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

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

12/01/2023

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. Així i tot, analitzarem quins valors buits disposem en el nostre conjunt de dades a partir d'aquesta temporada.

```
## fg3a_per_fga_pct             fg3             fg3a          fg3_pct
##              102             102              102              102
##           Season          League            Team             wins
##                0               0                0                0
##           losses    win_loss_pct               gb       pts_per_g
##                0               0                0                0
##    opp_pts_per_g        Playoffs              age       wins_pyth
##                0               0                0                0
##      losses_pyth             mov              sos              srs
##                0               0                0                0
##          off_rtg         def_rtg          net_rtg             pace
##                0               0                0                0
##  fta_per_fga_pct          ts_pct          efg_pct          tov_pct
##                0               0                0                0
##          orb_pct         ft_rate      opp_efg_pct      opp_tov_pct
##                0               0                0                0
##          drb_pct     opp_ft_rate                g               mp
##                0               0                0                0
##               fg              fga           fg_pct              fg2
##                0               0                0                0
##             fg2a         fg2_pct               ft              fta
##                0               0                0                0
```

```
##          ft_pct             orb             drb             trb
##               0               0               0               0
##             ast             stl             blk             tov
##               0               0               0               0
##              pf
##               0
```

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:

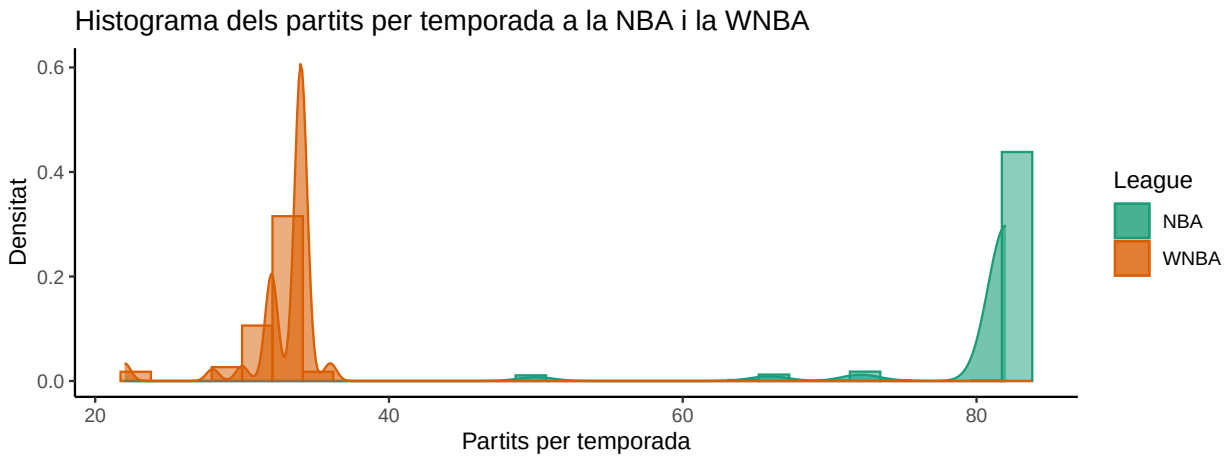
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   22.00   36.00   35.41   47.00   73.00
```

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?

