora comparemos a los tres modelos que generamos, para determinar con cual nos quedamos.

```
list("Metodo Manual" = summary(glm.trimmed),
     "Metodo Backward" = summary(backward_step),
     "Metodo Forward" = summary(forward_step))
```

```
## $`Metodo Manual`
```

```
##
## Call:
## glm(formula = No_College ~ age + player_height + oreb_pct + usg_pct +
##     ts_pct + ast_pct, family = "binomial", data = all_season_sampled[-c(1:3,
##     6:11, 22)])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4445  -1.0255   0.0605   1.0222   3.8068
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -24.337208   1.344957 -18.095  < 2e-16 ***
## age            -0.078172   0.008825  -8.858  < 2e-16 ***
## player_height   0.122618   0.006698  18.307  < 2e-16 ***
## oreb_pct       -6.361440   1.193232  -5.331 9.75e-08 ***
## usg_pct         2.155769   0.783523   2.751  0.00593 **
## ts_pct          1.777650   0.412712   4.307 1.65e-05 ***
## ast_pct         4.814332   0.580567   8.292  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4755.0  on 3429  degrees of freedom
## Residual deviance: 4171.5  on 3423  degrees of freedom
## AIC: 4185.5
##
## Number of Fisher Scoring iterations: 4
##
##
## $`Metodo Backward`
##
## Call:
## glm(formula = No_College ~ age + player_height + gp + pts + net_rating +
##     oreb_pct + usg_pct + ts_pct + ast_pct, family = "binomial",
##     data = all_season_sampled[-c(1:3, 6:11, 22)])
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.429  -1.020   0.056   1.010   3.649
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -25.562406   1.415432 -18.060  < 2e-16 ***
## age            -0.078193   0.008926  -8.760  < 2e-16 ***
## player_height   0.126359   0.006828  18.505  < 2e-16 ***
## gp              0.006492   0.001990   3.262 0.001105 **
## pts            -0.042870   0.011448  -3.745 0.000181 ***
## net_rating      0.010462   0.003791   2.760 0.005785 **
## oreb_pct       -7.036461   1.216436  -5.784 7.27e-09 ***
## usg_pct         4.922984   1.087502   4.527 5.99e-06 ***
## ts_pct          1.781353   0.483595   3.684 0.000230 ***
## ast_pct         4.990369   0.598660   8.336  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4755.0  on 3429  degrees of freedom
## Residual deviance: 4147.8  on 3420  degrees of freedom
## AIC: 4167.8
##
## Number of Fisher Scoring iterations: 4
##
##
## $`Metodo Forward`
##
## Call:
## glm(formula = No_College ~ age + player_height + gp + pts + reb +
##     ast + net_rating + oreb_pct + dreb_pct + usg_pct + ts_pct +
##     ast_pct, family = "binomial", data = all_season_sampled[-c(1:3,
##     6:11, 22)])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3347  -1.0216   0.0459   1.0071   3.6937
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -25.245788   1.483787 -17.014  < 2e-16 ***
## age            -0.078853   0.008963  -8.797  < 2e-16 ***
## player_height   0.124513   0.007288  17.084  < 2e-16 ***
## gp              0.006364   0.002030   3.135 0.001719 **
## pts            -0.053207   0.019091  -2.787 0.005320 **
## reb             0.003395   0.035467   0.096 0.923750
## ast             0.056477   0.058633   0.963 0.335429
## net_rating      0.010398   0.003827   2.717 0.006589 **
## oreb_pct       -7.406184   1.415460  -5.232 1.67e-07 ***
## dreb_pct        0.538011   1.035605   0.520 0.603403
## usg_pct         5.317728   1.249793   4.255 2.09e-05 ***
## ts_pct          1.826898   0.486987   3.751 0.000176 ***
## ast_pct         4.192627   1.008607   4.157 3.23e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4755.0  on 3429  degrees of freedom
## Residual deviance: 4146.4  on 3417  degrees of freedom
## AIC: 4172.4
##
## Number of Fisher Scoring iterations: 4
```

Analisis Comparativo Comparando todos los modelos, ahora escogeremos uno. El modelo con mas variables es el de **forward_step**, cual incluye tambien cinco variables no significativas **(Nivel de Significacion: 0.001)**. Para propositos de parsimonia, descartaremos el modelo forward_step ya que, por lo que

vemos, cuenta condemasiadas variables.

