PROJECTE KAGGLE

NBA FINALS TEAM STATS

Laia Rubio Castro

1600830

ÍNDEX

Índex

RESOLUCIÓ

Introducció

DATA BASE

Trobem dues bases de dades les quals tenen 24 variables i un total de 220 mostres respectivament.

Per una banda tenim la BD extreta del fitxer 'championsdata.csv' que conté un registre de totes les dades obtingudes de cada partit durant la final del campionat entre els anys 1980-2018 dels equips guanyadors.

Amb la funció describe() podem veure les 5 primeres mostres, que ens donen una idea del que ens trobarem a la base de dades.

	Year	Team	Game	Win	Home	MP	FG	FGA	FGP	TP	...	FTP	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
0	1980	Lakers	1	1	1	240	48	89	0.539	0	...	0.867	12	31	43	30	5	9	17	24	109
1	1980	Lakers	2	0	1	240	48	95	0.505	0	...	0.667	15	37	52	32	12	7	26	27	104
2	1980	Lakers	3	1	0	240	44	92	0.478	0	...	0.767	22	34	56	20	5	5	20	25	111
3	1980	Lakers	4	0	0	240	44	93	0.473	0	...	0.737	18	31	49	23	12	6	19	22	102
4	1980	Lakers	5	1	1	240	41	91	0.451	0	...	0.788	19	37	56	28	7	6	21	27	108

També mirem les seves estadístiques:

	Year	Game	Win	Home	MP	FG	FGA	FGP	TP	TPA	...	FTP	ORB	DRB	TRB	AST	STL
count	220.000	220.000	220.000	220.000	220.000	220.000	220.000	220.000	220.000	220.000	...	220.000	220.000	220.000	220.000	220.000	220.000
mean	1998.864	3.400	0.709	0.505	242.386	37.750	80.877	0.467	5.355	14.605	...	0.736	12.295	30.200	42.495	22.505	7.855
std	11.311	1.734	0.455	0.501	8.446	6.324	9.512	0.054	4.035	9.420	...	0.106	4.631	4.872	6.459	6.133	2.945
min	1980.000	1.000	0.000	0.000	240.000	25.000	62.000	0.289	0.000	0.000	...	0.368	3.000	16.000	22.000	11.000	1.000
25%	1989.000	2.000	0.000	0.000	240.000	33.000	75.000	0.430	2.000	6.750	...	0.667	9.000	27.000	38.000	18.000	6.000
50%	1999.000	3.000	1.000	1.000	240.000	37.000	80.000	0.467	5.000	15.000	...	0.740	12.000	30.000	42.000	22.000	8.000
75%	2009.000	5.000	1.000	1.000	240.000	42.000	87.000	0.500	8.000	20.000	...	0.816	15.000	33.250	47.000	27.000	10.000
max	2018.000	7.000	1.000	1.000	315.000	56.000	130.000	0.617	18.000	43.000	...	1.000	27.000	44.000	59.000	44.000	18.000

Un cop vistes les mostres, comprovem el tipus de dades contingudes a la BD i si hi ha valors nuls. Treballarem amb un total de 3 float64, 20 int64, 1 object i 6 valors nulls per a la mostra de TPP.

D'altra banda tenim la BD extreta del fitxer 'runnersupdata.csv' que conté un registre de totes les dades obtingudes de cada partit durant la final del campionat entre els anys 1980-2018 dels equips perdedors.

