PRACTICA 2: LIMPIEZA Y VALIDACIÓN DE LOS DATOS

Jose Ignacio Bengoechea Isasa

7 de enero 2018

Contents

Partimos de dos datasets, donde tenemos el total de minutos, partidos, puntos, rebotes, asistencias, tapones y robos efectuados por los jugadores de la NBA y las jugadores de la WNBA, que son la liga masculina y femenina de baloncesto de Estados Unidos. Estos datos corresponden a la temporada 2016-17.

Nuestro objetivo es unificar estas fuentes en un unico dataset, limpiarlo, normalizarlo si es necesario, y establecer visualizaciones que nos permitan obtener información sobre la brecha salarial existente entre ambas ligas.

```
# read data
nba_org <- read.csv("../data/nba-stats_in.csv")
nba_org["sex"] <- NA
nba_org$sex<-"0"
wnba_org <- read.csv("../data/wnba-stats_in.csv")
wnba_org["sex"] <- NA
wnba_org$sex<-"1"
t_nba <- rbind(nba_org, wnba_org)
```

```
## Warning in `[<-.factor`(`*tmp*`, ri, value = c(4565L, 2019L, 1544L,
## 6176L, : invalid factor level, NA generated
```

```
n.var <- names(t_nba)
```

---

1 Descripción del dataset.

Este dataset es interesante tanto para aficionados a la NBA y a la WNBA como para personas que quieran obtener datos acerca de la brecha salarial en los deportes profesionales. Este dataset fue generado en la Práctica 1 de las asignatura "Tipología y ciclo de vida de los datos".

El dataset y el código del mismo esta localizable en la siguiente dirección:

https://github.com/Bengis/nba-gap-cleaning

Si se consultan los datos de origen y se realizan visualizaciones de cuáles son las medias salariales de los jugadores y de las jugadoras,veremos que mientras los chicos tienen una media salarial de unos 10 millones de dolares las chicas tienen una media salarial de 100.000 dolares, lo cual representa un 1% del coste medio de cada jugadora de la NBA.

En esta práctica unificaremos los datos, los limpiaremos y trataremos de estimar un modelo que a partir de los datos nos pueda predecir el salario. Este modelo sera aplicado en el subconjunto de las chicas para ver si existe algun razonmiento productivo para que su salario sea tan bajo.

---

2 Limpieza de datos

El fichero de datos contiene 479 registros y 15 variables.

Contiene 314 jugadores y 165 jugadoras.

Las variables son player, games, minutes, points, rebds., assists, steals, blocks, salary, slry.pts., slry.rbds, slry.asts, slry.stls, slry.blks, sex.

2.1 Selección de variables

De estas variables nos interesa eliminar las siguientes, ya que son campos calculados:

- slry/pts. Es el resultado de salario/puntos.
- slry/rbds. Es el resultado de salario/rebotes.
- slry/asts. Es el resultado de salario/asistencias.
- slry/stls. Es el resultado de salario/robos.
- slry/blks. Es el resultado de salario/tapones.

Estos campos se van a ver modificados en las transformaciones que vamos a ir realizando, por lo que no tiene sentido mantenerlos y realmente no nos interesan en este trabajo donde queremos centrarnos en la creación de un modelo.

```
t_nba<-t_nba[,-10:-14]
```

2.2 Tipos de variables

La lectura del fichero con la función read.csv() ha realizado la siguiente asignación a cada variable, donde tenemos enteros en campos que van a ser transformados, para estos campos seria preferible usar un tipo numérico.

```r
# read data
res <- sapply(t_nba,class)
kable(data.frame(variables=names(res),clase=as.vector(res)))
```

variables	clase
player	factor
games	integer
minutes	integer
points	integer
rebds.	integer
assists	integer
steals	integer
blocks	integer
salary	integer
sex	character

Convertimos los atributos de estadisticas en tipo numérico.

```r
t_nba[2:8] <- lapply(t_nba[2:8], as.numeric)
res <- sapply(t_nba,class)
kable(data.frame(variables=names(res),clase=as.vector(res)))
```

variables	clase
player	factor
games	numeric
minutes	numeric
points	numeric
rebds.	numeric
assists	numeric
steals	numeric
blocks	numeric
salary	integer
sex	character

2.3 Eliminación de valores nulos, outliers y fringeliers

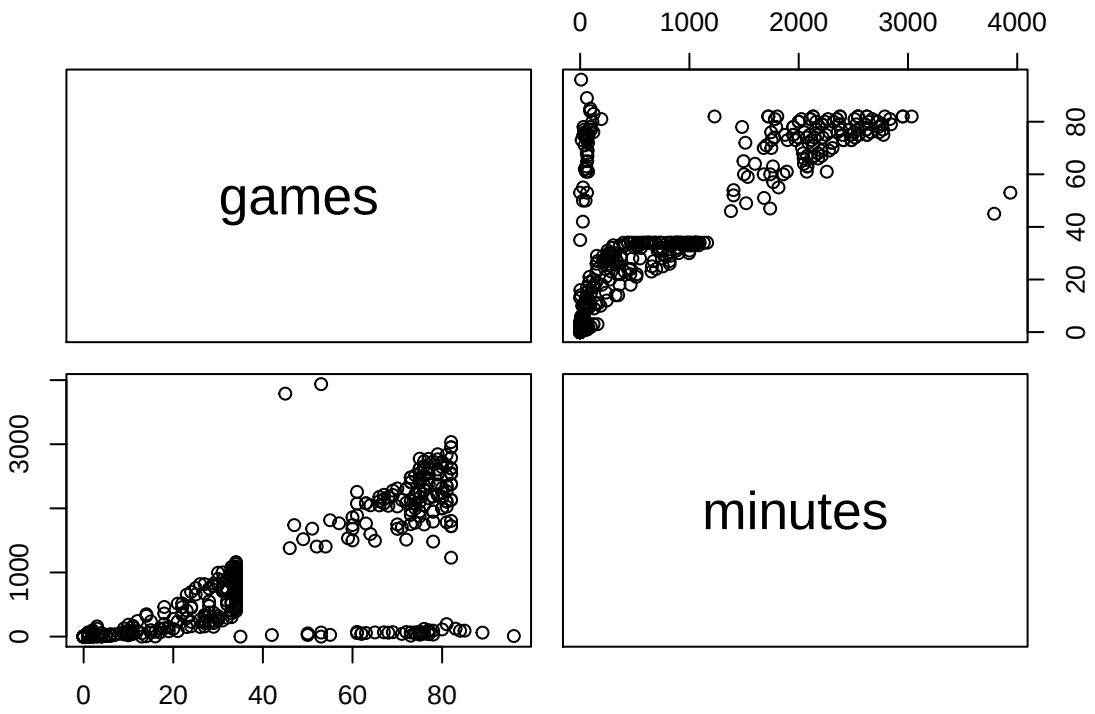
En primer lugar vamos a ver la cantidad de valores nulos que existen por cada atributo.

