



Universidade do Minho

Mestrado Integrado em Engenharia Informática

Unidade Curricular de Análise de Dados

Ano Letivo de 2020/2021

AD

Trabalho Prático de Análise de Dados **DATA Warehousing – 2019 European** **Football Market Values**

André Soares (a67654), Eduardo Costa (a85735), José
Rodrigues (a85501) e Ricardo Carvalho (a84261)

Janeiro, 2021

Introdução

Serve o presente relatório para explicitar as várias fases de desenvolvimento e diferentes componentes do trabalho realizado no âmbito da Unidade Curricular de Análise de Dados, cujo objetivo é o desenvolvimento de um sistema de *Data Warehousing*, bem como um sistema de *Business Intelligence* para suporte à decisão. Para o efeito, foi feito um trabalho de análise, planeamento e implementação, tendo como base um *Dataset* público de informação relativa aos futebolistas profissionais a atuar nas 9 principais ligas da Europa e seus valores de mercado, referente ao final do ano de 2019.

Para tal, foi seguido o modelo de desenvolvimento de Data Warehouses abordado tanto nas aulas teóricas como nas práticas, que passa pelas fases de ETL e *Business Intelligence*. Inicialmente, foi ainda necessária a seleção do *Dataset* que serviu de base para toda a implementação feita. Depois, na fase de ETL, foi feita a extração, transformação, tratamento e carregamento dos dados para um sistema multidimensional construído. Por fim, na fase de *Business Intelligence*, de forma a desenvolver um sistema de suporte à decisão, foram desenvolvidos vários tipos de indicadores que consideramos relevantes para a área de negócio em estudo.

Contextualização do *Dataset* escolhido

A escolha do *Dataset* público compreende a escolha de uma área de negócio, permitindo o desenvolvimento de indicadores relevantes para o caso de estudo. Começamos, então, nesta fase inicial do projeto, por tentar encontrar um *Dataset* de uma área que considerássemos interessante e do qual pensássemos ser possível extrair vários tipos de indicadores relevantes.

Através de plataformas de busca, como o Kaggle ou o *Dataset Search* da Google, encontramos vários tipos de *Datasets*, mas tentamos selecionar aquele nos desse o maior equilíbrio entre a representatividade daquilo que pode ser a quantidade de dados num caso real de implementação de um *Data Warehouse* e a quantidade de atributos acerca do caso de estudo. Optamos, então, por um *Dataset* público de informação relativa aos futebolistas profissionais a atuar nas 9 principais ligas europeias e seus valores de mercado, referente ao final do ano de 2019. Este tem todas as características que pensamos ser relevantes nesta escolha, para o projeto em questão, pois, apesar de conter cerca de 4500 linhas, contém, para cada uma delas, cerca de 35 atributos. Desta forma, pensamos ser perceptível que, construindo um sistema para este caso, mostraria que este estaria preparado para expandir para dados relativos a mais ligas, de mais continentes e a vários anos, sem que tenhamos que efetuar carregamentos da ordem dos milhões de linhas nas nossas máquinas, que não preparadas para esse efeito e considerando que não seria a quantidade de dados, mas sim a qualidade dos mesmos, o ponto crucial do contexto deste projeto. Por outro lado, a quantidade e variedade dos atributos dos futebolistas presentes no conjunto de dados selecionado, como exemplificado nas figuras abaixo, foi um ponto essencial para a sua escolha.

Summary

1 file

35 columns

String	26
Integer	6
DateTime	2
Other	1

Figura 1 - Número de colunas do *Dataset*

A FullName	A PlayerName	A Affiliation	A League	A Jersey	Birth Date	# Age	A birthPlace	# Height (m...)	A Citizenship 1	A Citizenshi...	A Position	A Position 2	A Foot
Emiliano Ariel Rigoni	Emiliano Rigoni	Sampdoria	Serie A	#18	2/4/1993	26	Colonía Caroya	1.8	Argentina	Italy	Forward	Right Winger	both
	Daniil Lesovoy	Arsenal Tula	Premier Liga	#22	1/12/1998	21	Moskau	1.75	Russia	Ukraine	Forward	Left Winger	right
	Kyllian Kaßbou	Montpellier	Ligue 1	#21	8/28/1998	21	Firminy	1.82	France		Midfielder	Defensive Midfield	left
Aarón Escandell Banalchoche	Aarón Escandell	Granada CF	LaLiga	#13	9/27/1995	24	Carcaixent	1.85	Spain		Goalkeeper	Goalkeeper	right
Leonardo Bonatini Lohner	Léo Bonatini	Vit. Guimarães	Liga NOS	#13	3/28/1994	25	Bejo Horizonte	1.84	Brazil	Italy	Forward	Centre-Forward	right

Figura 2 - Pequeno excerto do *dataset*

De forma mais específica, o *Dataset* escolhido tem vários tipos de informação acerca dos futebolistas que nele se encontram, desde os seus nomes, clube, liga, número de camisola utilizado, data de nascimento e idade, local de nascimento, altura, nacionalidades, posições em campo, pé preferido, agente ou agência que o representa, data de assinatura de contrato com o clube atual, última extensão de contrato e data de término do mesmo, patrocinador, até 7 clubes onde tenha feito

formação, número de jogos, valor de mercado e sua última revisão, valor acumulado de todas as suas transferências e maior transferência realizada, país que representa internacionalmente, número de internacionalizações e, finalmente, a sua mais recente lesão ou suspensão. Este conjunto variado de dados acerca de jogador pareceu-nos, desde logo, capaz de expressar vários tipos de indicadores acerca dos mesmos e das ligas e/ou clubes onde jogam, por serem os jogadores a unidade básica dos mesmos, ou seja, todas as informações acerca de cada liga e clube passam por aquilo que caracteriza os jogadores que lá jogam.

