Universidade do Minho
Escola de Engenharia
Mestrado Integrado em Engenharia Informática

Unidade Curricular de Análise de Dados

Ano Letivo de 2017/2018

Relatório do Trabalho Prático

Bruno Pereira (a75135), Luís Fernandes (a74748), Maria Ana de Brito (a73580)

Janeiro, 2018

AD

Índice

Índice de Figuras

Índice de Tabelas

1. Introdução

O projeto prático tem como objetivo a aplicação de métodos de *Business Intelligence* para a extração de informação de um conjunto de dados. A análise deve ser feita tendo em conta o ponto de vista de um analista, justificando todo o fluxo de informação presente no sistema.

O *dataset* escolhido, denominado *Bank Marketing,* está relacionado com campanhas de *marketing* de um determinado banco português, onde é apresentada informação acerca dos clientes contactados, que variam desde os seus dados pessoais, tais como a idade e o emprego, até ao número de vezes que já foi contactado previamente no âmbito de campanhas anteriores. Após uma análise dos dados, criou-se, então, uma base de dados relacional *MySQL* que é capaz de armazenar os dados do *dataset*. Desenvolveu-se, também, um *data warehouse* com as devidas dimensões.

Por fim, fez-se uma análise dos dados retirados e tiraram-se as respetivas conclusões dos mesmos.

Área de Aplicação: Análise de Dados, *Business Intelligence*, *Data warehouse*, Base de dados, *Talend*

Palavras-Chave: banco, dados, cliente, base de dados, *data warehouse*

2. Análise dos atributos do *dataset Bank Marketing*

Antes do desenho do modelo lógico da base de dados, foi necessário analisar os atributos tentando entender que relações existiam entre eles.

Este *dataset* refere-se a uma chamada efetuada a um determinado cliente. Sendo ele um *dataset* cujo propósito é ser avaliado por modelos de classificação, temos que existe um atributo que representa a classe desse modelo de classificação, sendo ele o ***y***, e a partir do qual podemos concluir se a chamada ocorreu com sucesso ou não.

Os atributos e o seu significado encontram-se especificados da seguinte forma:

- **age:** idade;
- **job:** tipo de emprego;
- **marital:** estado matrimonial;
- **education:** grau de educação;
- **default:** se possui crédito;
- **housing:** se tem empréstimo à habitação;
- **loan:** se tem algum empréstimo pessoal;
- **contact:** tipo de comunicação;
- **month:** mês do ultimo contacto;
- **day_of_week:** dia da semana do último contacto;
- **duration:** duração do último contacto;
- **campaign:** número de contactos realizados na campanha;
- **pdays:** número de dias passados desde um contacto relativamente a uma campanha anterior;
- **previous:** número de contactos realizados antes da campanha atual para este cliente;
- **poutcome:** resultado da última campanha;
- **emp.var.rate:** variação da taxa de empregabilidade (quadrimestral);
- **cons.price.idx:** índice de preços ao consumidor (mensal);
- **cons.conf.idx**: índice de confiança do consumidor (mensal);
- **euribor3m:** taxa Euribor trimestral;
- **nr.employed:** número de empregados (taxa quadrimestral);
- **y:** representa o resultado da campanha.

3. Modelo lógico da base de dados relacional

Através da análise do *dataset* em questão, conseguimos obter o seguinte modelo lógico.

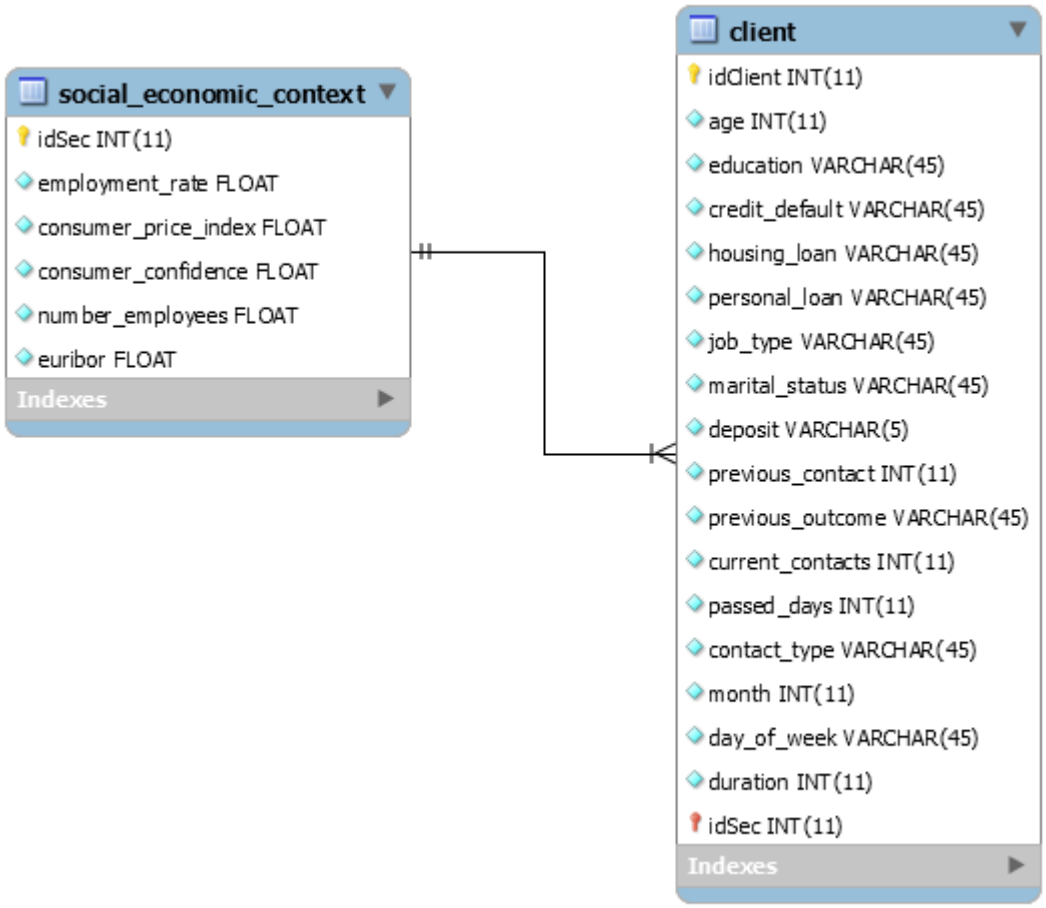

Figura 1 - Modelo lógico da base de dados relacional

A escolha de duas únicas entidades surgiu do facto de se ter concluído que os dados se agrupavam em dois grupos: um grupo com os dados relativos ao cliente (podendo ser eles dados pessoais, dados de campanhas anteriores, ou dados da campanha atual) e dados relacionados com contextos sociais e económicos.

3.1. Descrição das entidades

Tabela 1-Tabela das entidades da base de dados relacional

Nome da entidade	Descrição	Sinónimos	Contexto
client	Termo geral que representa todos os clientes da campanha de *marketing*	Consumidor, comprador	Cada cliente é representado pelos seus dados pessoais, bem como os dados da campanha atual e de campanhas anteriores
social_economic_context	Termo geral que representa os dados sociais e económicos	------	Cada contexto apresenta vários atributos referentes a ambientes sociais e económicos

3.2. Descrição dos relacionamentos

Tabela 2-Tabela dos relacionamentos da base de dados relacional

Entidade	Multiplicidade	Relação	Multiplicidade	Entidade
client	1	tem	N	social_economic_context

Um cliente apresenta pode apresentar mais do que um contexto, pois estes não são dependentes do cliente em si, mas sim de condições sociais e económicas, daí poderem surgir clientes que apresentam as mesmas características pessoais, mas com contextos sociais e económicos diferentes.

3.3. Descrição do cliente

Tabela 3-Tabela cliente

Entidade	Atributos	Descrição	Tipo de dados/Domínio	Nulo/Composto
	idClient	Identifica de forma única o cliente	Inteiro	Não/não
	age	Idade	Inteiro	Não/não

cliente	education	grau de educação	45 caracteres/ Só caracteres textuais	Não/não
	credit_default	se possui crédito	45 caracteres/ Só caracteres textuais	Não/não
	housing_loan	se tem empréstimo à habitação	45 caracteres/ Só caracteres textuais	Não/não
	personal_loan	se tem algum empréstimo pessoal	45 caracteres/ Só caracteres textuais	Não/não
	job_type	tipo de emprego	45 caracteres/ Só caracteres textuais	Não/não
	marital_status	estado matrimonial	45 caracteres/ Só caracteres textuais	Não/não
	Deposit	Resultado da campanha	5 caracteres/ Só caracteres textuais	Não/não
	previous_contact	número de contactos realizados em campanhas prévias	Inteiro	Não/não
	previous_outcome	resultado da última campanha	45 caracteres/ Só caracteres textuais	Não/não
	current_contacts	número de contactos realizados na campanha	Inteiro	Não/não
	passed_days	número de dias passados desde a última campanha	Inteiro	Não/não
	contact_type	tipo de comunicação;	45 caracteres/ Só caracteres textuais	Não/não
	month	mês do último contacto	Inteiro	Não/não
	day_of_week	dia da semana do último contacto	45 caracteres/ Só caracteres textuais	Não/não
	duration	duração do ultimo contacto;	Inteiro	Não/não

3.4. Descrição do social_economic_context

Tabela 4-Tabela social_economic_context

Entidade	Atributos	Descrição	Tipo de dados/Domínio	Nulo/Composto
	idSec	Identifica de forma única o	Inteiro	Não/não

Social_economic_context		contexto social e económico		
	employment_rate	variação da taxa de empregabilidade (quadrimestral)	Decimal	Não/não
	consumer_price_index	índice de preços ao consumidor (mensal);	Decimal	Não/não
	consumer_confidence	índice de confiança do consumidor (mensal)	Decimal	Não/não
	number_employees	número de empregados (taxa quadrimestral)	Decimal	Não/não
	euribor	taxa Euribor trimestral	Decimal	Não/não

4. Perguntas Iniciais

Uma das primeiras tarefas passou pela definição de perguntas (*queries*) que pretendíamos ver respondidas com a ajuda do Power BI, para que desta forma pudéssemos organizar a informação nos Data Marts convenientemente pensados para este efeito.

Foram então elaboradas as seguintes questões:

- Número de depósitos (‘yes’ e ‘no’ obtidos) por tipo de emprego (job_type), idade (age), estado civil (marital_status), grau de educação (education), empréstimos anteriores (housing_loan, personal_loan) e crédito em atraso (credit_default);

- Número de depósitos por respostas a anteriores campanhas (previous_outcome);

- Número de depósitos por tentativas de contacto da campanha em questão (previous_contacts), tipo de contacto (contact_type), mês do último contacto (month), dia da semana do último contacto (day_of_week)

- Número de depósitos por contexto socioeconómico, taxa de empregabilidade (employment_rate), Euribor.

