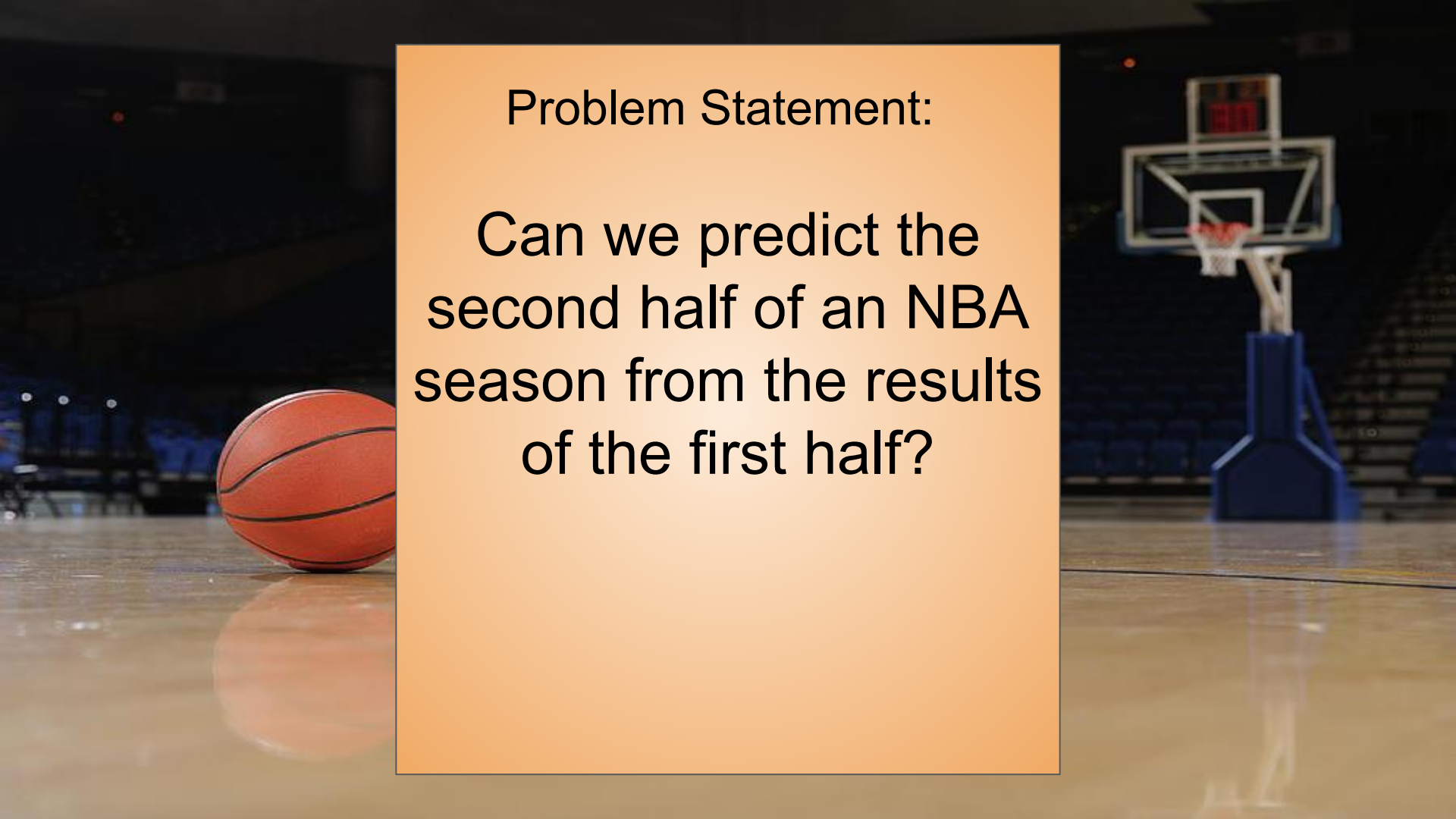
