

Relatório do Projeto de Data Mining

“Análise de dados da NBA em PowerBI”



Realizado por:

Eduardo Palma nº201900054;

Nuno Melo nº201700465;

Ricardo Santos nº201700524;

Docente: Ana Mendes

Unidade Curricular: Data Mining

Curso: Licenciatura em Bioinformática, 2º ano

Ano Letivo: 2020/2021

Instituição: Escola Superior de Tecnologia do Barreiro

Data: 21 de junho de 2021

Índice

1. Introdução	1
2. Apresentação do Caso de Estudo.....	1
3. Objetivos	2
4. Transformação dos Dados	3
4.1 <i>Dataset dos Games</i>	3
4.2 <i>Dataset Game Details</i>	5
5. Análise dos Dados	7
6. Dificuldades Sentidas.....	18
7. Conclusão	18
8. Bibliografia	20

Índice de Figuras

Figura 1 – Dataset dos Games Inicial	3
Figura 2 – Remoção da coluna GAME_STATUS_TEXT	3
Figura 3 – Substituição de Valores.....	4
Figura 4 – Alterou-se os tipos dos atributos	4
Figura 5 – Janela para trocar o tipo dos dados	4
Figura 6 – Valores em branco e nulos.....	5
Figura 7 – Filtro de seleção de dados	6
Figura 8 – Dataset do Game Details	6
Figura 9 – Gráfico de Barras dos Pontos Casa/Fora por Equipa	7
Figura 10 – Medida para calcular a diferença de Pontos Casa/Fora.....	7
Figura 11 – Gráfico de Barras da Diferença de Pontos Casa/Fora	8
Figura 12 – Gráfico de Barras da Percentagem de Vitórias	9
Figura 13 – Gráfico de Barras do Número de Épocas de cada Jogador	10
Figura 14 – Gráfico de Barras do Total de Jogos Disputados por Jogador	10
Figura 15 – Gráfico de Barras da Contagem de Ressaltos	11
Figura 16 - Média dos Ressaltos, Ressaltos Ofensivos e Ressaltos Defensivos	11
Figura 17 – Gráfico Circular das Médias dos diferentes tipos de Pontos	12
Figura 18 – Funil do Máximo de Bloqueios e Desarmes	13
Figura 19 – Gráfico de Dispersão dos Pontos versus Faltas	13
Figura 20 – Medida para calcular os Pontos	14
Figura 21 – Medida para calcular as Assistências	14
Figura 22 – Gráfico de Barras das Assistências e Pontos por Equipa	14
Figura 23 – Medida para calcular as Vitórias por Ano.....	15
Figura 24 – Gráfico de Barras da Média de Vitórias das Equipas	15
Figura 25 – Medida para calcular as Derrotas por Ano	15
Figura 26 – Gráfico de Barras da Média de Derrotas das Equipas	15
Figura 27 – Gráfico de linhas da Previsão da Média de Vitórias do Chicago Bulls	16
Figura 28 – Gráfico de linhas da Previsão da Média de Derrotas do Chicago Bulls	17

1. Introdução

Este projeto final que foi atribuído ao grupo, teve como principal objetivo utilizar uma ferramenta de análise de negócios da empresa *Microsoft Corporation* que é denominada por PowerBI. Este programa é concebido para fornecer visualizações e recursos de *business intelligence*, com a sua interface gráfica simples para que qualquer tipo de utilizador possa criar os seus próprios relatórios e *dashboards*. O grupo teve de escolher um *dataset* para o analisar no PowerBI sendo o escolhido retirado da plataforma Kaggle, com a temática de NBA (*National Basketball Association*).

2. Apresentação do Caso de Estudo

O conjunto dos cinco *dataset* escolhidos foi, como referido anteriormente, sobre os jogos da liga de basquetebol americano profissional NBA, com dados desde 2004 até ao último de 2020.

Os cinco *datasets* em uso estão repartidos em o “*Game*”, que contém dados específicos dos jogos que foram realizados ao longo dos anos, como por exemplo, o número de vitórias e o número de derrotas, o “*Game details*”, que dispõe de uma maior granularidade de dados estatísticos de cada jogador em cada um dos jogos que disputaram, como o número de pontos marcados, assistências, ressaltos, etc., o “*Players*”, contém informações sobre os jogadores de cada equipa que jogaram a nível profissional, o “*Teams*”, que possui os dados das equipas e dos seus donos, e por fim, o “*Ranking*”, que contém informação sobre as vitórias e derrotas de cada equipas por época.

3. Objetivos

Os principais objetivos que se propôs para a realização deste projeto foram:

- Aplicar as técnicas aprendidas no decorrer da unidade curricular de *Data Mining*;
- Utilização de ferramentas de tratamento de dados, nomeadamente, o *Power Query* do *PowerBI*, de forma a ter os dados prontos a utilizar;
- Interpretação dos gráficos obtidos de forma a retirar informação relevante.

4. Transformação dos Dados

Primeiro procede-se à transformação dos dados em cada um dos cinco *datasets*. Vai-se utilizar o *dataset* “Games”, *Figura 1*.

4.1 Dataset dos Games

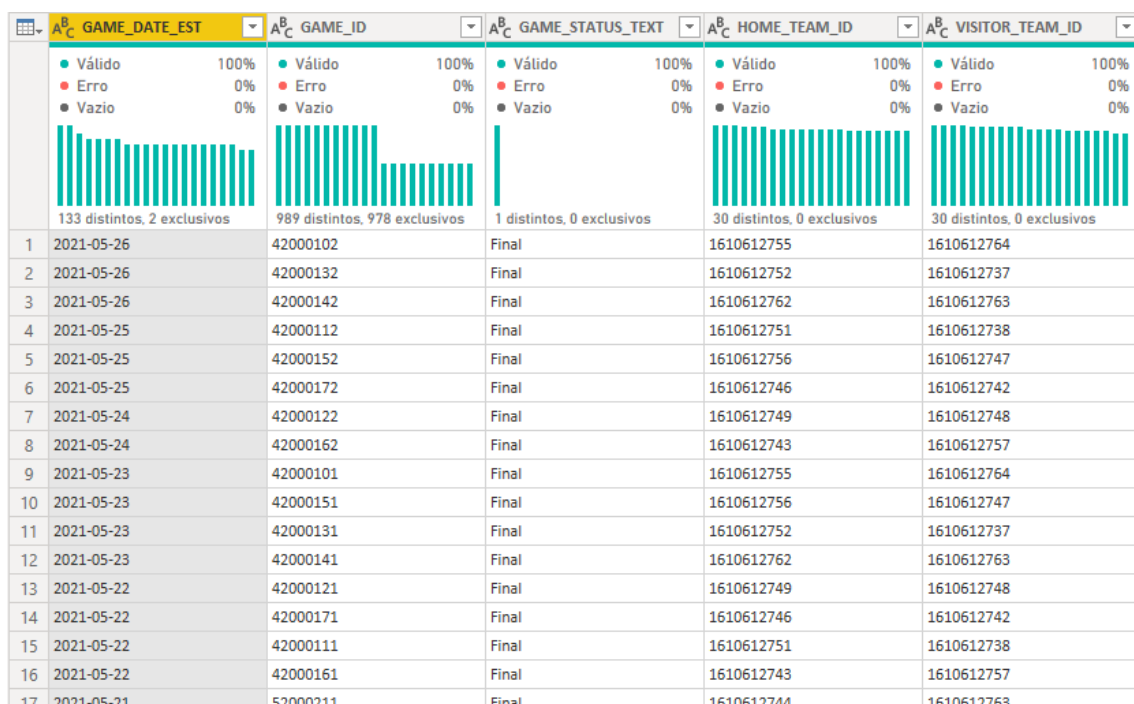


Figura 1 – Dataset dos Games Inicial

```
= Table.RemoveColumns("#Replaced Value",{"GAME_STATUS_TEXT"})
```

Figura 2 – Remoção da coluna GAME STATUS TEXT

Como mostra a *Figura 2*, removeu-se a coluna *GAME_STATUS_TEXT* pois apresentava sempre o mesmo valor e não seria útil para a análise em estudo.

