New Basketball Analysis

Bruno Vitorino Caio Pereira Pedro Lins

December 10, 2017

Abstract

A *National Basketball League* (NBA), liga de basquete profissional americana, é uma das ligas esportivas mais populares no mundo inteiro, com times e jogadores conhecidos nos mais diversos lugares. Tendo em vista a popularidade da liga, esse documento apresenta uma análise dos dados de times e jogadores da liga com o objetivo de tentar obter informações interessantes a respeito da mesma.

1 Introdução

O basquete é um dos esportes mais populares do mundo, tendo milhões de fãs espalhados pelo planeta. No centro de toda essa popularidade está a *National Basketball League*, também conhecida como NBA, principal liga do esporte nos Estados Unidos e conhecida por ter os melhores jogadores da modalidade. Observando a importância da NBA para o basquete, fizemos a coleta e análise dos dados sobre todos os times e jogadores nas quatro últimas temporadas regulares (três completas e a atual, que ainda não está finalizada).

Em seguida vamos detalhar o processo de cada uma das etapas do projeto: iniciando com a descrição dos dados e de como foi feita a coleta e o pré-processamento dos mesmos; as abordagens que utilizamos para fazer a análise: análises exploratórias, *clustering* dos times, *clustering* dos jogadores e, por fim, a predição das posições dos times na atual temporada; por fim faremos uma análise de cada uma dessas abordagens e apresentaremos a conclusão que podemos tirar a partir dos resultados obtidos.

2 Etapas

Nessa seção iremos descrever cada uma das etapas do projeto, da coleta de dados até as análises que podem ser feitas a partir dos mesmos. É importante informar que todas as etapas foram feitas usando a linguagem *Python* com o auxílio da ferramenta Jupyter Notebooks.

2.1 Dados

A coleta dos dados foi feita a partir do site *https://www.basketball-reference.com/*, que contém diversas estatísticas sobre todos os times e jogadores das principais ligas de basquete do mundo, com a ajuda da ferramenta BeautifulSoup, que auxilia na leitura e manipulação do HTML das páginas, facilitando assim a obtenção dos dados desejados. Decidimos coletar os dados das últimas quatro temporadas pois antes disso a liga possuía uma configuração diferente de times (franquias que se mudam de lugar e/ou nome), e nessas quatro a configuração se manteve a mesma, facilitando assim a obtenção dos dados.

2.1.1 Dados dos times

Para os dados dos times escolhemos pegar os dados que consistem das estatísticas médias dos times por jogo durante a temporada, consistindo de estatísticas como: cestas feitas, rebotes, *turnovers* (quantidade de vezes que o time perdeu a bola quando tinha a posse), faltas cometidas, entre outras. Isso foi organizado para cada time contendo esses dados para cada uma das quatro temporadas, como podemos ver na figura 1, que exemplifica a tabela criada para a equipe Boston Celtics.

Isso foi feito para cada uma das equipes e posteriormente criamos uma tabela para cada temporada, contendo os dados de cada time naquela temporada, o que pode ser visto na figura 2, que representa a tabela para a temporada 2016-2017.

	Team	Season	MP	FG	FGA	FG%	3P	3PA	3P%	2P	...	FT%	ORD	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
0	Boston Celtics	2014-2015	242.4	38.9	87.9	0.443	8.0	24.6	0.327	30.9	...	0.754	11.1	32.7	43.8	24.5	8.2	3.6	13.8	21.2	101.4
1	Boston Celtics	2015-2016	241.2	39.2	89.2	0.439	8.7	26.1	0.335	30.5	...	0.788	11.6	33.3	44.9	24.2	9.2	4.2	13.7	21.9	105.7
2	Boston Celtics	2016-2017	240.9	38.6	85.1	0.454	12.0	33.4	0.359	26.6	...	0.807	9.1	32.9	42.0	25.2	7.5	4.1	13.3	20.6	108.0
3	Boston Celtics	2017-2018	241.0	37.6	83.7	0.449	11.4	30.7	0.373	26.2	...	0.770	9.3	36.0	45.3	22.8	7.9	4.6	14.2	19.6	104.5

Figure 1: Exemplo de tabela contendo informações do time.

