# Fundamentos de Business Analytics

Gonçalo rosa

Nº 125684

Professor: Paulo Pereira

ISCTE – Instituto Universitário de Lisboa

Cadeira: Fundamentos de Business Analytics

# Introdução

Na era da informação, a capacidade de agregar grandes volumes de dados oferece oportunidades imensas para as organizações. No entanto, essa "tempestade de métricas" pode gerar confusão e dificultar a identificação de insights relevantes se não for tratada de forma adequada. O verdadeiro poder da análise de dados reside na sua capacidade de transformar dados brutos em informações valiosas.

Através do uso de ferramentas como o Power BI, conseguimos dar sentido a essa complexidade, permitindo não apenas a identificação de problemas previamente ocultos, mas também a revelação de rendimentos que não eram evidentes. Independentemente do resultado da análise, este é fundamental para fundamentar decisões estratégicas no negócio, garantindo que ações futuras sejam embasadas em dados concretos.

O objetivo deste relatório é suportar, do trabalho realizado com o hospital do *MASSACHUSETTS GENERAL HOSPITA.* Começaremos por analisar o dataset na sua totalidade as tabelas e os seus conteúdos, de seguida o processo ETL (*Extract-Transform-Load*) por fim iremos analisar os visuais contruídos em power bi para que possamos dar contexto aos dados e tirarmos conclusões sobre os nossos dados.

# Dataset

O dataset representa uma vista de um hospital ao longo de 11 anos com informações sobre os pacientes as visitas ao hospital, tendo por base as seguintes tabelas:

* Encounters

Tabela com dados sobre os encontros de cada paciente ao longo dos anos, com informação sobre o *Payer* (seguradora), valores de custo, código e descrição da razão da visita ao hospital, hora de entrada e hora de saída.

* Organizations

Informação sobre o hospital, nome, localização.

* Patients

Nome, data de nascimento, data de óbito (caso haja), onde nasceu e onde vive

* Payers

Informações sobre os *payers* como nome, localização

* Procedures

Os *procedures*, são todos os procedimentos ocorridos dentro de cada visita ao hospital, com chaves estrangeiras tanto para o paciente como para o encontro, hora de cada procedimento (início e fim)

Qualquer dúvida que tenhamos do nosso dataset, também foi nos disponibilizado um dicionário com todos os campos de cada tabela e uma breve explicação sobre cada um.

No nosso dataset irá nascer um modelo de dados, construído com tabelas de fatos e de dimensão, onde as tabelas de fato contêm dados numéricos que representam medições, transações ou eventos que ocorrem em um negócio. Elas são o núcleo de um modelo de dados analítico e as tabelas de dimensão fornecem o contexto para as medições contidas nas tabelas de fato. Elas permitem a análise dos dados de forma mais rica.

# Processo ETL

Os dados foram passados em ficheiros CSV, não tendo sido necessário nenhuma extração de uma base de dados, ou de outro tipo de fonte, oque facilita o processo em si.

O processo de transformação será replicado em todas as tabelas presentes no dataset, onde iremos primeiro verificar a integridade dos dados, de seguida os tipos de dados se encontram corretos. Por fim criação de novas colunas que iremos necessitar para o futuro.

## Extração

Os dados estão em formato CSV, com uma pasta tendo os encontros de todos os anos separados por intervalos de anos. A maioria das tabelas já se encontra separadas, oque facilita a visão futura do modelo de dados.

## Transformação

### Encounters

Ao carregar no power querypodemos logo dar o upload da pasta inteira dos encontros e agregar numa nova tabela ao à qual demos o nome de FCT\_FullEncounters, uma tabela de fato, de seguida podemos começar a filtrar a horizontal retirando dados que possam estar errados ou até mesmo vazios.