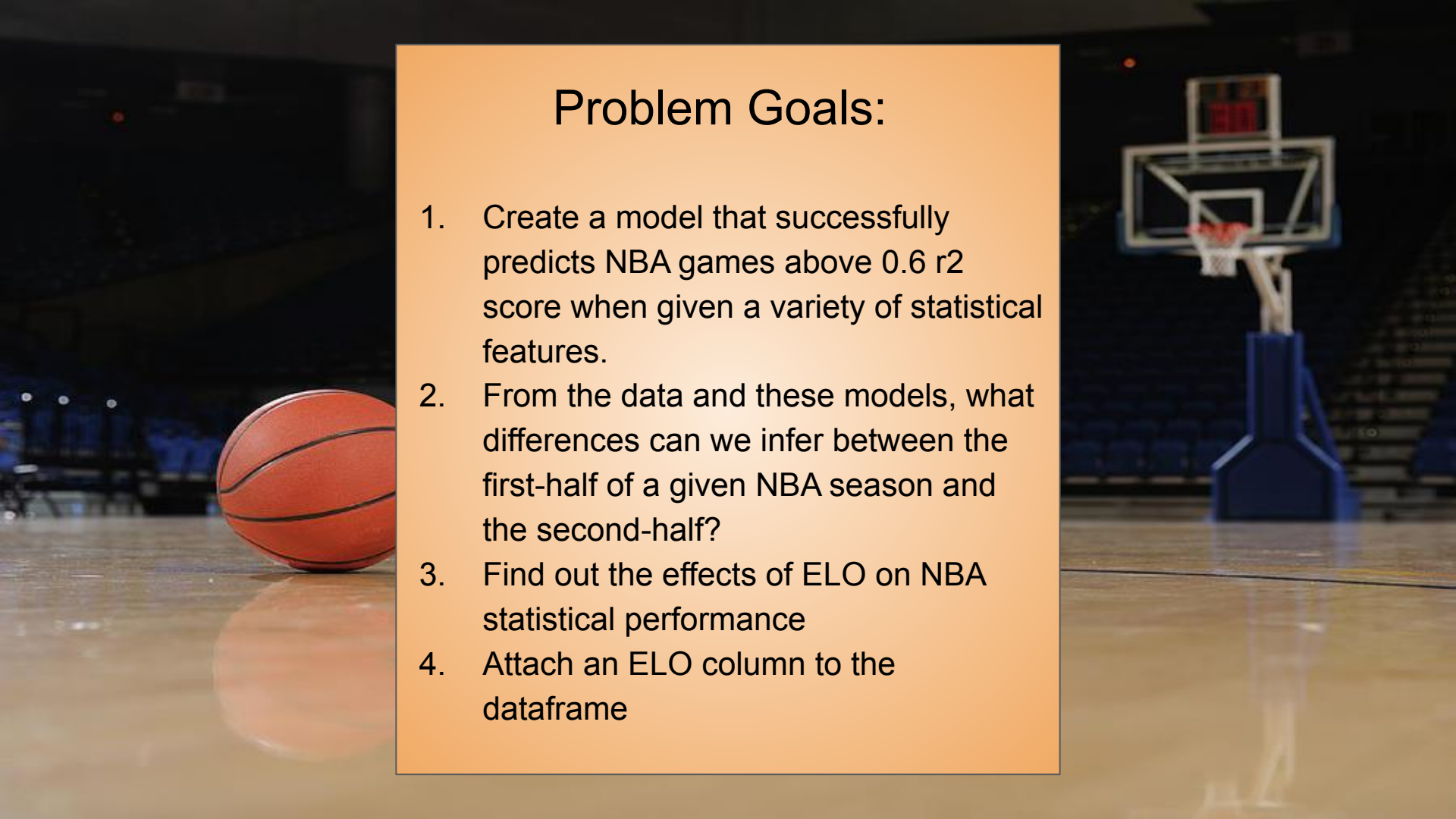
Predicting NBA Games