5. Modelo Lógico do Data Warehouse

5.1. Dimensões do Data Warehouse

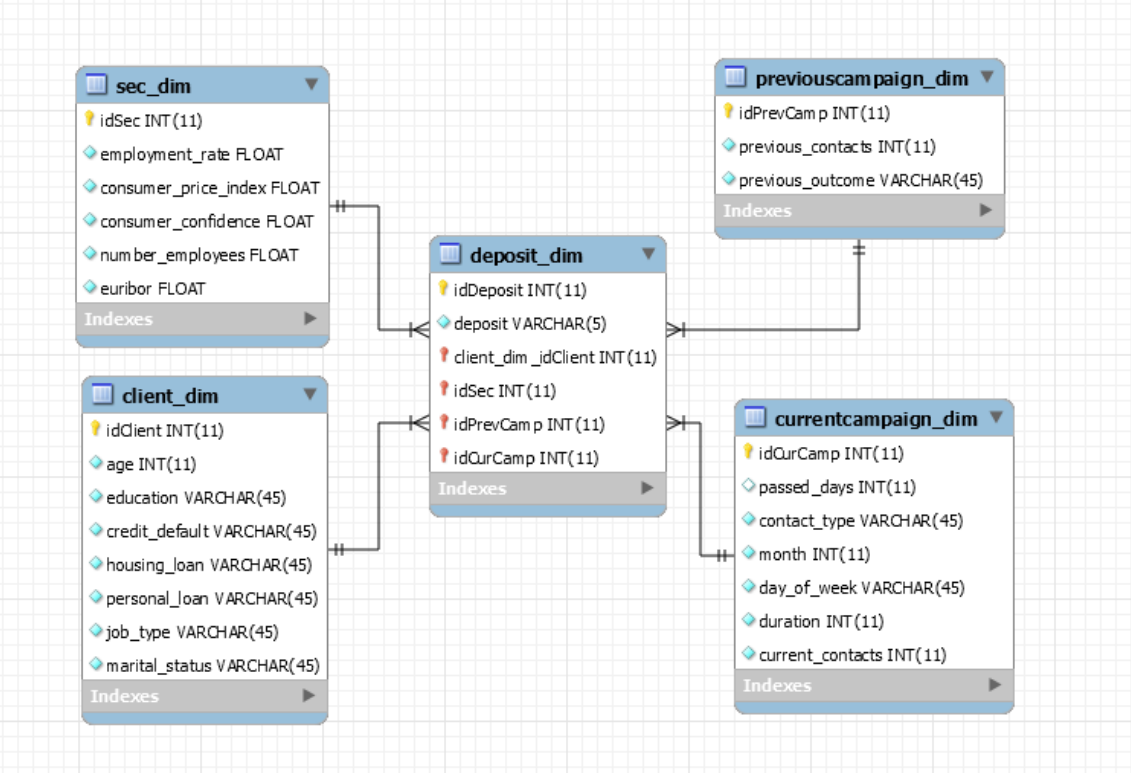

Figura 2 - Modelo lógico do data warehouse

Analisando os dados presentes no *dataset Bank Marketing* consegue-se distinguir claramente quatro dimensões:

- Informação pessoal do cliente:
  - Idade
  - Educação
  - Se tem crédito em atraso
  - Se tem um empréstimo da casa
  - Se tem um empréstimo pessoal

    - Emprego
    - Estado civil
- Campanha anterior:
    - Número de contactos feitos
    - Resultado da última campanha
- Campanha atual:
    - Número de dias que já passaram desde o último contacto
    - Tipo de contacto
    - Mês do contacto
    - Dia da semana do contacto
    - Duração da chamada
    - Número de contactos nesta campanha
- Dados socioeconómicos
    - Índice de confiança do consumidor
    - Índice de preço do consumidor
    - Taxa de emprego
    - Euribor
    - Média de número de empregados na empresa

Estas dimensões conjugam-se numa única dimensão, a dimensão do depósito, isto é, a resposta a esta campanha que o banco está a promover. Caso a resposta seja "yes", então o cliente concordou em aderir à campanha. Uma resposta negativa indica o contrário.

Esta divisão dos dados em dimensões revela uma clara separação em diferentes áreas. Assim, para uma posterior análise dos dados teremos em conta estas dimensões.

Podemos inquirir acerca do número de respostas positivas e negativas acerca de informações pessoais dos clientes para conhecer melhor o público. Por exemplo, podemos querer analisar qual a faixa etária que mais responde positivamente a este tipo de campanhas do banco e qual o seu nível de educação ou que tipo de emprego desempenha.

Por outro lado, podemos também querer perceber se o banco está a apostar numa nova clientela ou se contacta maioritariamente clientes passados. Além disso, também podemos querer saber qual o dia da semana ou qual o mês em que a campanha tem mais sucesso.

Estas e outras perguntas são possíveis de ser respondidas, graças a esta organização intuitiva e clara dos dados.

5.2. Dicionário de Dados

De forma a esclarecer quaisquer dúvidas que existam acerca da ambiguidade que os dados possam possuir, definimos um dicionário de dados relativo ao *data warehouse*, onde apresentamos as características das entidades envolvidas, bem como dos seus relacionamentos e atributos.

5.2.1 Descrição das Entidades

Tabela 5 - Descrição das entidades

Nome da Entidade	Descrição	Sinónimos	Contexto
Sec_dim	Termo geral que representa os dados socioeconómicos	Dimensão socioeconómica	Dados socioeconómicos que se verificaram ao longo do decorrer da campanha
Client_dim	Termo geral que representa os dados pessoais do utilizador	Dimensão do cliente	Informação pessoal dos clientes que foram contactados pelo banco
Previouscampaign_dim	Termo geral que representa os dados recolhidos das campanhas anteriores	Dimensão das campanhas anteriores	Dados referentes a campanhas anteriores promovidas pelo banco
Currentcampaign_dim	Termo geral que representa os dados recolhidos durante a campanha atual	Dimensão da campanha atual	Dados recolhidos acerca da campanha atual
Deposit_dim	Termo geral que representa a conjugação das dimensões anteriores e das respostas à campanha em vigor	Dimensão do depósito	Resultado da campanha atual do banco

Nesta tabela podemos verificar quais são as entidades que permitem o armazenamento dos dados no *data warehouse* e o que elas significam.

5.2.2 Descrição dos relacionamentos

Tabela 6 - Descrição dos relacionamentos

Entidade	Multiplicidade	Relação	Multiplicidade	Entidade
Sec_dim	1	Associa-se	N	Deposit_dim
Client_dim	1	Associa-se	N	Deposit_dim
Previouscampaign_dim	1	Associa-se	N	Deposit_dim
Currentcampaign_dim	1	Associa-se	N	Deposit_dim

5.2.3 Descrições das dimensões

Tabela 7 - Descrição da dimensão socioeconómica

Entidades	Atributos	Descrição	Tipo de Dados	Nulo
Sec_dim	idSec	Identifica cada registo de dados socioeconómicos	Number	Não
	employment_rate	Taxa de emprego	VARCHAR2	Não
	consumer_price_index	Índice de preço do consumidor	VARCHAR2	Não
	consumer_confidence	Índice de confiança do consumidor	VARCHAR2	Não
	number_employees	Média do número de empregados na empresa	VARCHAR2	Não

Nesta tabela podemos verificar quais os atributos da dimensão socioeconómica, assim como saber o que significam e descobrir quais são os seus tipos.

Tabela 8 - Descrição da dimensão do cliente

Entidades	Atributos	Descrição	Tipo de Dados	Nulo
Client_dim	idClient	Identifica unicamente cada cliente	Number	Não
	age	Idade do cliente	VARCHAR2	Não
	education	Nível de educação do cliente	VARCHAR2	Não
	credit_default	Verifica se cliente tem crédito em atraso	VARCHAR2	Não
	housing_loan	Verifica se tem empréstimo da casa	VARCHAR2	Não
	personal_loan	Verifica se tem empréstimo pessoal	VARCHAR2	Não
	job_type	Emprego do cliente	VARCHAR2	Não
	Marital_status	Estado civil do cliente	VARCHAR2	Não

Nesta tabela podemos verificar quais os atributos da dimensão do cliente, assim como saber o que significam e descobrir quais são os seus tipos.

Tabela 9 - Descrição da dimensão das campanhas anteriores

Entidades	Atributos	Descrição	Tipo de Dados	Nulo
Previouscampaign_dim	idPrevCamp	Identifica unicamente cada registo	Number	Não

	previous_contacts	Número de contactos feitos em campanhas anteriores	VARCHAR2	Não
	previous_outcome	Resultado da campanha anterior	VARCHAR2	Não

Nesta tabela podemos verificar quais os atributos da dimensão das campanhas anteriores, assim como saber o que significam e descobrir quais são os seus tipos.

Tabela 10 - Descrição da dimensão da campanha anterior

Entidades	Atributos	Descrição	Tipo de Dados	Nulo
Currentcampaign_dim	idCurCamp	Identifica unicamente cada registo	Number	Não
	passed_days	Número de dia que passaram desde o último contacto	VARCHAR2	Não
	contact_type	Tipo de contacto realizado	VARCHAR2	Não
	month	Mês em que foi realizado o contacto	VARCHAR2	Não
	day_of_week	Dia da semana em foi realizado o contacto	VARCHAR2	Não
	Duration	Duração do último contacto	VARCHAR2	Não

	Current_contacts	Número de contactos realizados durante a campanha atual	VARCHAR2	Não

Nesta tabela podemos verificar quais os atributos da dimensão da campanha atual, assim como saber o que significam e descobrir quais são os seus tipos.

Tabela 11 - Descrição da dimensão do depósito

Entidades	Atributos	Descrição	Tipo de Dados	Nulo
Deposit_dim	idDeposit	Identifica unicamente cada resultado da campanha	Number	Não
	deposit	Resultado da campanha atual	VARCHAR2	Não

Esta tabela é a recipiente das chaves estrangeiras das outras tabelas que correspondem às dimensões do problema. Aqui podemos verificar qual foi o resultado da campanha em vigor para cada cliente, tendo em conta os valores dos dados socioeconómicos presentes na altura e os resultados de campanhas anterior promovidas pelo mesmo banco.

6. Divisão dos dados em csv e sql

A divisão do *dataset* original de *bankmarketing* surge de um dos pontos do enunciado do problema exigir a obtenção dos dados, para constituir o *data warehouse,* de duas fontes diferentes, sendo elas uma base de dados relacional e um ficheiro do tipo *csv*.

A divisão efetuada passou por dividir o ficheiro *csv* original em duas partes de igual dimensão. Uma dessas partes seria utilizada para preencher a base de dados relacional, para depois se extraírem esses dados pelo *Talend.* Relativamente ao ficheiro *csv*, os dados deste serão obtidos diretamente através do *Talend*.

7. Utilização do Software Talend

A utilização do Talend teve vários objetivos (estando divididos por Jobs no projeto no software em causa), que serão explicitados de seguida.

7.1. Divisão em duas Fontes de Informação

Numa primeira abordagem, o Talend foi usado para colocar informação do ficheiro de dados CSV numa base de dados, servindo assim ao nosso objetivo de possuir duas fontes distintas de informação, base de dados e CSV diretamente, como anteriormente referido no ponto 6.

Fig. 1 – Passagem dos dados do ficheiro *.csv* para a base de dados (preenchimento da tabela SEC (social_economic_context))

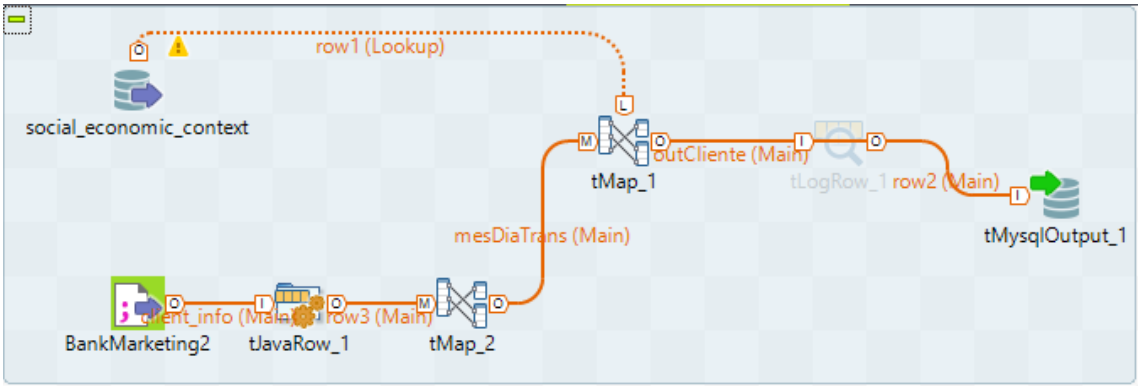

Fig. 2 – Passagem dos dados do ficheiro *.csv* para a base de dados (preenchimento da tabela Cliente)

Começou-se por preencher a tabela *social_economic_context* com os tuplos de valores únicos presentes no CSV, uma vez que na tabela *cliente* existe uma FK para esta tabela mencionada, necessitando desta previamente preenchida.

Neste primeiro processo, o tMap apresentado (tMap_1) é utilizado exclusivamente para seleção das colunas do ficheiro CSV a passar para a table *social_economic_context*.

De seguida foi então preenchida a tabela *cliente*, em que cada entrada irá corresponder a uma linha do ficheiro CSV original, possuindo ainda para cada entrada, como referido anteriormente, uma FK para a tabela *social_economic_context*.

Aqui podem ser encontrados dois tMap's, sendo que o tMap_2 tem a mesma função do mencionado no passo anterior, seleção de colunas, ao passo que o tMap_1 é utilizado para fazer *join* com a tabela *social_economic_context* previamente preenchida, com o intuito de retirar o *idSec* que corresponde aos valores procurados neste *join*.

