



**Ciências
ULisboa**

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

28 de Maio de 2017

Índice

1. Introdução	3
2. Análise das Fontes de Dados	4
2.1. Espectadores	4
2.1.1. Tratamento de Dados	4
2.2. Tipos de Programas Televisivos	8
2.3. Canais Vistos pelos Espectadores.....	10
2.3.1. Tratamento de Dados	10
2.4. Programação dos Canais Televisivos	14
2.4.1. Tratamento de Dados	15
2.5. Classes Sociais dos Espectadores	17
2.6. Fonte de Dados Adicionais.....	17
3. Relação entre Fontes de Dados	18
4. Modelação Dimensional.....	19
4.1. Processo de Negócio.....	19
4.2. Perguntas Analíticas.....	19
4.3. Definição do Grão	19
4.4. Dimensões do Negócio	20
4.4.1. Dimensão Espectador	20
4.4.2. Dimensão Programa.....	21
4.4.3. Dimensão Data	21
4.4.4. Dimensão Horário	22
4.5. Registo de Mudanças Lentas	23
4.6. Medidas Numéricas	24
4.7. Diagrama em Estrela do Data Warehouse.....	25
5. Sistema ETL.....	26
5.1. Extração dos Dados.....	27
5.2. Transformação dos Dados	28
5.3. Carregamento do Data Warehouse.....	30
5.3.1. SQL Server Management Studio	31
5.3.2. SQL Server Business Intelligence Development Studio.....	31
5.3.2.1. Integration Services.....	31
5.3.2.2. Analysis Services.....	33
6. Relatórios Analíticos Dinâmicos	36
6.1. Quais são os tipos de programas televisivos mais vistos por cada faixa etária e estatuto social ao longo da semana e durante o fim-de-semana?	37
6.2. Qual o número de espectadores por canal em cada dia, ao longo do semestre?	43
6.3. Quais os tipos de programa mais vistos, ao longo das horas de um dia?	46
6.4. Por cada região do país, quem passa mais e menos tempo a ver televisão, tendo em conta o estatuto social e se trabalha ou não em casa?	49
6.5. Tendo em conta os feriados nacionais/datas comemorativas no primeiro semestre de 1996, qual a média de espectadores por faixa etária e estatuto social?	51
7. Prospecção de Informação.....	54
7.1. Método de Classificação: Árvores de Decisão	55
7.2. Obtenção dos dados	55

7.3.	Resultados.....	56
8.	Conclusão	63
9.	Bibliografia	64
10.	Anexos.....	65
10.1.	Esquema de Gráficos.....	65
10.2.	Código SQL para Criação das Tabelas.....	66
10.3.	Hierarquias de Atributos	68
10.4	Código SQL para Prospeção de Informação.....	68

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, procedeu-se à modelação dimensional do *data warehouse*, onde foram definidos o grão e o tipo de tabela de factos, bem como as dimensões e medidas envolvidas, por forma a preparar a construção do sistema ETL.

O dito sistema ETL foi construído e demonstrado na terceira fase do projeto, permitindo numa quarta fase compor relatórios analíticos que visam responder às perguntas analíticas definidas na primeira etapa do trabalho.

Um método de prospeção de informação (isto é, *data mining*) foi, também numa quarta e última fase, definido para interpretar, analisar e encontrar padrões consistentes nos dados em questão. Este processo é importante pois permite tomar decisões mais eficientes em relação ao processo de negócio estudado, enquadrando-se assim no presente projeto.

2. Análise das Fontes de Dados

Nesta secção apresentam-se 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 espectadores e feriados e datas festivas.

2.1. Espectadores

O ficheiro “espetadores.csv” contém dados de espectadores. 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 espectadores, 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 espectador	3001
3	Região	Texto	Região do país da residência do espectador	"Gr. Lisboa"
4	Sexo	Texto	Masculino ou feminino	"Femin."
5	DonaDeCasa	Texto	Se o espectador trabalha em casa ou não	"DDC"
6	EscalãoEtário	Texto	Escalão etário do espectador	" +64"
7	Classe	Texto	Classe social do espectador	"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 espectadores 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 espectadores” 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”.

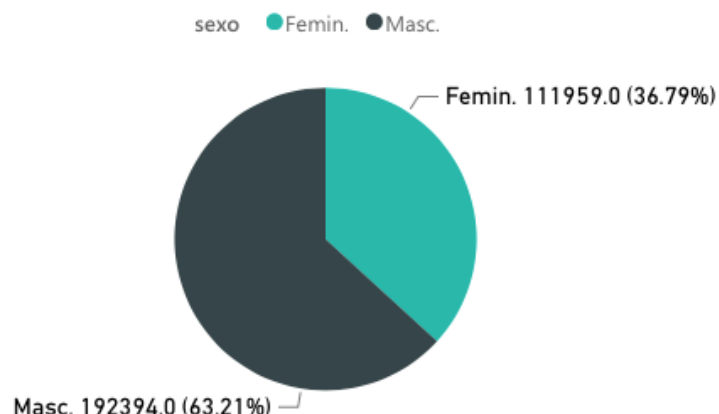
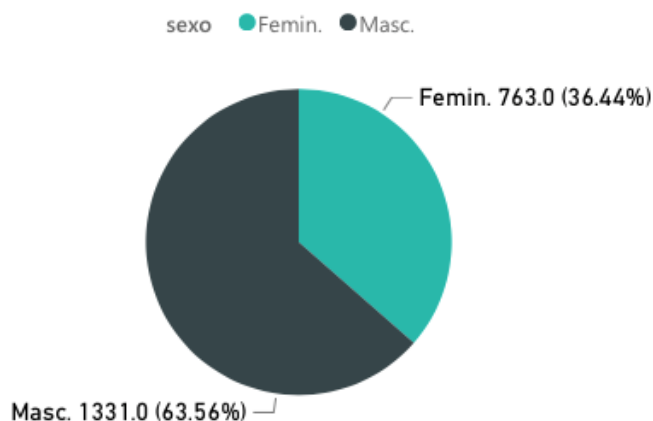


Gráfico 1 - Contagem e percentagem de códigos de espetadores e registos por sexo

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.

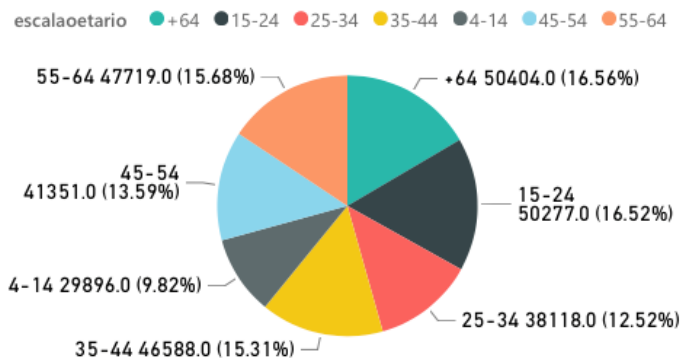
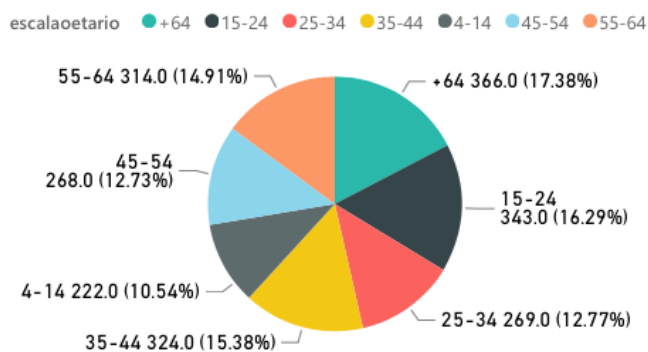


Gráfico 2 - Contagem e percentagem de códigos de espetadores e registos por escalão etário

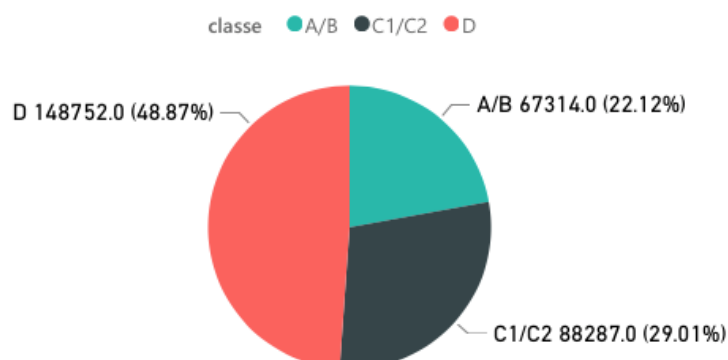
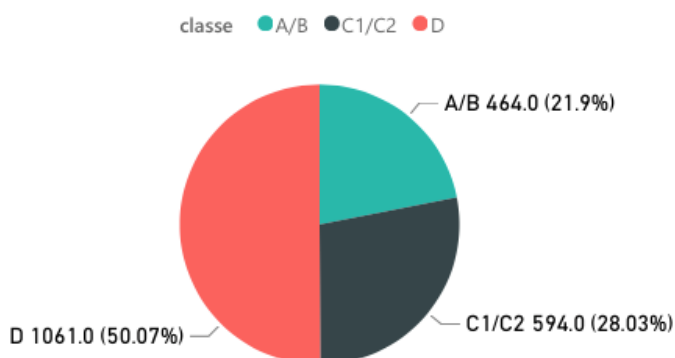


Gráfico 3 - Contagem e percentagem de códigos de espetadores e registos por classe social

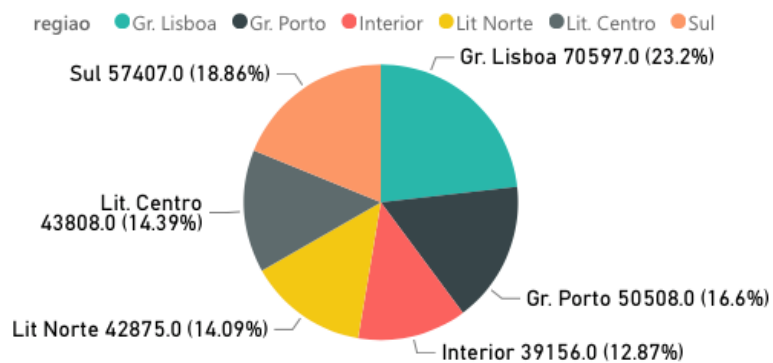
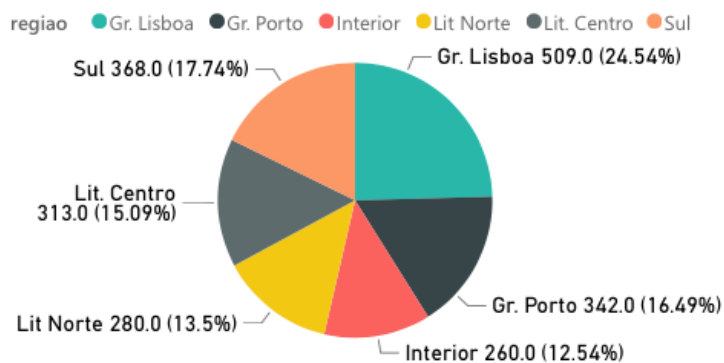


Gráfico 4 - Contagem e percentagem de códigos de espetadores e registos por região

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

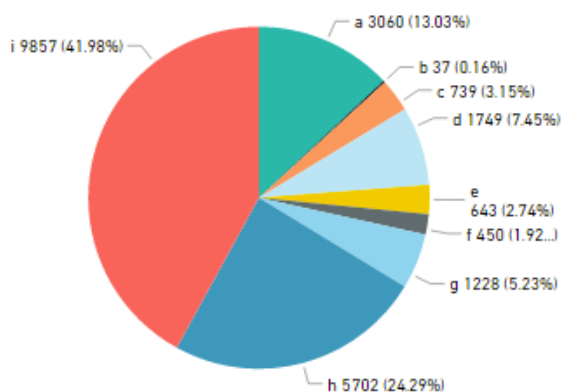
Tabela 5 - Designação dos Tipos Principais de Programas

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

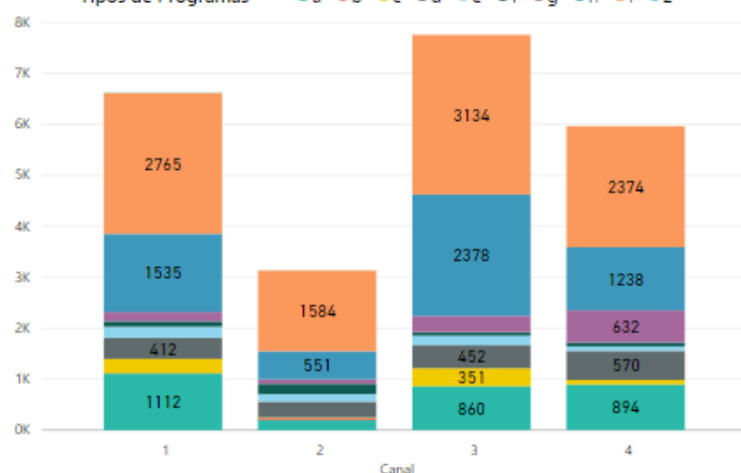


Gráfico 5 - Contagem dos Tipos Principais de Programas e Tipos de Programas por Canal

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 6.

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 7.

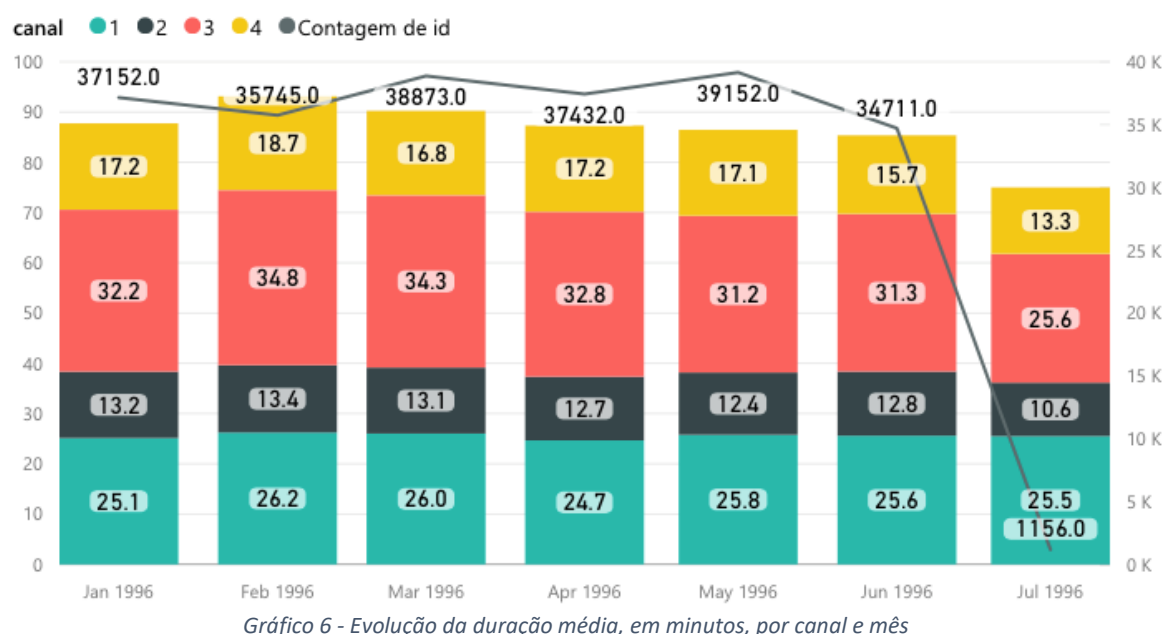
Tabela 7 - Análise de estatística descritiva sobre os campos da variável “audiencias”

	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
						“4”:	411176

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.



No gráfico 6 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.

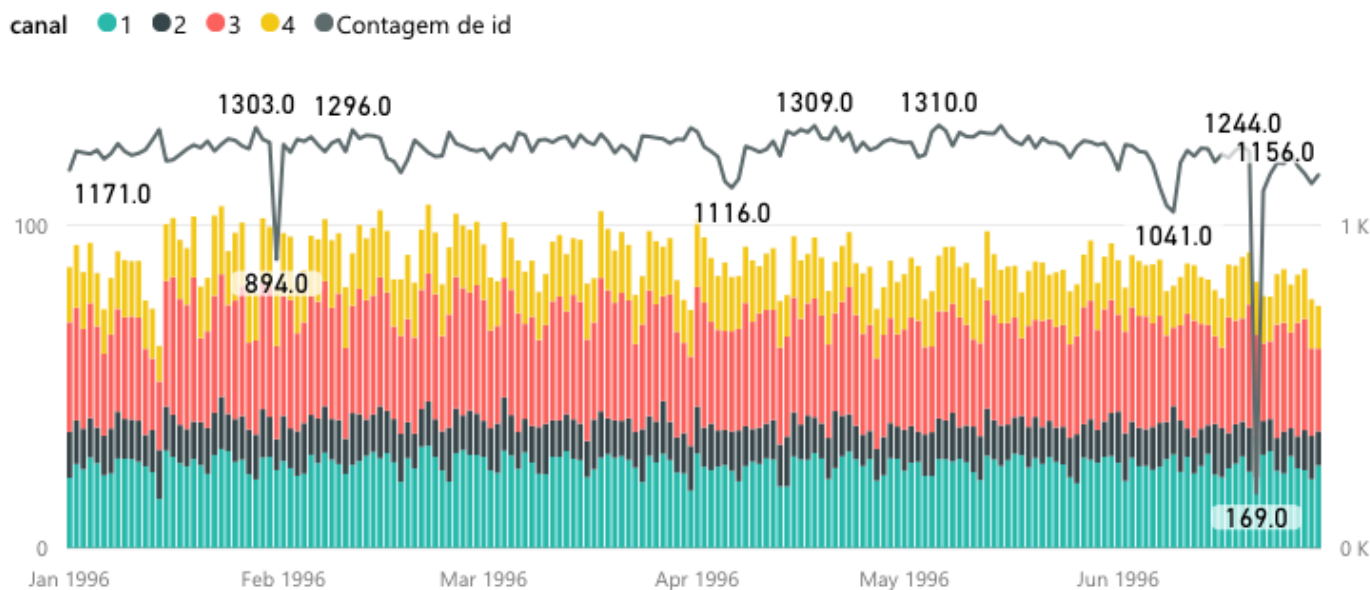


Gráfico 7 - Evolução da duração média, em minutos, por canal e dia (1 jan. a 30 junho)

No gráfico 7 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 8 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.