- Player. Es el nombre del jugador, no es una variables numérica.
- Games. Partidos. Hay 57 jugadores que no han jugado ningun partido.
- Minutes. Minutos. Hay 72 jugadores que no han jugado ni un minuto.
- Points. Puntos. Hay 89 jugadores que no han anotado.
- Rebds. Rebotes.Hay 83 jugadores que no han reboteado.
- Assists. Asistencias.Hay 86 jugadores que no han asistido.
- Steals. Robos. Hay 172 jugadores que no han robado jugadas.
- Blocks. Tapones.Hay 106 jugadores que no han taponado.
- Salary. Salario. Hay 0 jugadores que no tienen salario.
- Sex. Nos permite filtrar el sexo.

Aunque no se han limpiado los datos, ni normalizado, se ven una series de relaciones que son interesantes:

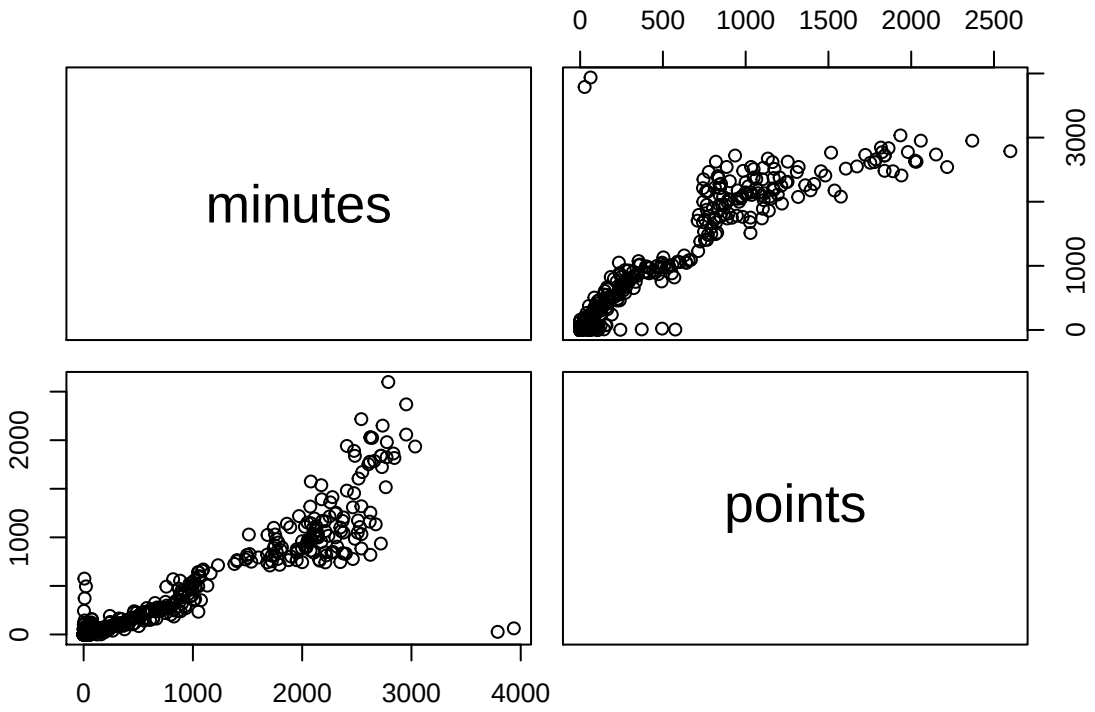
- Entre el número de minutos y partidos jugados la relación es lineal.

```
t_nba_reduced<-t_nba[,-4:-10]
t_nba_reduced<-t_nba_reduced[,-1:-1]
pairs(t_nba_reduced)
```



- Entre el número de minutos jugados y los puntos la relación es lineal.

```
t_nba_reduced<-t_nba[,-5:-10]
t_nba_reduced<-t_nba_reduced[,-1:-2]
pairs(t_nba_reduced)
```



Esto me lleva a pensar que los minutos son un atributo esencial, si un jugador no juega obviamente no va a tener oportunidad de conseguir ninguna estadistica.

Por otro lado, debemos tener en cuenta que tenemos dos origenes de datos distintos.

- Los jugadores de la NBA juegan un máximo de 82 partidos por temporada.
- Las jugadoras de la WNBA juegan un máximo de 34 partidos por temporada.

Por lo que los valores estadisticos totales discriminan a las jugadoras, que no tendran la misma cantidad de puntos, ni de rebotes.

