STATISTICAL SOFTWARE: R AND SAS
Practical exercise

Arnau García

December 27, 2023

1 Introduction

In this practical exercise we will be working with a data set of the GitHub repository of Victor Peña (teacher of the MESIO in the subject of Linear and Generalized Linear models). I am a loyal NBA (National Basket Association) fan. The NBA is currently the only sports league I follow on a daily basis. I keep up with everything going on in the league, and there are a number of players and teams that are arguably my favorites, even though I don't have a favorite team. In addition, the NBA, like other American leagues such as the baseball league, is a competition where statistics and data are key. Because of this, I have chosen a data set about the NBA. The reader can find the data set in `https://github.com/VicPena/data/tree/master/nba-elo`.

But it is not just any data set, it is a data set in which the ELO rating system (`https://en.wikipedia.org/wiki/Elo_rating_system`) is applied to study the quality of the teams and the probabilities of victory taking into account this scoring system. The ELO rating system was invented as a chess-rating system, this rating system is very interesting and has a lot of mathematics behind (the reader can find these details and other additional information about the ELO rating system in the previous link).

My goals in this practical exercise will be: use SAS for develop all the analysis and exploit the possible advantages of this programming language, and, asses if the ELO rating system is or not useful for understand how good are the NBA teams. In addition, I would like to study with this huge data set some personal curiosities. For instance, how good have been the teams I like the most throughout the history of the league. See if there really is a difference between home and away games. And study some aspects of the evolution of the league, such as the evolution of points scored over the years.

2 About the data set

The selected data set meets all the conditions of the statement of work. The data set has a total of 126314 observations, and a total of 23 variables. The data related with the ELO rating system in the data set is from `https://projects.fivethirtyeight.com/complete-history-of-the-nba/#nuggets`. And the data related with the games is from `https://www.basketball-reference.com/`. The variables of the data set are:

1. gameorder: Play order of game in NBA history.
2. game-id: Unique ID for each game.
3. lg-id: Which league the game was played in
4. -iscopy: Each row of data is tied to a single team for a single game, so -iscopy flags if this game-id has already occured for the opposing team in the same matchup.
5. year-id: Season id, named based on year in which the season ended.
6. date-game: Game date.
7. is-playoffs: Flag for playoff games.
8. team-id: Three letter code for team name, from Basketball Reference
9. fran-id: Franchise id. Multiple team-ids can fall under the same fran-id due to name changes or moves. Interactive is grouped by fran-id.
10. pts: Points scored by team.
11. elo-i: Team elo entering the game.
12. elo-n: Team elo following the game.
13. win-equiv: Equivalent number of wins in a 82-game season for a team of elo-n quality.
14. opp-id: Team id of opponent.
15. opp-fran: Franchise id of opponent.
16. opp-pts: Points scored by opponent.
17. opp-elo-i: Opponent elo entering the game.
18. opp-elo-n: Opponent elo following the game.
19. game-location: Home (H), away (A), or neutral (N).
20. game-result: Win or loss for team in the team-id column.

21. forecast: Elo-based chances of winning for the team in the team-id column, based on elo ratings and game location.
22. Notes: Additional information.

The identification variable is the Game id. Observe that, as the variable "-iscopy" indicates, in our data set we have the games duplicated, in the sense that each row corresponds to a single team and a single game. The variable "-iscopy" works as an indicator that is 1 if the game is a copy, and 0 if not. Due this, we can deduce some facts about the data set, for example, for those games duplicated, the "forecast" variable should sum 1. It is, for a given game, for instance, Denver Nuggets VS Los Angeles Lakers, playing in Denver, the elo-based chances of win for the Nuggets is p, and then, due that in the NBA there are no ties, the probability of win for the Lakers should be $1 - p$. If we observe several observations of the data set (see in Figure 1) we can see this.

It is important to comment that I have carried out this work using SAS on demand. And that the proc import was not working correctly for CSV files (which is the file type found on GitHub). I tried with other CSV, and it didn't work correctly either. So, I decided to convert the CSV to EXCEL, and then import in SAS the excel.

3 Methods and results

In this section of the report we will be referring many times to the SAS script. We start the analysis importing the EXCEL file (via proc import). Then, we prepare the data set so that it is suitable to work with. First we put several formats with the aim of put the correct labels to the categorical variables. Then, we use proc contents for check if all the variables are in the format that we need. We observe that there are several numerical variables that are in alphanumerical format! This is a problem for develop the analysis. Then, what we do is transform these variables to numeric format. In addition, we use the explanations of the variables exposed before in this report to put labels to the variables in the data set.