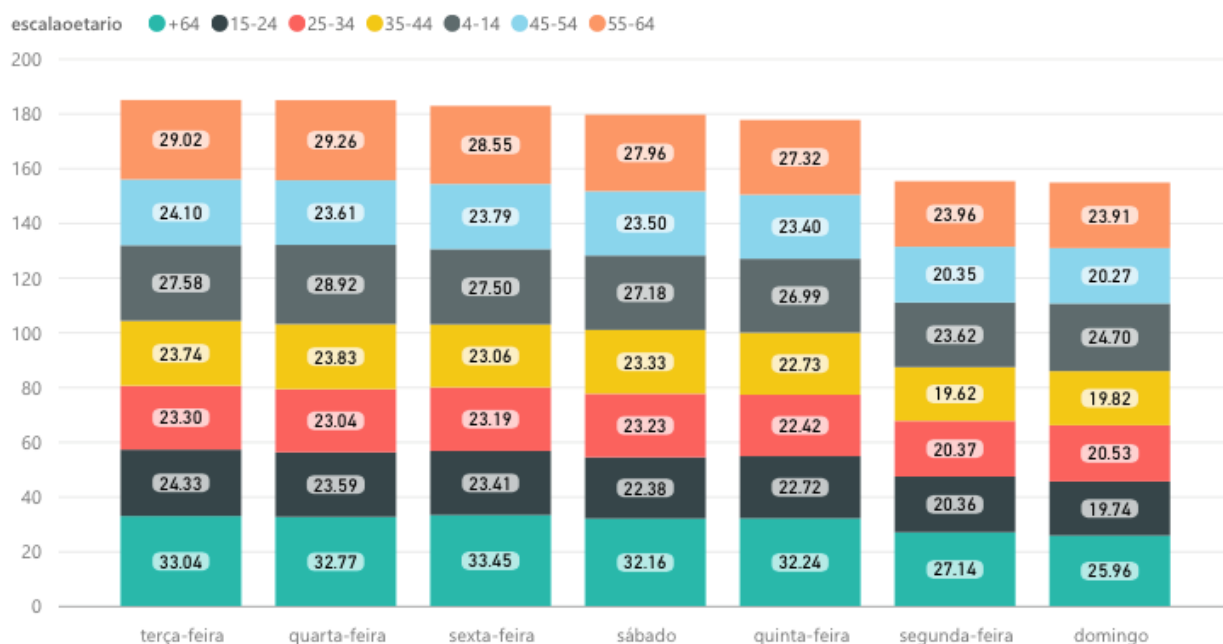


Gráfico 8 - Duração média, em minutos, por dia da semana e escala etária, ordenada segundo a grandeza total

Por fim, no gráfico 9 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).

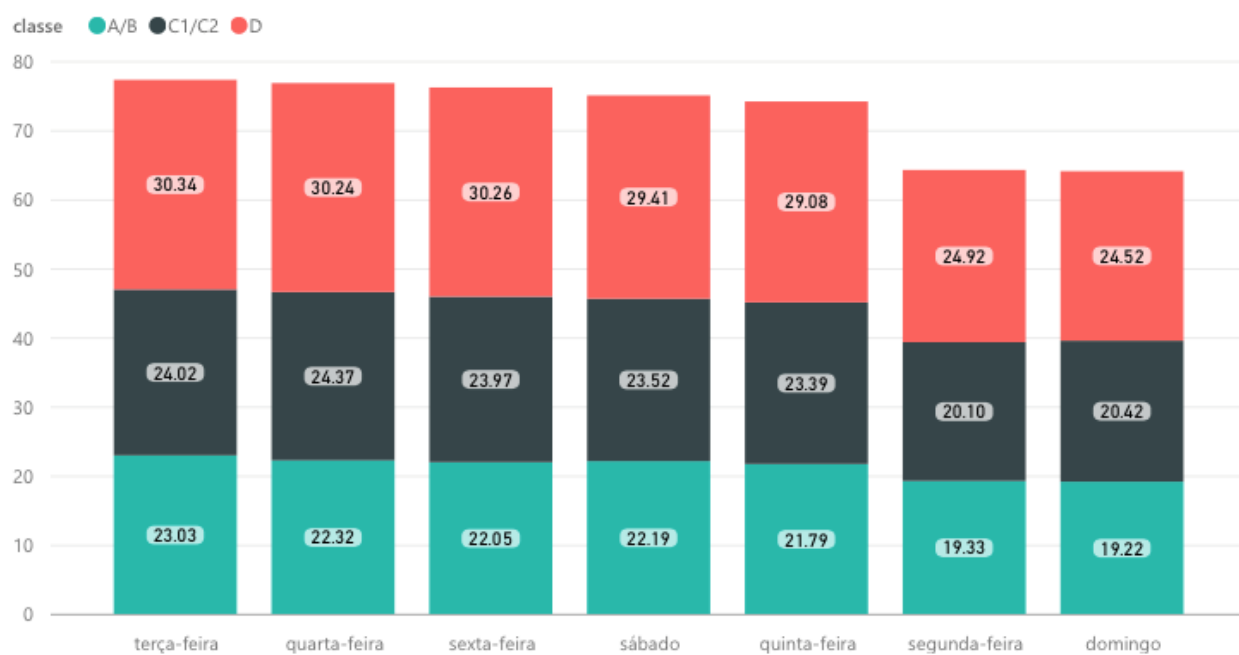


Gráfico 9 - Duração média, em minutos, por dia da semana e classe social, ordenada segundo a grandeza total

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 "HoralInicio", "Duracao", "Canal", "Classificacao" e "ParteTodo" da variável "todos.dados.pet" obtida através da função "summary"

	HoralInicio	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:

1. Vírgulas em falta

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

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

Correção: vírgula adicionada manualmente.

2. 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.

3. 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:

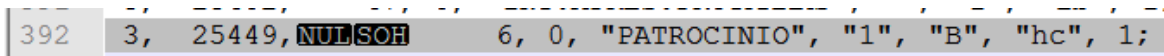
```
3, 24202, 43, 0, "PUBLICIDADE ASSOCIADA", "1", "B", "ha",
```

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

4. Caracteres inválidos

Exemplo: Ficheiro "19960610.pet", linha 392:

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



```
392 3, 25449, NULSOH 6, 0, "PATROCINIO", "1", "B", "hc", 1;
```

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

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

Por fim, foi criado um ficheiro "programas.tsv", com todos os dados extraídos dos ficheiros ".pet".

2.5. Classes Sociais dos Espetadores

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.

Tabela 12 - Listagem da fonte de dados adicional “calendário”

#	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

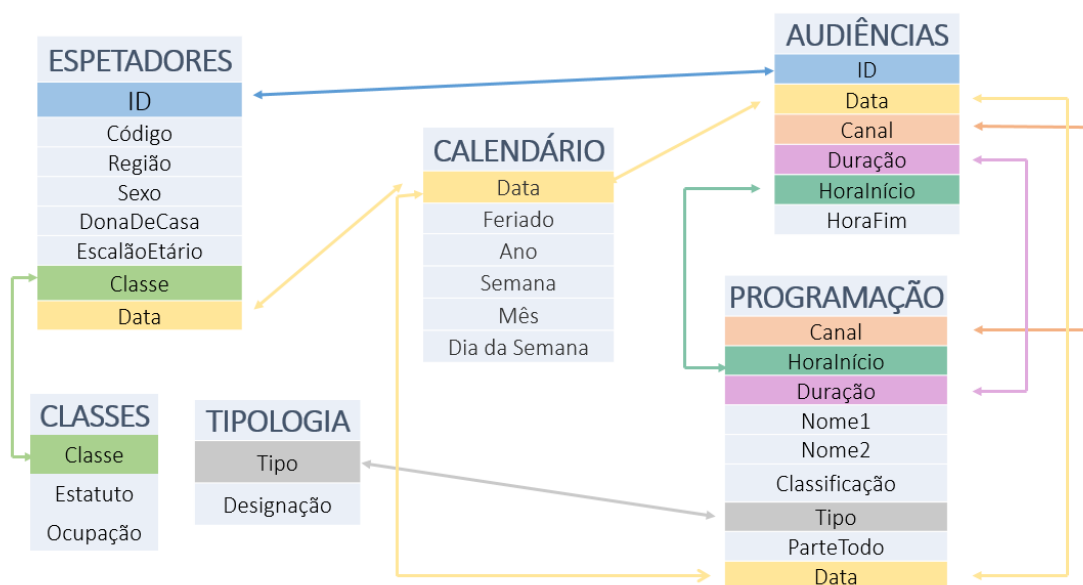
No diagrama de dados existe uma fonte de dados “Calendário” que apresenta outros atributos como ano, semana, mês e dia da semana, cuja representação neste ficheiro

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

não é necessária, visto que os seus respetivos valores podem ser obtidos recorrendo a funções em SQL ou Excel.

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 tipos de programas televisivos mais vistos por cada faixa etária e estatuto social ao longo da semana e durante o fim-de-semana?
- Qual o número de espectadores por canal em cada dia durante o primeiro semestre de 1996?
- Quais os tipos de programa mais vistos, ao longo das horas do dia?
- Por cada região do país, quem passa mais e menos tempo a ver televisão, tendo em conta o estatuto social e se trabalha ou não em casa?
- Tendo em conta os feriados nacionais e datas comemorativas, qual a média de espectadores por faixa etária e estatuto 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 determinada data e hora, durante um período de tempo.”

De notar que não há distinção entre as partes de um programa, sendo apenas analisada a duração do programa que o espetador vê.

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**.

4.4.1. Dimensão Espetador

A tabela Dimensão Espetador gera tantas linhas quanto o número de espetadores (isto é, tantas linhas quanto o número de códigos diferentes) existentes. Neste caso, a tabela de dimensão terá 2071 linhas.

Foi identificada uma hierarquia correspondente ao estatuto/classe social, que é definida da seguinte forma:

→ **Estatuto do Espetador**

- Estatuto Máximo do Espetador
 - Estatuto Mínimo do Espetador

O significado de cada atributo da tabela é sumariado na tabela 13.

Tabela 13 - Significado de cada atributo da tabela “Dimensão Espetador”

Atributos	Tipo de Dados	Descrição	Origem dos Dados	Exemplo
Chave Substituta Espetador	Número inteiro	Identificador único de espetador	Criado manualmente	1
Código (Chave Natural)	Número inteiro	Código identificador do espetador	Espetadores	6
Género	Texto	Masculino ou feminino		"Femin."
Escalão Etário	Texto	Escalão etário do espetador		" +64"
Região	Texto	Região do país da residência do espetador		"Gr. Lisboa"
Estatuto 1	Texto	Classe social principal do espetador		"Classe média/alta"
Ocupação 1	Texto	Descrição da ocupação do espetador, tendo em conta o seu estatuto		"Gestor, administrador, ou profissional de topo"

Estatuto 2	Texto	Classe social secundária do espetador, caso existente		"Estatuto não definido"
Ocupação 2	Texto	Descrição da ocupação do espetador, tendo em conta o seu estatuto, caso existente		"Ocupação não definida"
Dona de Casa	Texto	Se o espetador trabalha em casa ou não		"DDC"

4.4.2. Dimensão Programa

A tabela referente à Dimensão Programa gera tantas linhas quantos os programas registados, sendo que foi considerado "um programa" como a junção de todos os múltiplos registos de um mesmo programa, ou seja, consideraram-se os programas cuja "parte todo" era igual a 0.

Foram também considerados apenas os programas de classificação "P", ou seja, os intervalos comerciais ("B") e publicidades ao próprio canal ("I") foram ignoradas, por não serem de relevo para o processo de negócio em análise.

Nesta dimensão identificaram-se 2 hierarquias distintas: uma para os nomes dos programas, e outra para o tipo de programa. Estas hierarquias são demonstradas em seguida, e estão incluídas na descrição dos elementos da dimensão, na tabela 14.

→ Nome Programa

- Nome geral
 - Nome específico

→ Tipo Programa

- Tipo
 - Categoria
 - Género

Tabela 14 - Significado de cada atributo da tabela "Dimensão Programa"

Atributos	Tipo de Dados	Descrição	Origem dos Dados	Exemplo
Chave Substituta Programa	Número inteiro	Identificador único do programa	Criado manualmente	1
Nome Geral	Texto	Nome do conteúdo televisivo	Programação	"FILME"
Nome Específico	Texto	Um segundo nome do conteúdo televisivo		"CURTO CIRCUITO II"
Canal	Texto	Número do canal no ar		4
Tipo	Texto	Tipo do conteúdo		"FICÇÃO"
Categoria	Texto	Tipo do conteúdo		"FILME"
Género	Texto	Tipo do conteúdo		"Comédia"

4.4.3. Dimensão Data

A tabela "Dimensão Data" gera tantas linhas quanto o número de dias do período em análise; neste caso, serão 183 linhas.

Na dimensão Data identificaram-se 2 hierarquias possíveis a partir do ano, que são identificadas em seguida, assim como na tabela 15.

→ **Ano**

- Semana do ano
 - Dia da Semana
- Mês
 - Dia do mês

→ **Data Comemorativa**

- Feriado

Tabela 15 - Significado dos atributos da tabela “Dimensão Data”

Atributos	Tipo de dados	Descrição	Origem dos Dados	Exemplo
Chave Substituta de Data	Número inteiro	Identificador único da data	Criado manualmente	19960401
Data Completa (Chave Natural)	Data	Data completa		1996-04-01
Dia do Mês	Número inteiro	Número referente ao dia do mês	Calendário	4
Nome do Mês	Texto	Nome do mês do ano		“janeiro”
Ano	Número inteiro	Número referente ao ano em questão		1996
Dia da Semana	Texto	Nome do dia da semana		“Segunda-feira”
Semana do Ano	Número inteiro	Número referente à semana do ano, de 1 a 52		1
Fim-de-semana	Texto	Se o dia da semana é referente a fim-de-semana ou não		“dia de semana”
Indicador Data Comemorativa	Texto	Se o dia é uma data comemorativa ou não		“Data Não Comemorativa”
Nome Data Comemorativa	Texto	Nome da Data Comemorativa		“Data Não Comemorativa”
Indicador Feriado	Texto	Se o dia da semana é referente a um feriado ou não		“Não feriado”

4.4.4. Dimensão Horário

Por forma a controlar o crescimento da dimensão data, a dimensão referente às horas do dia foi separada desta, evitando tornar a dimensão data numa dimensão monstra.

Esta tabela de dimensão gera tantas linhas quanto o número de horas, minutos e segundos de um dia, isto é, 86 400 linhas.

Para este caso, identificámos apenas uma hierarquia, referente ao período do dia. Novamente, esta hierarquia é demonstrada em seguida, assim como na tabela 16, referente aos atributos da dimensão

→ **Período do dia**

- Hora

- Minutos
 - Segundos

Tabela 16 - Significado dos atributos da tabela “Dimensão Horário”

Atributos	Tipo de dados	Descrição	Origem dos Dados	Exemplo
Chave Substituta Horário	Número inteiro	Identificador de Horário	Criado manualmente	124751
Horário Completo	Número inteiro	Identificador único do horário		“12:47:51”
Período do Dia	Texto	Se a hora do dia corresponde ao período da manhã (7h-12h), tarde (12h-18h), noite (18h-00) ou madrugada (00h-7h)		“Manhã”
Hora	Número inteiro	Hora do dia		12
Minutos	Número inteiro	Minutos correspondentes à hora do dia		47
Segundos	Número inteiro	Segundos correspondentes à hora do dia		51

4.5. Registo de Mudanças Lentas

Para registar a mudanças de dados referentes aos espetadores, como por exemplo mudança de classe etária, classe social, região ou dona de casa, optámos pela técnica do Tipo 2 - **acrescentar uma linha** na tabela de Dimensão Espetadores. Esta estratégia é indicada para manter o histórico destas alterações.

Para além disto, foram acrescentadas 3 colunas: “Data Inicio”, “Data Fim” e “Em vigor”. Estas colunas permitem-nos organizar os dados de uma forma mais coerente e obter as respostas indicadas à pergunta analítica em questão.

Tabela 17 - Dimensão Espetador, tendo em consideração o registo de mudanças lentas.

Atributos	Tipo de Dados	Descrição	Origem dos Dados	Exemplo
Chave Substituta Espetador	Número inteiro	Identificador único de espetador	Criado manualmente	1
Chave Supernatural Espetador	Número Inteiro	Identificador único de espetador		
Código (Chave Natural)	Número inteiro	Código identificador do espetador	Espetadores	6
Género	Texto	Masculino ou feminino		"Femin."
Escalão Etário	Texto	Escalão etário do espetador		" +64"
Região	Texto	Região do país da residência do espetador		"Gr. Lisboa"
Estatuto Máximo de Espetador	Texto	Classe social de maior nível do espetador		"Classe média/alta"
Ocupação do Estatuto Máximo	Texto	Descrição da ocupação do espetador, tendo em conta o seu estatuto		"Gestor, administrador, ou profissional de topo"
Estatuto Mínimo do Espetador	Texto	Classe social secundária do espetador, caso existente		"Estatuto não definido"

Ocupação do Estatuto Mínimo	Texto	Descrição da ocupação do espetador, tendo em conta o seu estatuto, caso existente		"Ocupação não definida"
Dona de Casa	Texto	Se o espetador trabalha em casa ou não		"DDC"
Data Início	Data	Data respeitante ao dia em que começou a ser visualizado um programa	Criado manualmente	1996-01-01
Data Fim	Data	Data respeitante ao dia em que terminou de ser visualizado um programa		1996-07-01
Em Vigor	Texto	Se a linha se encontra em vigor, <i>ie</i> , possui a informação atualizada/corrente		"TRUE"

4.6. Medidas Numéricas

Para avaliar as tendências das audiências televisivas, foi adicionada uma medida aditiva "duração" à tabela de factos.

Esta medida refere-se ao período de tempo, em minutos, em que um espetador viu um programa, e permite o cálculo de valores agregados (ex. soma) ao longo de todas as dimensões.

O tempo mínimo considerado para a visualização de um canal é de 1 minutos (60 segundos), uma vez que esta é a frequência mínima de registo.

4.7. Diagrama em Estrela do *Data Warehouse*

De acordo com o processo de negócio tido em conta, o respetivo grão e dimensões de negócio, foi construído o diagrama demonstrado em seguida, com as correspondentes tabelas de dimensões e de factos.

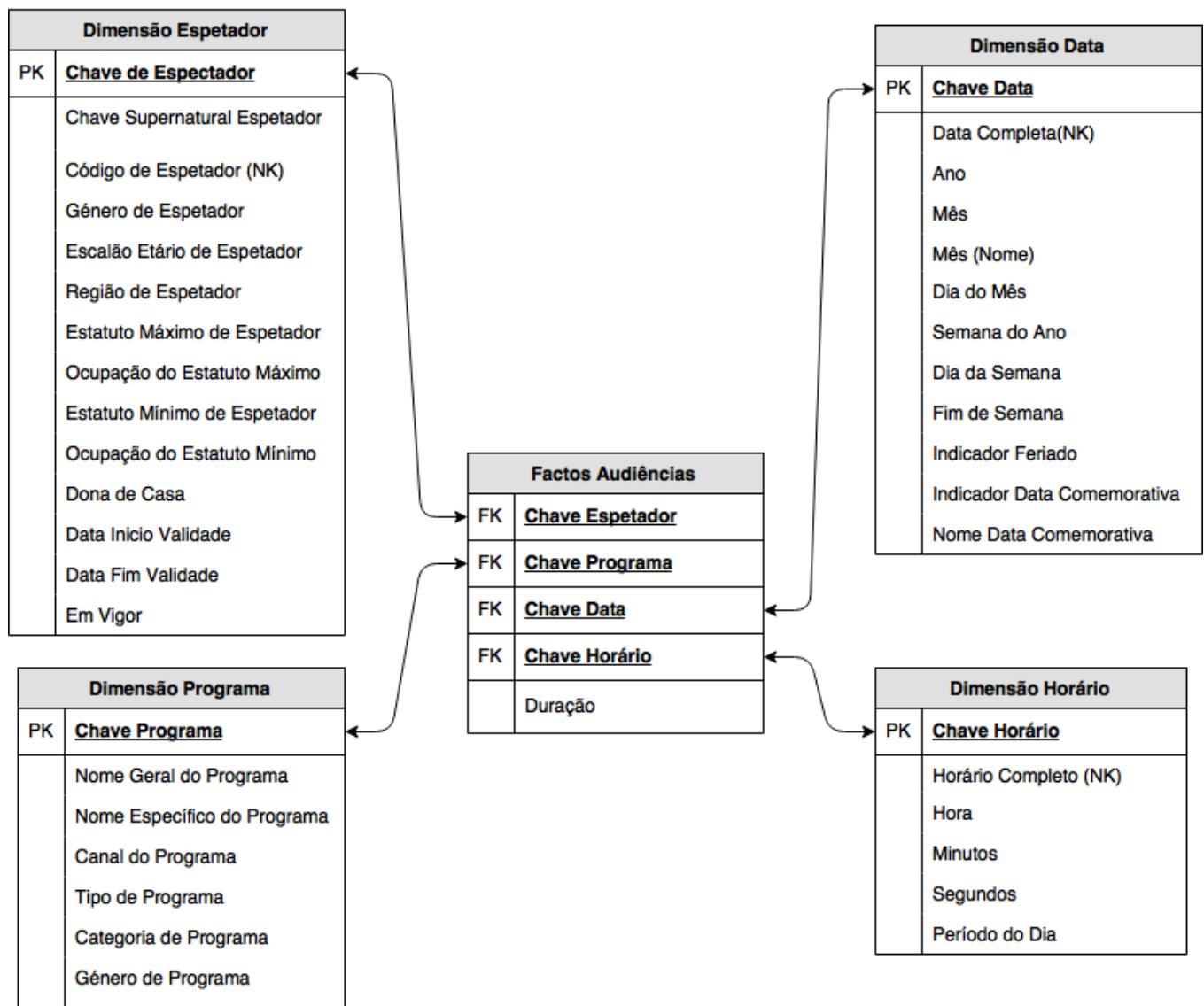


Diagrama 1 - Tabelas de Factos e Dimensões

5. Sistema ETL

Depois de definida a modelação dos dados e estruturado o *Data Warehouse*, é necessário povoá-lo com os dados organizados. Para tal, recorre-se ao sistema ETL. Este processo é uma das fases mais críticas na construção de um *data warehouse*, pois é nesta fase que grandes volumes de dados são processados.

