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

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

11/01/2023

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:

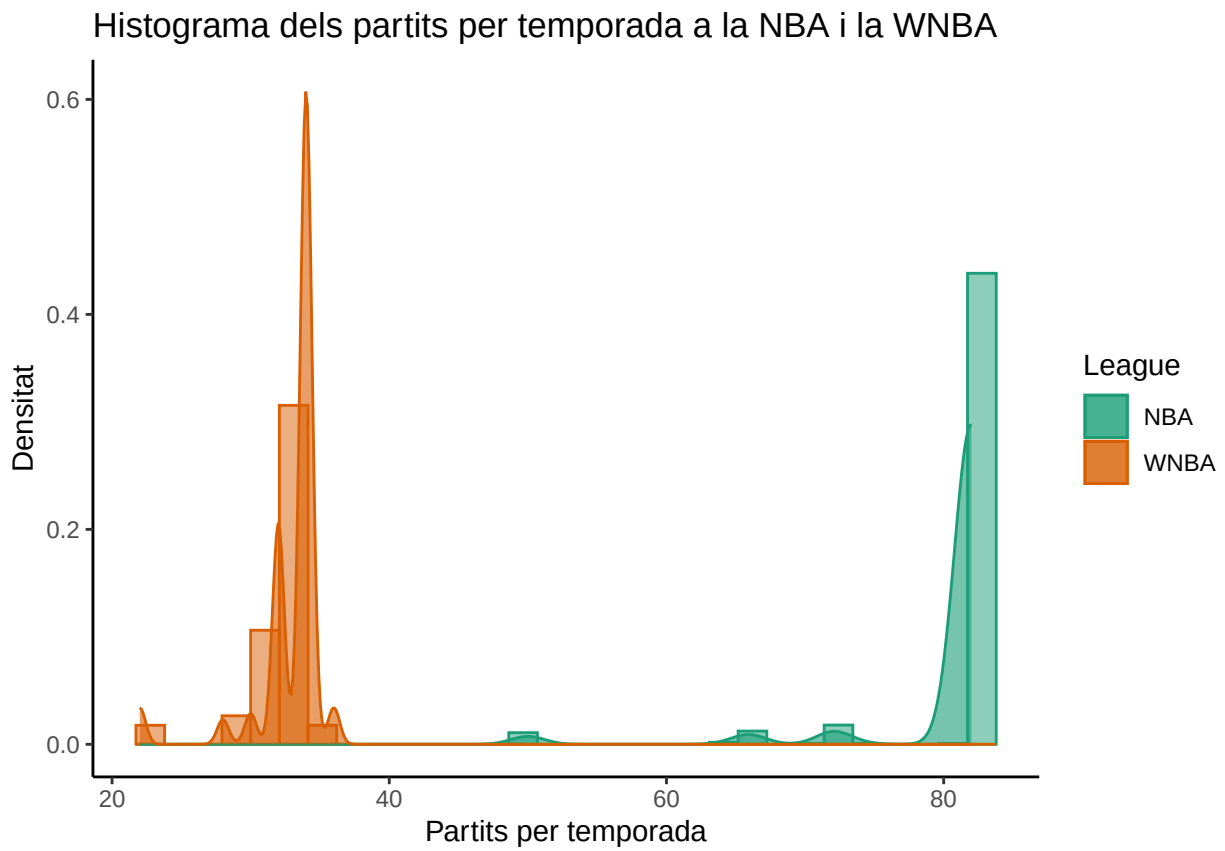
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   22.00   36.00   35.41   47.00   73.00
```

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.

4.1: Selecció dels grups de dades que es volen analitzar/comparar.

Per seleccionar els grups de dades a comparar, farem èmfasis en els mètodes d'anàlisi que es volen realitzar per tal d'escollir quines variables/camps ens poden ajudar per donar resposta a aquests.

- **Anàlisi estadística descriptiva**: Realitzarem una exploració de les dades per tal de veure quins equips presenten una mitjana de % de victòries més elevada durant els anys. Així doncs, ho compararem amb el percentatge de les vegades que ha estat classificat cada equip a playoffs per visualitzar la relació que poden tenir. A més, veurem l'evolució d'estadístiques bàsiques de cada temporada per veure la seva evolució i entendre quina tendència segueixen per avaluar com ha pogut canviar l'estil de joc de les grans lligues durant els anys.