	Team	Season	MP	FG	FGA	FG%	3P	3PA	3P%	2P	...	AST	STL	BLK	TOV	PF	PTS	WIN	LOSS	POS	CONF
0	Boston Celtics	2016-2017	240.9	38.6	85.1	0.454	12.0	33.4	0.359	26.6	...	25.2	7.5	4.1	13.3	20.6	108.0	53	29	1	E
1	Cleveland Cavaliers	2016-2017	242.4	39.9	84.9	0.470	13.0	33.9	0.384	26.9	...	22.7	6.6	4.0	13.7	18.1	110.3	51	31	2	E
2	Toronto Raptors	2016-2017	241.2	39.2	84.4	0.464	8.8	24.3	0.363	30.3	...	18.5	8.3	4.9	12.7	20.8	106.9	51	31	3	E
3	Detroit Pistons	2016-2017	241.5	39.9	88.8	0.449	7.7	23.4	0.330	32.2	...	21.1	7.0	3.8	11.9	17.9	101.3	37	45	10	E
4	Philadelphia 76ers	2016-2017	241.8	37.7	85.3	0.442	10.1	29.8	0.340	27.5	...	23.8	8.4	5.1	16.7	21.9	102.4	28	54	14	E
5	Milwaukee Bucks	2016-2017	241.2	38.8	81.9	0.474	8.8	23.7	0.370	30.0	...	24.2	8.1	5.3	14.0	20.2	103.6	42	40	6	E
6	Indiana Pacers	2016-2017	242.1	39.3	84.5	0.465	8.6	23.0	0.376	30.6	...	22.5	8.2	5.0	13.8	19.5	105.1	42	40	7	E
7	Washington Wizards	2016-2017	242.1	41.3	87.0	0.475	9.2	24.8	0.372	32.1	...	23.9	8.5	4.1	14.2	21.3	109.2	49	33	4	E
8	New York Knicks	2016-2017	242.1	39.6	88.5	0.447	8.6	24.7	0.348	31.0	...	21.8	7.1	5.5	13.9	20.3	104.3	31	51	12	E
9	Miami Heat	2016-2017	241.2	39.0	85.8	0.455	9.9	27.0	0.365	29.2	...	21.2	7.2	5.7	13.4	20.5	103.2	41	41	9	E
10	Orlando Magic	2016-2017	241.5	38.3	87.0	0.440	8.5	26.1	0.328	29.7	...	22.2	7.1	4.8	13.3	19.3	101.1	29	53	13	E
11	Brooklyn Nets	2016-2017	240.9	37.8	85.2	0.444	10.7	31.6	0.338	27.1	...	21.4	7.2	4.7	16.5	21.0	105.8	20	62	15	E
12	Charlotte Hornets	2016-2017	241.8	37.7	85.4	0.442	10.0	28.6	0.351	27.7	...	23.1	7.0	4.8	11.5	16.6	104.9	36	46	11	E
13	Atlanta Hawks	2016-2017	242.4	38.1	84.4	0.451	8.9	26.1	0.341	29.2	...	23.6	8.2	4.8	15.8	18.2	103.2	43	39	5	E
14	Chicago Bulls	2016-2017	241.2	38.6	87.1	0.444	7.6	22.3	0.340	31.0	...	22.6	7.8	4.8	13.6	17.7	102.9	41	41	8	E
15	Houston Rockets	2016-2017	241.2	40.3	87.2	0.462	14.4	40.3	0.357	25.9	...	25.2	8.2	4.3	15.1	19.9	115.3	55	27	3	W
16	Golden State Warriors	2016-2017	241.2	43.1	87.1	0.495	12.0	31.2	0.383	31.1	...	30.4	9.6	6.8	14.8	19.3	115.9	67	15	1	W
17	San Antonio Spurs	2016-2017	241.5	39.3	83.7	0.469	9.2	23.5	0.391	30.1	...	23.8	8.0	5.9	13.4	18.3	105.3	61	21	2	W
18	Minnesota Timberwolves	2016-2017	241.5	39.5	84.4	0.467	7.3	21.0	0.349	32.1	...	23.7	8.0	4.5	14.0	20.1	105.6	31	51	13	W
19	Denver Nuggets	2016-2017	240.9	41.2	87.7	0.469	10.6	28.8	0.368	30.6	...	25.3	6.9	3.9	15.0	19.1	111.7	40	42	9	W
20	Portland Trail Blazers	2016-2017	243.0	39.5	86.1	0.459	10.4	27.7	0.375	29.2	...	21.1	7.0	5.0	13.7	21.2	107.9	41	41	8	W
21	New Orleans Pelicans	2016-2017	242.7	39.1	87.0	0.450	9.4	26.8	0.350	29.8	...	22.8	7.8	5.5	12.9	18.2	104.3	34	48	10	W
22	Utah Jazz	2016-2017	240.9	37.0	79.5	0.466	9.6	26.0	0.372	27.3	...	20.1	6.7	5.0	13.6	18.8	100.7	51	31	5	W
23	Oklahoma City Thunder	2016-2017	241.5	39.5	87.4	0.452	8.4	25.8	0.327	31.0	...	21.0	7.9	5.0	15.0	20.9	106.6	47	35	6	W
24	Los Angeles Clippers	2016-2017	240.9	39.5	83.2	0.475	10.3	27.4	0.375	29.3	...	22.5	7.5	4.2	13.0	19.8	108.7	51	31	4	W
25	Los Angeles Lakers	2016-2017	240.3	39.3	87.4	0.450	8.9	25.7	0.346	30.4	...	20.9	8.2	3.9	15.2	20.7	104.6	26	56	14	W
26	Phoenix Suns	2016-2017	241.8	39.9	88.5	0.450	7.5	22.6	0.332	32.4	...	19.6	8.2	4.9	15.4	24.8	107.7	24	58	15	W
27	Memphis Grizzlies	2016-2017	242.7	36.4	83.6	0.435	9.4	26.5	0.354	27.0	...	21.3	8.0	4.2	12.9	22.4	100.5	43	39	7	W
28	Sacramento Kings	2016-2017	242.4	37.9	82.1	0.461	9.0	23.9	0.376	28.9	...	22.5	7.6	4.0	14.6	20.3	102.8	32	50	12	W
29	Dallas Mavericks	2016-2017	241.2	36.2	82.3	0.440	10.7	30.2	0.355	25.5	...	20.8	7.5	3.7	11.9	19.1	97.9	33	49	11	W

Figure 2: Exemplo de tabela contendo informações dos times na temporada.

2.1.2 Dados dos jogadores

Nos dados dos jogadores optamos por duas formas diferentes de dados: uma abordagem que reúne dados mais gerais (altura, peso, universidade em que jogou, entre outros) dos jogadores dos times e outra que possui as estatísticas de cada jogador do time na temporada. A primeira abordagem pode ser observada na figura 3 e a segunda na figura 4, exemplificando os dois tipos de tabela para cada time.

Com essas tabelas obtidas, optamos por usar as estatísticas dos jogadores para fazer uma tabela contendo as estatísticas de todos os jogadores que jogaram em uma determinada temporada, essa tabela foi posteriormente utilizada para fazer o clustering dos jogadores.