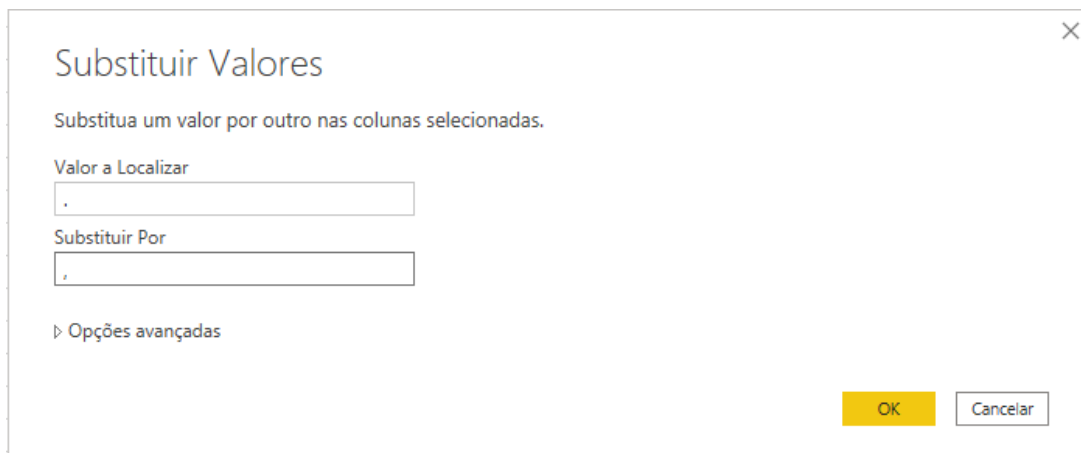


Figura 3 – Substituição de Valores

Substitui-se todos os pontos do *dataset* por vírgulas para, em seguida, alterar-se os tipos de dados de cada uma das colunas, Figura 3.

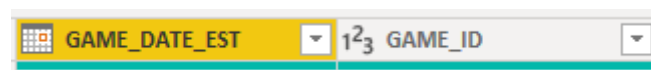


Figura 4 – Alterou-se os tipos dos atributos

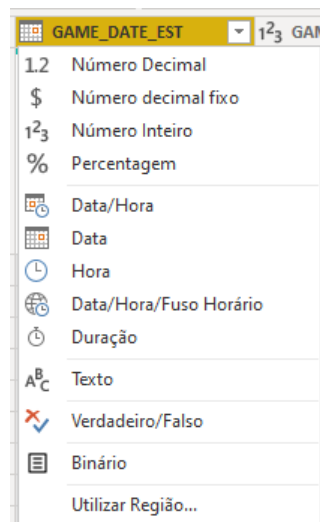


Figura 5 – Janela para trocar o tipo dos dados

Para cada uma das colunas do *dataset* repete-se o passo anterior para alterar para o tipo de dados correto, Figura 4 e 5.

As transformações aplicadas neste *dataset* foram também aplicadas aos *datasets* “Game details”, “Players”, “Ranking” e “Teams”, pelo que não serão apresentadas neste relatório.

4.2 Dataset Game Details

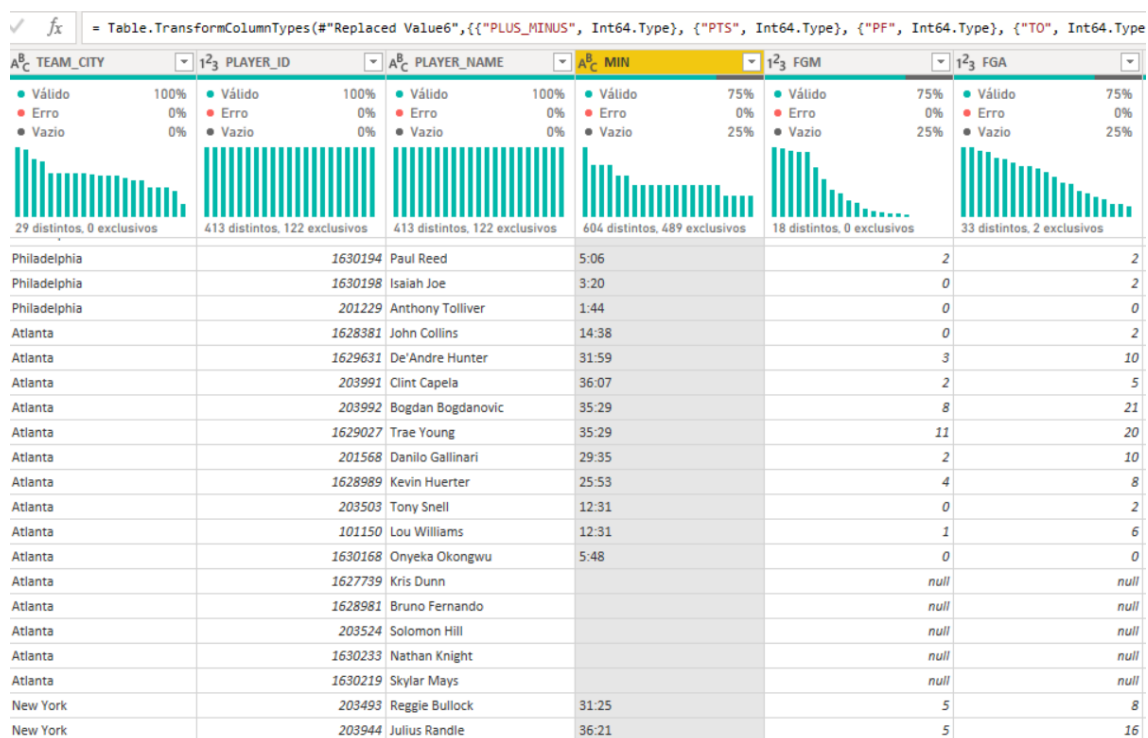


Figura 6 – Valores em branco e nulos

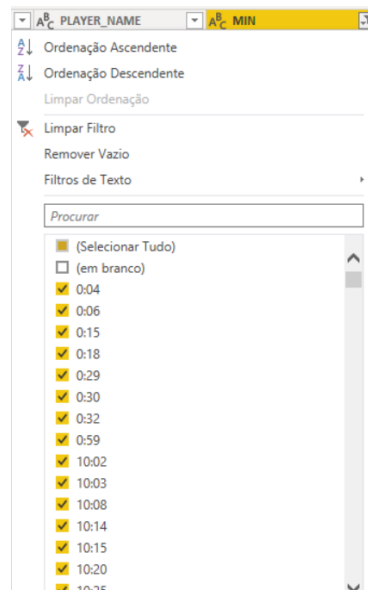


Figura 7 – Filtro de seleção de dados

Na presença de espaços vazios, verifica-se que as linhas que os continham eram de jogadores que nunca tinham jogado nenhum minuto em nenhum jogo, por isso inseriu-se um filtro onde remove-se as linhas em branco, Figura 6 e 7.