Extração, Transformação e Carga (Extract, Transform & Load - ETL) são etapas de uma técnica que permite às organizações extrair dados de fontes de informação diversas e reformulá-los e carregá-los para uma nova aplicação (Data Warehouse, com recurso a base de dados) para análise.

Através do diagrama 2, pode obter-se uma visão geral de todas as fases seguidas neste processo.

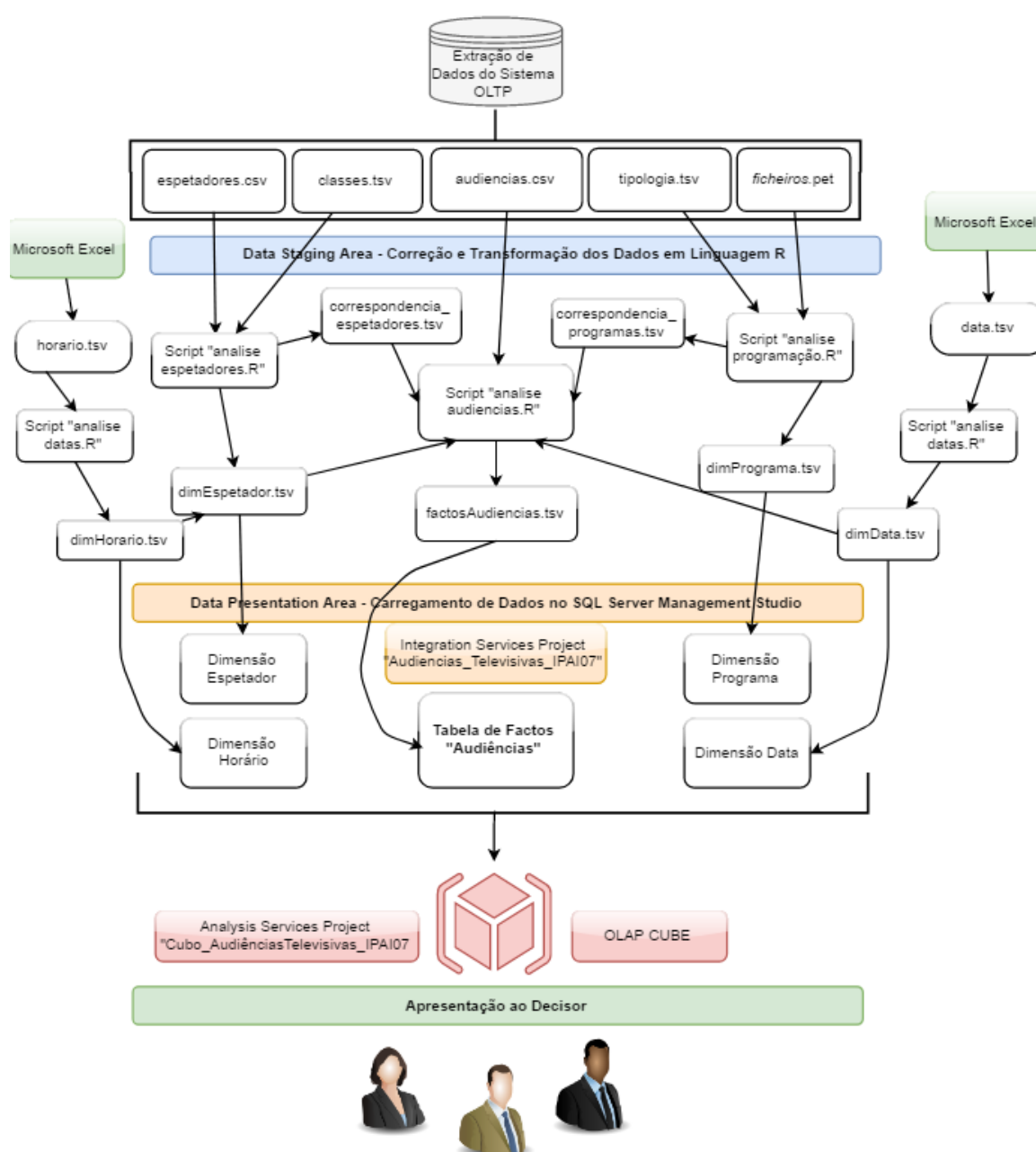


Diagrama 2 - Diagrama com fluxo de dados do Sistema ETL.

Os diversos programas elaborados e descritos no diagrama 2 são sumariados na tabela 18.

Tabela 18 - Responsabilidade da etapa "Extração" do sistema ETL e de cada programa

Programa	Responsabilidade	Entrada	Saída
analise datas.R	- Transforma a formação ficheiros de data e horários gerados em excel para dar o tipo correcto aos dados (nomeadamente remover 0 extra da chave).	data.tsv horario.tsv	dimData.tsv dimHorario.tsv
analise espetadores.R	- Substitui as abreviaturas pelas classes dos espetadores; - Corrige erros nos espetadores e remove dados sem sentido; - Processa mudanças lentas nos espetadores; - Torna os campos de espetador mais inteligíveis; - Cria a dimensão espetador; - Cria a tabela de correspondência entre espectadores extraídos e dimensão espetador.	espetadores.csv classes.tsv	dimEspecadores.tsv correspondencia_espetador.tsv
analise programas.R	- Substitui os tipos pela respetiva informação; - Seleciona os programas de interesse; - Cria a dimensão programa; - Cria a tabela de correspondência entre programas extraídos e dimensão programa.	ficheiros ".pet" tipologia.tsv	dimPrograma.tsv correspondencia_programa.tsv
analise audiencias.R	- Cruza a informação registada nas audiências com datas, horário, programas e espectadores; - Preenchimento de valores em falta; - Gera a tabela de factos; - Remoção de dados duplicados da tabela de factos.	audiências.csv dimData.tsv dimHorario.tsv correspondencia_espetador.tsv correspondencia_programa.tsv	factosAudiencias.tsv

5.1. Extração dos Dados

A **extração** é a primeira etapa no processo de obtenção de dados para o ambiente do *data warehouse*. Extrair significa ler e compreender a fonte de dados, e transcrever os dados necessários para a *Data Staging Area*, para futura manipulação.

As fontes de dados originais para extração de dados ("classes.tsv", "tipologia.tsv", "audiências.csv", "espetadores.csv" e ficheiros ".pet") foram fornecidas pelo docente. Para além destas, acrescentou-se uma fonte adicional "Calendário". Estas informações de data e hora foram geradas no Microsoft Excel e guardadas em ficheiros: "data.tsv" e "horario.tsv".

A estas fontes foi aplicada uma análise estatística (descrita na secção 2), que nos permitiu ter uma visão geral sobre como é que os dados se interligam e se comportam.

As diferentes ferramentas utilizadas para a extração dos dados e suas responsabilidades no processo são descritos na tabela 19.

Tabela 19 - Responsabilidades da etapa "Extração" do sistema ETL e de cada ferramenta

Ferramenta	Responsabilidade	Entrada	Saída
R Studio	→ Receção - Receção dos dados das várias fontes; - Detecção de alterações nos dados; - Aplicação de filtros para detecção; - Ordenação dos dados; - Primeira Limpeza dos dados; - Tratamento de exceções; - Eliminação de duplicados.	audiencias.csv espetadores.csv ficheiros ".pet" classes.tsv tipologia.tsv	-
MS Excel	→ Receção - Criação manual de fonte de dados.	-	data.tsv horario.tsv

5.2. Transformação dos Dados

Assim que os dados são extraídos para a *Data Staging Area* (área de trabalho para processar dados em bruto), existem inúmeras potenciais transformações (mais complexas que numa primeira análise), como: tradução de valores codificados, aplicação da transformação apenas a determinadas categorias de linhas e/ou colunas, conflitos de domínio, lidar com elementos ausentes, fusão (*merging*) ou agregação dos dados.

Os dados-fonte "classes.tsv" e "tipologia.tsv" são meramente informativos e, portanto, não foram sujeitos a nenhuma correção; relativamente às restantes fontes de dados, os erros detetados e corrigidos são descritos nos respetivos subcapítulos da secção 2.

Os dados fornecidos foram corrigidos e preparados para construção das dimensões e tabela de factos definidas anteriormente, recorrendo a duas ferramentas: Microsoft Excel, e R Studio.

As responsabilidades de ambas as ferramentas são descritas na tabela 20, e a transformação da tabela de factos e dimensões é descrita em respetivo subcapítulo.

Tabela 20 - Responsabilidades da etapa "Transformação" do sistema ETL e de cada ferramenta

Ferramenta	Responsabilidade	Entrada	Saída
R Studio	→ Integração - Segunda Limpeza dos Dados; - Geração de valores de chaves substitutas para todas as dimensões; - Fusão de duplicados; - Preenchimento de valores em falta; - Correspondência entre os diversos dados; - Criação das dimensões e tabela de factos.	audiencias.csv espetadores.csv ficheiros ".pet" classes.tsv tipologia.tsv data.tsv horário.tsv	factosAudencias.tsv dimEspectador.tsv dimPrograma.tsv correspondência_espetadores.tsv correspondência_programas.tsv dimHorario.tsv dimData.tsv

5.2.1. Dimensão Data e Dimensão Horário

As dimensões Data e Horário foram geradas automaticamente com recurso à ferramenta Microsoft Excel, sendo que cada célula corresponde a um dia (no caso da dimensão Data) e a uma hora, minuto e segundo (no caso da dimensão Horário).

Para a dimensão Data, para além das colunas referentes à data completa, dia, mês e ano, foram ainda adicionadas outras colunas descritivas, para uma melhor perceção por parte do decisor. São elas “nome do mês”, “semana do mês”, “dia da semana”, “fim de semana”, “indicador feriado”, “indicador data comemorativa”, “nome da data comemorativa”.

De igual forma, na dimensão Horário, para além das colunas referentes à hora completa, hora, minutos e segundos, foi adicionada uma coluna descritiva “período do dia”.

5.2.2. Dimensão Espetador

A dimensão Espetador foi criada com o recurso à fonte de dados “espetadores.csv”. Esta fonte continha 8 campos: “ID”, “Código”, “Género”, “Região”, “Classe Social”, “Escalão Etária”, “Dona de Casa” e “Data”.

Destes campos, 6 foram reaproveitados para a criação da dimensão: “Código”, “Género”, “Região”, “Classe Social”, “Escalão Etária” e “Dona de Casa”. Para além disto, foram também acrescentadas 3 colunas respetivas ao registo de mudanças lentas como descrito na secção 4.5.

Esta dimensão resultou do cruzamento entre os ficheiros “espetadores.csv” e “classes.tsv”.

5.2.3. Dimensão Programa

Relativamente à dimensão Programa, os dados que a constituem são provenientes de duas fontes iniciais:

- os ficheiros de formato “.pet” foram reunidos num só ficheiro, “programas.tsv”, de onde são retiradas todas as principais informações sobre cada programa visualizado;
- o ficheiro “tipologia.tsv”, que serviu para fazer a correspondência descritiva com os tipos, categorias e géneros de cada programa.

Assim, a dimensão Programa possui 6 campos, para além do campo referente à sua chave substituta: “nome geral”, “nome específico”, “canal”, “tipo”, “categoria” e “género”.

5.2.4. Tabela de Factos Audiências

A tabela de factos possui cinco atributos, sendo que quatro são chaves estrangeiras para as dimensões criadas (“Espetador”, “Programa”, “Data Início” e “Hora Início”), e o último atributo é a medida numérica definida, a duração.

Cada atributo corresponde à chave primária de cada dimensão criada, exceto o atributo “Duração”. Assim sendo:

- O atributo “Espetador” acomoda a chave primária da dimensão Espetador, ou seja, a chave substituta deste (“Chave Espetador”);
- O atributo “Programa” diz respeito à chave primária da dimensão Programa, também a sua chave substituta (“Chave Programa”);
- O atributo “Data Início” corresponde à chave primária da dimensão Data, ou seja, a sua chave substituta (“Chave Data”). Esta foi denominada “Data Início” porque refere-se à data que um espetador começou a visualizar um programa;
- O atributo “Hora Início” tem em consideração a chave primária da dimensão Horário, isto é, a sua chave substituta. Esta foi denominada “Hora Início” visto que se refere a uma certa hora (contêm horas, minutos e segundos) que um espetador começou a visualizar um programa.

Relativamente à “Duração”, que corresponde à medida numérica do data warehouse, é usada para avaliar o processo de negócio em causa. Esta medida soma os tempos de visualização de um programa ao longo de todas as dimensões, por exemplo, ao longo de um dia, de uma hora, por escalão etário ou até por região.

A duração foi obtida através do cálculo da diferença da hora de fim e hora de início do ficheiro “audiências.csv”.

Para complementar esta explicação, pode observar-se o Diagrama 1 da secção 4.7.

5.3. Carregamento do Data Warehouse

Esta etapa consiste em estruturar e carregar os dados modificados para uma base de dados relacional. O carregamento pode ser simples (reescrever dados novos por cima de antigos) ou mais completo em termos de dados históricos (mantendo um registo de todas as alterações efetuadas).

Para este efeito, foi utilizado o *SQL Server Management Studio*, mais concretamente a ferramenta *Integration Services*, seguindo o modelo dimensional. Posteriormente, utilizou-se a ferramenta *Analysis Services* do *SQL Server Business Intelligence* para o carregamento e construção do cubo de dados.

Esta ferramenta tem como principais vantagens admitir conjuntos de dados de grande dimensão, assim como permitir um controlo preciso sobre as dimensões e medidas, incluindo a definição de hierarquias de atributos.

Os programas utilizados nesta etapa e respetivas responsabilidades são sumarizados na tabela 21.

Tabela 21 - Responsabilidades da etapa “Carregamento” do sistema ETL e de cada programa

Programa	Responsabilidade	Entrada	Saída
Microsoft SQL Server Management Studio 2008	→ Distribuição - Criação das Tabelas em Linguagem SQL	Comandos SQL	Tabelas de factos e dimensões

SQL Server Business Intelligence Development Studio (Microsoft Visual Studio 2008)	→ Distribuição - Importação dos dados → Entrega - Transferência de dados para o cubo de dados	Tabelas de factos e dimensões	Cubo de Dados
--	--	-------------------------------	---------------

5.3.1. SQL Server Management Studio

Antes da realização de qualquer atividade de carregamento de dados, é indispensável a conexão à base de dados com as credencias fornecidas pelo docente.

Depois de criadas as tabelas de factos e dimensões, o passo seguinte passa por criar as tabelas correspondentes na base de dados relacional, usando a linguagem SQL, para depois ser efetuado o carregamento dos dados.

Os códigos SQL utilizados para a criação das diferentes tabelas são demonstrados em anexo, na secção 8.2.

Depois de criadas as tabelas em SQL, pode ser feito o carregamento de dados para a base de dados. O processo de carregamento dos dados foi realizado no *SQL Server Business Intelligence Development Studio*.

5.3.2. SQL Server Business Intelligence Development Studio

O processo de extração, transformação, e carregamento de dados para a base de dados relacional pode ser realizado com mais precisão e controlo através da criação de um *Integration Services Project*.

5.3.2.1. Integration Services

Recorrendo a esta ferramenta, começámos por criar um *Integration Services Project* denominado por “AudenciasTelevisivas”.

Uma vez criado, foram seguidos uma série de passos para configurar corretamente as propriedades do projeto:

1. Criação de uma **OLE DB Connection** (“Destino de dados SQL Server”) correspondente ao destino dos dados;

ConnectionManagerType	OleDb
ConnectionString	Data Source=CACILHEIRO\DIFCUL;User ID=IPAI07
DataSourceID	
DelayValidation	False
Description	
Expressions	
ID	{23E92B69-1DBA-4B2C-AB50-E6999ABA0AFC}
InitialCatalog	IPAI07BD
Name	Destino de dados SQL Server
Password	*****
RetainSameConnection	False
ServerName	CACILHEIRO\DIFCUL
SupportsDTCTransactions	True
UserName	IPAI07

Figura 1 - OLE DB Connection correspondente ao destino dos dados.

