Pràctica 2: Com realitzar la neteja i l'anàlisi de dades?

Carlos Martínez Torró (cmtorro@uoc.edu) i Xavier Roca Canals (xrocaca@uoc.edu)

13/01/2023

Índex de la pràctica

Exercici 1: Descripció del dataset

El nostre dataset recopila desenes de mètriques estadístiques de més de 70 temporades de la NBA i de 25 anys de la WNBA, és a dir, des dels inicis respectius de cada lliga. Les dades es van recollir del web Basketball Reference (propietat del grup Sports Reference) i estan classificades per equip i per temporada, amb un total de 1900 observacions i 53 variables. És a dir, a cada fila trobarem com li ha anat a un equip, des del punt de vista mètric, en una temporada en concret. L'ampla disponibilitat de variables ens permet anar des dels anàlisis més superficials (victòries, derrotes, punts a favor...) a altres més complexos (ritme de joc, eficiència dels llançaments...).

En una era com la nostra en la que les dades s'han infiltrat arreu, ens preguntem si també han arribat a la lliga de bàsquet per excel · lència. L'objectiu que perseguim amb la creació i l'actual anàlisi d'aquest dataset és ambiciós: es pot explicar l'evolució de la NBA a través de les estadístiques? Hi ha algun patró que segueixin aquells equips més exitosos? I estan dirigint aquests equips l'evolució de l'esport?

Exercici 2: Integració i selecció de les dades

A la Pràctica 1 ja vam realitzar una integració de diferents datasets creats a partir del *web scraping*, ja que estàvem interessats en diferents taules i vam optar per fer un dataframe per cadascuna d'elles i després unir els dataframes resultants per les columnes comunes (amb la funció `merge` de la llibreria `pandas`). Per tant, el dataset lliurat a la PRA1 és el que carregarem.

Com posar la sortida de les funcions `head`, `summary` o `str` pot allargar molt el document i resultar improductiu degut a l'elevat nombre de variables que tenim, podem crear una taula per veure la informació bàsica del dataset: quantes observacions tenim, quantes variables, quantes són numèriques, quantes categòriques, etc.

Table 1: Mètriques bàsiques del dataset

Mètrica	Valor
Nombre d'observacions	1900
Nombre de variables	53
Variables numèriques	48
Variables categòriques	5
Casos complets (%) (observacions sense NAs)	1522 / 1900 = 80.11 %
Variables completes (%) (variables sense NAs)	31 / 53 = 58.49 %

Com podem veure a la taula, hi ha un gran percentatge de casos complets (és a dir, equips amb tota la informació de l temporada completa), mentre que aproximadament el 60% de les variables no tenen cap valor NA. Podem fer una ullada a quines són les variables amb més valors perduts del nostre dataset.

Veiem que algunes d'aquestes variables estan relacionades amb el llançament de tres punts. Això és degut a que fins 1979 no existia aquest llançament, així que hi ha gairebé trenta anys de registres on aquestes variables tenen valor nul per definició. Per altra banda, algunes mètriques com els rebots ofensius (variable *orb*) o defensius (variable *drb*) i les pèrdues (com el percentatge del propi equip amb la variable *tov_pct* o del contrari amb *opp_tov_pct*) també van ser estadístiques que no es recollien originalment, així que és normal que hi hagi valors perduts en aquestes. De fet, si filtrem i només agafem les observacions posteriors a la temporada 1978-79 (el llançament de tres va començar a la 1979-80) veurem que el nombre de valors perduts és molt diferent:

```
## Valors perduts després de la temporada 1978-79: 0
```

Table 2: Variables amb més valors perduts

Variable	Valors NA
fg3a_per_fga_pct	378
fg3	378
fg3a	378
fg3_pct	378
tov_pct	259
orb_pct	259
opp_tov_pct	259
drb_pct	259
orb	259
drb	259

Exercici 3: Neteja de les dades

Com bé hem comentat en l'anterior apartat, trobem temporades en les quals ens falta informació sobre estadístiques bàsiques sobre el joc. D'aquesta forma, ens dificulta l'anàlisi sobre aquestes temporades, ja que no disposem dels elements bàsics per entendre com era l'estil o funcionament de joc de cada equip. En conseqüència, s'ha decidit descartar aquestes temporades.

3.1: Les dades contenen zeros o elements buits?

Així i tot, trobem temporades més antigues a l'aparició del tir de tres punts que si disposen d'aquestes estadístiques. Observant el conjunt de dades, trobem que a partir de la temporada 1973-74 disposem de tota la informació necessària. Tot i això, analitzarem quins valors buits disposem en el nostre conjunt de dades a partir d'aquesta temporada.

Table 3: Columnes amb valors NAs filtrant a partir de la temporada 1973-74

fg3a_per_fga_pct	fg3	fg3a	fg3_pct
102	102	102	102

Com era d'esperar, les úniques variables que observem amb valors buits són les que fan referència als tirs de tres punts. El que farem, és omplir aquests valors buits amb el valor 0, ja que s'ha decidit que és el valor que realment representa aquests camps. Com no existia el tir de tres no es van realitzar cap tir.

Un altre fet a tenir en compte, és el camp `gb`. Aquesta variable ens mostra quants partits té cada equip per darrere del rival que ocupa el primer lloc. En alguns casos, aquesta informació ve representada amb el valor `-`. Entenem que aquest valor, fa referència al fet que no té cap partit per darrere dels rivals (és a dir, és l'equip que va primer de la seva divisió) i el seu valor real és 0.

3.2: Identifica i gestiona els valors extrems.

En aquest cas, a partir de la funció 'summary' podem observar els mínims i màxims valors de cada variable. No s'exemplificarà en el document, ja que el gran volum de variables dificultaria la lectura del document de la pràctica. Així i tot, un cop comprovat aquest, sí que podem trobar valors molt petits en comparació a la mitjana, com pot ser el cas de victòries que ha assolit un equip en una temporada.