ETL (*Extract-Transform-Load*)

Num processo de ETL, começa-se pela extração de informação a partir de diferentes fontes de dados, neste caso, de um ficheiro Excel. De seguida, devemos proceder à transformação e tratamento desses dados, gerando consistência nos mesmos. Depois, a modelação multidimensional: primeiro perceber o modelo de negócio e aplicá-lo na construção e projeção do modelo lógico e passar, posteriormente, para modelo físico, carregando, finalmente, os dados para o mesmo.

As principais dificuldades que se podem encontrar neste processo são no tratamento dos dados a trabalhar, pois os *Data Warehouses* são caracterizados pela integração e consistência dos dados neles contidos. É também muito importante identificar claramente a área de negócio em estudo, para que o projeto a desenvolver seja consistente e tenha o significado desejado para a organização. Posteriormente, também na modelação é necessário perceber qual o modelo de dados multidimensional adequado à situação e implementá-lo.

Consistência de dados

Toda a informação do nosso *Dataset* está contida no ficheiro *MarketValues.xlsx*. Aquando do início do desenvolvimento do nosso trabalho, tivemos de corrigir erros e incoerências deste ficheiro, de forma a ficarmos com um conjunto de dados conciso e consistente.

Para tal, alteramos o formato das datas e, desta forma, ficaram no formato correto para serem tratadas por *SQL*. Tivemos também de corrigir diversas células que se encontravam vazias, substituindo o seu conteúdo pela palavra *NA*, de forma a podermos criar, para os casos necessários, um valor na respetiva tabela representativo desta palavra.

Outro tratamento que realizamos com intuito de atingir consistência de dados, foi alterar caracteres com *encoding* especial, para o respetivo caracter, sendo que temos de seguida um exemplo de uma célula de excel com erro de *encoding*, seguido da mesma corrigida.

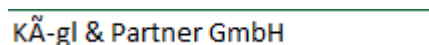
A screenshot of an Excel cell containing the text 'KÃ-gl & Partner GmbH'. The text is displayed with a green border around the cell, and the characters 'Ã' and 'gl' are clearly visible as an encoding error.

Figura 3 - Célula com erro de *encoding*


A screenshot of an Excel cell containing the text 'Kögl & Partner GmbH'. The text is displayed with a green border around the cell, and the characters 'ö' and 'gl' are correctly rendered, indicating the encoding error has been fixed.

Figura 4 - Célula corrigida

Modelo de negócio

A área em estudo, neste caso, estende-se, então, a vários assuntos acerca do futebol europeu, como as características e comportamento dos jogadores, o seu valor de mercado e transferências, agência, contratos e patrocínios, entre outros, relacionando-os com os clubes e respetivas ligas. Estas considerações, tal como se pode ver na secção seguinte, foram consideradas na modelação realizada, de forma a agrupar os dados de forma lógica e consistente.

Modelo lógico

Após uma análise do ficheiro de dados, avaliamos quais as tabelas e relacionamentos a desenvolver e chegámos ao seguinte modelo lógico, modelo este que consiste num esquema em floco de neve. A tabela de factos tem o nome de *fact_table* e tem ligação a seis tabelas, sendo que o seu sétimo atributo é o seu *id*, que é também a sua *primary_key*.

