Universidade do Minho

Mestrado Integrado em Engenharia Informática

Unidade Curricular de

Análise de Dados

Ano Letivo de 2020/2021

AIBM

Trabalho Prático de Aplicações Informáticas na Biomedicina

DATA Warehousing – 2018 Episódios de Urgência

Peter Vala e Ricardo Carvalho (a84261)

Fevereiro, 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 Aplicações Informáticas Na BioMedicina, cujo objetivo é o desenvolvimento de um sistema de *Data Warehousing*, bem como um sistema de *Business Intelligence* para suporte à decisão clínica. Para o efeito, foi feito um trabalho de análise, planeamento e implementação, tendo como base um *Dataset* público de informação relativa aos incidentes de urgência, referente ao ano de 2018.

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, 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*

O DataSet fornecido para a realização deste projeto consiste num conjunto que contêm dados de episódios de urgência do ano de 2018. Estes dados encontram-se divididos por 5 ficheiros Excel, os quais vamos explicar de seguida.

O ficheiro principal, de onde retiramos os id's da tabela de factos, é o ficheiro *urgency_episodes_new.* Este ficheiro é composto por diversos dados, como dados de admissão, triagem e de saída (alta hospitalar). Devido à elevada extensão de cada linha do ficheiro, apenas vamos apresentar parte deste ficheiro.

URG_EPISODE	DATE_OF_BIRTH	SEX	DISTRICT	DT_ADMITION_URG	ID_	DESC_EXT	ID_PROF_ADMITION	DT_ADMITION_TRAIGE	ID_PROF_	PAIN_SCA	ID_COLOR	DESC_COL
1781367	20/11/1974 00:00	M	PORTO	01/01/2018 00:14	5	Doenca	4710	01/01/2018 00:22	3115	4	3	Amarelo
1781368	04/01/1970 00:00	F	PORTO	01/01/2018 00:15	5	Doenca	5357	01/01/2018 00:25	3115	4	3	Amarelo
1781369	13/12/1970 00:00	F	PORTO	01/01/2018 00:13	5	Doenca	4789	01/01/2018 00:21	3115	4	3	Amarelo
1781371	13/12/1992 00:00	F	SETUBAL	01/01/2018 00:49	8	Queda	5229	01/01/2018 00:51	4325	4	3	Amarelo
1781372	07/11/1991 00:00	F	SETUBAL	01/01/2018 00:50	5	Doenca	4926	01/01/2018 00:53	4325	4	3	Amarelo
1781373	14/02/1951 00:00	F	PORTO	01/01/2018 00:51	5	Doenca	3716	01/01/2018 00:54	4325	4	3	Amarelo
1781375	21/06/1983 00:00	F	PORTO	01/01/2018 01:05	5	Doenca	4926	01/01/2018 01:07	3115	4	3	Amarelo
1781376	27/06/1971 00:00	M	PORTO	01/01/2018 01:08	1	Acidente	4710	01/01/2018 01:13	3115	4	3	Amarelo

1: Ficheiro urgency_episodes_new

Os restantes 4 ficheiros fornecem outros dados relacionados com episódios de urgência, sendo estes o ficheiro *urgency_exam*, que contem os registos de todos os exames realizados nos incidentes, o ficheiro *urgency_procedures*, que contem os registos de todos os procedimentos realizados nos incidentes deste ano e o ficheiro *urgency_prescriptions*, que contem todos os dados de prescrições de medicamentos realizadas nos episódios de urgência do *Dataset*. O quarto ficheiro, icd9_hierarchy, contem uma lista hierárquica de códigos para diversos diagnósticos.

