



Universidade de Lisboa
Faculdade de Ciências
Mestrado em Bioinformática e Biologia Computacional - MBBC
Mestrado em Informática - MI

Integração e Processamento Analítico de Informação - IPAI

PROJECTO:
Audiências Televisivas

Docente: António Ferreira

André Oliveira, 45648 – MI
Maria Móteiro, 43178 - MBBC
Tânia Maldonado, 44745 - MBBC

19 de março de 2017

Índice

1. Introdução	2
2. Análise das Fontes de Dados	3
2.1. Espetadores	3
2.1.1. Tratamento de Dados	3
2.2. Tipos de Programas Televisivos	7
2.3. Canais Vistos pelos Espetadores	9
2.3.1. Tratamento de Dados	9
2.4. Programação dos Canais Televisivos	13
2.4.1. Tratamento de Dados	14
2.5. Classes Sociais dos Espetadores	16
2.6. Fonte de Dados Adicionais	16
3. Relação entre Fontes de Dados	17
4. Modelação Dimensional	18
4.1. Processo de Negócio	18
4.2. Perguntas Analíticas	18
4.3. Definição do Grão	18
4.4. Dimensões do Negócio	19
4.5. Medidas Numéricas	20
4.6. Diagrama em Estrela do Data Warehouse	20
5. Conclusão	21
6. Bibliografia	21
7. Anexos	22

1. Introdução

Os dados das audiências são um importante elemento de gestão das estações televisivas, estando na base da definição do preço da publicidade a cobrar aos anunciantes, assim como da tomada de decisões sobre as grelhas de programação, entre outros.

O projeto desta unidade curricular envolve a modelação e construção de um *data warehouse* que incorpore dados de audiências televisivas no período de tempo que vai desde o dia 1 de janeiro de 1996, até 30 de junho de 1996. Este *data warehouse* será, idealmente, capaz de dar resposta a um leque de cenários de tomada de decisão, mas neste contexto será relativo a um único processo de negócio.

Propusemo-nos, no presente relatório, traçar algumas tendências e hábitos de consumo de televisão durante o primeiro semestre de 1996. Na primeira fase do projeto foi feita uma análise dos dados existentes e do negócio, com o objetivo de compreender as fontes de dados disponibilizadas assim como as interligações entre as diversas fontes de dados e, por fim, descrever um processo de negócio que possa utilizar estes dados. O processo de negócio escolhido pretende analisar os programas visto por um espetador durante um período de tempo.

Na segunda fase do trabalho iremos proceder à modelação dimensional do Data Warehouse, onde definimos o grão e o tipo de tabela de factos bem como as dimensões e medidas envolvidas.

Estas análises preliminares, bem como a deteção e correção de erros irão facilitar o trabalho aquando da construção do sistema ETL na etapa 3.

2. Análise das Fontes de Dados

Nesta secção apresentamos todos as fontes de dados (adicionais e disponibilizadas), bem como a identificação de erros executada e sua correção e posterior análise dos dados. As fontes de dados disponíveis são: espectadores, canais vistos pelos espectadores, programação dos canais, tipos de programas, classes sociais dos espetadores e feriados e datas festivas.

2.1. Espetadores

O ficheiro “espetadores.csv” contém dados de espetadores. Cada registo contém oito campos separados por vírgulas, como por exemplo:

6,3001,"Gr. Lisboa","Femin.,"DDC","+64","D",#1996-01-01#

O significado de cada um dos campos de um registo, segundo a ordem em que aparecem, é explicado na Tabela 1.

Tabela 1 - Significado de cada um dos campos de um registo de espetadores, segundo a ordem em que aparecem

#	Campo	Tipo de Dados	Descrição	Exemplo
1	ID	Número inteiro	Identificador único de registo	6
2	Código	Número inteiro	Identificador único de espetador	3001
3	Região	Texto	Região do país da residência do espetador	"Gr. Lisboa"
4	Sexo	Texto	Masculino ou feminino	"Femin."
5	DonaDeCasa	Texto	Se o espetador trabalha em casa ou não	"DDC"
6	EscalãoEtário	Texto	Escalão etário do espetador	" +64"
7	Classe	Texto	Classe social do espetador	"D"
8	Data	Data	Data de criação do registo	#1996-01-01#

2.1.1. Tratamento de Dados

Para analisar o ficheiro “espetadores.csv” foi utilizado o software RStudio, aliado a um script R fornecido inicialmente pelo docente, tendo sido por nós adaptado posteriormente.

Através da linha de código “print(NA_values <- sapply(espetadores, function(x) which(is.na(x))))” não foram encontrados “missing values”. Esta informação foi contestada pela função “summary”, que no campo “região” encontrou duas entradas de “Região – Z”. Estas linhas foram eliminadas manualmente.

Na Tabela 2, Gráfico 1,2,3 e 4, encontram-se os dados de análise estatística a parte das variáveis em estudo, já com os dados corrigidos.

Tabela 2 - Análise de estatística descritiva sobre os campos “id”, “codigo” e “data” da variável “espetadores” obtida através da função “summary”

	id	codigo	data
Min.:	1	240	1996-01-01

Mean:	152177	20892256	1996-03-31
Max.:	304353	34105603	1996-06-30

Tabela 3 - Análise de estatística descritiva sobre os campos "regiao", "sexo", "donadecasa", "escalaoetario" e "classe" da variável "espetadores" obtida através da função "summary"

regiao		sexo		dona de casa		escalao etario		classe	
Gr. Lisboa:	70597	Femin.:	111957	DDC:	145020	4-14:	29865	A/B:	67314
Gr. Porto:	50508	Masc.:	192394	nDDC:	159331	15-24:	50277	C1/C2:	88287
Interior:	39156					25-34:	38118	D:	148750
Lit. Norte:	42875					35-44:	46587		
Lit. Centro:	43808					45-54:	41351		
Sul:	57407					55-64:	47719		
						+64:	50404		

Para além dos erros referidos acima, também foram encontrados alguns **dados incoerentes**, nomeadamente no que toca a espetadores que mudam de atributos (como por exemplo de sexo). Estas linhas foram eliminadas.

Por exemplo: espetadores com o mesmo código apresenta valores discordantes para os atributos género e dona de casa.

153943	240	Lit. Centro	Masc.	DDC	35-44	C1/C2	#1996-04-01#
167460	240	Lit. Centro	Femin.	nDDC	35-44	C1/C2	#1996-04-09#

Imagem 1 - Exemplo de linhas com erros detetados

Os dados foram analisados visualmente com recurso à ferramenta Power BI, da Microsoft, na qual foram obtidos os gráficos 1, 2, 3 e 4. Para efeitos de “número de espetadores” foram tidos em consideração o número de registos individuais.

Pode verificar-se pelos gráficos que há uma maioria de registos do sexo masculino, de classe social D (trabalhadora) e da zona da Grande Lisboa. Há também uma ligeira predominância de pessoas da faixa etária “+64”.