Com en el cas anterior utilitzem la funció describe() per veure les dades i estadístiques d'aquestes:

	Year	Team	Game	Win	Home	MP	FG	FGA	FGP	TP	...	FTP	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
0	1980	Sixers	1	0	0	240	40	90	0.444	0	...	0.786	14	26	40	28	12	13	14	17	102
1	1980	Sixers	2	1	0	240	43	85	0.506	0	...	0.778	5	29	34	34	14	11	20	21	107
2	1980	Sixers	3	0	1	240	45	93	0.484	1	...	0.588	13	24	37	34	12	8	13	25	101
3	1980	Sixers	4	1	1	240	41	79	0.519	0	...	0.885	5	29	34	31	5	10	14	20	105
4	1980	Sixers	5	0	0	240	42	94	0.447	0	...	0.792	13	29	42	32	9	7	12	25	103

RESOLUCIÓ

	Year	Game	Win	Home	MP	FG	FGA	FGP	TP	TPA	...	FTP	ORB	DRB	TRB	AST	STL
count	220.000	220.000	220.000	220.000	220.000	220.000	220.000	220.000	220.000	220.000	...	220.000	220.000	220.000	220.000	220.000	220.000
mean	1998.868	3.400	0.286	0.495	241.477	36.350	81.791	0.445	4.750	14.568	...	0.748	12.073	28.759	40.832	21.182	7.368
std	11.316	1.734	0.453	0.501	16.047	6.334	8.761	0.060	4.036	9.772	...	0.088	4.441	4.405	5.995	5.927	2.757
min	1980.000	1.000	0.000	0.000	40.000	21.000	60.000	0.299	0.000	0.000	...	0.500	2.000	20.000	26.000	8.000	0.000
25%	1989.000	2.000	0.000	0.000	240.000	32.000	76.750	0.409	1.000	7.000	...	0.684	9.000	26.000	37.000	17.000	5.000
50%	1999.000	3.000	0.000	0.000	240.000	36.000	81.000	0.443	4.000	13.500	...	0.750	12.000	29.000	41.000	21.000	7.000
75%	2009.000	5.000	1.000	1.000	240.000	40.000	87.000	0.484	7.000	20.250	...	0.806	15.000	32.000	44.000	25.000	9.000
max	2018.000	7.000	1.000	1.000	315.000	62.000	105.000	0.625	24.000	45.000	...	1.000	28.000	44.000	63.000	43.000	15.000

Per a aquesta BD treballarem amb un total de 3 float64, 20 int64, 1 object i 3 valors nulls per a la mostra de TPP.

El primer pas que farem un cop visualitzades les dades es combinar les dues BD per obtenir una amb totes les dades recopilades. Això ho fem mitjançant la funció de panda pd.concat(), on afegim al final de les dades del fitxer dels guanyadors, les dades dels perdedors.

Aleshores la BD sobre la que treballarem té un total de 24 variables i 220 mostres amb unes estadístiques finals de:

	Year	Game	Win	Home	MP	FG	FGA	FGP	TP	TPA	...	FTP	ORB	DRB	TRB	AST	STL
count	440.000	440.000	440.000	440.000	440.000	440.000	440.000	440.000	440.000	440.000	...	440.000	440.000	440.000	440.000	440.000	440.000
mean	1998.866	3.400	0.498	0.500	241.932	37.050	81.334	0.456	5.052	14.586	...	0.742	12.184	29.480	41.664	21.843	7.611
std	11.301	1.732	0.501	0.501	12.816	6.360	9.145	0.058	4.042	9.587	...	0.097	4.533	4.695	6.280	6.060	2.860
min	1980.000	1.000	0.000	0.000	40.000	21.000	60.000	0.289	0.000	0.000	...	0.368	2.000	16.000	22.000	8.000	0.000
25%	1989.000	2.000	0.000	0.000	240.000	32.000	75.000	0.416	2.000	7.000	...	0.677	9.000	26.000	37.000	17.000	6.000
50%	1999.000	3.000	0.000	0.500	240.000	37.000	81.000	0.455	4.000	14.000	...	0.744	12.000	30.000	41.000	21.000	7.000
75%	2009.000	5.000	1.000	1.000	240.000	41.000	87.000	0.493	8.000	20.000	...	0.811	15.000	33.000	45.000	26.000	9.000
max	2018.000	7.000	1.000	1.000	315.000	62.000	130.000	0.625	24.000	45.000	...	1.000	28.000	44.000	63.000	44.000	18.000

On les dades son 3 float64, 20 int64, 1 object i 9 valors nulls per a la mostra de TPP.

