3 Pontos Sobre NBA

Angelo Carmignani, Gabriel Bortoli, Wesley Maia

2023-07-13

Contents

Prefácio

É com grande satisfação que apresento este trabalho de visualização de informação sobre dados da NBA, desenvolvido no âmbito da disciplina MAI5017 - Visualização de Informação, como parte integrante do Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria (MECAI) do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo.

A visualização de dados desempenha um papel crucial na análise e compreensão de conjuntos complexos de informações. No contexto do esporte, em particular, a visualização de dados tem se mostrado uma ferramenta poderosa para explorar e comunicar insights valiosos a partir de estatísticas e padrões relacionados aos jogos e aos jogadores de basquete.

Este trabalho tem como objetivo principal explorar os dados da NBA e utilizar técnicas de visualização para revelar informações significativas e interessantes sobre as partidas, os jogadores, as equipes e as tendências ao longo do tempo. Para tanto, foram utilizadas ferramentas e técnicas de programação em Python, um ambiente amplamente adotado no campo da ciência de dados e análise estatística.

A abordagem adotada neste trabalho segue o formato bookdown, uma estrutura que permite combinar narrativa, código e gráficos interativos de maneira integrada e coesa. Dessa forma, os resultados obtidos são apresentados de maneira clara e acessível, facilitando a compreensão e a exploração dos dados pelos leitores.

O estudo da visualização de informação aplicada à NBA não se restringe apenas ao interesse acadêmico, mas também possui um potencial significativo na indústria esportiva, na tomada de decisões estratégicas e no desenvolvimento de estratégias competitivas. Por meio da visualização eficaz de dados, é possível identificar padrões ocultos, analisar desempenhos individuais e coletivos, e extrair insights relevantes para apoiar a tomada de decisões informadas.

Agradeço a todos os professores e colegas do MECAI pelo apoio e incentivo ao longo desta jornada de aprendizado. Espero que este trabalho contribua para o avanço do conhecimento na área de visualização de informação e inspire pesquisas futuras no campo da análise de dados esportivos.

Desejo a todos uma leitura proveitosa e enriquecedora.

Sobre os Autores

FAZER

Chapter 1

Introdução

A NBA (National Basketball Association) é uma das ligas de basquete mais populares e prestigiadas do mundo, com uma rica história que se estende por 77 anos. Desde sua fundação em 1946, a NBA tem sido palco de inúmeras façanhas atléticas, rivalidades intensas e momentos memoráveis que cativaram os fãs de basquete em todo o mundo.

Neste trabalho de Visualização de Dados, exploraremos um conjunto abrangente de estatísticas dos últimos 71 anos da NBA. Utilizando o Jupyter Notebook, mergulharemos nesses dados para extrair insights valiosos sobre as equipes, jogadores e padrões que moldaram a liga ao longo das décadas.

1.1 Objetivo

O objetivo desta análise é investigar diversas facetas do basquete profissional, desde o desempenho das equipes até as estatísticas individuais dos jogadores. Por meio de técnicas de análise de dados e visualização, buscaremos responder a perguntas como:

Quais equipes dominaram a NBA ao longo dos anos? Quais jogadores tiveram as melhores performances estatísticas em diferentes épocas? Existem tendências ou padrões significativos nas estatísticas da NBA ao longo das décadas? Como o jogo evoluiu em termos de estilo de jogo, pontuação média e estilos de arremesso? Ao responder a essas perguntas, esperamos obter uma compreensão mais profunda da evolução da NBA e das dinâmicas que impulsionam o sucesso das equipes e dos jogadores ao longo do tempo. Esses insights não apenas fornecerão informações interessantes sobre a história da liga, mas também poderão ajudar a prever tendências futuras e orientar estratégias para equipes e jogadores no presente.

Chapter 2

Dados

O projeto tem um conjunto de dados fornecido pelo Kaggle chamado nba.csv. NA base apresenta os dados dos jogadores de todas as temporadas de 1951 a 2022, com um total de 33330 ocorrências

2.1 Descrição da base

As colunas são descritas a seguir:

POR JOGADOR:

Variável	Descrição
Rank	A classificação do jogador (ordenado por pontos marcados a cada temporada)
Year	O ano da temporada (por exemplo, "2018-19")
Season Start Year	O ano de início da temporada (por exemplo, 2018)
Season Type	Temporada regular ou playoffs
Player ID	Um ID gerado para cada jogador
Player	O nome do jogador
Team ID	ID gerado para cada equipe
Team	A equipe do jogador na respectiva temporada
Games Played	Jogos disputados na respectiva temporada
Minutes Played	Minutos jogados na respectiva temporada
FG Made	Cestas de campo convertidas (Field Goals Made)
FG Attempts	Tentativas de cestas de campo (Field Goals Attempted)
FG %	Porcentagem de acertos de cestas de campo (Field Goal Percentage)

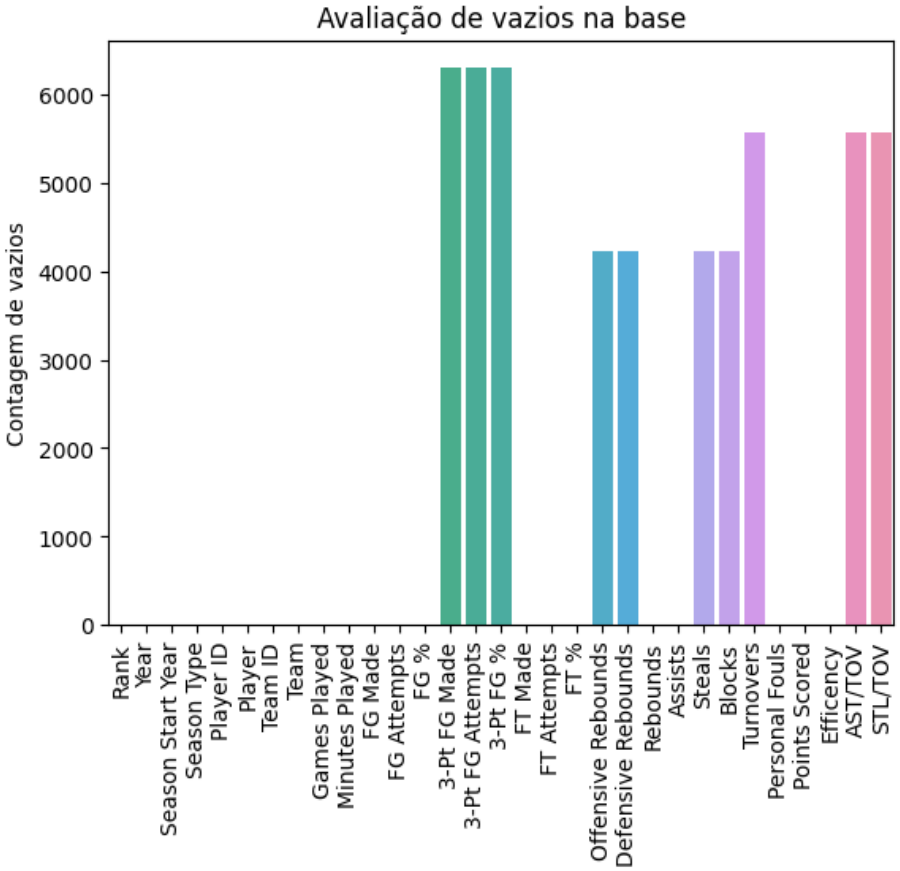
Variável	Descrição
3-Pt FG Made	Cestas de três pontos convertidas (3 Point Field Goals Made)
3-Pt FG Attempts	Tentativas de cestas de três pontos (3 Point Field Goals Attempted)
3-Pt FG %	Porcentagem de acertos de cestas de três pontos (3 Point Field Goal Percentage)
FT Made	Lances livres convertidos (Free Throws Made)
FT Attempts	Tentativas de lances livres (Free Throws Attempted)
FT %	Porcentagem de acertos de lances livres (Free Throw Percentage)
Offensive Rebounds	Rebotes ofensivos
Defensive Rebounds	Rebotes defensivos
Rebounds	Total de rebotes (ofensivos + defensivos)
Assists	Assistências
Steals	Roubos de bola
Blocks	Tocos (bloqueios de arremessos)
Turnovers	Perdas de bola (erros)
Personal Fouls	Faltas pessoais
Points Scored	Pontos marcados
Efficiency	Eficiência calculada como (Pontos Marcados + Rebotes + Assistências + Roubos de Bola + Tocos - Chutes de Campo Perdidos - Lances Livres Perdidos - Perdas de Bola) dividido por Jogos Disputados
AST/TOV	Taxa de assistências para turnovers (Assist-to-Turnover ratio)
STL/TOV	Taxa de roubos de bola para turnovers (Steal-to-Turnover ratio)