1 ² 3 GAME_ID	1 ² 3 TEAM_ID	A ^B TEAM_ABBREVIATION	A ^B TEAM_CITY	1 ² 3 PLAYER_ID	A ^B PLAYER_NAME
<div> <div>Válido 100%</div> <div>Erro 0%</div> <div>Vazio 0%</div> </div> <div>47 distintos, 0 exclusivos</div>	<div> <div>Válido 100%</div> <div>Erro 0%</div> <div>Vazio 0%</div> </div> <div>30 distintos, 0 exclusivos</div>	<div> <div>Válido 100%</div> <div>Erro 0%</div> <div>Vazio 0%</div> </div> <div>30 distintos, 0 exclusivos</div>	<div> <div>Válido 100%</div> <div>Erro 0%</div> <div>Vazio 0%</div> </div> <div>30 distintos, 0 exclusivos</div>	<div> <div>Válido 100%</div> <div>Erro 0%</div> <div>Vazio 0%</div> </div> <div>378 distintos, 85 exclusivos</div>	<div> <div>Válido 100%</div> <div>Erro 0%</div> <div>Vazio 0%</div> </div> <div>378 distintos, 85 exclusivos</div>
1	42000102	1610612764 WAS	Washington	203078	Bradley Beal
2	42000102	1610612764 WAS	Washington	1629060	Rui Hachimura
3	42000102	1610612764 WAS	Washington	203458	Alex Len
4	42000102	1610612764 WAS	Washington	203526	Raul Neto
5	42000102	1610612764 WAS	Washington	201566	Russell Westbrook
6	42000102	1610612764 WAS	Washington	202722	Davis Bertans
7	42000102	1610612764 WAS	Washington	202397	Ish Smith
8	42000102	1610612764 WAS	Washington	1629655	Daniel Gafford
9	42000102	1610612764 WAS	Washington	1628990	Chandler Hutchison
10	42000102	1610612764 WAS	Washington	201577	Robin Lopez
11	42000102	1610612764 WAS	Washington	1630264	Anthony Gill
12	42000102	1610612764 WAS	Washington	1629726	Garrison Mathews
13	42000102	1610612764 WAS	Washington	1630216	Cassius Winston
14	42000102	1610612764 WAS	Washington	1629067	Isaac Bonga
15	42000102	1610612755 PHI	Philadelphia	201980	Danny Green
16	42000102	1610612755 PHI	Philadelphia	202699	Tobias Harris
17	42000102	1610612755 PHI	Philadelphia	203954	Joel Embiid
18	42000102	1610612755 PHI	Philadelphia	203552	Seth Curry
19	42000102	1610612755 PHI	Philadelphia	1627732	Ben Simmons
20	42000102	1610612755 PHI	Philadelphia	1629680	Matisse Thybulle
21	42000102	1610612755 PHI	Philadelphia	201588	George Hill

Figura 8 – Dataset do Game Details

Depois de aplicadas as transformações em todos os *datasets* verifica-se que em nenhuma coluna existem valores em falta ou erros, por isso está-se pronto para começar a analisar os dados, Figura 8.

5. Análise dos Dados

Nesta etapa do projeto foram criados vários gráficos para serem analisados. Nas seguintes páginas estes serão apresentados com uma breve descrição.

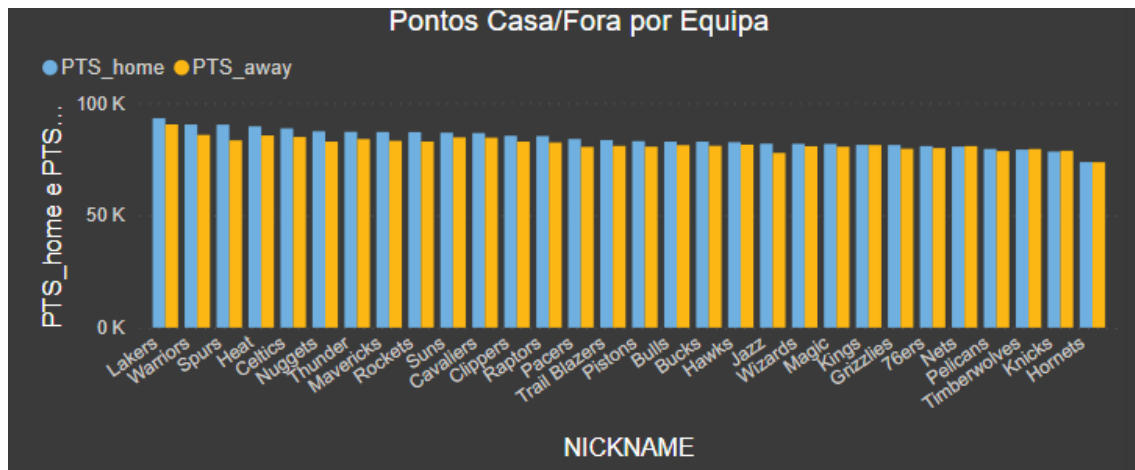


Figura 9 – Gráfico de Barras dos Pontos Casa/Fora por Equipa

Como se consegue visualizar na Figura 9, existe uma pequena diferença entre os pontos marcados nos estádios pertencentes a cada equipa dos pontos marcados de quando jogam fora.

```
1 DifferenceBetweenPointsHomeAway = SUMX('Games', 'Games'[PTS_home] - 'Games'[PTS_away])
```

Figura 10 – Medida para calcular a diferença de Pontos Casa/Fora

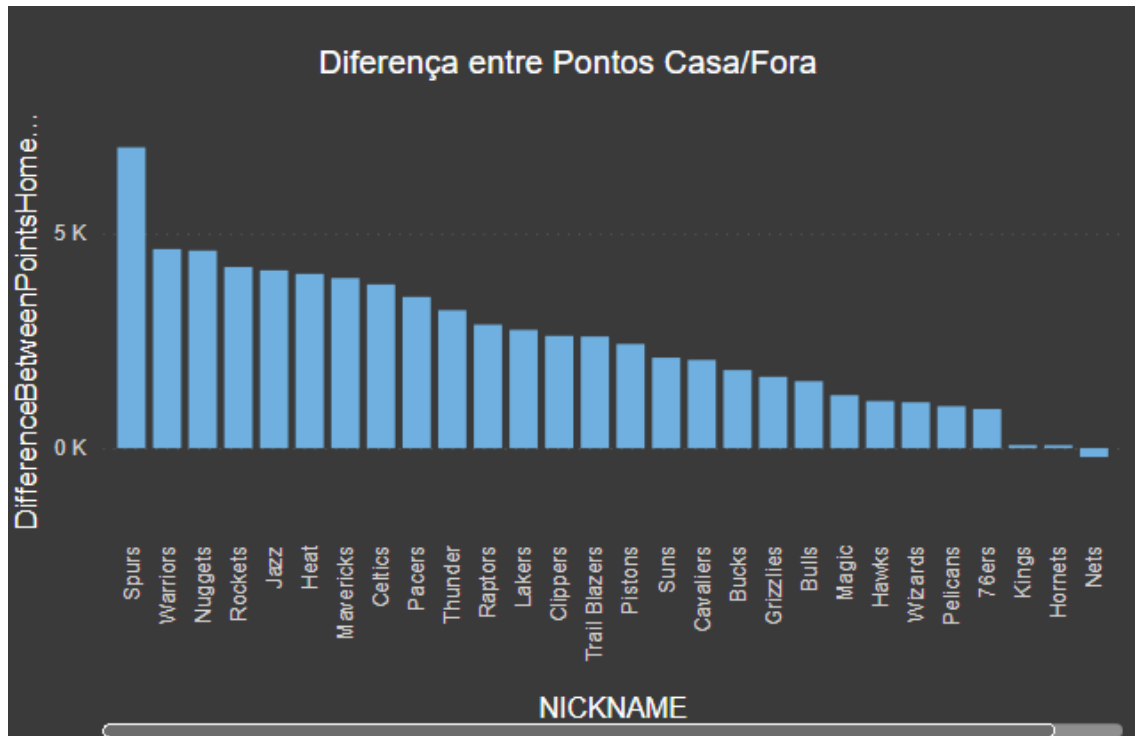


Figura 11 – Gráfico de Barras da Diferença de Pontos Casa/Fora

Após a análise do gráfico da relação dos pontos marcados em casa/fora pelas equipas, Figura 9, procedeu-se à criação de uma medida para calcular a diferença entre os pontos marcados quando jogam no seu “território” *versus* quando jogam fora, Figura 10.