Por lo que realizaremos los siguientes ajustes:

- Se elimina la columna de total de partidos.

```
t_nba<-t_nba[which(t_nba$games!="0"),]
```

- Se dividen todos los estadisticos de productividad por los partidos jugados. Asi tenemos estadisticos por partido, de minutos, puntos, rebotes, asistencias, robos, tapones y salario. De esta forma unificamos los datos de chicos y chicas.

```
games_m=82
games_w=34
for(i in 2:8) {
    t_nba[which(t_nba$sex=="0"),i] <- t_nba[which(t_nba$sex=="0"),i]/games_m
    t_nba[which(t_nba$sex=="1"),i] <- t_nba[which(t_nba$sex=="1"),i]/games_w
}
```

- Se elimina la columna de total de partidos.

```
t_nba<-t_nba[,-2:-2]
```

- Eliminamos los registros de jugadores que no han jugado ningún minuto.

```
t_nba<-t_nba[which(t_nba$minutes!="0"),]
```

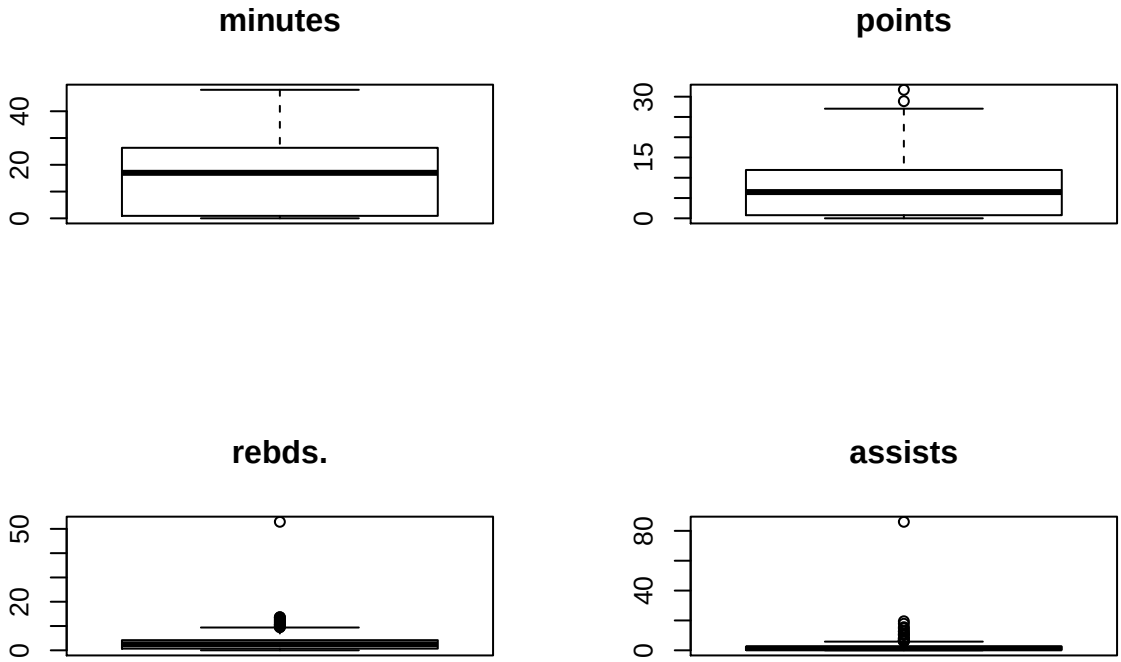
- Eliminamos los registros de jugadores que tengan un número de minutos por partido que podamos catalogar como fringelier, es decir que se alejen 3 veces la desviación estandard de la media.
- Al haber eliminado los fringeliers habremos eliminado tambien los outliers.

```
remove_outliers <- function(x, limit = 3) {
    mn <- mean(x, na.rm = T)
    out <- limit * sd(x, na.rm = T)
    x < (mn - out) | x > (mn + out)
}
t_nba<-t_nba[remove_outliers(t_nba$minutes,3)==FALSE,]
```

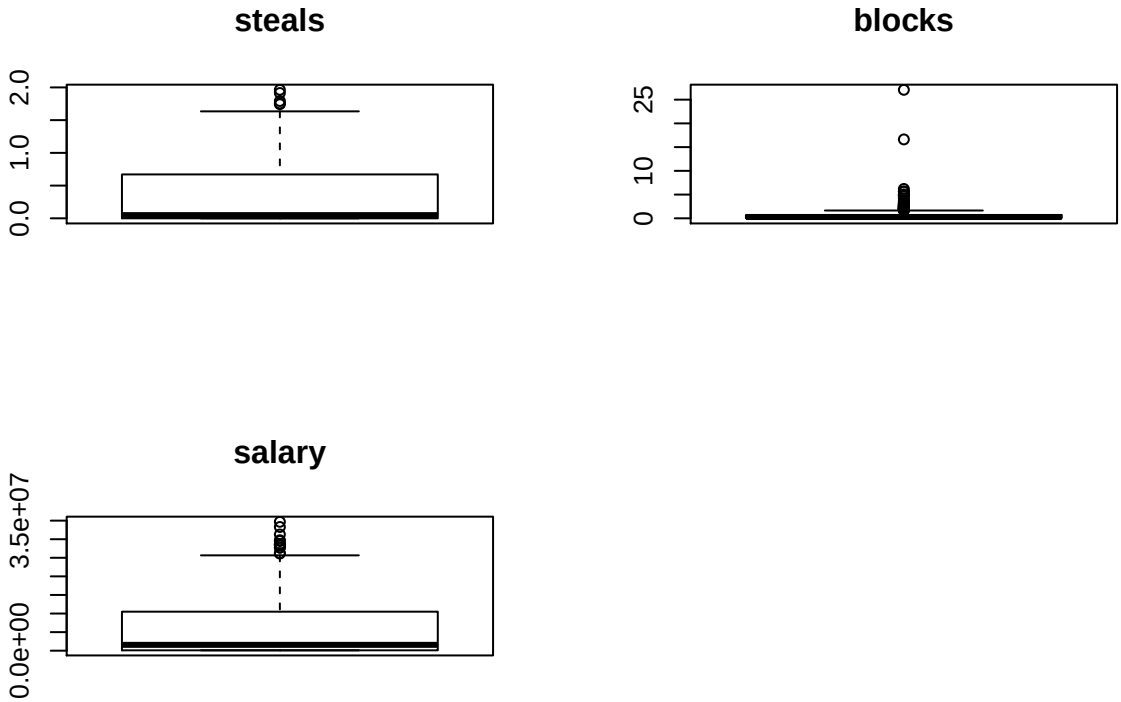
Vemos que con la definicion de fringelier no existen valores asociados a los minutos. Esto se debe a que la desviación estandard es lo bastante grande como para poder abarcar la practica totalidad de valores sin que se consideran fringeliers.