Chapter 3

Tratamento dos Dados

Para analisar os perfis de usuário da NBA, Pandas será usados para carregar o conjunto de dados em um DataFrame para que possa ser explorado e visualizado com Python.

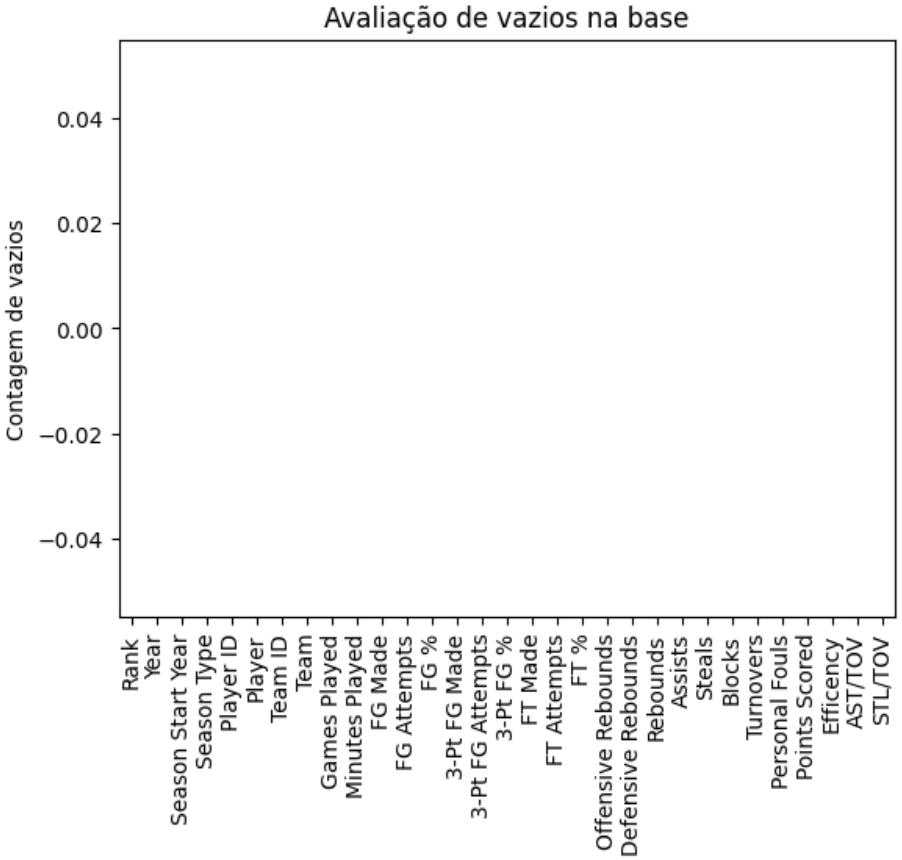
Existem algum valores vazios o que é um problema para as análises que serão feitas. Vamos conferir quantas linhas do dataset existem esse tipo de dados.



O DataFrame apresenta valores nulos em algumas colunas, como 3-Pt FG Made, 3-Pt FG Attempts, 3-Pt FG %, Offensive Rebounds, Defensive Rebounds, Steals, Blocks, Turnovers, AST/TOV e STL/TOV. Esses valores nulos podem indicar a ausência de dados ou informações faltantes para algumas estatísticas específicas dos jogadores em determinadas temporadas.

Para manter a consistência e garantir a confiabilidade da análise, optou-se por filtrar o DataFrame, excluindo as temporadas anteriores a 1978. Dessa forma, as colunas mencionadas estarão preenchidas a partir desse ano, permitindo uma análise mais completa e precisa das estatísticas dos jogadores da NBA.

Essa decisão foi tomada para evitar distorções nos resultados devido à ausência de dados em períodos anteriores, garantindo que a análise seja baseada em informações mais completas e recentes.


Avaliação de vazios na base
Contagem de vazios
0.04
0.02
0.00
−0.02
−0.04
Rank
Year
Season Start Year
Season Type
Player ID
Player
Team ID
Team
Games Played
Minutes Played
FG Made
FG Attempts
FG %
3-Pt FG Made
3-Pt FG Attempts
3-Pt FG %
FT Made
FT Attempts
FT %
Offensive Rebounds
Defensive Rebounds
Rebounds
Assists
Steals
Blocks
Turnovers
Personal Fouls
Points Scored
Efficency
AST/TOV
STL/TOV

Chapter 4

O que define um jogador bom?

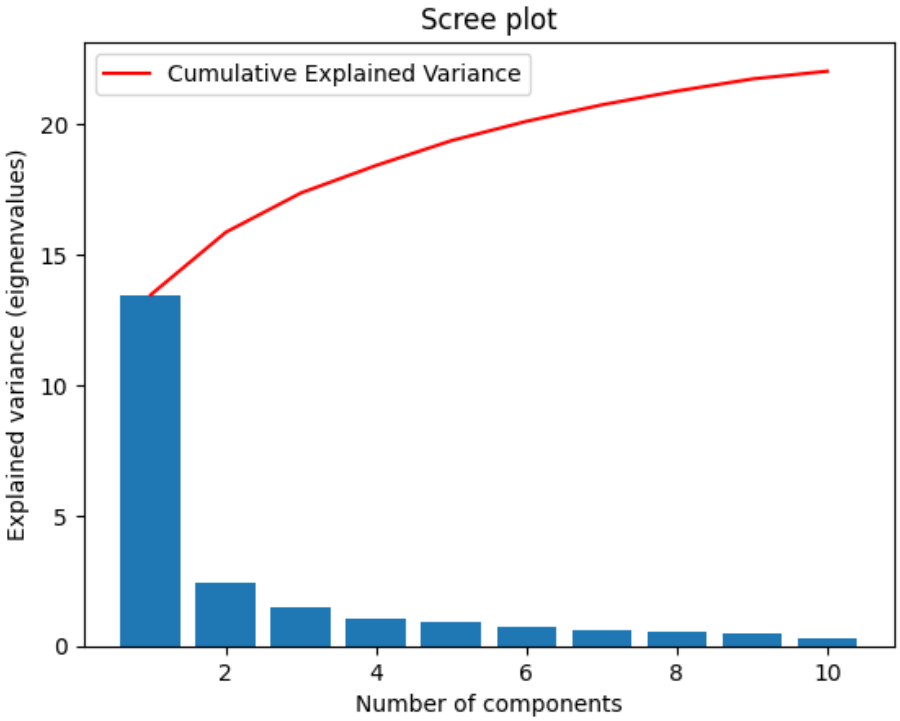
Para simplificação vamos utilizar o rank apresentado no dataset, que seriam os jogadores ordenados de acordo com a pontuação por temporada.