Ho podem veure en aquest petit exemple:

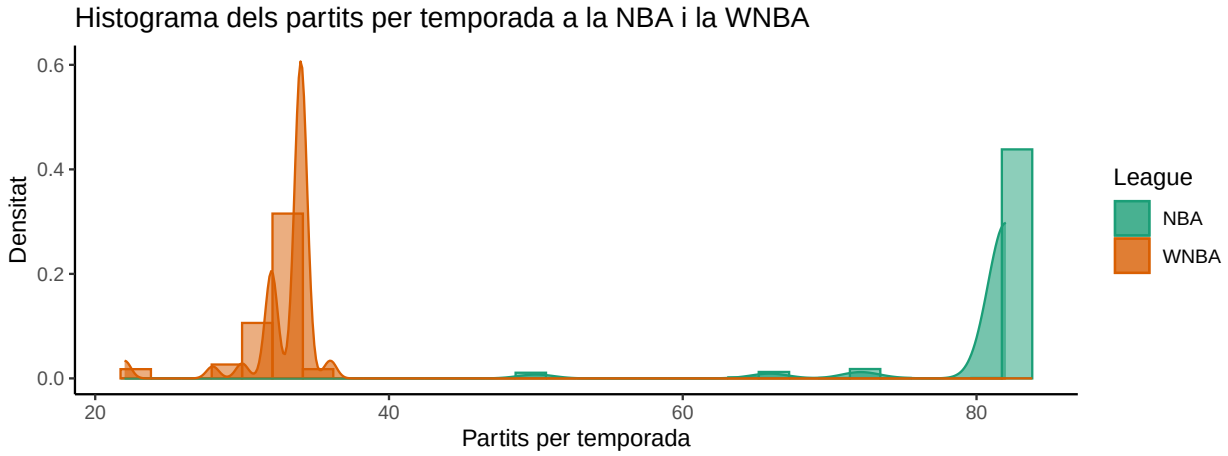
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   22.00   36.00   35.41   47.00   73.00
```

És normal, que ens podem trobar aquests casos en múltiples variables, ja que en una mateixa temporada, els equips guanyadors presentaran valors màxims en victòries i els perdedors mínims en aquestes. En conseqüència, també es veurà reflectit en les estadístiques d'aquests equips.

El que s'ha decidit és no realitzar cap modificació, ja que aquests valors són correctes perquè es basen en dades reals sobre les temporades i les estadístiques de joc de cada un dels equips. Per tant, hem d'assumir que serà possible trobar valors que siguin *outliers*, però que haurem de tractar com un valor més.

Exercici 4: Anàlisi de les dades

Abans de començar amb l'anàlisi, haurem de tenir en compte que els valors absoluts ens poden portar a error i que haurem d'optar, en general, pels relatius (percentatges o mètriques ajustades per partit o possessions). Això és degut a que la NBA i la WNBA tenen un nombre de partits totals diferents, però també perquè la pròpia NBA ha evolucionat en aquest aspecte: en el seu inici es jugaven molts menys partits. Veiem-ho amb un gràfic, on utilitzarem les dades de partits des de 1973:



Com es pot veure, en general les temporades a la NBA han tingut més de 80 partits, però no sempre ha sigut així. Pel que fa a la WNBA, hi ha més variació en aquest número de partits, però es troba al voltant dels 30. És per això que, en cas d'utilitzar valors absoluts, haurem de comprovar prèviament que estiguem comparant entre temporades amb el mateix número de partits.

Pel que fa als subapartats de l'exercici en sí, hem decidit que, per facilitar la lectura, els farem consecutius per a cadascuna de les tres preguntes que ens hem plantejat. És a dir, farem la selecció, comprovació de normalitat i les aplicacions dels estadístics de forma contínua per a cada anàlisi, amb la idea de mantenir el fil i els raonaments particulars fins acabar l'anàlisi.

Anàlisi 1: ha canviat la preferència dels llançaments?

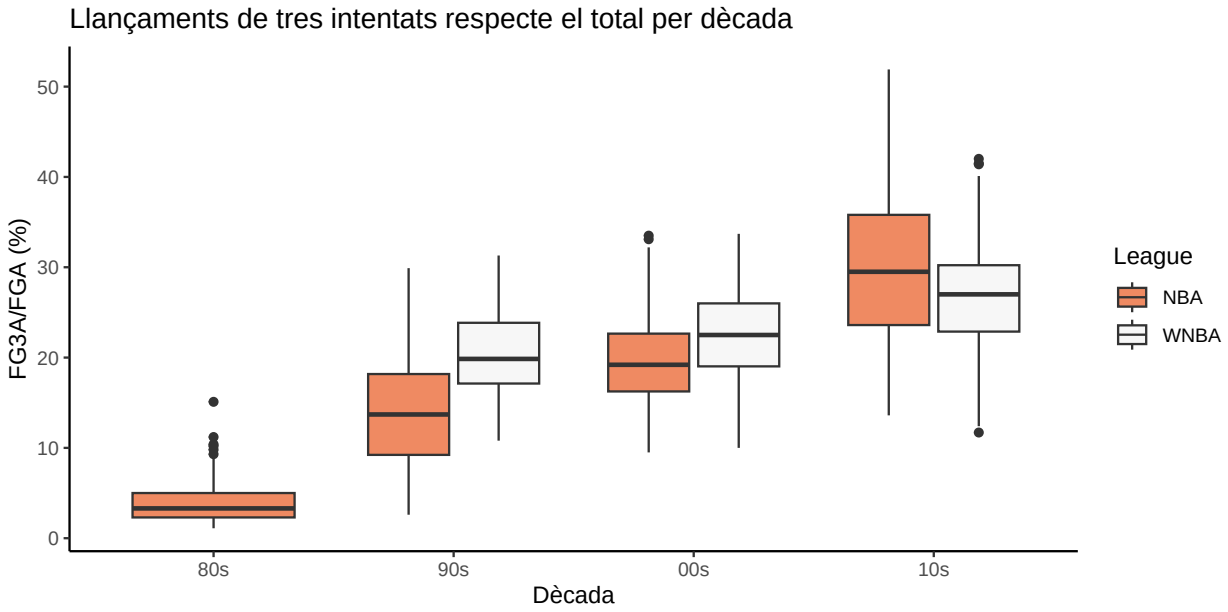
La introducció dels llançaments de tres a la lliga l'any 1979 va suposar un abans i un després a l'hora de jugar l'esport. No obstant, un espectador que no hagi vist res de bàsquet en els darrers 20 anys es podria sorprendre amb la quantitat aparent de llançaments de tres que es practiquen avui en dia. Amb l'aparició de fenòmens com Stephen Curry, el joc sembla haver canviat en els darrers anys.

Anàlisi 1.1: Selecció de les dades que es volen analitzar/comparar

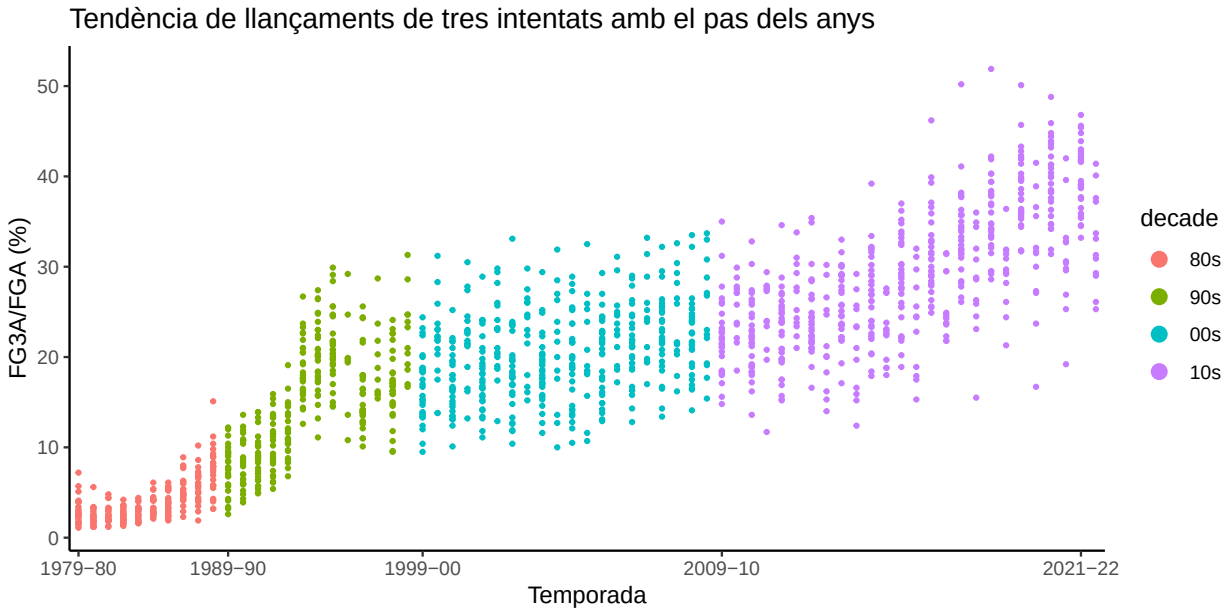
Per analitzar-ho millor, farem un anàlisi dècada per dècada dels llançaments de tres. Ens fixarem en una variable en concret: `fg3a_per_fga_pct`. Aquesta variable indica (en percentatge) quants llançaments de tres ha realitzat un equip de tots els llançaments intentats. És a dir, si de 10 tirs de camp, 4 són triples, parlarem d'un 40% (en la variable estaria codificat com a 0.4). Triem aquesta variable en comptes del nombre de triples perquè ens pot donar una idea millor de si ha variat la selecció de llançaments.

Així doncs, farem quatre grups diferents: del 79 fins el 90, del 90 fins el 2000, del 2000 fins el 2010 i del 2010 fins l'actualitat. Ho codificarem tot en una nova variable, que anomenarem `decade`:

Amb un boxplot podem comparar les èpoques a primera vista:



Veiem que, efectivament, hi ha una tendència a llançar més de tres amb els anys. Utilitzem ara un gràfic de dispersió per a observar aquesta tendència amb el pas de les temporades:



Notablement, sembla que aquest percentatge es manté en un rang similar des de mitjans dels 90 fins mitjans de la dècada dels 2010s, on comença a pujar. Hi ha un equip, fins i tot, que va sobrepassar el 50% de llançaments de tres del total de llançaments.

Table 4: Equip amb el valor màxim de llançaments de tres intentats respecte el total de llançaments

Season	Team	wins	losses	Playoffs	fg3a_per_fga_pct	fg3_pct
2018-19	Houston Rockets	53	29	Yes	0.519	0.356

Anàlisi 1.2: Comprovació de la normalitat i homogeneïtat de la variància

Passarem ara a fer l'anàlisi estadístic d'aquestes dades. Abans, però, haurem de comprovar la normalitat per saber si haurem d'aplicar un test paramètric o un no paramètric.