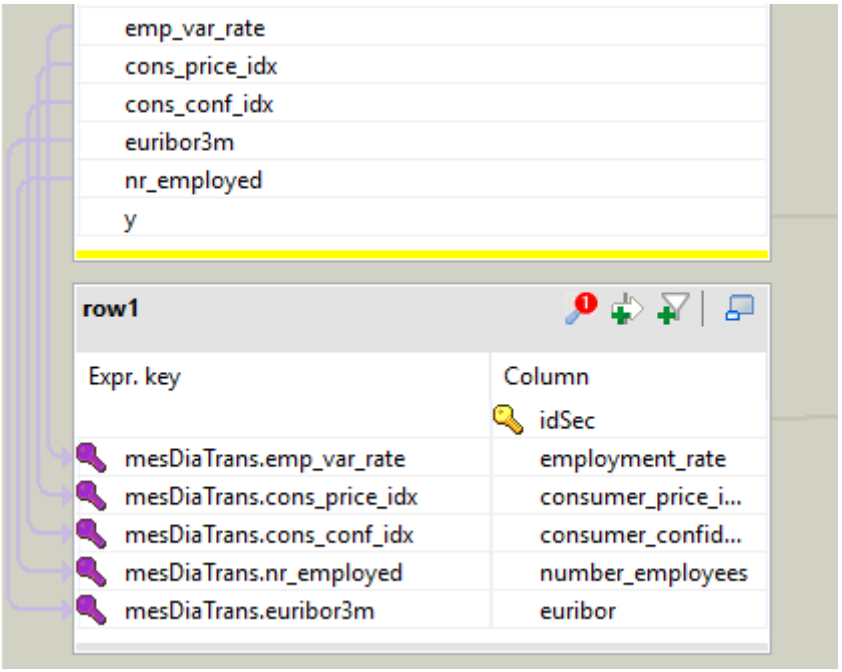

Fig. 3 – Join presente no tMap_1

7.2. Tratamento dos Dados

O Talend foi também peça fundamental no que diz respeito ao tratamento dos dados, uma vez que permitiu, por exemplo, a passagem dos valores do atributo *month* de {“jan”, “feb”, …} para {1, 2, …} respetivamente ou dos valores do *atributo day_of_week* de {“mon”, “tue”, …} para {“monday”, “tuesday”, …} respetivamente.

```
if(current1.month.equals("jan"))
row5.month = 1;
if(current1.month.equals("feb"))
row5.month = 2;
if(current1.month.equals("mar"))
row5.month = 3;
if(current1.month.equals("apr"))
row5.month = 4;
if(current1.month.equals("may"))
row5.month = 5;
if(current1.month.equals("jun"))
row5.month = 6;
if(current1.month.equals("jul"))
row5.month = 7;
if(current1.month.equals("aug"))
row5.month = 8;
if(current1.month.equals("sep"))
row5.month = 9;
if(current1.month.equals("oct"))
row5.month = 10;
if(current1.month.equals("nov"))
row5.month = 11;
if(current1.month.equals("dec"))
row5.month = 12;
if(current1.day_of_week.equals("mon"))
row5.day_of_week = "monday";
if(current1.day_of_week.equals("tue"))
row5.day_of_week = "tueday";
if(current1.day_of_week.equals("wed"))
row5.day_of_week = "wednesday";
if(current1.day_of_week.equals("thu"))
row5.day_of_week = "thursday";
if(current1.day_of_week.equals("fri"))
row5.day_of_week = "friday";
if(current1.day_of_week.equals("sat"))
row5.day_of_week = "saturday";
if(current1.day_of_week.equals("sun"))
row5.day_of_week = "sunday";
```

Fig. 4 – Tratamento dos dados

7.3. Preenchimento dos *Data Marts*

Mais tarde, o software em questão, foi ainda utilizado para efetuar o preenchimento dos vários *Data Marts* que compõem o nosso *Data Warehouse*, sendo os dados provenientes tanto do CSV diretamente, como da base de dados, ficando este pronto a ser estudado com recurso à ferramenta Power BI.

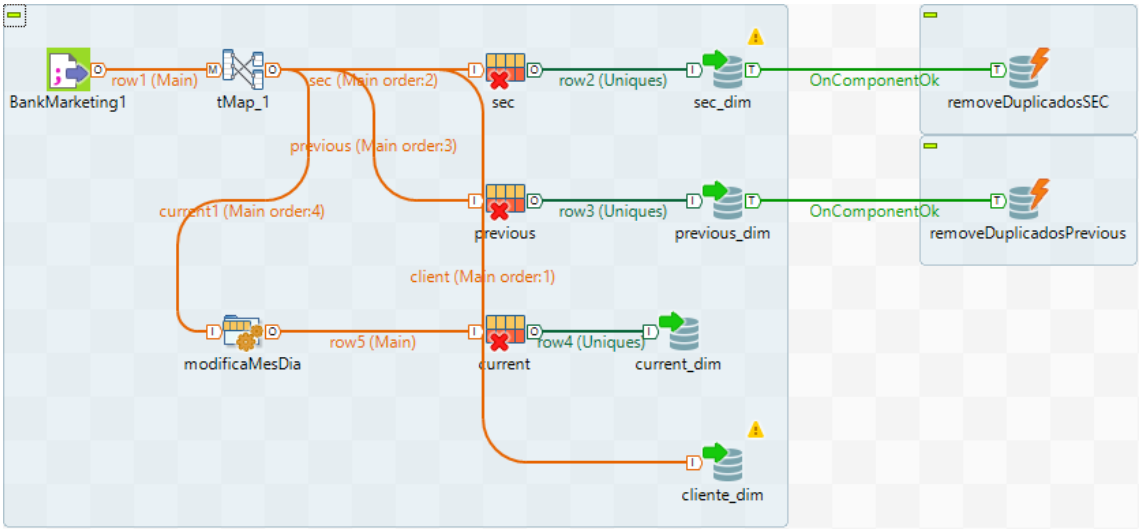

Fig. 5 – Preenchimento dos *Data Marts* com dados provenientes do *CSV*

O processo passa por utilizar o *tMap_1* para selecionar as colunas que seguirão para cada uma das dimensões a preencher, sendo que no caso do preenchimento das dimensões *sec_dim*, *previous_dim* e *current_dim* foi necessário o uso de um *tUniqueRow* para que não fossem colocadas informações duplicadas nas tabelas.

Salientar ainda o uso de um *tJavaRow* para proceder à transformação dos dados, como mencionado no ponto anterior, no caso do preenchimento da dimensão *current_dim*.

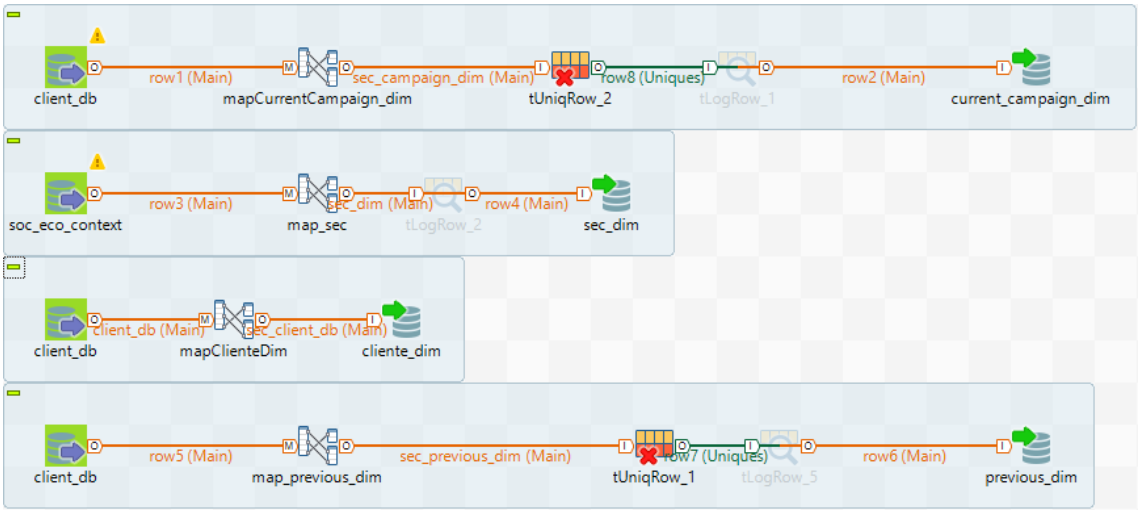

Fig. 6 – Preenchimento dos *Data Marts* com dados provenientes da base de dados

Aqui o processo assemelha-se ao anteriormente descrito, diferenciando-se apenas no facto da informação ser proveniente da base de dados ao invés do CSV.

De notar que, as restantes diferenças que poderão ser vislumbradas acontecem pela preocupação de não inserir informação duplicada na tabela *social_economic_context*, daí a não existência de um *tUniqueRow* nesta abordagem, bem como da preocupação da transformação dos dados aquando da inserção na tabela *cliente*, logo a inexistência do *tJavaRow*.

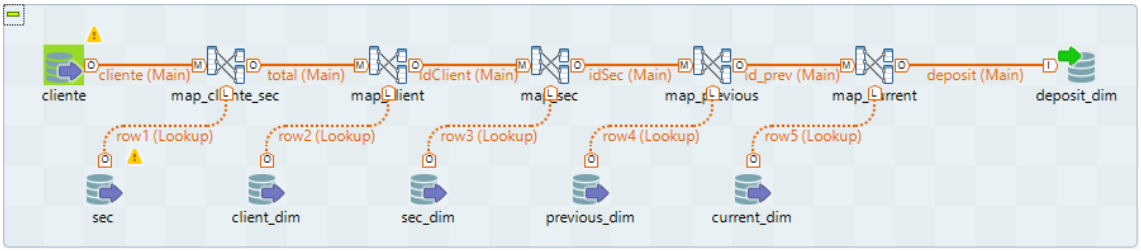

Fig. 7 – Preenchimento da dimensão central do *Data Warehouse*

Uma vez que a dimensão *deposit_dim* é composta, maioritariamente, por chaves estrangeiras das restantes dimensões, é neste *job* executado todo o processo de obtenção dessas chaves a inserir na dimensão dita.

Cada *tMap* apresentado vai proceder a um *join* com parte da informação que segue do momento imediatamente anterior, à exceção do primeiro, que serve para compor cada entrada com todos os valores da tabela *cliente*, bem como os presentes na tabela *social_economic_context* correspondestes à entrada nessa tabela com o *idSec* que se apresenta na primeira referida. Os restantes *tMap* utilizarão, como referido, as informações que seguem dos momentos anteriores, isto é, no segundo *tMap* apresentado, será executado um *join* com todas as informações relativas à *client_dim* de forma a encontrar o id associado, passando para a próxima fase as informações não utilizadas, e o id encontrado. Este processo é repetido em cada *tMap* para que no fim constem apenas os id's pretendidos para preencher a *deposit_dim* e o atributo *y* (sendo que este é o único atributo que segue de início a fim o processo, uma vez que é pretendido que seja guardado na dimensão referida, sendo o único atributo que não é chave estrangeira).

8. Criação e Explicitação dos *Reports*

Após o preenchimento do *data warehouse* com os dados provenientes tanto da base de dados como do ficheiro CSV, procedeu-se à criação de relatórios (*reports*) contendo diversos gráficos que ilustram os dados acerca dos resultados da campanha do banco relativa a depósitos.

8.1. Report 1

Figura 3 - Report 1

Neste *report* apresentamos cinco gráficos distintos e uma contagem. Esta revela o número de resultados negativos da campanha, assim como o número de resultados positivos

que se obteve. Consegue-se identificar um número de respostas negativas muito mais elevado do que respostas positivas, pois temos quase oito vezes mais “no” do que “yes”.

8.1.1 Gráfico 1

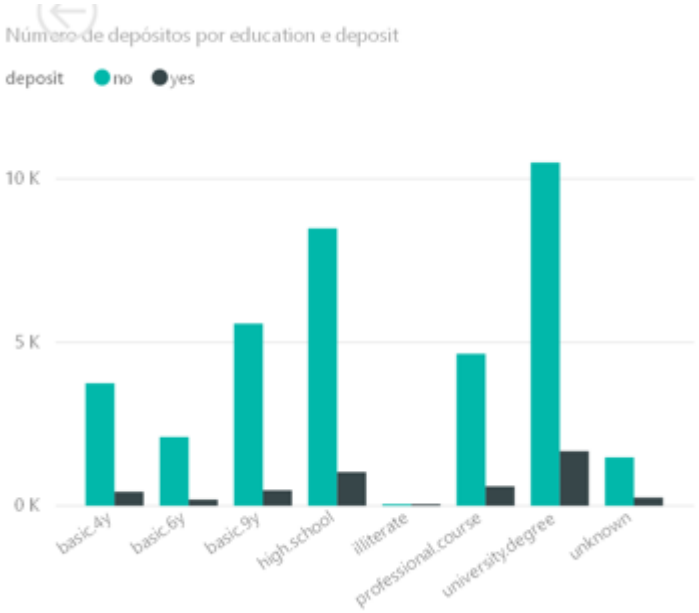

Figura 4 – Depósitos por educação

Por outro lado, o primeiro gráfico mostra a relação existente entre a proporção de respostas “yes” e ”no” e a educação do cliente, onde se verifica que os clientes com uma educação até ao 9º ano possuem uma maior tendência para recusar a adesão à campanha bancário. Dos 6045 clientes contactados com esta educação, 5572 recusaram fazer um depósito no banco, o que constitui cerca de 92.1% de clientes dessa categoria.