By: Adam Klesc



Problem Statement:

Can we predict the second half of an NBA season from the results of the first half?



Problem Goals:

1. Create a model that successfully predicts NBA games above 0.6 r2 score when given a variety of statistical features.
2. From the data and these models, what differences can we infer between the first-half of a given NBA season and the second-half?
3. Find out the effects of ELO on NBA statistical performance
4. Attach an ELO column to the dataframe

Who am I?

I am a data science student studying at General Assembly, I've been an avid fan of both sports and sports analytics my entire life. Specifically, the league to my left

Implementing the Data Science Process with the NBA

1) Gather Data
2) Exploratory Data Analysis (EDA)
3) Data Cleaning and Processing
4) Building Machine Learning models
5) Evaluating our Models
6) Making Conclusions about our Findings

Gathering Schedule Information

October Schedule

Share & Export ▾

Date	Start (ET)	Visitor/Neutral	PTS	Home/Neutral	PTS			Attend.	Notes
Tue, Oct 19, 2021	7:30p	Brooklyn Nets	104	Milwaukee Bucks	127	Box Score		17,341	
Tue, Oct 19, 2021	10:00p	Golden State Warriors	121	Los Angeles Lakers	114	Box Score		18,997	
Wed, Oct 20, 2021	7:00p	Indiana Pacers	122	Charlotte Hornets	123	Box Score		15,521	
Wed, Oct 20, 2021	7:00p	Chicago Bulls	94	Detroit Pistons	88	Box Score		20,088	
Wed, Oct 20, 2021	7:30p	Boston Celtics	134	New York Knicks	138	Box Score	2OT	19,812	
Wed, Oct 20, 2021	7:30p	Washington Wizards	98	Toronto Raptors	83	Box Score		19,800	
Wed, Oct 20, 2021	8:00p	Cleveland Cavaliers	121	Memphis Grizzlies	132	Box Score		15,975	
Wed, Oct 20, 2021	8:00p	Houston Rockets	106	Minnesota Timberwolves	124	Box Score		16,079	
Wed, Oct 20, 2021	8:00p	Philadelphia 76ers	117	New Orleans Pelicans	97	Box Score		12,845	
Wed, Oct 20, 2021	8:30p	Orlando Magic	97	San Antonio Spurs	123	Box Score		16,697	
Wed, Oct 20, 2021	9:00p	Oklahoma City Thunder	86	Utah Jazz	107	Box Score		18,306	
Wed, Oct 20, 2021	10:00p	Sacramento Kings	124	Portland Trail Blazers	121	Box Score		17,467	
Wed, Oct 20, 2021	10:00p	Denver Nuggets	110	Phoenix Suns	98	Box Score		16,074	
Thu, Oct 21, 2021	7:30p	Dallas Mavericks	87	Atlanta Hawks	113	Box Score		17,162	
Thu, Oct 21, 2021	8:00p	Milwaukee Bucks	95	Miami Heat	137	Box Score		19,600	
Thu, Oct 21, 2021	10:00p	Los Angeles Clippers	113	Golden State Warriors	115	Box Score		18,064	

Gathering Individual Box Scores

Denver Nuggets (20-19)