Apesar dessa rank não levar em consideração fatores defensivos, será feito uma avaliaçào se os maiores "cestinhas" também apresentam características defensivas acima da média.

A primeiro momento vamos avaliar como as informações estatísticas de cada jogador por temporada varia e como está relacionado com o rank. Para tanto será utiizado a técnica do PCA para entender melhor esse comportamento.

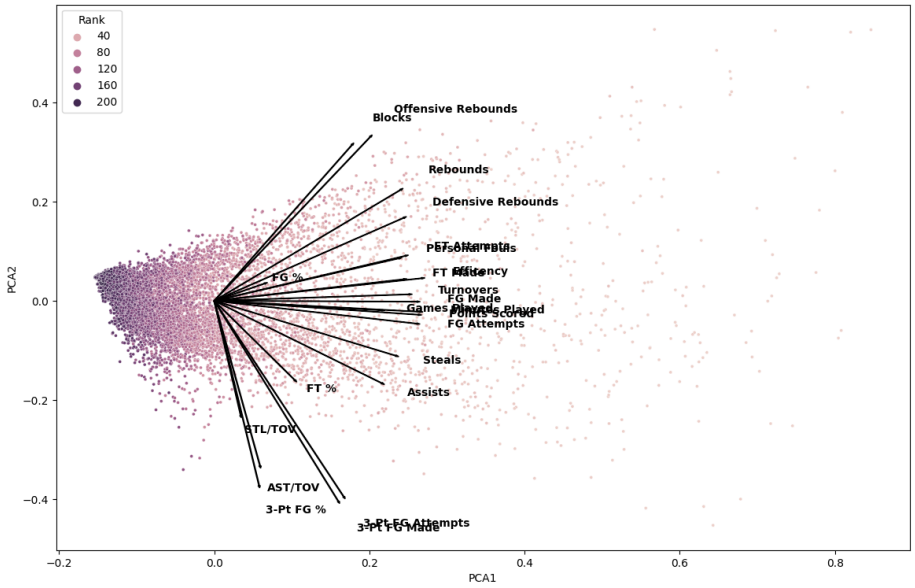
Para definir o número de componentes ideal foi availiada a variância explicada por cada componente e adotada a regra do cotovelo.

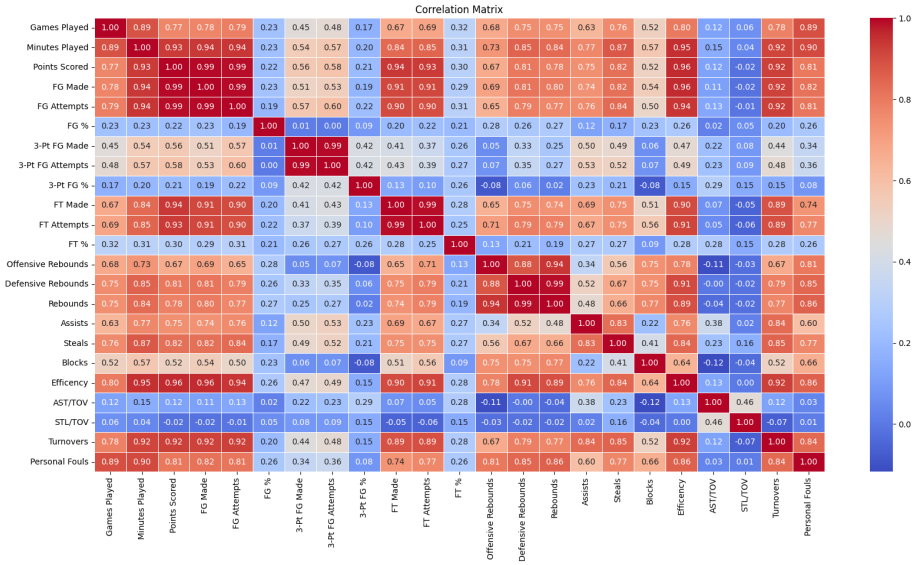
Pelo gráfico abaixo a partir da segunda componente já é explicada quase toda variância, sendo duas componentes suficiente para ter uma descrição dos dados, além de uma maior facilidade de interpretação visual dos dados.



Na primeira análise do PCA é possível notar que as variáveis conseguem separar os jogadores mais bem ranqueados. Além disso, esse ranque pode seguir por três caminhos, todos com a variável eficiência sendo a principal, mas um mais voltado para cestas de dois, outros de três e por fim um voltado rebotes e bloqueios. Esse último principalmente para jogadores na função de pivôs e alas-pivôs.

Além disso, é possível notar que existem muitas variáveis correlacionadas, para futuras análises serão retiradas as que tiverem maior correlação e menor peso no PCA.





A matriz de correlação gerada mostra os coeficientes de correlação entre diferentes variáveis. Os valores na matriz variam de -1 a 1 e indicam a força e direção do relacionamento entre as variáveis.

Aqui estão algumas observações a partir da matriz de correlação:

Há uma forte correlação positiva entre “Games Played” e diversas outras variáveis como “Minutes Played”, “FG Made”, “FG Attempts”, “FT Made” e “FT Attempts”. Isso faz sentido, pois jogadores que jogam mais partidas tendem a

acumular mais minutos e ter mais tentativas e sucessos em arremessos de quadra e lances livres.

Existe uma forte correlação positiva entre “Offensive Rebounds” e “Defensive Rebounds”, o que é esperado, já que ambos contribuem para o número total de rebotes.