Veamos una representación mediante boxplot de las variables numéricas:

```
par(mfrow=c(2,2))
for(i in 1:ncol(t_nba)) {
  if (is.numeric(t_nba[,i])){
    boxplot(t_nba[,i], main = colnames(t_nba)[i], width = 100)
  }
}
```




```
par(mfrow=c(1,1))
```



Mediante los boxplots vemos la presencia de outliers en puntos, rebotes, asistencias, robos, tapones e incluso en el salario.

Al revisar los datos vemos que han habido errores durante el proceso de scrapping, que han generado estos outliers en las variables rebotes, asistencias y tapones. En el resto no son errores sino valores de productividad por encima de la media en jugadores que son muy productivos los cuales me resisto a eliminar ya que considero que son validos.

El objetivo de la práctica es mostrar la desigualdad salarial entre hombres y mujeres, y para ello es necesario que se mantengan los estadisticos que han producido estos jugadores sin eliminar los que no sean validos.

```
filas_bro<-nrow(t_nba)
t_nba<-t_nba[remove_outliers(t_nba$rebds.,3)==FALSE,]
t_nba<-t_nba[remove_outliers(t_nba$assists,3)==FALSE,]
t_nba<-t_nba[remove_outliers(t_nba$blocks,3)==FALSE,]
filas_aro<-nrow(t_nba)
```

Hemos pasado de 407 filas a 395 filas.

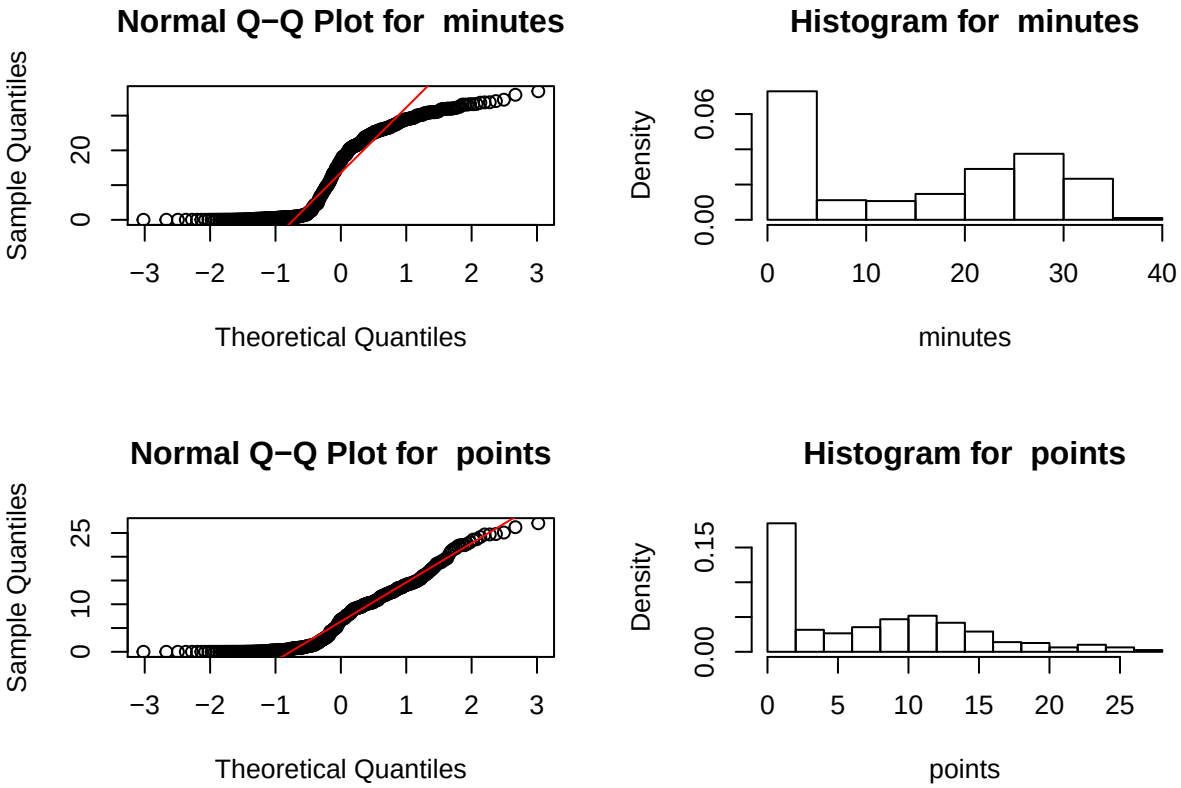
---

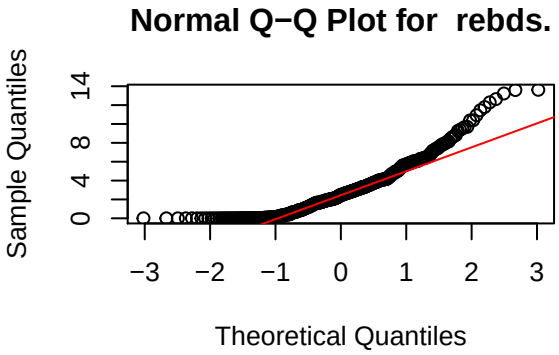
3 Normalización de datos

3.1 Revisión de datos normalizados

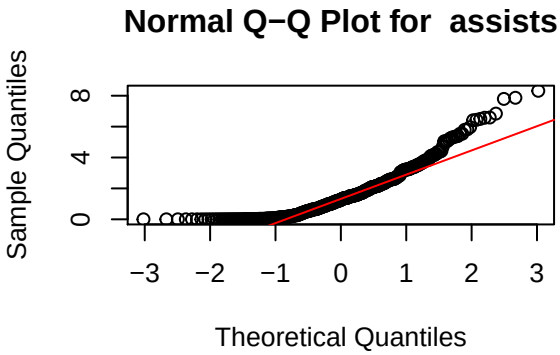
Para revisar si las variables pueden ser candidatas a la normalización miramos las graficas de quantile-quantile plot y el histograma.