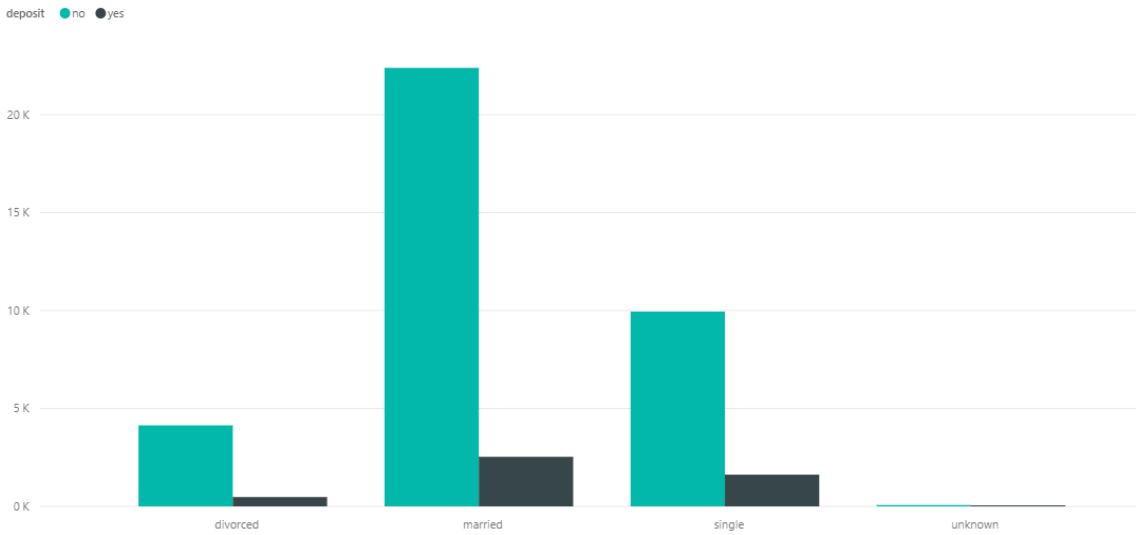

Figura 5 - Depósitos por estado civil

Este gráfico apresenta os resultados por estado civil do cliente. Como era de esperar, a maioria dos clientes pertencem à classe *married*. Por estas razões, registam o maior número de chamadas.

8.1.2 Gráfico 2

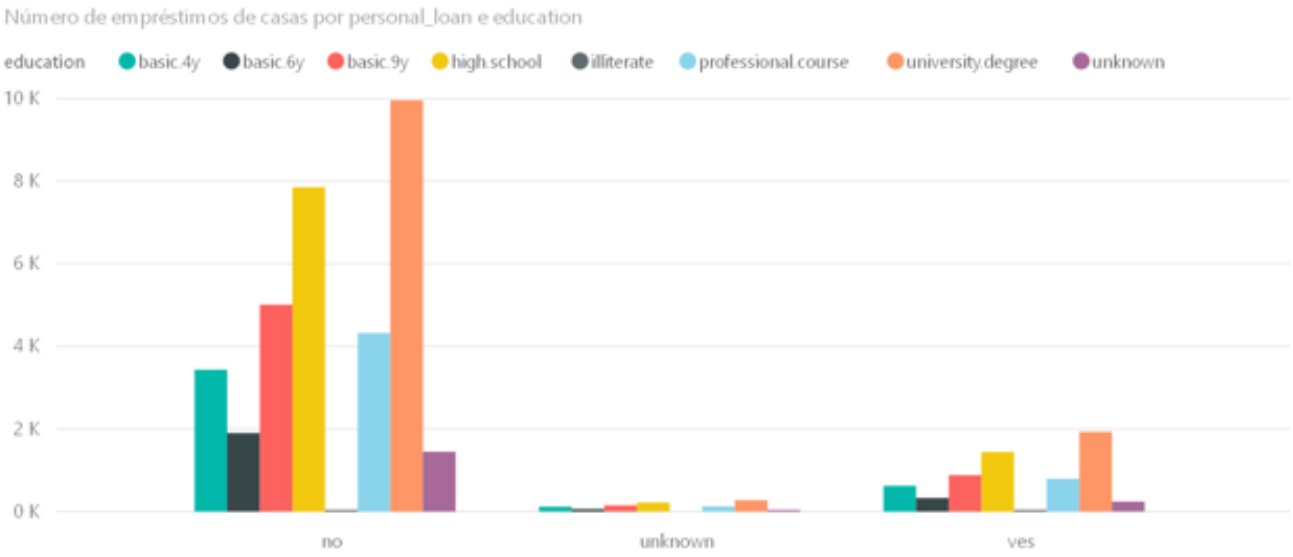

Figura 6 – Empréstimos de casas por empréstimos pessoais e educação

Este gráfico tem como objetivo conhecer a população de clientes que o banco contactou. Assim, apresentamos o número de clientes que possuem um empréstimo de habitação segundo um empréstimos pessoais e educação. Verifica-se que os clientes com

licenciatura que não possuem um empréstimo de habitação verificam o maior número de empréstimos pessoais.

Também apresentamos no *report* o cenário de empréstimos sobre a casa por educação.

8.1.3 Gráfico 3

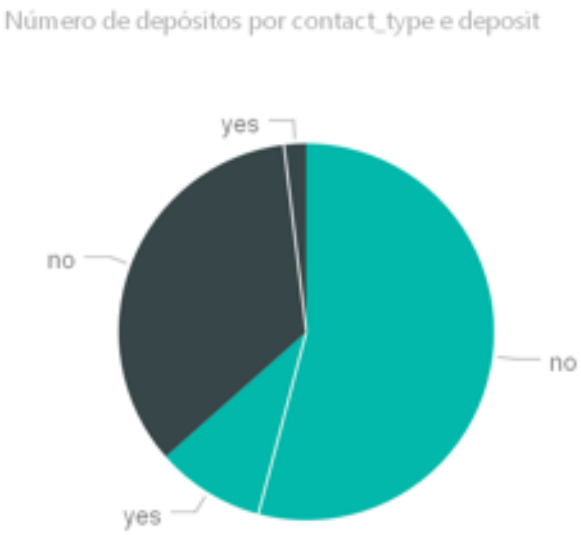

Figura 7 - Depósitos por tipo de contacto

Neste gráfico apresentamos os depósitos (categorizados em “yes” e “no”) segundo o tipo de contacto (como se verifica na figura a maior parte dos contactos são feitos através do telemóvel – azul claro). Também variamos a categoria, apresentando os depósitos por mês e dia da semana.

8.1.4 Gráfico 4

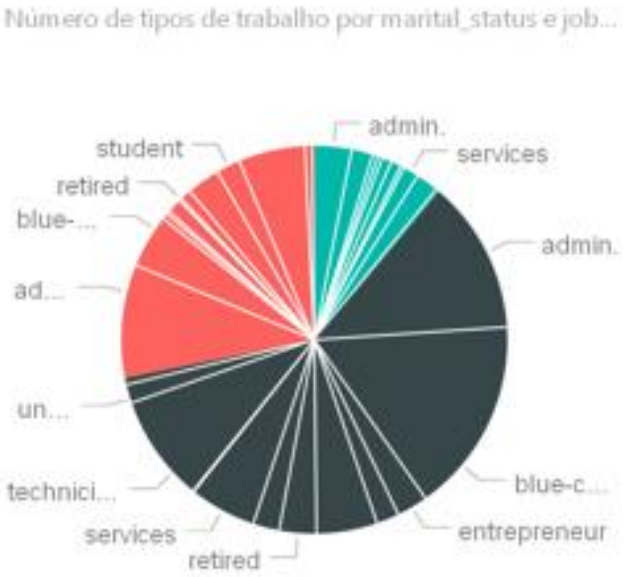

Figura 8 - Tipo de emprego por estado civil

Podemos observar o tipo de emprego que os clientes têm segundo o estado civil. Como se consegue verificar a maior parte dos clientes são casados (cor preta), sendo que os clientes solteiros constituem a segunda maior parcela (cor vermelha), enquanto que os divorciados são os menos frequentes (cor azul). Portanto, um cliente casado tem maior probabilidade de ter um trabalho de “colarinho branco”, ou seja, pertence à classe trabalhadora que executa tarefas semiprofissionais, tais como, tarefas administrativas e burocráticas.

8.2. Report 2

Figura 9 - Report 2

Neste *report* apresentamos vários gráficos acerca dos resultados dos depósitos, dependendo de vários fatores, que podemos filtrar por estado civil do cliente usando as opções no lado direito.

8.2.1 Gráfico 1

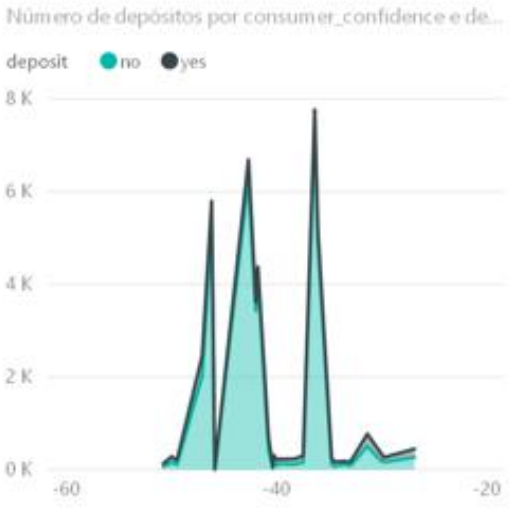

Figura 10 - Depósitos por índice de confiança do consumidor

Neste gráfico apresentamos os resultados da campanha por confiança do consumidor. Regista-se um número elevado de chamadas quando o índice de confiança do consumidor é de cerca de -36.40, pois, como o valor do índice é baixo, os consumidores (clientes do banco) estão menos otimistas em relação ao futuro e gastam menos, logo o banco tem necessidade de cativar novos clientes.

8.2.2 Gráfico 2

Figura 11 - Número de chamadas por taxa de emprego

Este gráfico mostra a relação entre o número de chamadas feitas durante a campanha segundo a taxa de emprego registada. Quando a taxa de emprego aumenta, nota-se uma subida do número de chamadas realizado, uma vez que existem mais pessoas empregadas, logo o banco identifica uma boa oportunidade de aderir mais cliente.

8.2.3 Gráfico 3

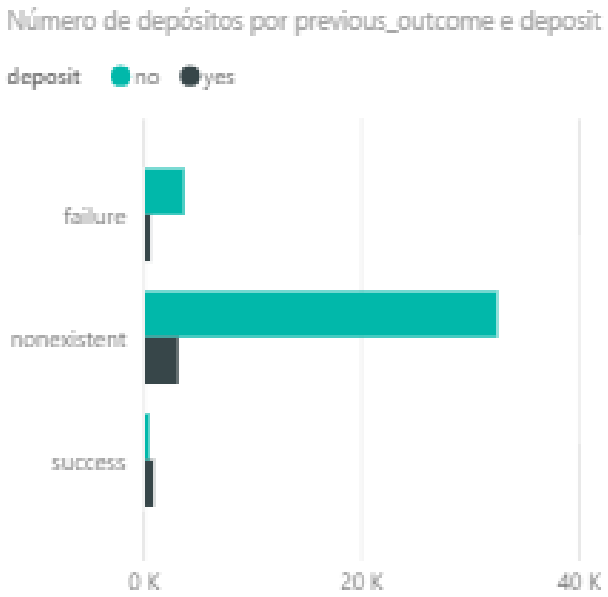

Figura 12 - Depósitos por resultado da campanha anterior

Neste gráfico apresentamos os resultados dos depósitos (“yes”/”no”) categorizados pelo resultado da campanha anterior. Verifica-se que a maioria dos contactados nunca tinha sido alvo das campanhas deste banco.