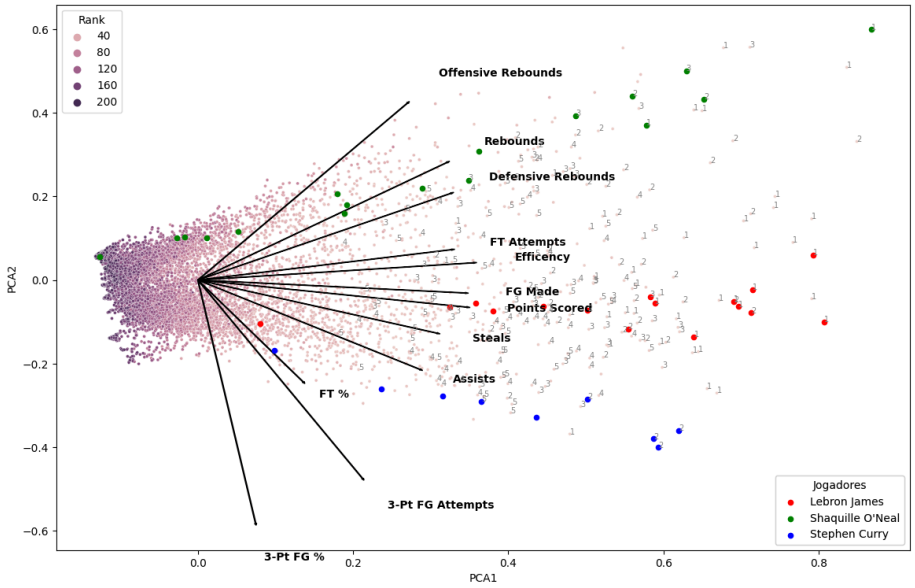
Há uma forte correlação positiva entre “Points Scored” e diversas outras variáveis como “FG Made”, “FT Made” e “Minutes Played”. Isso indica que jogadores que marcam mais pontos também tendem a converter mais arremessos de quadra, lances livres e atuam por mais minutos.

Após a redução de dimensionalidade é possível notar que ainda as estatísticas apresentam uma separação razoável na determinação do rank dos jogadores. Aliás, os maiores ranques são pontos bem distante do centroide. Na figura abaixo foi destacado os top 5 jogadores de cada temporada, além de trazer um exemplo para três jogadres: Lebron James, Shaquille O’Neal e Stephen Curry.

Interessante na avaliação dos três como o Curry se destaca pelas cestas dos três, O’Neal pela presença no garrafão, com elevado número de rebotes e o Lebron com uma das maiores eficiências já vistas na história da NBA.

Em resumo, para definição de um jogador bom é necessário uma maior eficiência como um todo, sendo que essa pode desviar de acordo com a posição e característica dos jogadores, seja sendo um armador com facilidade em cestas de 3 (Curry), um ala-pivô com característica mais equilibradas (Lebron) ou um pivô com uma presença no garrafão (O’Neal).

Por fim, mesmo o ranque utilizado ser de acordo com a pontuação, nota-se que os mais bem ranqueados nesse quesito não se destacam apenas por essa característica.



Para corroborar com a avaliação do PCA foi utilizada uma outra técnica a fim de verificar o poder de separação dos atributos da base, as Curvas de Andrews. Elas são úteis porque permitem identificar quais variáveis têm um maior impacto na separação dos grupos, ajudando a entender a importância relativa de cada característica. Além disso, elas também podem ser usadas para detectar a presença de outliers ou padrões incomuns nos dados.

As Curvas de Andrews são construídas utilizando a série de Fourier para transformar as variáveis originais em uma combinação de funções seno e cosseno. A série de Fourier é uma representação matemática de uma função periódica como uma soma infinita de funções seno e cosseno com diferentes frequências.

A fórmula da série de Fourier utilizada para construir as Curvas de Andrews é a seguinte:

Fórmula de Fourier para Curvas de Andrews

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos(nx) + b_n \sin(nx))$$

Nesta fórmula, f(x) representa a função que descreve a curva de Andre para uma determinada variável. Os coeficientes a_0, a_n e b_n são calculados com base nos dados originais e representam a amplitude e a fase das funções seno e cosseno em diferentes frequências.

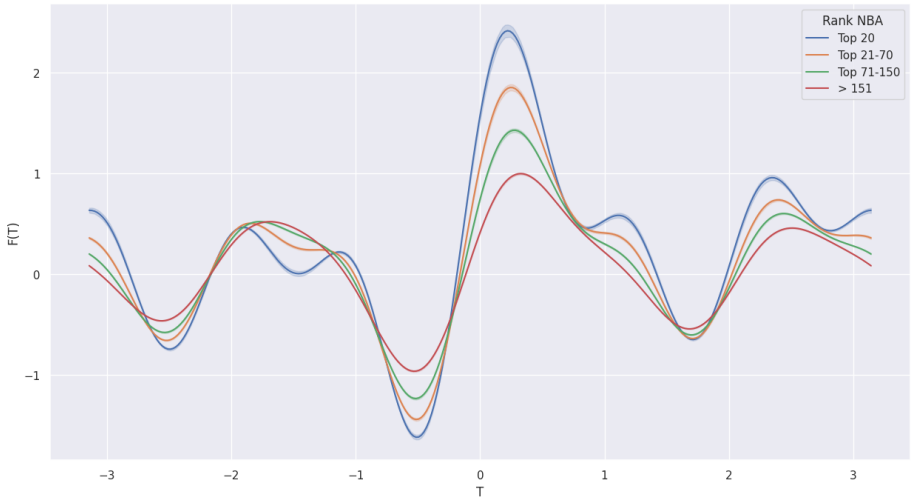
Para cada variável do conjunto de dados, a série de Fourier é aplicada e os coeficientes a_0, a_n e b_n são determinados. Em seguida, as séries de Fourier são somadas para criar a curva de Andrews para cada grupo no conjunto de dados.

Essa abordagem permite representar as variáveis em termos de frequências harmônicas, revelando padrões e relações entre elas. As Curvas de Andrews re-

sultantes são plotadas em um gráfico para visualização e análise da separação entre os grupos.

A interpretação das Curvas de Andrews é baseada na análise da forma das curvas e na distância entre elas. Se as curvas de diferentes grupos estão próximas umas das outras, isso indica que as variáveis têm um poder de separação menor. Por outro lado, se as curvas estão bem separadas, isso sugere que as variáveis têm um alto poder de separação entre os grupos.

Quando é observada a curva, é possível notar que as variáveis selecionaas no PCA apresentam uma boa separação entre os grupos top 20,top 21-70, 71-150 e maior que 150, indicando que esses atributos separam bem não somente os melhores jogadores, como destacados no PCA, mas também entre jogadores menos bem ranqueados.