Once we have our SAS data set ready for work, what we do is carry out a descriptive analysis to check that there are no values of any variable out of range or extreme observations which could be considered outliers. Indicate in the script if everything was ok or not (problem 3 of the statements). We want to asses several things related with categorical variables:

- Check if we obtain 50% of games at home and 50% away. We do it via proc freq, and indeed we obtain what we expect.
- Check if we obtain 50% of games win and 50% lost. We do it via proc freq, and indeed we obtain what we expect. Notice that we have the data duplicated, it is each game duplicated, and then we expect the half of the games win and the other half lost.

miércoles, 27 de diciembre de 2023 12:46:52 **1**

Obs	gameorder	game_id	lg_id	_iscopy	year_id	date_game	seasongame	is_playoffs	team_id
1	1	194611010TRH	NBA	0	1947	11/1/1946	1	0	TRH
2	1	194611010TRH	NBA	1	1947	11/1/1946	1	0	NYK
3	2	194611020CHS	NBA	0	1947	11/2/1946	1	0	CHS
4	2	194611020CHS	NBA	1	1947	11/2/1946	2	0	NYK
5	3	194611020DTF	NBA	0	1947	11/2/1946	1	0	DTF
6	3	194611020DTF	NBA	1	1947	11/2/1946	1	0	WSC
7	4	194611020PRO	NBA	1	1947	11/2/1946	1	0	BOS
8	4	194611020PRO	NBA	0	1947	11/2/1946	1	0	PRO
9	5	194611020STB	NBA	1	1947	11/2/1946	1	0	PIT
10	5	194611020STB	NBA	0	1947	11/2/1946	1	0	STB

Obs	fran_id	pts	elo_i	elo_n	win_equiv	opp_id	opp_fran	opp_pts	opp_elo_i	opp_elo_n
1	Huskies	66	1300	1293.2767	40.29483	NYK	Knicks	68	1300	1306.7233
2	Knicks	68	1300	1306.7233	41.70517	TRH	Huskies	66	1300	1293.2767
3	Stags	63	1300	1309.6521	42.012257	NYK	Knicks	47	1306.7233	1297.0712
4	Knicks	47	1306.7233	1297.0712	40.692783	CHS	Stags	63	1300	1309.6521
5	Falcons	33	1300	1279.6189	38.864048	WSC	Capitols	50	1300	1320.3811
6	Capitols	50	1300	1320.3811	43.135952	DTF	Falcons	33	1300	1279.6189
7	Celtics	53	1300	1294.8458	40.459381	PRO	Steamrollers	59	1300	1305.1542
8	Steamrollers	59	1300	1305.1542	41.540619	BOS	Celtics	53	1300	1294.8458
9	Ironmen	51	1300	1295.3092	40.50798	STB	Bombers	56	1300	1304.6908
10	Bombers	56	1300	1304.6908	41.49202	PIT	Ironmen	51	1300	1295.3092

Obs	game_location	game_result	forecast	notes
1	H	L	0.64006501	
2	A	W	0.35993499	
3	H	W	0.63110125	
4	A	L	0.36889875	
5	H	L	0.64006501	
6	A	W	0.35993499	
7	A	L	0.35993499	
8	H	W	0.64006501	
9	A	L	0.35993499	
10	H	W	0.64006501	

Figure 1: Ten first observations of the dataset exported from SAS.

Procedimiento MEANS

Variable	Etiqueta	N	Media	Desv. est	Mínimo	Máximo
gameorder	Play order of game in NBA history	126314	31579.00	18231.93	1.0000000	63157.00
_iscopy	If game_id has already occured for the opposing team in the same matchup	126314	0.5000000	0.5000020	0	1.0000000
year_id	Season id, named based on year in which the season ended	126314	1988.20	17.5823095	1947.00	2015.00
seasongame	seasongame	126314	43.5337334	25.3751779	1.0000000	108.0000000
is_playoffs	Flag for playoff games	126314	0.0638567	0.2444985	0	1.0000000
pts	Points scored by team	126314	102.7299824	14.8148447	0	186.0000000
opp_pts	Points scored by opponent	126314	102.7299824	14.8148447	0	186.0000000
elo_i2	Team elo entering the game	126314	1495.24	112.1399451	1091.64	1853.10
elo_n2	Team elo following the game	126314	1495.24	112.4616874	1085.77	1853.10
win_equiv2	Equivalent number of wins in a 82-game season for a team of elo_n quality	126314	79.1392752	1277.11	10.1525010	62966.00
opp_elo_i2	Opponent elo entering the game	126314	1495.24	112.1399451	1091.64	1853.10
opp_elo_n2	Opponent elo following the game	126314	1495.24	112.4616874	1085.77	1853.10
forecast2	Elo-based chances of winning for the team in the team_id column, based on elo ratings and game location	126314	0.5000000	0.2152522	0.0204468	0.9795532

Figure 2: Result of Proc means of all the numerical variables in the SAS data set.

Obs	gameorder	game_id	lg_id	_iscopy	year_id	date_game	seasongame	is_playoffs	team_id	fran_id	pts	opp_id	opp_fran	opp_pts	game_location	game_result	elo_i2	elo_n2	win_equiv2	opp_elo_i2	opp_elo_n2	forecast2
26685	13343	197210260VIR	ABA	Copy	1973	10/26/1972	7	Regular season game	DNR	Nuggets	0	VIR	Squires	2	Away	Lost	1460.3387	1457.4464	40.40886	1484.1907	1487.0830	.32894754

Figure 3: Game with 0 points.

- Check if we obtain 50% of games as copy and 50% as not copy. We do it via proc freq, and indeed we obtain what we expect.
- Check if there are more regular season games than playoff games. We do it again via proc freq, and indeed, we obtain what we expect.

Thus, form the descriptive analysis of the categorical variables we obtain that everything is how we expect. Now, using proc means we study the numeric variables. We obtain the table in Figure 2.

We can observe that we have games with 0 points! This is an strange result and we should study this in deep in order to see if it is an outlier or not. Via proc print and where instruction we print the observations in the data set with $pts = 0$, and we obtain the table in Figure 3.