Share & Export ▼ Glossary

	Basic Box Score Stats																		
Starters	**MP**	**FG**	**FGA**	**FG%**	**3P**	**3PA**	**3P%**	**FT**	**FTA**	**FT%**	**ORB**	**DRB**	**TRB**	**AST**	**STL**	**BLK**	**TOV**	**PF**	**PTS**	**+/-**
Nikola Jokić	36:38	9	14	.643	1	3	.333	2	2	1.000	5	8	13	8	1	0	5	2	21	+6
Aaron Gordon	34:39	11	16	.688	1	4	.250	7	7	1.000	2	10	12	1	0	0	4	4	30	+6
Monte Morris	34:11	5	11	.455	1	6	.167	0	0		0	3	3	5	1	0	3	1	11	+6
Austin Rivers	29:43	0	4	.000	0	4	.000	4	6	.667	1	4	5	2	0	0	2	4	4	+13
JaMychal Green	17:36	0	2	.000	0	1	.000	0	0		0	3	3	1	1	1	0	1	0	+15
Reserves	**MP**	**FG**	**FGA**	**FG%**	**3P**	**3PA**	**3P%**	**FT**	**FTA**	**FT%**	**ORB**	**DRB**	**TRB**	**AST**	**STL**	**BLK**	**TOV**	**PF**	**PTS**	**+/-**
Facundo Campazzo	28:09	0	6	.000	0	5	.000	2	2	1.000	0	5	5	5	0	1	1	2	2	-19
Jeff Green	21:11	5	16	.313	0	4	.000	2	3	.667	3	1	4	0	0	0	3	0	12	-13
Davon Reed	13:43	0	2	.000	0	2	.000	0	0		1	6	7	1	0	0	0	1	0	-8
Zeke Nnaji	11:22	2	4	.500	1	1	1.000	0	0		1	0	1	0	1	0	0	1	5	-8
Bones Hyland	8:28	0	5	.000	0	3	.000	0	0		0	3	3	1	0	0	0	2	0	0
James Ennis	4:20	0	0		0	0		0	0		0	0	0	0	0	0	1	1	0	-8
Team Totals	**240**	**32**	**80**	**.400**	**4**	**33**	**.121**	**17**	**20**	**.850**	**13**	**43**	**56**	**24**	**4**	**2**	**19**	**19**	**85**	

Dataframe After Cleaning

	TEAM_HOME	TEAM_AWAY	FG_HOME	FG_AWAY	FGA_HOME	FGA_AWAY	FG_PCT_HOME	FG_PCT_AWAY	FG(3)_HOME	FG(3)_AWAY	...
0	IND	ORL	34	36	71	93	0.479	0.387	7	9	...
1	MIA	CHI	37	35	72	83	0.514	0.422	11	7	...
2	LAL	LAC	42	41	93	83	0.452	0.494	14	8	...
3	CLE	BRK	35	33	84	82	0.417	0.402	5	9	...
4	TOR	BOS	38	32	86	66	0.442	0.485	5	3	...
...	...	...	...	...	...	...	...	...	...	...	...
7375	PHI	CHI	52	45	93	95	0.559	0.474	12	9	...
7376	SAS	DAL	41	37	88	91	0.466	0.407	8	11	...
7377	DEN	MIN	39	39	87	91	0.448	0.429	10	13	...
7378	LAC	UTA	54	47	106	106	0.509	0.443	12	14	...
7379	POR	SAC	53	50	91	96	0.582	0.521	14	18	...

7380 rows × 98 columns

Dataframe to Model

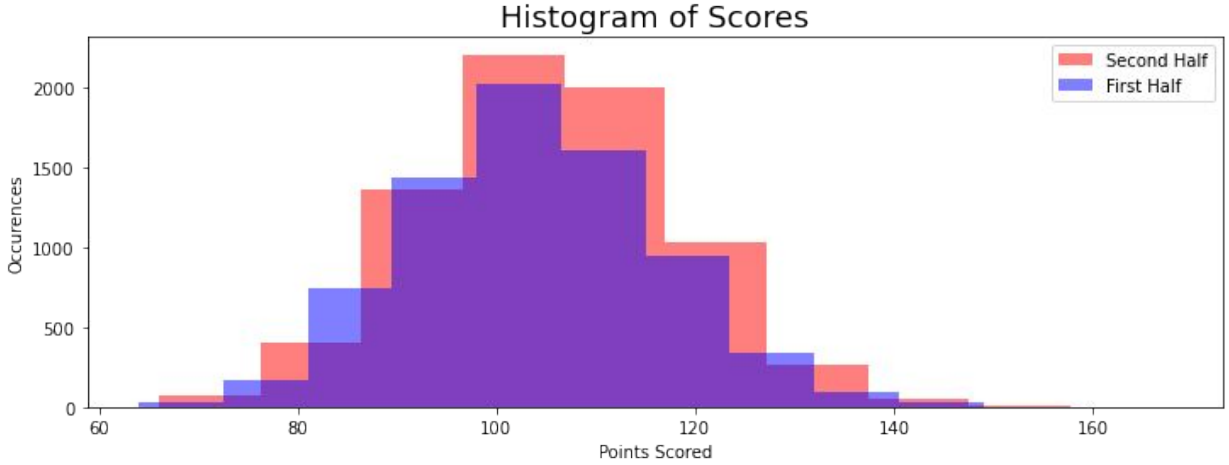
	FG_LAST_10_GAMES_HOME	FGA_LAST_10_GAMES_HOME	FG_PCT_LAST_10_GAMES_HOME	FG(3)_LAST_10_GAMES_HOME	FGA(3)_LAST_10_GAMES_H
0	36.0	76.0	0.4740	9.0	
1	43.0	83.0	0.5180	4.0	
2	33.0	81.0	0.4070	8.0	
3	38.5	91.0	0.4225	11.0	
4	32.0	70.5	0.4540	8.5	
...	...	...	...	...	
7181	41.3	91.0	0.4552	10.2	
7182	42.9	89.7	0.4794	10.2	
7183	39.3	89.2	0.4425	9.0	
7184	41.6	84.6	0.4923	11.4	
7185	44.0	93.1	0.4762	10.4	