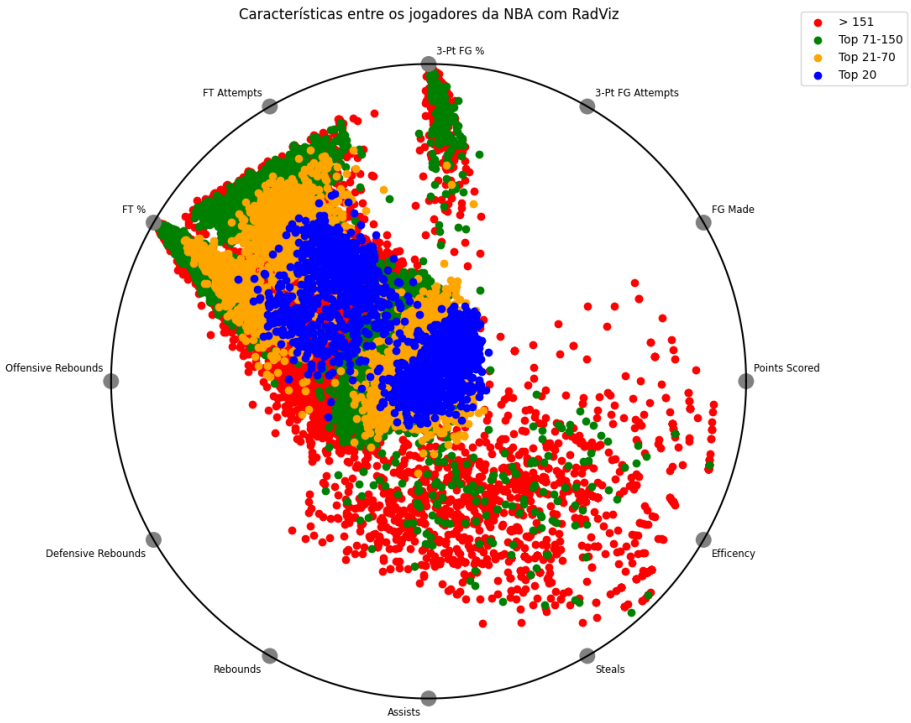
Por fim a última avaliação para os jogadores foi utilizando a técnica Radviz. Ele é um método de visualização utilizado para representar dados multivariados. Ele consiste em um gráfico circular no qual os pontos de dados são dispostos em torno de um círculo, sendo a posição de cada ponto determinada pelas suas variáveis. Cada variável é representada por um eixo radial, e a localização do ponto no gráfico indica a contribuição relativa de cada variável para o ponto de dados. Dessa forma, o Radviz permite identificar padrões e relacionamentos entre as variáveis de forma intuitiva e compacta.

Para o caso da NBA, uma propriedade interessante é quando um jogador se destaque em vários atributos, ele ficará mais ao centro do círculo, enquanto jogadores que se destacam mais por algum atributo específico tenderam a essa âncora.

Quando é observado os top 20 jogadores, é possível notar que estão mais concentrados ao centro, indicando que seus atributos não se limitam a uma habilidade específica, mas a apresentar um resultado razoável em todos os atributos. Além

disso, conforme o ranque vai aumentando, os pontos vão ficando mais dispersos ao longo do círculo, indicando algum valor mais extremo.

Em resumo pelo Radviz, pode-se concluir que para ser bom na NBA não basta ser o cestinha, mas se destacar de maneira considerável em outros quesitos, como defesa, rebotes, variação de arremesos e assistências.

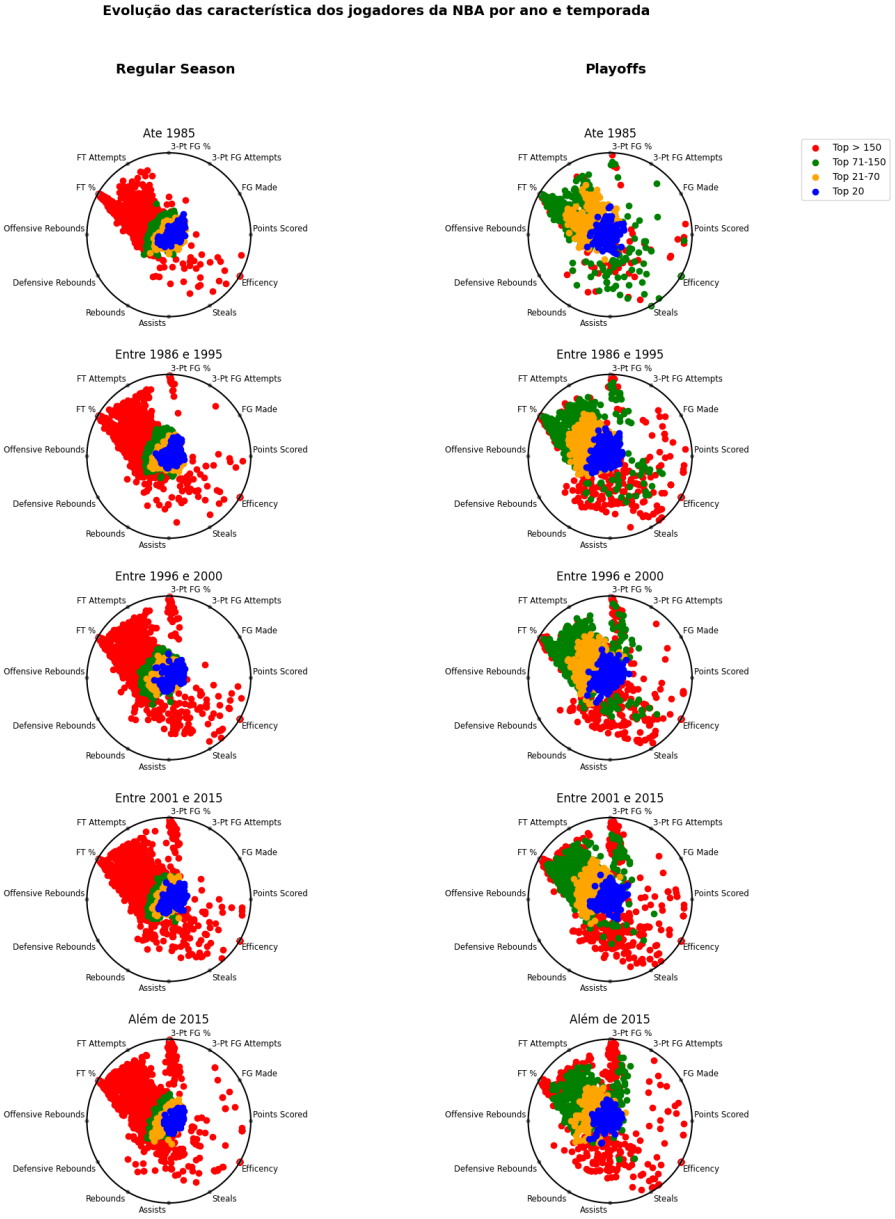


Chapter 5

Evolução do jogo ao longo dos tempos e temporadas.

Ao avaliar o Radviz aberto por ano e temporada (regular e playoffs), é possível obter algumas conclusões:

- Na década de 80/90 os chutes de 3 eram menos frequentes. A variação desses dados é bem menor nesse período. É possível verificar essa análise ao avaliar as âncoras 3-pt FG%, 3Pt FG Attempts e 3-pt FG Made, os quais não apresentam quase nenhum indicador tendendo a eles. Diferentemente das próximas décadas em que há uma dispersão bem maior tendendo a essas âncoras.
- As informações são bem mais dispersas na temporada regular que nos playoffs, logo é mais difícil ter uma separação entre os jogares mais bem ranqueados, principalmente nos top 150.
- Nos playoffs a dispersão dos dados é menor, sendo possível obter uma melhor separação para os top 20 jogadores. Importante notar que alguns extremos são jogadores até top 150, fato que não acontece na temporada regular.
- Os top 20 jogadores da NBA, independente do ano ou da temporada sempre apresentam uma característica bem definida: são constante em todos indicadores, sem apresentar extremos.
- Um exemplo disso é o Stephen Curry, muito conhecido pelos arremessos de 3 pontos, sempre bem ranqueado, porém se mantém próximo ao centro, uma vez que também é bem avaliado nos outros quesitos.