Table 5: Test de Shapiro-Wilk per avaluar normalitat de la mostra

decade	statistic	p.value
80s	0.8818854	0.0000000
90s	0.9804219	0.0003230
00s	0.9900004	0.0045421
10s	0.9875335	0.0001305

Com es pot veure a la taula, el p-valor és en tots els casos molt inferior a 0.05, el que permet refutar la hipòtesi nul · la que assumeix una distribució normal. Per tant, podem dir que cap de les quatre mostres avaluades té una distribució normal per la variable `fg3a_per_fga_pct`.

Pel que fa a la variància, podem comparar les variàncies amb el test de Bartlett o el test de Levene.

Table 6: Avaluació de l'homoscedasticitat amb els tests de Bartlett i Levene

Test	p.valor
Test de Bartlett	0
Test de Levene	0

Podem comprovar mirant els p-valors d'ambdós tests que les mostres no tenen la mateixa variància, ja que en ambdós casos refutem la hipòtesi nul · la de igualtat de variàncies.

Anàlisi 1.3: Aplicació de proves estadístiques per comparar els grups de dades.

Per tant, tenim mostres on no hi ha una distribució normal dels valors i tampoc tenim una situació de homoscedasticitat. No obstant, com tenim una mostra força gran, podem aplicar el **teorema del límit central**, ja que, degut a la mida de la mostra, podem assumir que si fem les mitjanes aritmètiques de diferents mostrejos aleatoris, la distribució d'aquestes mitjanes aritmètiques serà gaussiana.

Pel que fa a la variància, podem aplicar el test de Welch, que és una alternativa a l'ANOVA clàssic quan no hi ha homoscedasticitat. Així doncs, mirarem si hi ha diferències entre les diferents dècades pel que fa al percentatge de llançaments de tres respecte el total de llançaments fets.

El p-valor de 0 ens indica que hi ha diferències. Ara bé, entre quins grups? Per saber-ho, necessitem fer un test *post hoc*. Podem utilitzar el test de Games-Howell, similar al test de Tukey (un dels més comuns), però aquest no assumeix igualtat de variàncies.

Table 7: Nombre d'observacions que tenim per dècada

Dècada	Observacions
Pre-three era	378
80s	231
90s	308
00s	437
10s	546

Table 8: Test de Welch per avaluar diferència entre els llançaments de tres respecte el total per cada dècada

Test	p.valor
ANOVA de Welch	0

Table 9: Diferències entre dècades en el percentatge de llançament de tres respecte als llançaments total

Grup 1	Grup 2	p-Valor ajustat
80s	90s	0
80s	00s	0
80s	10s	0
90s	00s	4.9e-10
90s	10s	0
00s	10s	1.98e-13

Així, veiem que hi ha diferències significatives entre tots els grups. No obstant, veiem que les diferències entre els 90s i els 2000s i entre els 2000s i els 2010s són menors que entre les altres dècades, un fet que podíem intuir amb la representació gràfica d'aquesta variable.

Anàlisi 2: ha canviat l'estil de joc durant els anys?

Dins de l'NBA, cadascun dels equips que han competit o encara competeixen han desenvolupat un estil de joc característic a partir d'aprofundir i millorar en certes capacitats ofensives/defensives que els ha permès aconseguir victòries i obtenir el rendiment desitjat. A més, cada època en concret té les seves peculiaritats pròpies, ja sigui per les regles del moment, pels equips dominadors o per grans estrelles que obliguen als equips a evolucionar per adaptar-se.

Anàlisi 2.1: Selecció de les dades que es volen analitzar/comparar

Com disposem d'un gran nombre de variables, ens hem preguntat si podríem veure, amb l'ajuda d'aquestes, si ha canviat la forma de jugar en les últimes dècades de la lliga. Amb aquesta idea al cap, volem saber si la correlació entre algunes de les variables més importants ha anat canviant amb els anys o si, pel contrari, s'ha mantingut constant. A més, també ens preguntem quina és la correlació entre aquestes variables i les victòries aconseguides durant la temporada en funció de l'època, amb l'objectiu final de veure si quines variables es correlacionen més fortament amb les victòries.

Per tant, les variables a utilitzar es tracten de les estadístiques més bàsiques que s'han aconseguit sobre el joc.

Table 10: Variables seleccionades per la correlació

Variable	Descripció de la variable
win_loss_pct	percentatge de victòries
age	mitjana d'edat
off_rtg	classificació ofensiva
def_rtg	classificació defensiva
fg	mitjana de tirs de camp
fg2	mitjana de tirs de dos punts
fg3	mitjana de tirs de tres punts
ft	mitjana de tirs lliures
orb	mitjana de rebots ofensius