	Team	Season	Name	Height	Weight	Birthdate	Experience	College
0	Chicago Bulls	2017-2018	Quincy Pondexter	6-7	220	March 10, 1988	5	University of Washington
1	Chicago Bulls	2017-2018	Zach LaVine	6-5	189	March 10, 1995	3	University of California, Los Angeles
2	Chicago Bulls	2017-2018	Cameron Payne	6-3	185	August 8, 1994	2	Murray State University
3	Chicago Bulls	2017-2018	Jerian Grant	6-4	195	October 9, 1992	2	University of Notre Dame
4	Chicago Bulls	2017-2018	Bobby Portis	6-11	230	February 10, 1995	2	University of Arkansas
5	Chicago Bulls	2017-2018	Robin Lopez	7-0	255	April 1, 1988	9	Stanford University
6	Chicago Bulls	2017-2018	Denzel Valentine	6-6	212	November 16, 1993	1	Michigan State University
7	Chicago Bulls	2017-2018	Paul Zipser	6-8	215	February 18, 1994	1	NaN
8	Chicago Bulls	2017-2018	Kris Dunn	6-4	210	March 18, 1994	1	Providence College
9	Chicago Bulls	2017-2018	Kay Felder	5-9	176	March 29, 1995	1	Oakland University
10	Chicago Bulls	2017-2018	David Nwaba	6-4	209	January 14, 1993	1	California Polytechnic State University, San L...
11	Chicago Bulls	2017-2018	Lauri Markkanen	7-0	230	May 22, 1997	0	University of Arizona
12	Chicago Bulls	2017-2018	Cristiano Felicio	6-10	275	July 7, 1992	2	NaN
13	Chicago Bulls	2017-2018	Justin Holiday	6-6	185	April 5, 1989	4	University of Washington
14	Chicago Bulls	2017-2018	Nikola Mirotic	6-10	220	February 11, 1991	3	NaN
15	Chicago Bulls	2017-2018	Ryan Arcidiacono	6-3	188	March 26, 1994	0	Villanova University
16	Chicago Bulls	2017-2018	Antonio Blakeney (TW)	6-4	197	October 4, 1996	0	Louisiana State University

Figure 3: Exemplo de tabela contendo informações dos jogadores do time.

	Team	Season	Name	Age	G	GS	MP	FG	FGA	FG%	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS/G
0	Chicago Bulls	2017-2018	Justin Holiday	28	22	22	34.4	4.9	13.3	0.366	...	0.778	0.5	3.7	4.2	2.1	1.2	0.6	1.2	2.0	14.3
1	Chicago Bulls	2017-2018	Lauri Markkanen	20	23	23	30.6	5.0	12.8	0.390	...	0.825	1.1	6.8	7.9	1.4	0.6	0.6	1.4	1.8	14.3
2	Chicago Bulls	2017-2018	Robin Lopez	29	23	23	29.9	5.9	11.3	0.519	...	0.773	2.8	2.7	5.5	2.1	0.3	0.8	1.8	2.1	13.3
3	Chicago Bulls	2017-2018	Denzel Valentine	24	23	12	29.3	3.7	9.8	0.382	...	0.714	0.8	4.6	5.4	3.1	0.8	0.1	1.0	2.0	10.0
4	Chicago Bulls	2017-2018	Kris Dunn	23	19	10	27.4	5.2	11.7	0.439	...	0.577	0.7	4.0	4.7	4.5	1.9	0.3	3.2	3.2	12.2
5	Chicago Bulls	2017-2018	Jerian Grant	25	23	14	24.7	3.1	7.5	0.413	...	0.776	0.6	2.2	2.8	4.9	0.8	0.1	1.5	1.8	8.9
6	Chicago Bulls	2017-2018	Bobby Portis	22	14	0	22.1	4.6	10.3	0.444	...	0.774	1.9	5.1	7.0	1.7	0.2	0.4	1.3	1.1	12.0
7	Chicago Bulls	2017-2018	David Nwaba	25	11	3	20.8	2.6	4.8	0.547	...	0.654	0.5	4.6	5.2	1.5	0.7	0.6	0.8	1.7	7.0
8	Chicago Bulls	2017-2018	Paul Zipser	23	20	7	16.6	1.7	5.3	0.324	...	0.333	0.3	3.1	3.4	1.1	0.4	0.2	0.9	1.6	4.1
9	Chicago Bulls	2017-2018	Cristiano Felicio	25	20	0	15.3	1.8	2.9	0.621	...	0.556	1.2	1.8	3.0	0.7	0.2	0.2	0.8	2.1	4.1
10	Chicago Bulls	2017-2018	Antonio Blakeney	21	11	0	14.8	2.5	7.0	0.364	...	0.737	0.3	1.5	1.8	0.6	0.2	0.0	0.7	0.7	7.1
11	Chicago Bulls	2017-2018	Quincy Pondexter	29	15	1	10.5	0.9	2.8	0.310	...	0.824	0.5	0.8	1.3	0.4	0.5	0.1	0.6	1.4	2.9
12	Chicago Bulls	2017-2018	Kay Felder	22	13	0	10.0	1.5	5.0	0.308	...	0.917	0.1	1.0	1.1	1.5	0.2	0.1	1.2	0.8	4.2
13	Chicago Bulls	2017-2018	Ryan Arcidiacono	23	1	0	5.0	0.0	2.0	0.000	...	NaN	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0

Figure 4: Exemplo de tabela contendo estatísticas dos jogadores do time.

2.2 Abordagens

Ao ver os dados, optamos por algumas abordagens que seguem caminhos distintos. Uma delas é fazer análise exploratória diretamente nos dados. Outra abordagem é verificar os grupos por meio de algoritmos de clustering: a similaridade entre os times e entre os jogadores. Além disso, utilizamos o uso de alguns algoritmos para experimentar sobre a predição da temporada atual, como Random Forest, além de analisarmos a importância de cada atributo por meio de seus p-values.

2.2.1 Análise

A análise exploratória feita envolve a construção de algumas visualizações de densidade das pontuações de cada temporada.

A partir das mesmas, e de alguns cálculos de centralidade (média e variância), conseguimos tirar a seguinte conclusão: A temporada 2017-2018, até então, foi a com maior pontuação média dos times.Isso se deve ao fato de não termos todos os dados da temporada 2017-2018, por ela ainda não ter acabado. Outra análise interessante feita foi a de porcentagens de jogadores por universidade.