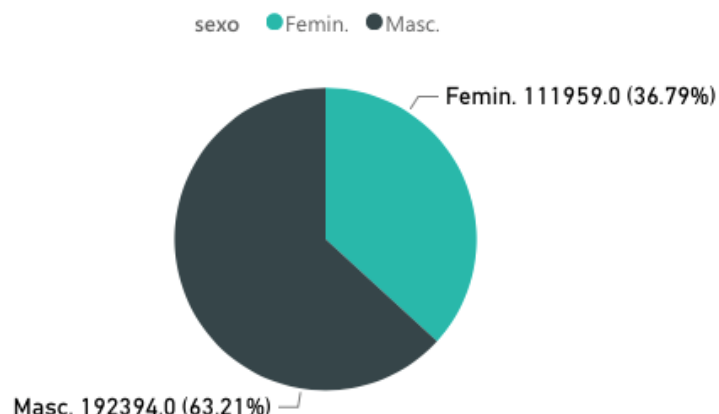
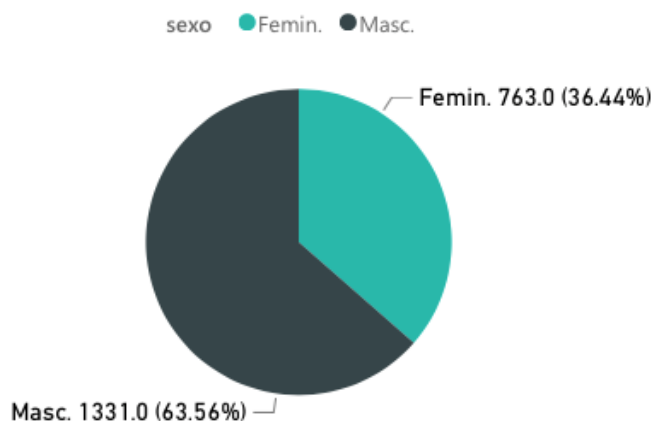


Gráfico 1 - Contagem e percentagem de códigos de espetadores e registos por sexo

O gráfico à esquerda diz respeito à totalidade de espetadores (2071), enquanto o segundo não tem em conta esta totalidade, mas sim a quantidade de registos, ou seja, o facto de um mesmo espetador com um certo código poder ter “id”s diferentes. Observa-se que a proporção é muito semelhante.

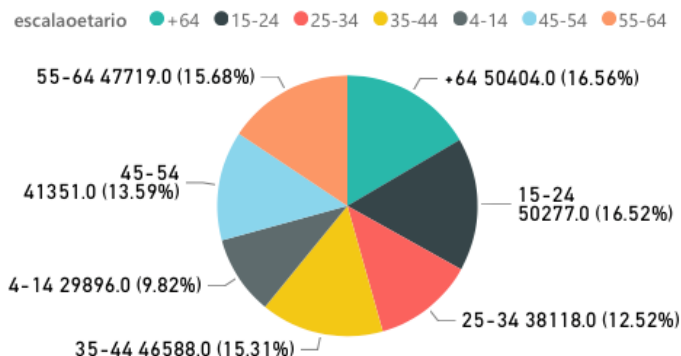
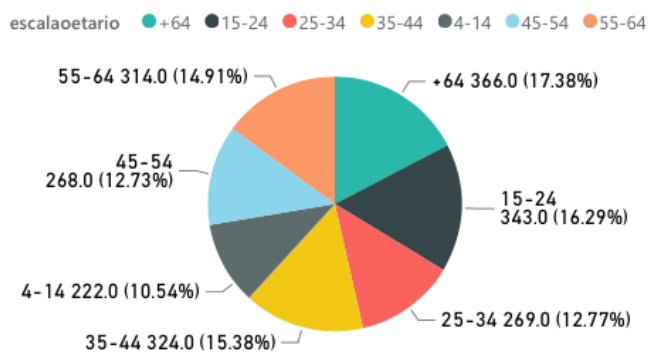


Gráfico 2 - Contagem e percentagem de códigos de espetadores e registos por escalão etário

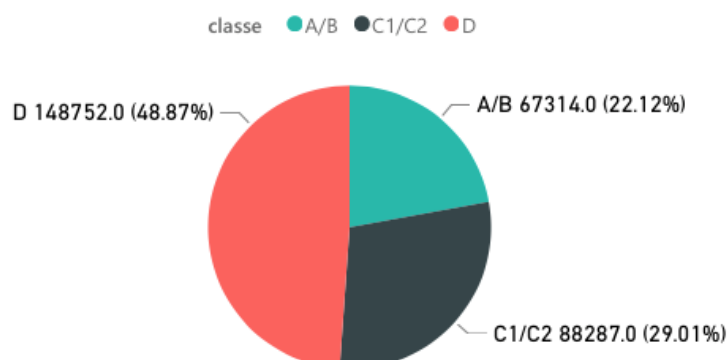
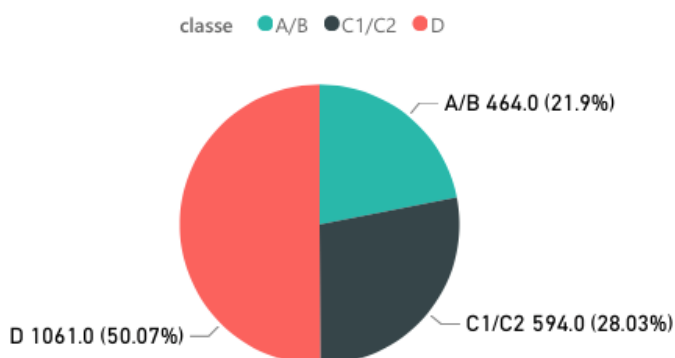


Gráfico 3 - Contagem e percentagem de códigos de espetadores e registos por classe social

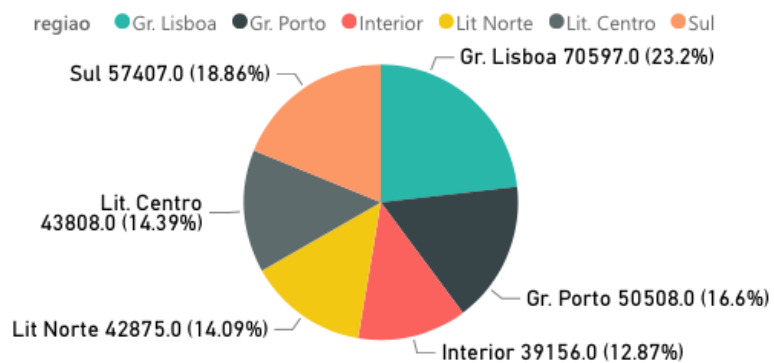
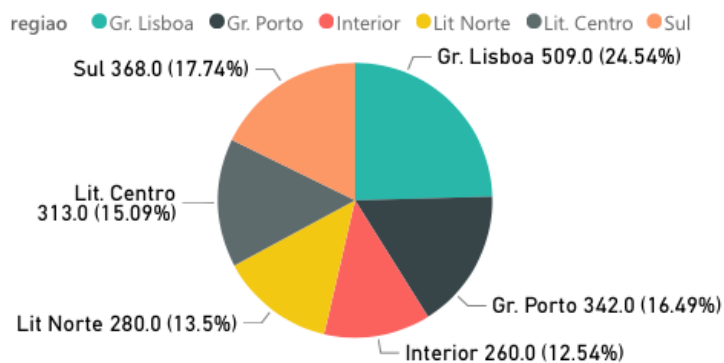