- **Anàlisi estadística inferencial**: En aquest cas, realitzarem diferents regressions entre les estadístiques bàsiques en comparació a les victòries per tal d'analitzar la importància d'aquestes en el rendiment dels equips.
- **Model supervisat**: Finalment, realitzarem un model predictor que ens identifiqui si un equip té la possibilitat d'arribar a playoffs segons certes variables més avançades que intervenen en el joc.

Per tant, el conjunt de dades final tindria les següents variables:

Faltaria ficar la taula. Pendent del que vulguis fer

4.2: Comprovació de la normalitat i homogeneïtat de la variància.

Per tal de realitzar la comprovació de la normalitat, ens basarem en el test de Shapiro-Wilk, ja que es tracta dels més potents per contrastar aquesta. Assumirem com a hipòtesi nul · la que la població està distribuïda normalment. Direm que $\alpha = 0.05$, per tant, si p-valor és major a α, assumirem que les dades segueixen una distribució normal.

```
##
##  Shapiro-Wilk normality test
##
## data:  df3$fg3
## W = 0.9643, p-value < 2.2e-16
```

Seguidament, per comprovar l'homogeneïtat de la variància, ens basarem en els tests de Levene, per les que segueixen una distribució normal i Fligner-Killeen, per les que no les compleixen. Assumirem com a hipòtesi nul · la la igualtat de variàncies en els grups de dades. Direm que $\alpha = 0.05$, per tant, si p-valor és major a α, no podrem refutar aquesta hipòtesi nul · la i, per tant, podrem dir que hi haurà igualtat de variàncies.

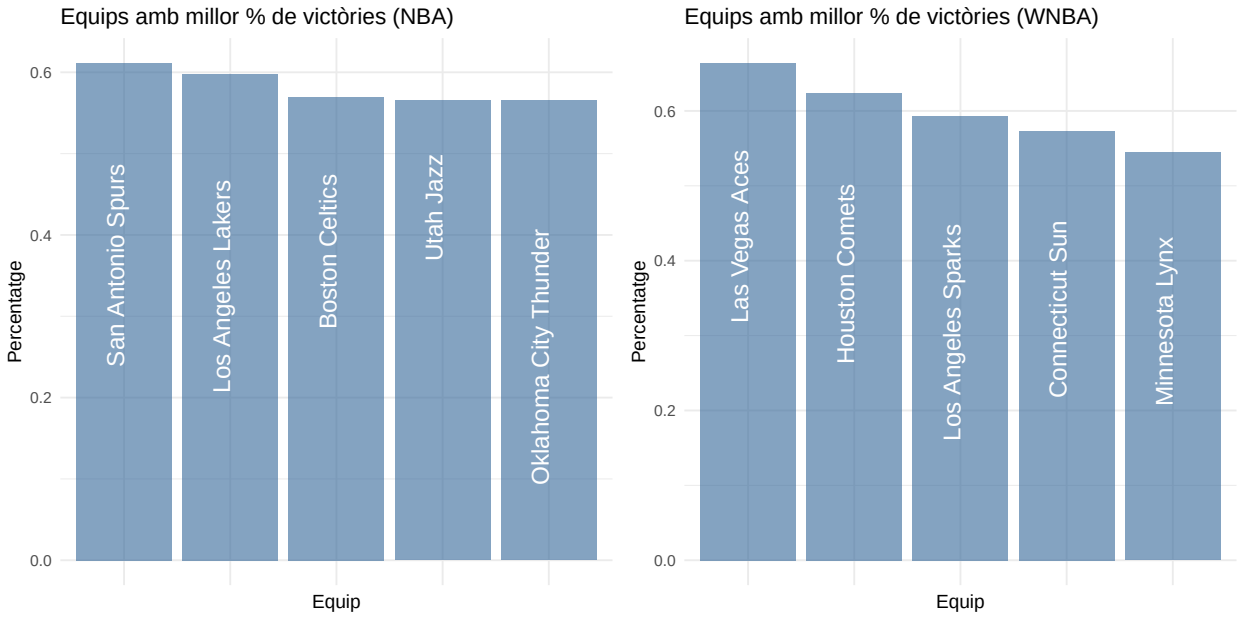
```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  wins by fg3
## Fligner-Killeen:med chi-squared = 194.45, df = 151, p-value = 0.009865
```

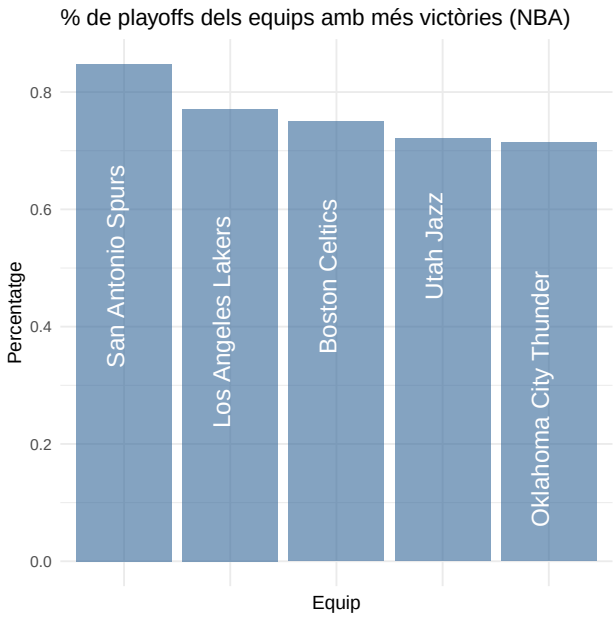
4.3: Aplicació de proves estadístiques per comparar els grups de dades. Aplicar almenys tres mètodes d'anàlisi diferents.

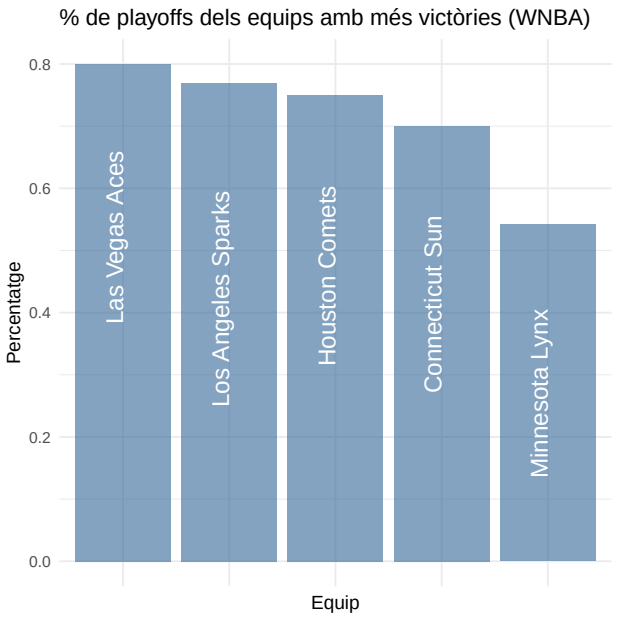
Seguidament, aplicarem les proves estadístiques descrites anteriorment.

Anàlisi estadística descriptiva

Començarem agrupant el percentatge de victòries i percentatge de vegades que ha entrat a playoffs cada equip i realitzarem la mitjana a partir del nombre de temporades. D'aquesta forma, estudiarem quins equips són els equips més guanyadors durant els anys i com es veu influenciat en les classificacions als playoffs.