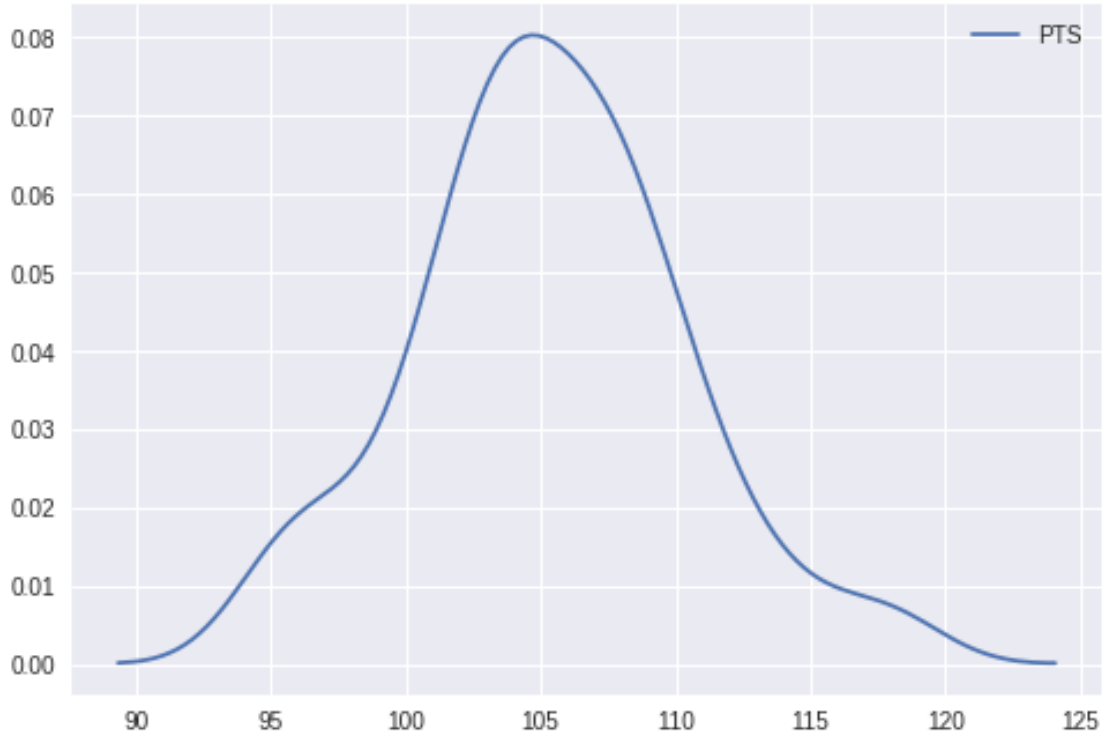

Figure 5: Exemplo de gráfico de densidade das pontuações. No caso, 2017-2018

Na mesma, analisamos quais universidades tem mais jogadores jogando na liga atualmente. Para isso, construímos alguns gráficos de barras, que nos determinam a universidade na qual saem mais jogadores da liga: Universidade do Kentucky, conhecida por ser uma grande potência no torneio da NCAA (a liga de esportes universitária americana), porta de entrada para a maioria dos jogadores da NBA.

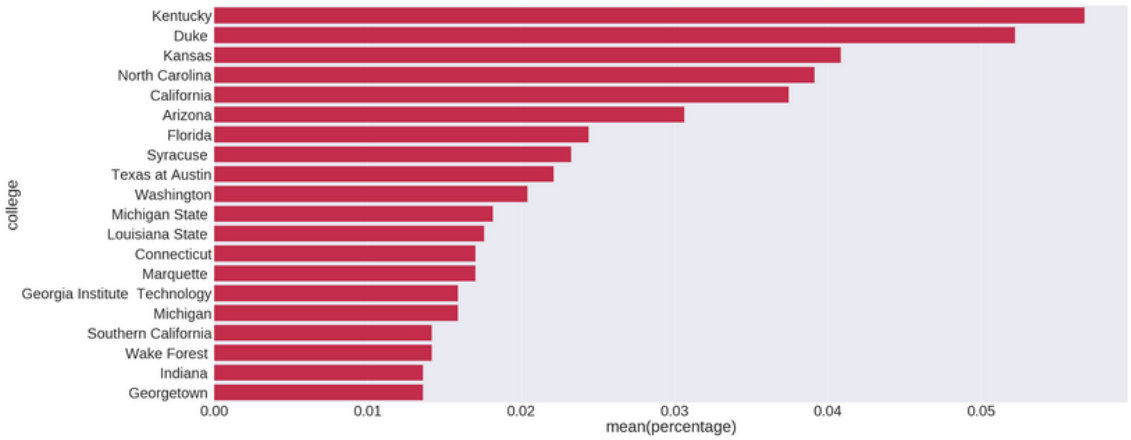

Figure 6: Gráfico de barra das porcentagens médias de jogadores de cada universidade

2.2.2 Clustering dos Jogadores

Para o clustering dos jogadores, utilizamos o algoritmo KMeans com parâmetro K = 100. Pelo alto número de jogadores, assumimos que um K alto seria ideal, para termos grupos mais específicos e homogêneos. As análises dos clusters observados tiveram abordagens distintas: observamos quais jogadores se parecem com um jogador bem sucedido (ex: LeBron James) e o que fez alguns jogadores pertencerem a diferentes clusters. Tomamos como exemplo Stephen Curry e DeMarcus Cousins, por serem casos interessantes.

2.2.3 Clustering dos Times

No clustering de times, houve maior facilidade, pois são poucos times em relação a jogadores, então utilizamos o mesmo algoritmo, porém com K = 15. Esse número se mostrou eficiente pelos resultados apresentados, apesar de podermos observar alguns grupos um pouco heterogêneos. Analisamos os clusters mais interessantes novamente, sendo esses os que apresentaram agrupamentos como: times da parte de baixo da tabela ou times intermediárimos, ou mesmo um único time, em diversas temporadas.

2.2.4 Predição

Para a predição, fizemos uma abordagem de predição na temporada atual utilizando os dados de temporadas anteriores. Além disso, usamos alguns modelos lineares para medir os p-values, as importâncias de cada atributo na predição.