1803883	RXU.300.2018.9972	Coluna dorsal, duas incidencias
1788088	RXU.300.2018.3022	Coluna lombar, duas incidencias
1810834	RXU.300.2018.12965	Anca unilateral, uma incidencia
1878269	RXU.514.2018.864	TC, suplemento de contraste endovenoso
1795365	RXU.503.2018.198	Ecografia abdominal superior
1792589	RXU.511.2018.1273	TC, contraste oral

2: Ficheiro urgency_exams

URG_EPIS	COD_PRES	ID_PROF_	DT_PRESC	COD_DRU	QT	PVP	COMPART	POSOLOG	DESC_DRUG
1811774	16482290	57676	########	8796	1	0	0	1 aplicaÃ§	Enoxaparina sÃ³dica, [Lovenox], 40 mg/0.4 ml, SoluÃ§Ã£o injetÃ¡vel, Sering
1812299	16483079	38425	########	185400	1	0	0	1 cp 8/8 h(	Paracetamol, 1000 mg, Comprimido, Blister - 18 unidade(s)
1786132	16388254	44273	########	2032261	1	0	0	1 cp 12/12	Amoxicilina + Ãcido clavulÃ¢nico, 875 mg + 125 mg, Comprimido revestido, E
1794743	16423083	49152	########	81341	1	0	0	1 cp de 8/i	Gabapentina, 100 mg, CÃ¡psula, Blister - 20 unidade(s)
1794888	16423102	29117	########	4689	1	0	0	1 comprin	Etodolac, [Dualgan], 300 mg, Comprimido revestido, Blister - 20 unidade(s)

3: Ficheiro urgency_prescriptions

URG_EPIS	ID_PROFE	DT_PRESC	ID_PRESCI	DT_BEGIN	NOTE
1795050	5437	########	3259760	########	Procedimento associado a administracao de Cloreto Sodio 0,9% (Sol.Inj.,1000ml,---,I.V.,unitaįrio).
1794522	5830	########	3258401	########	Procedimento associado a administracao de Cloreto Sodio 0,9% (Sol.Inj.,100ml,---,I.V.,unitaįrio) ; Tram
1794562	5993	########	3258430	########	
1794566	5435	########	3258474	########	Procedimento associado a administracao de Metoclopramida (Sol.Inj.,10mg,---,I.V.,unitaįrio).
1796026	2901	########	3262221	########	Procedimento associado a administracao de Enoxaparina (H.B.P.M.) (Sol.Inj.,40mg,---,S.C.,unitaįrio).
1796091	5420	########	3262322	########	Procedimento associado a administracao de Combivent (Sol.,5mcg,00:20/00:20h,Inal.,normal).

4: Ficheiro urgency_procedures

level_1_c	level_1_d	level_2_c	level_2_d	level_3_c	level_3_d	level_4_c	level_4_d	level_5_c	level_5_desc
1-139	INFECTIOU	01/set	INTESTINA	1	Cholera	1.9	Cholera, unspecified		
1-139	INFECTIOU	01/set	INTESTINA	1	Cholera	1.1	Due to Vibrio cholerae el tor		
1-139	INFECTIOU	01/set	INTESTINA	1	Cholera	1.0	Due to Vibrio cholerae		
1-139	INFECTIOU	01/set	INTESTINA	2	Typhoid a	2.9	Paratyphoid fever, unspecified		

5: Ficheiro icp9_hierarchy

ETL (*Extract – Transform - Load*)

Num processo de ETL, começa-se pela extração de informação a partir de diferentes fontes de dados, no nosso caso, de um ficheiro Excel. De seguida, devemos proceder à transformação e tratamento desses dados, gerando consistência nos mesmos. Depois, vem 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 DW 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, como já referido, em 5 ficheiros de excel e, aquando do início do desenvolvimento do nosso trabalho, tivemos de corrigir erros e incoerências destes ficheiros, 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 com problemas de *encoding,* sendo que o conteúdo destas continha caracteres especiais, apenas tivemos de mapear os diversos caracteres com problemas e substituir pelos corretos.

Em relação a certas células que aparecem vazias nos ficheiros, optamos por tratar das mesmas na parte de criação dos ficheiros de povoamento.

De seguida, apresentamos como exemplo, o ficheiro urgency_exams original, seguido do mesmo após o tratamento de dados.