Ahora la decision caera entre el metodo de **backward_step** o **glm.trimmed**. Tanto el residual deviance como el AIC de ambos modelos es bastante similar. Lo cual hace que estos factores en esta circunstancia no sean tan crucial a la hora de tomar la decision. Aunque **backward_step** cuenta con menos variables que forward_step, todavia cuenta con mas variables que glm.trimmed. No solamente eso, pero en **backward_step** tres de estas variables no caen dentro de los parametros de significacion que hemos definido.

Sin embargo, en **glm.trimmed** vemos que hasta ahora es el modelo mas parsimonio de las tres opciones– ya que cuenta con la menor cantidad de variables. Otra ventaja de **glm_trimmed** es que de todas sus variables, solamente 1 no cumple con nuestro criterio de significacion.

Todos estos factores nos llevan a concluir que de los 3 modelos provistos, el modelo **glm.trimmed** es el ideal para el proposito de nuestro analisis.

Prediciendo la Experiencia Universitaria de Jugadores

Ya que tenemos nuestro modelo logistico listo, vamos a empezar el proceso de entrenamiento para el modelo.

Deseo recordar que, aunque anteriormente sacamos una muestra de los datos para poder comparar jugadores tanto universitarios como no universitarios, tendremos que realizar un **2ndo Muestreo**. Esto se debe a que, aunque en el primer muestreo se saco muestra de los datos universitarios, en el grupo no-universitario nos quedamos con la poblacion.

Esto era adecuado para el tipo de analisis que realizamos anteriormente. Pero para poder aplicar un modelo de prediccion, tendremos que cambiar esto

Fase 1: Prediccion del Sample Dataset

```
ID_sample_data <- all_season_sampled %>% # Creando una columna de ID para facilitar nuestro anti_join
    mutate(ID = row_number()
      )

train_No_College <-  ID_sample_data %>%
    filter(No_College == 1) %>%
    slice_sample(n = 1715 * 0.5)

train_Yes_College <- ID_sample_data %>%
    filter(No_College == 0) %>%
    slice_sample(n = 1715 * 0.5)

train <- rbind(train_Yes_College, train_No_College) %>%
    slice(sample(1:n()))  # Esto se hace para que los factores esten distribuidos aleatoriamente

test <- ID_sample_data %>%
      anti_join(train, by = "ID")
```

```
glm.trimmed.phase1 <- glm(
    No_College ~age + player_height + oreb_pct + usg_pct + ts_pct + ast_pct, data = train[-c(1:3,6:11,2
    family = "binomial")
```

```
glm.probs <- predict(glm.trimmed.phase1, test,
type = "response")
```

Preparando el Modelo de Prediccion Ahora creamos la tabla

```
glm.pred <- rep("FALSE", dim(test)[1])
glm.pred[glm.probs > .5] <- "TRUE"
list("Table" = table(glm.pred, test$No_College),
"Proporciones" = prop.table(table(glm.pred, test$No_College)),
"Porciento Correcto" = (554+595)/1716, "Error" = 1-((554+595)/1716))
```

```
## $Table
##
## glm.pred FALSE TRUE
##    FALSE   581  271
##    TRUE    277  587
##
## $Proporciones
##
## glm.pred     FALSE      TRUE
##    FALSE 0.3385781 0.1579254
##    TRUE  0.1614219 0.3420746
##
## $`Porciento Correcto`
## [1] 0.6695804
##
## $Error
## [1] 0.3304196
```

Analisis Mirando los resultados, por ser nuestro primer modelo, es bastante impresionante. El modelo salio correcto **67%** de las veces. Esto significa que el modelo es mejor que meramente escoger al azar por **17%**.