The game with 0 points is a game between the Denver Rockets and the Virginia Squires on October 26, year 1972. If we search on the internet we can find the following New York Times new. It seems that the referee of the game, due to the large number of fouls committed by the Denver Rockets, decided to give the game to the Squires by 0-2.

Now, it is important to discuss what we have to do with this historical match with 0 points. In my opinion, this observation should not be deleted. In this data set our objective is rate the NBA teams using as a criteria in this case the ELO rating system. Thus, we expect that the ELO scores explain the dynamics of the different teams. It is clear that in terms of a game statistics, a game with a result of $0 - 2$ has no sense. Notwithstanding, the fact that the Denver Rockets committed such a high number of fouls may be related to the season this team was having. And this is exactly what we want to explain using the ELO score, how NBA team dynamics are and the valuation that these teams have.

The reader can return to the Figure 2 and observe that we have an obs with min "win

Obs	gameorder	game_id	lg_id	_iscopy	year_id	date_game	seasongame	is_playoffs	team_id	fran_id	pts	opp_id	opp_fran	opp_pts	game_location	game_result	elo_i2	elo_n2	win_equiv2	opp_elo_i2	opp_elo_n2	forecast2
70451	35226	199303020MIL	NBA	Copy	1993	3/2/1993	53	Regular season game	DAL	Mavericks	86	MIL	Bucks	120	Away	Lost	1132.1041	1128.0035	10.94723	1394.6218	1398.7224	.11038452
70465	35233	199303030CHI	NBA	Copy	1993	3/3/1993	54	Regular season game	DAL	Mavericks	97	CHI	Bulls	125	Away	Lost	1128.0035	1127.3315	10.91465	1668.5016	1669.1736	.02443492
70496	35248	199303050DAL	NBA	Not copy	1993	3/5/1993	55	Regular season game	DAL	Mavericks	86	HOU	Rockets	105	Home	Lost	1127.3315	1124.9786	10.80124	1614.0183	1616.3712	.09744443
70509	35255	199303060DAL	NBA	Not copy	1993	3/6/1993	56	Regular season game	DAL	Mavericks	102	PHO	Suns	109	Home	Lost	1124.9786	1124.1189	10.76005	1675.6212	1676.4810	.06951863
70553	35277	199303090SAS	NBA	Copy	1993	3/9/1993	57	Regular season game	DAL	Mavericks	84	SAS	Spurs	119	Away	Lost	1124.1189	1122.8372	10.69888	1585.7332	1587.0149	.03794581
70557	35279	199303100DAL	NBA	Not copy	1993	3/10/1993	58	Regular season game	DAL	Mavericks	96	POR	Trailblazers	124	Home	Lost	1122.8372	1118.5309	10.49556	1551.8792	1556.1854	.13077566
70596	35298	199303120PHO	NBA	Copy	1993	3/12/1993	59	Regular season game	DAL	Mavericks	98	PHO	Suns	116	Away	Lost	1118.5309	1118.1117	10.47595	1685.1344	1685.5536	.02109748
70605	35303	199303130SAC	NBA	Copy	1993	3/13/1993	60	Regular season game	DAL	Mavericks	94	SAC	Kings	122	Away	Lost	1118.1117	1114.2330	10.29593	1365.6101	1369.4888	.11916494
70623	35312	199303150DAL	NBA	Not copy	1993	3/15/1993	61	Regular season game	DAL	Mavericks	96	UTA	Jazz	109	Home	Lost	1114.2330	1111.1008	10.15250	1505.4294	1508.5616	.15759116
70649	35325	199303170DAL	NBA	Not copy	1993	3/17/1993	62	Regular season game	DAL	Mavericks	102	ORL	Magic	96	Home	Win	1111.1008	1127.5350	10.92450	1485.6344	1469.2002	.17074701

Figure 4: Observations with win equiv variable less than 11.

Procedimiento MEANS

Variable	Etiqueta	Mediana	25º Pctl	50º Pctl	75º Pctl	Media	Desv. est	Mínimo	Máximo
elo_i2	Team elo entering the game	1500.95	1417.24	1500.95	1576.06	1495.24	112.1399451	1091.64	1853.10
elo_n2	Team elo following the game	1500.95	1416.99	1500.95	1576.29	1495.24	112.4616874	1085.77	1853.10
opp_elo_i2	Opponent elo entering the game	1500.95	1417.24	1500.95	1576.06	1495.24	112.1399451	1091.64	1853.10
opp_elo_n2	Opponent elo following the game	1500.95	1416.99	1500.95	1576.29	1495.24	112.4616874	1085.77	1853.10
forecast2	Elo-based chances of winning for the team in the team_id column, based on elo ratings and game location	0.5000000	0.3279858	0.5000000	0.6720142	0.5000000	0.2152522	0.0204468	0.9795532

Figure 5: Complete description via proc means of several variables.

equiv" 10.1525010, which is much less than the median (1277.11). We study in deep the observations with very low "win equiv" and we try to explain why we have obtained these values. Using proc print and the where instruction we find the observations with low "win equiv". See the observations obtained in Figure 4. We can see that this low win equiv is due the Dallas mavericks of the 1993, see more info about this team in `https://www.basketball-reference.com/teams/DAL/1994.html` we can see that the Dallas Mavericks was the worst team in the NBA! They had a final record of $13-69$. Then, these observations with low win equiv are not an outlier and are correct. Thus in this case, it is clear that we do not have to remove these observations. Are observations with win equiv results unusual, because this team was unusual bad, but this is not incorrect.