URG_EPISODE	NUM_EXAM	DESC_EXAM
1796039	RXU.300.2018.6547	Joelho, duas incidÃªncias
1798249	RXU.300.2018.7564	Punho, duas incidÃªncias
1799772	RXU.300.2018.8147	Punho, duas incidÃªncias
1781772	RXU.300.2018.183	Bacia
1795222	RXU.514.2018.107	TC do crÃ¢nio
1800420	RXU.300.2018.8435	Abdomen Simples 1 Incidencia

6: Ficheiro urgency_exams original

URG_EPISODE	NUM_EXAM	DESC_EXAM
1796039	RXU.300.2018.6547	Joelho, duas incidencias
1798249	RXU.300.2018.7564	Punho, duas incidencias
1799772	RXU.300.2018.8147	Punho, duas incidencias
1781772	RXU.300.2018.183	Bacia
1795222	RXU.514.2018.107	TC do cranio
1800420	RXU.300.2018.8435	Abdomen Simples 1 Incidencia

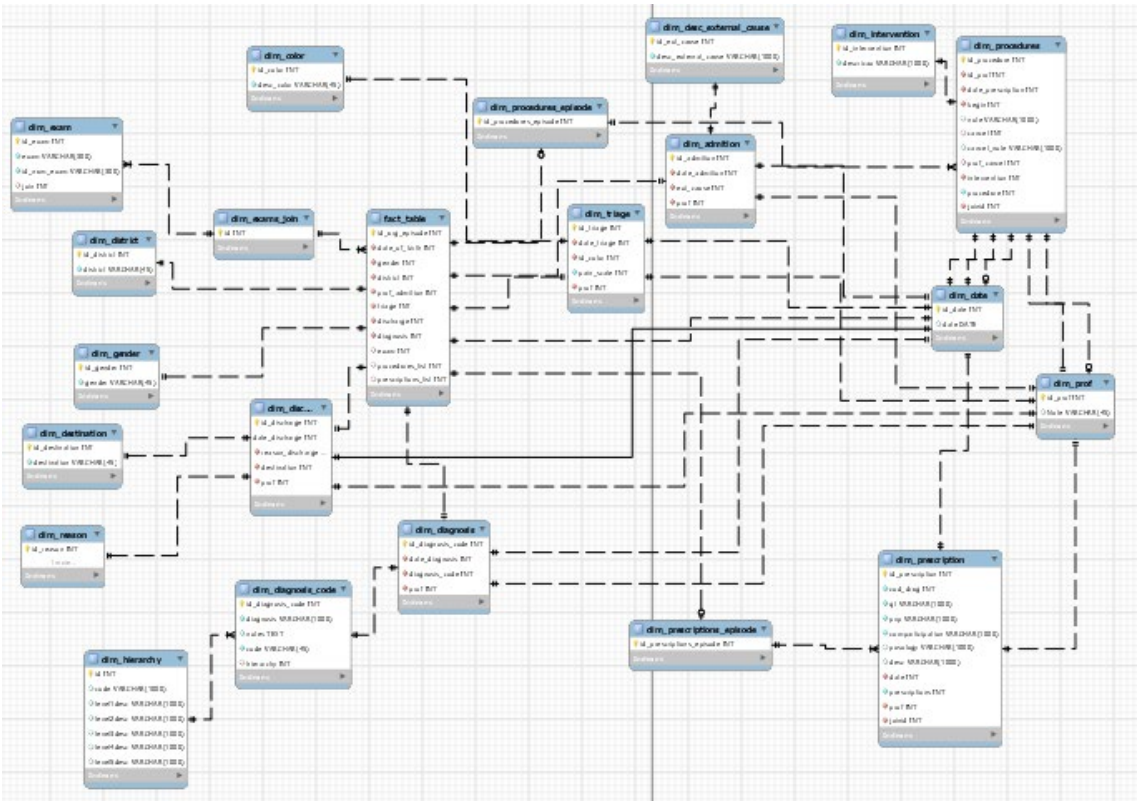
7: Ficheiro urgency_exams após tratamento