Gráfico 4 - Contagem e percentagem de códigos de espetadores e registos por região

Tal como no gráfico 1, os gráficos 2, 3 e 4 encontram-se separados: à esquerda é apresentado o gráfico que respeito à totalidade de espetadores (2071), enquanto o gráfico da direita se refere à quantidade de registos únicos. Novamente, a proporção entre os dois gráficos mantém-se semelhante.

2.2. Tipos de Programas Televisivos

O ficheiro “tipologia.tsv” guarda uma classificação dos vários tipos de programas televisivos.

A interpretação de cada um dos dois campos que formam um registo é esclarecida na Tabela 4, e uma listagem completa dos tipos de programa principais é apresentada na Tabela 5.

O tipo de programa é representado por uma sequência de até três letras (por exemplo abc), em que a primeira letra representa o tipo mais genérico do programa (ie, o tipo principal); a segunda letra indica o subtipo de programa; a terceira letra indica o último nível hierárquico do tipo de programa.

Tabela 4 - Significado de cada um dos campos de um registo de programas televisivos, segundo a ordem em que aparecem

#	Campo	Tipo de Dados	Descrição	Exemplo
1	Tipo	Texto	Identificador hierárquico do tipo de programa	abc
2	Designação	Texto	Designação do nível hierárquico do tipo de programa	Desenho Animado

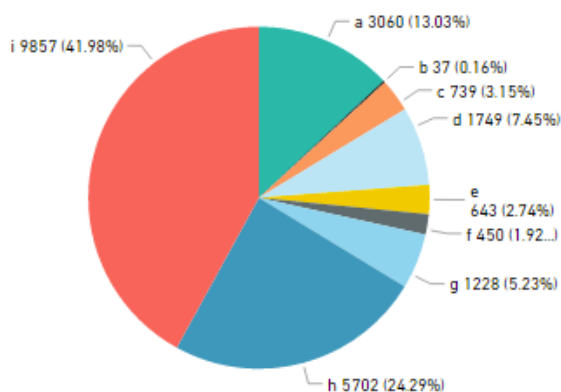
Tabela 5 - Designação dos Tipos Principais de Programas

Identificador	Tipo Principal
a	Ficção
b	Eruditos
c	Variedades/divertim.
d	Informação
e	Cultura/conhecimento
f	Desporto
g	Juventude
h	Publicidade
i	Diversos
z	Outros

Para uma melhor compreensão dos dados, os mesmos foram analisados visualmente tendo sido obtido o conjunto de gráficos 5. A contagem total dos tipos de programa foi obtida através da leitura exclusiva da primeira letra do campo “tipo” dos dados referentes à programação.

Contagem dos Tipos Principais de Programas

Tipos de Programas a b c d e f g h i z



Tipos de Programa Principal por Canal

Tipos de Programas a b c d e f g h i z

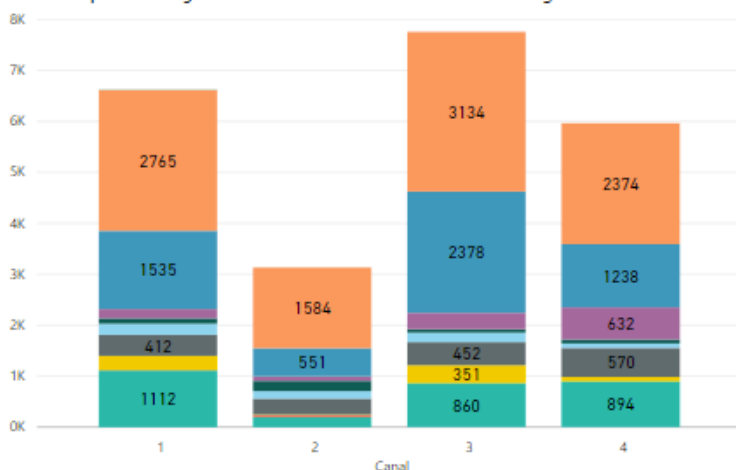


Gráfico 5 - Contagem dos Tipos Principais de Programas e Tipos de Programas por Canal

Através destes gráficos pode verificar-se que a maioria de visualizações é referente a programas do tipo “diversos”, sendo as duas categorias seguintes mais vistas as referentes a “publicidade” e “ficção”.

Pode ainda observar-se que as proporções dos tipos de programas se mantêm semelhantes nos vários canais.

2.3. Canais Vistos pelos Espectadores

O ficheiro “audiencias.csv” contém registos dos tempos de visualização de canais por cada espetador, estando os vários campos separados por vírgulas, como no exemplo seguinte:

57,#1996-01-01#,1,6,#1996-01-01 14:47:00#,1996-01-01 14:53:00#

O significado de cada um dos seis campos que formam um registo, pela ordem em que surgem, é apresentado na Tabela 5.

Tabela 6 - Significado de cada um dos campos de um registo dos canais observados, segundo a ordem em que aparecem

#	Campo	Tipo de Dados	Descrição	Exemplo
1	ID	Número inteiro	Identificador de registo do espetador	57
2	Data	Data	Data de criação do registo	#1996-01-01#
3	Canal	Número inteiro	Número do canal visto pelo espetador	1
4	Duração	Número inteiro	Tempo de visualização do canal, em minutos	6
5	HoraInício	Data	Hora de início da visualização do canal	#1996-01-01 14:47:00#
6	HoraFim	Data	Hora de fim da visualização do canal	#1996-01-01 14:53:00#

De notar que os valores dos identificadores de registos de espetadores (primeiro campo, ID) estão relacionados com o campo ID do ficheiro “espetadores.csv”, descrito no capítulo 2.1. Assim, o valor “57” mencionado nesta secção é o mesmo que consta na secção sobre os espetadores de televisão.

2.3.1. Tratamento de Dados

Através da linha de código “print(NA_values <- sapply(audiencias, function(x) which(is.na(x))))” obtivemos todos os índices correspondentes a variáveis com um **output “NA”**. Estes valores não definidos referem-se à não presença de horas no valor da data, tal como foi confirmado no Excel.

A informação obtida pela linha suprarreferida é corroborada pela linha de código “summary(audiencias);”, que devolve as informações presentes na Tabela 6.