2. Criação de uma **File Connection** correspondente à fonte dos dados;
3. **Execute SQL Task**;

General	
Name	Recriar Tabelas
Description	Execute SQL Task
Options	
TimeOut	0
CodePage	1252
Result Set	
ResultSet	None
SQL Statement	
ConnectionType	OLE DB
Connection	Destino de dados SQL Server
SQLSourceType	File connection
FileConnection	TabelasFacto_e_Dimensao.sql
IsQueryStoredProcedure	False
BypassPrepare	True

Figura 2 - Detalhes da caixa referente à opção "Execute SQL Task"

4. **Data Flow Task** para as Dimensões e Tabela de Factos (só é possível transferir dados para as tabelas de factos depois de transferir para as tabelas de dimensões);

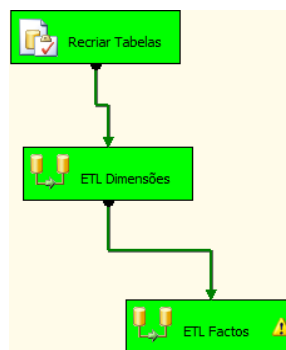


Figura 3 – Data Flow

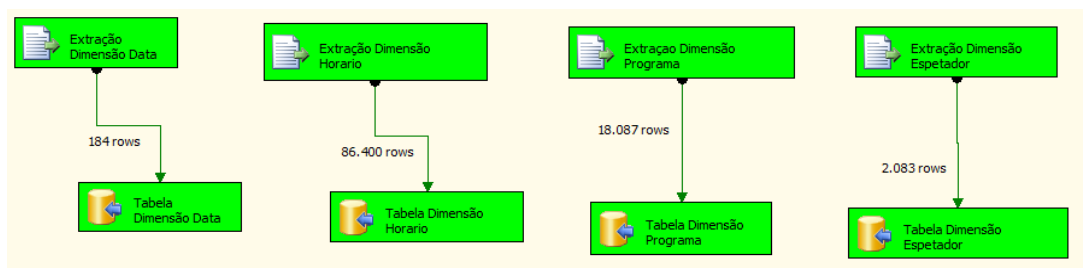


Figura 4 – Data Flow da extração das dimensões

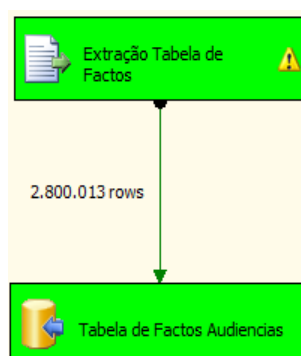


Figura 5 – Data Flow da extração da tabela de factos

Finalizados estes passos, as tabelas criadas no SQL Server foram carregadas com os dados provenientes dos ficheiros “.tsv”: “**dimHorario.tsv**” com **86400** linhas, “**dimEspetador.tsv**” com **2083** linhas, “**dimPrograma.tsv**” com **18087** linhas, “**dimData.tsv**” com **184** linhas e “**factosAudiencias.tsv**” com **2800013** linhas.

Com os dados carregados, foi possível proceder à construção do cubo de dados.

5.3.2.2. Analysis Services

Antes de executar os passos relativos à construção do cubo de dados propriamente dito, é necessário **criar um novo projeto** e **definir as permissões de uso** do cubo. Este novo projeto, de tipo *Analysis Services*, deve ser criado no âmbito da solução que contém o projeto de *Integration Services*.

Uma vez criado o novo projeto de tipo *Analysis Services*, de seu nome “Cubo AudienciasTelevisivas IPAI07” e definidas as permissões de segurança, começámos por **identificar as fontes de dados** que irão sustentar o cubo de dados, fontes essas que irão fornecer os dados atómicos sobre medidas e dimensões.

Neste caso, existe apenas uma fonte, que é a base de dados SQL Server criada com os comandos descritos em 5.3.1. A esta fonte de dados foi dada o nome “FonteDados IPAI07”, e o servidor é “CACILHEIRO\DIFCUL”.

Definida a fonte de dados e configurado o repositório do cubo de dados, o próximo passo consiste em **selecionar as tabelas que fornecem os dados** às tabelas de dimensões e de factos. Para este efeito, foi **criada uma vista sobre as tabelas SQL**, denominada “IPAI07BD”, onde são visíveis as tabelas, respetivos atributos, e as ligações formadas pelas chaves estrangeiras, sendo que o aspeto do esquema é em estrela, com a tabela de factos no centro e as de dimensões em redor.

De seguida, foram dados nomes mais inteligíveis sobre os elementos do esquema em estrela para que, mais tarde, os atributos das dimensões e as medidas da tabela de factos possam ser incluídos num relatório dinâmico com o máximo de informação de contexto possível. Foi ainda aplicado um “*New named calculation*”, denominada “Dado Atuais de Espetadores”, à coluna de dados “em vigor”, para uma melhor interpretação dos dados.

Estas especificações são demonstradas na figura 6.

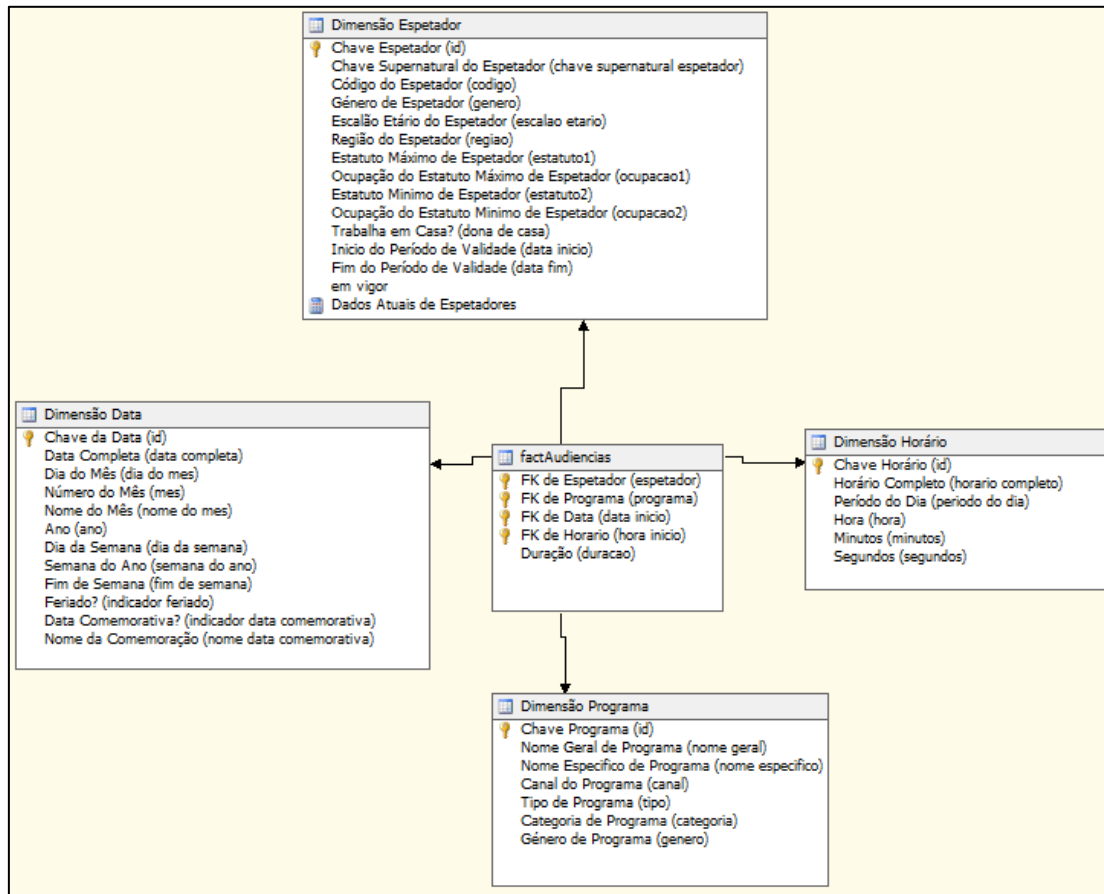


Figura 6 - Esquema em estrela gerado pelo SQL Server Business Intelligence Development Studio

Posteriormente foram **definidas as dimensões** que vão permitir navegar no cubo de dados. Embora o cubo não estivesse ainda criado, nem identificadas as medidas do negócio, já nos foi possível especificar as hierarquias entre os atributos das dimensões (disponíveis na secção 8.3).

Depois de definidas as dimensões e respetivas hierarquias, seguiu-se a etapa de **criação do cubo de dados**, para o qual também vai ser necessário identificar a tabela que guarda os factos e as respetivas medidas. Ao cubo de dados foi dado o nome “Cubo IPAI07”.

Por fim, foi feito o **deploy do cubo**, que serviu para copiar os dados atómicos das tabelas de factos e dimensões em SQL para dentro do cubo, bem como para calcular e armazenar os valores agregados resultantes de todas as combinações possíveis entre atributos das dimensões. O cubo já finalizado é demonstrado na figura 7.

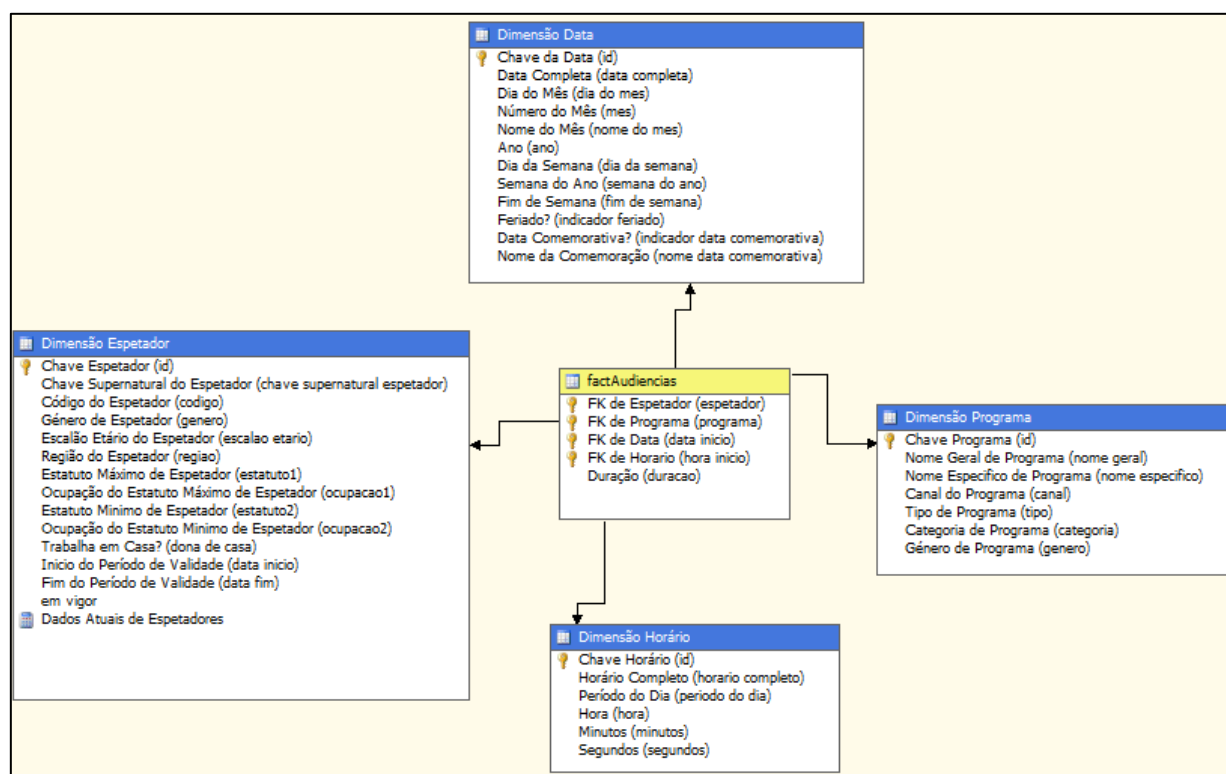


Figura 7 - Cubo de dados gerado pelo SQL Server Business Intelligence Development Studio

Assim, pode resumir-se o processo de criação do cubo de dados pela ferramenta *Analysis Services* nos seguintes passos:

- 1) criação do projeto e definição de permissões de segurança;
- 2) identificação da fonte de dados e configuração do repositório para o cubo de dados;
- 3) criação de vistas mais inteligíveis sobre os dados, caso necessário;
- 4) definição de dimensões e hierarquias de atributos;
- 5) criação do cubo de dados.

6. Relatórios Analíticos Dinâmicos

Depois de criado o cubo de dados no *Analysis Services*, tornou-se possível compor relatórios dinâmicos. Para tal, no campo “*Browser*” do cubo, basta arrastar medidas e atributos de dimensões para dentro do respetivo campo do relatório. Podem ainda ser aplicados diversos filtros, para apenas serem mostrados alguns valores do leque de disponíveis.

Como o *Analysis Services* apenas mostra a informação na forma de tabelas, e como já havia sido utilizado numa primeira fase, decidiu-se voltar a utilizar a ferramenta da Microsoft Power BI para dar resposta às perguntas analíticas definidas no capítulo 4.2.

Uma vez que não foi possível aceder diretamente ao *Analysis Services* pelo Power BI (apesar de haver essa opção), a alternativa encontrada foi fazer o *upload* manual das várias dimensões e da tabela de factos para o projeto do Power BI, fazendo depois as devidas ligações entre as dimensões e os factos na secção “*Relações*”, tal como é demonstrado na figura 8.

Para facilitar a análise nesta ferramenta, duas colunas foram acrescentadas em duas tabelas distintas. Na dimensão Data, acrescentou-se a coluna “Dia da Semana Ordenado” e na tabela de factos, a coluna “Duração minutos”.



Figura 8 - Ligação das várias dimensões à tabela de factos no Power BI

Uma vez definidas as ligações, pôde começar-se a produzir os vários relatórios analíticos correspondentes às questões colocadas na primeira fase.

O relatório completo está disponível na web³, para uma melhor visualização e análise interativa dos dados.

3

<https://app.powerbi.com/view?r=eyJrIjojNGU4MjBmMDktNWU2NS00YjJmLWFKNzktYmUwNTZmY2NIN2VlIiwidCI6IjBiZmE4NTAwLWlxZjltNDU2Ni1iYWYxLTZmNTkzNzA4OTNINyIsImMiOiJh9>

6.1. Quais são os tipos de programas televisivos mais vistos por cada faixa etária e estatuto social ao longo da semana e durante o fim-de-semana?

Para dar resposta a esta questão, foi criada uma série de gráficos de barras empilhadas (gráficos 8 e 9). No caso do gráfico 8, cada barra representa um dos sete dias da semana e a representação (em percentagem) dos vários tipos de programa durante esse dia da semana (em média).

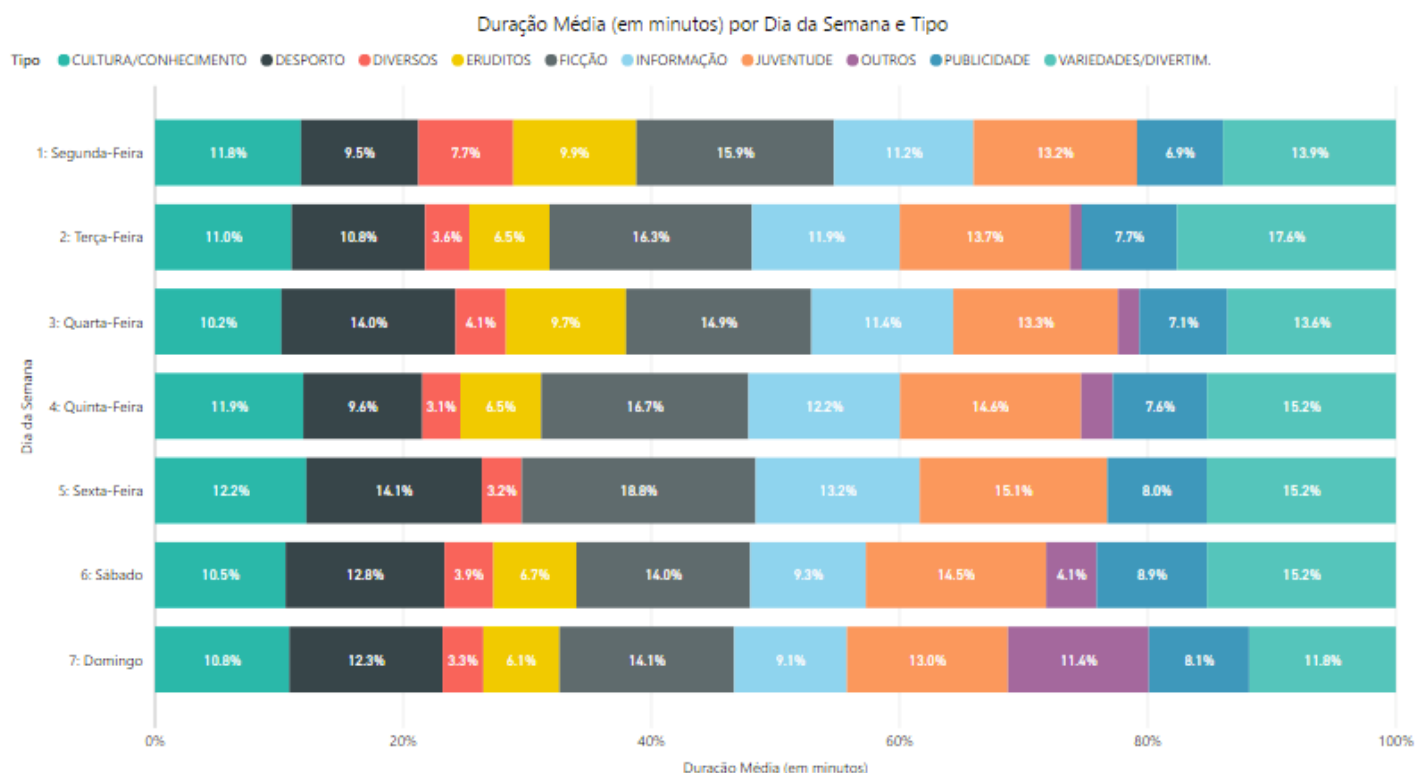


Gráfico 8 - Programas mais vistos tendo em conta a duração por dia da semana, escalão etário e classe social (estatuto).

Esta primeira representação gráfica não tem ainda em conta nem o escalão etário nem o estatuto social do espetador, razão pela qual foi adicionada no relatório em Power BI, à parte, uma caixa de seleção para cada um dos atributos que se pretende explorar. Programas sem tipo definido foram excluídos do gráfico, para uma melhor visualização das restantes proporções temporais.

Assim, e de um modo geral, pode afirmar-se que as os tipos de programas mais vistos pertencem às categorias de ficção (excetuando sábados e domingos), juventude (de uma forma constante ao longo da semana), variedades/divertimento, e em ligeiramente menor percentagem, cultura/conhecimento, informação e desporto.

Por sua vez, os tipos de programas menos vistos inserem-se nas categorias de eruditos, publicidade, diversos e outros.

Olhando para o gráfico 9, outra conclusão que se pode retirar é que ao fim de semana, os espetadores (no geral) dedicam menos tempo a programas de ficção e informação, face aos restantes dias da semana. Também para este gráfico é possível selecionar atributos específicos, embora não seja mostrado na imagem.