Modelo de negócio

Não faço a mínima o que por aqui

Modelo lógico

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



Todas as foreign_keys da fact_table, com a exceção de três, são ligações simples de id. As outras três, são foreign_keys de tabelas de relacionamento de um episódio de urgência para os exames, procedimentos e prescrições. Esta tabela foi necessária pois ao contrário de, por exemplo, as admissões, que têm uma relação um para um com os episódios de urgência, os exames, as prescrições e os procedimentos têm uma relação de 1 para N, pois cada episódio pode ter vários destes, tal como pode não ter nenhum, sendo que as tabela de relacionamento têm o id 1 reservado para estes casos.

Criação dos ficheiros de povoamento

Após termos construído o modelo lógico, tivemos então a necessidade de passar os dados para ficheiros de povoamento e, para isso, desenvolvemos um programa em java que nos permite ler o ficheiro Excel. Nesse programa, lemos cada linha do ficheiro e criamos os diversos ficheiros de povoamento, tirando partido de objetos representativos das linhas de cada tabela. Ao ler cada células dos diversos ficheiros, decidimos se a mesma tem de ter valor (caso seja uma foreign key) e, para tal, não só tratamos os valores nulos como id's específicos (regra geral, id 1), como criamos *records* correspondentes a nulo, por exemplo, para objetos como as datas, criamos uma linha que mapeia o valor nulo, isto é, não ter data associada, ao id 1 e, desta forma, podemos mapear as datas que não têm valor a esta linha, sendo que especialmente nas datas de cancelamento dos procedimentos, esta opção se mostrou bastante viável.

Temos de seguida um exemplo de um objeto representativo de uma linha da tabela *dim_admission.*

```
public class Admition {
    private int date;
    private int cause;
    private int prof;
```

Após passarmos todas as linhas e colunas do ficheiro para os respetivos objetos, utilizamos uma classe que contêm todos os métodos capazes de passar objetos para ficheiros povoamento.

Este programa gera seis ficheiros de povoamento com os nomes sequenciais, de forma a ser mais fácil saber qual a ordem de ficheiros a correr para povoar o datawarehouse.

De seguida, temos o exemplo de escrita e do resultado da mesma para o ficheiro povoamento4.

```
public static void dim_Admition(List<Admition> pr) {
    try {
        FileOutputStream fos = new FileOutputStream(new File( pathname: "povoamento4.sql"),  append: true);
        int i = 1;

        for(Iterator var4 = pr.iterator(); var4.hasNext(); ++i) {
            Admition e = (Admition)var4.next();
            String query = "insert into dim_admition values (" + i + ", " + e.getDate() + "," + e.getCause() + ", " +
            fos.write(query.getBytes());
            fos.flush();
        }

        fos.flush();
        fos.close();
```

```
insert into dim_admition values (1, 109,5, 4710);
insert into dim_admition values (2, 109,5, 5357);
insert into dim_admition values (3, 109,5, 4789);
insert into dim_admition values (4, 109,8, 5229);
insert into dim_admition values (5, 109,5, 4926);
insert into dim_admition values (6, 109,5, 3716);
insert into dim_admition values (7, 109,5, 4926);
insert into dim_admition values (8, 109,1, 4710);
insert into dim_admition values (9, 109,8, 5229);
insert into dim_admition values (10, 109,5, 3550);
insert into dim_admition values (11, 109,5, 4923);
insert into dim_admition values (12, 109,5, 5174);
insert into dim_admition values (13, 109,5, 4787);
```

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_admition	date_admition	ext_cause	prof
1	109	5	4710
2	109	5	5357
3	109	5	4789
4	109	8	5229
5	109	5	4926
6	109	5	3716
7	109	5	4926
8	109	1	4710
9	109	8	5229
10	109	5	3550
11	109	5	4923
12	109	5	5174