drb	mitjana de rebots defensius
ast	mitjana d'assistències
stl	mitjana de pilotes robades
blk	mitjana de bloquejos
tov	mitjana de pilotes perdudes
pf	mitjana de faltes personals

Anàlisi 2.2: Correlació entre les estadístiques més bàsiques

Seguidament, comprovarem quina correlació poden tenir totes aquestes variables entre si, i entendre quin estil de joc predomina durant cada època dins de l'NBA i WNBA. No obstant, abans de començar amb les correlacions haurem de veure la distribució d'aquestes variables, ja que depenent de si hi ha normalitat o no haurem d'aplicar un mètode o un altre (per exemple, la correlació de Pearson és paramètrica, mentre que la τ de Kendall i la ρ de Spearman són no paramètriques). Com la nostra intenció és veure l'evolució en diferents eres, mirarem la distribució d'aqustes mostres en funció de l'era i la lliga:

Table 11: Normalitat de les variables en funció de la dècada i la lliga (primeres files)

Variable	Descripció de la variable	f	f	p-valor
orb	90s	WNBA	0.9843325	0.9252866
def_rtg	90s	WNBA	0.9833122	0.9050902
ast	80s	NBA	0.9966187	0.9021855
pf	90s	NBA	0.9968436	0.8611868
off_rtg	90s	WNBA	0.9813831	0.8610215
pf	80s	NBA	0.9962430	0.8531362

Note:
Ordenats per p-valor descendent

Table 12: Resum de les distribucions de les variables en funció del seu p-valor

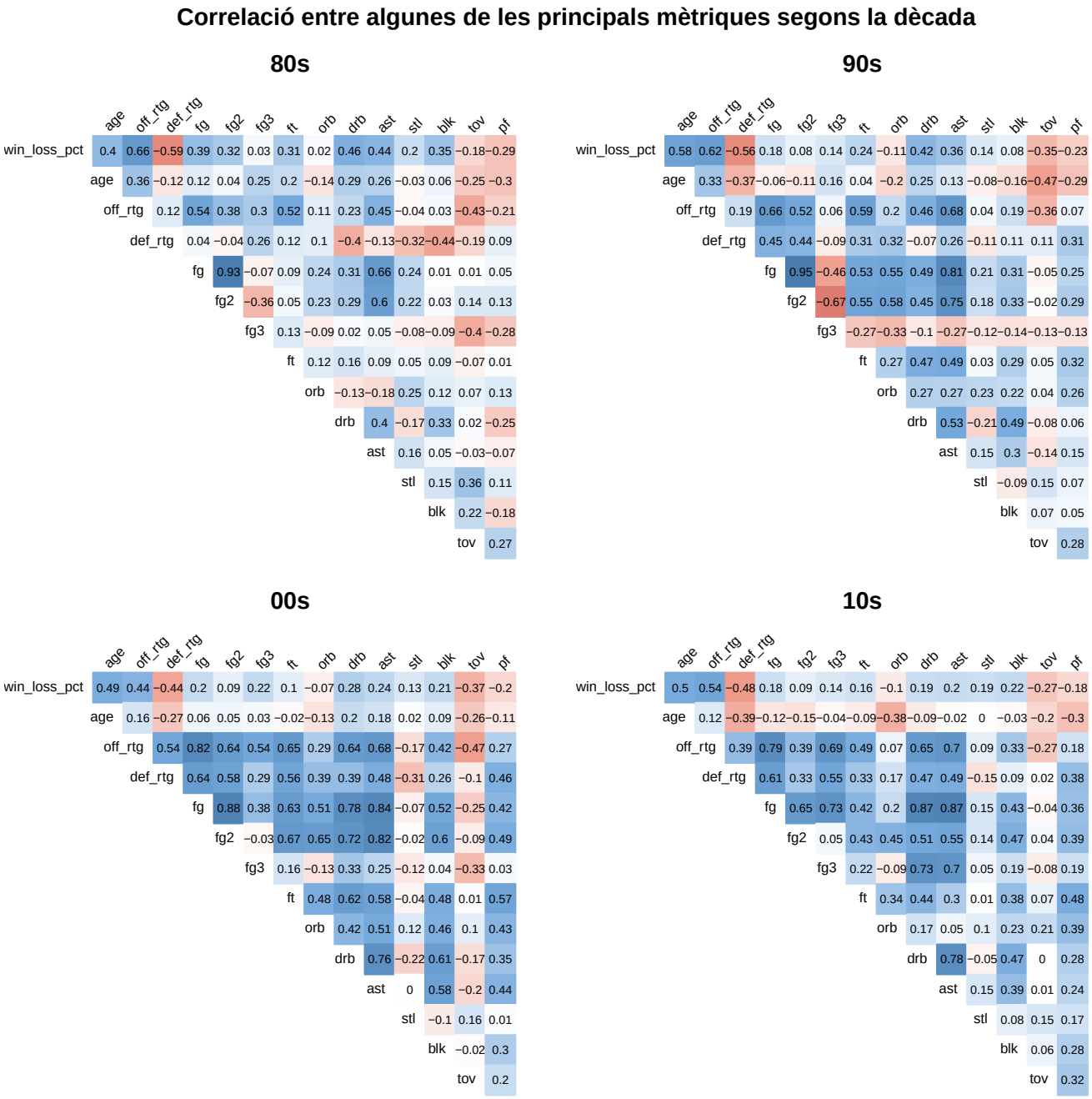
Paràmetre	Nombre d'observacions
Total d'observacions	105
Superior a 0.05	73
Inferior a 0.05	32

Veiem que aproximadament un terç de les distribucions no són gaussianes, ja que refutem la hipòtesi nul · la

en 32 dels 105 casos.

Anàlisi 2.3: Aplicació de proves estadístiques

Com algunes de les mostres que volem comparar no tenen distribució normal, hem d'utilitzar un mètode no paramètric, en comptes d'utilitzar la típica correlació de Pearson. En aquest cas, farem servir el mètode de correlació per rangs de Spearman, que és un mètode no paramètric.



Destaquem algunes coses d'aquestes correlacions:

- **Anys 80**: la correlació positiva més forta amb el percentatge de victòries és amb el *rating* ofensiu, mentre que la més negativa és amb el *rating* defensiu (la qual cosa té sentit, ja que les millors defenses tindran *ratings* defensius menors, mentre que les pitjors tindran un valor més elevat d'aquesta mètrica). En aquesta dècada també trobem la correlació més forta entre aquest percentatge i l'encert en els llançaments (tant totals com de dos, representats a les variables *fg* i *fg2*, respectivament). Hi ha un parell d'aspectes curiosos d'aquesta època quan la comparem amb les altres:

  - El *rating* ofensiu està correlacionat positivament sobretot amb l'encert en els llançaments i les assistències, però la resta de correlacions són dèbils (traient les pèrdues de pilotes o *tov*, que correlacionen negativament amb aquest *rating* a totes les èpoques). En altres èpoques, veiem per exemple una relació més forta amb el *rating* defensiu (especialment a partir dels 2000), el que ens suggeriria que els millors atacs també impliquen encaixar més punts. També hi ha una correlació més gran de les assistències en les èpoques posteriors als 80, el que podria tenir sentit ja que el joc es torna més col · laboratiu i menys individualista amb els anys.
  - El *rating* defensiu no té cap correlació positiva forta amb cap altra variable. Té algunes negatives que tenen cert sentit: els rebots defensius, les pilotes robades i els taps. És a dir, estadístiques de caire molt defensiu, motiu pel qual és lògic que aquesta correlació sigui negativa. En altres èpoques, hi ha correlacions positives més fortes amb altres variables, com és el cas del percentatge de llançaments encertats. Com hem dit anteriorment, tenir els millors atacs pot significar també encaixar més punts (pel ritme de partit que imposen, per exemple).

- **Anys 90**: l'edat i els *ratings* ofensius i defensius tenen les correlacions més fortes amb el percentatge de victòries. L'edat és un factor que correlaciona positivament amb les victòries en totes les èpoques, però en aquesta és on la correlació és més forta. En aquesta dècada ja veiem que el percentatge de llançaments encertats (especialment de dos punts) sembla influir menys en les victòries que en la dècada anterior. Comencem a veure les diferències pel que fa a les correlacions amb els ratings ofensius i defensiu si ho comparem amb els 80. Comencem a veure, per exemple, que capturar més rebots defensius i fer més taps correlaciona positivament amb un *rating* ofensiu superior. Pel que fa al *rating* defensiu, veiem que els percentatges d'encert en el llançament ja són un factor molt influent en aquest, mentre que estadístiques defensives com els rebots o les pilotes robades gairebé ja no tenen relació amb ell. Podem començar a intuir canvis en el joc.