Esto nos promete mucho en el futuro. Sin embargo, se debe recalcar que todavia hay otras mejoras que se le podrian aplicar al modelo. Por ejemplo, si se acuerdan bien, en nuestro modelo **glm.trimmed** (el que escogimos) habia una de las variables que no caia dentro de nuestro nivel de significacion **0.01**. Aqui incluire el codigo para demostrarlo.

```
summary(glm.trimmed)
```

```
##
## Call:
## glm(formula = No_College ~ age + player_height + oreb_pct + usg_pct +
##     ts_pct + ast_pct, family = "binomial", data = all_season_sampled[-c(1:3,
##     6:11, 22)])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4445  -1.0255   0.0605   1.0222   3.8068
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)   -24.337208   1.344957 -18.095  < 2e-16 ***
## age            -0.078172   0.008825  -8.858  < 2e-16 ***
## player_height   0.122618   0.006698  18.307  < 2e-16 ***
## oreb_pct       -6.361440   1.193232  -5.331 9.75e-08 ***
## usg_pct         2.155769   0.783523   2.751  0.00593 **
## ts_pct          1.777650   0.412712   4.307 1.65e-05 ***
## ast_pct         4.814332   0.580567   8.292  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4755.0  on 3429  degrees of freedom
## Residual deviance: 4171.5  on 3423  degrees of freedom
## AIC: 4185.5
##
## Number of Fisher Scoring iterations: 4
```

Esta variable viene siendo **usg_pct**, lo cual es el porcentaje de jugadas de equipo que el jugador obtuvo. Posiblemente si lo eliminamos del modelo, tendremos mejor precision?

Veamos

```
glm.trimmed.phase1 <- glm(
    No_College ~age + player_height + oreb_pct + ts_pct + ast_pct, data = train[-c(1:3,6:11,22,24)],
    family = "binomial")

glm.probs <- predict(glm.trimmed.phase1, test,
type = "response")
```

Eliminando usg_pct

```
glm.pred <- rep("FALSE", dim(test)[1])
glm.pred[glm.probs > .5] <- "TRUE"
list("Table" = table(glm.pred, test$No_College),
  "Proporciones" = prop.table(table(glm.pred, test$No_College)),
  "Porciento Correcto" = (557+601)/1716, "Error" = 1-((557+601)/1716))
```

La Tabla

```
## $Table
##
## glm.pred FALSE TRUE
##    FALSE   583  268
##    TRUE    275  590
##
## $Proporciones
##
```

```
## glm.pred     FALSE      TRUE
##    FALSE 0.3397436 0.1561772
##    TRUE  0.1602564 0.3438228
##
## $`Porciento Correcto`
## [1] 0.6748252
##
## $Error
## [1] 0.3251748
```

El incremento parece ser un marginal. Siguiendo los principios de la parsimonia, si con este modelo con menos variables adquirimos un resultado similar o marginalmente mejor, favoreceremos el modelo mas simple

Fase 2

Ahora veamos si podemos aplicar el modelo a los datos originales completos. Ojo, tendremos que modificar los datos originales para que tambien tengan la columna de No_College. Esto es debido que, aunque creamos dicha columna en la muestra original que tomamos, no fue algo que aplicamos a los datos originales.

```
all_season_mod <- all_season %>%
  mutate(No_College = ifelse(college == "None",TRUE, FALSE))

all_season_test <- all_season_mod %>%
  anti_join(train, by =c("...1"))

# Esta vez, utilizamos '...1' para el anti-join
# debido a que los datos originales no tienen el id column que creamos


glm.probs <- predict(glm.trimmed.phase1, all_season_test,
type = "response")
```

```
glm.pred <- rep("FALSE", dim(all_season_test)[1])
glm.pred[glm.probs > .5] <- "TRUE"
list("Table" = table(glm.pred, all_season_test$No_College),
"Proporciones" = prop.table(table(glm.pred, all_season_test$No_College)),
"Porciento Correcto" = (6096+590)/dim(all_season_test)[1],
"Error" = 1-((6096+590)/dim(all_season_test)[1]) )
```