OBJECTIU

L'objectiu d'aquesta pràctica es utilitzar els coneixements apresos fins a la data per a analitzar les dades proporcionades per tal de, un cop aplicats tots els processos necessaris, poder identificar quins sons els factors que porten a guanyar.

HIPÒTESIS I PREDICCIONS

Deixant de banda el fet de que l'equip que més punts finals obté serà el guanyador, ja que això es irrefutable. Volem saber quines variables influeixen en l'obtenció de la victòria.

Les hipòtesis a comprovar seran:

- El fet de tenir minuts de pròrroga, es a dir, jugar més temps beneficia.
- Jugar a casa aporta algun avantatge.
- La intensitat del partir afecta a les possibilitats de guanyar.
- Hi ha diferencia entre fer tirs de 2 punts i triples.

RESOLUCIÓ

Plantejament de dades

SELECCIÓ DE DADES

Ara que ja sabem que volem comprovar hem de triar les variables que creiem que ens aportaran aquesta informació.

Per una part, tenim l'atribut Win, que es de tipus binari i es el que volem predir.

D'altra banda, utilitzarem l'atribut Year com a variable base per a fer les comparatives entre els diferents atributs.

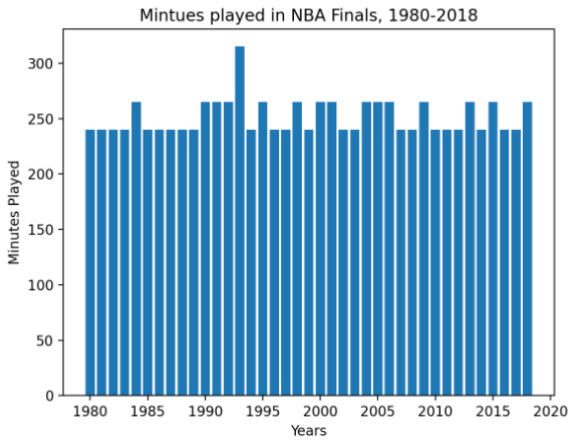
De totes les variables, per a realitzar l'anàlisi de les qüestions mencionades anteriorment prioritzarem les següents:

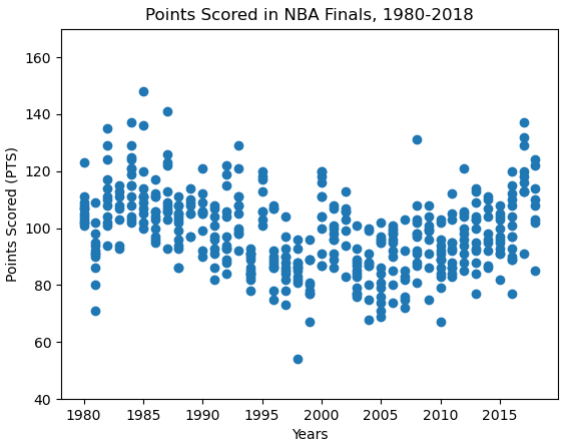
- MP (Minuts played)
- Home
- PTS (Points scored)
- TOV (Turnovers)
- PF (Personal fouls)
- FG (Field goals made)
- FGA (Field goals attempts)
- FGP (Field goals percentatge)
- TP (3 Point Field Goals Made)
- TPA (3 Point Field Goals attempts)
- TPP (3 Point Field Goals percentage)

ANÀLISI DE DADES

El primer de tot que volem saber es si el fet de tenir mes temps de joc beneficia en l'obtenció de punts i per tant, en la possibilitat de guanyar.

Per fer això creem una gràfica que mostra els minuts de joc per a cada any i comparem amb els punts finals marcats anualment.