- **Anys 2000**: segueix la tendència observada als 90 pel que fa a l'edat i els *ratings* com principals correlacions amb les victòries. El canvi més important respecte a la dècada passada és la correlació entre els percentatges d'encert en llançaments de tres (variable *fg3*) i el percentatge d'encert de dos (variable *fg*). Aquesta correlació és prou negativa als 80 i força més als 90s. És curiós perquè això ens indicaria que a major percentatge d'encert des del triple, menys percentatge d'encert en els dos punts. Ens podria suggerir que en els 90 van començar a aparèixer els especialistes de tres, és a dir, jugadors que anaven molt bé des de la línia de tres, però no tan bé des de distàncies més curtes. Aquesta correlació és pràcticament 0 als 2000, el que ens diu que no hi ha cap relació entre aquests dos paràmetres. Relacionat amb això, als 2000 és on comença a tenir un pes molt important el percentatge de llançaments de tres en el *rating* ofensiu i també en el percentatge d'encerts en general. Per tant, en aquesta època, hem analitzat com la capacitat ofensiva evoluciona a partir de l'aparició dels tirs de tres punts.

- **Anys 2010s fins l'actualitat**: el que més destaca és la importància del triple. Veiem que és l'era on menys correlació hi ha entre l'encert del llançament de dos i l'encert global, el que es pot explicar per la preponderància del tir de tres i la millora en l'encert d'aquest llançament respecte a dècades passades. Destaquem també la correlació entre les assistències i el percentatge d'encert de tres, que és la més positiva en aquesta era. Podríem dir, doncs, Per tant, l'estil de joc de l'NBA, ha evolucionat a partir d'un joc ofensiu que es basava en tirs de dos punts, cap a un estil on els tirs de tres punts tenen un rol central.

Anàlisi 3: És possible predir si un equip arribarà o no a playoffs en funció de les seves estadístiques?

En aquest últim estudi, mirarem de realitzar un predictor a partir d'un model supervisat per tal de predir si un equip es pot classificar a playoffs segons les estadístiques recollides durant la temporada. Per aquest cas, tornarem a dividir el conjunt de dades per èpoques, per continuar entenent l'evolució històrica de l'NBA i la WNBA.

Anàlisi 3.1: Selecció de les dades que es volen analitzar/comparar

Per a seleccionar les variables a utilitzar, mirarem d'afegir el % d'encerts per les estadístiques de tirs de camp. Aquesta ens permet entendre de millor el rendiment de l'equip sobre aquesta estadística. Així i tot, continuarem usant certes variables d'anteriors anàlisis i n'afegirem algunes de noves.

Table 13: Variables seleccionades pel model lineal

Variable	Descripció de la variable
age	edat mitjana
fg_pct	percentatge de llançaments de camps encertats
fga	total de llançaments intentats per partit
fg3_pct	percentatge de llançaments de 3 encertats
fg3a	total de llançaments de 3 intentats per partit
ft_pct	percentatge de tirs lliures encertats
orb	rebots ofensius per partit
drb	rebots defensius per partit
ast	assistències per partit
stl	pilotes robades per partit
tov	pilotes perdudes per partit
off_rtg	rating ofensiu de l'equip
def_rtg	rating defensiu de l'equip
pace	Estimació de posessions per partit
ts_pct	Percentatge True Shooting (tenint en compte tot tipus de llançaments)
efg_pct	Percentatge de tirs de camp efectius (ajustant pel valor del triple
opp_efg_pct	Percentatge de tirs de camp efectius de l'oponent
sos	Dificultat del calendari (nivell dels equips rivals)

Anàlisi 3.2: Preparació dels subconjunts de dades d'entrenament i prova.

Seguidament, haurem de crear els diferents subconjunts de dades per tal d'entrenar el nostre model i posteriorment avaluar-lo. Per realitzar la partició, ens basarem en el mètode d'exclusió (holdout) que ens permet dividir aleatòriament les dades en dos conjunts independents. El que farem serà utilitzar dos terços pel conjunt d'entrenament i el terç restant per les dades de prova.