8.2.4 Gráfico 4

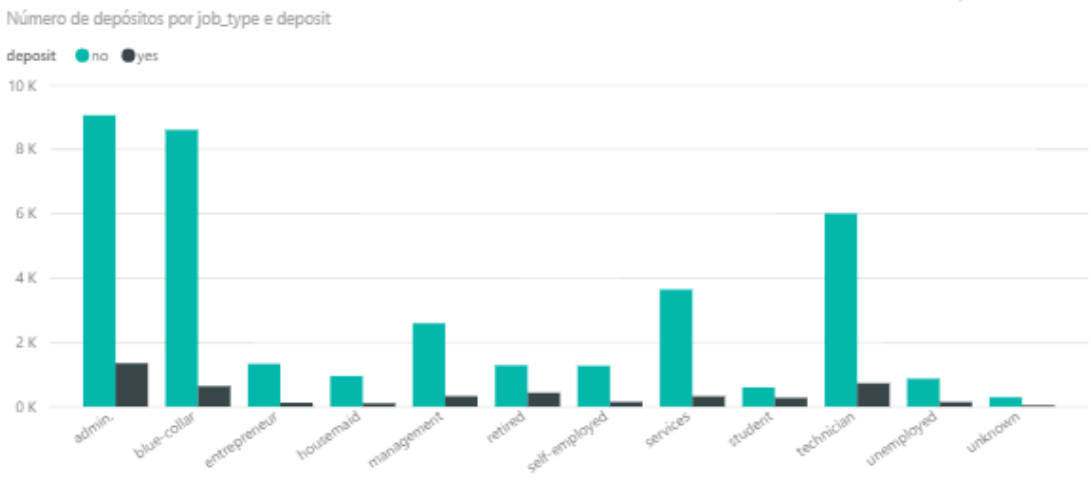

Figura 13 - Depósitos por tipo de emprego

Neste *report* apresentamos os resultados da campanha de depósitos dependendo do tipo de emprego dos clientes. Regista-se que a maior parte dos clientes contactados são administrativos ou “blue-collars”.

8.3. Report 3

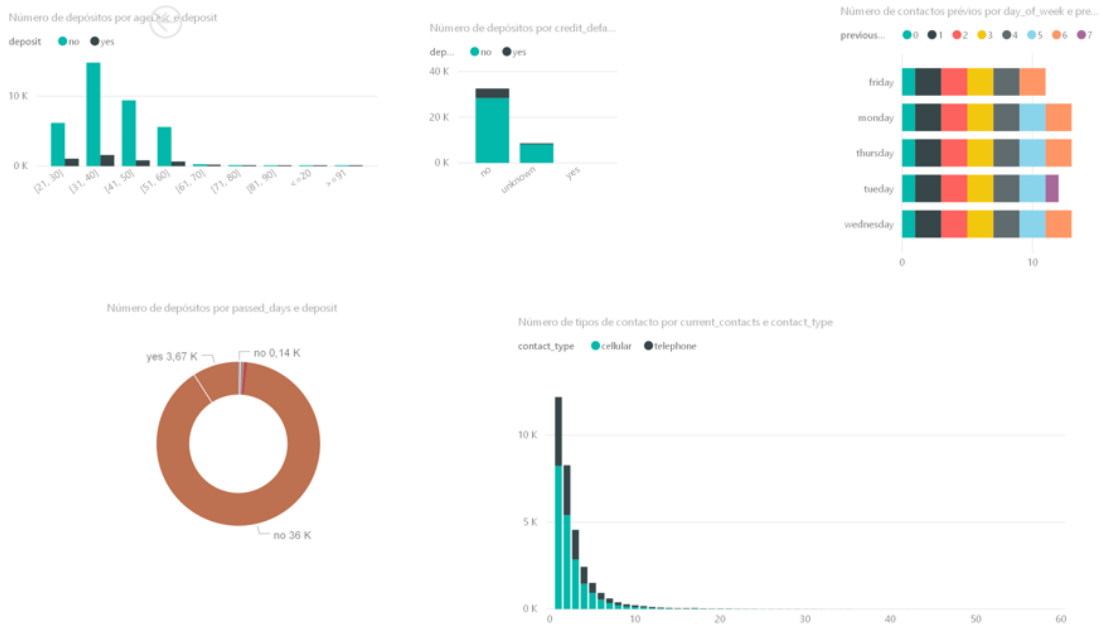

Figura 14 - Report 3

8.3.1 Gráfico 1

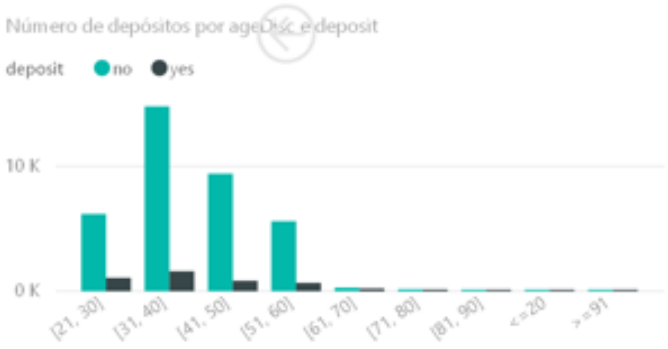

Figura 15 - Depósitos por faixa etária

Neste gráfico apresentamos os resultados dos depósitos por faixa etária. Regista-se um número mínimo de contactos a pessoas cuja idade é inferior ou igual a vinte anos ou superior ou igual a 91 anos, o que faz sentido, visto que se as pessoas têm 20 anos ou menos, só podem fazer depósitos no banco se forem maiores de idade (>=18 anos, logo existem menos contactados com essas idades) e pessoas com mais de 91 anos normalmente já não fazem depósitos em bancos. Assim, a maior parte dos contactados situa-se entre os 21 e os 40 anos. É de notar que a maioria das pessoas contactadas recusam a campanha. Também

apresentamos noutro gráfico os resultados sem ter intervalos de idades, ou seja, sem que a idade esteja discretizada.

8.3.2 Gráfico 2

Figura 16 - Depósitos por crédito em atraso

Neste gráfico apresentamos os resultados da campanha por crédito em atraso. Conseguimos verificar que a maior parte não tem crédito em atraso e que a maioria disse que não à campanha do banco. Também temos fatores como empréstimos sobre a casa e empréstimos pessoais.

8.3.3 Gráfico 3

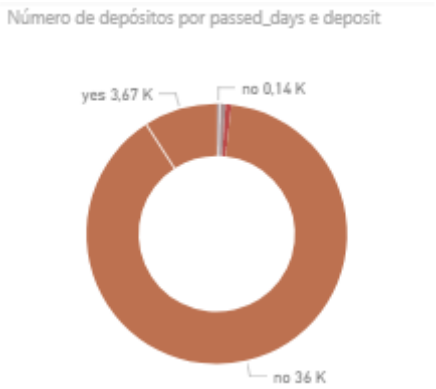

Figura 17 - Depósitos por dias passados

Neste gráfico apresentamos os resultados dos depósitos segundo o número de dias que se passaram desde o último contacto. Quase todos os clientes ainda não tinham sido contactados, sendo que a maioria respondeu “não” à campanha.

8.3.4 Gráfico 4

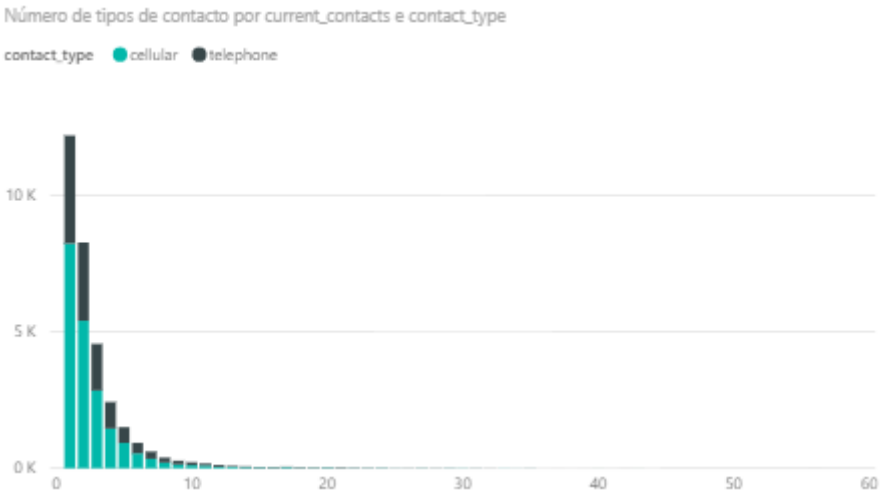

Figura 18 - Tipos de contacto por dias passados

Neste gráfico apresentamos a relação existente entre o tipo de contacto efetuado e o número de dias passados desde a última chamada. No *report* também apresentamos outros fatores, tais como o mês do contacto, dia da semana e número de contactos realizados na campanha atual.

8.4. Report 4

Figura 19 - Report 4

8.4.1 Gráfico 1

Figura 20 - Depósitos por dias passados

Este gráfico mostra o número de depósitos pelo número de dias passados desde a última chamada.

8.4.2 Gráfico 2

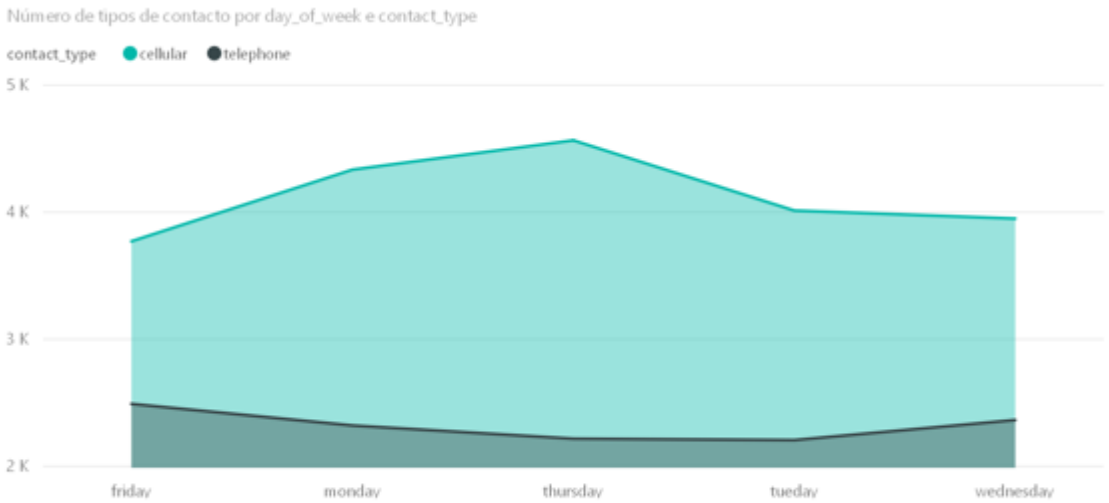

Figura 21 - Tipos de contacto por dia da semana

Aqui apresentamos o tipo de contacto feito durante a campanha, dependendo do dia da semana. Regista-se um grande uso do telemóvel como meio de contacto principalmente no início da semana. Outro fator com que podemos analisar o tipo de contacto é o mês.

8.4.3 Gráfico 3

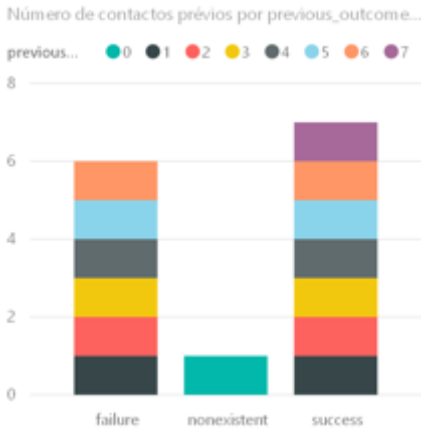

Figura 22 - Número de contactos prévios por resultado da última campanha

O gráfico acima apresenta o número de contactos prévios de campanhas anteriores em relação ao resultado da última campanha.

8.4.4 Gráfico 4

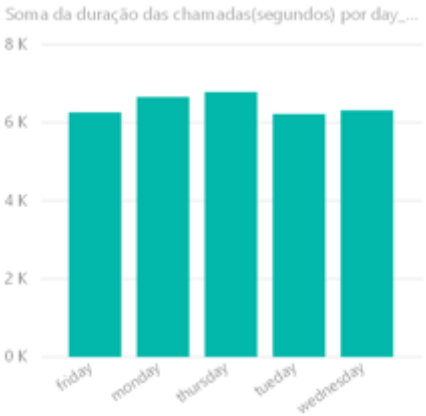

Figura 23 - Soma da duração das chamadas por dia da semana

Neste gráfico é apresentada a soma da duração das chamadas por dia da semana em que foram feitas. Pode-se identificar que se despendeu um maior número de segundos em contacto com o cliente nas terças-feiras. No entanto, o tempo de chamadas em cada dia da semana foi muito semelhante. Também podemos apresentar estes resultados tendo por base o dia da semana e o mês.