Tabela 7 - Análise de estatística descritiva sobre os campos da variável “audiencias”

	id	data	duracao	horainicio	horafim		Canal
Min.:	57	1996-01-01	0.00	1996-01-01 02:00:00	1996-01-01 02:01:00	“1”:	695774
Mean:	149897	1996-03-29	24.93	1996-03-30 14:39:29	1996-03-30 15:04:59	“2”:	268942
Max.:	304353	1996-06-30	1055	1996-07-01 01:54:00	1996-07-01 01:55:00	“3”:	772158

Através do software Excel foi feita uma ordenação das colunas "horainicio" e "horafim" (à vez) de acordo com o número de caracteres para isolar as entradas que não se encontravam no formato "AAAA-MM-DD HH:MM:SS".

Verificou-se que as entradas neste formato correspondiam à hora "00:00:00", pelo que 3221 entradas da coluna "horainicio" e 3858 entradas da coluna "horafim" foram corrigidas para conter a hora "00:00:00".

Tal como para os dados do capítulo 2.1, também aqui foram elaborados gráficos para uma melhor interpretação dos dados.

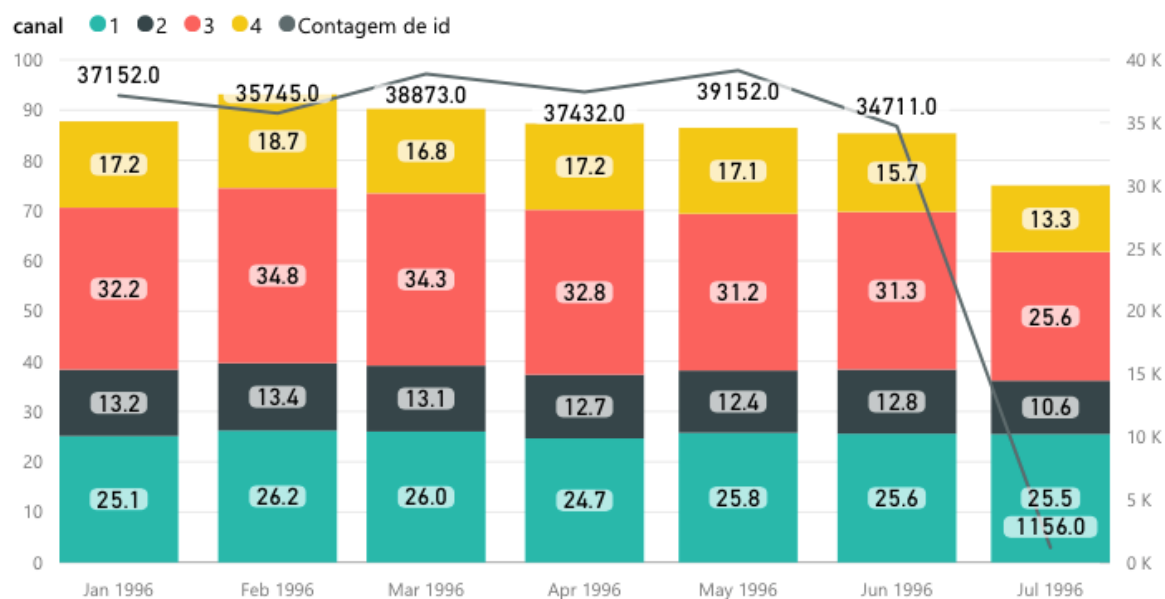


Gráfico 6 - Evolução da duração média, em minutos, por canal e mês

No gráfico 4 pode observar-se a evolução da duração média de visualização de cada canal, ao longo dos vários meses do semestre em análise. Em todos os meses, a maioria do *share*¹ vai para o canal 3.

De notar uma certa homogeneidade que apenas é quebrada em julho, devida ao facto de que neste mês apenas registadas 1156 visualizações pertencentes às primeiras horas da madrugada do dia 1 de julho.

¹ Percentagem de audiência de um canal/programa relativamente à audiência do total de televisão, para o mesmo período.

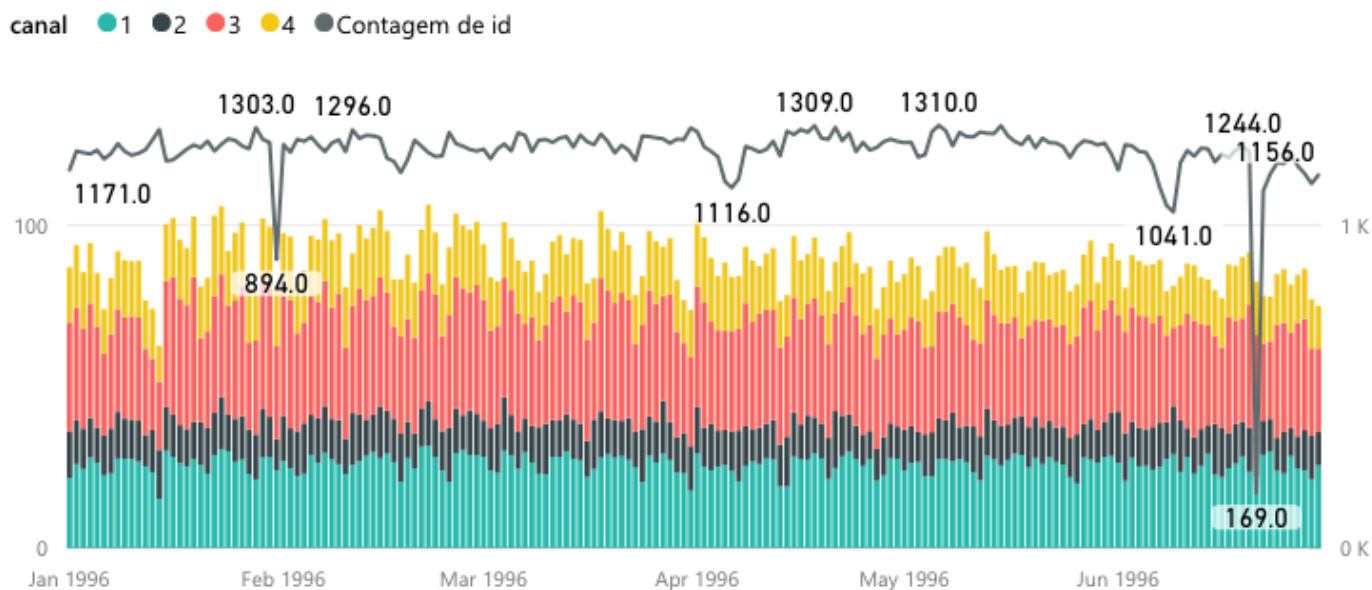


Gráfico 7 - Evolução da duração média, em minutos, por canal e dia (1 jan. a 30 junho)

No gráfico 6 apresenta-se a evolução diária dos registos de visualizações, assim como a evolução das durações por canal. Continua a verificar-se, assim como no gráfico 4, que a maioria do share vai para o canal 3.