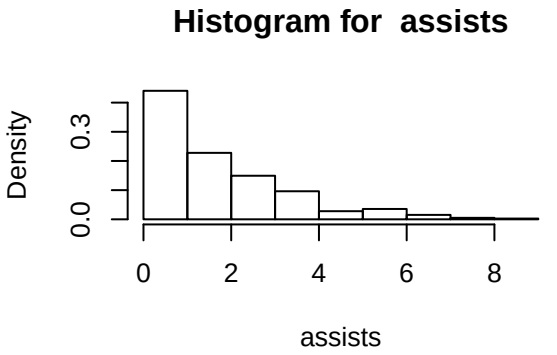
```
par(mfrow=c(2,2))
for(i in 1:ncol(t_nba)) {
  if (is.numeric(t_nba[,i])){
    qqnorm(t_nba[,i],main = paste("Normal Q-Q Plot for ",colnames(t_nba)[i]))
    qqline(t_nba[,i],col="red")
    hist(t_nba[,i],
      main=paste("Histogram for ", colnames(t_nba)[i]),
      xlab=colnames(t_nba)[i], freq = FALSE)
  }
}
```

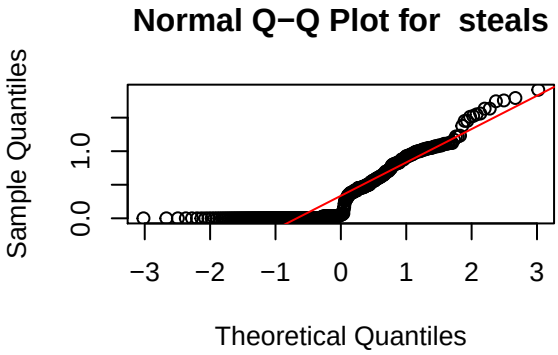


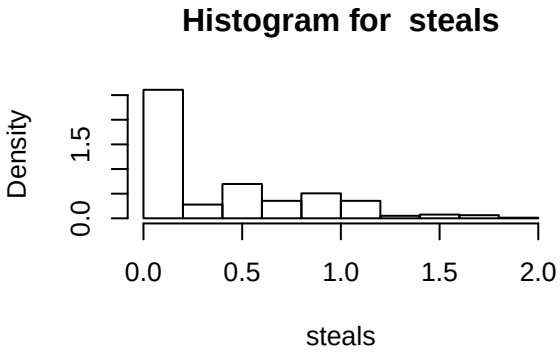

Normal Q−Q Plot for rebds.
Sample Quantiles
0
4
8
14
−3
−2
−1
0
1
2
3
Theoretical Quantiles

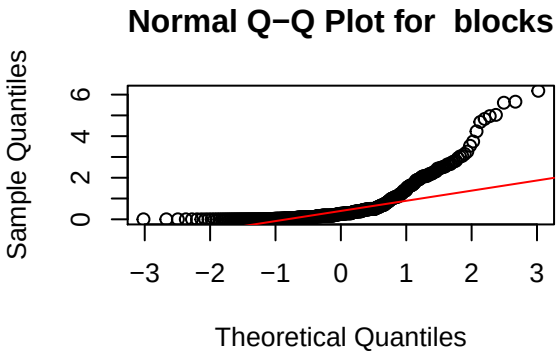

Histogram for rebds.
Density
0.00
0.20
0
2
4
6
8
10
12
14
rebds.

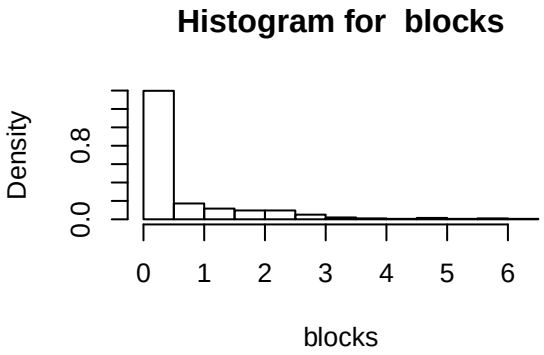

Normal Q−Q Plot for assists
Sample Quantiles
0
4
8
−3
−2
−1
0
1
2
3
Theoretical Quantiles

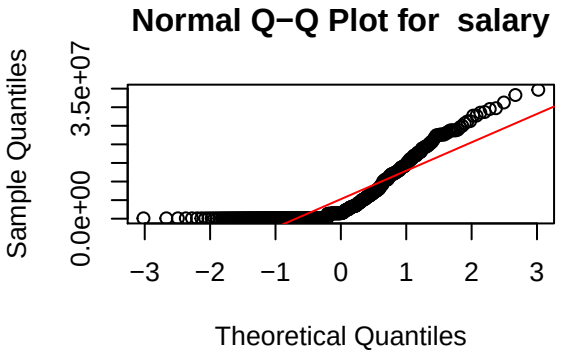

Histogram for assists
Density
0.0
0.3
0
2
4
6
8
assists

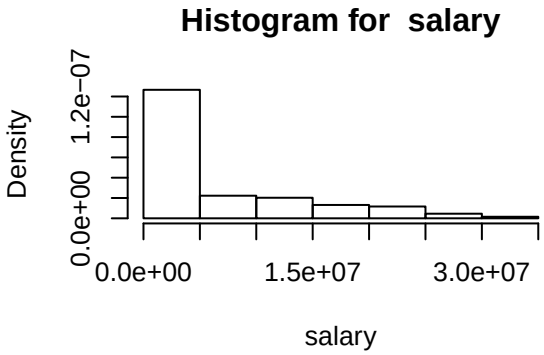

Normal Q−Q Plot for steals
Sample Quantiles
0.0
1.0
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Histogram for steals
Density
0.0
1.5
0.0
0.5
1.0
1.5
2.0
steals


Normal Q−Q Plot for blocks
Sample Quantiles
0
2
4
6
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Histogram for blocks
Density
0.0
0.8
0
1
2
3
4
5
6
blocks





Los resultados del quantile-quantile plot nos indica que las variables pueden ser candidatas a la normalización si es necesario.