Now, we use proc means and additional instructions for develop a complete description of several variables of our interest. These variables are all the variables related with the ELO score, and the forecast variable. We obtain with the proc means the table in the Figure 5. The results obtained for the variables "elo-i", "elo-n", "opp-elo-i", "opp-elo-n" are what we expect. Remember that the games are duplicated, and then we expect to obtain the same results for the variables "elo-i" and "opp-elo-i", and the same for "elo-n" and "opp-elo-n". The distribution of the "forecast" variable is also as we expect. Notwithstanding, notice that the minimum and the maximum obtained for the variable "forecast" seems very low and high, respectively. In the following explanations we will be analyzing this more in deep.

In order to be able to classify the classes of teams in the NBA according to ELO rating, I thought it would be interesting to carry out a categorization using the "United States Chess Federation ratings". So, via proc format, and using the categories of the United states

Procedimiento FREQ

elo_i_rate	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
Class E	406	0.32	406	0.32
Class D	25658	20.31	26064	20.63
Class C	76588	60.63	102652	81.27
Class B	23613	18.69	126265	99.96
Class A	49	0.04	126314	100.00

Figure 6: Frequencies of the different ELO categories created for the Team ELO entering the game variable.

chess rating (see the categories proposed for United States Chess Federation in `https://en.wikipedia.org/wiki/Elo_rating_system`), I have made a categorization of the variables related to the ELO. In this way, I have created new variables that tell us, for each observation (i.e. for each match), what is the rating of the team and the opponent team according to the categories proposed by the United States Chess Federation. The frequencies of classes obtained are exposed in Figure 6. We can see that in most of games the Team ELO entering the game is of Class C (notice that Class A is better than Class B, which is better than Class C, etc).

In addition I have developed a categorization about the win probabilities. And, in addition, I have also added a categorization on the years of the games, so that they can be divided by decades. In order to asses if the variable forecast is working well, it is, if the probability to win is predicting well the game results, we used a proc freq for develop a bivariate table. We can see the results in Figure 7. We can observe that, for low probabilities (< 0.25), the 82.56% of the games are lost. For high probabilities (> 0.75) the 82.56% of the games are win. For probabilities between 0.25 and 0.5, the 62% of games are lost. And for probabilities $0.5 - 0.75$, the 62% of the games are win. Notice that the symmetry obtained in the results are due the data is doubled. But, we can observe that it seems that the Elo-based chances of winning for the team in the team-id column, based on elo ratings and game location are explaining well the probability of win or loose a game.

Now, we cross the continuous variable forecast with the categorical variable game result for

Procedimiento FREQ

Frecuencia Porcentaje Pct fila Pct col	Tabla de win_prob por game_result		
	game_result(Win or loss for team in the team_id column)		
win_prob	Lost	Win	Total
Probabilities of win<25%	15435 12.22 82.56 24.44	3260 2.58 17.44 5.16	18695 14.80
Probabilities of win 25%-50%	27572 21.83 62.01 43.66	16890 13.37 37.99 26.74	44462 35.20
Probabilities of win 50%-75%	16890 13.37 37.99 26.74	27572 21.83 62.01 43.66	44462 35.20
Probabilities of win >75%	3260 2.58 17.44 5.16	15435 12.22 82.56 24.44	18695 14.80
Total	63157 50.00	63157 50.00	126314 100.00

Figure 7: Frequencies of the created variable win prob (categorization of the forecast variable) with respect the game result variable.

Procedimiento MEANS

Analysis Variable : forecast2 Elo-based chances of winning for the team in the team_id column, based on elo ratings and game location									
Win or loss for team in the team_id column	N Obs	Mediana	25º Pctl	50º Pctl	75º Pctl	Media	Desv. est	Mínimo	Máximo
Lost	63157	0.3859397	0.2531275	0.3859397	0.5501131	0.4074622	0.1943464	0.0204468	0.9729074
Win	63157	0.6140603	0.4498869	0.6140603	0.7468725	0.5925378	0.1943464	0.0270926	0.9795532

Figure 8: Descriptive statistics for the continuous variable forecast with respect the categorical variable game result.

study the same (if the forecast variable is predicting well the game results). We obtain the table in Figure 8. We see that the general trend is what we expect, and that indicates that the forecast variable is working well. Notwithstanding, observe that the maximum for the lost category is 0.9729074, and the minimum for the win category is 0.0270926. These results seems strange! Let's study these observations in deep. Using proc print and the where instruction, as we did before, we print the observations such that forecast < 0.03 and game-result $= W$. We can observe that the two cases obtained of very low probability of win, and then win, are both in the same season (1993), and both for the Dallas Mavericks playing away. See the results of this team in `https://www.basketball-reference.com/teams/DAL/1994.html` we can see that the Dallas Mavericks was the worst team in the NBA! But this is sport, and the bad teams also win. We proceed similar for those observations such that forecast > 0.96 and game-result $= L$. We obtain three observations. Two of these three observations are explained again for the 1993 Dallas Mavericks, because this two observations are two good teams (Houston Rockets and Seattle Thunder, both class B teams) loosing with the worst team in 1993 (Dallas Mavericks, Class E team). The remaining observation is the game Chicago Bulls VS Orlando Magic in 21/3/1992. In 1992, the Chicago Bulls (Class B), lead by the greatest of all time, Micheal Jordan, won the NBA and was the best team in the league. On the other hand, Orlando Magic (Class D) was one of the worst teams in the league. But, these exceptions are normal. How boring basketball would be if the best teams always won!