2.3 Resultados

Os resultados observados para cada análise feita foram interessantes sob várias perspectivas. Na análise, podemos explorar os dados de forma a fazer algumas interpretações em mesmo análises rasas, como observar um gŕafico de densidade e calcular medidas centrais. Já nos clusters criados, permitiu com que observássemos consistências nas predições feitas por analistas e crenças gerais do público do esporte, como unanimidade na superioridade do Golden State Warriors, bem como a performance de LeBron James. Nas predições, pudemos ver os p-values de diversas features, sendo as mais importantes usadas no modelo de predição final.

2.3.1 Análise

Os resultados da análise, como mencionado anteriormente, foi o de gráficos aproximadamente normais para todas as temporadas, tendo a pontuação média de cada temporada aumentado ao longo dos anos, como podemos ver na figura 7.

```
média e variancia 2014-2015: 100.01666666666667,17.67272222222222
média e variancia 2015-2016: 102.66333333333333,13.876322222222225
média e variancia 2016-2017: 105.58999999999999,16.479566666666667
média e variancia 2017-2018: 105.29666666666667,24.255655555555553
```

Figure 7: Medidas para cada temporada

Além de vermos a distribuição de jogadores por universidade, que, como vimos anteriormente, é majoritariamente pela universidade do Kentucky, fazemos também uma análise da altura dos jogadores em relação a porcentagem relativa, similarmente a um histograma, visto na figura 8.

2.3.2 Clustering

Os resultados apresentados em cada abordagem feita usando o K-Means foram interessantes tanto por serem reveladores quanto por serem confirmadores do que se esperava dos jogadores ou times. Como foi dito, para os jogadores utilizamos K = 100, pelo alto número de jogadores na base coletada e para os times utilizamos K = 15, pelo número reduzido de times(em relação aos jogadores).

Para o de jogadores, realizamos o K-Means, e logo após isso, vimos a quais clusters o jogador LeBron James, que, por ser um jogador de destaque, a análise se torna interessante. Podemos ver todos os clusters aos quais LeBron pertenceu nas quatro temporadas na figura 9. Como a temporada 2017-2018 ainda não acabou, não temos dados suficientes para alocar LeBron ou qualquer outro jogador em um cluster relevantemente parecido com o mesmo. Por isso, só analisamos os 3 primeiros resultados. Neles, vemos que LeBron pertenceu aos clusters 84 e 29. Como o ápice da carreira do jogador foi em 2015-2016, tomamos o próximo passo da análise vendo os jogadores do cluster 84, ao qual LeBron pertenceu no ápice de sua carreira. Os jogadores do cluster podem ser vistos na figura 10. Os jogadores mostrados no cluster, como esperado, são jogadores de grande sucesso na NBA atualmente, o que mostra coerência da criação dos grupos. Um exemplo dos jogadores vistos é Stephen Curry, este que joga pelos atuais campeões Golden State Warriors.

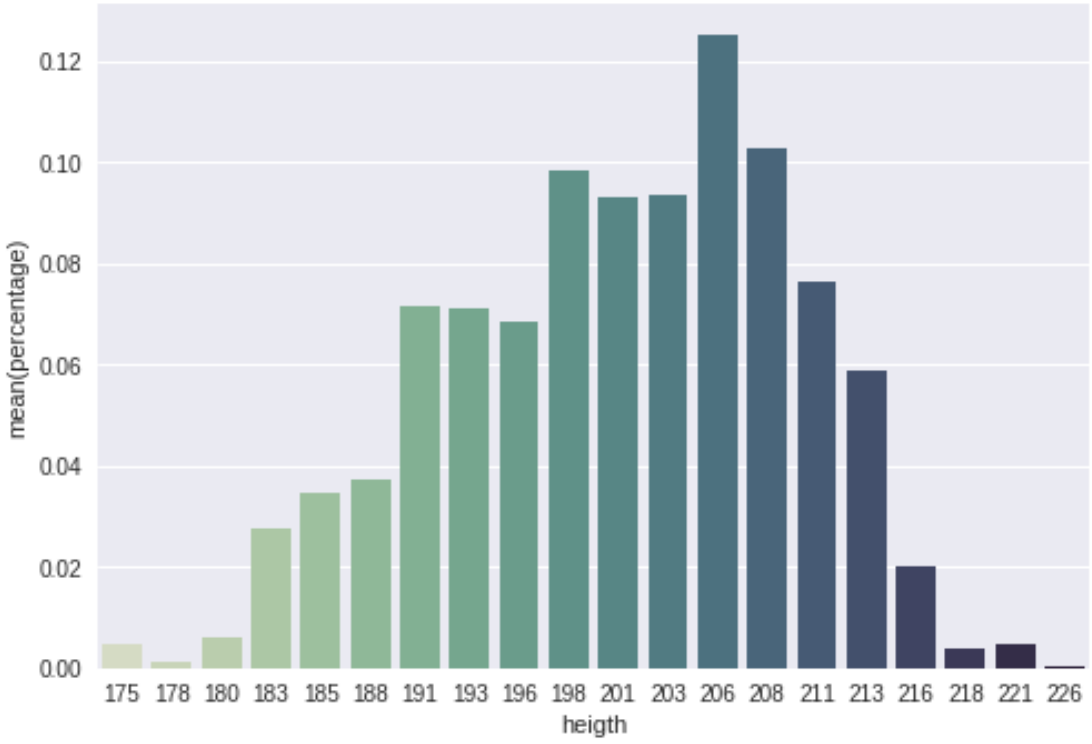

Figure 8: Gráfico de barras frequência por altura

Tomemos uma análise do mesmo: vejamos a quais clusters o mesmo pertence. Na figura 11, vemos todos os clusters aos quais o mesmo pertence. Como podemos ver, o mesmo pertence sempre ao cluster 4, exceto no ano em que seu time perdeu para os Cleveland Cavaliers, time de LeBron. Isso pode ser interpretado como um maior investimento do time em Stephen, pois nessa temporada, o mesmo tem estatísticas mais altas que em outras, e assim isso fez com que o time tivesse maior fraqueza em outras áreas.