La tabla

```
## $Table
##
## glm.pred FALSE TRUE
##    FALSE  6096  268
##    TRUE   3027  590
##
## $Proporciones
##
```

```
## glm.pred      FALSE       TRUE
##    FALSE 0.61076044 0.02685102
##    TRUE  0.30327622 0.05911231
##
## $`Porciento Correcto`
## [1] 0.6695374
##
## $Error
## [1] 0.3304626
```

Fase 3

Aqui, segun la sugerencia de un estudiante durante la presentacion, se pondra un factor para determinar si un jugador tiene experiencia universitaria o no. Dicho factor sera el factor de nacionalidad.

```
all_season_mod2 <- all_season %>%
  mutate(No_College = ifelse(college == "None",TRUE, FALSE),
         Foreigner = ifelse(country == "USA", FALSE, TRUE)
  )

train_No_College2 <-  all_season_mod2 %>%
    filter(No_College == 1) %>%
    slice_sample(n = 1715 * 0.5)

train_Yes_College2 <- all_season_mod2 %>%
    filter(No_College == 0) %>%
    slice_sample(n = 9980 * 0.5)

train_2 <- rbind(train_Yes_College2, train_No_College2) %>%
    slice(sample(1:n()))

all_season_test2 <- all_season_mod2 %>%
  anti_join(train_2, by =c("...1"))
```

```
glm.trimmed.phase3 <- glm(
    No_College ~age + player_height + oreb_pct + ts_pct + ast_pct + factor(Foreigner), data = train_2,
    family = "binomial")

glm.probs <- predict(glm.trimmed.phase3, all_season_test2,
type = "response")
```

```
glm.pred <- rep("FALSE", dim(all_season_test2)[1])
glm.pred[glm.probs > .5] <- "TRUE"
list("Table" = table(glm.pred, all_season_test2$No_College),
"Proporciones" = prop.table(table(glm.pred, all_season_test2$No_College)),
"Porciento Correcto" = (4730+491)/dim(all_season_test2)[1],
"Error" = 1-((4730+491)/dim(all_season_test2)[1]) )
```

La tabla

```
## $Table
##
## glm.pred FALSE TRUE
##    FALSE  4744  360
##    TRUE    246  498
##
## $Proporciones
##
## glm.pred      FALSE       TRUE
##    FALSE 0.81121751 0.06155951
##    TRUE  0.04206566 0.08515732
##
## $`Porciento Correcto`
## [1] 0.8920212
##
## $Error
## [1] 0.1079788
```

Fase 3.5

Que es esta fase? Pues, aqui buscaremos cuan grande es el efecto de las metricas de baloncesto versus solamente prediciendo a base de la nacionalidad. Asi veremos si los elementos de las metricas valen la pena incluir.

```
glm.trimmed.phase3.5 <- glm(
    No_College ~factor(Foreigner), data = train_2,
    family = "binomial")

glm.probs <- predict(glm.trimmed.phase3.5, all_season_test2,
type = "response")
```

```
glm.pred <- rep("FALSE", dim(all_season_test2)[1])
glm.pred[glm.probs > .5] <- "TRUE"
list("Table" = table(glm.pred, all_season_test2$No_College),
"Proporciones" = prop.table(table(glm.pred, all_season_test2$No_College)),
"Porciento Correcto" = (4633+557)/dim(all_season_test2)[1],
"Error" = 1-((4633+557)/dim(all_season_test2)[1]) )
```