Para revisar si las variables estan normalizadas se aplica el test de Shapiro Wilk en cada variables numérica.

```
shapiro.test(t_nba$minutes)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  t_nba$minutes
## W = 0.86357, p-value < 2.2e-16
```

```
shapiro.test(t_nba$points)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  t_nba$points
## W = 0.89537, p-value = 8.279e-16
```

```
shapiro.test(t_nba$rebds.)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  t_nba$rebds.
## W = 0.88758, p-value < 2.2e-16
```

```r
shapiro.test(t_nba$assists)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  t_nba$assists
## W = 0.86715, p-value < 2.2e-16
```

```r
shapiro.test(t_nba$steals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  t_nba$steals
## W = 0.80845, p-value < 2.2e-16
```

```r
shapiro.test(t_nba$blocks)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  t_nba$blocks
## W = 0.66718, p-value < 2.2e-16
```

```r
shapiro.test(t_nba$salary)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  t_nba$salary
## W = 0.76198, p-value < 2.2e-16
```

El test nos indica que ninguna variable esta normalizada, ya que el p-valor es inferior al coeficiente 0.05, por lo que se puede rechazar la hipotesis nula y entender que no es normal.

Que no sea normal no quiere decir que no pueda ser normalizable, ya que segun el teorema del limite central al tener mas de 30 elementos en las observaciones podemos aproximarla como una distribución normal de media 0 y desviación estandard 1.

3.2 Transformación de datos normalizados

Es posible normalizar todos los datos, pero en la representación final de los mismos me gustaria que se pudieran ver los valores reales de los estadisticos. Esto implica que no se realice la normalización, ya que nos permitira ver e identificar de una manera mas realista la brecha salarial, no solo en porcentaje sino con los valores reales.

3.3 Aplicación de pruebas estadisticas

Como hemos indicado anteriormente existe una correlación entre los valores de las estadisticas y de los minutos, un mayor numero de minutos debe representar un valor mayor en las estadisticas.

Asi mismo debe existir una relación entre el salario y las estadisticas, ya que cuanto mejor sea el jugador mas se le va a pagar. Los salarios de los jugadores de la NBA son reales en el dataset. Sin embargo, los salario de las jugadores de la WNBA son estimaciones obtenidas a partir del salario maximo posible.

Para poder determinar entonces el modelo del salario podemos usar un modelo de regresión simple que se base en el estadistico mas importante dentro del juego del baloncesto que son los puntos.

Este modelo se implementa en un set dividido en dos subconjuntos, uno para entrenarlo llamado train y otro para evaluarlo llamado test. La caracteristica de este dataset es que solo hay jugadores de la NBA, de esta forma podemos valorar los salarios que se predicen dentro del entorno de estos jugadores.

Luego usaremos ese modelo para predecir cual deberia ser el salario en el dataset de las chicas y compararlo con el real.

```
t_nba_glm<-t_nba[which(t_nba$sex=="0"),]
ntrain <- nrow(t_nba_glm)*0.8
ntest <- nrow(t_nba_glm)*0.2
set.seed(1)
index_train<-sample(1:nrow(t_nba_glm),size = ntrain)
train<-t_nba_glm[index_train,]
test<-t_nba_glm[-index_train,]
modelo<-lm(formula = salary ~ points, data=train)
summary(modelo)
```

```
##
## Call:
## lm(formula = salary ~ points, data = train)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -16701654  -3653767   -592110   4995122  16119777
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4828039     711031    6.79 1.53e-10 ***
## points        675464      63982   10.56  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6511000 on 182 degrees of freedom
## Multiple R-squared:  0.3798, Adjusted R-squared:  0.3764
## F-statistic: 111.5 on 1 and 182 DF,  p-value: < 2.2e-16
```

El modelo no es muy bueno, el coeficiente R^2 ajustado es 0.3764, bastante alejado del optimo. Sin embargo el uso de la variables de los puntos parece adecuada de acuerdo al p-valor de la misma.

Podemos comprobar la validez del modelo realizando la predicción y comparando los valores predecidos con los reales.

```
prob_sl<-predict(modelo, test, type="response")
mc_sl<-data.frame(
  real=test$salary,
  predicted= prob_sl,
  dif=ifelse(test$salary>prob_sl, -prob_sl*100/test$salary,prob_sl*100/test$salary)
  )
colnames(mc_sl)<-c("Real","Predecido","Dif%")
kable(mc_sl)
```

	Real	Predecido	Dif%
3	6261395	23090272	368.77200
9	33285709	21138017	-63.50478

	Real	Predecido	Dif%
13	22471910	20174245	-89.77539
28	17000450	17027573	100.15954
31	14500000	16294447	112.37550
32	28703704	16030852	-55.84942
36	19000000	15157691	-79.77732
39	1471382	14869384	1010.57262
44	20061728	14457515	-72.06516
45	23000000	14465753	-62.89458
55	5000000	13938562	278.77123
67	25656667	13320759	-51.91929
68	12307692	13304285	108.09732
70	5504420	13197199	239.75640
71	4328000	13131300	303.40342
79	4149242	12867705	310.12181
96	10500000	11895696	113.29234
98	13168750	11838034	-89.89490
104	11747890	11673287	-99.36497
108	2334528	11574439	495.79352
118	10162921	11170808	109.91729
119	1577230	11187283	709.29937
120	5000000	11137858	222.75717
125	16000000	11039010	-68.99381
127	12500000	10940162	-87.52129
132	10505000	10734228	102.18208
144	6270000	5404654	-86.19863
145	1471382	9564522	650.03664
152	16400000	4836276	-29.48949
157	3704160	4828039	130.34098
188	1471382	5009260	340.44595
190	9259260	5239907	-56.59099
192	2500000	5091634	203.66537
193	9003125	5569401	-61.86076
194	4956480	5157533	104.05637
199	1312611	5437603	414.25855
208	8393000	5676487	-67.63359
210	2000000	5141058	257.05292
232	3477600	4852751	139.54309
245	11422536	4828039	-42.26766
247	6000000	4828039	-80.46731
249	1524305	4836276	317.27744
255	1312611	5305806	404.21767
258	6655325	5058685	-76.00958
283	2301360	5832997	253.45868
298	1700640	4828039	283.89539

Sin embargo esta práctica no se basa en tratar de establecer el salario a partir de las estadisticas, sino en ver la brecha salarial. Para ello vamos a predecir el salario de las chicas con este mismo modelo y compararlo con los datos reales.