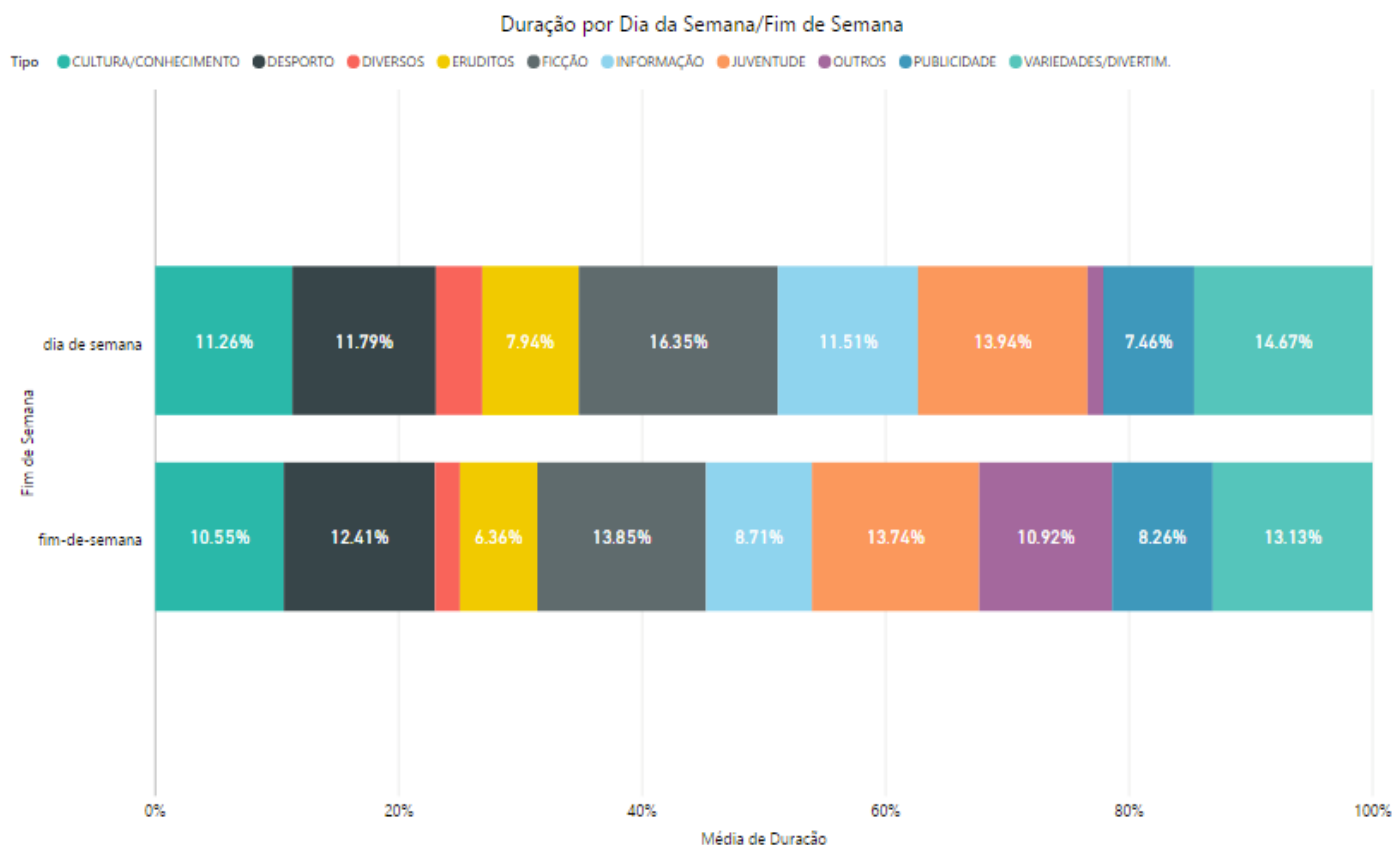


Gráfico 9 - Programas mais vistos tendo em conta a duração por dia da semana ou fim de semana

Quando se faz *drill-down* numa das três categorias de estatuto, verifica-se o seguinte:

- Selecionando a **classe média/alta**, verifica-se imediatamente um aumento percentual no tipo eruditos, em detrimento de uma diminuição percentual em quase todos os outros tipos exceto os relativos à ficção (que mantém a mesma proporção) e juventude (que, por sua vez, aumenta percentualmente);
- Ao optar pela **classe média/baixa**, o tipo de programas que regista um maior aumento percentual de duração pertence à publicidade, enquanto o tipo erudito regista uma diminuição e os restantes tipos mantêm as mesmas proporções;
- Relativamente à **classe trabalhadora**, verifica-se um ligeiro aumento dos tipos “outros” e “tipo não definido”, mantendo-se as restantes proporções.

Se, por sua vez, se optar por particionar as durações apenas relativamente às **faixas etárias**, podem retirar-se as seguintes conclusões:

- Para uma faixa etária dos **4 aos 14 anos**, os tipos de programas mais vistos são, como seria de esperar, pertencentes às categorias de ficção, juventude e variedades/divertimento;
- Passando à faixa etária seguinte, dos **15 aos 24 anos**, há uma acentuada diminuição percentual nos programas de tipo “juventude” (entre 2 a 4 pontos percentuais), mas um aumento nos tipos relativos ao desporto, ficção e

informação. Os tipos mais vistos dizem respeito à ficção e variedades/divertimento, enquanto os menos vistos pertencem aos tipos diversos e eruditos;

- Na faixa etária dos **25 aos 34 anos**, os tipos de programa mais vistos continuam a ser os relativos à ficção e variedades/divertimento;
- Passando para a faixa etária seguinte, dos **35 aos 44 anos**, há um claro aumento do tipo “eruditos”, acompanhado de uma descida no tempo despendido nos tipos que caracterizam a faixa etária anterior (ficção e variedades/divertimento). Pode também observar-se que há uma distribuição do tempo relativamente semelhante pelos tipos relativos a ficção, informação, juventude e divertimento, e que o tipo “eruditos” tem uma componente temporal superior, relativamente à observada nas restantes faixas etárias;
- Dos **45 aos 54 anos**, verifica-se que a semelhança nas proporções dos vários tipos se desfaz, isto é, começa novamente a haver uma superioridade temporal dos tipos ficção e variedades/divertimento relativamente aos restantes tipos de programas, como informação e eruditos;
- A tendência observada na faixa etária anterior é seguida também dos **55 aos 64 anos**, sendo observada uma ligeira diminuição percentual relativa ao tipo “eruditos”;
- Por fim, **a partir dos 64 anos** de idade, o panorama geral mantém-se: há uma superioridade dos tipos de programa relativos a ficção e variedades/divertimento, seguido pelos tipos “juventude”, “cultura/conhecimento”, “desporto” e “informação”, em detrimento de outros tipos como “eruditos”, “diversos” e “outros”.

Um outro tipo de conclusões pode retirar-se quando se misturam os dois tipos de fatores (faixa etária e estatuto).

Selecionando, por exemplo, um escalão etário a partir dos 64 anos e um estatuto de classe média/alta, há uma grande percentagem de tempo dedicada a programas de tipo cultura/conhecimento e eruditos (quer durante o fim de semana, quer durante a semana – embora esta última com maior percentagem).

Por sua vez, a mesma faixa etária, mas de classe média/baixa, dedica muito pouco tempo durante a semana a programas de tipo erudito, fazendo-o apenas ao fim de semana (ao contrário do que se verifica na classe média/alta).

Por fim, para tentar perceber quais as categorias de programas mais vistas dentro de cada um dos três tipos mais vistos (ficção, variedades/divertimento e cultura/conhecimento), foram elaborados três gráficos circulares (com que também se pode interagir através das caixas de seleção).

No gráfico 10 é possível visualizar a distribuição, em percentagem, das várias categorias pertencentes aos programas de tipo “ficção”. É ainda possível perceber que aproximadamente três quartos desta distribuição são relativos a telenovelas, telefilmes e filmes.

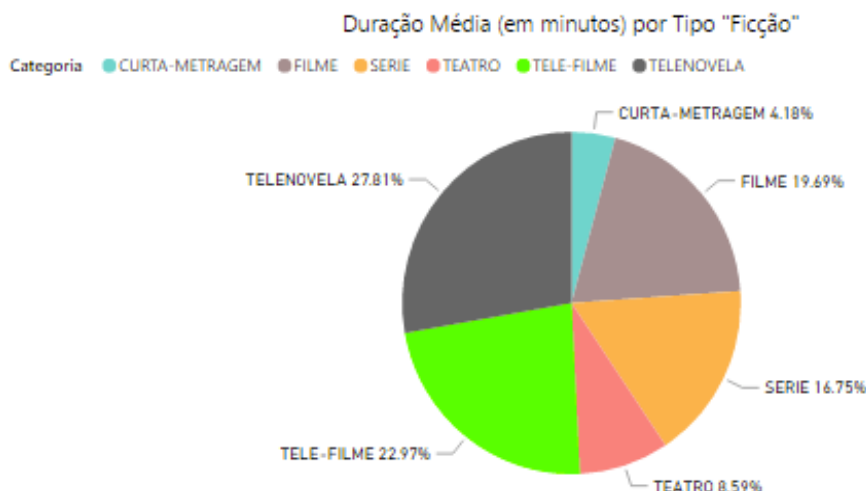


Gráfico 10 - Distribuição do tempo médio de visualização, em percentagem, das categorias de programas pertencentes ao tipo “ficção”

Ao contrário do que seria de esperar, ao selecionar (uma a uma) cada uma das várias faixas etárias, não se verifica uma grande discrepância entre as várias categorias de programas.

Por exemplo, a percentagem de jovens (dos 15 aos 24 anos) que ocupa o seu tempo a ver telenovelas é de 25,73%, enquanto que os idosos (a partir dos 64 anos) dedicam apenas mais 4 pontos percentuais, ou seja, 29,2%.

Conclusão semelhante se tira relativamente aos diferentes estatutos sociais: nas diferentes categorias de programas do tipo “ficção”, a distribuição temporal mantém-se aproximadamente idêntica.

A mesma análise foi feita para o tipo “variedades/divertimento”, e o respetivo gráfico pode ser observado em seguida.

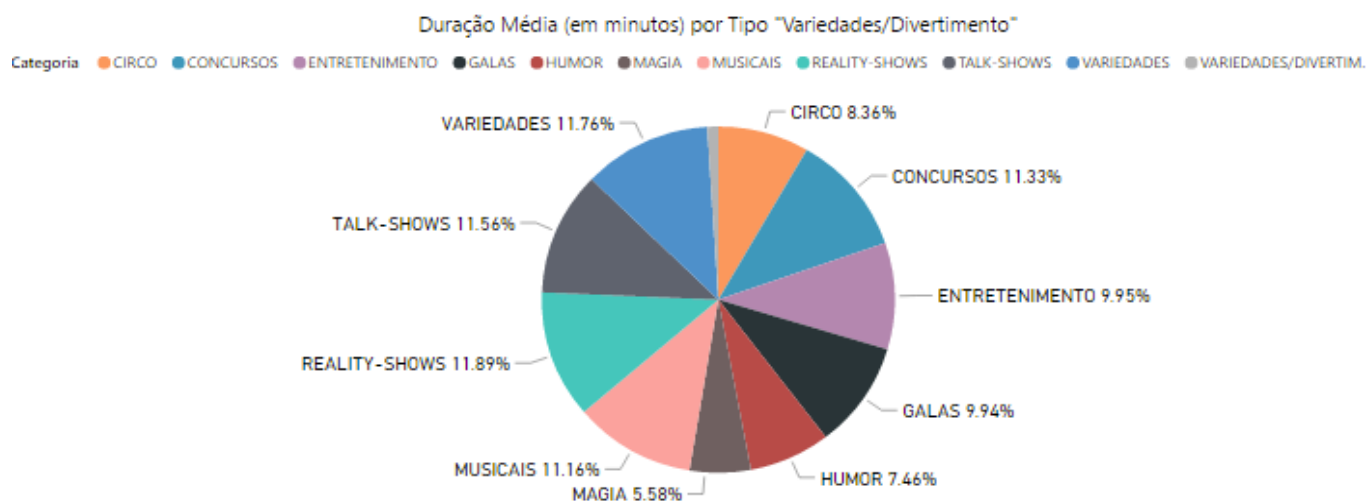


Gráfico 11 - Distribuição do tempo médio de visualização, em percentagem, das categorias de programas pertencentes ao tipo “variedades/divertimento”

Pelo gráfico 11 pode-se observar que, de uma forma geral, há uma distribuição equilibrada entre as diferentes categorias deste tipo. Ainda assim, verifica-se uma ligeira preferência por talk-shows, reality-shows, musicais, concursos e variedades.

Através do estudo das várias faixas etárias, uma conclusão que se pode retirar é que as faixas etárias dos 25 aos 34 anos e dos 35 aos 44 têm uma ligeira preferência pela categoria “magia”, relativamente a outras faixas etárias (nomeadamente a dos 4 aos 14 anos, como seria de esperar).

Por sua vez, e à semelhança do que tem sido feito até agora, seleccionando a faixa etária a partir dos 64 anos e comparando a classe média/alta com a classe média/baixa, verifica-se que há uma preferência da primeira por musicais e galas, enquanto que a segunda prefere concursos, humor e outras variedades.

Por fim, no gráfico 12 são visíveis as várias distribuições de categorias do tipo “cultura/conhecimento”.

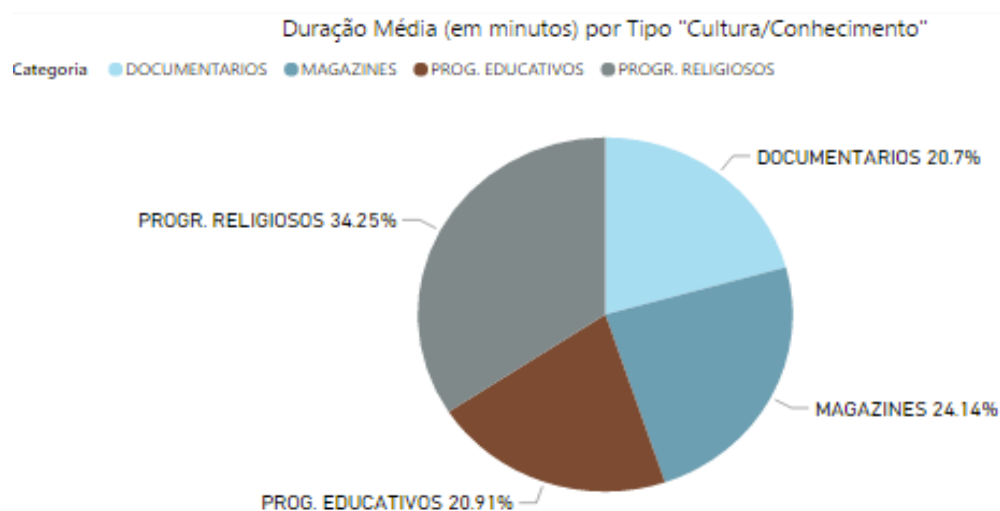


Gráfico 12 - Distribuição do tempo médio de visualização, em percentagem, das categorias de programas pertencentes ao tipo “cultura/conhecimento”

Neste gráfico é possível observar uma clara maioria de programas de carácter religioso, o que não é de admirar: embora Portugal seja considerado um país laico, é do conhecimento geral a importância dada pelos portugueses à religião.

Neste tipo de programas, e ao contrário dos casos anteriores (mas como é de esperar), a distribuição das várias categorias mostra-se mais discrepante consoante a faixa etária em estudo.

Um exemplo será o caso dos espetadores na faixa etária entre os 4 e os 14 anos, que ocupam 24,26% do seu tempo a visualizar programas religiosos, face a 38,47% relativamente a pessoas na faixa etária a partir dos 64 anos.

Pegando ainda na faixa etária entre os 4 e os 14 anos, outra conclusão pode ser tirada combinando este atributo com as classes sociais: as crianças da classe média/alta

passam, em média, 18,97% do seu tempo a ver programas religiosos, face a 26,07% de crianças da classe média/baixa.

Já os jovens adultos (entre os 25 e os 34 anos) de classe média/alta passam 26,59% do seu tempo a ver documentários e 27,78% a ver programas educativos, enquanto espetadores da mesma idade, mas de classe média/baixa dedicam, respetivamente, 26,61% e 24,76% aos mesmos programas.

6.2. Qual o número de espetadores por canal em cada dia, ao longo do semestre?

Para dar resposta a esta questão, foi criado o gráfico 13, de áreas empilhadas, que representa a contagem de espetadores e duração média, em minutos, por data e canal. No eixo dos XX encontra-se a data completa (dia, mês e ano) correspondente ao primeiro semestre do ano 1996 e no eixo dos YY o número de espetadores que viu um canal com uma certa duração por dia.

A esta representação gráfica foi adicionada, à parte, uma caixa de seleção para cada um dos atributos que se pretende explorar.

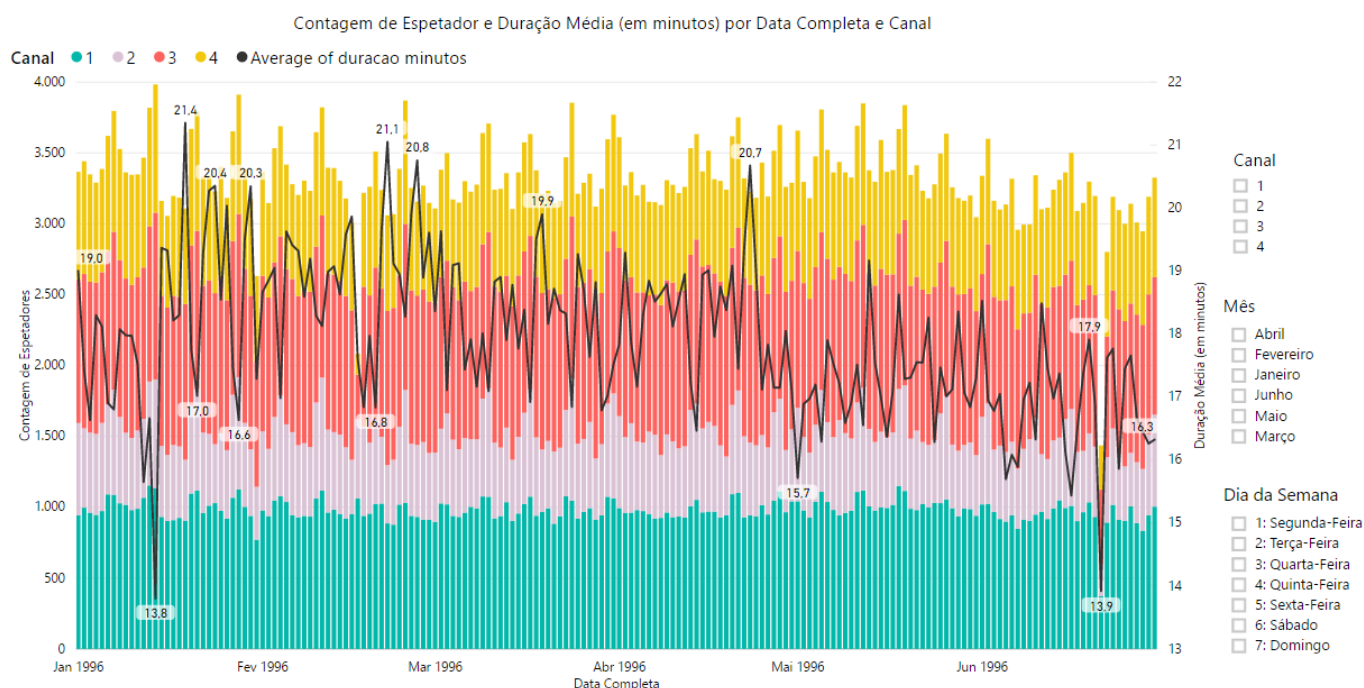


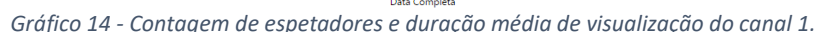
Gráfico 13 - Contagem de espetadores e duração média, em minutos, por data e canal

Analisando o gráfico 13, pode constatar-se que, no total, a soma das visualizações dos 4 canais pode variar entre os 1400 e os 4000 espetadores por dia. O dia 14 de Janeiro de 1996 (Domingo) foi o que obteve, somando as visualizações dos 4 canais, uma maior contagem de espetadores (com 3984 espetadores).

Para além disto, ainda se verificam três reduções significativas em relação à contagem de espetadores, nomeadamente no dia 31 de Janeiro (Véspera de Ano Novo) com 2632 espetadores, 17 de Fevereiro (Sábado) com 2082 espetadores e 21 de Junho (Sexta-Feira) com 1436 espetadores.