Observam-se 2 reduções significativas no número de contagens de espetadores: a primeira no dia 31 de janeiro de 1996 (onde se verificaram apenas 894 registos), e a segunda no dia 21 de junho (onde se verificaram apenas 169 registos).

No gráfico 7 é visível a duração média (em minutos) segundo os dias da semana a que os registos se referem, e as faixas etárias dos respetivos espetadores.

De registar que os dias da semana com maior número de espetadores (em média) são terça e quarta-feira, sendo que os dias onde seria de esperar um maior número de espetadores (sábado e domingo) se encontram a meio e no fim da lista, respetivamente.

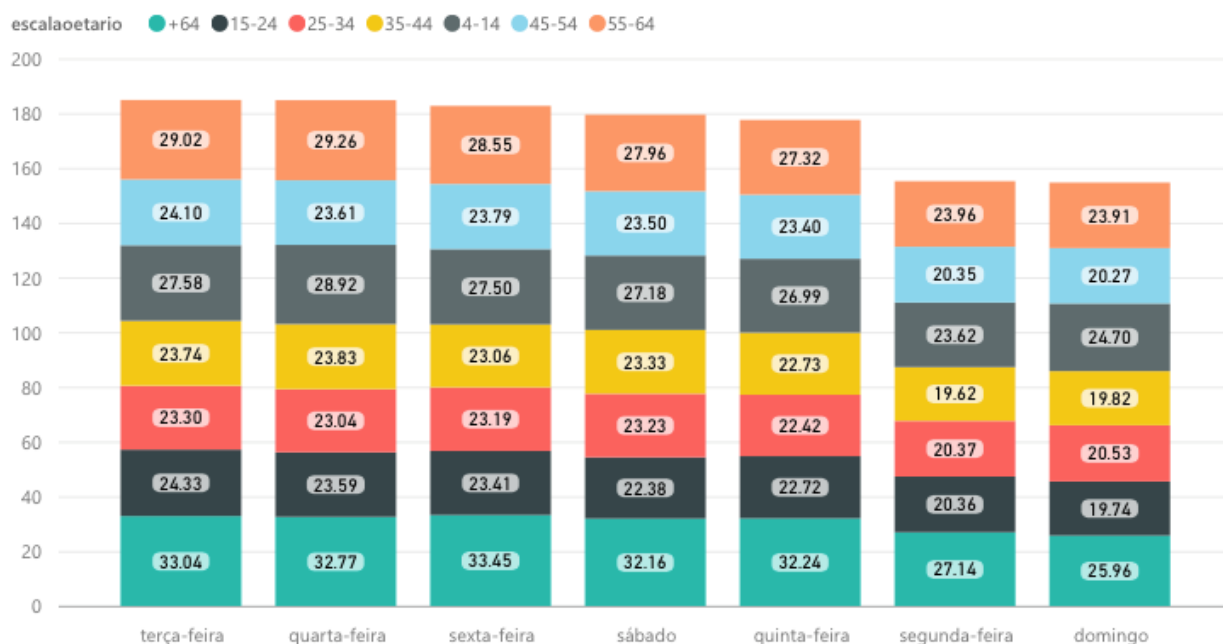


Gráfico 7 - Duração média, em minutos, por dia da semana e escala etária, ordenada segundo a grandeza total

Por fim, no gráfico 8 é visível a duração média de visualização ao longo dos dias da semana, por cada classe social. Tal como no gráfico 3, verifica-se uma maioria da classe D (trabalhadora), seguida da classe C1/C2 (média/média baixa).

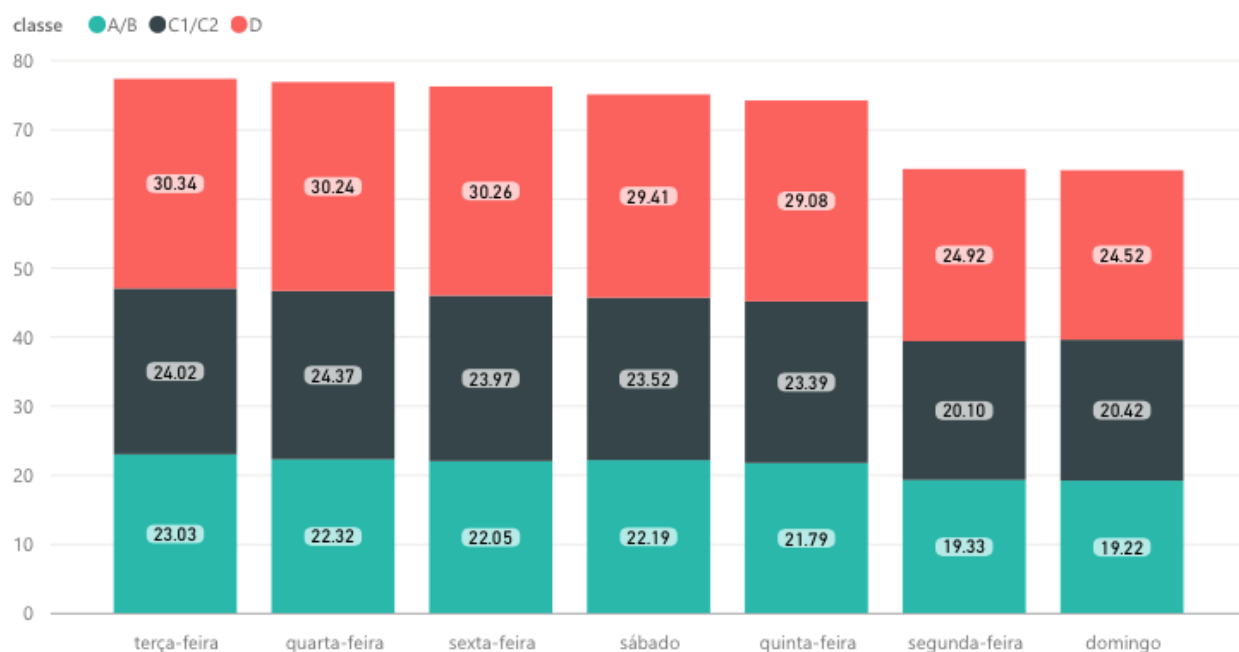


Gráfico 8 - Duração média, em minutos, por dia da semana e classe social, ordenada segundo a grandeza total

2.4. Programação dos Canais Televisivos

Esta fonte de dados é constituída por vários ficheiros, cada um guardando a programação televisiva tal qual foi emitida num determinado dia. Estão disponíveis tantos ficheiros quantos os dias do primeiro semestre de 1996, com nomes que identificam univocamente o dia a que respeitam.

A extensão PET vem de origem, mas o conteúdo é semelhante ao de um ficheiro com campos separados por vírgulas (.csv), com a diferença de cada linha terminar com o símbolo “;”.