O processo de transformação, após o filtro horizontal, podemos então começar a olhar para as colunas que achamos necessárias, outras que podem ser transformadas em tabelas de dimensão.   
Ao percorrer as tabelas podemos alterar o tipo de dados em certas tabelas que contem dados monetários para fixed decimal points, como é o caso da coluna *BASE\_ENCOUNTER\_COST, PAYER\_COVERAGE* e *TOTAL\_CLAIM\_COST.* Outras transformações possíveis que possamos realizar é transformar os códigos em tipo numérico, isto beneficiamos porque um número ocupa menos espaço em memória que texto*.*

Agora que estamos com os dados já no tipo correto, podemos criar uma coluna com o valor que o paciente pagou do próprio bolso OUT\_OF\_POCKET\_COST, fazendo uma subtração do *TOTAL\_CLAIM\_COST* e da *PAYER\_COVERAGE.*

Por fim podemos remover todas as colunas que ou sã redundantes ou não serão utilizadas nos nossos visuais.

### Procedures

Outra tabela de fatos, FCT\_Procedures, para que possamos entender oque ocorre em cada encontro, aplicando o mesmo processo de transformação à da tabela de FCT\_FullEncounters.  
Os dados obtidos estão em bom estado não havendo erros, podemos então passar para a verificação e transformação dos tipos de dados, transformamos as colunas de START e STOP no tipo correto DATE/TIME, e as chaves estrangeiras para os PATIENTS e FCT\_FULLENCOUNTERS para texto visto que foi passado em GUID, não sendo possível transformar em numérico, mas podemos transformar os códigos em número visto que seguem a mesma estrutura que a tabela de FCT\_FULLENCOUNTERS, e transformamos o BASE\_COST para decimal visto que representa o custo base do procedimento.

### Patients A tabela DIM\_PATIENTS contem informação sobre todos os pacientes que deram entrada no hospital, podemos ver que existem dados que estão corrompidos e que necessitam de ser limpos, como as colunas FIRST e LAST, contendo o nome dos pacientes, apesar de estarem corrompidos, nos queremos manter o anonimato dos pacientes, dai removermos as colunas que não necessitemos como a FIRST, LAST, MAINDEN, PREFIX, SUFIX, LATITUDE, LONGITUDE, ADDRESS, entre outras. A remoção destas colunas serve para que também possamos otimizar a nossa memória, e não tenhamos informação desnecessária.

### Payers

Tabela com informações referentes às seguradoras, mais uma vez nesta tabela contem informações que para mostrar no nosso dashboard, não são necessárias logo podemos descartar neste processo de ETL.

Codes   
Teremos várias tabelas com códigos tanto de procedimentos, como diagnósticos, onde teremos as descrições como o código associado.

### Calendário

Criamos usando DAX, uma nova tabela com todas as datas existentes dentro da tabela FCT\_FULLENCOUNTERS, podemos assim extrair informações extras sobre as datas de encontros.

## Carregamento

O carregamento é a última parte do processo ETL, onde podemos colocar todas as transformações em prática e visualizar todas as novas tabelas de dimensão e de fatos.

## Relações

As relações realizadas nesta etapa irá ditar a estrutura dos modelos de dados, criando assim ligações entre as tabelas de fatos e dimensão, podemos assim criar visuais com informações agregadas de todos os lados.

O meu modelo de dados é o seguinte:

Uma imagem com texto, captura de ecrã, Retângulo, diagrama

Descrição gerada automaticamente

Figura 1 Modelo de dados

As duas tabelas que não possuem nenhum tipo de ligação, é a tabela das medidas que foram criadas ao longo do dashboard, assim a sua localização é facilitada não sendo necessária sua pesquisa em todas as outras possíveis tabelas.   
A tabelas das ORGANIZATIONS, apenas contem a informação relativa ao hospital, sendo todas entradas no mesmo hospital não é necessário ter esse desperdiço de memória alocado.

# Dashboard

O dashboard, terá informações relativa ao hospital e aos seus pacientes, quantos homens, mulheres, as suas origens e etnicidades, no seguimento teremos informações mais informações e métricas sobre diversas questões impostas.

### Quantos doentes foram admitidos ou readmitidos ao longo do tempo?

### Quanto tempo, em média, os doentes permanecem no hospital?

### Qual o custo médio por visita?

### Quantos procedimentos estão cobertos pelo seguro?