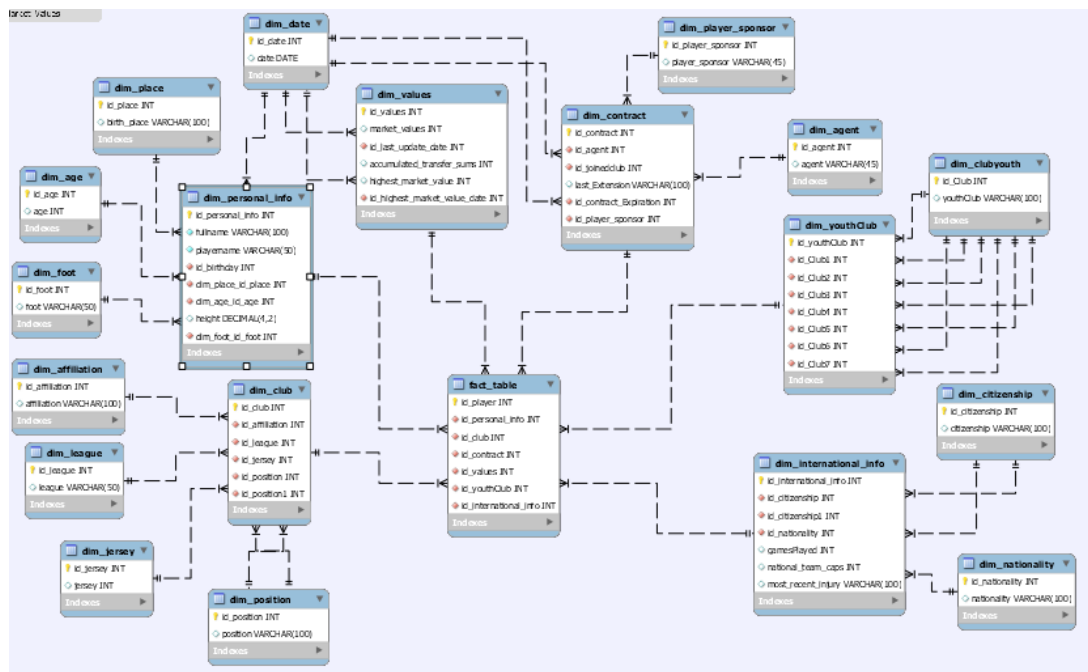


Figura 5 - Modelo lógico desenvolvido

Criação dos ficheiros de povoamento

Após termos construído o modelo lógico, decidimos, então, desenvolver um programa em Java que nos permitisse ler o ficheiro Excel e passar os dados para ficheiros SQL de inserção na base de dados. Neste programa, lemos cada linha do ficheiro e criamos os diversos ficheiros de povoamento, tirando partido de objetos representativos de cada tabela e dos seus atributos. Cada coluna de cada linha do nosso ficheiro Excel foi ser tratada separadamente, de forma a serem gerados os IDs necessários para cada linha de cada tabela.

Temos, de seguida, um exemplo de um objeto representativo da tabela *dimClub*.

```
public class dimClub {  
  
    private int id_club;  
    private int id_affiliation;  
    private int id_league;  
    private int id_jersey;  
    private int id_position;  
    private int id_position1;  
}
```

Figura 6 - Exemplo de um objeto representativo da tabela *dimClub*

Após passarmos todas as linhas e colunas do ficheiro para os respetivos objetos, utilizamos uma classe que contém todos os métodos capazes de passar objetos para os ficheiros SQL referidos. Existe um método para cada tabela do nosso *Data Warehouse* que vai gerar um ficheiro de povoamento por tabela.

De seguida, temos o exemplo de escrita e do resultado da mesma, para o ficheiro *povoarPlace.sql*, referente à tabela com o mesmo nome.

```
try {
    FileOutputStream fos = new FileOutputStream(new File( pathname: "povoarPlace.sql"));

    int i = 1;
    for (String s : dim_place) {

        String query = "insert into dim_place values (" + i++ + ", \"" + s + "\");\n";
        fos.write(query.getBytes());
        fos.flush();
    }
}
```

Figura 7 - Exemplo de código para gerar ficheiros povoamento

```
insert into dim_place values (2, "München");
insert into dim_place values (3, "Oostende");
insert into dim_place values (4, "Petaluma, California");
insert into dim_place values (5, "Köln");
insert into dim_place values (6, "Skopje");
insert into dim_place values (7, "Bonn");
insert into dim_place values (8, "Lunel");
insert into dim_place values (9, "Graz");
insert into dim_place values (10, "Stuttgart");
insert into dim_place values (11, "Istmina");
insert into dim_place values (12, "Saarbrücken");
insert into dim_place values (13, "Oviedo");
insert into dim_place values (14, "NA");
insert into dim_place values (15, "Hamburg");
```

Figura 8 - Ficheiro povoamento da tabela dim_place

Tendo realizado as fases de *Extract* e o *Transform* do processo de ETL, passamos para o *Load* dos dados para o *Data Warehouse*, que consistiu em correr os ficheiros de povoamento, que geraram tabelas como as apresentadas nos exemplos seguintes.

	id_player	id_personal_info	id_club	id_contract	id_values	id_youthClub	id_international_info
	1333	1333	1333	1333	1333	4	844
	1334	1334	1334	66	1334	1334	276
	1335	1335	1335	230	1335	4	1335
	1336	1336	1336	1336	1336	1336	853
	1337	1337	1337	1337	1337	1337	853
	1338	1338	1338	1338	1338	1338	1338
	1339	1339	1339	66	1339	4	853

Figura 9 - Tabela fact_table após povoamento

	id_contract	id_agent	id_joinedclub	last_Extension	id_contract_Expiration	id_player_sponsor
	6	6	10	NA	20	2
	7	7	2	NA	7	1
	8	8	25	NA	3	1
	9	9	2	Jan 18, 2019	7	1
	10	10	2	NA	15	1
	11	11	34	NA	15	3

Figura 10 - Tabela dim_contract após povoamento

Business Intelligence

Um desenvolvimento de sistemas de suporte à decisão na área em que um Data Warehouse está inserido pode ser realizado através de processos de análise e extração de conhecimento como *Data mining*, *Data Science* ou desenvolvimento de indicadores com *software* como Tableau ou Microsoft Desktop PowerBI, por exemplo. Por fim, também a identificação de quais os indicadores a desenvolver é crucial para a criação de sistemas de suporte à decisão relevantes na área em que o data warehouse está inserido.