O gráfico criado a partir desta medida, Figura 11, pode levar a uma análise posterior para se identificar quais são as equipas que tem um melhor desempenho quando jogam em casa.

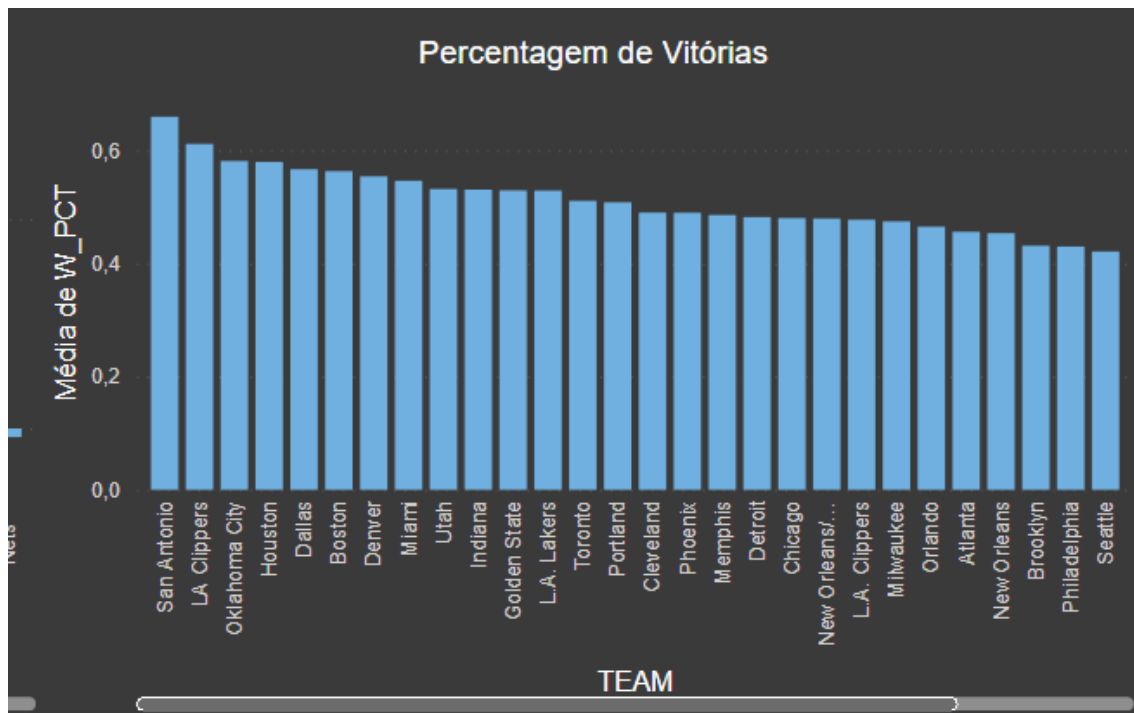


Figura 12 – Gráfico de Barras da Percentagem de Vitórias

No gráfico de “Percentagem de Vitórias”, Figura 12, observa-se as equipas que têm uma maior percentagem de vitórias, o que pode levar a um estudo de previsão de desempenho para as próximas épocas.

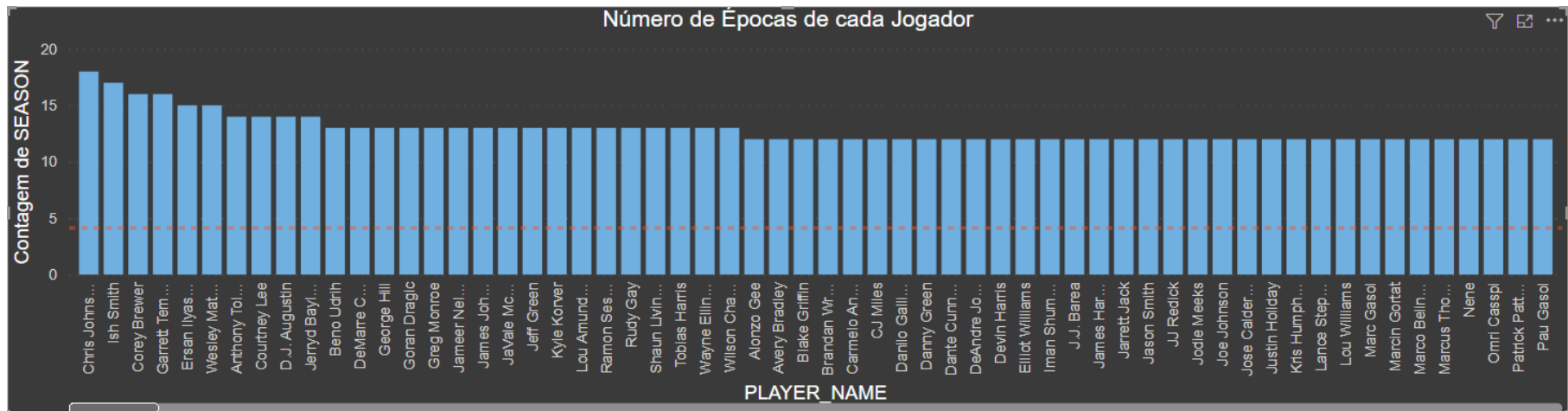


Figura 13 – Gráfico de Barras do Número de Épocas de cada Jogador

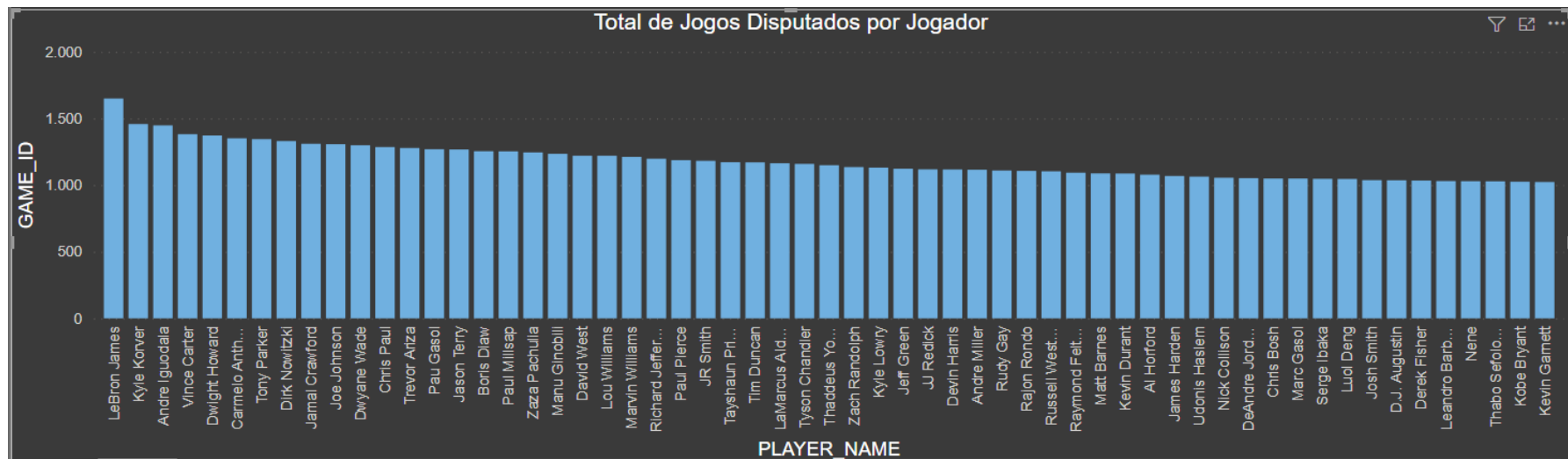


Figura 14 – Gráfico de Barras do Total de Jogos Disputados por Jogador

Foi feita também uma análise aos jogadores, em específico, ao número de épocas e ao número de jogos de cada jogador, Figura 13 e 14.