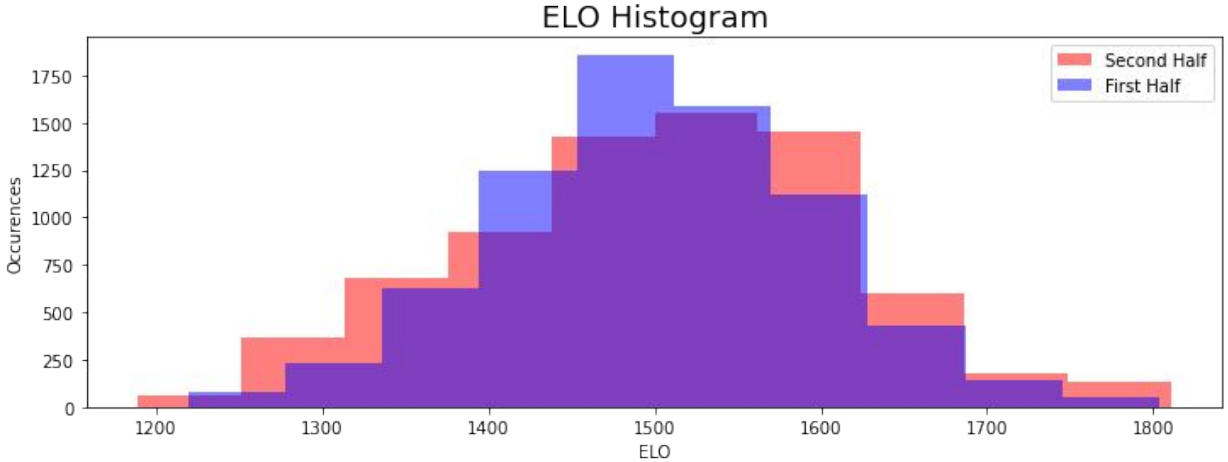
7186 rows × 108 columns

Added Features

- ELO Rating (Home and Away)
- Win Record at Home
- Win Record Away
- Win-Rate
- Game Number
- ELO Home Odds