Para este projeto, o sistema de *Business Intelligence* passou pela construção de vários indicadores considerados relevantes para a área de negócio em questão, optando pela plataforma Power BI devido à sua grande popularidade junto das grandes empresas. O objetivo desta fase é dar dados consistentes a um possível cliente na área, como, por exemplo, neste caso, a UEFA, FIFA, qualquer clube europeu, agência ou patrocinador.

Nesse sentido, criamos as seguintes páginas de representação gráfica de vários indicadores, organizados por diferentes temas.

Nestas duas primeiras, apresentamos vários dados acerca dos jogadores e das suas características, assim como algumas estatísticas gerais dos futebolistas presentes no *Dataset*.

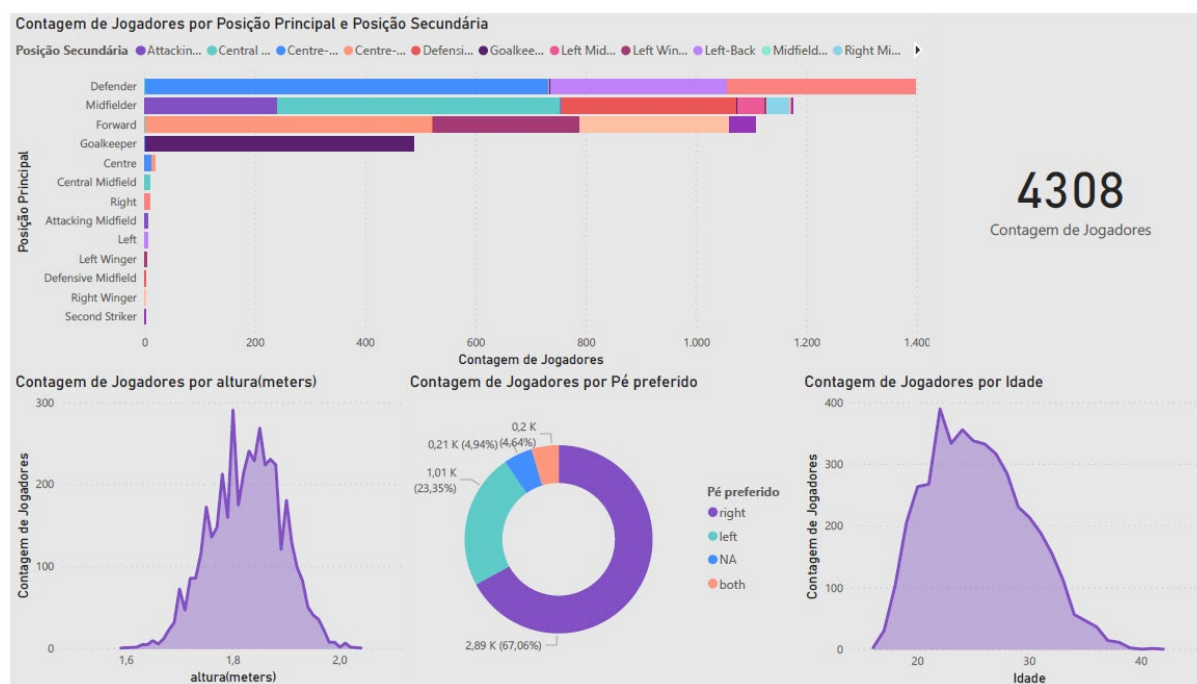


Figura 11 - Página de estatísticas de jogadores

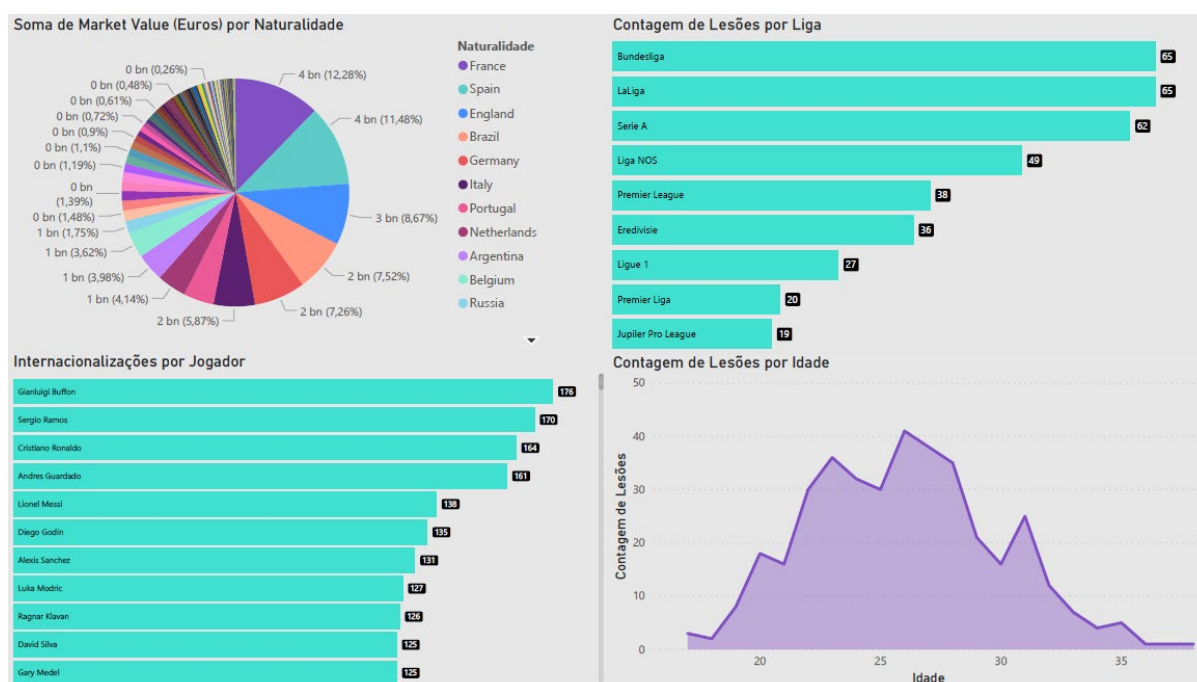


Figura 12 - Página de estatísticas gerais de jogador

De seguida, devido à possibilidade de cada jogador poder ter duas nacionalidades, decidimos representar esta distribuição nos diagramas e mapas apresentados abaixo.

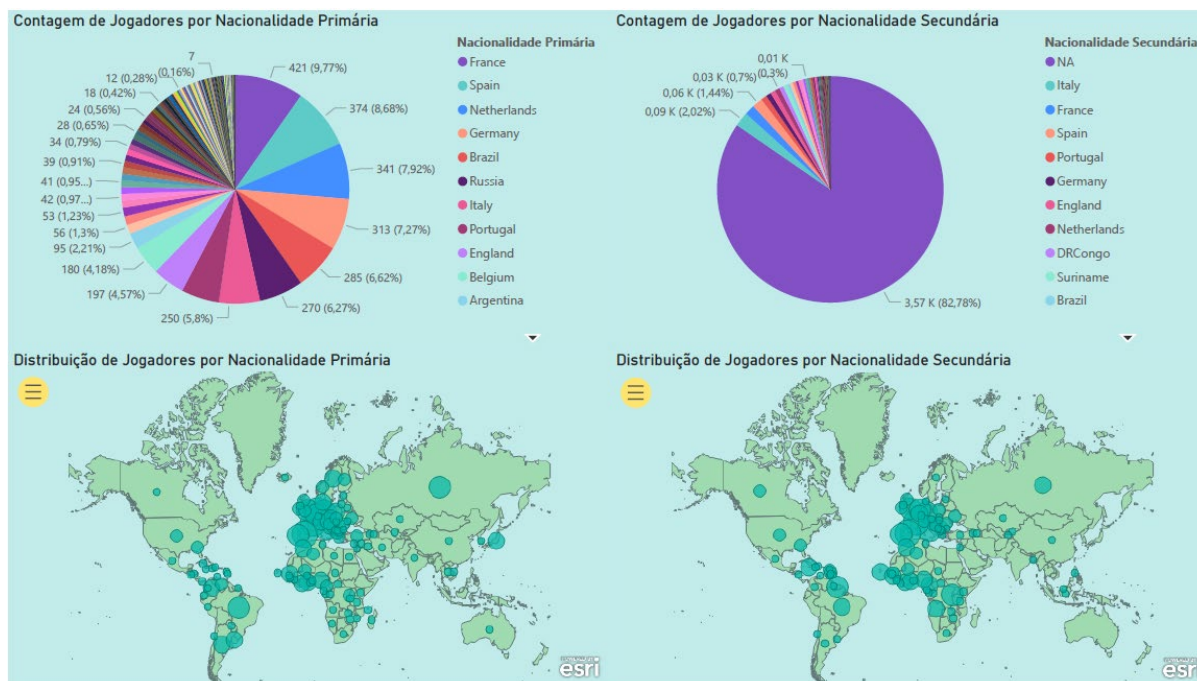


Figura 13 - Página de estatísticas de nacionalidade

Uma das informações centrais do *Dataset* abordado são os valores de mercado dos jogadores no momento da sua construção. Sendo assim, decidimos organizá-los pelas várias variantes de valores encontradas, calculando, também, as médias por idade.