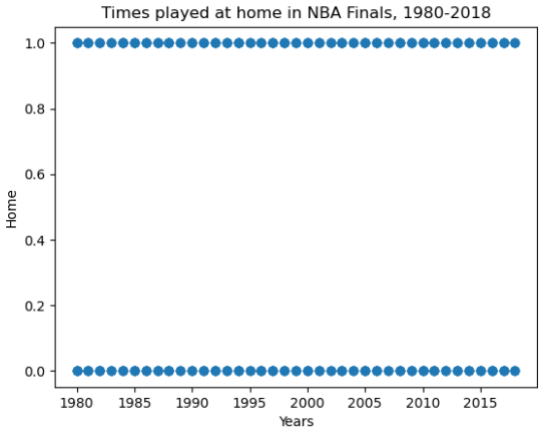


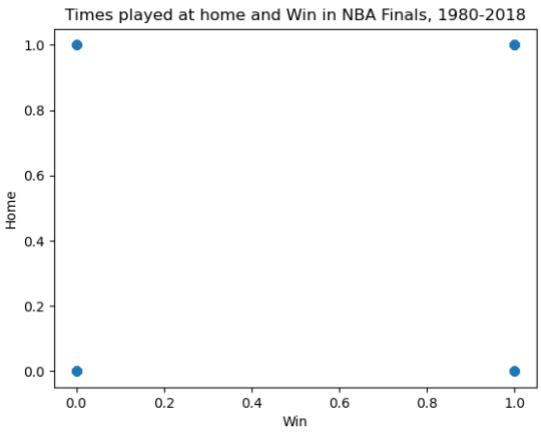


RESOLUCIÓ

Podem observar com haver jugat una pròrroga (gràfiques superiors a 240) no implica haver marcar mes punts. Aleshores podem contestar a la pregunta inicial que aquest no es un factor vinculant a la victòria.

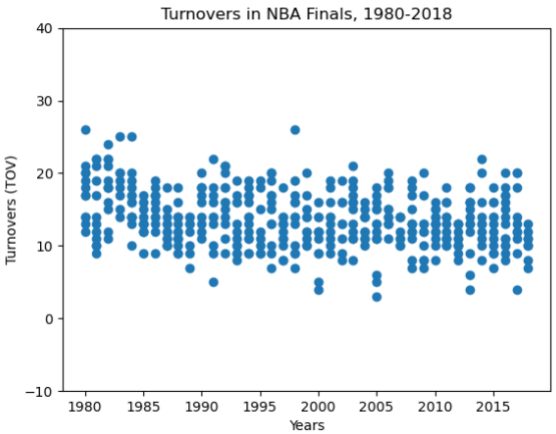
El segon punt que volem saber es si jugar a casa dona algun tipus d'avantatge. El primer que podríem pensar es que si, ja que es juga en terreny conegut. Però, fora dels avantatges psicològics que es puguin tenir les dades no aporten cap informació que validi aquesta creença. Com estem parlant d'una variable binaria només prendrà valors de 1/0 (s'ha jugat a casa o s'ha jugat a fora) per a cada any. I, en cas de voler veure la seva gràfica respecte la victòria, com aquesta variable també es binaria no ens aporta cop informació. Simplement ens esta mostrant totes les possibilitats (s'ha jugat a casa i s'ha perdut, s'ha jugat a casa i s'ha guanyat o s'ha jugat a fora i s'ha guanyat, s'ha jugat a fora i s'ha perdut).

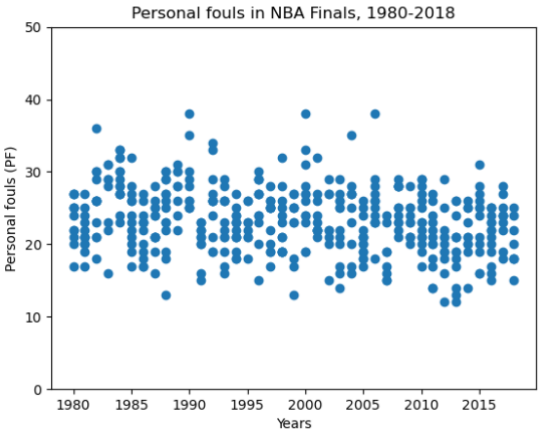




El tercer punt a saber es si la intensitat del partit afecta a les possibilitats de guanyar. Amb intensitat ens referim als possibles casos que poden portar a que l'equip contrari guanyi. Es a dir, perdre la pilota (TOV) o realitzar faltes personals (PF). Aquestes accions poden portar a una certa tensió que regirà les accions dels jugadors afectant la productivitat del partit, però volem saber si les dades reflexen això o no.