First and Second Half Distributions

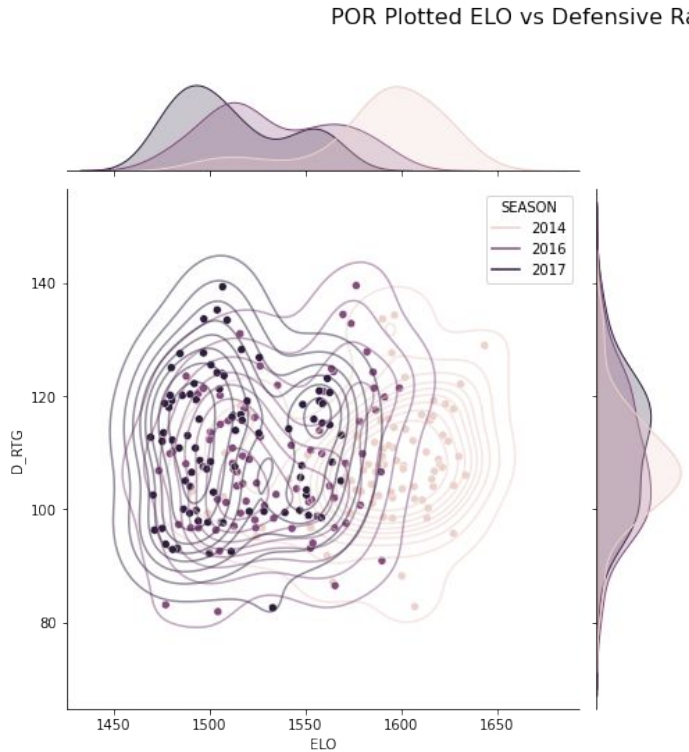


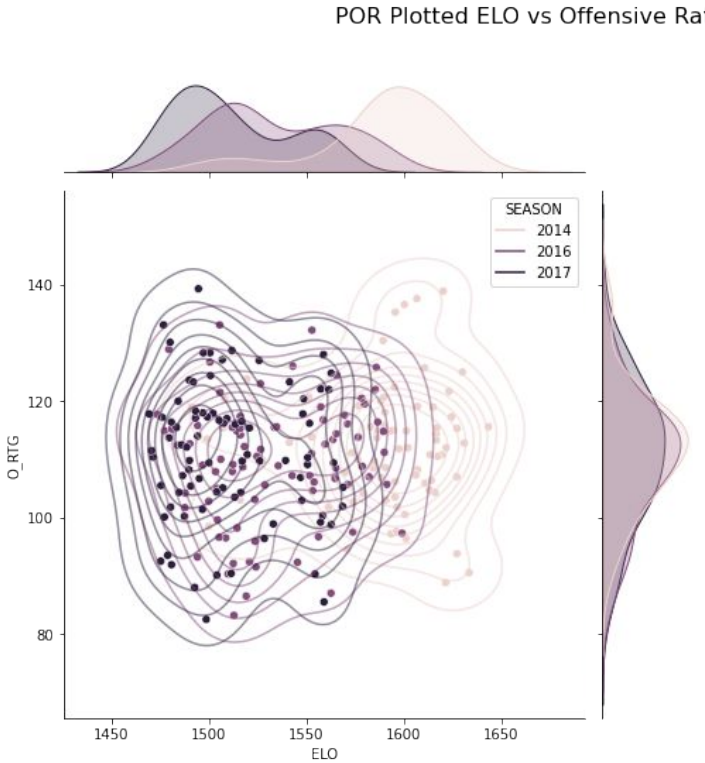


Takeaways:

- ELO skews to the second half of a given season
- Offensive and defensive performance do not vary depending on half

ELO Correlations





Testing Models

MODEL	FEATURE-SET	SCORE (r2)
Logistic Regression	PCA (size: 35)	0.5137646386985866
K-Nearest Neighbors	Kitchen Sink (Unscaled)	0.517689742411017
Decision Tree	Kitchen Sink	0.5093243709787587
Bagging Classifier	Kitchen Sink (Unscaled)	0.50708774502779
Random Forest	Kitchen Sink (Unscaled)	0.5165558741219263
Adaboost Model	Kitchen Sink (Unscaled)	0.5065554212427624
Gradient Boost	Kitchen Sink	0.5076400497794851

Feature Sets: Kitchen Sink (102), Kitchen Sink (unscaled), PCA (35), PCA (30), PCA(40), Variance (40)

Results so far…

- It is unlikely I will be able to produce a good model that accurately predicts the second half of the season when only given the first half of the season as training data.
- The distributions of second-half and first-half stats inform us that the distributions are fairly similar outside of ELO, which is a stat that carries over from the first half's games played.
- ELO is correlated with points scored, defensive performance, and offensive performance

What's Next …

- Current model is lackluster, could tune hyperparameters or add more features.
- Potentially spin off a new project from the existing data and features
- Further EDA that utilizes more features and explores specific teams and their performance
- Feature engineering and EDA using player data as well

The End.