Outra análise adequada é sobre o jogador DeMarcus Cousins, pois o mesmo muda de time no meio de uma das temporadas. O que é intrigante é que quando isso acontece, o jogador muda de cluster. Vendo os clusters mais de perto, podemos observar que antes da mudança, o jogador estava em um cluster de extrema homogeneidade, além de possuir alguns dos melhores jogadores da NBA, como Dwyane Wade. Já quando o mesmo muda, observamos maior hetereogeneidade no grupo, pois o mesmo apresenta jogadores de diversas posições na quadra. Isso pode ser interpretado como a alocação malfeita ou tentativa de colocar o jogador em uma posição nova, o que o desloca para outro cluster.

	Team	Season	Name	Age	G	GS	MP	FG	FGA	FG%	...	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS/G	Cluster
23	Cleveland Cavaliers	2014-2015	LeBron James	30	69	69	36.1	9.0	18.5	0.488	...	0.7	5.3	6.0	7.4	1.6	0.7	3.9	2.0	25.3	29
592	Cleveland Cavaliers	2015-2016	LeBron James	31	76	76	35.6	9.7	18.6	0.520	...	1.5	6.0	7.4	6.8	1.4	0.6	3.3	1.9	25.3	84
1118	Cleveland Cavaliers	2016-2017	LeBron James	32	74	74	37.8	9.9	18.2	0.548	...	1.3	7.3	8.6	8.7	1.2	0.6	4.1	1.8	26.4	84
1660	Cleveland Cavaliers	2017-2018	LeBron James	33	25	25	37.2	11.0	18.7	0.587	...	1.2	6.8	8.0	8.6	1.4	1.2	4.1	1.7	28.2	36

Figure 9: Clusters de LeBron James

Já para os clusters de times, as análises foram mais diretas. Como temos um número baixo de clusters, bastou analisarmos alguns clusters (quase todos) de perto, para que pudéssemos ver o sucesso ou fracasso do K-Means. Inicialmente, vimos que os times da parte de baixo ficaram, em sua grande maioria, em dois clusters: o 0 e o 11. Podemos ver o cluster 0 na figura 13. Após isso, vimos que os times da parte de cima da tabela também ficaram concentrados em alguns clusters, como por exemplo o 2 e o 14. Podemos ver o cluster 2 na figura 14. O mais interessante dessa análise se dá quando vemos o cluster 10: o mesmo só tem o time Golden State Warriors, o que mostra que o time teve um desempenho ímpar, diferente de todos os outros, nas temporadas analisadas, e isso

	Team	Season	Name	Age	G	GS	MP	FG	FGA	FG%	...	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS/G	Cluster
278	Houston Rockets	2014-2015	James Harden	25	81	81	36.8	8.0	18.1	0.440	...	0.9	4.7	5.7	7.0	1.9	0.7	4.0	2.6	27.4	84
592	Cleveland Cavaliers	2015-2016	LeBron James	31	76	76	35.6	9.7	18.6	0.520	...	1.5	6.0	7.4	6.8	1.4	0.6	3.3	1.9	25.3	84
610	Toronto Raptors	2015-2016	DeMar DeRozan	26	78	78	35.9	7.9	17.7	0.446	...	0.8	3.7	4.5	4.0	1.0	0.3	2.2	2.1	23.5	84
831	Houston Rockets	2015-2016	James Harden	26	82	82	38.1	8.7	19.7	0.439	...	0.8	5.3	6.1	7.5	1.7	0.6	4.6	2.8	29.0	84
850	Golden State Warriors	2015-2016	Stephen Curry	27	79	79	34.2	10.2	20.2	0.504	...	0.9	4.6	5.4	6.7	2.1	0.2	3.3	2.0	30.1	84
917	Portland Trail Blazers	2015-2016	Damian Lillard	25	75	75	35.7	8.2	19.7	0.419	...	0.6	3.4	4.0	6.8	0.9	0.4	3.2	2.2	25.1	84
971	Oklahoma City Thunder	2015-2016	Kevin Durant	27	72	72	35.8	9.7	19.2	0.505	...	0.6	7.6	8.2	5.0	1.0	1.2	3.5	1.9	28.2	84
972	Oklahoma City Thunder	2015-2016	Russell Westbrook	27	80	80	34.4	8.2	18.1	0.454	...	1.8	6.0	7.8	10.4	2.0	0.3	4.3	2.5	23.5	84
1103	Boston Celtics	2016-2017	Isaiah Thomas	27	76	76	33.8	9.0	19.4	0.463	...	0.6	2.1	2.7	5.9	0.9	0.2	2.8	2.2	28.9	84
1118	Cleveland Cavaliers	2016-2017	LeBron James	32	74	74	37.8	9.9	18.2	0.548	...	1.3	7.3	8.6	8.7	1.2	0.6	4.1	1.8	26.4	84
1140	Toronto Raptors	2016-2017	DeMar DeRozan	27	74	74	35.4	9.7	20.9	0.467	...	0.9	4.3	5.2	3.9	1.1	0.2	2.4	1.8	27.3	84
1227	Washington Wizards	2016-2017	John Wall	26	78	78	36.4	8.3	18.4	0.451	...	0.8	3.4	4.2	10.7	2.0	0.6	4.1	1.9	23.1	84
1355	Chicago Bulls	2016-2017	Jimmy Butler	27	76	75	37.0	7.5	16.5	0.455	...	1.7	4.5	6.2	5.5	1.9	0.4	2.1	1.5	23.9	84
1373	Houston Rockets	2016-2017	James Harden	27	81	81	36.4	8.3	18.9	0.440	...	1.2	7.0	8.1	11.2	1.5	0.5	5.7	2.7	29.1	84
1408	San Antonio Spurs	2016-2017	Kawhi Leonard	25	74	74	33.4	8.6	17.7	0.485	...	1.1	4.7	5.8	3.5	1.8	0.7	2.1	1.6	25.5	84
1459	Portland Trail Blazers	2016-2017	Damian Lillard	26	75	75	35.9	8.8	19.8	0.444	...	0.6	4.3	4.9	5.9	0.9	0.3	2.6	2.0	27.0	84
1474	New Orleans Pelicans	2016-2017	Anthony Davis	23	75	75	36.1	10.3	20.3	0.505	...	2.3	9.5	11.8	2.1	1.3	2.2	2.4	2.2	28.0	84
1515	Oklahoma City Thunder	2016-2017	Russell Westbrook	28	81	81	34.6	10.2	24.0	0.425	...	1.7	9.0	10.7	10.4	1.6	0.4	5.4	2.3	31.6	84