1, 20000, 2775, 0, "SESSAO DUPLA I", "CLASSE", "P", "aae", 0;

O significado de cada um dos nove campos de um registo é ilustrado na Tabela 8.

Tabela 8 - Significado de cada um dos campos de um registo de programação, segundo a ordem em que aparecem

#	Campo	Tipo de Dados	Descrição	Exemplo
1	Canal	Número inteiro	Número do canal no ar	1
2	HoraInício	Número inteiro	Hora inicial do programa, no formato hhmmss	20000
3	Duração	Número inteiro	Duração do conteúdo televisivo, em segundos	2775
4	Zero	Número inteiro	Sem significado	0
5	Nome1	Texto	Nome do conteúdo televisivo	"SESSAO DUPLA I"
6	Nome2	Texto	Um segundo nome do conteúdo televisivo	"CLASSE"
7	Classificação	Texto	Classificação do conteúdo, detalhada a seguir	"P"
8	Tipo	Texto	Tipo do conteúdo, de acordo com a tipologia em cima	"aae"
9	ParteTodo	Número inteiro	Se representa o conteúdo todo ou uma das suas partes	0

Todos os ficheiros PET guardam registos com hora de início às 2h00 da noite (valor 20000 em HoraInício) e registam a sequência de conteúdos televisivos emitida num período de 24 horas; por exemplo, o valor 253015 representa "25" horas, 30 minutos, e 15 segundos, ou seja, cerca da uma e meia da noite do dia seguinte.

Relativamente ao campo “Classificação”, este pode tomar três valores distintos: "P" para programa, "B" para intervalo comercial, e "I" para publicidade ao próprio canal.

Por fim, o campo “ParteTodo” indica se o registo diz respeito a um programa como um todo (valor 0) ou a uma das partes (valor 1), sendo que um valor 0 pode incluir intervalos.

Em alguns ficheiros PET existem registos cujo valor da variável “ParteTodo” é 2, o que significa que se trata de uma parte de um todo com código 1, em vez de 0, isto é uma parte integrada dentro de outra parte. Nestas circunstâncias, a duração total do programa (com ParteTodo = 0) corresponde à soma das durações dos subprogramas

(com ParteTodo = 1), e a duração destes será a soma das dos sub-subprogramas (com ParteTodo = 2).

2.4.1. Tratamento de Dados

Tal como para as fontes de dados referentes aos espetadores e canais vistos por estes, também aqui foi utilizado um script R para auxílio à análise de estatística descritiva dos dados em estudo.

Uma vez que algumas das variáveis da presente fonte de dados possuem valores descritivos em grande quantidade, apenas algumas variáveis são apresentadas na Tabela 9, de acordo com a função "summary(todos.dados.pet);".

Tendo em conta que a coluna "Zero" apresentada nos dados originais não tinha significado, e para uma melhor visualização e entendimento dos dados, esta foi eliminada.

Tabela 9 - Análise de estatística descritiva sobre os campos "Horalnicio", "Duracao", "Canal", "Classificacao" e "ParteTodo" da variável "todos.dados.pet" obtida através da função "summary"

	Horalnicio	Duracao		Canal		Classificacao		ParteTodo
Min.:	20000	2.0	1	6625	B	6221	0	9543
Mean:	168543	573.8	2	3133	I	9121	1	12895
Max.:	255957	22478.0	3	7757	P	8137	2	1041
			4	5964				

Usando um script R para ordenar os dados por valores e procura por valores NA, foram encontrados os seguintes erros:

- **Vírgulas em falta**

Exemplo: Ficheiro "19960607.pet", linha 399:

```
3, 22934, 5, 0, "PATROCINIO", "1""B", "hc", 1;
```

Correção: vírgula adicionada manualmente.

- **Aspas em falta**

Exemplo: Ficheiro "19960217.pet", linha 384:

```
3, 90044, 97, 0, "INT.APRES.PROGRAMAS",", "I", "ib", 0;
```

Correção: aspa adicionada manualmente, mas quando o atributo não tem valor (como no exemplo acima) os atributos foram ignorados.

- **Valores de atributos em falta**

Exemplo 1: Ficheiro "19960605.pet", linha 404:

```
, 43030, 2576, 0, "OUTRO CINEMA", "KRUSH GROOVE", "P", "aak", 1;
```

Correção: falta o número do canal, portanto as linhas com este tipo de erro foram eliminadas uma vez que o canal não pode ser inferido por outros dados.

Exemplo 2: Ficheiro “19960508.pet”, linha 396:

```
3, 24202, 43, 0, "PUBLICIDADE ASSOCIADA", "1", "B", "ha",
```

Correção: falta o atributo “ParteTodo”, pelo que as linhas detetadas com este tipo de erro foram eliminadas.

- **Caracteres inválidos**

Exemplo: Ficheiro “19960610.pet”, linha 392:

```
3, 25449, 6, 0, "PATROCINIO", "1", "B", "hc", 1;
```

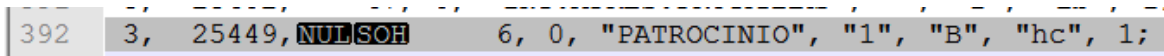
A screenshot of a text editor showing line 392. The line contains the text: 3, 25449, NULSOH 6, 0, "PATROCINIO", "1", "B", "hc", 1;. The characters 'NUL' and 'SOH' are highlighted in red, indicating they are invalid characters.

Imagem 2 - Representação gráfica do exemplo supracitado para exemplificação dos caracteres inválidos

Correção: os caracteres inválidos foram removidos, tendo os restantes valores sido preservados.

2.5. Classes Sociais dos Espectadores

O ficheiro “classes.tsv” descreve o significado das letras A, B, ..., que identificam classes sociais.

O significado de cada um dos campos de um registo é explicado na Tabela 10, e a lista completa das diferentes classes sociais e sua descrição é apresentada na Tabela 11.

Tabela 10 - Significado de cada um dos campos de um registo das classes sociais dos espetadores, segundo a ordem em que aparecem

#	Campo	Tipo de Dados	Descrição	Exemplo
1	Classe	Texto	Classe social	A
2	Estatuto	Texto	Estatuto social	Classe média/alta
3	Ocupação	Texto	Ocupações representativas	Gestor, administrador, ou profissional de topo

Tabela 11 - Listagem das classes sociais e respetivos estatutos e ocupações

Classe	Estatuto	Ocupação
A	Classe média/alta	Gestor, administrador, ou profissional de topo
B	Classe média	Gestor, administrador, ou profissional intermédio
C1	Classe média/baixa	Supervisor ou empregado de escritório, gestor, administrador, ou profissional júnior
C2	Classe trabalhadora qualificada	Trabalhador manual qualificado
D	Classe trabalhadora	Trabalhador manual pouco ou não qualificado
E	Aqueles com menor nível de subsistência	Pensionistas sem outros rendimentos, trabalhadores temporários

De notar que o campo “Classe” é também usado no ficheiro “espetadores.csv” (ver capítulo 2.1), embora neste caso existam várias ocorrências em que num mesmo valor são concatenados dois identificadores de classe (por exemplo, A/B).