Tendo em conta a linha da duração média, em minutos, ocorrem várias oscilações ao longo do semestre. Podem ser destacadas duas reduções significativas: no dia 14 de Janeiro (correspondente a um domingo, com uma duração média de 13,8 minutos e com uma contagem de espetadores igual a 3984) e no dia 21 de Junho (correspondente a uma sexta-feira, com uma duração média de 13,92 minutos e com uma contagem de 1436 espetadores). Contrariamente, o dia 19 Janeiro representa o dia com uma duração

O dia 21 de Junho foi o dia com menos visualizações (372 espetadores) e com uma duração média de 11,8 minutos. Verificou-se que foi no dia 18 Maio que os espetadores passaram mais tempo, em média, a visualizar este canal (23,3 minutos). Estes detalhes podem ser visualizados no gráfico 14.



Os dias com durações médias mais baixas de visualizações (7,6 minutos por espetador) correspondem a 25 de Abril e 11 de Junho e com a duração mais elevada (14,9 minutos) corresponde a 4 de Março. Estas características podem ser visualizadas no gráfico 15.

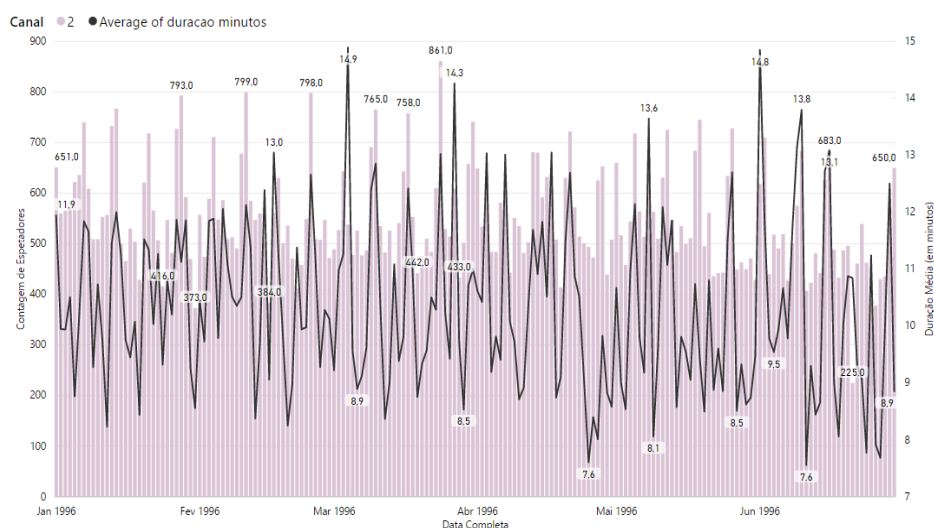
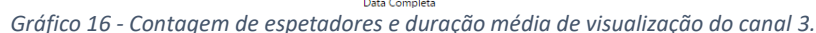
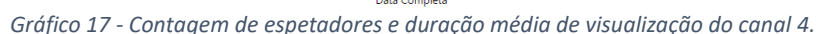


Gráfico 15 - Contagem de espectadores e duração média de visualização do canal 2.

Verificou-se que foi no dia 19 Janeiro que os espetadores passaram mais tempo, em média, a ver este canal (27,8 minutos). Estes pormenores podem ser visualizados no gráfico 16.



Verificou-se que foi no dia 13 Maio que os espetadores passaram mais tempo, em média, a visualizar este canal (18 minutos). Estes detalhes podem ser visualizados no gráfico 17.



6.3. Quais os tipos de programa mais vistos, ao longo das horas de um dia?

Para dar resposta a esta questão, foi produzido o gráfico 18, que mostra quais os tipos de programas mais vistos tendo em conta uma determinada hora. Cada barra representa uma hora do dia e a representação (em percentagem) dos vários tipos de programa durante essa hora (em média).

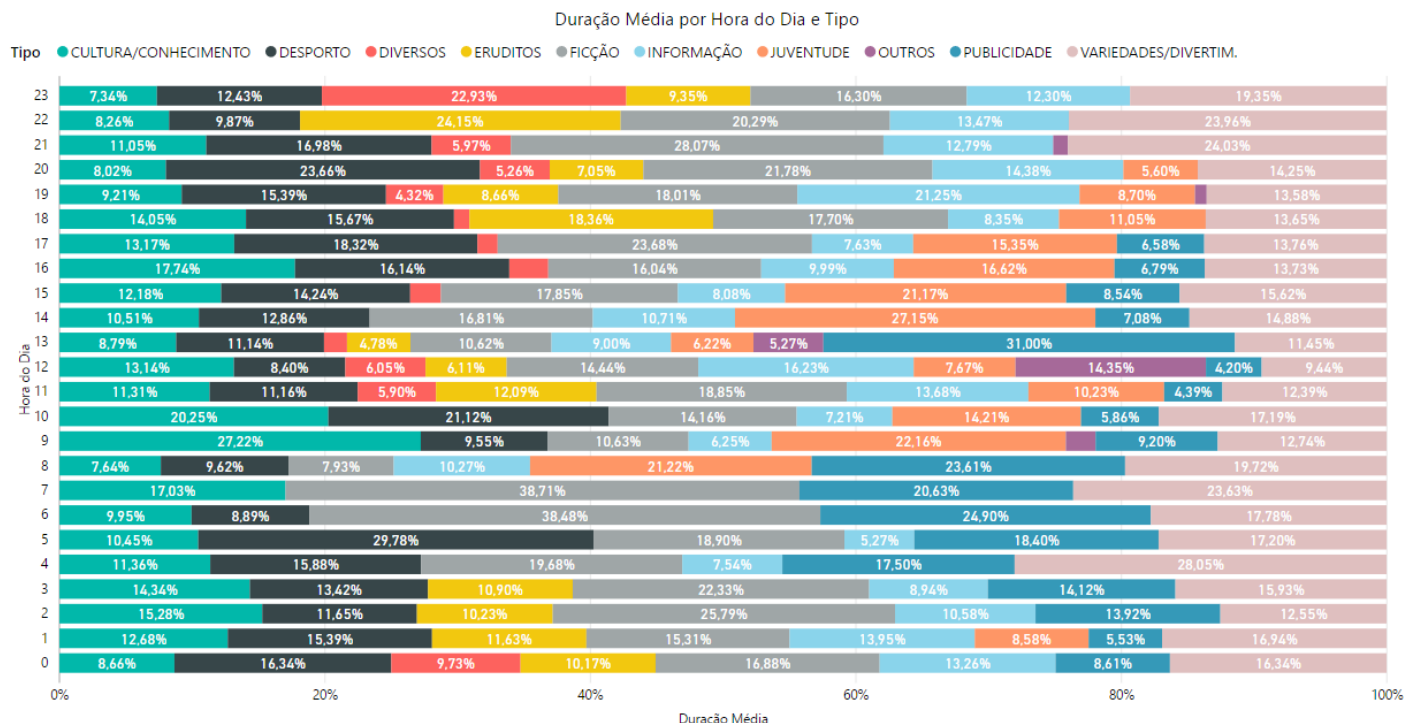


Gráfico 18 – Duração média dos tipos de programas ao longo das horas do dia.

Deste gráfico podem retirar-se diversas conclusões, tendo em conta as diferentes horas e períodos do dia:

- **00h00:** a esta hora os tipos de programas mais visualizados, com uma percentagem entre os 16-17%, são Desporto, Ficção e Variedades/Divertimento. Os menos visualizados correspondem a programas de carácter cultural e publicidade (9%). Não se encontram registadas visualizações de programas como Juventude e Outros;
- **01h00:** os tipos de programas mais visualizados, com uma percentagem entre os 15-16%, são Desporto, Ficção e Variedades/Divertimento. Os menos visualizados correspondem à Publicidade (6%). Não se encontram registadas visualizações de programas como Diversos e Outros;
- **02h00-07h00:** não se encontram registadas visualizações de programas como Diversos, Outros, Juventude e Eruditos, exceto às 02h00 e 03h00 que registam visualizações de programas Eruditos. No geral, programas como Desporto, Ficção, Publicidade e Variedades/Divertimento, são os mais vistos com percentagens entre os 20-40%;

- **Manhã (08h00-12h00):** programas relacionados com Cultura, Juventude, Desporto e Ficção são os mais optados pelos espetadores enquanto que programas do tipo Publicidade, Diversos e Eruditos são os menos vistos;
- **Tarde (13h00-18h00):** durante estas horas a maioria dos programas tem semelhantes percentagens de visualização exceto Diversos, Eruditos e Publicidade;
- **Noite (18h00-23h00):** os programas mais vistos durante este período correspondem a Desporto, Ficção e Variedades/Divertimento com percentagens entre os 18-30%.

Para além desta interpretação, podemos ainda analisar o gráfico por tipo de programa, verificando-se o seguinte:

- **Cultura/Conhecimento:** este tipo de programa tem maiores visualizações às 07h00 (17,03%), 09h00 (27,22%), 10h00 (20,25%) e 16h00 (17,14%);
- **Desporto:** este tipo de programa tem maiores visualizações às 05h00(29,78%), 10h00(23,66%) e 20h00(21,12%);
- **Diversos:** não existem registos desde a 01h00 até as 10h00 de visualizações deste programa. Curiosamente, a maior percentagem de visualização corresponde às 23h, com 22,93%;
- **Eruditos:** não existem registos desde as 04h00-10h00 e das 14h00-17h00. A maior percentagem de visualização corresponde às 18h com 18,36%;
- **Ficção:** este tipo de programa tem maior percentagem de visualizações às 06h00 (38,98%), 07h00 (38,71%) e 21h00 (28,07%). As mais baixas percentagens registam-se às 08h00, 09h00 e 13h00;
- **Informação:** não existem registos de visualizações entre as 06h00 e as 07h00. A maior percentagem de visualização corresponde às 19h, com 21,25%. As percentagens de visualizações das restantes horas encontram-se entre 5-14%;
- **Juventude:** não existem registos entre as 02h00 e as 07h00, nem entre as 21h00-00h00. As maiores percentagens de visualização correspondem às 9h, com 22,16%, às 14h, com 27,15%, e às 15h, com 22,16%;
- **Outros:** em relação a este tipo de programas, só existem registos às 09h00, 12h00, 13h00, 19h00 e 21h00, sendo às 12h que se observa a maior percentagem de visualização (14,35%);
- **Publicidade:** este tipo de programa tem maiores visualizações às 06h00 (24,90%), 08h00 (23,61%) e 13h00 (31%). Não existem registos desde as 18h00 até às 23h00;

- **Variedades/Divertimento:** todas as horas apresentam uma percentagem de visualizações que varia deste os 11% até aos 24%, exceto as 04h00, que registam uma percentagem de 28,05%.

Olhando para o panorama geral, a maior parte dos tipos de programas apresentam a mesma proporção em termos de visualizações, exceto os Diversos e Outros.

Parece também não existir uma correlação evidente entre a preferência de visualização de determinado tipo de programa numa hora específica, exceto para o tipo Ficção que mostra as maiores percentagens registadas às 06h00 e 07h00.

6.4. Por cada região do país, quem passa mais e menos tempo a ver televisão, tendo em conta o estatuto social e se trabalha ou não em casa?

Para dar resposta a esta questão, foi elaborado um gráfico de colunas agrupadas. Cada coluna representa a duração de visualização de um programa (em minutos); cada conjunto de colunas agrupa as diferentes regiões do país, de acordo com o perfil do espetador (se trabalha ou não em casa) e o seu estatuto social.

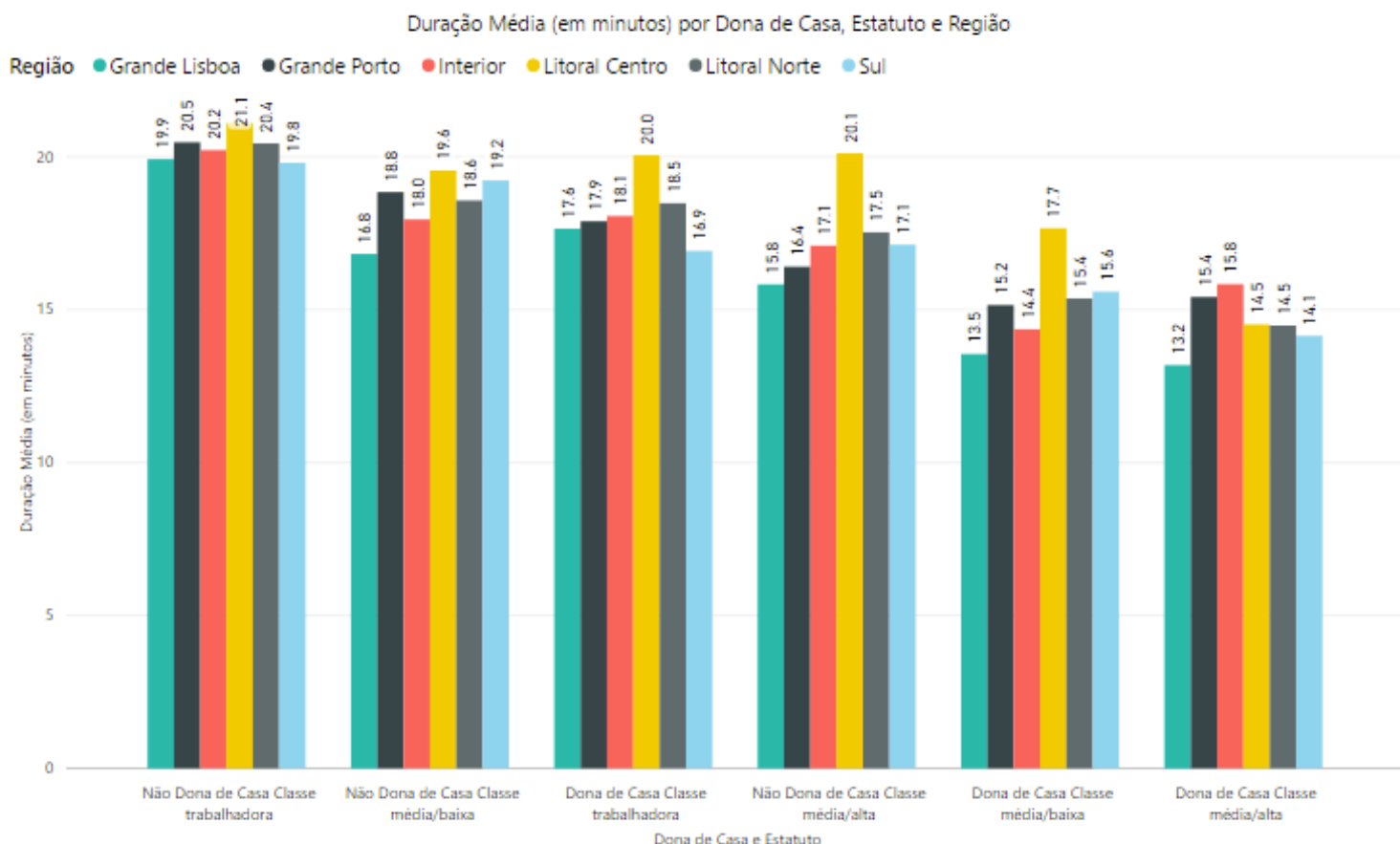


Gráfico 19 - Média de duração (em minutos) por região, dona de casa e estatuto social

Para facilitar a retirada de informação, os conjuntos de colunas foram agrupados de acordo com a duração de visualização: à esquerda o conjunto com maior média de duração de visualização, à direita o grupo que possui menor média de duração de visualização.

Desta forma, é fácil concluir que o grupo com maior média de duração de visualização de um programa é o grupo de espetadores com perfil “não dona de casa” da classe trabalhadora, onde a duração média ronda os 20 minutos, enquanto que o conjunto com menor média de duração de visualização se refere ao perfil de dona de casa da classe média/alta, com uma duração média de 14 minutos.

Outra conclusão atingida rapidamente é que os espetadores da região litoral centro são os que passam, em média, mais tempo a ver um programa, independentemente do seu perfil e estatuto social.

Por outro lado, a região do país que passa, em média, menos tempo a ver um programa, é a região da grande Lisboa.

Ao restringir a análise a apenas uma região, por exemplo a região Sul, obtivemos o gráfico 20.

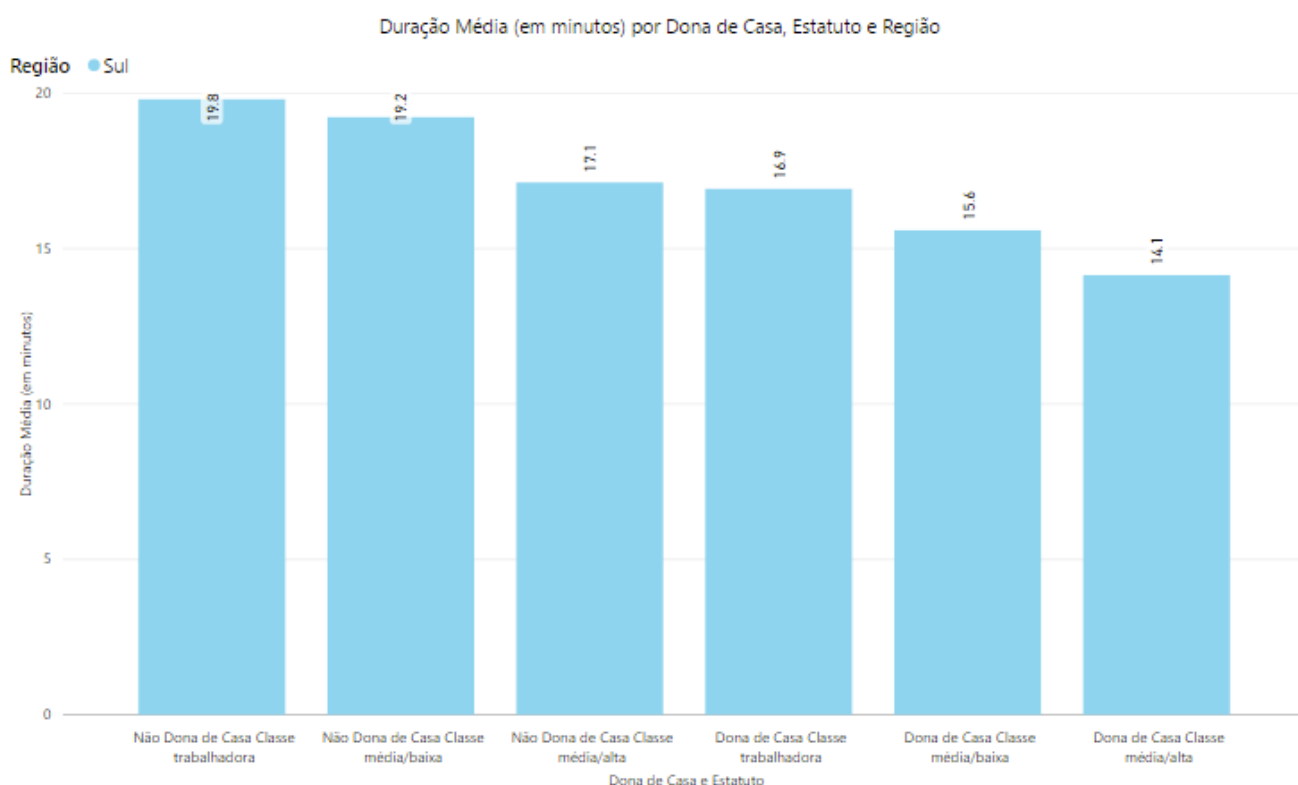


Gráfico 20 - Duração média, em minutos, relativa à região sul e aos diferentes perfis e estatutos sociais

Através deste gráfico pode observar-se que o subconjunto de espetadores que passa mais tempo, em média, a ver um programa, é o subconjunto com o perfil de não-dona de casa da classe trabalhadora, com 19,8 minutos de duração média.