Figure 10: Cluster 84 - Ápice da carreira de LeBron James

	Team	Season	Name	Age	G	GS	MP	FG	FGA	FG%	...	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS/G	Cluster
298	Golden State Warriors	2014-2015	Stephen Curry	26	80	80	32.7	8.2	16.8	0.487	...	0.7	3.6	4.3	7.7	2.0	0.2	3.1	2.0	23.8	4
850	Golden State Warriors	2015-2016	Stephen Curry	27	79	79	34.2	10.2	20.2	0.504	...	0.9	4.6	5.4	6.7	2.1	0.2	3.3	2.0	30.1	84
1392	Golden State Warriors	2016-2017	Stephen Curry	28	79	79	33.4	8.5	18.3	0.468	...	0.8	3.7	4.5	6.6	1.8	0.2	3.0	2.3	25.3	4
1891	Golden State Warriors	2017-2018	Stephen Curry	29	23	23	32.6	8.3	17.7	0.473	...	0.7	4.4	5.1	6.6	1.7	0.2	2.9	2.5	26.3	36

Figure 11: Clusters de Stephen Curry

	Team	Season	Name	Age	G	GS	MP	FG	FGA	FG%	...	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS/G	Cluster
539	Sacramento Kings	2014-2015	DeMarcus Cousins	24	59	59	34.1	8.4	18.1	0.467	...	3.1	9.5	12.7	3.6	1.5	1.7	4.3	4.1	24.1	59
1073	Sacramento Kings	2015-2016	DeMarcus Cousins	25	65	65	34.6	9.2	20.5	0.451	...	2.4	9.1	11.5	3.3	1.6	1.4	3.8	3.6	26.9	59
1475	New Orleans Pelicans	2016-2017	DeMarcus Cousins	26	17	17	33.8	8.4	18.5	0.452	...	2.2	10.2	12.4	3.9	1.5	1.1	3.6	4.4	24.4	36
1602	Sacramento Kings	2016-2017	DeMarcus Cousins	26	55	55	34.4	9.2	20.3	0.452	...	2.1	8.5	10.6	4.8	1.4	1.3	3.8	3.7	27.8	59
1964	New Orleans Pelicans	2017-2018	DeMarcus Cousins	27	25	25	35.6	8.9	19.3	0.462	...	2.2	10.5	12.6	5.1	1.6	1.7	5.0	3.5	25.9	36

Figure 12: Clusters de DeMarcus Cousins

confere com a realidade visto que esse time foi o melhor time da NBA nas temporadas regulares analisadas, inclusive na temporada em que foi derrotado nas finais. Vemos o cluster 10 na figura 15.

	2P	2P%	2PA	3P	3P%	3PA	AST	BLK	CONF	DRB	...	PF	POS	PTS	STL	Season	TOV	TRB	Team	WIN	Cluster
8	28.3	0.454	62.3	6.8	0.347	19.7	21.3	4.7	E	29.8	...	21.6	15.0	91.9	7.0	2014-2015	14.7	40.4	New York Knicks	17.0	0
10	30.8	0.486	63.3	6.8	0.347	19.5	20.6	3.8	E	31.8	...	20.9	13.0	95.7	7.9	2014-2015	14.9	41.8	Orlando Magic	25.0	0
18	31.5	0.461	68.3	5.0	0.332	14.9	21.6	4.0	W	29.3	...	19.2	15.0	97.8	8.1	2014-2015	15.0	40.9	Minnesota Timberwolves	16.0	0
25	30.8	0.461	66.8	6.5	0.344	18.9	20.9	4.5	W	32.3	...	21.2	14.0	98.5	7.0	2014-2015	13.2	43.9	Los Angeles Lakers	21.0	0
41	31.8	0.481	66.0	6.5	0.352	18.4	22.3	4.0	E	31.9	...	18.0	14.0	98.6	7.6	2015-2016	14.8	42.4	Brooklyn Nets	21.0	0

Figure 13: Cluster da parte de baixo da tabela

	2P	2P%	2PA	3P	3P%	3PA	AST	BLK	CONF	DRB	...	PF	POS	PTS	STL	Season	TOV	TRB	Team	WIN	Cluster
13	28.1	0.506	55.5	10.0	0.380	26.2	25.7	4.6	E	31.8	...	17.8	1.0	102.5	9.1	2014-2015	14.2	40.6	Atlanta Hawks	60.0	2
47	33.2	0.515	64.4	7.0	0.375	18.5	24.5	5.9	W	34.5	...	17.5	2.0	103.5	8.3	2015-2016	13.1	43.9	San Antonio Spurs	67.0	2
77	30.1	0.500	60.2	9.2	0.391	23.5	23.8	5.9	W	33.9	...	18.3	2.0	105.3	8.0	2016-2017	13.4	43.9	San Antonio Spurs	61.0	2

Figure 14: Cluster da parte de cima da tabela

	2P	2P%	2PA	3P	3P%	3PA	AST	BLK	CONF	DRB	...	PF	POS	PTS	STL	Season	TOV	TRB	Team	WIN	Cluster
16	30.8	0.514	60.0	10.8	0.398	27.0	27.4	6.0	W	34.3	...	19.9	1.0	110.0	9.3	2014-2015	14.5	44.7	Golden State Warriors	67.0	10
46	29.4	0.528	55.7	13.1	0.416	31.6	28.9	6.1	W	36.2	...	20.7	1.0	114.9	8.4	2015-2016	15.2	46.2	Golden State Warriors	73.0	10
76	31.1	0.557	55.8	12.0	0.383	31.2	30.4	6.8	W	35.0	...	19.3	1.0	115.9	9.6	2016-2017	14.8	44.4	Golden State Warriors	67.0	10

Figure 15: Cluster 10 - Golden State Warriors

2.3.3 Predição

Já na parte de predição, realizamos duas tarefas: Treinar um Random Forest com todas as features dos times visando prever a posição dos times na atual temporada; Selecionar as features mais importantes usando como critério os p-values retornados por modelos lineares e também pelo Random Forest.