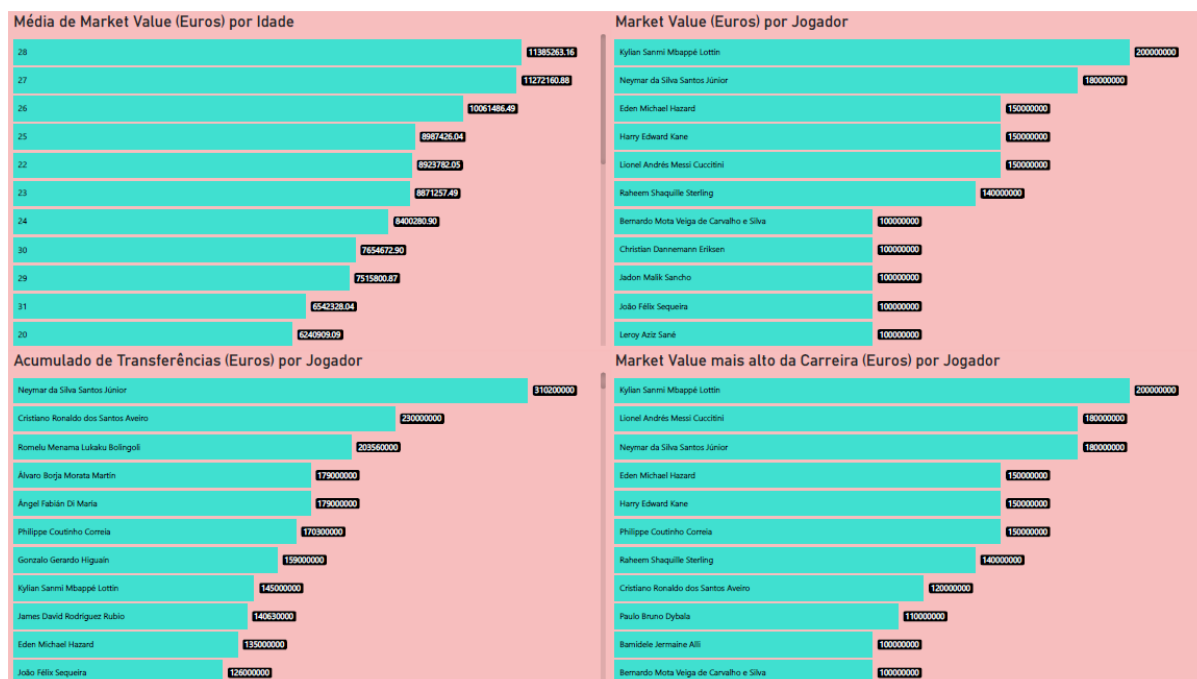


Figura 14 - Página de valores de mercado

Relativamente aos clubes, consideramos relevante saber algumas das estatísticas acerca das idades, valores de mercado e número de jogadores de cada um, assim como a distribuição da quantidade de jogadores em final de contrato a partir do momento de construção do *Dataset*.

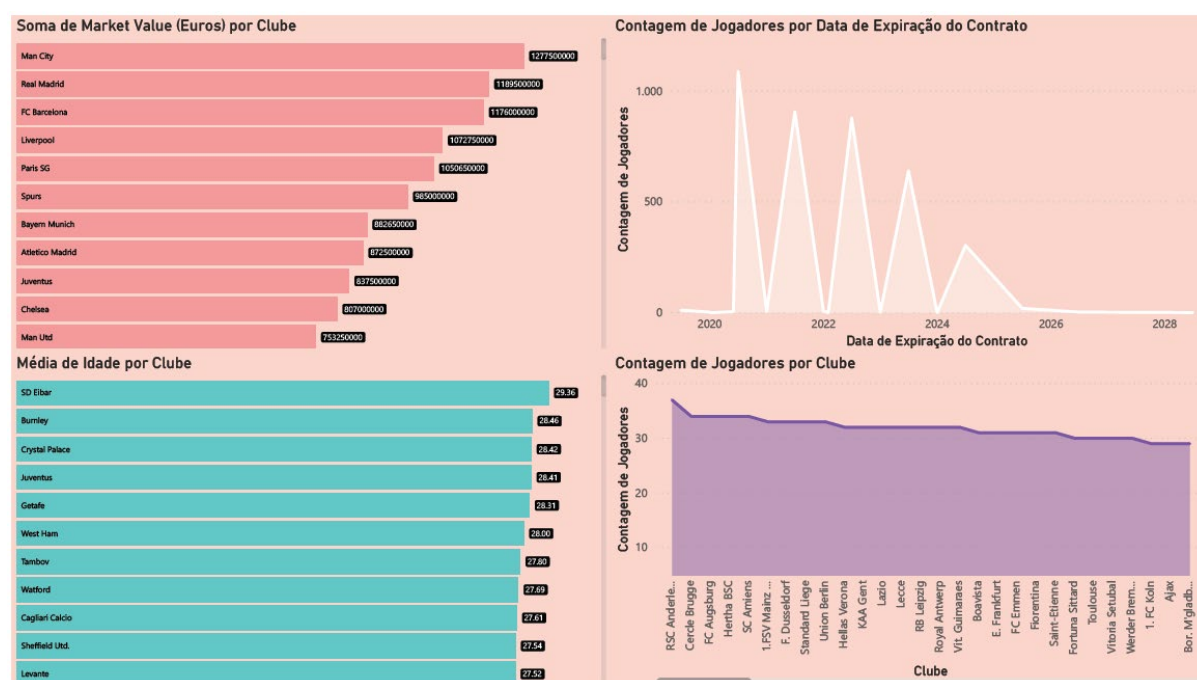


Figura 15 - Página de estatísticas de clubes

À semelhança do caso anterior, também para as ligas foram consideradas as estatísticas de cada uma delas, através dos jogadores e clubes que nelas competem.