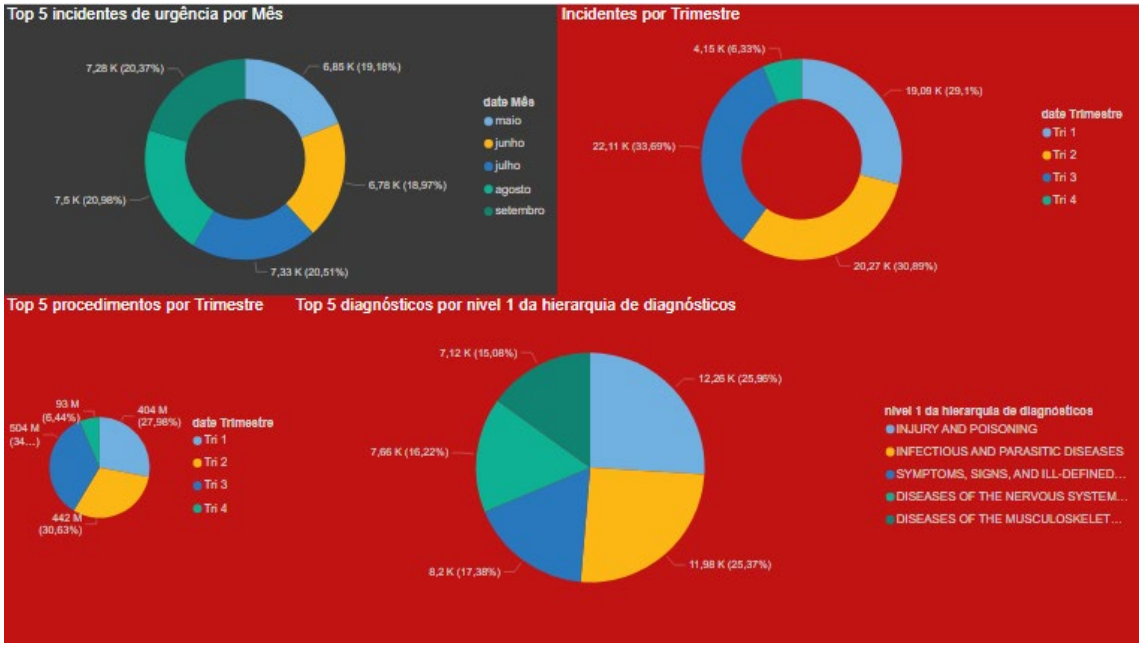
id_urg_episode	date_of_birth	gender	district	prof_admition	triage	discharge	diagnosis	exam	procedures_list	prescriptions_list
1781367	371	1	2	1	1	1	1	3858	16555	1
1781368	372	2	2	2	2	2	2	1	1	1
1781369	373	2	2	3	3	3	3	1	1	1
1781371	374	2	3	4	4	4	4	3488	1	35249
1781372	375	2	3	5	5	5	5	28032	20650	10399
1781373	376	2	2	6	6	6	6	7763	5979	6379
1781375	377	2	2	7	7	7	7	59320	10348	1
1781376	378	1	2	8	8	8	8	2276	1	18063

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 indicares a desenvolver é crucial para a criação de sistemas de suporte à decisão relevantes na área em que o data warehouse está inserido.