Evolução das característica dos jogadores da NBA por ano e temporada
Regular Season
Playoffs
Ate 1985
Entre 1986 e 1995
Entre 1996 e 2000
Entre 2001 e 2015
Além de 2015
3-Pt FG %
3-Pt FG Attempts
FG Made
Points Scored
Efficency
Steals
Assists
Rebounds
Defensive Rebounds
Offensive Rebounds
FT %
FT Attempts
Top > 150
Top 71-150
Top 21-70
Top 20

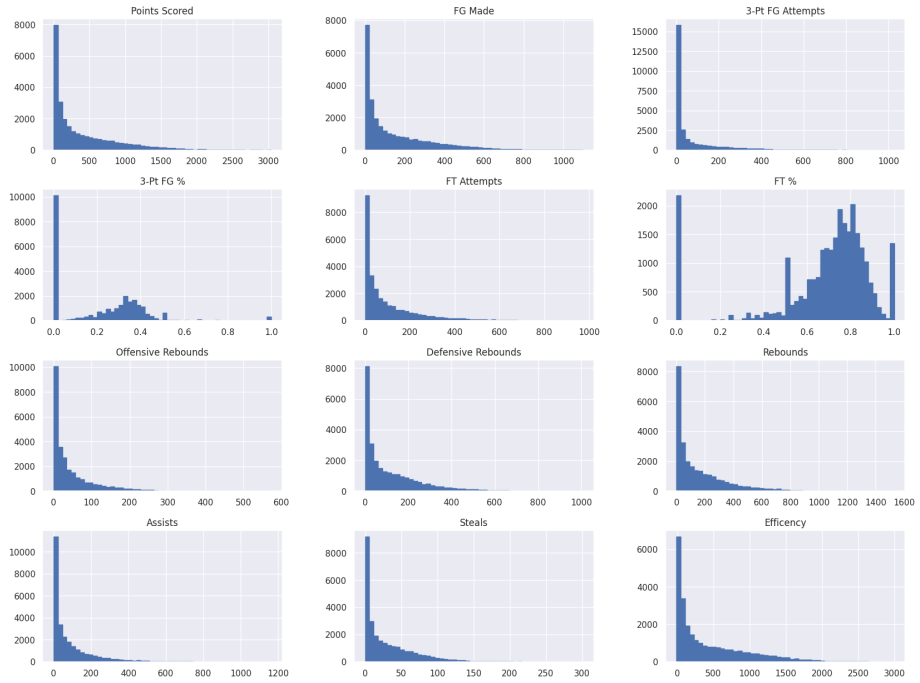
Chapter 6

Principais estatística do jogo

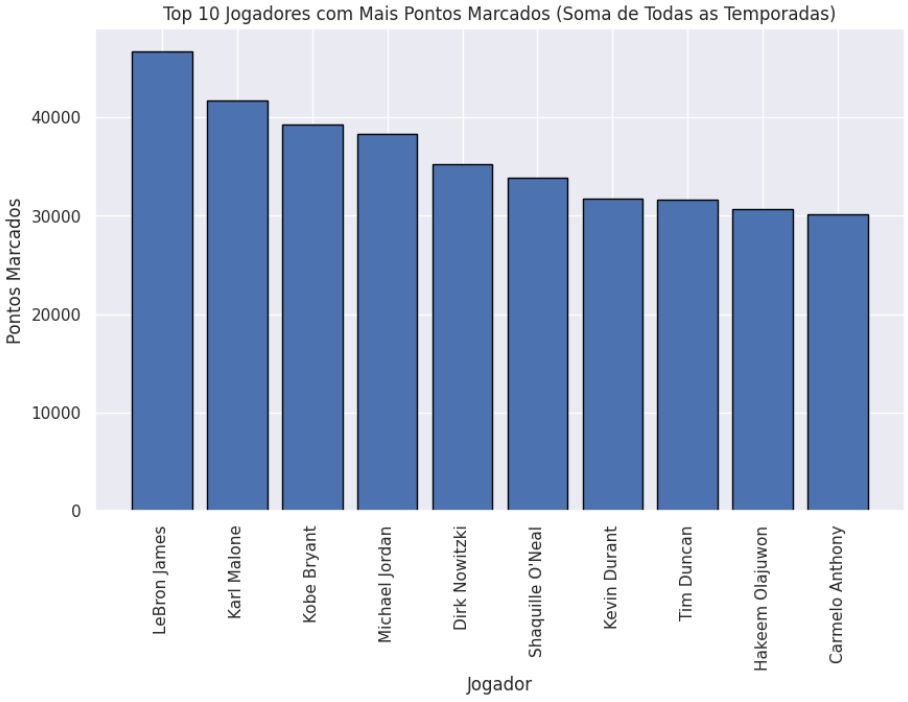
Um maneira simples para entender melhor como um jogo de basquete funciona e como são os jogadores é avaliar as distribuições dos principais atributos. Quando os atributos são de características como, marcou mais pontos, arremesou mais, ou mais roubadas de bola, todos apresentam um formato exponencial, indicando que a tendência é que ao aumentar a frequência deles, uma menor quantidade de jogadores conseguem tal marca.

Além disso, é possível interpretar da forma que essa curva decai. Ao comparar arremessos de três frente a arremessos de dois pontos, a curva de três apresenta uma queda muito mais acentuada, indicando que é um atributo com menor frequência (ou de maior dificuldade). A mesma analogia pode ser aplicada para rebotes ofensivos e defensivos, sendo o primeiro muito menos frequente.

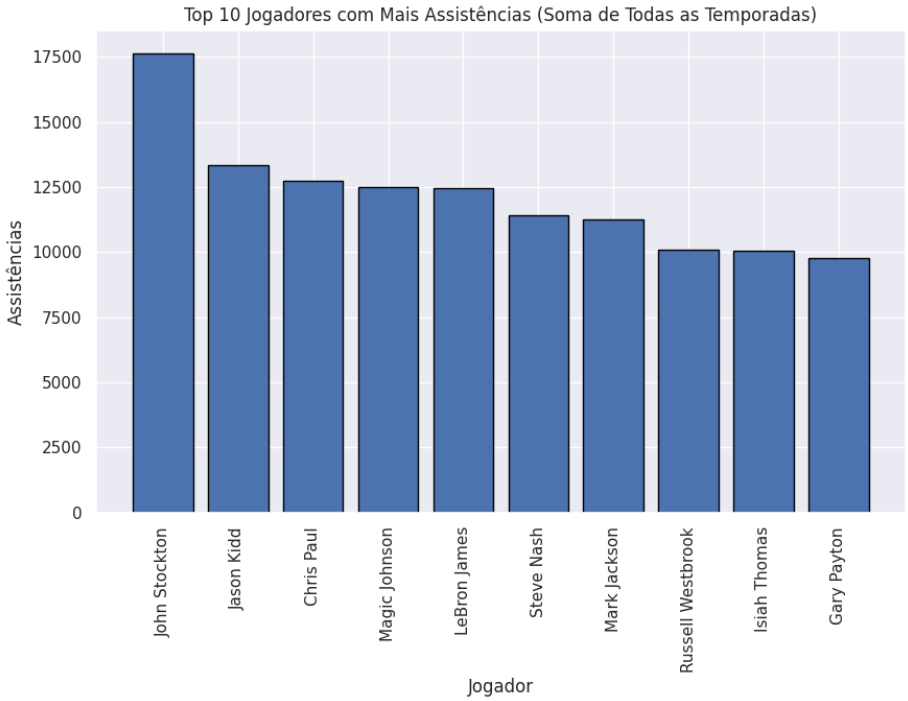
Por fim, os atributos de percentual de acerto de três pontos ou lances livres se assemelham a uma normal, indicando a característica do percentual de acerto de cada jogador. Importante destacar que a média de acertos em cestas de três é em torno de 35%, enquanto de lances livres é de 80%, evidenciando a maior dificuldade no acerto em cestas de três.