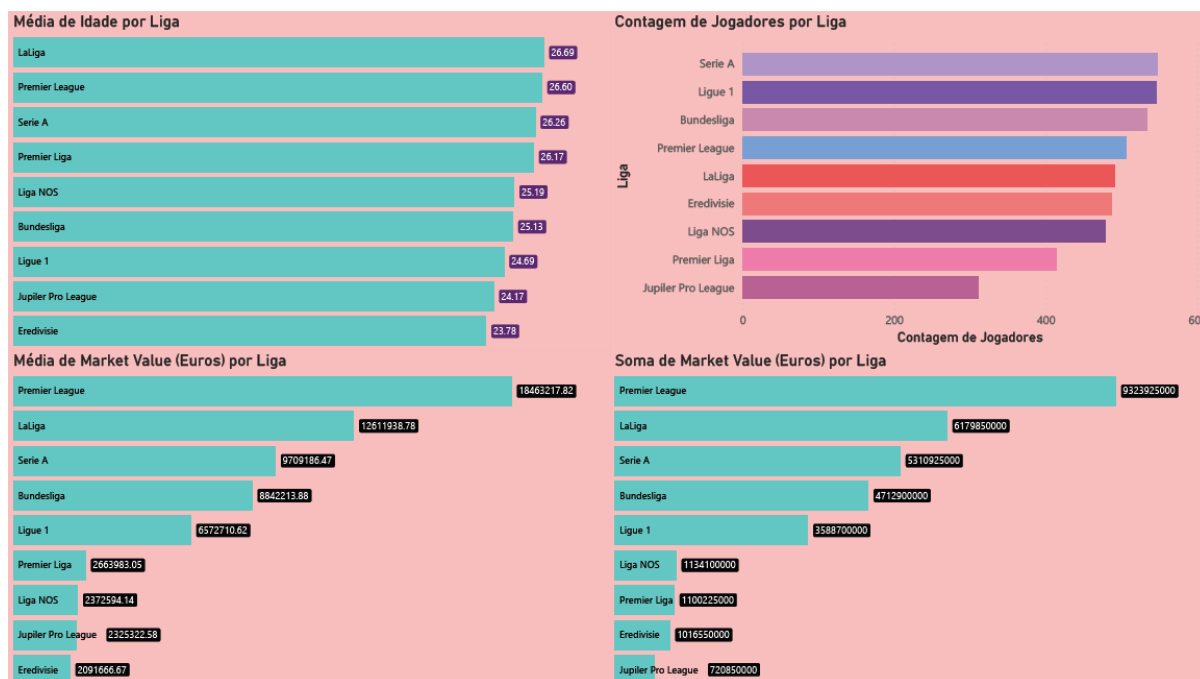


Figura 16 - Página de estatísticas de ligas

Também acerca dos agentes, consideramos importante visualizar a distribuição dos jogadores pelos mesmos, assim como o valor de mercado total e o acumulado de valores de transferências para cada um dos jogadores por eles agenciados.