8.4.5 Gráfico 5

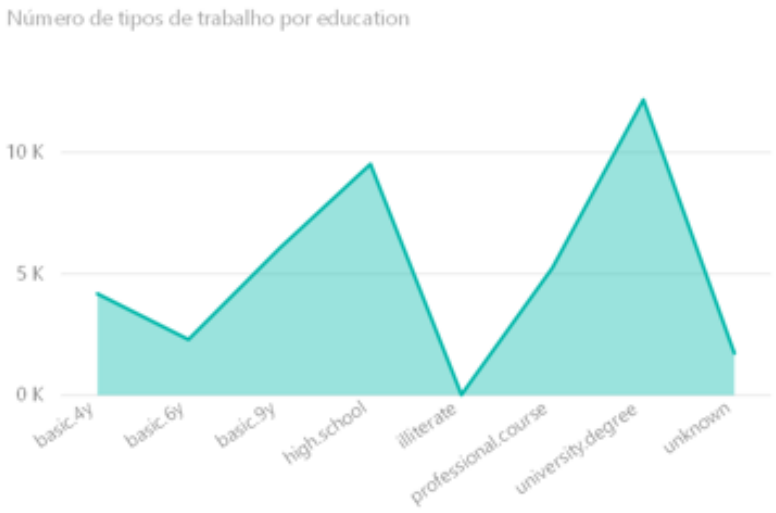

Figura 24 - Número de tipos de emprego por educação

Neste gráfico podemos verificar o número de tipos de emprego por educação do cliente. Podemos registar que a maior parte de pessoas com empregos tem uma licenciatura ou uma educação ao nível do secundário.

8.5. Report 5

Figura 25 - Report 5

8.5.1 Gráfico 1

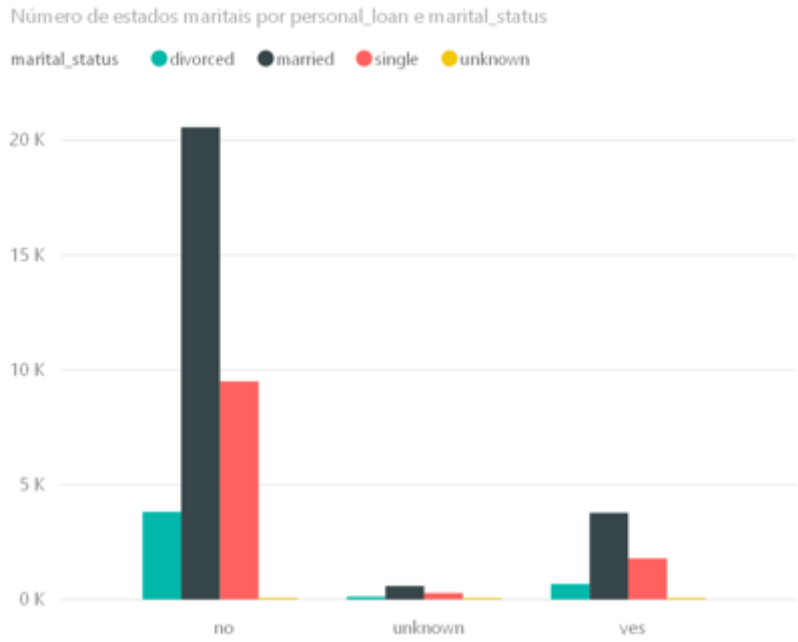

Figura 26 - Estados civis por empréstimos pessoais

Neste gráfico apresentamos os estados civis dos clientes por quem tem empréstimos pessoais ou não (não têm ou não se sabe). Verificamos que a maior parte dos clientes que não possuem empréstimos pessoais são casados. Outro fator são os empréstimos de habitação.

8.5.2 Gráfico 2

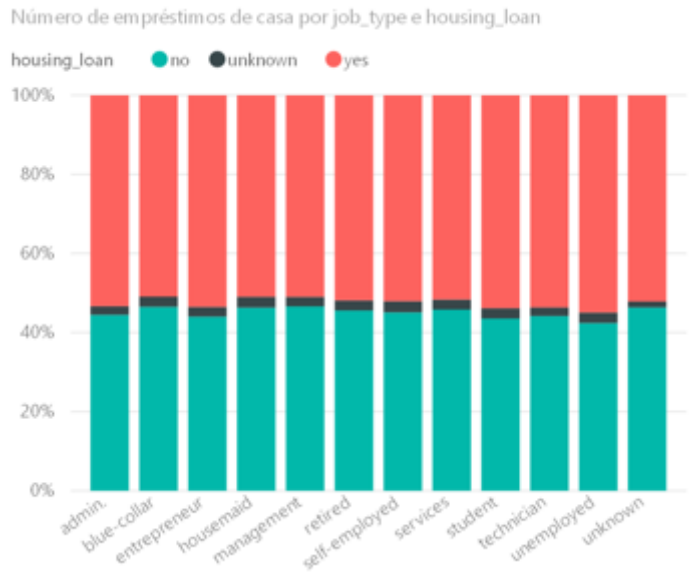

Figura 27 - Empréstimos de habitação por tipo de emprego

Neste gráfico apresentamos os empréstimos de habitação por tipo de emprego. Verifica-se que os resultados são bastante equilibrados, pois para cada tipo de emprego há um número semelhante entre clientes que têm empréstimos e os que não têm empréstimos.

8.5.3 Gráfico 3

Figura 28 - Empréstimos pessoais por tipo de emprego

Neste gráfico apresentamos os clientes que têm (ou não) empréstimos pessoais por tipo de emprego. Assim, verificamos que as pessoas cujo emprego é administrativo, burocrático ou técnico são mais suscetíveis a não terem um empréstimo pessoal.

8.5.4 Gráfico 4

Figura 29 - Depósitos por empréstimos de habitação

Neste gráfico apresentamos o número de depósitos/chamadas feitas (sem distinção entre “yes” e “no”) dependendo dos clientes apresentarem um empréstimo à habitação. Assim, verificamos que existe um maior número de indivíduos que possuem empréstimos a serem contactados. Outros fatores são o tipo de trabalho, estado civil, faixa etária, educação e empréstimos pessoais.

8.5.5 Gráfico 5

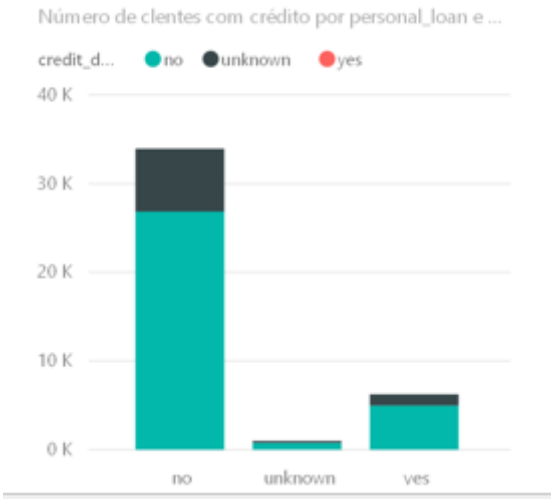

Figura 30 - Clientes com crédito em atraso por empréstimos pessoais

Neste gráfico apresentamos o número de pessoas com crédito em atraso por empréstimos pessoais. Verificamos que a maioria não tem nenhum empréstimo pessoal e que também não crédito em atraso. Outro fator são os empréstimos de habitação.

8.5.6 Gráfico 6

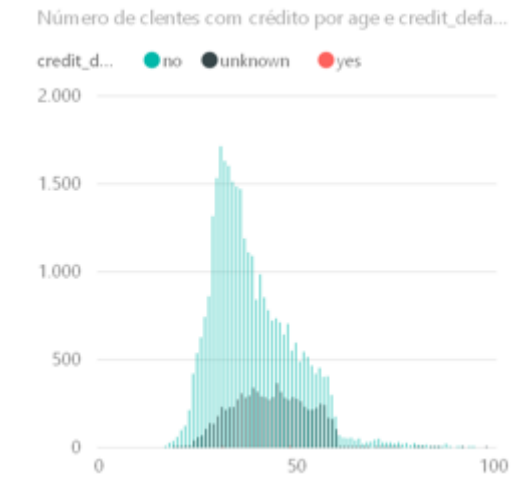

Figura 31 - Clientes com crédito em atraso por idade

Neste caso apresentamos o número de clientes com crédito em atraso por idade (não discretizada). Verificamos que existe um maior número de pessoas sem crédito em atraso com idades que rondam os 30 anos. Outros fatores são o estado civil, o tipo de trabalho e educação.

8.6. Report 6

Figura 32 - Report 6

8.6.1 Gráfico 1

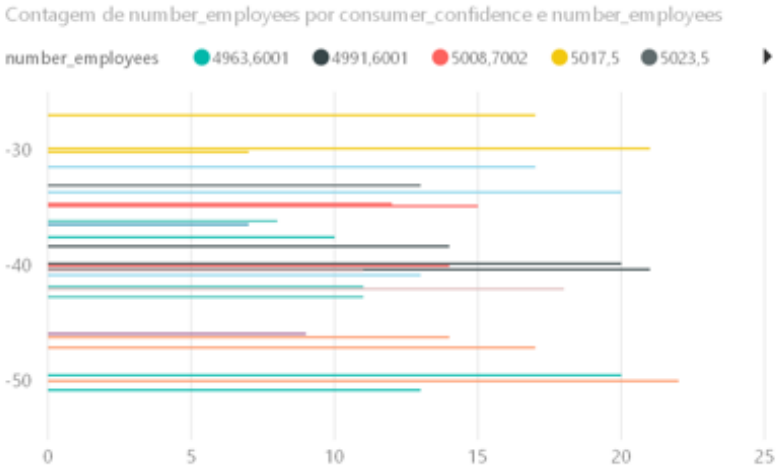

Figura 33 - Número de empregados por índice de confiança do consumidor

Neste gráfico apresentamos a média de empregados no banco por índice de confiança do consumidor. Registamos que, em geral, quando o índice apresenta valores baixos a medianos, o número de pessoas empregados no banco é maior.

8.6.2 Gráfico 2

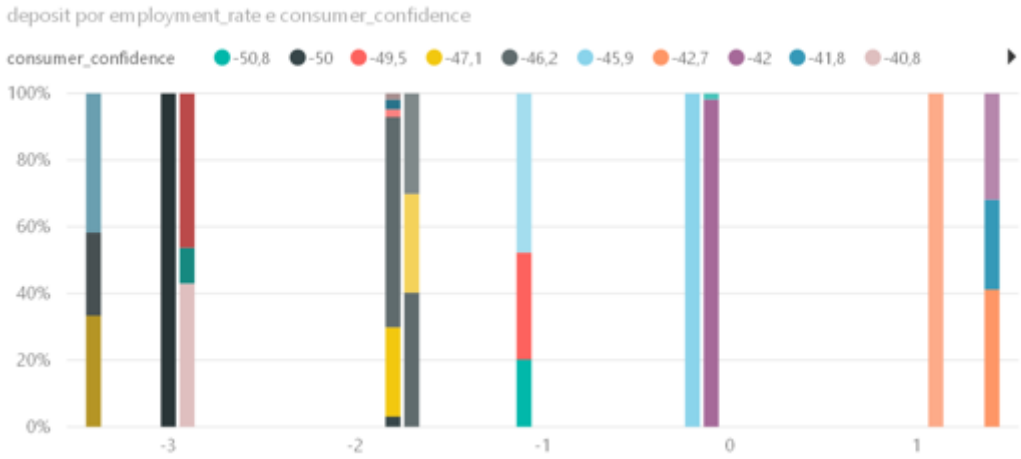

Figura 34 - Depósitos por taxa de emprego e índice de confiança do consumidor

Neste gráfico temos o número de contactos feitos a campanha do banco por taxa de emprego e índice de confiança do consumidor. Outros fatores são a Euribor, o número de empregados no banco, índice de preço de consumidor e taxa de emprego.

8.6.3 Gráfico 3

Figura 35 - Depósitos por número de empregados

Neste gráfico apresentamos os resultados dos depósitos por número de empregados a trabalhar no banco. Quando o número de empregados aumenta, também o número de clientes contactados aumenta, tendo que estes têm maior tendência a declinar o depósito.