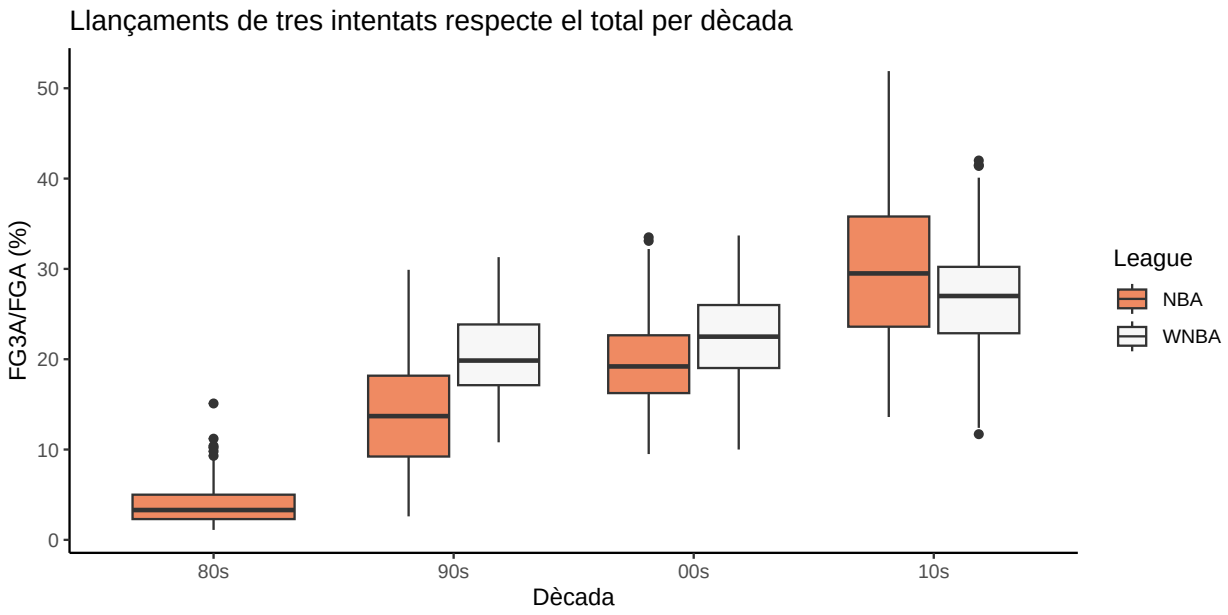
La introducció dels llançaments de tres a la lliga l'any 1979 va suposar un abans i un després a l'hora de jugar l'esport. No obstant, un espectador que no hagi vist res de bàsquet en els darrers 20 anys es podria sorprendre amb la quantitat aparent de llançaments de tres que es practiquen avui en dia.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

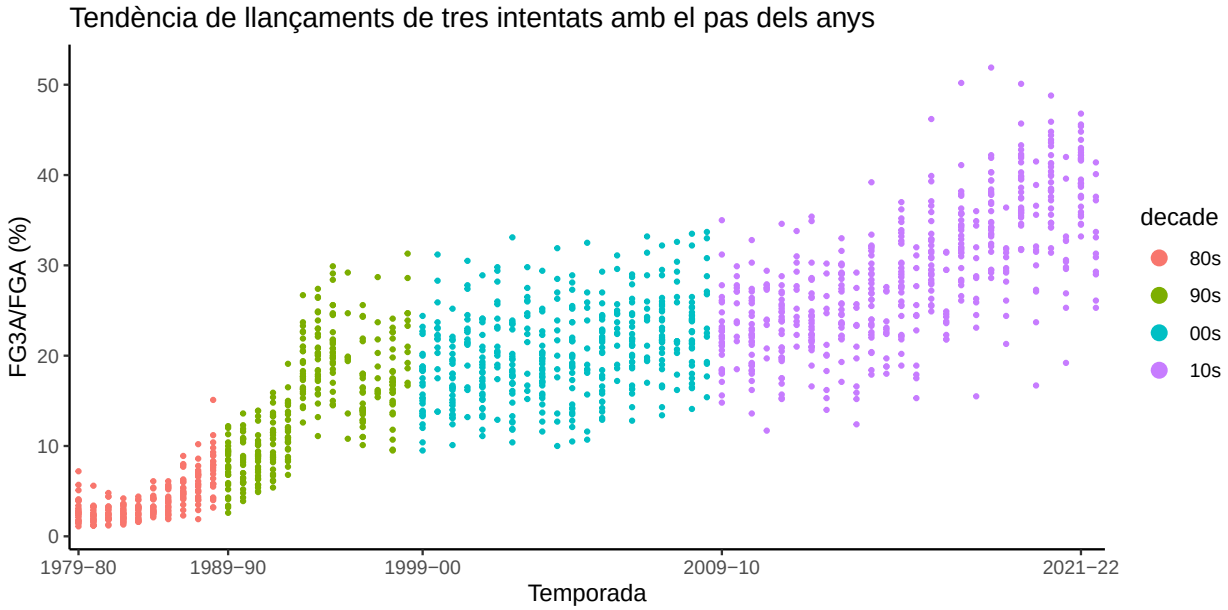
Per analitzar-ho millor, farem un anàlisi dècada per dècada dels llançaments de tres. Ens fixarem en una variable en concret: `fg3a_per_fga_pct`. Aquesta variable indica (en percentatge) quants llançaments de tres ha realitzat un equip de tots els llançaments intentats. És a dir, si de 10 tirs de camp, 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 3: 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 4: 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 5: 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.