3.4 Comparación de datos con el dataset de las jugadoras de la WNBA

¿Que ocurriria si aplicasemos el mismo modelo con las chicas?

```
test<-t_nba[which(t_nba$sex=="1"),]
prob_sl<-predict(modelo, test, type="response")
mc_sl<-data.frame(
  real=test$salary,
  predicted= prob_sl,
  dif=ifelse(test$salary>prob_sl, -prob_sl*100/test$salary,prob_sl*100/test$salary)
  )
colnames(mc_sl)<-c("Real","Predecido","Dif%")
kable(mc_sl)
```

	Real	Predecido	Dif%
315	105000	18118781	17255.982
316	105000	17860515	17010.015
317	105000	17582383	16745.127
318	105000	17522784	16688.365
319	105000	17324118	16499.160
320	105000	16767853	15969.384
321	105000	16509588	15723.417
322	105000	16132123	15363.927
323	105000	15893724	15136.880
324	105000	15834124	15080.118
325	105000	15496392	14758.469
326	105000	15218260	14493.581
327	105000	14920261	14209.773
328	105000	14820929	14115.170
329	105000	14661996	13963.806
330	105000	14582530	13888.123
331	105000	14602396	13907.044
332	105000	14522930	13831.362
333	105000	14324264	13642.156
334	105000	14205065	13528.633
335	105000	14006399	13339.427
336	105000	13847466	13188.063
337	105000	13609067	12961.016
338	105000	13191869	12563.685
339	105000	13132269	12506.923
340	105000	12854137	12242.035
341	105000	12794537	12185.274
342	105000	12715071	12109.591
343	105000	12516405	11920.386
344	105000	12039607	11466.293
345	105000	11840941	11277.087
346	105000	11840941	11277.087
347	105000	11741609	11182.484
348	105000	11701875	11144.643
349	105000	11503210	10955.438
350	105000	11264811	10728.391
351	105000	11244944	10709.471
352	105000	11145611	10614.868
353	105000	10986679	10463.503

	Real	Predecido	Dif%
354	105000	10907212	10387.821
355	105000	10728413	10217.536
356	105000	10629080	10122.933
357	105000	10350948	9858.046
358	105000	10350948	9858.046
359	105000	10231748	9744.522
360	105000	10231748	9744.522
361	105000	10192015	9706.681
362	105000	10231748	9744.522
363	105000	10092682	9612.078
364	105000	10092682	9612.078
365	105000	10052949	9574.237
366	105000	10013216	9536.396
367	105000	10013216	9536.396
368	105000	9993349	9517.476
369	105000	9774817	9309.350
370	105000	9754951	9290.429
371	105000	9675484	9214.747
372	105000	9615884	9157.985
373	105000	9615884	9157.985
374	105000	9576151	9120.144
375	105000	9536418	9082.303
376	105000	9556285	9101.224
377	105000	9456952	9006.621
378	105000	9476818	9025.541
379	105000	9417219	8968.780
380	105000	9337752	8893.097
381	105000	9337752	8893.097
382	105000	9278153	8836.336
383	105000	9198686	8760.654
384	105000	9000020	8571.448
385	105000	8880821	8457.925
386	105000	8821221	8401.163
387	105000	8622555	8211.958
388	105000	8503356	8098.434
389	105000	8324557	7928.149
390	105000	8324557	7928.149
391	105000	8264957	7871.388
392	105000	8264957	7871.388
393	105000	8125891	7738.944
394	105000	8125891	7738.944
395	105000	8106024	7720.023
396	105000	8066291	7682.182
397	105000	8046425	7663.262
398	105000	7966958	7587.579
399	105000	7788159	7417.294
400	105000	7748426	7379.453
401	105000	7788159	7417.294
402	105000	7768292	7398.374
403	105000	7728559	7360.533
404	105000	7649093	7284.850
405	105000	7390827	7038.883

	Real	Predecido	Dif%
406	105000	7192162	6849.678
407	105000	7172295	6830.757
408	105000	7172295	6830.757
409	105000	7172295	6830.757
410	105000	7112695	6773.996
411	105000	6914029	6584.790
412	105000	6914029	6584.790
413	105000	6834563	6509.108
414	105000	6794830	6471.267
415	105000	6774963	6452.346
416	105000	6794830	6471.267
417	105000	6755097	6433.426
418	105000	6655764	6338.823
419	105000	6596164	6282.061
420	105000	6596164	6282.061
421	105000	6556431	6244.220
422	105000	6536564	6225.299
423	105000	6496831	6187.458
424	105000	6457098	6149.617
425	105000	6437232	6130.697
426	105000	6377632	6073.935
427	105000	6298165	5998.253
428	105000	6318032	6017.173
429	105000	6178966	5884.730
430	105000	6139233	5846.888
431	105000	6099500	5809.047
432	105000	6079633	5790.127
433	105000	6020033	5733.365
434	105000	5980300	5695.524
435	105000	5880967	5600.921
436	105000	5841234	5563.080
437	105000	5801501	5525.239
438	105000	5781634	5506.318
439	105000	5741901	5468.477
440	105000	5761768	5487.398
441	105000	5722035	5449.557
442	105000	5682301	5411.716
443	105000	5602835	5336.033
444	105000	5563102	5298.192
445	105000	5503502	5241.431
446	105000	5523369	5260.351
447	105000	5404169	5146.828
448	105000	5384303	5127.907
449	105000	5384303	5127.907
450	105000	5344570	5090.066
451	105000	5344570	5090.066
452	105000	5344570	5090.066
453	105000	5324703	5071.146
454	105000	5185637	4938.702
455	105000	5185637	4938.702
456	105000	5106171	4863.020
457	105000	5086304	4844.099