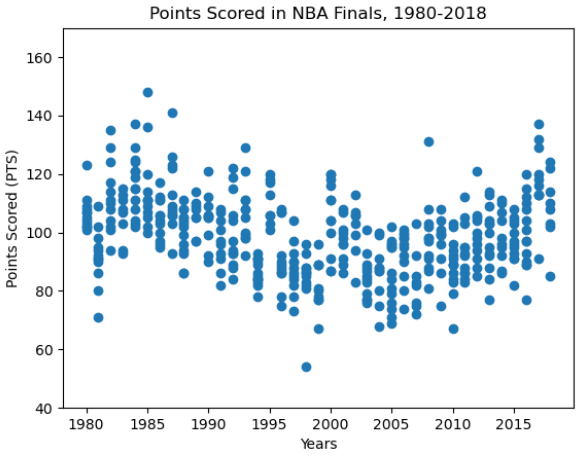
Visualitzem les gràfiques per a poder analitzar els resultats.





RESOLUCIÓ

A primera vista podem veure que els dos gràfics son bastant lineals i tenen una certa tendència a decréixer però ara compararem els dos gràfics amb el gràfic dels punts finals obtinguts per veure si tenen algun tipus de relació.



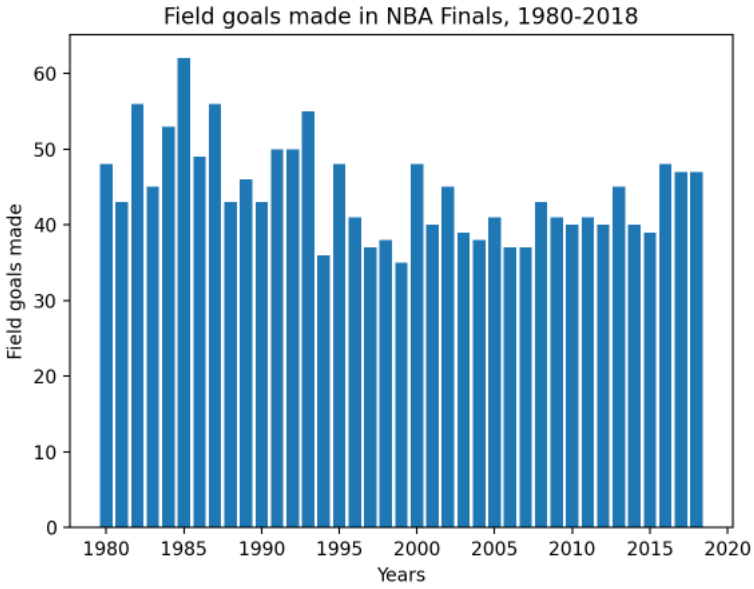
Com acabem de mencionar ambdós gràfics mostren una tendència bastant lineal amb una certa inclinació a la baixa. El gràfic del PTS tot i que baixa fins als finals dels 90, a principis dels 2000 comença a créixer.

Això ens indica que a menor intensitat major puntuació. Per tant, l'existència de TOV o PF afecta als PTS.