Table 6: Nombre d'observacions que tenim per dècada

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

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: 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 8: $Diferènciesentredècadesalavariablefg3a_per_fga_pct$

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: Estudi sobre les estadístiques bàsiques i generals sobre el joc durant les diferents èpoques

Una altra cosa que ens preguntem és quins són els factors que més influeixen en el joc de manera històrica. És a dir, quines variables tenen més pes quan, a final de temporada, mirem les victòries i derrotes d'un equip.

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

Com disposem d'un gran nombre de variables, farem una correlació amb una selecció d'elles. Com hem comentat abans, no seria raonable utilitzar el valor absolut de victòries, ja que aquest pot variar en funció del nombre de partits totals que es juguen; per tant, utilitzarem el percentatge de victòries (variable *win_loss_pct*). A continuació, detallem quines variables seleccionarem per l'estudi:

Així, hem fet una selecció d'estadístiques clàssiques de la NBA per a veure quines d'elles influeixen en el percentatge de victòries a final de temporada. Abans de tot, però, haurem de transformar algunes variables.

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,

Table 9: Variables seleccionades pel model lineal

Variable	Descripció de la variable
age	és l'edat mitjana de l'equip un factor influent?
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

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

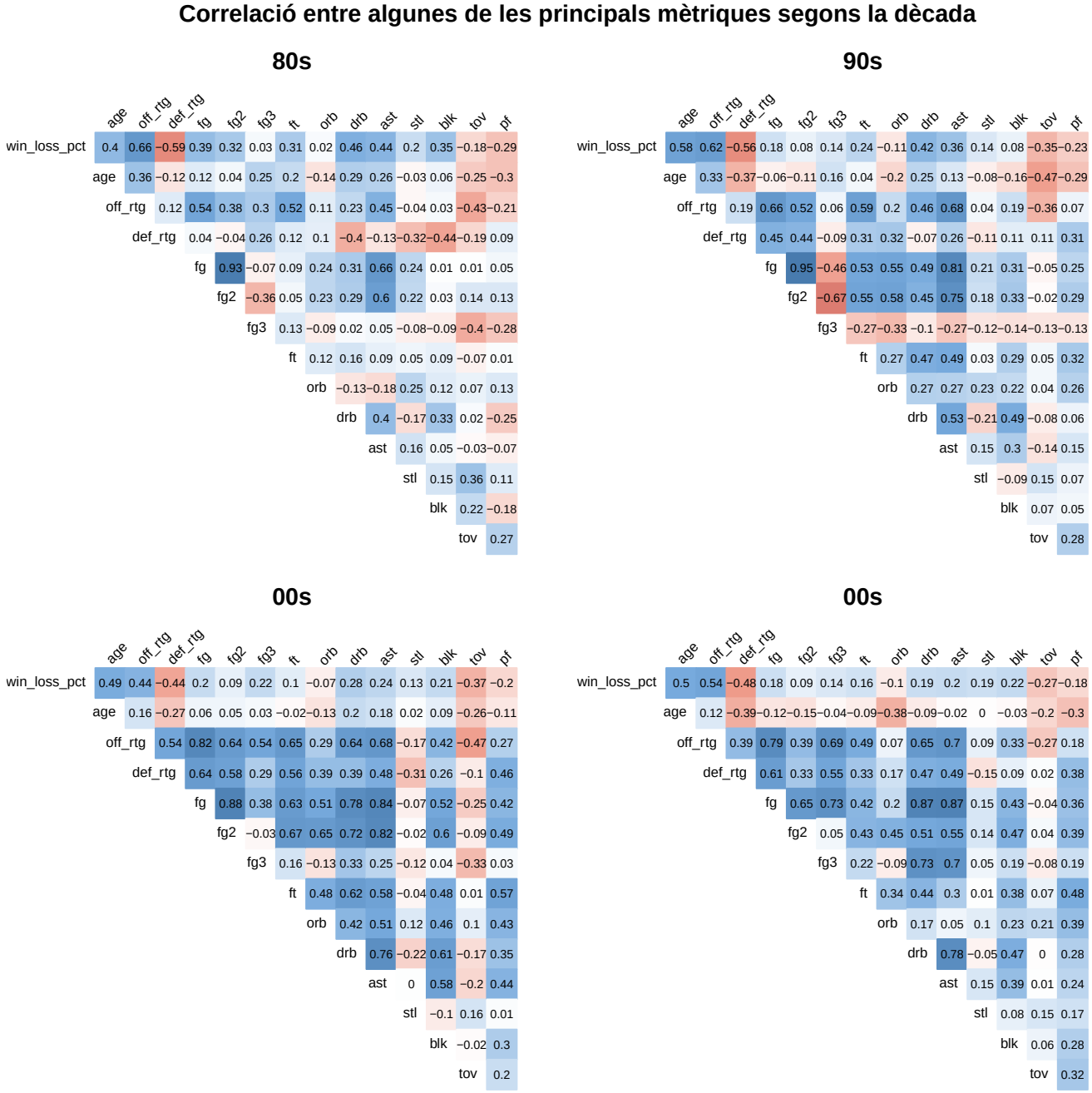
Table 12: Resum de les distribucions de les variables en funció del seu p-valor

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.