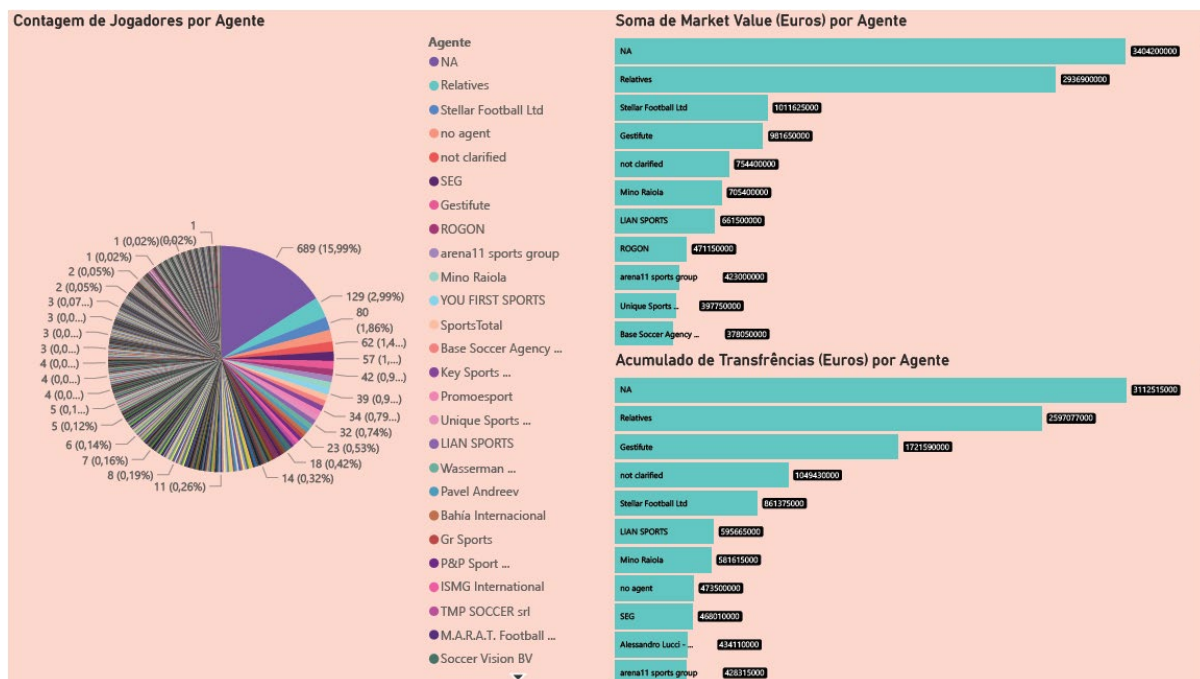


Figura 17- Página de estatísticas de agentes

Por último, achamos importante, também, referir a distribuição dos jogadores por cada patrocinador principal, assim como por nacionalidade (das 5 mais influentes, neste caso) e valores de mercados dos jogadores.

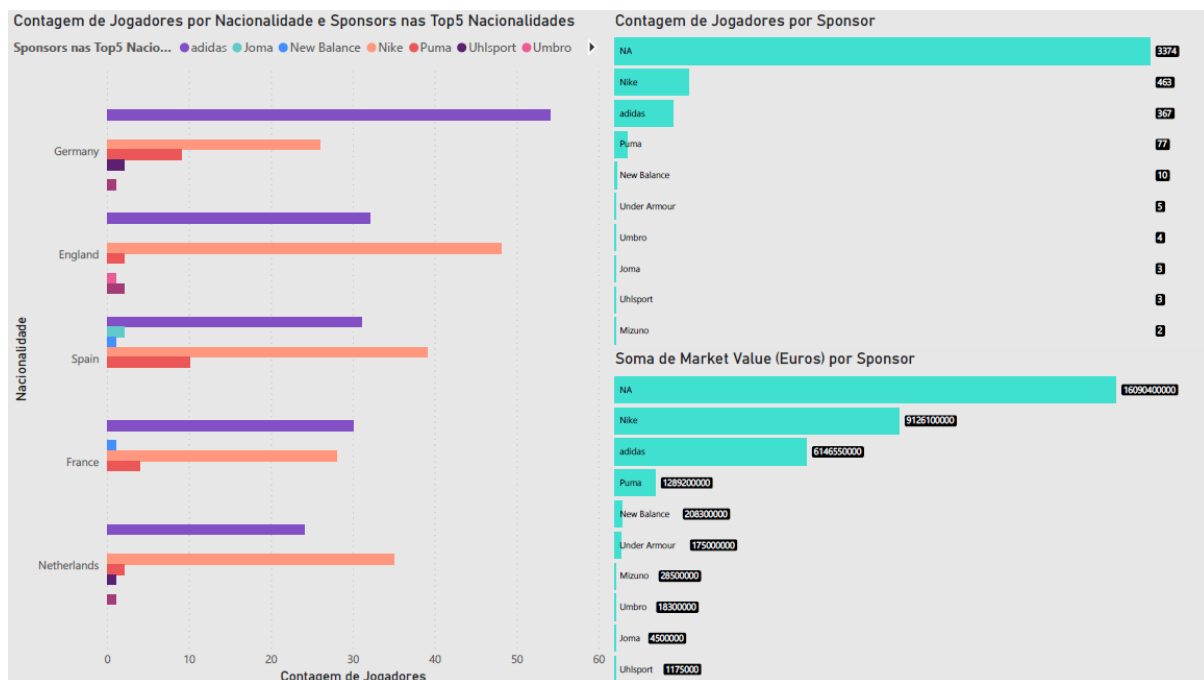


Figura 18 - Página de estatísticas de patrocinadores

Conclusão

Apesar de algumas dificuldades, o grupo dá por concluído com sucesso o desenvolvimento deste projeto de *Data Warehousing*, dando-se satisfeito por colocar em prática todos os conhecimentos adquiridos durante as aulas práticas e teóricas, e nas mais diversas plataformas, tais como MySQL e PowerBI.

Acreditamos ter desenvolvido um sistema capaz de processar dados de forma a permitir um sistema de povoamento inicial e de ter desenvolvido um *Data Warehouse* com dados consistentes, que permitem uma melhor análise e que proporcionam um sistema capaz de apoiar melhores decisões para o negócio, com base em dados do passado.

Em relação à parte de *Business Intelligence*, desenvolvemos diversas representações gráficas, em Microsoft Power BI, que acreditamos serem úteis para uma análise deste *Dataset* e que cumprem os fundamentos requeridos para este trabalho prático. Através dos mesmos, conseguimos disponibilizar meios extremamente concisos e fáceis de analisar para extrair diversos tipos de conclusões acerca dos assuntos relacionados com a área de negócio em questão, desde os jogadores e suas características e valores de mercado, aos seus clubes e ligas, passando, também, pelos patrocinadores e agentes. Desta forma, qualquer entidade responsável ou interessada, poderá, facilmente, tirar vários tipos de elações que lhe sejam benéficas.