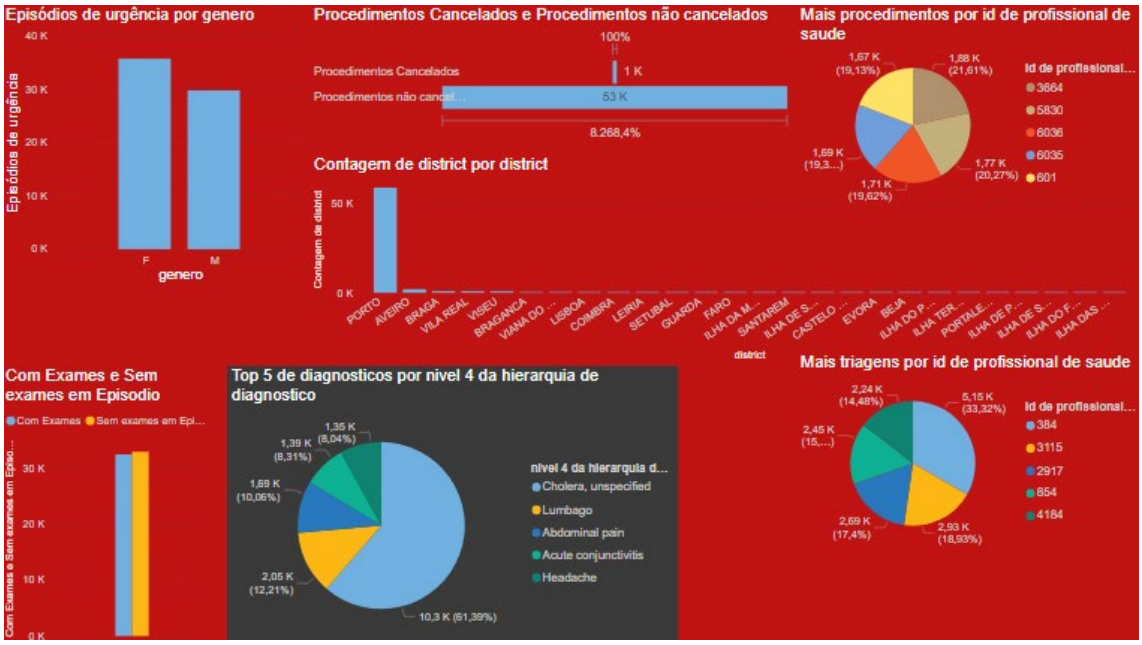
Para este projeto, o sistema de *Business Intelligence* passou pela construção de vários indicadores considerados relevantes para a área de negócio em questão, optando pela plataforma Power BI devido à sua grande popularidade junto das grandes empresas. O objetivo desta fase é dar dados consistentes a administrador do hospital ao qual este *dataset* corresponde.

Nesse sentido, criamos as seguintes páginas de representação gráfica de vários indicadores.