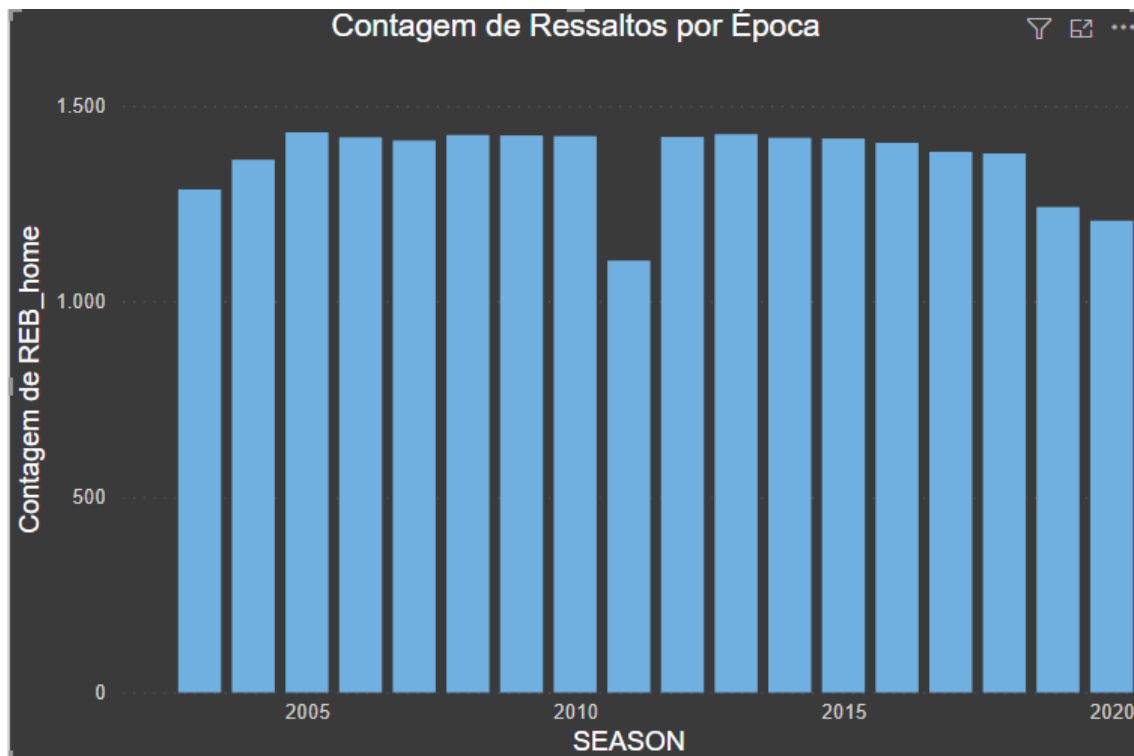


Figura 15 – Gráfico de Barras da Contagem de Ressaltos

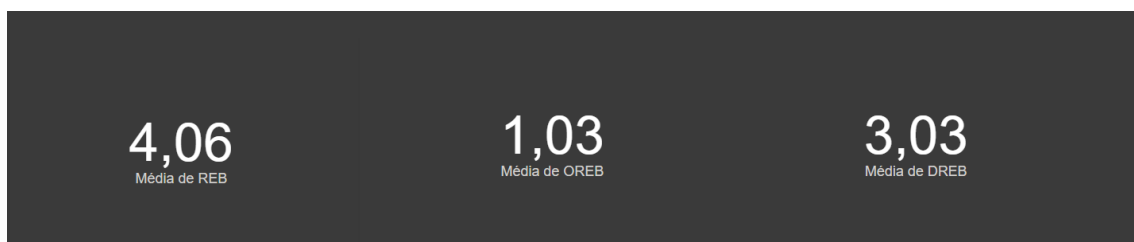


Figura 16 - Média dos Ressaltos, Ressaltos Ofensivos e Ressaltos Defensivos

Os ressaltos são considerados um dos fatores mais importantes para as equipas atingirem a vitória, por isso, prosseguiu-se a uma análise dos mesmos por época, Figura 15. E de seguida, fez-se uma pequena análise para saber a média de ressaltos ofensivos e defensivos, Figura 16.

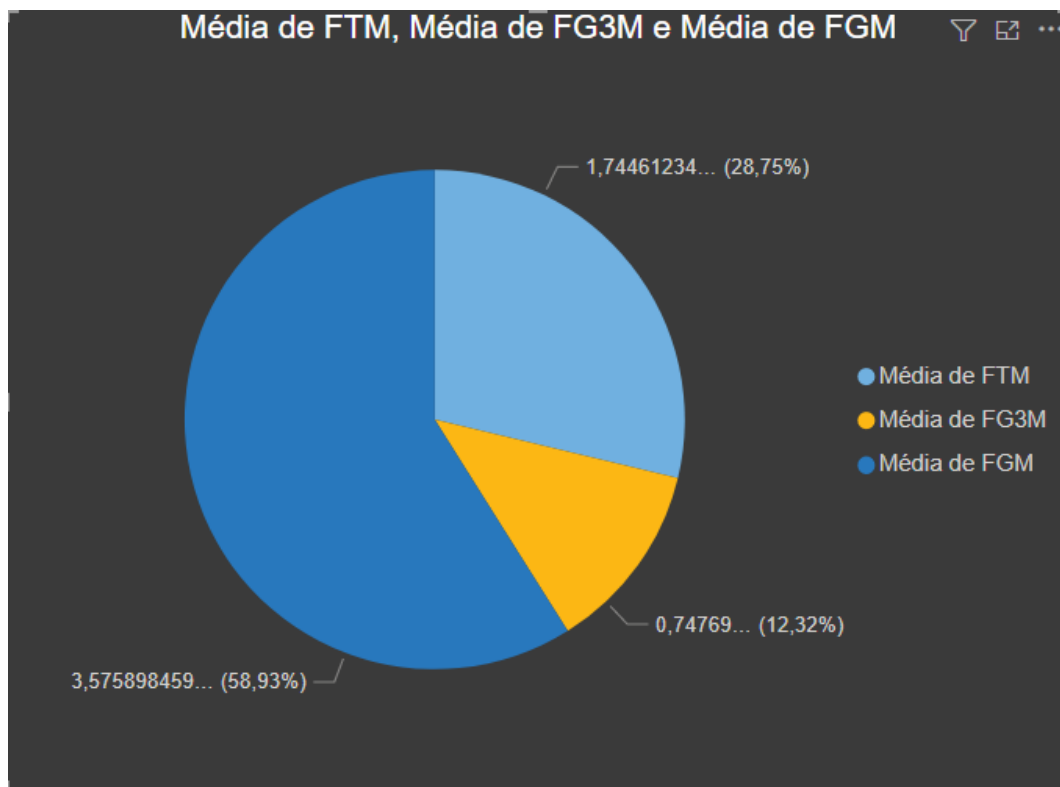


Figura 17 – Gráfico Circular das Médias dos diferentes tipos de Pontos

Na Figura 17, está representada a média de lançamentos ao cesto bem-sucedidos, divididos em lançamentos livres marcados (FTM – 1 ponto), lançamentos marcados (FGM – 2 pontos) e os lançamentos triplos (FG3M – 3 pontos).