We have also done a boxplot with the aim of asses the behaviour of the continuous variable forecast with respect the categorical variable game result. See the boxplot obtained in the Figure 9. The trend that we observe is what we expect: the win category has higher values of the forecast variable than the lost probability. We add to the analysis the game location category, and we do a boxplot again. This boxplot is exposed in Figure 10. We observe that Home category has higher forecast values than Away category. Which indicates a very reasonable hypothesis: it is more probable to win a game at home that away.

Now, using the ELO categorizations done in the previous section, we analyze the team ELO entering the game vs the forecast variable. We want to study if the better the ELO class in the chess ranking, the higher the probability of victory. We use proc means and proc sgplot for develop a table with summary statistics and a boxplot. The results obtained are in Figure

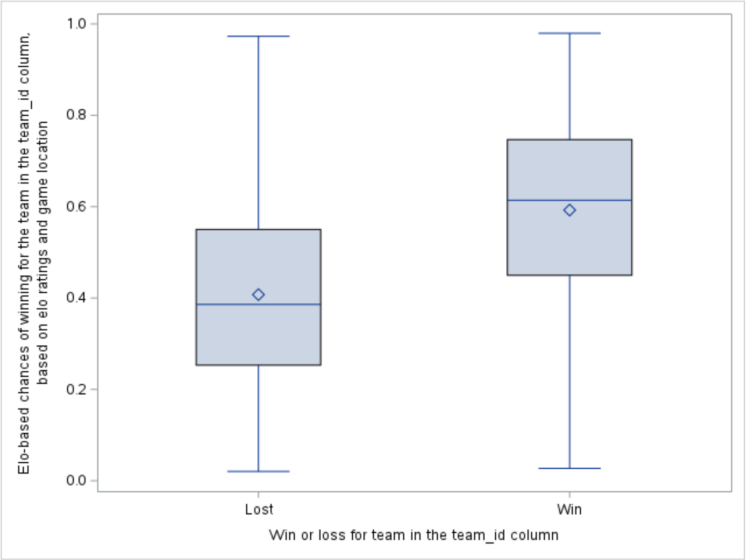


Figure 9: Boxplot developed with proc sgplot, the continuous variable is forecast and the categorical variable is game result.

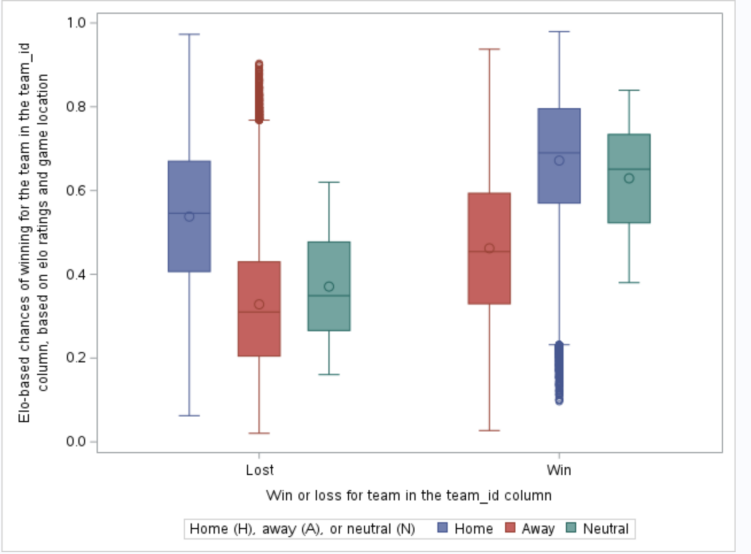


Figure 10: Boxplot developed with proc sgplot, the continuous variable is forecast and the categorical variables are game result and game location.

Procedimiento MEANS

Analysis Variable : forecast2 Elo-based chances of winning for the team in the team_id column, based on elo ratings and game location

elo_i_rate	N Obs	Mediana	25º Pctl	50º Pctl	75º Pctl	Media	Desv. est	Mínimo	Máximo
Class E	406	0.1833074	0.1067716	0.1833074	0.3307473	0.2264530	0.1523389	0.0204468	0.7154450
Class D	25658	0.3093987	0.1871417	0.3093987	0.4574497	0.3337436	0.1791686	0.0242707	0.9005104
Class C	76588	0.5074090	0.3492649	0.5074090	0.6608626	0.5063529	0.1949435	0.0548183	0.9620542
Class B	23613	0.6890909	0.5387844	0.6890909	0.8054551	0.6642443	0.1752798	0.1640383	0.9795532
Class A	49	0.7492212	0.6128564	0.7492212	0.8546149	0.7451421	0.1380375	0.4494200	0.9757294