La tabla

```
## $Table
##
## glm.pred FALSE TRUE
##    FALSE  4642  290
##    TRUE    348  568
##
## $Proporciones
##
## glm.pred      FALSE       TRUE
##    FALSE 0.79377565 0.04958960
```

```
##    TRUE  0.05950752 0.09712722
##
## $`Porciento Correcto`
## [1] 0.8867248
##
## $Error
## [1] 0.1132752
```

Analisis Esta iteracion del modelo parece dar resultados similares en la poblacion. Sin embargo, no es impresionante. Esto se debe a que, la gran mayoria de los datos son de observaciones de jugadores universitarios. Y el asunto con los datos es que la variable de interes eran los **No** Universitarios.

Tomando esto en cuenta, el **6.3%** de veces en donde se predijo correctamente que un estudiante era No Universitario es un resultado pesimo. Posiblemente esto significa que en los datos originales, no hay suficientes observaciones de jugadores **no* universitarios para realizar un proceso de entrenamiento adecuado.

Addendum

En la Fase 3 que se puso en el estudio, donde se agrego la nacionalidad, esto nos dio un modelo excelente. El porcentaje de bases en que el modelo predijo a los no universitarios correctamente era **8.4%**. Tomandondo en cuenta que los datos no universitarios son solo **15%** de los datos, esto quiere decir que el modelo salio correcto mas de mitad de las veces.

Sin embargo, en la fase 3.5, llegamos a una conclusion aun mas sorprendente. Solamente predecir a base de nacionalidad rinde iguales o mejores resultados, con un **9.5%** veces que salio correcto con este metodo la prediccion de si carecia de experiencia universitaria o no. Con esto, podemos decir conclusivamente que, a la de classificacion de experiencia universitaria, las metricas de juego **no son tan utiles**.

Conclusion

Cuando hacemos un 'breakdown' de los datos originales, nuestros intentos con el modelo de prediccion parecen ser relativamente eficaz. Se debe notar que en las **dos** versiones del modelo se pudo obtener una prediccion correcta que era significativamente mas alta que el proceso de meramente escoger al azar. En esa circumstancia en particular, vemos que prediccion a base de nuestro modelo logistico resulto en resultados buenas.

Sin embargo, a la hora de aplicar el modelo a la poblacion, el modelo tuvo mucha dificultad con predicciones buenas. Por lo que vemos, esto se debe a que, aunque si hay datos significativamente grandes de jugadores sin experiencia universitaria, no hay de lo suficiente **para entrenar un modelo de prediccion**. Se tendra que esperar al futuro cuando se generen mas datos. Ahi, se podran correr modelos mas robustos que cuentan con mas datos que se pueden incluir a la hora de prediccion.

Aunque, tambien se debe notar, como vimos en los **boxplots**, al nivel grafico/descriptivo, se puede ver que hay algunas diferencias entre las metricas de basketball entre el grupo Universitario y el No-Universitario. Por lo general, los No-Universitarios solian demostrar tener una ventaja decente sobre los Universitarios al nivel promedio. Sin embargo, cuando se buscaban lo que son los outliers, el grupo Universitario solia tener los outliers mas favorables. Mientres que el no-universitario tenia outliers en la dirrecion opuesta.

Esto es parte de la razon de porque entedemos que todavia hay algo de interes que se debe mirar mas a fondo en estos datos. Pero por ahora, se tendra que esperar a que tengamos mas datos y registros disponibles para nuestro uso.

Nota: Con los datos adquiridos en el addendum, podemos decir que el factor nacionalidad tuvo un efecto bastante grande e inesperado en nuestro modelo, en una direccion positiva. Con la fase 3, vimos que las metricas de juego no son tan efectivas a la hora de predecir si un jugador tiene experiencia universitaria o no–estas solo servian para esconder la variable que **si** importaba. . . la nacionalidad del jugador.