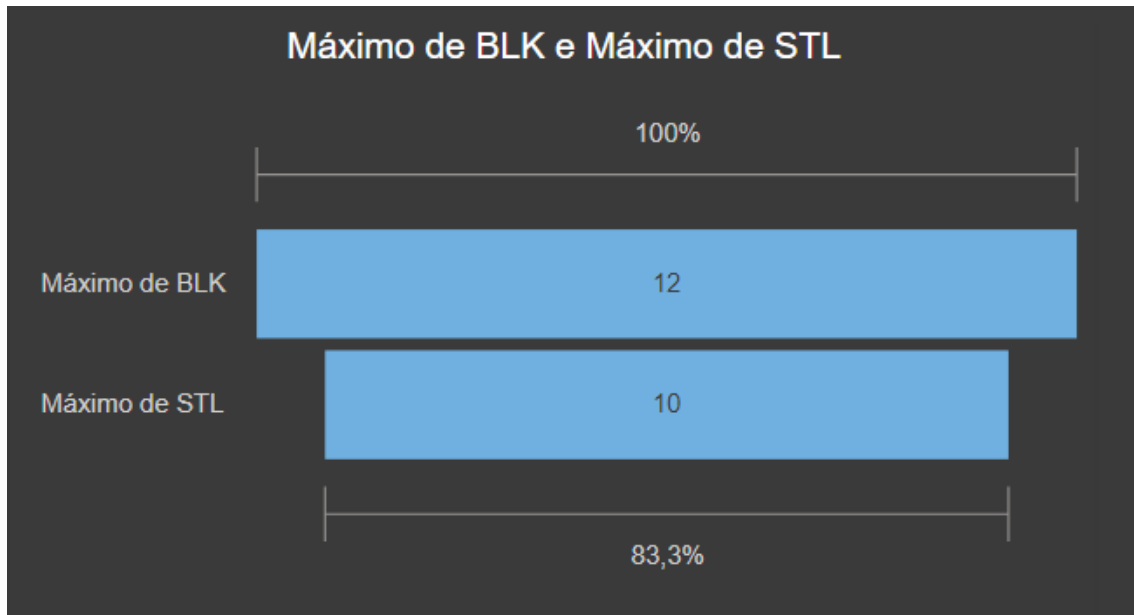


Figura 18 – Funil do Máximo de Bloqueios e Desarmes

Para além de marcar pontos também é importante saber defender, por isso utiliza-se os bloqueios (BLK) e os desarmes (STL) para impedir que a equipa adversária tenha vantagem, logo verificou-se qual o máximo de bloqueios e desarmes realizados num único jogo, Figura 18.

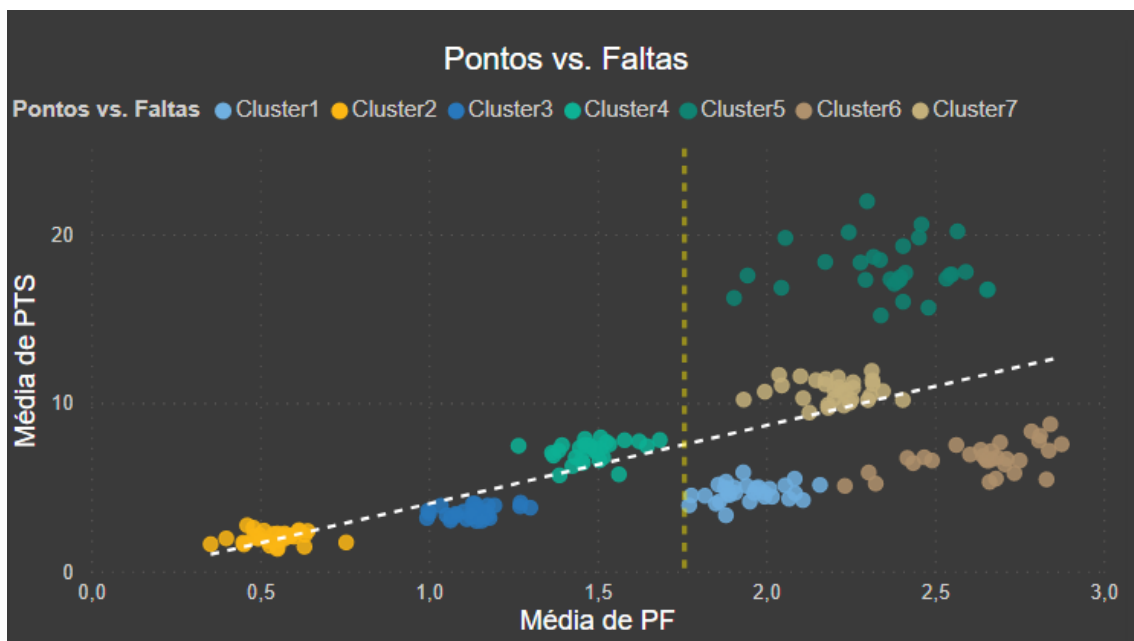


Figura 19 – Gráfico de Dispersão dos Pontos versus Faltas

Criou-se um gráfico de Dispersão, Figura 19, para identificar o conjunto de equipas que num jogo ao cometerem uma falta pessoal, essa falta contribuiu para que elas consigam marcar pontos.

```
1 POINTS = CALCULATE(COUNT('Game details'[PTS]), FILTER('Game details', 'Game details'[TEAM_ID]= 'Teams'[TEAM_ID]))
```

Figura 20 – Medida para calcular os Pontos

```
1 Assists = CALCULATE(SUM('Game details'[AST]))
```

Figura 21 – Medida para calcular as Assistências

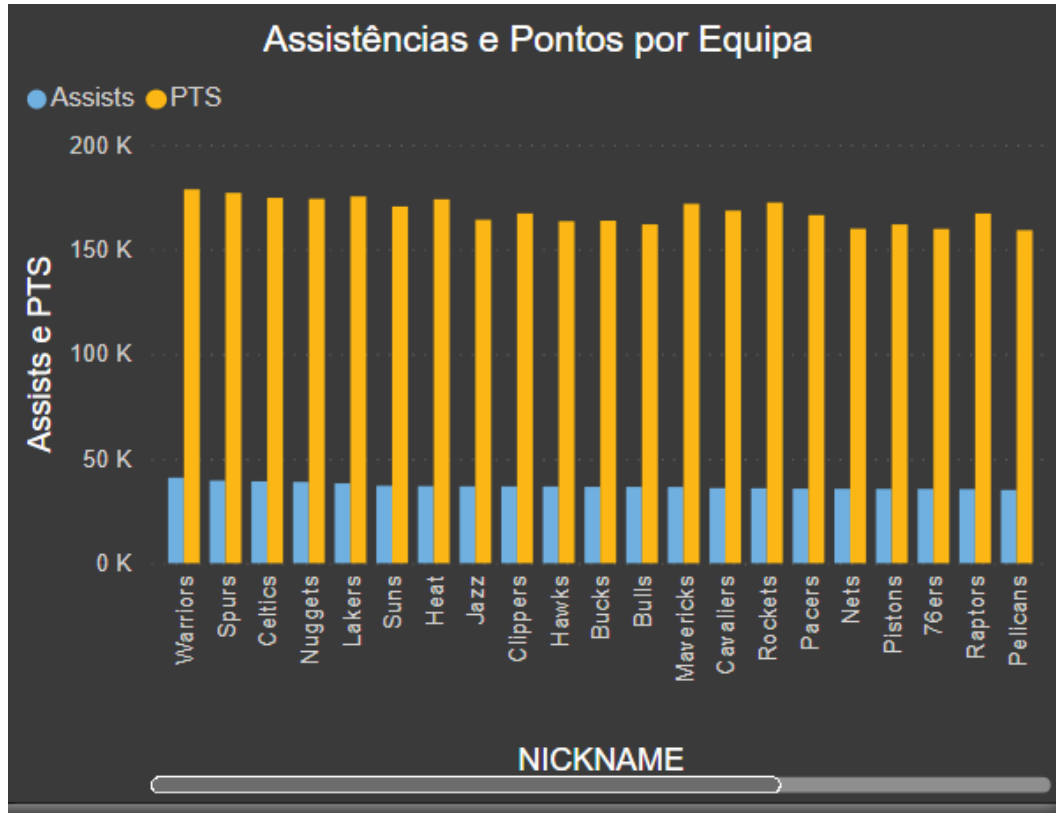


Figura 22 – Gráfico de Barras das Assistências e Pontos por Equipa

Um jogo não se consegue ganhar individualmente, é necessário saber jogar em equipa, por isso procedeu-se à criação de um gráfico de barras onde é possível visualizar as assistências de cada equipa em relação aos pontos marcados, Figura 20, 21 e 22.

```
1 WinYear = TOTALYTD(AVERAGE(Ranking[W]),Ranking[STANDINGSDATE].[Date])
```