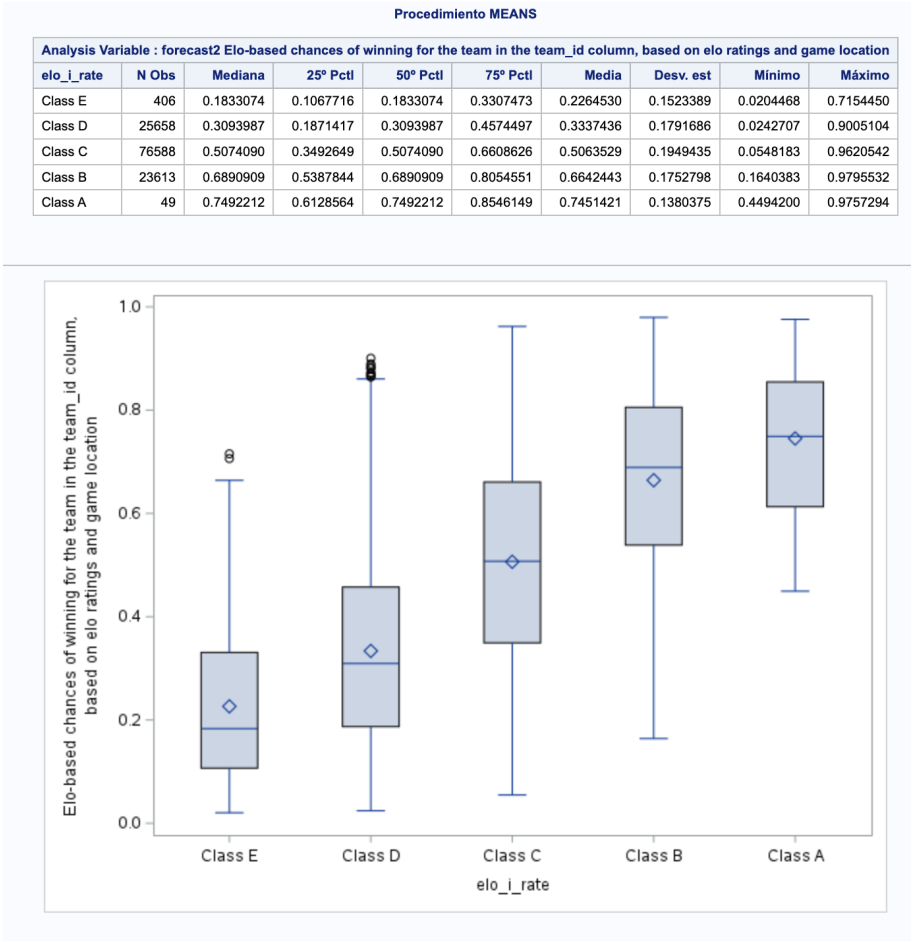


Figure 11: Summary statistics and boxplot of the continuous variable forecast with respect the categorical variable ELO rating.

11, and we obtain the expected result, the better the class the better the win probability.

Now, we want to check if there are more wins in home games than in away games. It is a common hypothesis in sports that is easier to win a game playing it at home that away. Using the huge data set that we have we will try to asses this. Using proc freq we build a table with the bivariate frequencies of the variables game result and game location. See the table in Figure 12. The results obtained support the hypothesis. In the Home the 62.25% of games are won, and the 37.75% are lost.

Now, we want to study the evolution in scoring in the different decades of the league. For this we use the categorization of the years by decades, and look at the points of the teams. We have obtained the boxplot in Figure 13. We can see that the evolution of the league in terms of points is quite stable. Keep in mind that we only have data up to 2015, and that

Procedimiento FREQ

Frecuencia
Porcentaje
Pct fila
Pct col

Tabla de game_location por game_result

game_location(Home (H), away (A), or neutral (N))	game_result(Win or loss for team in the team_id column) Lost	Win	Total
Away	39305 31.12 62.25 62.23	23833 18.87 37.75 37.74	63138 49.98
Home	23833 18.87 37.75 37.74	39305 31.12 62.25 62.23	63138 49.98
Neutral	19 0.02 50.00 0.03	19 0.02 50.00 0.03	38 0.03
Total	63157 50.00	63157 50.00	126314 100.00

Figure 12: Table of bivariate frequencies of the categorical variables game result and game location.

it was more or less from those years onwards that the league had a big paradigm shift with the three-point shot.

Finally, our objective is to study how good various nba teams have been over the course of history. With this purpose, we have created a macro that given the name of one team returns a frequency table of the game results per decade, and another frequency table with its ELO class per decade. In addition, we plot an histogram of the points scored by the team. This several operations will be useful for asses how good was this team in the whole NBA history, and for see if the ELO ratings per decade reflects good the team quality.

We use the macro for study how good have been the Boston Celtics, an historic franchise of the NBA. The tables obtained for the Boston Celtics are in Figure 14. We can see that Celtics was a nice team in 1960s, they had a 70.45% of wins and in the 79% of the games (before play the game) they were a Class A team in the ELO rating. The Celtics of the 1970s was also nice. In most of the games during the league history the Celtics was Class B or Class C team, which indicates the good quality of this team.