2.6. Fonte de Dados Adicionais

Foi criado um ficheiro de extensão “tsv” (separado por tabulações), que contém feriados e datas festivas de 1996, com base na informação disponível na Internet².

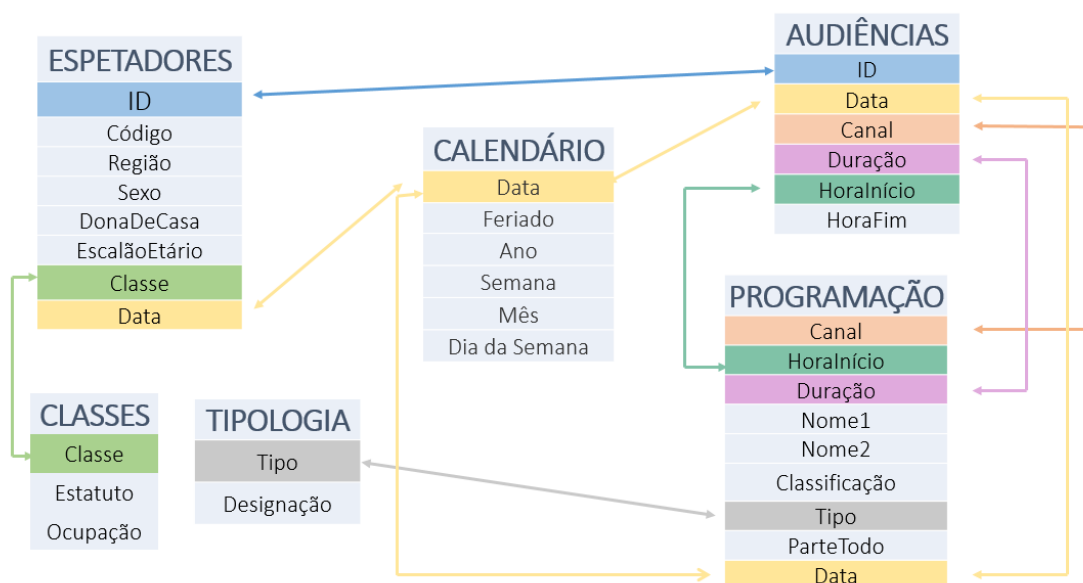
#	Campo	Tipo de Dados	Descrição	Exemplo
1	Data	Data	Dia do feriado	01/05/1996
2	Feriado	Texto	Indica se é feriado ou não	Feriado
3	Descrição	Texto	Nome do feriado e nomes alternativos	Dia do Trabalhador

No diagrama de dados existe uma fonte de dados “Calendário” que apresenta outros atributos como ano, semana, mês e dia da semana, cuja representação neste ficheiro não é necessária, visto que os seus respetivos valores podem ser obtidos recorrendo a funções em SQL ou Excel.

² <https://www.calendarr.com/portugal/calendario-1996/>

3. Relação entre Fontes de Dados

Para uma melhor compreensão dos dados disponibilizados, foi realizado um diagrama que relaciona as diferentes fontes de dados, direta ou indiretamente.



Esquema 1 - Diagrama das Fontes de Dados e suas relações.

Analisando o resultado final podemos retirar algumas informações:

- O ficheiro referente às audiências contém os dados de cada registo feito pelo espetador, que é o mesmo registado no ficheiro espetadores. Para além disto, contém a data desse registo que também pode se obter no ficheiro relativo aos espetadores;
- No ficheiro alusivo audiências, um dos parâmetros é o canal, ou seja, podemos ter acesso ao que um espetador viu cruzando os dados com o ficheiro referente à programação e posteriormente saber o tipo de programa através do ficheiro que se refere às diferentes tipologias;
- Finalmente, para saber a classe social de cada espetador pode-se fazê-lo cruzando o ficheiro alusivo aos espetadores com o ficheiro referente às classes.

Uma fonte de dados extra foi acrescentada, “Calendário”, que nos permite aceder aos feriados e datas comemorativas do primeiro semestre de 1996. Estes dados são cruzados com o ficheiro das audiências e da programação.

É possível observar um exemplo destas relações no esquema 3 (na seção Anexos), onde carregando na seção “Femin.” do gráfico circular, no canto superior direito, se verifica a resposta das variáveis com as quais interage nos restantes gráficos.

A resposta de maior destaque, neste caso, será uma clara superioridade da duração média de visualização por raparigas da faixa etária dos 4 aos 14 anos, à sexta feira, com uma média de duração de 94 minutos, conforme se pode verificar no gráfico do canto superior esquerdo, a cinzento.

4. Modelação Dimensional

4.1. Processo de Negócio

O consumo de televisão varia ao longo do ano. Um dos fatores diretamente relacionado com essa variação do consumo é a sazonalidade (ao longo do ano). Entender os espetadores que contactam com este meio em diferentes períodos do ano e também do dia são avaliações importantes para a compreensão da evolução do consumo televisivo.

Nas estações televisivas é necessário analisar os hábitos dos telespetadores para adaptar os conteúdos transmitidos. Assim, baseando-nos na amostra disponibilizada, o processo de negócio em foco será as **tendências das audiências televisivas tendo em conta um determinado período de tempo**, nomeadamente durante o primeiro semestre de 1996.

A análise deste processo de negócio consistirá nomeadamente em estudar o que acontece aos diferentes parâmetros, por exemplo faixa e classe etária, região, duração da visualização.

4.2. Perguntas Analíticas

Para facilitar este estudo, elaborámos algumas questões específicas para este processo:

- Quais são os programas televisivos mais vistos por cada faixa etária e classe social ao longo da semana e durante o fim-de-semana?
- Qual a média de espetadores por canal em cada dia durante o primeiro semestre de 1996?
- Quanto tempo cada espetador vê televisão durante a semana e ao fim-de-semana?
- Por cada região do país, quem (grupo social e se trabalha ou não em casa) passa mais e menos tempo a ver televisão tendo em conta a ocupação?
- Tendo em conta os feriados nacionais e datas comemorativas no primeiro semestre de 1996, qual a média de espetadores por faixa etária e classe social?

4.3. Definição do Grão

Tendo em conta o processo de negócio em estudo, referente às tendências das audiências televisivas para um determinado período de tempo, é necessário registar eventos detalhado, tendo sido definido o seguinte grão para a tabela de factos:

“Um espetador vê um programa, numa data, durante um período de tempo.”

Factos com este grão podem ser utilizados para responder a questões como a seguinte:

“Quais os programas mais vistos entre as 15h e 16h?”