Figura 23 – Medida para calcular as Vitórias por Ano

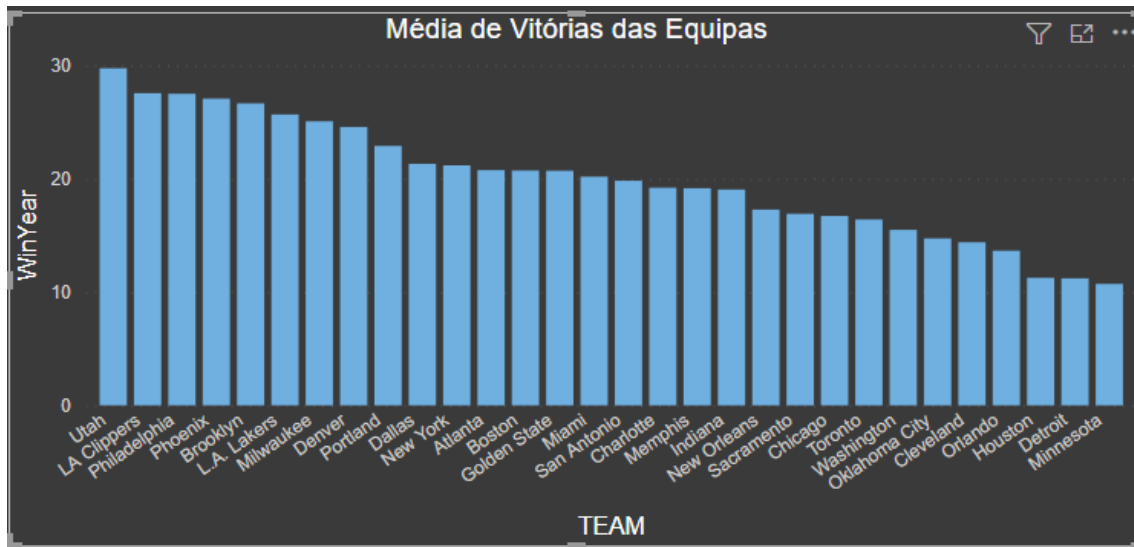


Figura 24 – Gráfico de Barras da Média de Vitórias das Equipas

```
1 LostYear = TOTALYTD(AVERAGE(Ranking[L]),Ranking[STANDINGSDATE].[Date])
```

Figura 25 – Medida para calcular as Derrotas por Ano

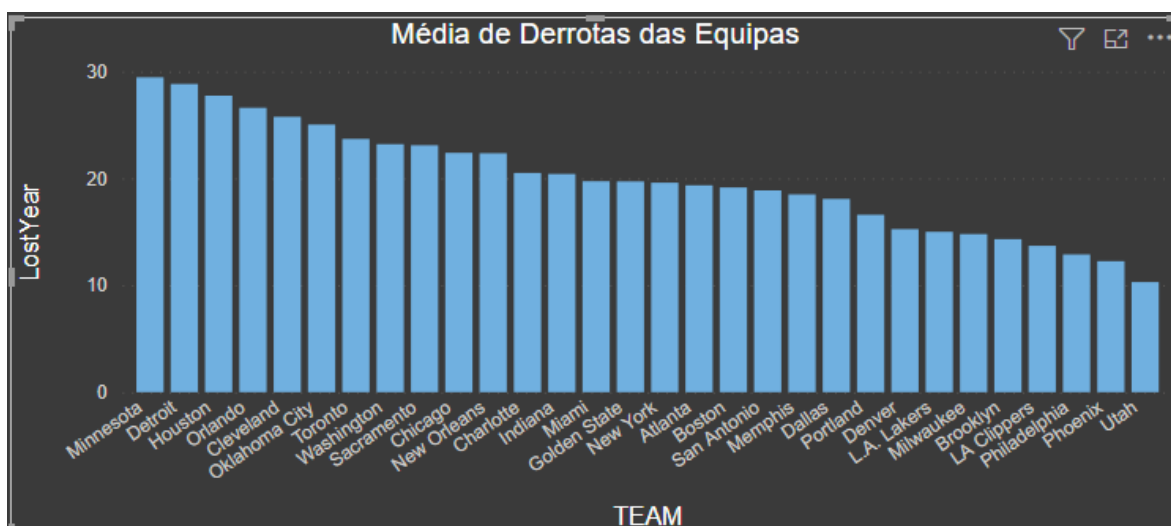


Figura 26 – Gráfico de Barras da Média de Derrotas das Equipas

Para visualizar o número médio de vitórias e o número médio de derrotas de cada equipa por ano, Figura 24 e 26, foi necessário recorrer a duas medidas, Figura 23 e 25.

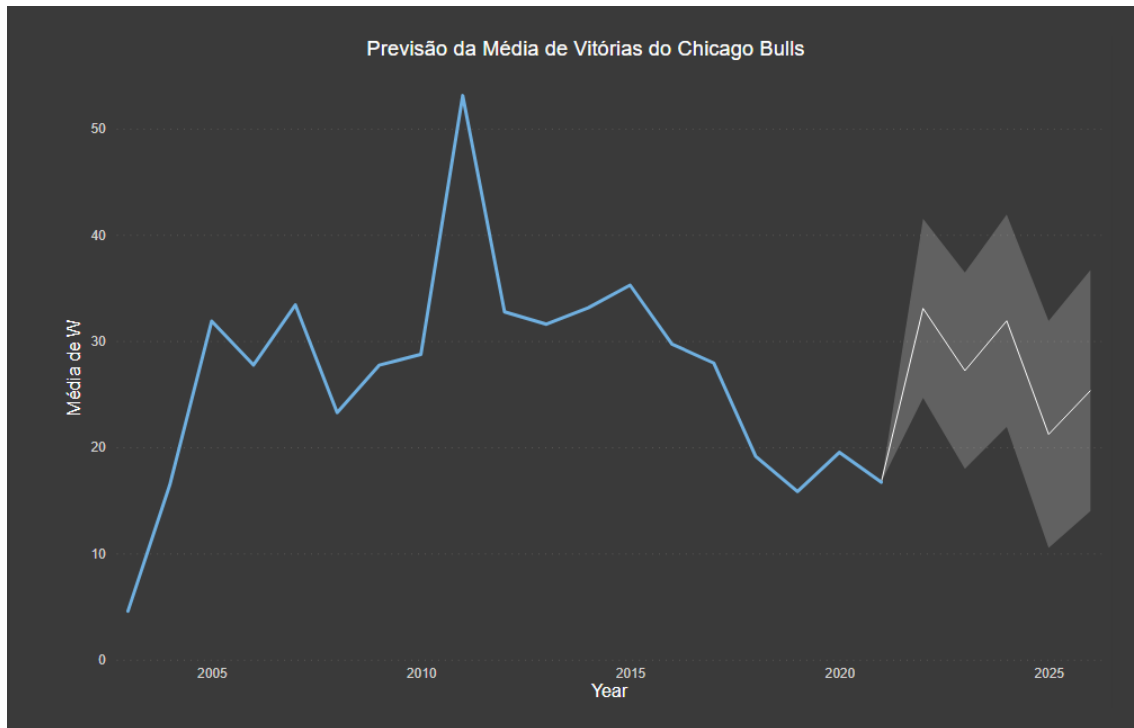


Figura 27 – Gráfico de linhas da Previsão da Média de Vitórias do Chicago Bulls

Com as funcionalidade do *PowerBI* é possível criar um modelo de previsão a partir de um gráfico de linhas, por isso optou-se por prever qual seria a média de vitórias e de derrotas da equipa Chicago Bulls utilizando a média de jogos ganhos a partir de 2003 até 2020, Figura 27 e 28.