Os resultados podem ser observados nas figuras 16 - 18. A figura 16 mostra as importâncias de cada feature observada no conjunto de treino, obtidas através do Random Forest. A figura 17 mostra as regressões lineares da variável y com as features selecionadas. Nesse caso, fizemos uma filtragem das features, excluindo algumas como tempo de jogo, além de separarmos os modelos lineares para as conferências leste e oeste, que mostraram diferenças significativas.

Por fim, na figura 16 vemos o erro médio das predições realizadas pelo Random Forest com seleção de features. O erro mostrado é aproximadamente 2.86, o que indica um modelo que não possui overfitting, pois aparenta não ter muito viés. Por outro lado, é uma taxa de erro interessante, visto que estamos usando como conjunto de teste os próprios dados da temporada 2017-2018.

```
Feature ranking:
1. feature 9 (0.070643)
2. feature 3 (0.060250)
3. feature 6 (0.057904)
4. feature 19 (0.054016)
5. feature 0 (0.049957)
```

Figure 16: Importância de cada atributo usando Random Forest

3 Conclusão

Como podemos ver através da análise dos resultados, todas as abordagens resultaram em informações relevantes e interessantes a respeito dos assuntos ao qual estavam relacionadas. Observamos que as análises exploratórias são importantes para conseguir observar fatos que não são possíveis de serem observados só com as tabelas de dados. Já na realização do *clustering* tanto dos jogadores quanto dos times pudemos ver de que modo os jogadores/times podem ser agrupados utilizando um dos algoritmos mais conhecidos e respeitados da aprendizagem não supervisionada, e como falado na seção anterior, algumas relações interessantes são observadas nos resultados dos algoritmos. E na parte de predição descobrimos quais são as estatísticas mais importantes para determinar o sucesso de um time na liga, tanto na conferência leste quanto oeste e, a partir disso, realizamos a predição das posições dos times na temporada atual com base nesses dados das temporadas anteriores, obtendo resultados coerentes com a posição atual dos times na temporada, apesar dela estar só no começo.

Para o futuro, o que podemos melhorar na análise exploratória é fazer uma análise mais profunda das estatísticas obtidas a fim de encontrar mais padrões que podem ser analisados. Na

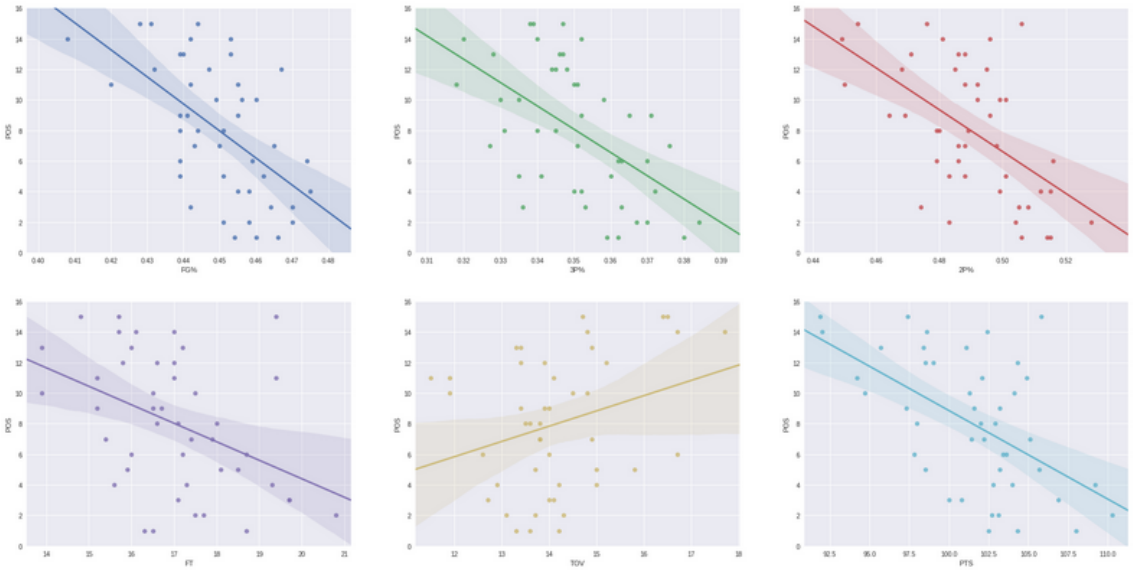

Figure 17: Regressões lineares de y com as variáveis selecionadas na conferência leste

```
2.86666666667
```

Figure 18: Erro médio do modelo final

realização dos clusters uma possível melhora é realizar uma validação cruzada para escolher o número de clusters ótimo e garantir, assim, resultados ainda melhores e também utilizar outros algoritmos e ver como eles se comportam em relação ao K-Means, algoritmo mais famoso e reconhecido de clustering. Já na parte das predições podemos buscar estatísticas diferentes para que tenhamos um número maior de atributos para serem analisados, o que pode resultar em atributos que inicialmente eram considerados não muito importantes que na verdade, com a ajuda dos dados para mostrar, são sim importantes.

Por fim, concluímos que o projeto foi uma experiência muito boa. Além de aumentar ainda mais o nosso conhecimento das ferramentas que foram utilizadas, nos deu uma boa visão prática de como é trabalhar num projeto de *data science* do começo ao fim, da coleta dos dados, passando pelo processo de observar o que pode ser feito com os dados, para a realização das técnicas e por fim a análise dos resultados obtidos a partir . Todas as dificuldades que foram encontradas durante a realização do projeto e o que tivemos que fazer para superá-las com certeza foram um aprendizado muito importante tanto para a carreira acadêmica quanto para a carreira profissional.