We analyze the Memphis Grizzlies (the Grizzlies franchise was created in the 1990s, first the team was located in Vancouver, currently the team is located in Memphis). The tables obtained are exposed in Figure . Grizzlies in 2010s were good, they had 59.55% of wins. In more than 95% of games in the 2010s the Grizzlies was a Class C or Class B team. Marc Gasol, an historic player of this franchise, probably helped in this good results. Notwithstanding, in the first decade of the franchise they had a 81.08% of lost games. They was a Class D, Class E team.

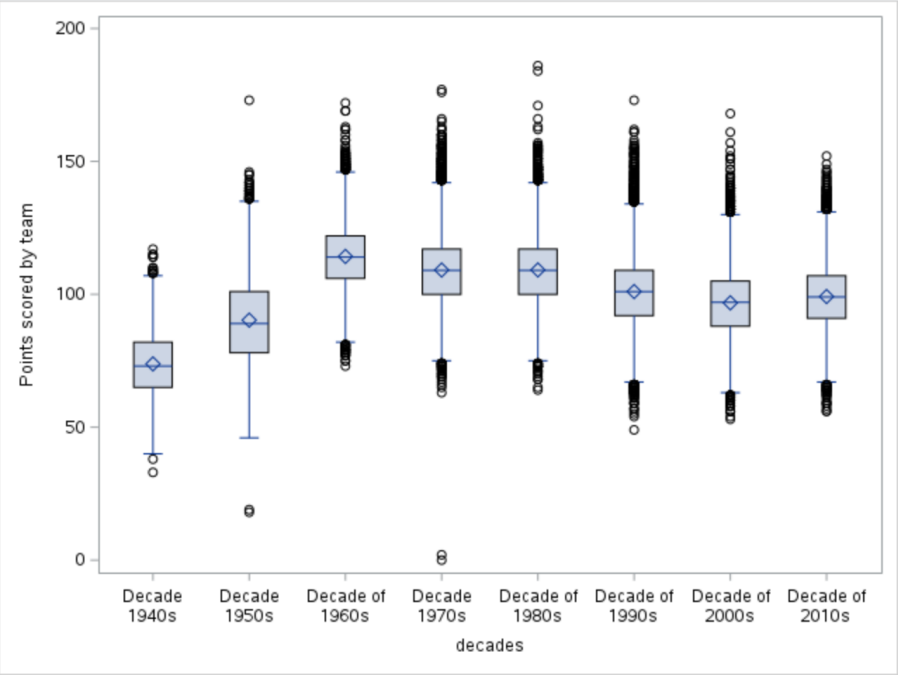
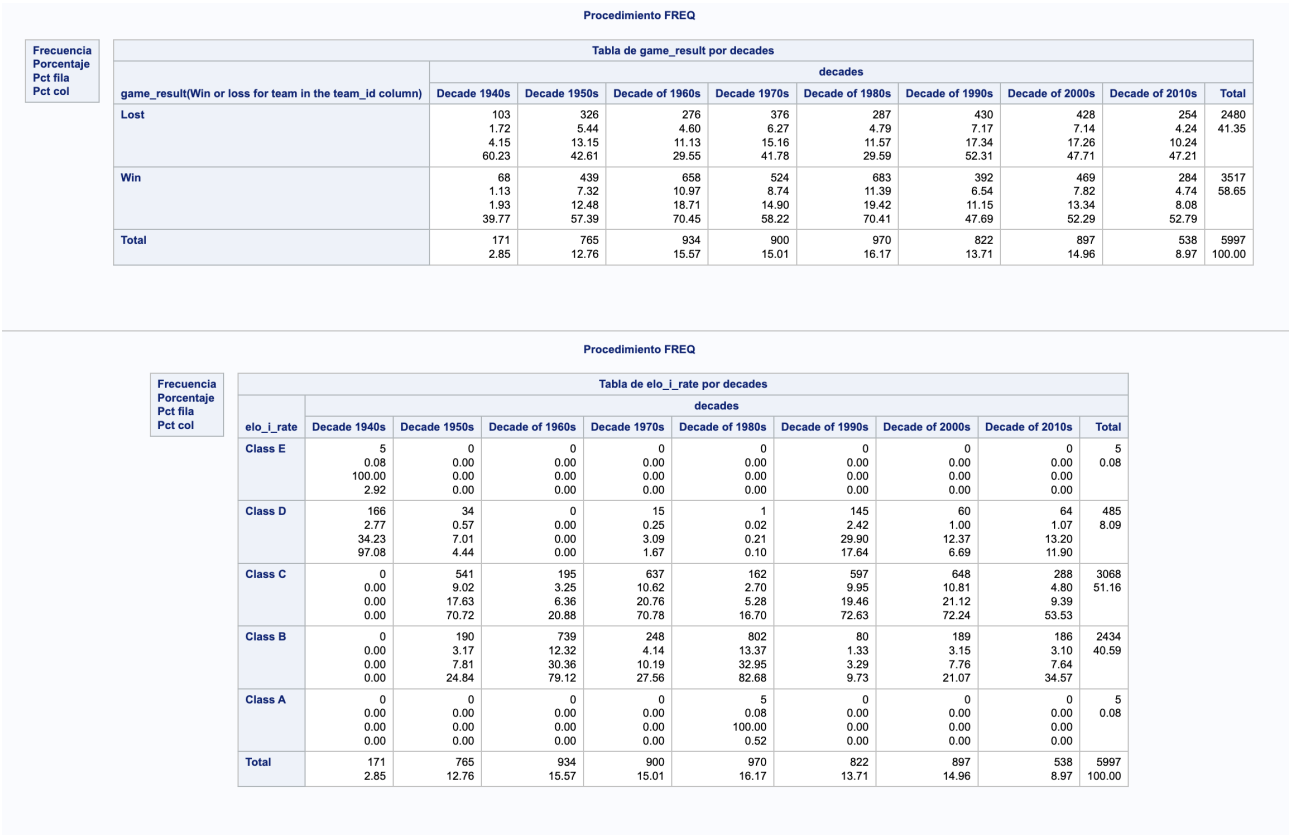


Figure 13: Boxplot of the continuous variable points with respect the categorical variable decades.