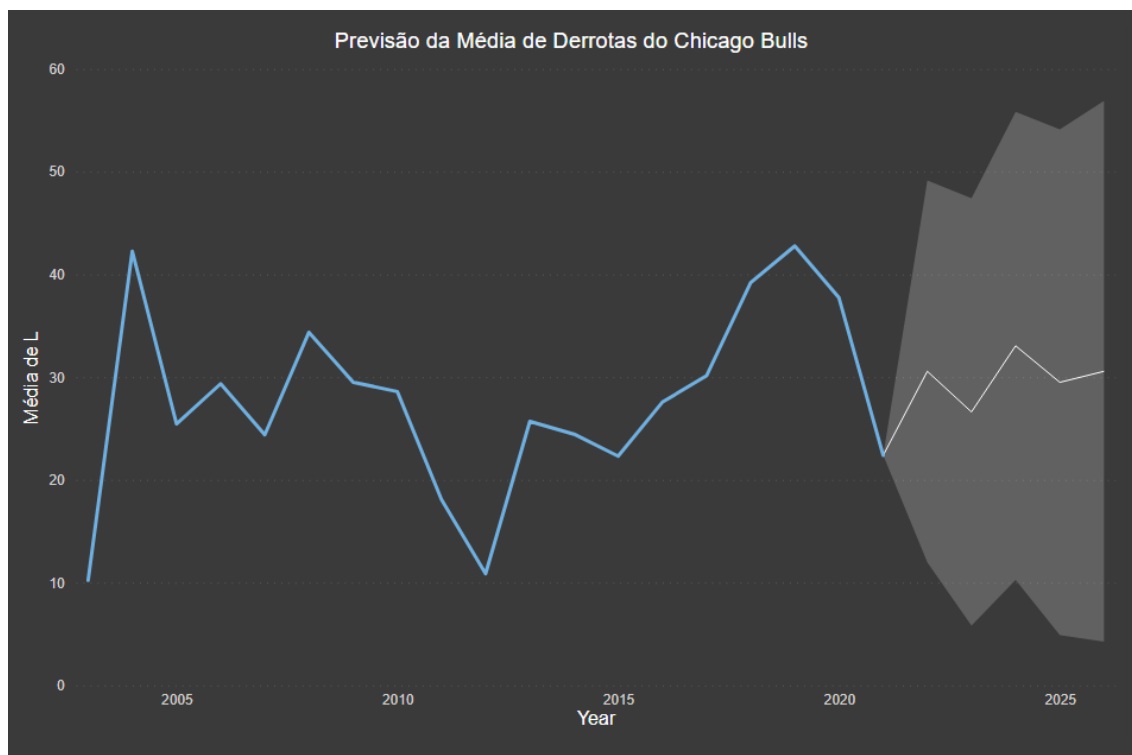


Figura 28 – Gráfico de linhas da Previsão da Média de Derrotas do Chicago Bulls

6. Dificuldades Sentidas

Na realização deste projeto foram sentidas algumas dificuldades, nomeadamente, dificuldades na manipulação do programa *PowerBI* que levaram a alguns atrasos para a conclusão do mesmo. Mas estas dificuldades foram superadas pelo conhecimento que nos foi transmitido nas aulas pela docente e por iniciativa própria.

7. Conclusão

Para concluir este projeto consegue-se extrair informação útil dos gráficos que foram criados. Conseguiu-se identificar que os San Antonio Spurs são os que marcam mais pontos em casa (7009 pontos) e são também os que têm uma maior percentagem de vitórias relativamente às outras equipas (66%).

LeBron James é o jogador que tem mais jogos disputados (1649) mas, só jogou 11 épocas, em contraste com Chris Johnson jogou 18 épocas, sendo o jogador que tem mais épocas, este apenas disputou 264 jogos.

Em termos de ressaltos, a época onde existiu mais foi em 2005 (1432) e onde existiu menos foi em 2011 (1104). Existiu uma média de 1 ressalto ofensivo e 3 ressaltos defensivos por jogo o que faz uma média de 4 ressaltos por jogo.

De todos os lançamentos ao cesto que são possíveis de se realizar, os lançamentos marcados constituíam 58,93%, o que significa que houve pelo menos 3 bem-sucedidos por jogo. Os lançamentos livres possuem 28,75% dos lançamentos bem-sucedidos, o que indica que existiu 1 a 2 por jogo. Por fim, os lançamentos triplos marcados fazem 12,32% dos lançamentos e apenas 1 bem-sucedido por jogo.

Os bloqueios máximos efetuados num jogo foram 12, mas os desarmes foram apenas 10, o que significa que 16,7% de cada bloqueio volta para a equipa adversária.

As equipas que efetuaram mais faltas e conseguiram marcar mais pontos devidos a estas foram Oklahoma City Thunder e os Utah Raptors, enquanto, os Orlando Magic e os Philadelphia 76ers apesar de terem feito o mesmo número de faltas não as conseguiram converter em pontos.

Os Boston Celtics foram uma equipa que fez 39138 assistências das quais contribuíram para um total de 174581 pontos, o que demonstra que os seus jogadores são mais propícios a jogar em equipa, em comparação os LA Lakers fizeram 38180 assistências, mas fizeram 175267 pontos.

A LA Clippers é a segunda equipa com maior média de vitórias por ano, com 28, e apenas a quarta melhor em número médio de derrotas, com apenas 13. Os Phoenix Suns apesar de terem menos uma vitória, 27, também têm menos uma derrota, 12.

Na previsão média de vitória para a equipa Chicago Bulls utilizou-se um método incorporado na análise de dados do *PowerBI*, no intervalo de confiança de 95% e utilizando uma sazonalidade de 17 obteve-se uma previsão para os próximos 5 anos, para o ano de 2022 obteve um resultado de 33 ± 8 , para o ano de 2023 obteve um resultado de 27 ± 9 , para o ano de 2024 obteve um resultado de 31 ± 10 , para o ano de 2025 obteve um resultado de 21 ± 10 e para o ano de 2026 obteve um resultado de 25 ± 11 .

Na previsão média de derrotas para a mesma equipa utilizou-se exatamente os mesmos parâmetros e obteve-se para 2022 30 ± 18 , para 2023 26 ± 20 , para 2024 33 ± 22 , para 2025 29 ± 24 e para 2026 30 ± 26 .

Para finalizar este projeto foi muito interessante no contexto da unidade curricular e ajudou os analistas a aplicar os seus conhecimentos que foram adquiridos ao longo da sua vida académica e espera-se que com isto seja possível no futuro também vir a ter utilidade nas suas vidas profissionais.

8. Bibliografia

<https://www.kaggle.com/nathanlauga/nba-games>

Consultado a 20 de junho de 2021

<https://docs.microsoft.com/pt-pt/power-bi/fundamentals/power-bi-overview>

Consultado a 20 de junho de 2021