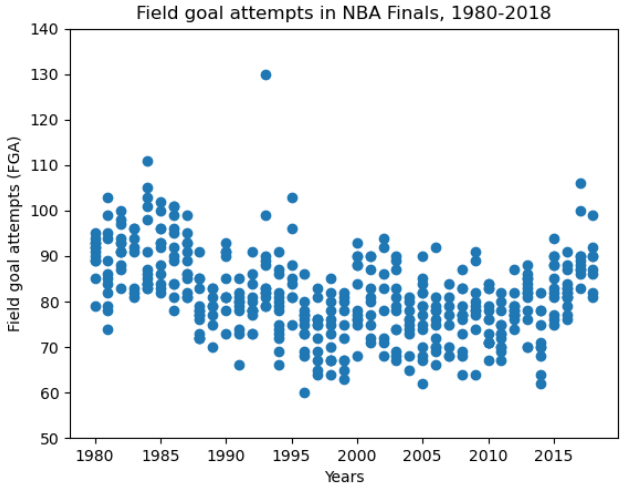
Finalment volem saber si hi ha alguna diferència entre fer tirs de 2 punts o triples. Per a fer això primer analitzarem la importància de cada cas per separat.

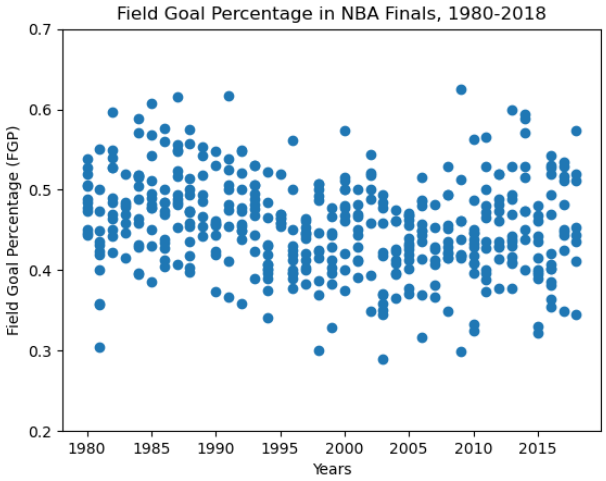
TIRS DE DOS PUNTS:

Per a veure la relació d'aquest tirs respecte la victòria mirarem la quantitat de tirs realitzats, el numero total d'intents a cistella i finalment el percentatge de tirs, tots tres casos representats per a cada any.