8.6.4 Gráfico 4

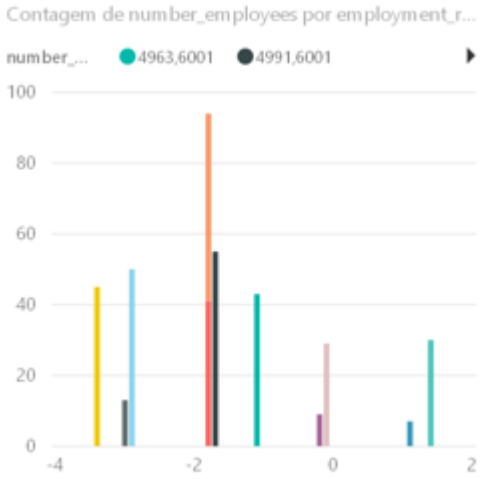

Figura 36 - Número de empregados por taxa de emprego

Neste gráfico explicitamos a relação existente entre o número de empregados no banco com a taxa de emprego do país. Notamos que quando a taxa de emprego tem um valor elevado, existe um maior número de pessoas a trabalhar no banco.

9. Dashboard do Microsoft Power BI

Uma dashboard do Microsoft Power BI funciona como uma página onde se irão armazenar os gráficos considerados mais relevantes dos reports. De seguida, apresenta-se uma dashboard do Power BI, constituído por tiles (gráficos, ou outra qualquer informação) que constituem os reports.

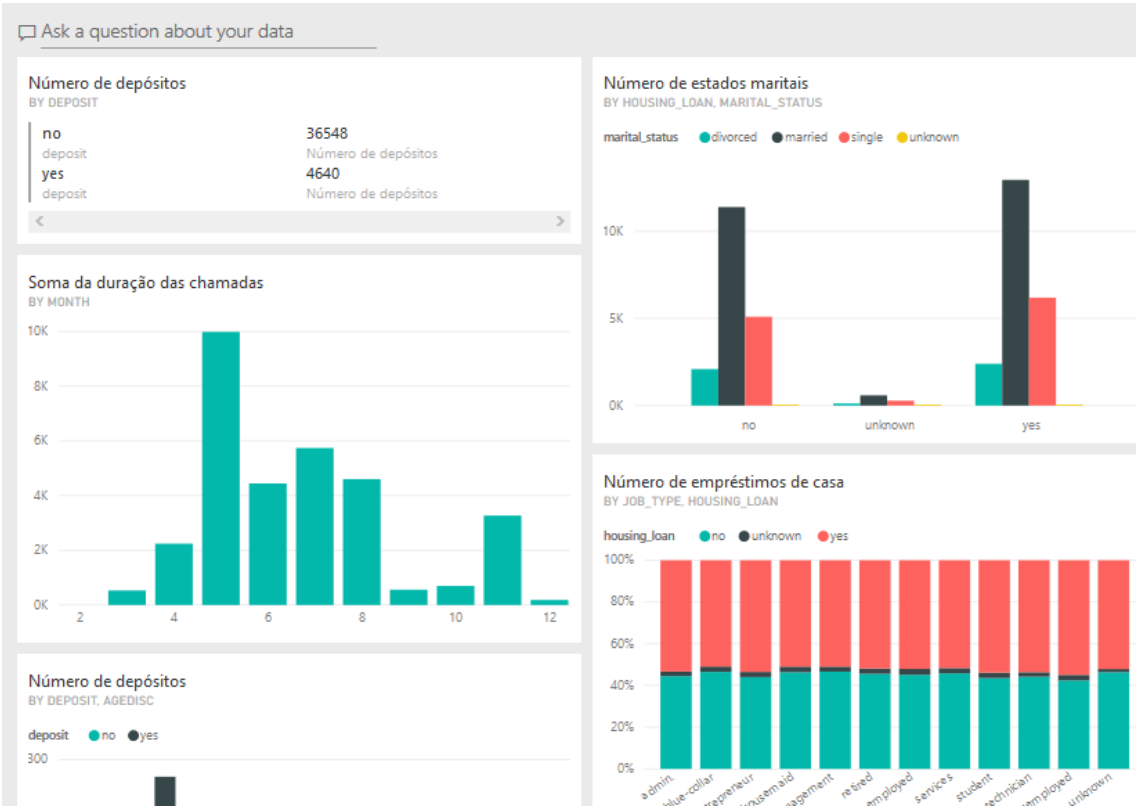

Figura 37-Dashboard do Power BI

Além da possível afixação das tiles dos reports, também é possível efetuar perguntas ao Microsoft Power BI, como por exemplo, qual a soma das durações para cada tipo de depósitos e estados civis.

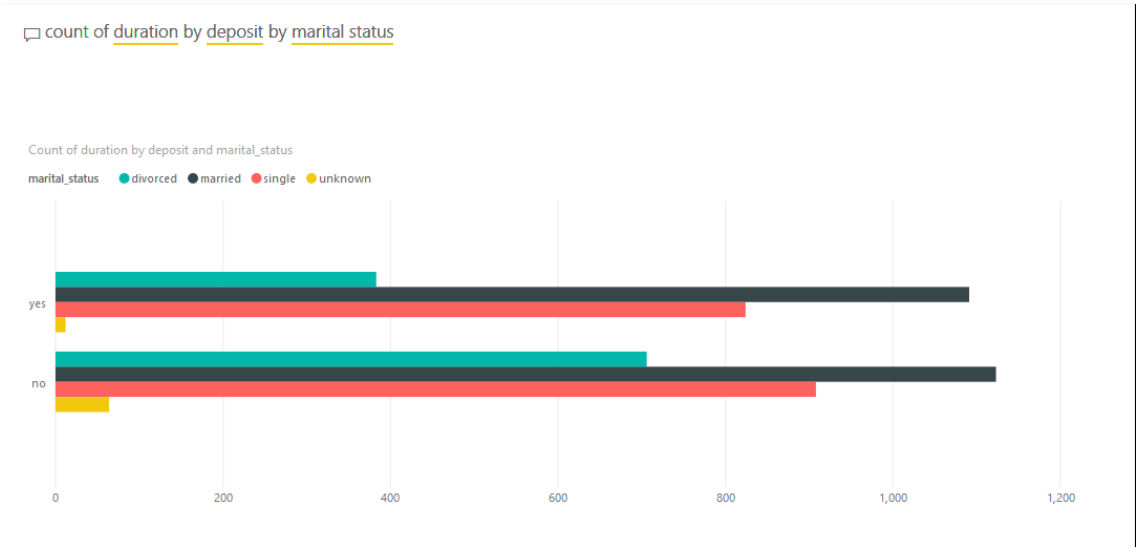

Figura 38-Exemplo de uma pergunta ao Power BI

9.1. Análise dos tiles da dashboard

9.1.1 Número total de depósitos de cada tipo

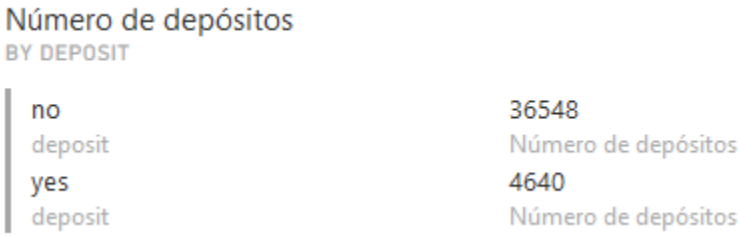

Figura 39-Número total de depósitos

Esta tile permite analisar as estatísticas dos depósitos de todas as campanhas. Analisando a figura, podemos concluir que existem 36548 depósitos rejeitados e 4640 depósitos que foram aceites pelo cliente.

9.1.2 Estado civil por crédito pessoal

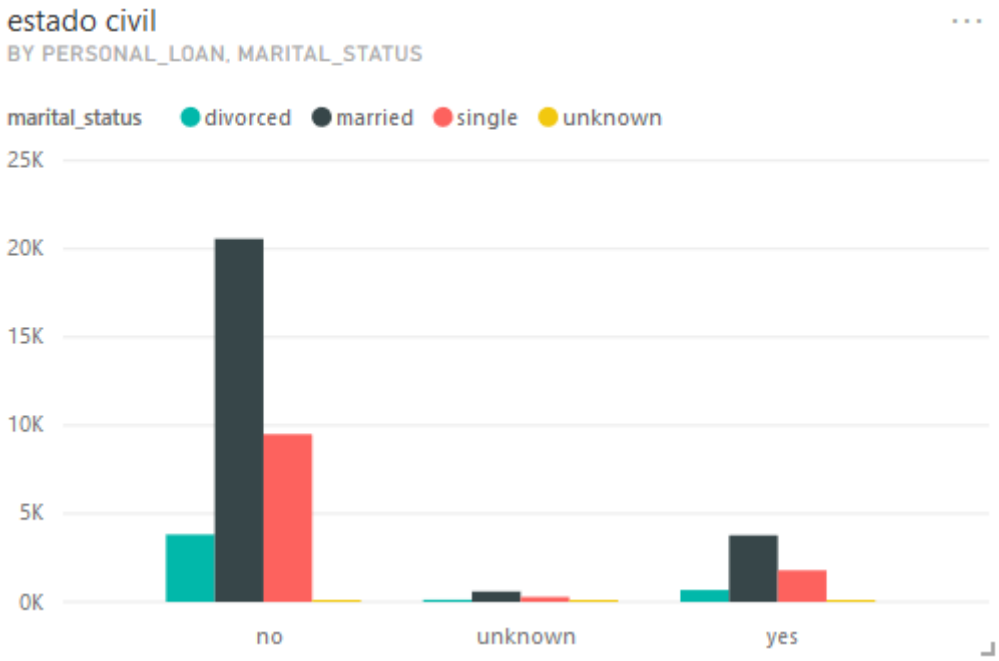

Figura 40-Estado civil por crédito pessoal

Esta tile permite analisar o número de clientes de cada estado civil (solteiro, casado, divorciado ou desconhecido) que possui, ou não, crédito de habitação. Analisando a figura, podemos concluir que é superior o número de pessoas com crédito de habitação do que os que não têm. Também se pode verificar que as pessoas casadas têm mais créditos de habitação do que qualquer outro grupo.

9.1.3 Soma da duração das chamadas por mês

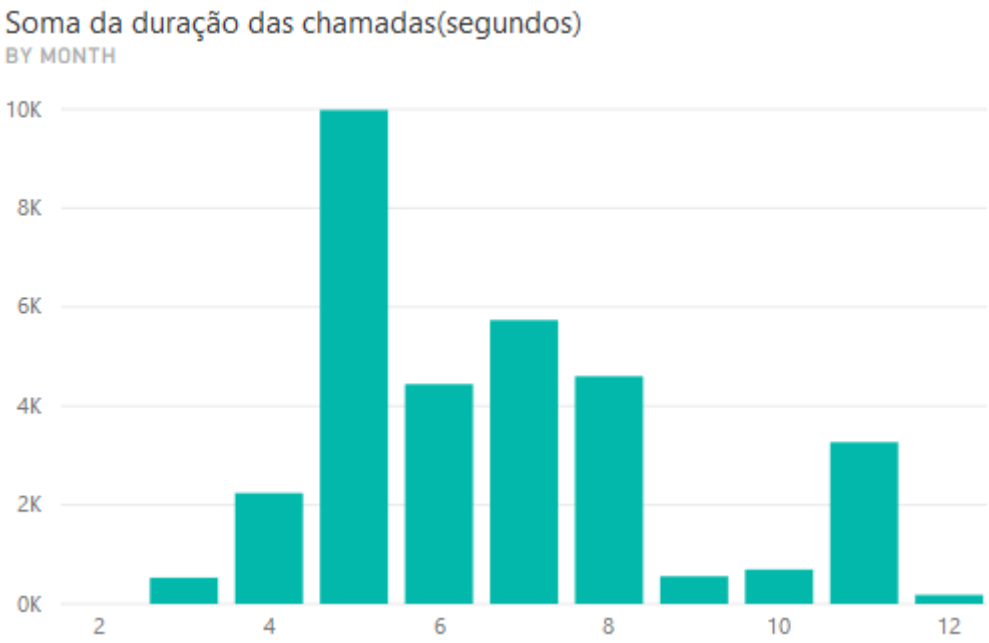

Figura 41-Soma da duração das chamadas por mês

Esta tile permite analisar a soma da duração (em segundos) das chamadas das campanhas relativamente aos meses do ano. Analisando o gráfico conseguimos verificar que o quinto mês (maio) foi o mês em que foi possível falar mais tempo com os clientes, enquanto que a soma da duração das chamadas em fevereiro foi praticamente nula.