Table 14: Observacions i variables dels conjunts d'entrenament i test dels anys 80

Conjunt	Nombre d'observacions	Nombre de variables
Entrenament	154	19
Test	77	19

En aquest cas, podem observar les dimensions dels primers subconjunts que fan referència a l'època dels 80. Trobem per al subconjunt d'entrenament 154 files amb un total de 19 variables. Així doncs, per al subconjunt de prova, trobem 77 files i 19 variables.

Anàlisi 3.3: Creació del model i avaluació a partir de la utilització de la regressió logística.

El mètode de model supervisat seleccionat es tracta del model de regressió logística. Aquest ens permetrà predir i classificar la variable categòrica *Playoffs* per tal de determinar si un equip pot entrar a aquests segons les seves estadístiques. No avaluarem en aquest cas la normalitat ni la variància de les variables ja que els models logístics no assumeixen que les dades segueixen una distribució normal ni tampoc homoscedasticitat.

Primer de tot, crearem el model a partir del subconjunt d'entrenament i finalment, mirarem de realitzar la predicció amb el subconjunt de prova per tal d'avaluar-lo. Per avaluar, crearem una matriu de confusió per cada una de les èpoques.

A continuació, presentem el resum de cada model, amb la seva matriu de confusió, les seves mesures de qualitat (sensitivitat, especificitat i precisió) i la corba ROC associada.

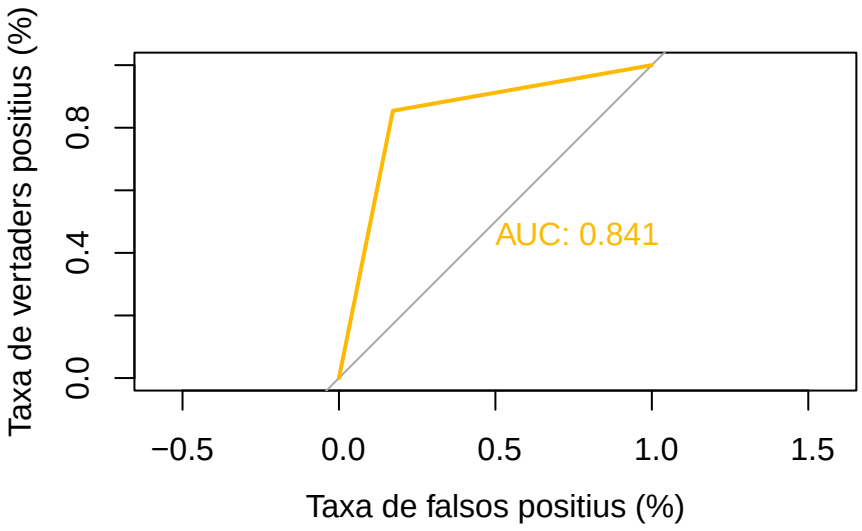
Comencem analitzant els anys 80:

Table 15: Matriu de confusió amb el model pels anys 80

	No	Yes
No	24	7
Yes	5	41

Table 16: Mesures de la qualitat del model per la dècada dels 80

Mesura	Descripció	Valor al model (%)
Sensitivitat	Proporció de registres positius correctament identificats	82.8
Especificitat	Proporció de registres negatius correctament identificats	85.4
Precisió	Proporció de registres classificats com a positius que ho són	77.4



Podem comprovar, a través de totes les mètriques i també veient la corva ROC, que el nostre model funciona força bé per a predir si un equip arribaria o no a playoffs als anys 80.

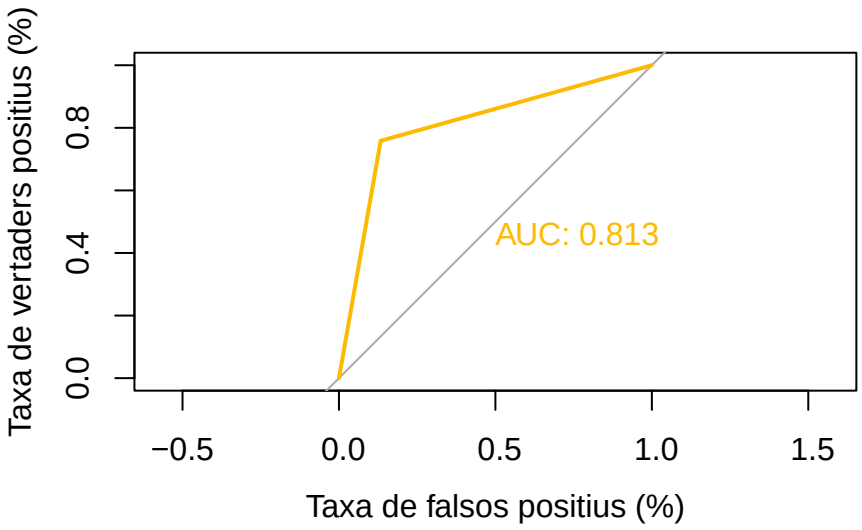
Passem a l'època dels 90.

Table 17: Matriu de confusió amb el model pels anys 90

	No	Yes
No	39	14
Yes	6	44

Table 18: Mesures de la qualitat del model per la dècada dels 90

Mesura	Descripció	Valor al model (%)
Sensitivitat	Proporció de registres positius correctament identificats	86.7
Especificitat	Proporció de registres negatius correctament identificats	75.9
Precisió	Proporció de registres classificats com a positius que ho són	73.6



En el cas dels anys 90, el model funciona una mica pitjor que a la dècada anterior, però tot i així encara pot predir força bé la possibilitat d'arribar a playoffs.

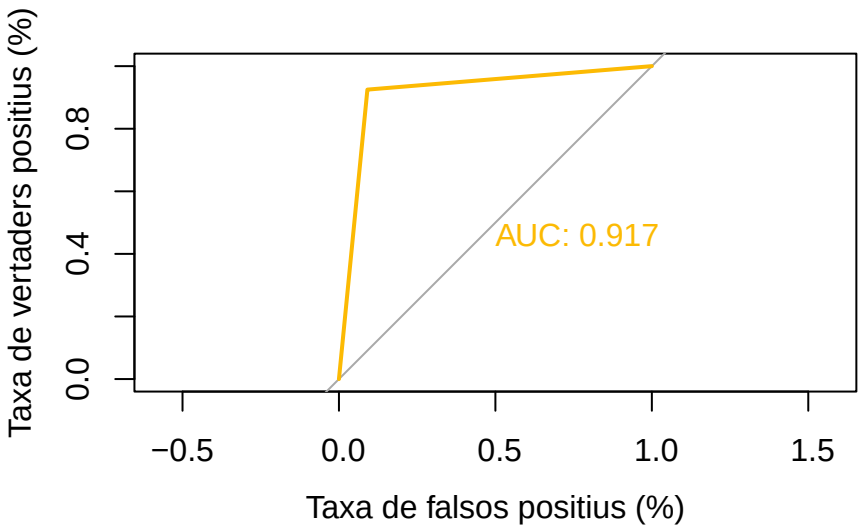
És el torn dels anys 2000:

Table 19: Matriu de confusió amb el model pels anys 2000

	No	Yes
No	60	6
Yes	6	74

Table 20: Mesures de la qualitat del model per la dècada dels 2000

Mesura	Descripció	Valor al model (%)
Sensitivitat	Proporció de registres positius correctament identificats	90.9
Especificitat	Proporció de registres negatius correctament identificats	92.5
Precisió	Proporció de registres classificats com a positius que ho són	90.9



Veiem una gran millora en el model per aquesta època, amb totes les mètriques per sobre del 90%, el que ens indica que és un model excel · lent per a predir l'arribada a playoffs.

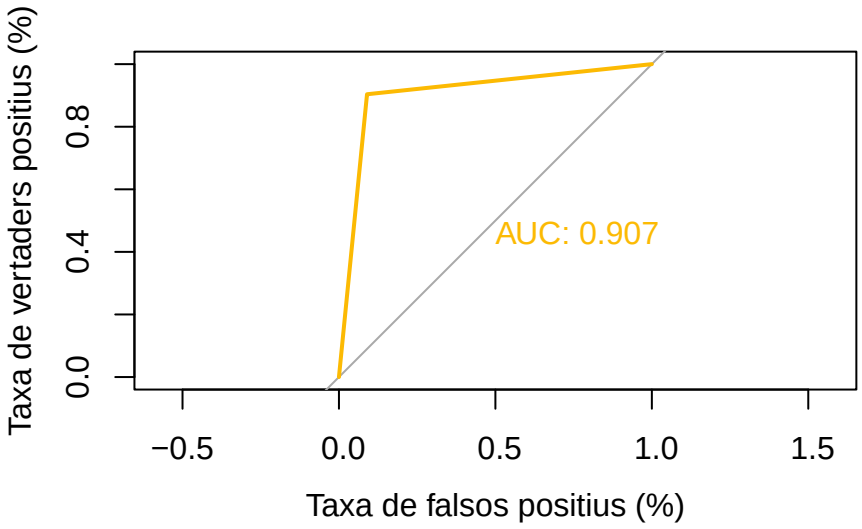
Finalment, analitzem el model de la dècada del 2010 fins l'actualitat:

Table 21: Matriu de confusió amb el model pels anys 2010 fins ara

	No	Yes
No	71	10
Yes	7	94

Table 22: Mesures de la qualitat del model per la dècada dels 2010 fins l'actualitat

Mesura	Descripció	Valor al model (%)
Sensitivitat	Proporció de registres positius correctament identificats	91.0
Especificitat	Proporció de registres negatius correctament identificats	90.4
Precisió	Proporció de registres classificats com a positius que ho són	87.7



Tenim un model molt similar al dels anys 2000, tot i que una mica inferior en la seva qualitat. Tot i així, aquí també hi ha més dades, així que una precisió de gairebé el 90% també ens permet afirmar que és un gran model per a predir l'arribada a playoffs.

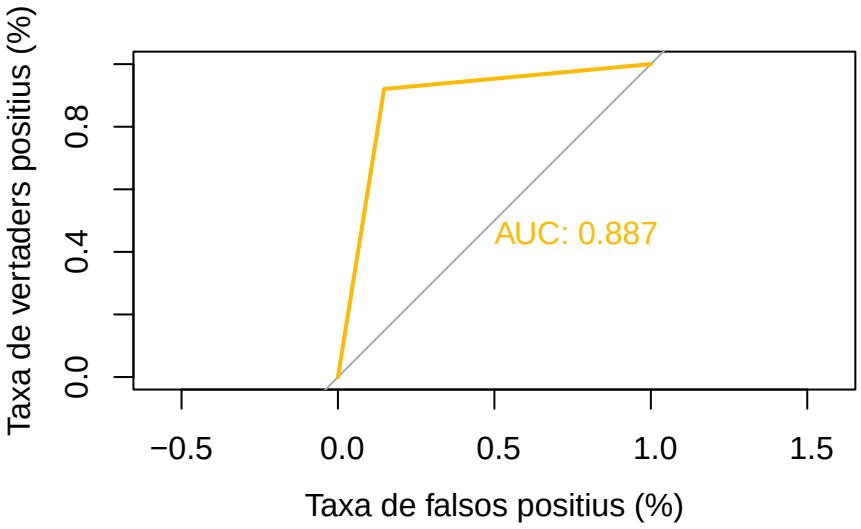
Per últim, realitzarem un model per a totes les èpoques analitzades.