Com es pot observar, veiem una alta relació en la mitjana de victòries aconseguides amb el percentatge de vegades que ha entrat l'equip a playoffs. Aquest fet pot resultar evident, ja que amb més victòries obtingudes en una temporada més possibilitats d'entrar a playoffs. Així i tot, si veiem a l'NBA, l'equip TOP 5 (Oklahoma City Thunder) té un percentatge proper a 60% de victòries i supera un 70% de vegades que ha classificat a playoffs. Per altra banda, si ho comparem amb el top 5 de la WNBA (Minnesota Lynx). Aquest té un percentatge de victòries proper al 50% i el seu % de vegades que ha entrat a playoffs és de quasi un 50% també. Fet que ens podria assegurar que aquest equip pot resultar irregular (diferència elevada de victòries/derrotes en les temporades) o que frega els límits de classificació cada temporada, fet que desemboca en aquesta irregularitat d'arribada a playoffs.

Anàlisi estadística descriptiva

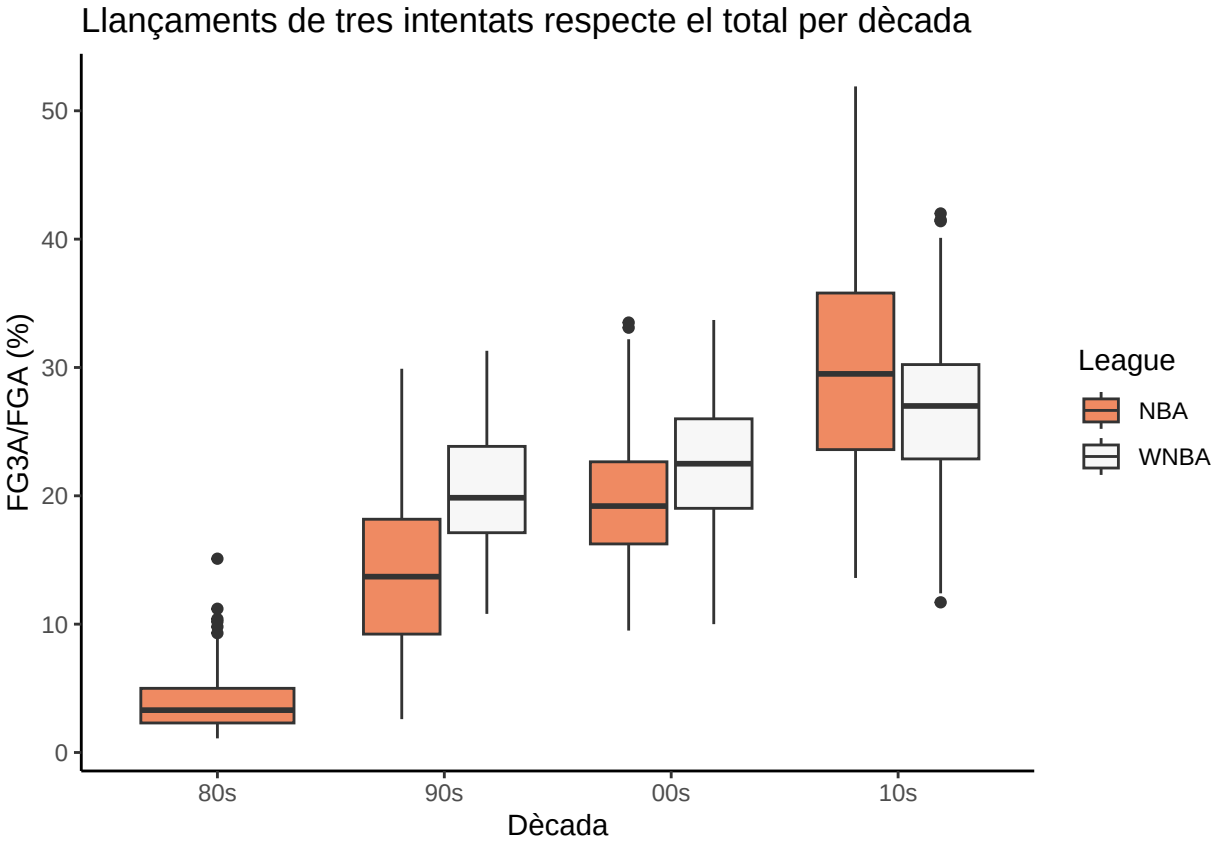
La introducció dels llançaments de tres a la lliga l'any 1979 va suposar un abans i un després a l'hora de jugar l'esport. No obstant, un espectador que no hagi vist res de bàsquet en els darrers 20 anys es podria

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.