Por sua vez, uma resposta hipotética poderia ser:

“O programa do tipo “c” (“variedades/diver.”) do canal 4, visto por x espetadores, com uma duração média de y minutos.”

Neste caso, o **tipo de tabela de factos** utilizado é do tipo **transaccional**, ou seja, a tabela que registam eventos que ocorrem em determinados momentos – cada facto aconteceu num ponto no tempo. Em que cada linha corresponde ao registo de um novo evento conforme o grão definido.

Recorrendo ao nosso grão, cada linha da tabela de factos irá registar o que um espetador visualizou num dado momento ou período de tempo.

4.4. Dimensões do Negócio

De acordo com os atributos definidos para a tabela de factos, foram identificadas quatro dimensões: Espetador, Programa, Data e Horário.

A **dimensão Espetador** possui como atributos a chave substituta, a chave natural referente a um registo único de cada espetador, assim como o seu género, escalão etário, região, classe social e se trabalha em casa ou não.

Relativamente à **dimensão Programa**, os atributos definidos foram a chave substituta, o canal ao qual se refere o registo, assim como o tipo de programa, e os nomes do programa (nome geral e nome específico).

Nesta dimensão identificámos 2 hierarquias distintas, uma para os nomes dos programas e outra para o tipo de programa:

Nome Programa

- Nome geral
 - Nome específico

Tipo Programa

- Tipo Nível 1
 - Tipo Nível 2
 - Tipo Nível 3

Para a **dimensão Data** foram definidos os atributos: chave substituta, o valor da data (chave natural), dia do mês, nome do mês, ano, dia da semana, semana do ano, se é fim-de-semana ou não, e se é feriado ou não.

Na dimensão Data identificou-se 2 hierarquias possíveis a partir do ano:

→ Ano

- Semana do ano
 - É Fim-de-Semana?
 - Dia da semana
- Mês
 - Dia do mês

Por ultimo, foi definida a **dimensão Horário** que contém informações sobre as horas de um dia nomeadamente: chave substituta, hora, minutos, segundos e período de dia que admite como valores: manhã (7h-12h), tarde(12h-18h), noite(18h-00) ou madrugada (00h-7h). Para este caso, identificámos apenas uma hierarquia:

- Período do dia
 - Hora
 - Minutos
 - Segundos

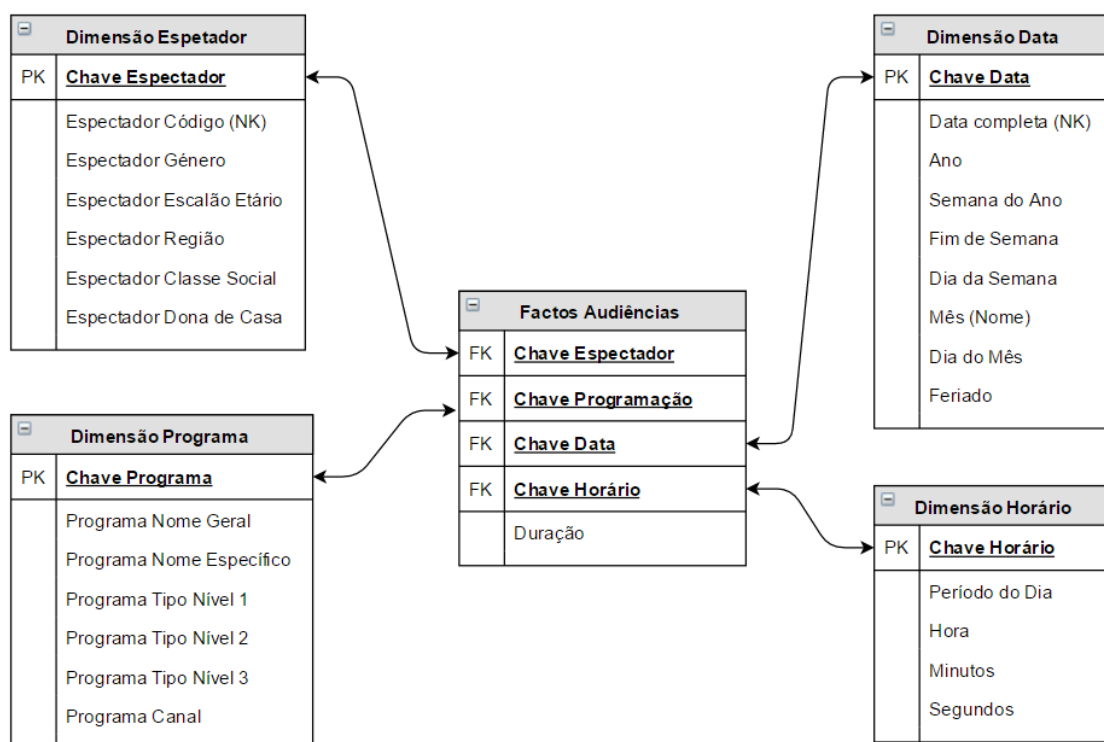
A dimensão com as horas foi separada da dimensão data, para controlar o crescimento da última dimensão, que se poderia tornar numa dimensão monstra. Assim achamos que seria preferível modelar como entidade separada.

4.5. Medidas Numéricas

Para avaliar as tendências das audiências televisivas foi adicionada uma medida aditiva “duração” à tabela de factos, esta permite valores agregados (ex. soma) ao longo de todas as dimensões. Esta medida refere-se ao período de tempo, em minutos e segundos, em que um espetador viu um programa.

4.6. Diagrama em Estrela do *Data Warehouse*

Tendo em conta o que foi mencionado atrás construi-se o seguinte diagrama com as tabelas de dimensões e de factos:



Esquema 2 - Tabelas de Factos e Dimensões

5. Conclusão

Neste relatório preparamos a construção do nosso data warehouse para auxiliar a tomada de decisão no contexto das audiências televisivas.

Analizamos os dados que temos disponíveis para verificar a existência de erros que devem ser tidos em conta aquando do processo de extração, transformação e carregamento para o data warehouse.

Este trabalho tem como foco analisar as tendências televisivas dos espectadores, sendo esse o nosso processo prioritário.

Nesta fase focamo-nos na modelação dimensão do data warehouse. Assim ficou estabelecido que o mais adequado seria uma tabela de factos do tipo transacional e o grão seria um espectador ver um dado programa, numa dada data e hora durante um determinado período de tempo.

6. Bibliografia

Romeu, A. M. (2014). *A MEDIÇÃO DAS AUDIÊNCIAS TELEVISIVAS EM PORTUGAL: NOVAS PRÁTICAS, NOVOS CONSUMOS, NOVOS DESAFIOS*. Universidade Católica Portuguesa-Faculdade de Ciências Humanas.

7.1. Esquema de Gráficos



Esquema 3 - Conjunto de gráficos relativos a todas as fontes de dados, onde é destacado o conjunto de dados relativo aos registos do sexo feminino