Table 23: Matriu de confusió amb el model per totes les èpoques (post creació tir de 3)

	No	Yes
No	185	23
Yes	32	267

Table 24: Mesures de la qualitat del model per totes les èpoques (post creació tir de 3)

Mesura	Descripció	Valor al model (%)
Sensitivitat	Proporció de registres positius correctament identificats	85.3
Especificitat	Proporció de registres negatius correctament identificats	92.1
Precisió	Proporció de registres classificats com a positius que ho són	88.9



Notablement, el nostre model funciona de manera excel · lent per a totes les èpoques, amb mètriques molt properes (o superiors) al 90%. Analitzarem una mica més en detall aquest model.

Table 25: Resum del model per a totes les èpoques

Paràmetre	Estimate	Error estàndard	Estadístic	p-valor
(Intercept)	-6.5369395	7.4994204	-0.8716593	0.3833943
age	0.2348006	0.0838267	2.8010224	0.0050941
fg_pct	-1.2461021	65.6910864	-0.0189691	0.9848657
fga	-0.3717743	0.2764528	-1.3448020	0.1786892
fg3_pct	-4.3502207	4.5043605	-0.9657799	0.3341543
fg3a	-0.0096376	0.1422790	-0.0677374	0.9459946
ft_pct	18.8800082	10.6225080	1.7773588	0.0755092
orb	1.1569256	0.4635890	2.4955847	0.0125750
drb	-0.0368648	0.1242169	-0.2967773	0.7666365
ast	0.0616610	0.0930573	0.6626132	0.5075783
stl	0.2421971	0.2033813	1.1908522	0.2337116
tov	-1.2051315	0.4169613	-2.8902715	0.0038491
off_rtg	0.1014617	0.3715366	0.2730868	0.7847865
def_rtg	-0.7562916	0.1012898	-7.4666146	0.0000000
pace	0.3488325	0.2415946	1.4438752	0.1487741
ts_pct	66.5365804	119.7577321	0.5555932	0.5784890
efg_pct	59.1117316	80.6525265	0.7329185	0.4636081
opp_efg_pct	-11.9751861	14.3687777	-0.8334172	0.4046095
sos	-1.5763221	0.3326560	-4.7385957	0.0000022

Com podem veure al model, tenim cinc predictors amb significació estadística, el que ens suggereix que hi ha relació entre aquestes variables i la possibilitat d'arribar a playoffs. Aquestes variables són:

- **Edat**: sembla que hi ha una relació entre l'edat mitjana de l'equip i arribar a playoffs. Com té un valor de *estimate* positiu, això ens indica que els equips més veterans tindrien més possibilitats d'arribar a playoffs que els més joves segons el nostre model.

- **Rebots ofensius**: com en el cas de l'edat, el valor positiu de *estimate* indica que els equips amb més rebots ofensius per partit ho tindran millor per arribar a playoffs.
- **Pèrdues**: perdre molt la pilota disminuirà les possibilitats d'arribar a *playoffs*, tal com indica el valor negatiu de *estimate*.
- **Rating defensiu**: és el predictor amb el p-valor més baix, el que ens indica la seva forta relació amb arribar a playoffs. Veiem que té un valor *estimate* negatiu, el que vol dir que els equips amb valors més alts d'aquesta variable tindran menys probabilitats d'arribar a playoffs, ja que un valor més petit d'aquesta variable indica una millor defensa. Notablement, el rating ofensiu sembla no tenir cap relació amb arribar a playoffs. Això ens suggereix que el nostre model premia una gran defensa abans que un gran atac.
- **Dificultat del calendari**: com era d'esperar, tenir rivals més complicats dificulta l'arribada a playoffs. Recordem que a la NBA els equips juguen més amb els de la seva divisió que amb els d'altres divisions. Si tens a la teva divisió rivals complicats, això complicarà els playoffs. El que ens indica el seu *estimate* negatiu és que a major dificultat de calendari (és un percentatge calculat en funció de les victòries dels rivals), menys probabilitats d'arribar a playoffs.

Destaquem també la variable *ft_pct*, és a dir, el percentatge de tirs lliures encertats. Amb un valor de *estimate* molt positiu, es queda prop de tenir un p-valor significatiu (0.078). Això dóna rellevància a l'encert d'aquest tipus de llançaments, que de vegades semblen valorar-se menys que els llançaments de dos o els triples. Per finalitzar, també parlarem de les mètriques avançades *ts_pct* (o *True Shooting Percentage*, una mètrica d'eficiència en el tir que té en compte l'encert en tot el tipus de llançaments, amb fórmula *FGA + 0.44 * FTA*) i *efg_pct* (tirs de camp efectius, on es dóna més valor al llançament de tres i que té com a fórmula (FG + 0.5 * 3P) / FGA). Són dues mètriques amb valors *estimate* molt positius, el que clarament posa de manifest la seva importància, però amb p-valor molt per sobre de 0.5, el que ens diu que no són bons predictors segons el nostre model.