RESOLUCIÓ

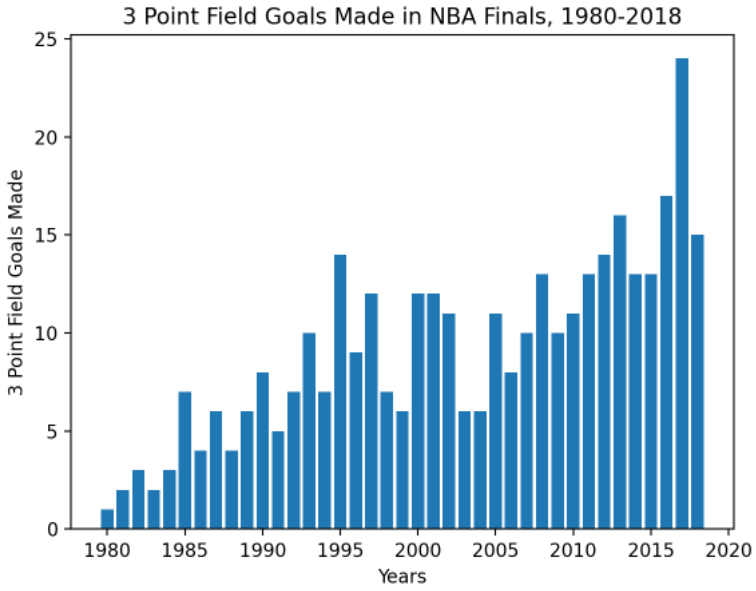




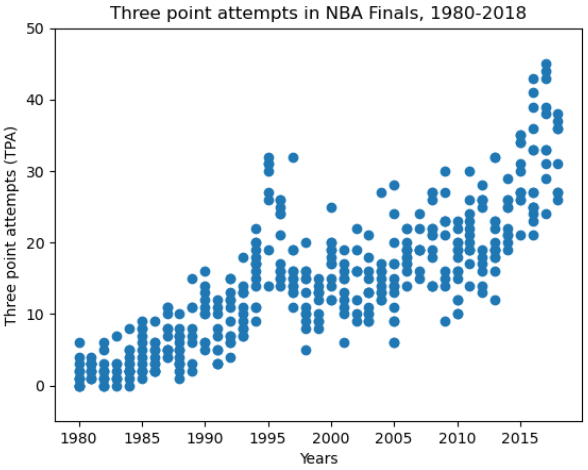
Les conclusions que podem obtenir de l'anàlisi dels tres gràfics es que el numero d'intents a principis dels 80 era mes elevat però va començar a caure fins als 2000. I, a partir d'aquell moment va començar a augmentar lleugerament. Tot i això, si mirem el gràfic dels punts aquests segueixen un sentit decreixent a nivell general i, tot i que es cert que creix en els darrers anys, no arriba als valors inicials en cap altre moment. Al gràfic de probabilitats podem veure que conté moltes mostres amb valors al voltat de 0.5 i tot i que la gran majoria son menors que aquest també conté molts valors majors. Això ens indica que de mitja com a mínim el 50% dels punts marcats son per tirs dobles.

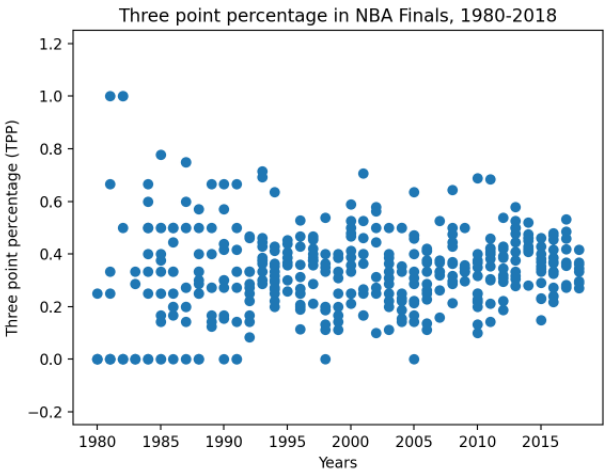
TIRS TRIPLES:

Realitzarem el mateix procés que pel cas anterior.



RESOLUCIÓ





Clarament podem veure que el nombre de triples marcats ha augmentat considerablement al llarg dels anys. Això ha estat perquè, com ens mostra la gràfica d'intents, als primers anys no es realitzaven gaires i aquesta tendència ha anat creixent als darrers anys. Això també es veu reflexat al gràfic de percentatges on els primers anys contenen molts 0s i a finals dels anys 90 això ja va desapareixent. Es pot veure que els valors no son gaire alts, ja que la majoria no passen el 0.5. Per tant, tot i que cada vegada es facin més triples aquests continuen sent menors que els dobles.

Ara que ja hem realitzat els estudis independents, podem arribar a la conclusió de que la tendència ha anat canviant al llarg dels anys. On, al principi era no apostar pels triples i tirar tirs de dos punts fins que actualment es tiren molts mes triples, disminuint així el nombre de tirs dobles.

RESOLUCIÓ

Creació del regressor

Un cop analitzades les dades, com ja sabíem en un principi el factor clau per a la victòria es tenir el màxim valor de punts finals possibles. Com els atributs de FG i TP estan directament relacionats amb els punts finals obtinguts no els tindrem en compte sinó que utilitzarem els nombre d'intents ambdós casos.

Com el que podem predir es un valor binari (Win) utilitzarem un regressor logístic.

Per tant, com a atributs predictors utilitzarem els valors de TOV, PF, TPA i FGA i com a atribut a predir Win.

RESOLUCIÓ

Conclusions

De tots els atributs seleccionats inicialment els únics rellevants han sigut TOV, PF, TPA i FGA. Aquests son els que tindran influencia sobre el resultat final.

Després de reiterats intents no s'han trobat resultats concloents. Tot i això degut als anàlisis obtinguts fins aquest punt sembla que la tendència més important resideix a la puntuació feta durant el temps de joc.