Per tant, explicarem quines correlacions trobem en cada època i així doncs, entendre quin estil de joc predominava en cada època.

Durant l'època dels 80, trobem que el fet que augmentava el percentatge de victòries es tractava de la capacitat ofensiva, en comptes de la defensiva. On trobem una correlació negativa per la segona. Altres factors que afavorien un millor percentatge de victòries eren les assistències, els rebots defensius i els tirs de camp. Si ens fixem en aquests últims, predominaven els tirs de dos punts. A més, les assistències presenten una forta correlació amb aquesta distància/puntuació. Finalment, observem com els rebots ofensius no eren gaire importants en aquesta època. Per tant, trobem un estil de joc on es necessitava una gran capacitat ofensiva basada en tirs de dos punts i assistències, on podem també destacar els rebots defensius com a

capacitat defensiva.

En l'època dels 90, trobem resultats bastant similars. Un punt diferenciador seria l'edat com a estadística amb més importància que en l'anterior època per aconseguir un millor percentatge de victòries. És a dir, afegim una mitjana d'edat més elevada, una plantilla amb més experiència per tal de guanyar partits. Pels tirs de camp, encara predominen els tirs de dos i les assistències en aquesta distància. Seguidament, podem observar com els rebots ofensius comencen a prendre més valor i aquests presenten una millor correlació amb els tirs de dos punts. Finalment, trobem com la mitja de pilotes perdudes presenta una correlació amb faltes personals. Per tant, podríem assumir que els equips comencen a cometre més faltes personals conjuntament amb les pilotes perdudes, és a dir, introduirien el concepte de faltes tàctiques. En aquesta època, encara trobem el factor ofensiu i afegim l'experiència dels jugadors com a variables claus per assolir victòries. Aquests trets ofensius són força similars als de l'anterior època. Un fet interessant és que es comença a tenir en compte la importància d'obtenir rebots ofensius i cometre faltes tàctiques en pèrdues de pilota.

En l'època dels 00, comencem a observar una evolució de l'estil de joc. Si parlem de percentatge de victòries, trobem les mateixes correlacions observades en les anteriors èpoques. En aquest cas, els tirs de camp, es veuen diferenciats en els tirs de dos punts i tres punts. Per tant, iniciem una època on es comença a donar importància a aquesta nova distància. Així i tot, el joc associatiu es continua basant en els tirs de dos punts, deixant el tir de tres per individualitats. Per als rebots ofensius, veiem com tenen més correlació amb els tirs de dos que amb els de tres, podríem assumir que en realitzar un tir de dos punts, els jugadors presenten una millor col · locació a la pista contrària. Per tant, en aquesta època, hem analitzat com la capacitat ofensiva evoluciona a partir de l'aparició dels tirs de tres punts.