Exercici 6: Resolució del problema.

El conjunt de dades que presentem es basa en desenes de mètriques estadístiques sobre les temporades de l'NBA i de la WNBA. Aquestes es van recollir durant la primera pràctica a partir de l'ús de tècniques de *web scrapping*. Aquest conjunt presenta 1900 observacions i 53 variables. En aquest, podem trobar des d'estadístiques bàsiques fins a altres més complexes sobre el joc d'un equip durant la temporada regular.

Un cop hem començat a analitzar el conjunt de dades, hem trobat que a partir de la temporada 1973-74 enrere no trobem els registres sobre estadístiques dels equips. Per tant, ens dificulta la seva anàlisi per falta d'informació. Així doncs, hem descartat aquestes temporades. Un altre fet a tenir en compte, és que els tirs de tres punts no s'inclou fins als anys 1979. Això no suposa un problema, ja que podem omplir els valors buits amb el valor 0, perquè es tracta d'un valor que representa perfectament l'estadística. Un cop realitzar els canvis, ja disposàvem d'un conjunt de dades on podíem garantir la qualitat d'aquestes. Així doncs, hem començat a analitzar aquest.

La primera anàlisi que s'ha realitzat es basava en l'evolució dels tirs de camp, fent èmfasis en els tirs de tres punts. En el conjunt, disposem d'una estadística *fg3a_per_fga_pct* que ens permet recollir el percentatge de tirs intentats des de la distància de tres punts segons el total de tirs intentats. Així doncs, podíem veure a partir de les dades si el tir de tres punts prenia importància segons el pas de les dècades. Hem utilitzat mètodes estadístics com el test de Welch, que és una alternativa a l'ANOVA clàssic quan no hi ha homoscedasticitat i finalment, el test de Games-Howell per acabar de veure les diferències significatives entre tots els grups. Els resultats obtinguts han estat similars als resultats obtinguts en un inici amb la representació visual de la variable.

El segon anàlisi consistia a avaluar quines variables presentaven una major importància en l'estil de joc de cada dècada a partir de les correlacions entre aquestes. Per aquest cas, hem utilitzat les estadístiques més

bàsiques sobre el joc, com per exemple, els tirs encertats de qualsevol distància, assistències, rebots, entre altres. Primerament, hem comprovat que una part de les variables no presentaven una distribució normal. En conseqüència hem optat per un mètode no paramètric com és el cas de la correlació de Pearson. Així doncs, com a conclusió general, hem detallat com l'estil de joc dels equips evoluciona a partir d'un joc pràcticament basat en els tirs de dos fins a arribar a un estil de joc més tàctic a partir de la combinació del joc associatiu (assistències) amb els tirs de tres punts, donant-li importància als rebots i faltes personals.

L'última anàlisi es basava en la creació d'un model que fos capaç de predir si un equip era capaç de classificar a playoffs segons les estadístiques que presentava sobre la temporada. Hem fet un model per a cada època i un general i hem comprovat, al calcular les seves mesures i veure la seva corva ROC, que són bons models predictius. Després, hem analitzat el model general i hem vist que hi havia certes variables que eren bones predictores, com tenir una bona defensa, no perdre gaire pilotes o tenir un calendari més fàcil.

Finalment, podem dir que el conjunt de dades presentat ens permet assolir els objectius plantejats a l'inici d'aquesta pràctica. Tot i que és cert que no disposem de tota la informació sobre les temporades inicials, aquest estudi ens permet entendre com han evolucionat les diferents lligues segons les dècades a través de les estadístiques recollides. Hem pogut destacar diferents patrons en els estils de joc i la direcció que presenta l'estil de joc actual. És a dir, els equips estan deixant de banda els tirs de dos pels tirs de tres punts. Amb aquests s'obtenen una major puntuació, però presenten una major dificultat. Així i tot, podríem assumir que els equips estan optant per jugadors més tècnics i tiradors, en lloc d'optar per un joc físic per jugar a prop de la cistella.

Exercici 7: Codi.

El codi de la pràctica es troba en l'arxiu R Markdown `PRA2cmtorroxrocaca.Rmd` pujat al GitHub. No el reproduïm aquí per falta d'espai.

Exercici 8: Vídeo.

L'enllaç al vídeo es troba a FFFFFFFFFF