Por sua vez, a menor duração (14,1 minutos) corresponde ao subconjunto das donas de casa da classe média/alta, seguindo a tendência geral.

Olhando apenas para os diferentes estatutos, pode ainda verificar-se que, de uma forma geral, os espetadores de perfil “não-dona de casa” dedicam mais tempo a cada programa de televisão.

Através da interação com os diferentes atributos, confirma-se que a tendência geral é para os espetadores de classe trabalhadora passarem a maior quantidade de tempo a ver um programa, seguidos da classe média/baixa, e por fim da classe média/alta.

6.5. Tendo em conta os feriados nacionais/datas comemorativas no primeiro semestre de 1996, qual a média de espetadores por faixa etária e estatuto social?

Para dar resposta a esta questão, uma série de gráficos foram traçados. É importante realçar que, para esta análise, foi restringido o conjunto de dias a observar: mantiveram-se apenas os dias considerados como “datas comemorativas”, sejam eles considerados feriados ou não.

O primeiro gráfico desenhado foi um gráfico de barras, em que cada barra corresponde a um destes dias, e a sua altura corresponde à contagem de espetadores com registos televisivos para o dia em questão.

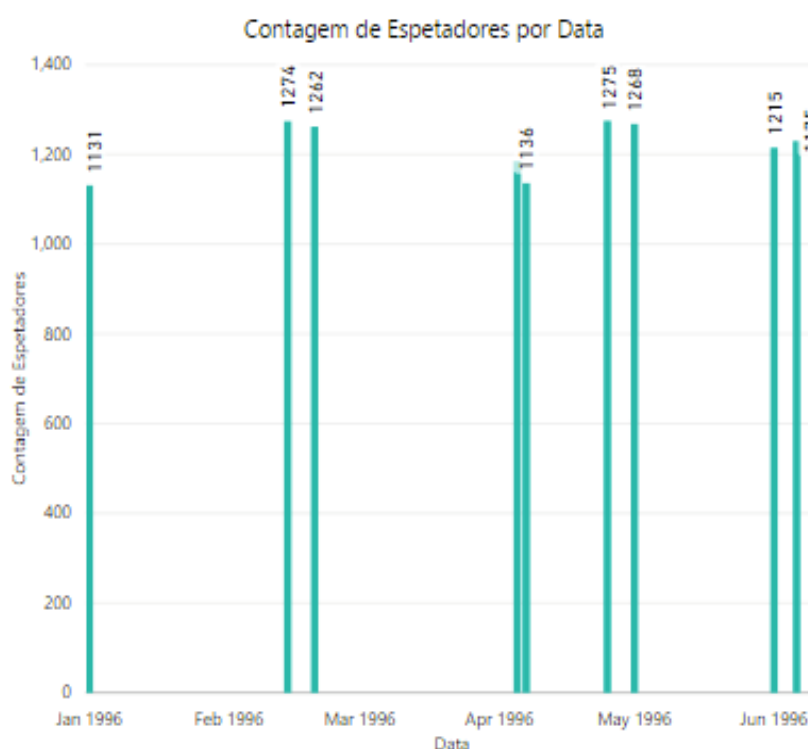


Gráfico 21 - Quantidade de espetadores que vê televisão em feriados e datas comemorativas.

Através do gráfico 21 observa-se que os dias com uma menor contagem de espetadores são o dia 1 de Janeiro, que sucede as comemorações de Ano Novo, e os dias 5 e 7 de Abril, Sexta-feira Santa e Domingo de Páscoa, respetivamente.

Por serem datas que em Portugal são amplamente celebradas, justifica-se desta forma a discrepância na contagem de espetadores, relativamente aos restantes dias.

Embora este gráfico possa não conter muito mais informação, é importante para realçar quais os dias em análise nos gráficos que se seguem.

Para prosseguir a análise, traçou-se um novo gráfico de colunas agrupadas. Desta vez, cada conjunto de colunas corresponde às diferentes datas comemorativas, isto é, se são feriado ou não feriado. Dentro de cada conjunto, cada coluna corresponde a um dos três estatutos diferentes.

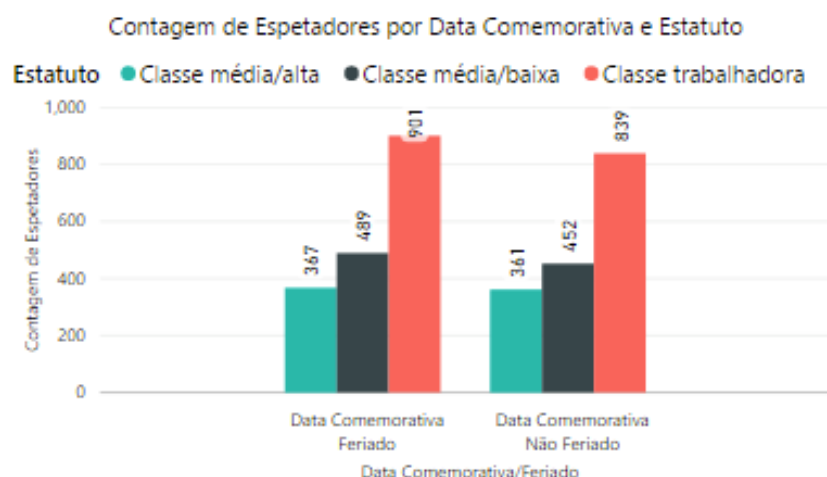


Gráfico 22 - Quantidade de espectadores, por classe social, que vê televisão em feriados e datas comemorativas.

Através do gráfico 22 rapidamente se conclui que nas datas comemorativas (sejam elas feriado ou não), o maior número de espectadores pertence à classe trabalhadora. Por sua vez, a classe média/alta é sempre a que registra um menor número de espectadores. Esta tendência mantém-se, qualquer que seja o mês (e, portanto, as datas comemorativas) em estudo.

Análise semelhante foi efetuada para relativamente aos escalões etários, produzindo o gráfico 23. Nele, cada uma das barras do conjunto corresponde a um diferente escalão etário.

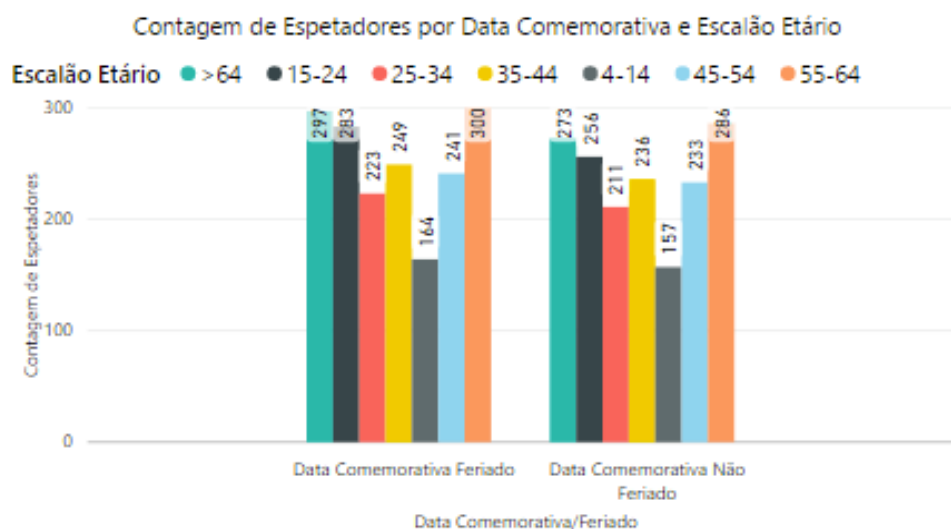


Gráfico 23 - Quantidade de espectadores, por escalão etário, que vê televisão em feriados e datas comemorativas.

É visível que, das sete faixas etárias em análise, aquela que regista em ambos os tipos de diferentes datas comemorativas um menor número de espectadores é a que se refere a espectadores entre os 4 e os 14 anos. Esta faixa etária regista um pico especialmente baixo no feriado de dia 1 de Janeiro, com apenas 80 espectadores.

De uma forma geral, isto é, ao longo das várias datas comemorativas, é o conjunto de faixas etárias que vai dos 4 aos 14 e dos 25 aos 44 anos que regista o menor número

de espetadores. No entanto, esta tendência inverte-se se especificarmos a qual das classes sociais pertence o público em análise.

Ao seleccionar espetadores considerados da classe média/alta, verifica-se que uma clara maioria é pertencente à faixa etária dos 35 aos 44 anos, enquanto que a faixa etária a partir dos 64 anos (que na análise anterior era uma das que registava maior número de espetadores) regista o menor número de espetadores, como se pode verificar no gráfico 24.

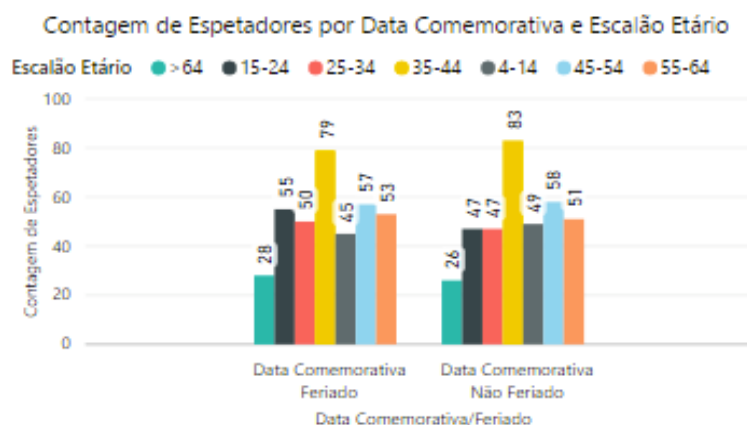


Gráfico 24 - Contagem de espetadores por escalão etário, referentes à classe média/alta

Embora este gráfico reflita uma contagem referente a todas as datas comemorativas, esta tendência verifica-se também se se optar por analisar cada data comemorativa em separado.

Já para a classe média/baixa, o maior número de espetadores continua a pertencer à faixa etária dos 35 aos 44 anos, mas é também acompanhada pela faixa etária dos 15 aos 24 anos. A faixa etária a partir dos 64 anos continua a ser a que regista menos público.

No entanto, é relativamente à classe trabalhadora que se verifica um grande número de espetadores nas faixas etárias a partir dos 55 anos, como se pode ver no gráfico 25.

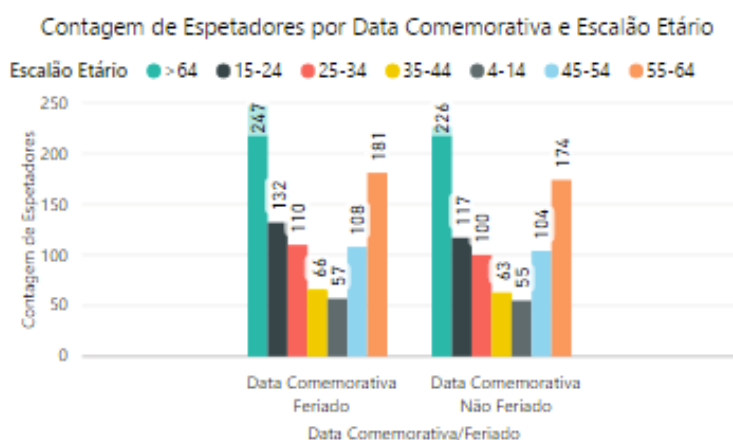


Gráfico 25 - Contagem de espetadores por escalão etário, referentes à classe trabalhadora

7. Prospeção de Informação

A complexidade de um negócio impõe às empresas e organizações a necessidade de aumentar a eficiência na obtenção do conhecimento, processamento, análise e interpretação de informações devido às grandes quantidades de dados envolvidas.

A prospeção de informação tem como objetivo a descoberta de padrões em grandes quantidades de dados, úteis para os processos de negócio em estudo.

No contexto do *Data Warehouse*, os sistemas de *data mining*, tarefa integrada no processo de tomada de decisão, são considerados a “Inteligência”. Isto porque, exploram dados históricos para gerar informação útil no futuro, detetam padrões e sugerem novas regras de negócio e permitem tomar decisões informadas e traçar planos de ação.

Aplicar métodos de *data mining* aos nossos dados normalmente implica um conjunto de tarefas representadas na figura 9.

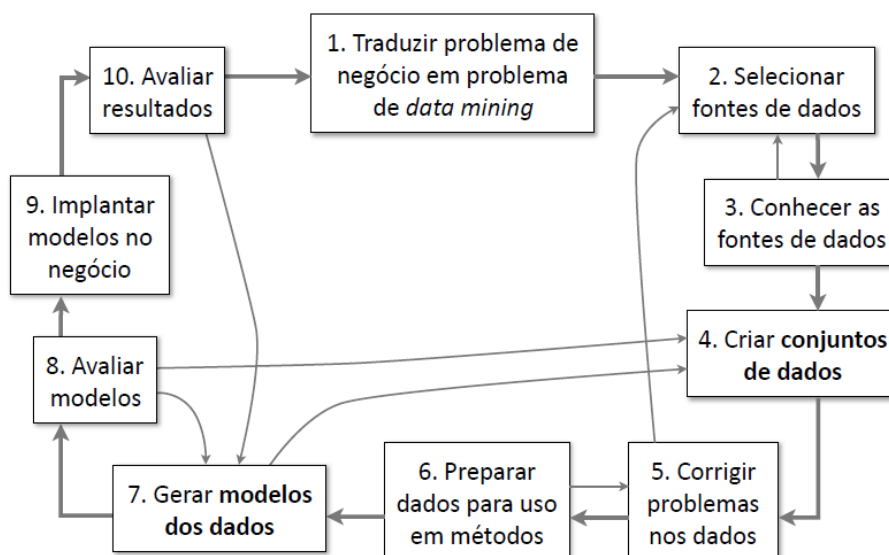


Figura 9 - Ciclo virtuoso de data mining

O esquema 5 pode ser resumido em quatro passos principais:

1. **Identificar um problema** do negócio, com base na análise dos dados no data warehouse;
2. Aplicar **métodos de data mining** para transformar dados e construção e modelos de dados;
3. Agir sobre a informação obtida – **tomada de decisões** sobre o funcionamento do negócio;
4. Medir resultados de decisão – **avaliar** se o problema no negócio foi resolvido.

Neste caso não se fará a avaliação dos modelos com intuito de gerar modelos mais otimizados, mas sim apenas uma prova de conceito de prospeção de informação neste contexto.

O processo de data mining envolve seis tarefas principais: classificação, estimação, predição, associação, agrupamento e interpretação.

A **classificação** é a tarefa mais utilizada e consiste em examinar as características de um objeto e atribuí-lo a um conjunto predefinido de classes, ou seja, gerar modelos que organizam dados em classes pré-determinadas.

Este tipo de tarefa, associada ao método das **Árvores de Decisão**, será o utilizado para encontrar padrões nos dados de audiências televisivas, mais especificamente num subconjunto de dados.

7.1. Método de Classificação: Árvores de Decisão

Árvores de Decisão são métodos supervisionados usadas para dividir um grande conjunto de dados heterogêneos em conjuntos sucessivamente mais pequenos de dados homogêneos, aplicando uma sequência de simples regras de decisão.

Em relação ao processo típico deste método, podemos dividi-lo em cinco fases:

1. Selecionar fontes de dados **pré-classificados**;
2. Preparar dados para análise;
3. Gerar modelo dos dados;
4. Avaliar modelo **comparando com dados pré-classificados**;
5. Aplicar modelo a novos dados.

Dos dados do projeto de “Audiências Televisivas”, foi escolhido um subconjunto de dados para responder a uma necessidade de classificação específica, que será descrito na secção seguinte.

7.2. Obtenção dos dados

Dado o contexto do nosso processo de negócio seria interessante desenvolver um classificador que baseado num conjunto de características do espectador prevê-se qual a sua categoria de programas favorita.

Esta preferência é inferida através da categoria de programas que o espetador despende mais tempo a ver, ou seja, a categoria que totaliza uma maior soma de durações de programas observados.

Os dados foram obtidos diretamente do SQL Server utilizando *queries* SQL devido à sua performance (e rapidez) ser mais alta que as restantes alternativas. Os comandos executados estão disponíveis na secção anexos, mais concretamente na secção 10.4.

Foram limitados os resultados à primeira semana do ano (entre 1996-01-01 e 1996-01-07), pois trata-se apenas de uma prova de conceito, e não era desejável que os métodos de aprendizagem demorassem a treinar.

Sumariamente, foram obtidos os registos de visualização de programas, nas data referidas apenas referente a 3 categorias: Telenovela ("TELENOVELA"), Programas Infanto-Juvenil ("PROG. INF. JUVENIL") e Noticiário ("NOTICIÁRIO"). Foram escolhidas

estas 3 para limitar o número de categorias e porque se previu que iriam ter públicos-alvo distintos.

De seguida, foi feito um cruzamento dos factos de visualização de interesse com alguma informação do programa (canal, e categoria do programa visto), bem como com a maioria da informação dos espectadores (género, escalão etário, região, estatuto 1, ocupação 1, estatuto 2, ocupação 2 e se é dona de casa), e guardou-se toda esta informação, juntamente com o tempo de visualização, numa tabela única.

Para cada espectador foi calculado qual o seu programa favorito (que mais tempo passa a ver), e apenas foram mantidos registos de espectadores cujo programa favorito se enquadrava numa das 3 categorias supra-citadas.

Estes dados foram transferidos do SQL Server para um ficheiro separado por tabulações (.tsv) pela ferramenta "Import and Export Data" para poder ser usado pelo R Studio.

Na ferramenta R Studio não tiveram de ser efetuadas nenhuma alteração aos dados, pois os valores foram interpretados como fatores, e as árvores de decisão conseguem lidar com fatores para fazer classificação.

Decidiu-se usar o pacote "rpart", pois este permite gerar gráficos das árvores mais informativos (esses gráficos não são mostrados no relatório).

7.3. Resultados

Nesta seção são apresentados os resultados obtidos através do script "*data_mining.R*", no qual se aplicou o método de aprendizagem automática escolhido (árvores de decisão) sobre o conjunto de dados.

À partida, todos os resultados obtidos devem ser reproduzíveis, visto que o uso de números aleatórios foi controlado pela semente fornecida aos geradores de números pseudo-aleatórios. Teoricamente o único resultado que pode variar é o *sampling* diferencial do conjunto de dados (será explicado adiante).

Inicialmente foi construída uma árvore de decisão com todos os dados. Apesar desta abordagem não ser correta (pois não permite avaliação), tal permite visualizar os dados e perceber se o modelo será adequado.

A árvore obtida é apresentada na figura 10, juntamente com a tabela 22, que descreve a comparação entre o número real de espectadores que mais viam uma dada categoria com número de espectadores que o modelo previu para cada categoria.