	Real	Predecido	Dif%
458	105000	5086304	4844.099
459	105000	5066437	4825.179
460	105000	5066437	4825.179
461	105000	5046571	4806.258
462	105000	5026704	4787.337
463	105000	5006838	4768.417
464	105000	5006838	4768.417
465	105000	4967105	4730.576
466	105000	4947238	4711.655
467	105000	4947238	4711.655
468	105000	4947238	4711.655
469	105000	4927371	4692.735
470	105000	4887638	4654.894
471	105000	4867772	4635.973
472	105000	4867772	4635.973
473	105000	4828039	4598.132
474	105000	4828039	4598.132
475	105000	4828039	4598.132
476	105000	4828039	4598.132
477	105000	4828039	4598.132
478	105000	4828039	4598.132
479	105000	4828039	4598.132

Pasariamos de diferencias medias del 100% a diferencias del 10000%. Las chicas estarian encantadas de que se valorase su productividad incluso con un modelo que se ajuste tan mal ya que supondria un incremento superior al 1000% de su sueldo.

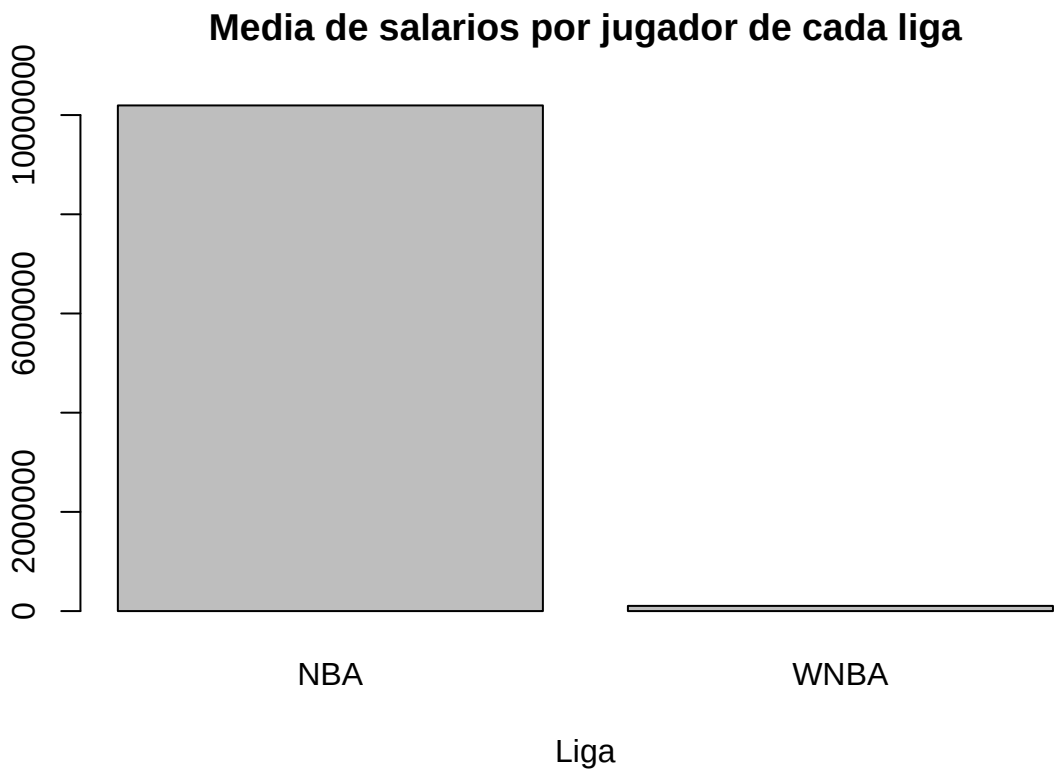
Lo que podemos ver es que el modelo de predicción de salario de los chicos aunque no es optimo es mucho mejor que el modelo real que se esta utilizando en la WNBA.

---

4 Representación de los resultados a partir de tablas y gráficas.

Hay muchas representaciones que nos dan una idea de lo enorme que es la brecha salarial, pero creo que la siguiente, que muestra la cantidad media de dinero que ganan por jugador y jugadora es la mas significativa.

```
options(scipen=5)
nba_salary=sum(as.numeric(t_nba$salary[which(t_nba$sex==0)])/length(which(t_nba$sex==0)))
wnba_salary=sum(as.numeric(t_nba$salary[which(t_nba$sex==1)])/length(which(t_nba$sex==1)))
counts <- c(nba_salary, wnba_salary)
barplot(counts, names=c("NBA", "WNBA"), main="Media de salarios por jugador de cada liga",
    xlab="Liga")
```



Una diferencia tremenda, posiblemente una de las mayores brechas salariales que existe en el deporte profesional, e incluso me atrevo a decir que en gran parte de las profesiones.

---

5 Resolución del problema y conclusiones.

Nos queda que:

- El salario medio para un jugador de la NBA es de 10194261 dolares.
- El salario medio para una jugadora de la WNBA es de 105000 dolares.

Estamos hablando de que cada jugadora gana de salario medio un 1% del salario medio de cada jugador. De hecho con solo dos jugadores de la NBA basta para que sus salarios sean superiores al de total de jugadoras de la WNBA.

Por úiltimo procedemos a la exportación de datos en el dataset de salida.

---

6 Exportación del código en R y de los datos producidos.

El código en R esta incluido en este fichero con extensión rmd y tambien se puede descargar en GitHub desde la siguiente dirección:

https://github.com/Bengis/nba-gap-cleaning/blob/master/code/nba-gap-cleaning.r

Los datos de salida se exportan mediante el siguiente comando y pueden ser descargados desde en GitHub desde la siguiente dirección:

https://github.com/Bengis/nba-gap-cleaning/blob/master/data/nba_out.csv

```
write.csv(t_nba, file = "../data/nba_out.csv")
```