Procedimiento FREQ

Frecuencia
Porcentaje
Pct fila
Pct col

Tabla de game_result por decades

game_result(Win or loss for team in the team_id column)	decades								
	Decade 1940s	Decade 1950s	Decade of 1960s	Decade 1970s	Decade of 1980s	Decade of 1990s	Decade of 2000s	Decade of 2010s	Total
Lost	103 1.72 4.15 60.23	326 5.44 13.15 42.61	276 4.60 11.13 29.55	376 6.27 15.16 41.78	287 4.79 11.57 29.59	430 7.17 17.34 52.31	428 7.14 17.26 47.71	254 4.24 10.24 47.21	2480 41.35
Win	68 1.13 1.93 39.77	439 7.32 12.48 57.39	658 10.97 18.71 70.45	524 8.74 14.90 58.22	683 11.39 19.42 70.41	392 6.54 11.15 47.69	469 7.82 13.34 52.29	284 4.74 8.08 52.79	3517 58.65
Total	171 2.85	765 12.76	934 15.57	900 15.01	970 16.17	822 13.71	897 14.96	538 8.97	5997 100.00

Procedimiento FREQ

Frecuencia
Porcentaje
Pct fila
Pct col

Tabla de elo_i_rate por decades

elo_i_rate	decades								
	Decade 1940s	Decade 1950s	Decade of 1960s	Decade 1970s	Decade of 1980s	Decade of 1990s	Decade of 2000s	Decade of 2010s	Total
Class E	5 0.08 100.00 2.92	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	5 0.08
Class D	166 2.77 34.23 97.08	34 0.57 7.01 4.44	0 0.00 0.00 0.00	15 0.25 3.09 1.67	1 0.02 0.21 0.10	145 2.42 29.90 17.64	60 1.00 12.37 6.69	64 1.07 13.20 11.90	485 8.09
Class C	0 0.00 0.00 0.00	541 9.02 17.63 70.72	195 3.25 6.36 20.88	637 10.62 20.76 70.78	162 2.70 5.28 16.70	597 9.95 19.46 72.63	648 10.81 21.12 72.24	288 4.80 9.39 53.53	3068 51.16
Class B	0 0.00 0.00 0.00	190 3.17 7.81 24.84	739 12.32 30.36 79.12	248 4.14 10.19 27.56	802 13.37 32.95 82.68	80 1.33 3.29 9.73	189 3.15 7.76 21.07	186 3.10 7.64 34.57	2434 40.59
Class A	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	5 0.08 100.00 0.52	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	5 0.08
Total	171 2.85	765 12.76	934 15.57	900 15.01	970 16.17	822 13.71	897 14.96	538 8.97	5997 100.00

Figure 14: Results of the macro for the Boston Celtics.

Procedimiento FREQ

Frecuencia
Porcentaje
Pct fila
Pct col

Tabla de game_result por decades

game_result(Win or loss for team in the team_id column)	decades			
	Decade of 1990s	Decade of 2000s	Decade of 2010s	Total
Lost	240 14.48 24.54 81.08	524 31.62 53.58 62.98	214 12.91 21.88 40.45	978 59.02
Win	56 3.38 8.25 18.92	308 18.59 45.36 37.02	315 19.01 46.39 59.55	679 40.98
Total	296 17.86	832 50.21	529 31.93	1657 100.00

Procedimiento FREQ

Frecuencia
Porcentaje
Pct fila
Pct col

Tabla de elo_i_rate por decades

elo_i_rate	decades			
	Decade of 1990s	Decade of 2000s	Decade of 2010s	Total
Class E	25 1.51 100.00 8.45	0 0.00 0.00 0.00	0 0.00 0.00 0.00	25 1.51
Class D	271 16.35 36.09 91.55	464 28.00 61.78 55.77	16 0.97 2.13 3.02	751 45.32
Class C	0 0.00 0.00 0.00	334 20.16 52.68 40.14	300 18.11 47.32 56.71	634 38.26
Class B	0 0.00 0.00 0.00	34 2.05 13.77 4.09	213 12.85 86.23 40.26	247 14.91
Total	296 17.86	832 50.21	529 31.93	1657 100.00

Figure 15: Results of the macro for the Grizzlies.

4 Final observations

In this final section I would like to add several final comments and observations. All the tables in this report are screenshots from the results of the SAS procedures. Although I know how to export the tables (see in Figure 1 and in the SAS script) in pdf format, using ODS pdf, and other ODS options, I preferred to take screenshots because it was simpler, cleaner in the sense that I didn't have to save many pdf documents.

As I final comment I would like to mention that I enjoyed this work, and that I found it very interesting to explore a data set on a topic of my interest.