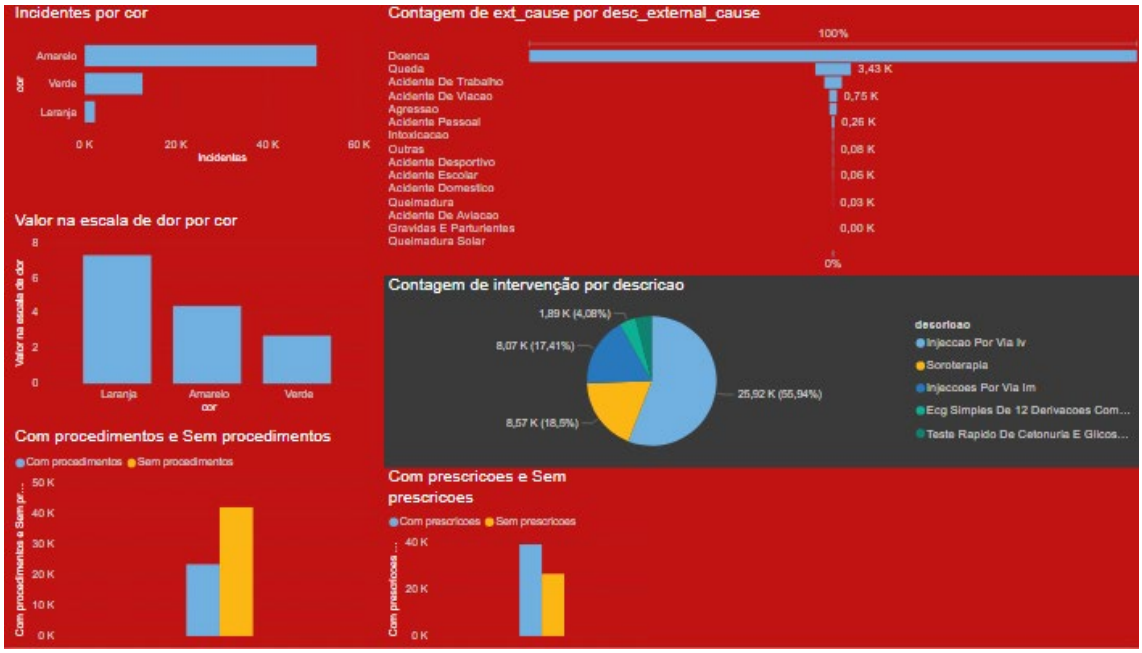


Devido à necessidade de os profissionais do hospital terem férias e, devido à possibilidade de haver contratos de emprego de diferentes durações, criamos esta página, que permita ao administrador do hospital saber diversos dados estatísticos relativos ao número de incidentes por trimestre e aos meses com mais afluência de pacientes.

Temos também os trimestres com mais procedimentos, o que ajuda a saber quais os meses onde profissionais especializados são mais requisitados e, nesse seguimento, temos os diagnósticos mais comuns neste ano, o que mais uma vez pode ajudar a saber que especialidades necessitam de ser reforçadas e quais estão com profissionais a mais.