Bibliografia

1. Asif, R., Taha, M., Izhar, S., & Hasan, M. (2016). Football(Soccer) Analytics: A Case Study on the Availability and Limitations of Data for Football Analytics Research. International Journal of Computer Science and Information Security, 14(11).

2. Berri, D. J., Brook, S. L., & Fenn, A. J. (2011). From college to the pros: predicting the NBA amateur player draft. Journal of Productivity Analysis, 35(1), 25–35. https://doi.org/10.1007/s11123-010-0187-x

3. Cao, C. (2012). Sports Data Mining Technology Used in Basketball Outcome Prediction. Dublin Institute of Technology.

4. Evans, B. A. (2018). From college to the NBA: what determines a player's success and what characteristics are NBA franchises overlooking? Applied Economics Letters, 25(5), 300–304. https://doi.org/10.1080/13504851.2017.1319551

5. Fernandez, J., Camerino, O., Anguera, M. T., & Jonsson, G. K. (2009). Identifying and analyzing the construction and effectiveness of offensive plays in basketball by using systematic observation. Behavior Research Methods, 41(3), 719–730. https://doi.org/10.3758/brm.41.3.719

6. Franks, A., Miller, A., Bornn, L., & Goldsberry, K. (2015). Characterizing the spatial structure of defensive skill in professional basketball. The Annals of Applied Statistics, 9(1). https://doi.org/10.1214/14-aoas799

7. Hamdad, L., Benatchba, K., Belkham, F., & Chrairi, N. (2018). Basketball Analytics. Data Mining for Acquiring Performances. 6th IFIP International Conference on Computational Intelligence and Its Applications (CIIA), 13–24. https://hal.inria.fr/IFIP-AICT-522/hal-01913896

8. Hausman, J. A., & Leonard, G. K. (1997). Superstars in the National Basketball Association: Economic Value and Policy. Journal of Labor Economics, 15(4), 586–624. https://doi.org/10.1086/209839

9. Hollinger, J. (2003). Pro Basketball Prospectus: All-New 2003–04 Edition. Brassey's.

10. Hollinger, J. (2004). Pro Basketball Forecast: 2004–05 Edition (Pro Basketball Prospectus). Potomac Books.

11. Hollinger, J. (2005). Pro Basketball Forecast: 2005–06 Edition. Brassey's analytics for professional speed skating. Data Mining and Knowledge Discovery, 31(6), 1872–1902. https://doi.org/10.1007/s10618-017-0512-3

12. Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. T. (2007). A Starting Point for Analyzing Basketball Statistics. Journal of Quantitative Analysis in Sports, 3(3). https://doi.org/10.2202/1559-0410.1070

13. Lewis, M. (2004). Moneyball: The Art of Winning an Unfair Game (1st ed.). W. W. Norton & Company. Morgulev, E., Azar, O. H., & Lidor, R. (2018). Sports analytics and the big-data era. International Journal of Data Science and Analytics, 5(4), 213–222. https://doi.org/10.1007/s41060-017-0093-7

14. Moxley, J. H., & Towne, T. J. (2015). Predicting success in the National Basketball Association: Stability & Potential. Psychology of Sport and Exercise, 16, 128–136. https://doi.org/10.1016/j.psychsport.2014.07.003

15. Oliver, D. (2004). Basketball on Paper: Rules and Tools for Performance Analysis (Illustrated ed.). POTOMAC BOOKS.

16. Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. 2013 International Conference on Collaboration Technologies and Systems (CTS). https://doi.org/10.1109/cts.2013.6567202

17. Spurr, S. J. (2000). The Baseball Draft. Journal of Sports Economics, 1(1), 66–85. https://doi.org/10.1177/152700250000100106

18. Staffo, D. (1998). The Development of Professional Basketball in the United States with an Emphasis on the History of the NBA to its 50th Anniversary Season in 1996–97. Physical Educator, 55(1).

19. Walters, C., & Williams, T. (2012). To tank or not to tank? Evidence from the NBA. MIT Sloan Sports Conference March.