9.1.4 Crédito de habitação por tipo de emprego

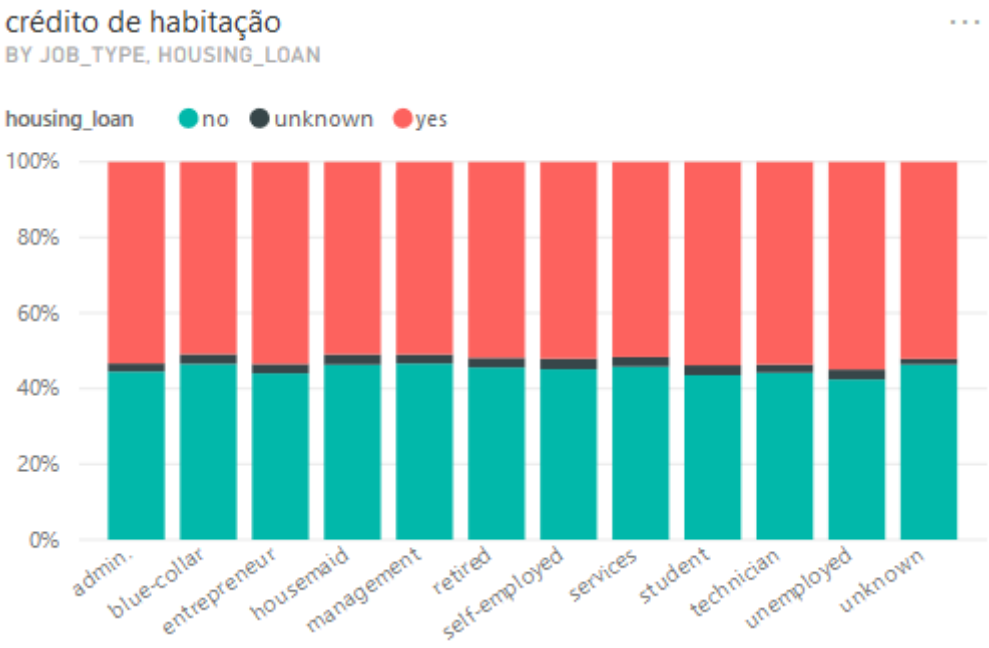

Figura 42-Crédito de habitação por tipo de emprego

Esta tile permite analisar a percentagem de existência de crédito de habitação para cada tipo de trabalho. Analisando o gráfico, podemos verificar que cerca de 40% dos estudantes não possuem crédito de habitação, enquanto que cerca de 60% dos estudantes possui crédito de habitação.

9.1.5 Número de depósitos por grau de educação

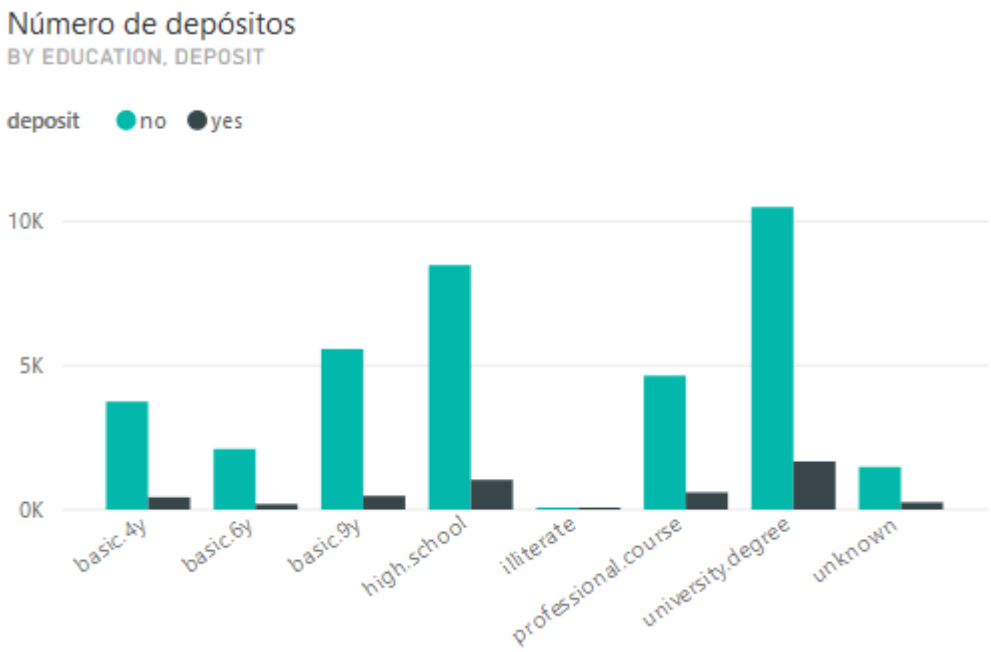

Figura 43-Número de depósitos por grau de educação

Esta tile permite analisar o número de depósitos ocorridos para diferentes níveis de grau de educação. Analisando o gráfico, podemos concluir que o maior número de chamadas ocorreu para os clientes que possuem um grau de educação universitária, provavelmente, porque este grupo apresenta rendimentos maiores e maior probabilidade de investir no crédito a depósito.

9.1.6 Número de depósitos por tipo de emprego

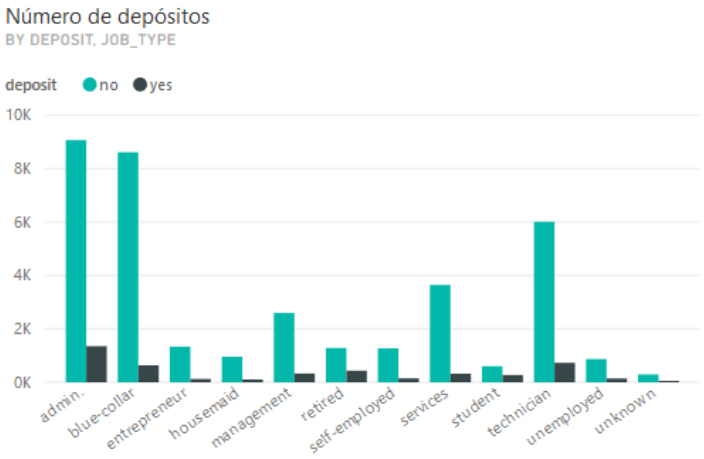

Figura 44-Número de depósitos por tipo de emprego

Esta tile permite analisar o número de depósitos ocorridos para diferentes tipos de emprego. Analisando o gráfico, podemos concluir que o maior número de chamadas ocorreu para os clientes que possuem cargos de administração, possivelmente pelos seus rendimentos serem mais elevados e ser mais fácil vender o crédito.

9.1.7 Número de depósitos por número de empregados

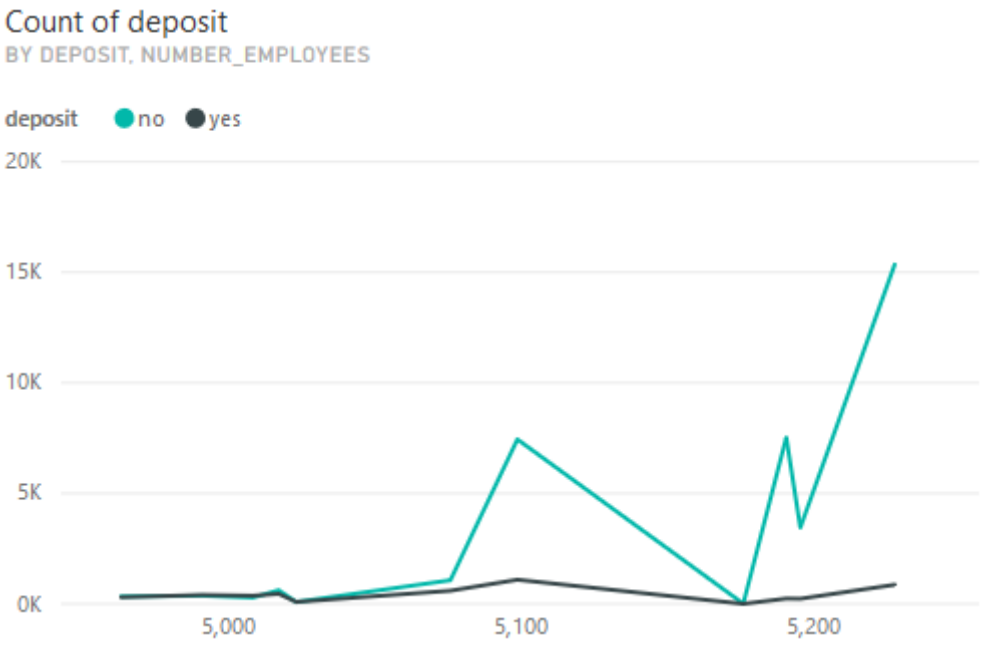

Figura 45-Número de depósitos por número de empregados

Esta tile permite analisar o número de depósitos ocorridos para diferentes valores do número de empregados. Analisando o gráfico, podemos concluir que a tendência é de haver mais depósitos com maior número de empregados, apesar de existir um caso perfeitamente visível no gráfico em que isso não se verifica. Também se consegue vislumbrar uma zona em que, contrariamente à tendência, ocorreram mais depósitos com valor positivo do que com valor negativo.

9.1.8 Número de depósitos por dia da semana

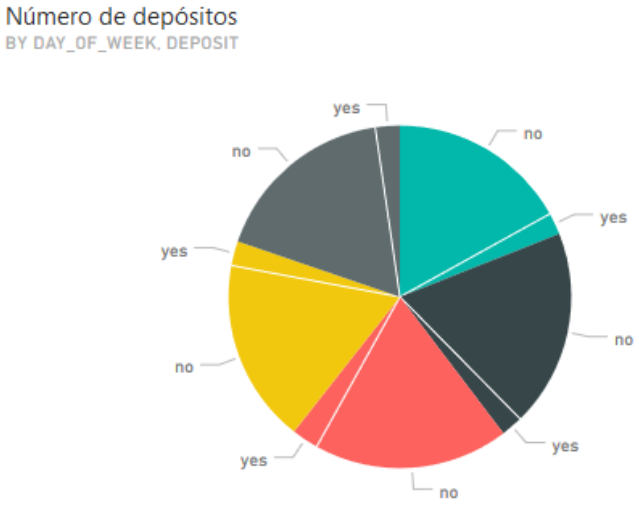

Figura 46-Número de depósitos por dia da semana

Esta tile permite analisar o número de depósitos ocorridos para diferentes dias da semana. Analisando o gráfico, podemos verificar que o dataset só continha valores para cinco dias da semana, não existindo qualquer tipo de chamadas no fim de semana. Conseguimos ainda visualizar como estão distribuídos os depósitos para cada dia da semana, concluindo que, independentemente do dia, houve mais clientes a dizerem não ao depósito do que sim.

Denote-se que é possível visualizar os valores referentes a cada dia da semana na *dashboard* do *Power BI* sobrepondo o cursor ao gráfico, como se pode visualizar na seguinte figura:

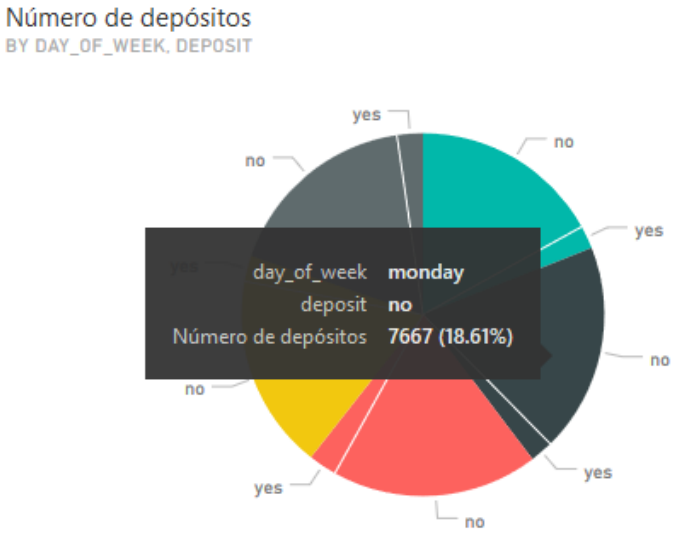

Figura 47-Gráfico com os detalhes dos depósitos e o dia da semana

10. Conclusões

O desenvolvimento deste projeto prático permitiu ao grupo analisar os dados referentes a uma campanha de *marketing* de um banco nas várias dimensões identificadas: informação do cliente, dados da campanha anterior e da atual, bem como dados socioeconómicos. Vários *reports* foram compostos com dados relacionados com estas dimensões e selecionaram-se os mais importantes para serem apresentados num *dashboard*.

Antes da criação do *data warehouse*, o grupo definiu algumas *queries* que auxiliaram o seu desenvolvimento, como, por exemplo, saber quantas respostas positivas à campanha indivíduos numa certa faixa etária deram. Estas perguntas apoiaram, pois, as decisões tomadas em relação à análise de dados feita.

Em suma, a incorporação de dados provenientes de diferentes fontes num único *data warehouse* permitiu uma análise dos dados, através das *queries* definidas previamente.