Finalment, en l'època dels 10, seguim observant l'evolució de l'estil de joc. Per aconseguir victòries, trobem les mateixes correlacions. En aquest cas, quan parlem de tirs de camp, trobem una forta correlació amb els tirs de tres punts, deixant de banda els de dos punts. A més, les assistències també es veuen orientades cap a aquesta distància. És a dir, els equips donen una major importància al tir de tres com a tret característic de l'estil de joc i presenten un major joc associatiu dirigit a aquesta distància. Per tant, podem dir que 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 de joc més tàctic (pèrdues de pilota i faltes) que es basa en els tirs de tres punts.

```
data <- wnba_data %>% filter(decade == "90s")
factors = data %>% select(all_of(n))
res <- cor(factors)
corrplot(res,method="color",tl.col="black", tl.srt=30, order = "original",
   number.cex=0.75,sig.level = 0.01, addCoef.col = "black")

data <- wnba_data %>% filter(decade == "00s")
factors = data %>% select(all_of(n))
res <- cor(factors)
corrplot(res,method="color",tl.col="black", tl.srt=30, order = "original",
   number.cex=0.75,sig.level = 0.01, addCoef.col = "black")

data <- wnba_data %>% filter(decade == "10s")
factors = data %>% select(all_of(n))
res <- cor(factors)
corrplot(res,method="color",tl.col="black", tl.srt=30, order = "original",
   number.cex=0.75,sig.level = 0.01, addCoef.col = "black")
```

Durant l'època dels 90, trobem com a variables importants per aconseguir un millor percentatge de victòries, l'edat, la classificació ofensiva, els tirs de tres punts i els tirs lliures. En canvi, si observem les correlacions segons els tirs, observem com es presenta una forta correlació amb els tirs de dos punts. A més, també podem correlacionar aquests amb els rebots defensius i les assistències. En aquesta època, trobem un predomini dels tirs de dos punts en la capacitat ofensiva, però observem que tenen una major importància els tirs de

tres a l'hora de guanyar partits. A més, el factor d'experiència també afavoreix a un millor percentatge de victòries.

En l'època dels 00, veiem una petita evolució en l'estil de joc. Per assolir millors resultats, trobem la classificació ofensiva i les assistències. És a dir, el joc associatiu, en aquesta època presenta un millor rendiment. Si observem els tirs de camp, veiem una distribució entre els tirs de dos i tres punts. A més, les assistències també es distribueixen en aquests dos atributs. Finalment, tornem a observar com en l'NBA, com s'incorpora la importància dels rebots ofensius en els tirs de dos punts i cometre faltes tàctiques en pèrdua de pilotes. Per tant, hem analitzat l'evolució cap a un joc amb més recursos ofensius, tirs de dos i tres punts i un joc associatiu que presenta una gran importància a l'hora d'obtenir victòries.

Finalment, a l'època dels 10, trobem com torna a evolucionar la capacitat ofensiva. En aquest cas, els tirs de dos punts tornen a ser importants per aconseguir un millor percentatge de victòries. Tornen a predominar els tirs de dos punts en els tirs de camp, en canvi, les assistències es veuen distribuïdes entre les dues distàncies. Per part dels rebots ofensius ja no presenten gaire importància. Així doncs, hem analitzat una altra evolució de l'estil de joc, on tornen a predominar els tirs de dos punts que tornen a ser claus per a obtenir victòries.

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

Una altra cosa que ens preguntem és quins són els factors que més influeixen en el joc de manera històrica. És a dir, quines variables tenen més pes quan, a final de temporada, mirem les victòries i derrotes d'un equip.

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

Com disposem d'un gran nombre de variables, farem una correlació amb una selecció d'elles. Com hem comentat abans, no seria raonable utilitzar el valor absolut de victòries, ja que aquest pot variar en funció del nombre de partits totals que es juguen; per tant, utilitzarem el percentatge de victòries (variable *win_loss_pct*). A continuació, detallem quines variables seleccionarem per l'estudi:

Table 13: Variables seleccionades pel model lineal

Variable	Descripció de la variable
age	és l'edat mitjana de l'equip un factor influent?
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

Així, hem fet una selecció d'estadístiques clàssiques de la NBA per a veure quines d'elles influeixen en el percentatge de victòries a final de temporada. Abans de tot, però, haurem de transformar algunes variables.

Exercici 5: Representació dels resultats.

Exercici 6: Resolució del problema.

Exercici 7: Codi.

Exercici 8: Vídeo.