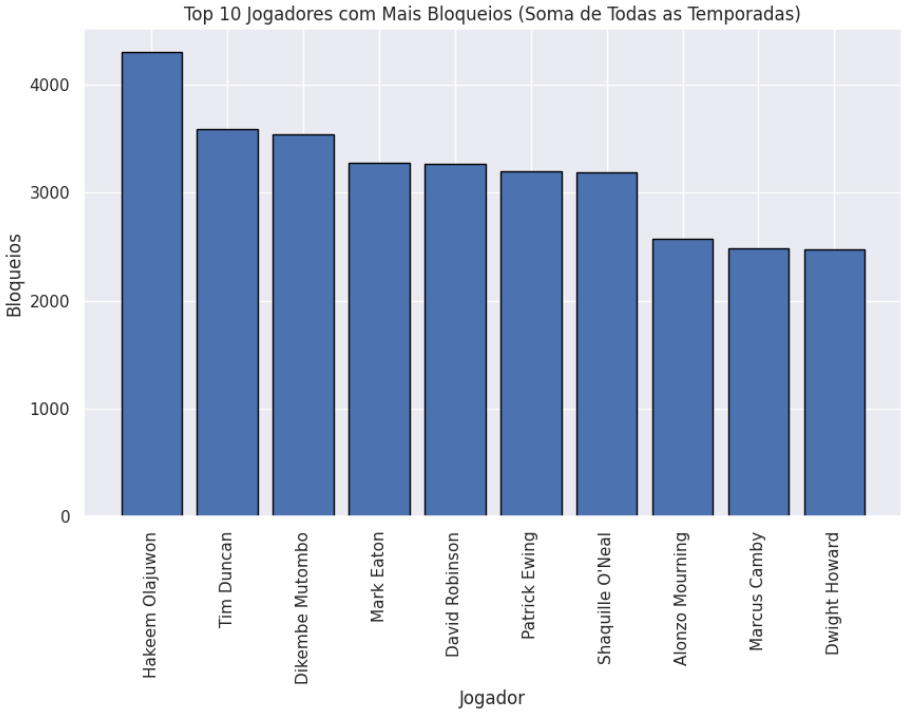
Principais pontuadores de todos os tempos



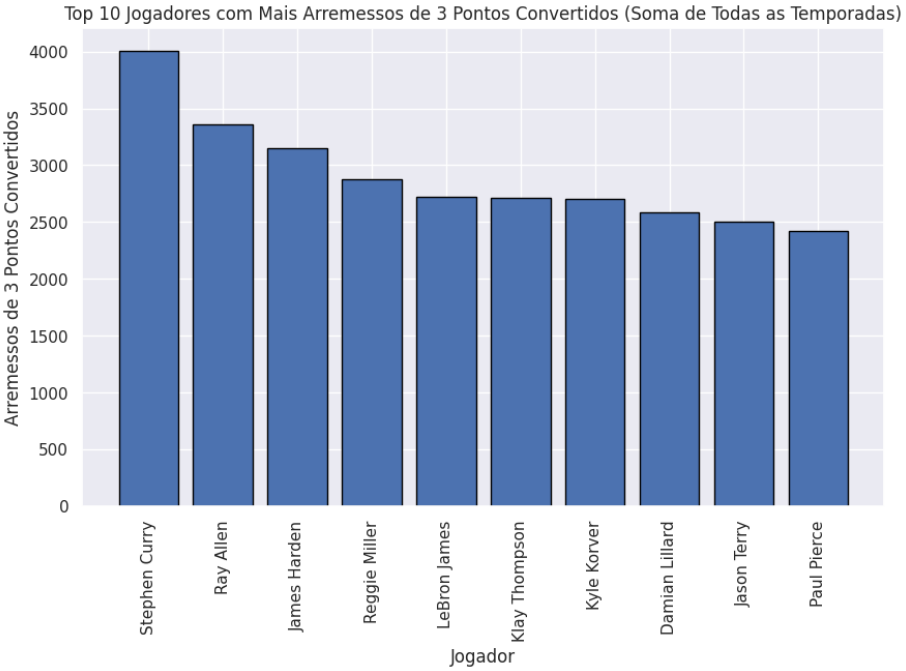
Principais “garçons” de todos os tempos



Principais bloqueadores de todos os tempos



Principais pontuadores dos três de todos os tempos



Chapter 7

O que define um time bom?

7.1 Chernoff Faces

Últimos campeões e vices da NBA:

2013 - San Antonio Spurs | Miami Heat

2014 - Golden State **Warriors** | Cleveland Cavaliers

2015 - Cleveland Cavaliers | Golden State **Warriors**

2016 - Golden State **Warriors** | Cleveland Cavaliers

2017 - Golden State **Warriors** | Cleveland Cavaliers

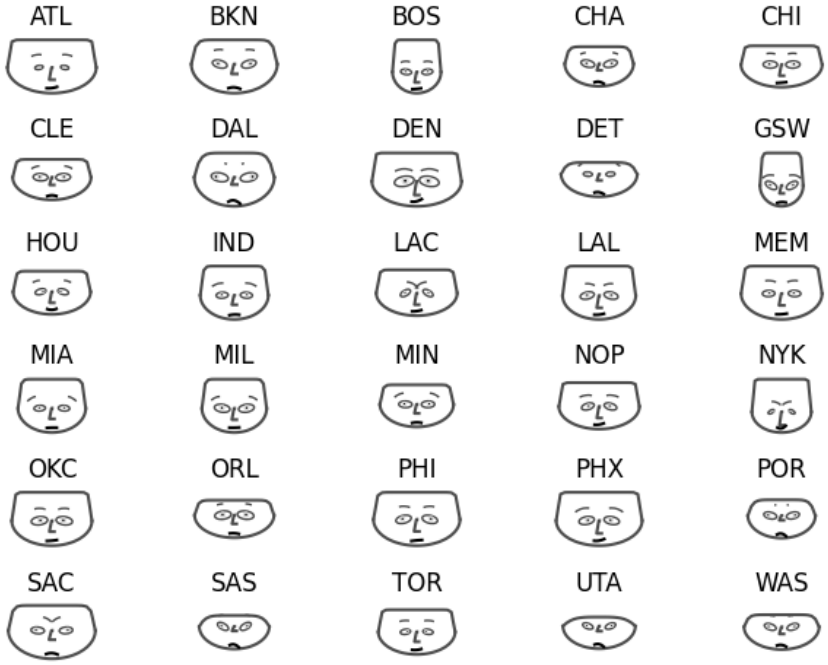
2018 - Toronto Raptors | Golden State **Warriors**