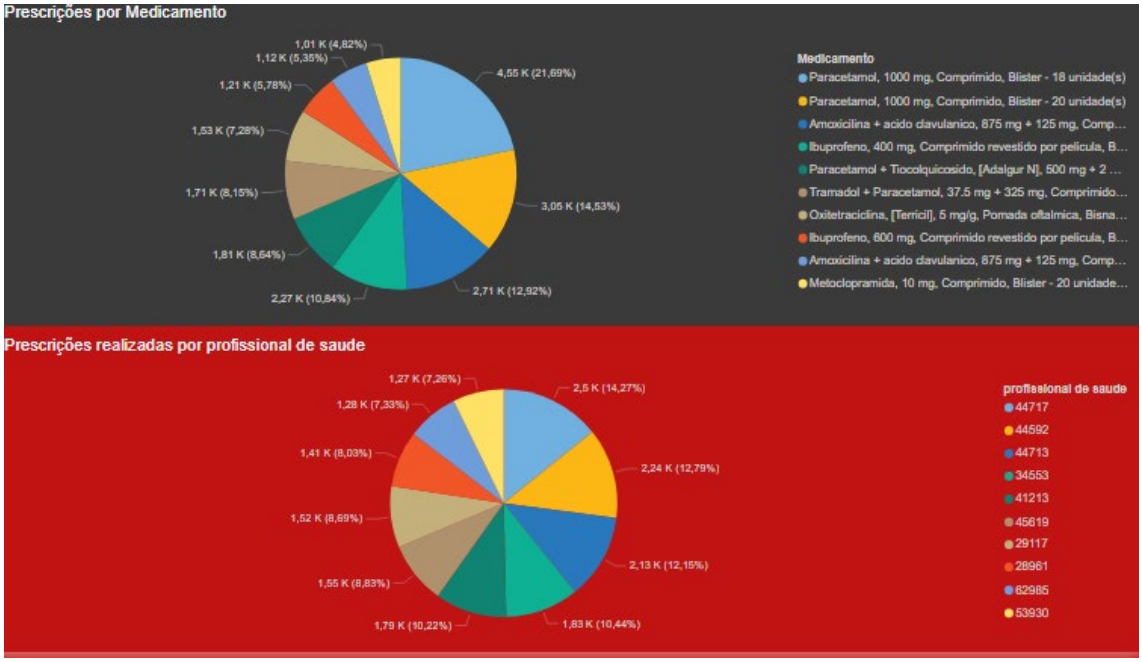


Achamos relevante para a análise estatística desenvolver uma página de relatório com informações dos pacientes, como saber se houve mais homens ou mulheres com episódios de urgência e inclusive de onde são estes pacientes, se na grande parte dos episódios, houve exames realizados ou não, o que pode indicar a necessidade de investir na área de exames. Desenvolvemos também indicadores dos 5 profissionais de saúde que realizaram mais procedimentos e mais triagens. Tendo em conta que há profissionais de saúde com valores bastante mais elevados de procedimentos ou triagens que os restantes colegas, pode o gestor deste hospital, por exemplo, tirar a conclusão de que certos profissionais estão sobrecarregados.

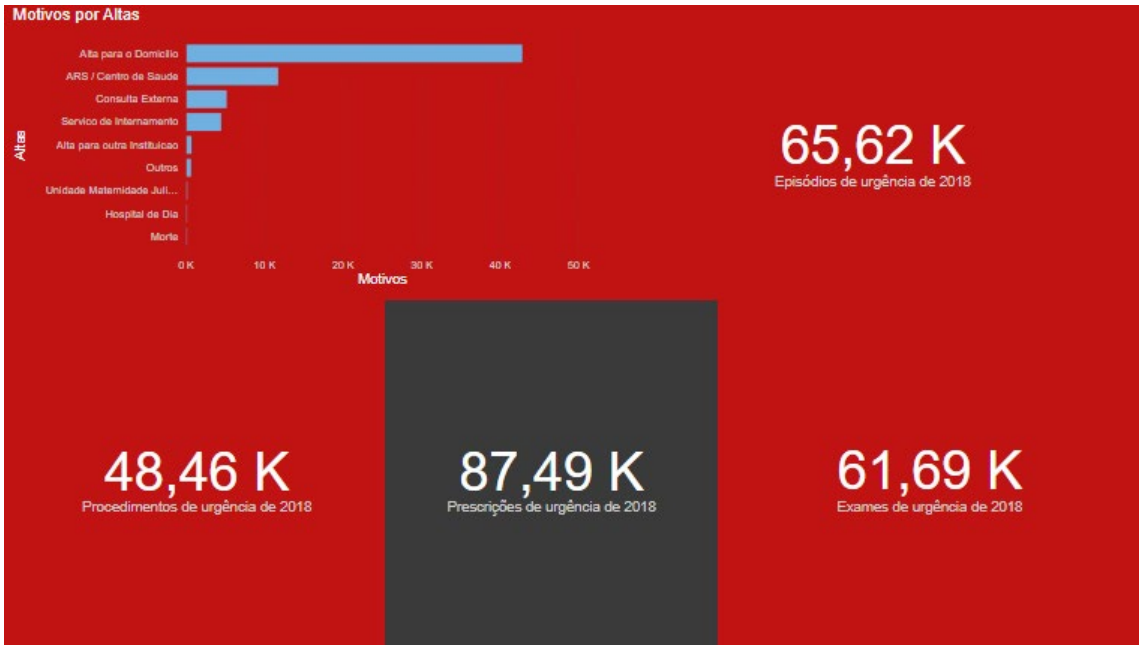


Criamos uma página que permita também quais a pulseiras mais atribuídas aos pacientes e, qual a dor sentida pela maior parte destes pacientes. Fornecemos também dados estatísticos dos procedimentos e das prescrições, associados a episódios.

No seguimento da página anterior, fornecemos quais as principais causas externas que levaram os pacientes ao hospital.



De forma a facilitar a gestão de medicamentos do hospital, desenvolvemos um gráfico que permite saber quais os medicamentos mais receitados e, damos também a informação de quais os profissionais de saúde que realizam mais prescrições.



Nesta última página, fornecemos dados sobre os principais motivos de altas hospitalares e, diversos dados gerais sobre o número de episódios, procedimentos, prescrições e exames totais realizados neste ano.

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 hospitalares, 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.