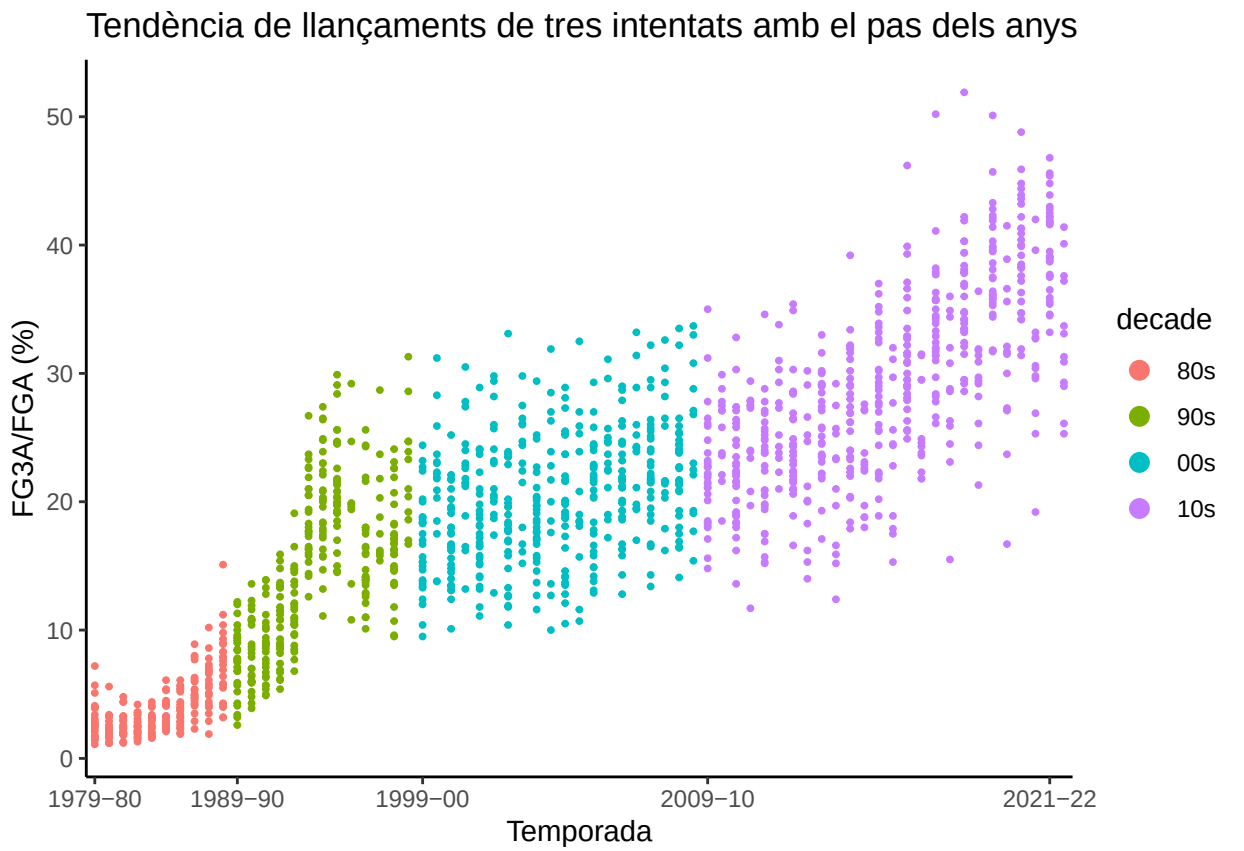
Per analitzar-ho millor, farem un anàlisi dècada per dècada dels llançaments de tres. Ens fixarem en una variable en concret: `fg3a_per_fga_pct`. Aquesta variable indica (en percentatge) quants llançaments de tres ha realitzat un equip de tots els llançaments intentats. És a dir, si de 10 tirs de camp,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

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.

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:

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

Exercici 5: Representació dels resultats.

Exercici 6: Resolució del problema.

Exercici 7: Codi.

Exercici 8: Vídeo.