2019 - Los Angeles Lakers | Miami Heat

2020 - Milwaukee Bucks | Phoenix Suns

2021 - Golden State **Warriors** | Boston Celtics

2022 - Denver Nuggets | Miami Heat

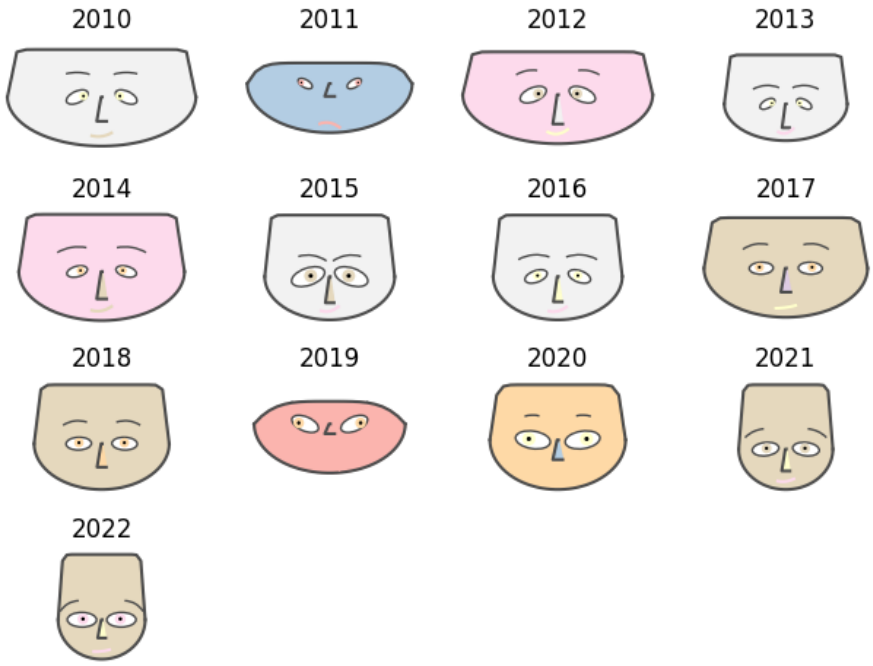

ATL
BKN
BOS
CHA
CHI
CLE
DAL
DEN
DET
GSW
HOU
IND
LAC
LAL
MEM
MIA
MIL
MIN
NOP
NYK
OKC
ORL
PHI
PHX
POR
SAC
SAS
TOR
UTA
WAS

Eastern Conference | Western Conference

Team	W	L	Pct	GB	Conf	Home	Away	L10	Strk
1 Bucks	58	24	.707	-	35-17	32-9	26-15	6-4	L2
2 Celtics	57	25	.695	1.0	34-18	32-9	25-16	8-2	W3
3 76ers	54	28	.659	4.0	34-18	29-12	25-16	5-5	W2
4 Cavaliers	51	31	.622	7.0	34-18	31-10	20-21	7-3	L1
5 Knicks	47	35	.573	11.0	32-20	23-18	24-17	5-5	L2
6 Nets	45	37	.549	13.0	30-22	23-18	22-19	6-4	L1
7 Hawks	41	41	.500	17.0	26-26	24-17	17-24	5-5	L2
8 Heat	44	38	.537	14.0	24-28	27-14	17-24	6-4	W1
9 Raptors	41	41	.500	17.0	26-26	27-14	14-27	6-4	W1
10 Bulls	40	42	.488	18.0	27-25	22-19	18-23	6-4	W2
11 Pacers	35	47	.427	23.0	24-28	20-21	15-26	3-7	W1
12 Wizards	35	47	.427	23.0	21-31	19-22	16-25	3-7	L1
13 Magic	34	48	.415	24.0	20-32	20-21	14-27	5-5	L4
14 Hornets	27	55	.329	31.0	15-37	13-28	14-27	5-5	W1
15 Pistons	17	65	.207	41.0	8-44	9-32	8-33	1-9	L1

Eastern Conference | **Western Conference**

Team	W	L	Pct	GB	Conf	Home	Away	L10	Strk
1 Nuggets	53	29	.646	-	34-18	34-7	19-22	5-5	W1
2 Grizzlies	51	31	.622	2.0	30-22	35-6	16-25	6-4	L1
3 Kings	48	34	.585	5.0	32-20	23-18	25-16	5-5	L3
4 Suns	45	37	.549	8.0	30-22	28-13	17-24	7-3	L2
5 Clippers	44	38	.537	9.0	27-25	23-18	21-20	6-4	W3
6 Warriors	44	38	.537	9.0	30-22	33-8	11-30	8-2	W3
7 Lakers	43	39	.524	10.0	27-25	23-18	20-21	8-2	W2
8 Timberwolves	42	40	.512	11.0	29-23	22-19	20-21	7-3	W3
9 Pelicans	42	40	.512	11.0	29-23	27-14	15-26	7-3	L1
10 Thunder	40	42	.488	13.0	25-27	24-17	16-25	4-6	W2
11 Mavericks	38	44	.463	15.0	28-24	23-18	15-26	2-8	L2
12 Jazz	37	45	.451	16.0	24-28	23-18	14-27	2-8	L1
13 Trail Blazers	33	49	.402	20.0	23-29	17-24	16-25	1-9	L4
14 Rockets	22	60	.268	31.0	12-40	14-27	8-33	4-6	W3
15 Spurs	22	60	.268	31.0	10-42	14-27	8-33	3-7	W1



Últimas posições na temporada regular do Golden State Warriors:

2010 - 3º

2011 - 4º

2012 - 2º

2013 - 2º

2014 - 1º

2015 - 1º

2016 - 1º

2017 - 1º

2018 - 1º

2019 - 5º

2020 - 4º

2021 - 2º

2022 - 4º

Concluindo, é possível observar tendencias de padrão e comportamento semelhantes em anos ou times que são parecidos, porém devido a natureza do gráfico

e a falta de informação, principalmente para o "usuário" final, sobre os componentes representados por cada item (tamanho dos olhos, posição, inclinação da sobrancelha, etc.), não é possível fazer análises mais profundas.

Vemos como um gráfico de apoio para facilitar algumas visualizações ou fomentar questionamentos, por exemplo: os gráficos de 2015 e 2017 são significativamente diferentes, apesar do time ter sido campeão em ambas as temporadas. O desempenho nesses anos foi muito diferente? Ou então, os Pacers (IND) e os Wizards (WAS) tiveram o mesmo aproveitamento na temporada (0.427), porém seus gráficos são distintos. O que será que aconteceu? Sabemos que o resultado final de um jogo é apenas vitória ou derrota, porém a margem de pontos pode ser de 1 ou de 20, bem como a quantidade de arremesos, roubadas de bola, entre outros. Um time pode ter tido um desempenho comparável a times pior classificados e ter tido "sorte" de ganhar a quantidade suficiente de jogos, mesmo que por uma margem pequena e perder por muito, enquanto outro sempre tinha jogos parelhos, porém perdeu em mais ocasiões.