Prever a categoria de programa mais visto por um espectador (Todos os dados)

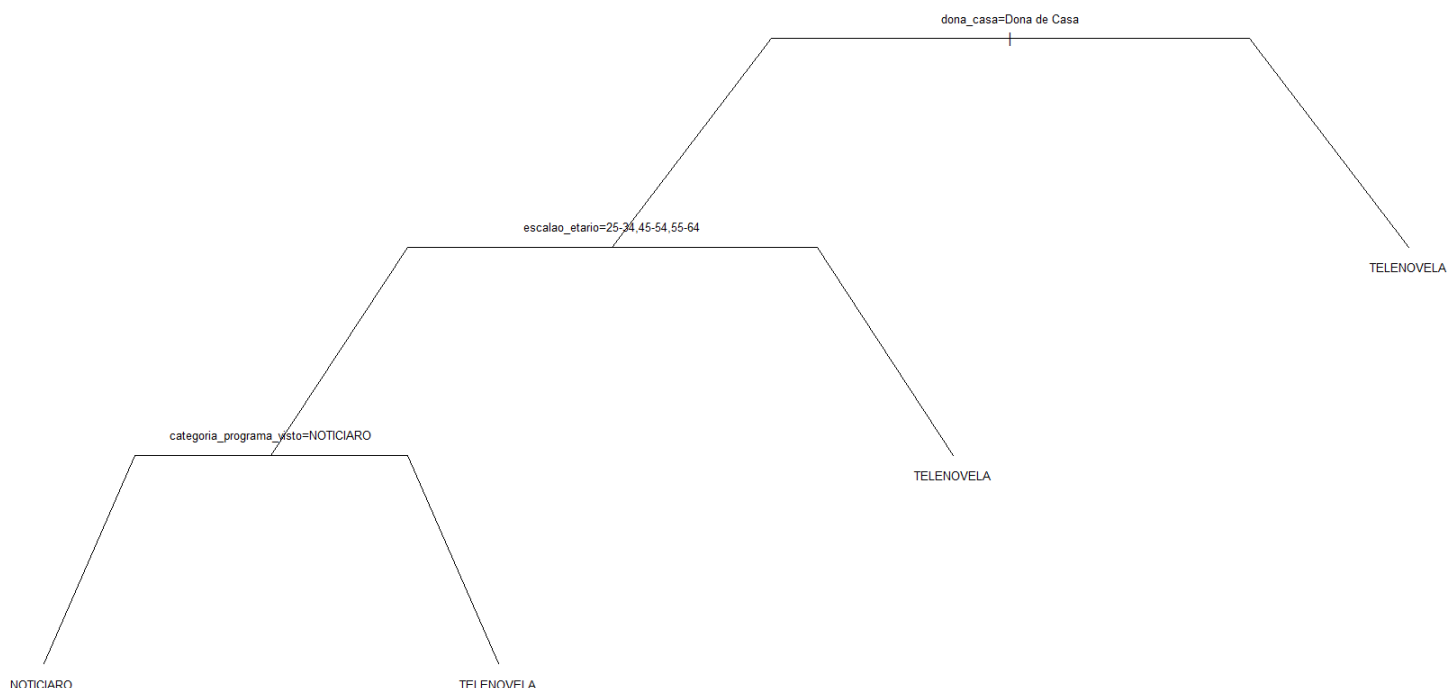


Figura 10 - Árvore de decisão construída com todos os dados em análise

Tabela 22 - Matriz de confusão comparando o valor real da classificação e a classificação prevista pelo modelo, usando todos os dados em análise

		Previsão			
Real		NOTICIÁRIO	PROG. INF. JUVENIL	TELENOVELA	Soma
	NOTICIÁRIO	2279	0	7636	9915
	PROG. INF. JUVENIL	54	0	1408	1462
	TELENOVELA	1301	0	26024	27325

Através da tabela 22, pode verificar-se que a árvore de decisão nunca classifica como “PROG. INF. JUVENIL”.

De seguida procedeu-se à partição do conjunto de dados em dados de teste e dados para treino do modelo, e foi criado um novo modelo partindo destes dados. O novo modelo, apresentado na figura 11 (na página seguinte), permitiu tirar conclusões semelhantes.

A tabela seguinte apresenta a comparação entre a classificação real presente no conjunto de teste, e a previsão do modelo para esses mesmos dados. Mais uma vez, não há previsão de classificação como “Prog. Inf. Juvenil”.

Tabela 23 - Matriz de confusão comparando o valor real da classificação e a classificação prevista pelo modelo, usando os dados de teste

		Previsão			
Re		NOTICIÁRIO	PROG. INF. JUVENIL	TELENOVELA	Soma

NOTICIÁRIO	647	0	2684	3331
PROG. INF. JUVENIL	4	0	480	484
TELENOVELA	380	0	8706	9086

Prever a categoria de programa mais visto por um espectador (Dados de Treino)

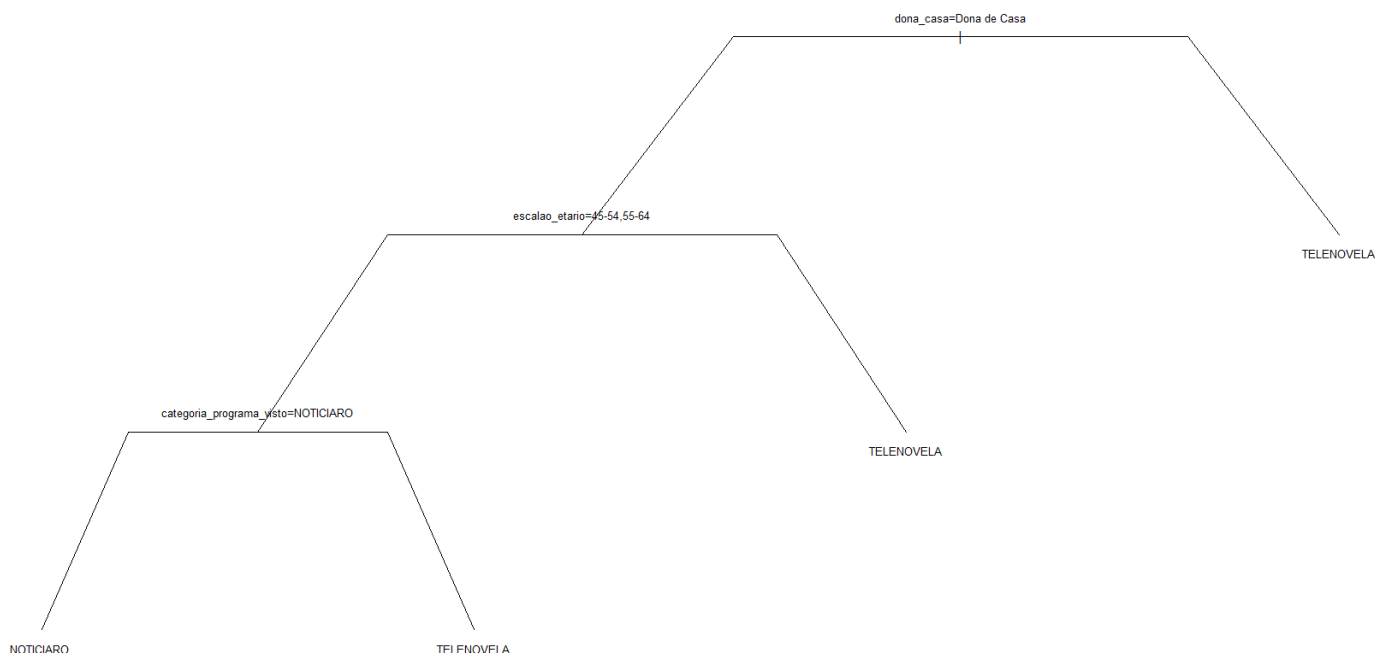


Figura 11 - Árvore de decisão construída com os dados de treino.

A qualidade do modelo foi avaliada com recurso ao *F score* dos dados de teste. Com os dados da tabela anterior obtemos os resultados da tabela 24.

Pode concluir-se que o modelo apresenta bom desempenho para a categoria “Telenovela” (*F score* elevado), no entanto não consegue prever a categoria “Prog. Inf. Juvenil”.

A visão geral do modelo é obtida fazendo a média de cada uma das métricas (que não é possível obter neste caso, pois não existem dados para a última categoria citada).

Tabela 24 - Valores de precisão, rechamada e *F score*, calculados para avaliar o modelo usando os dados de teste

Categoria	Precisão	Rechamada	F1 Score
NOTICIÁRIO	0.628	0.194	0.297
PROG. INF. JUVENIL	NaN	0	NaN
TELENOVELA	0.733	0.958	0.831
Média	NaN	0.384	NaN

O facto de uma das classes de classificação ser ignorada pelo modelo é um acontecimento corrente em aprendizagem automática. Tal deve-se ao enviesamento dos conjuntos de dados para uma das classes, em que existe maior abundância de dados de uma classe do que das restantes.

Assim, a estratégia adotada pelo modelo consiste em classificar com a classe mais abundante, pois desta forma terá uma taxa de erro muito baixa. No entanto, isso não é útil (não conseguirmos prever uma dada classe), especialmente se estivermos especialmente interessados nessa classe.

Uma solução possível para este problema é criar um conjunto de dados equilibrado a partir dos dados desequilibrados, através de técnicas como *oversampling* das categoriais mais raras e *undersampling* das categorias mais comuns.

Desta forma, este problema foi resolvido com ferramentas disponíveis para a ferramenta R Studio, em que se passou de um conjunto de dados desequilibrado, para um conjunto de dados mais equilibrado, com as proporções de cada classe representadas na tabela 25.

Tabela 25 - Comparação da proporção de valores de cada categoria nos dados de treino iniciais e no resultado do *sampling diferencial*

	NOTICIÁRIO	PROG. INF. JUVENIL	TELENOVELA	Total
Proporção Inicial	0.255	0.038	0.707	1
Proporção Final	0.308	0.346	0.346	1

Usando esta reamostragem dos dados, foi criado um novo modelo, representado na figura 12, que foi também avaliado para poder ser comparado com o anterior.

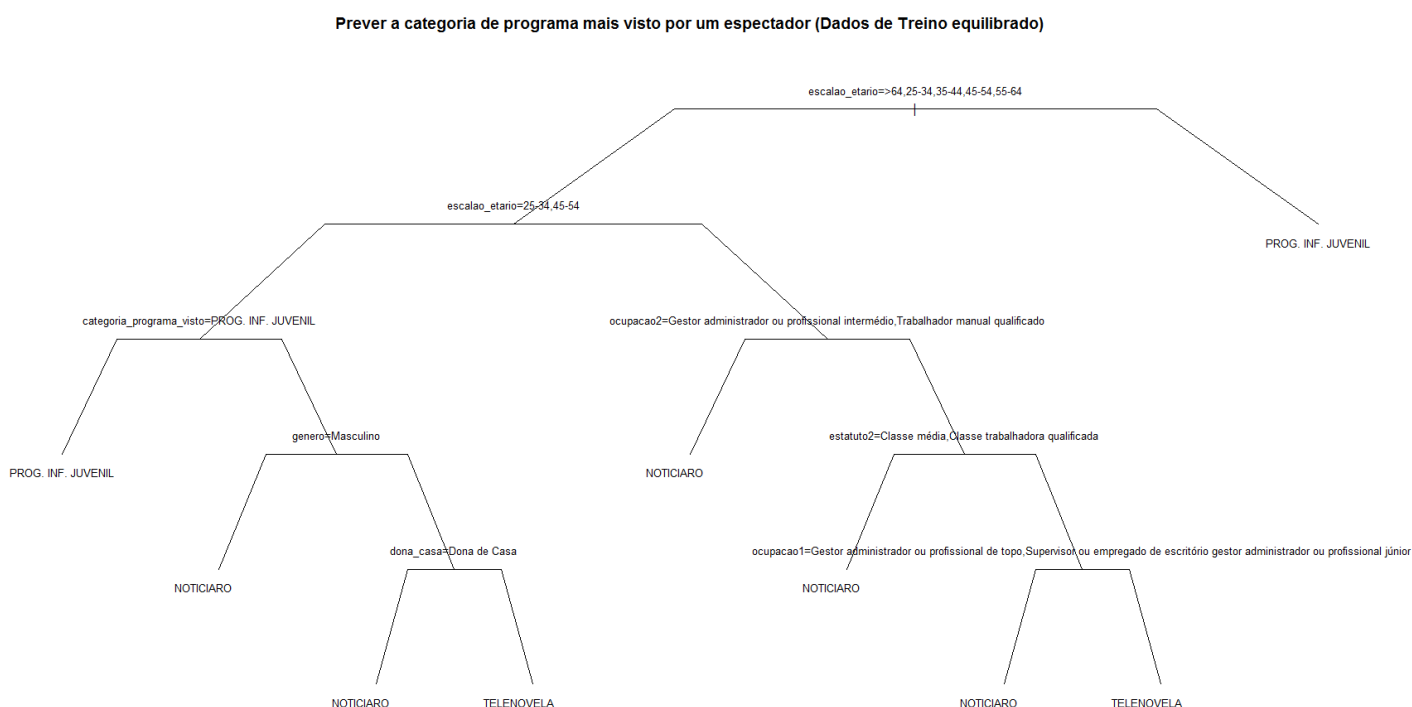


Figura 12 - Árvore de decisão construída usando dados de treino resultantes do *sampling diferencial*

Com estes novos dados já é possível prever com sucesso a maioria dos casos de “Prog. Inf. Juvenil”, no entanto o modelo não é perfeito, e esta reamostragem teve como consequência a pior performance do modelo para a categoria “Telenovela”.

Visto que o objetivo deste trabalho não era obter um classificador com elevado desempenho, mas sim apreender a aplicar métodos de aprendizagem automática, neste contexto, acredita-se que os resultados obtidos são satisfatórios.

Nas seguintes tabelas são apresentadas a comparação entre o número real de espectadores de cada categoria com o número previsto, bem como a avaliação do modelo através do *F score*.

Neste caso, já foi possível obter um *F score* global do modelo (visto que foi possível obter o valor para “Prog. Inf. Juvenil”), no entanto observa-se uma diminuição do desempenho na classificação de “Telenovela”.

Tabela 26 - Matriz de confusão comparando o valor real da classificação e a classificação prevista pelo modelo (gerado através de dados com *sampling diferencial*), para os dados de teste

		Previsão			
Real		NOTICIÁRIO	PROG. INF. JUVENIL	TELENOVELA	Soma
	NOTICIÁRIO	1486	349	1496	3331
	PROG. INF. JUVENIL	27	445	12	484
	TELENOVELA	2342	1765	4979	9086

Tabela 27 - Valores de precisão, chamada e *F score*, calculados para avaliar o modelo (gerado através de dados com *sampling diferencial*), para os dados de teste

Categoria	Precisão	Rechamada	F1 Score
NOTICIÁRIO	0.385	0.446	0.414
PROG. INF. JUVENIL	0.174	0.919	0.292
TELENOVELA	0.768	0.548	0.639
Média	0.442	0.638	0.448

Observando a figura 12, verifica-se que o modelo tem demasiados ramos, pelo que pode ser mais complexo do que o necessário. Para isso, analisou-se o ganho obtido no erro por cada separação que é efetuada nos dados.

Tabela 28 - Separação dos dados na árvore de decisão (gerada através de dados com *sampling diferencial*) e respetivo erro associado a cada divisão

nsplit	CP	rel error	xerror	xstd
0	0.297	1	1.014	0.006
1	0.057	0.703	0.714	0.007
3	0.039	0.589	0.592	0.006
4	0.031	0.550	0.552	0.006
5	0.020	0.519	0.521	0.006
6	0.012	0.498	0.500	0.006
8	0.010	0.475	0.477	0.006

Pode-se verificar que a partir da 5ª divisão a diminuição do erro deixa de ser muito significativa. Assim, seria possível reduzir a complexidade da árvore criando outro

O resultado da poda pode ser observado na figura 13, na página seguinte, que apresenta o modelo simplificado.



Tabela 29 - Separação dos dados na árvore de decisão após poda do modelo construído com sampling diferencial. Valores até ao limite definido na poda (cp = 0.02)

É importante também perceber se esta diminuição da complexidade teve algum impacto relevante no desempenho do modelo. Usando as mesmas métricas que anteriormente verifica-se que o impacto é negligenciável, com uma ligeira diminuição do *score* geral, sendo a simplificação mais evidente no desempenho a classificar “Telenovela”.

Tabela 30 - Matriz de confusão comparando o valor real da classificação e a classificação prevista pelo modelo podado, para os dados de teste

		Previsão			
Real		NOTICIÁRIO	PROG. INF. JUVENIL	TELENOVELA	Soma
	NOTICIÁRIO	1894	349	1088	3331
	PROG. INF. JUVENIL	32	445	7	484
	TELENOVELA	3349	1765	3972	9086

Tabela 31 - Valores de precisão, chamada e F score, calculados para avaliar o modelo podado para os dados de teste

Categoria	Precisão	Rechamada	F1 Score
NOTICIÁRIO	0.359	0.569	0.440
PROG. INF. JUVENIL	0.174	0.919	0.292
TELENOVELA	0.784	0.437	0.561
Média	0.434	0.642	0.431

Analisando o último modelo obtido, pode concluir-se que o mesmo faz sentido. Primeiro, este verifica o escalão etário do espectador; se este for jovem (“4-14” ou “15-24”) classifica como “Prog. Inf. Juvenil”, caso contrário volta a dividir os restantes escalões etários.

Num dos casos analisa a categoria do último programa que o espectador viu, e usa essa informação para prever o tipo mais visto como “Prog. Inf. Juvenil” ou “Noticiário”. No outro caso, baseia-se na informação da ocupação 2 (que está relacionada com o estatuto 2, pelo que o modelo ainda podia ser mais simplificado, removendo a separação por este último atributo) e caso esta informação esteja definida (“Gestor administrador ou profissional intermédio” ou “Trabalhador manual qualificado”) prevê que a classificação seja “Noticiário”, caso contrário verifica o valor do atributo estatuto 2, e caso seja negativo classifica como “Telenovela”.

As regras definidas por este modelo fazem sentido com o que uma pessoa podia pensar empiricamente, pelo que apesar do desempenho do modelo não ser muito alto, este parece fazer sentido e estar de acordo com os dados.

Fazendo o *summary* dos modelos verifica-se que os atributos que estes consideram mais relevantes são o escalão etário, categoria do último programa visto, bem como estatuto e ocupação.

No entanto talvez esses não sejam os atributos mais relevantes, pois outros métodos mais avançados (*random forest*) consideram que outros atributos como escalão etário, região, duração do último programa visto e se é dona de casa ou não, são mais relevantes.

8. Conclusão

Este trabalho teve como foco analisar as tendências televisivas dos espectadores, sendo esse o nosso processo de negócio prioritário. Assim sendo, o intuito deste projeto é a construção de um *data warehouse*, com o objetivo de auxiliar a tomada de decisão no contexto da análise das audiências televisivas.

Foram analisados os dados 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*.

Depois de estabelecido que o mais adequado seria uma tabela de factos do tipo transacional, e que o grão seria “um espectador que vê um dado programa, numa dada data e hora, durante um determinado período de tempo”, passou-se à construção do *data warehouse*.

Numa terceira fase do projeto, focámo-nos em construir (e demonstrar) o *data warehouse*, mais especificamente todo o processo de extração, transformação e carregamento de dados. E posteriormente, obter o cubo de dados.

Na última etapa deste projeto, as respostas às perguntas analíticas colocados foram respondidas com recurso à ferramenta Microsoft Power BI. Para além disto, foi aplicado um método de *data mining*, mais propriamente o método de árvores de decisão, para encontrar padrões úteis entre um subconjunto de dados, permitindo prever o programa favorito de um espectador com base nas características desse espectador e do tipo do ultimo programa visualizado.

Ao longo deste projeto foram encontradas algumas dificuldades, nomeadamente no que diz respeito às ferramentas utilizadas e à grande quantidade de dados em análise, principalmente no que toca ao tempo de processamento de tão grande quantidade de informação.

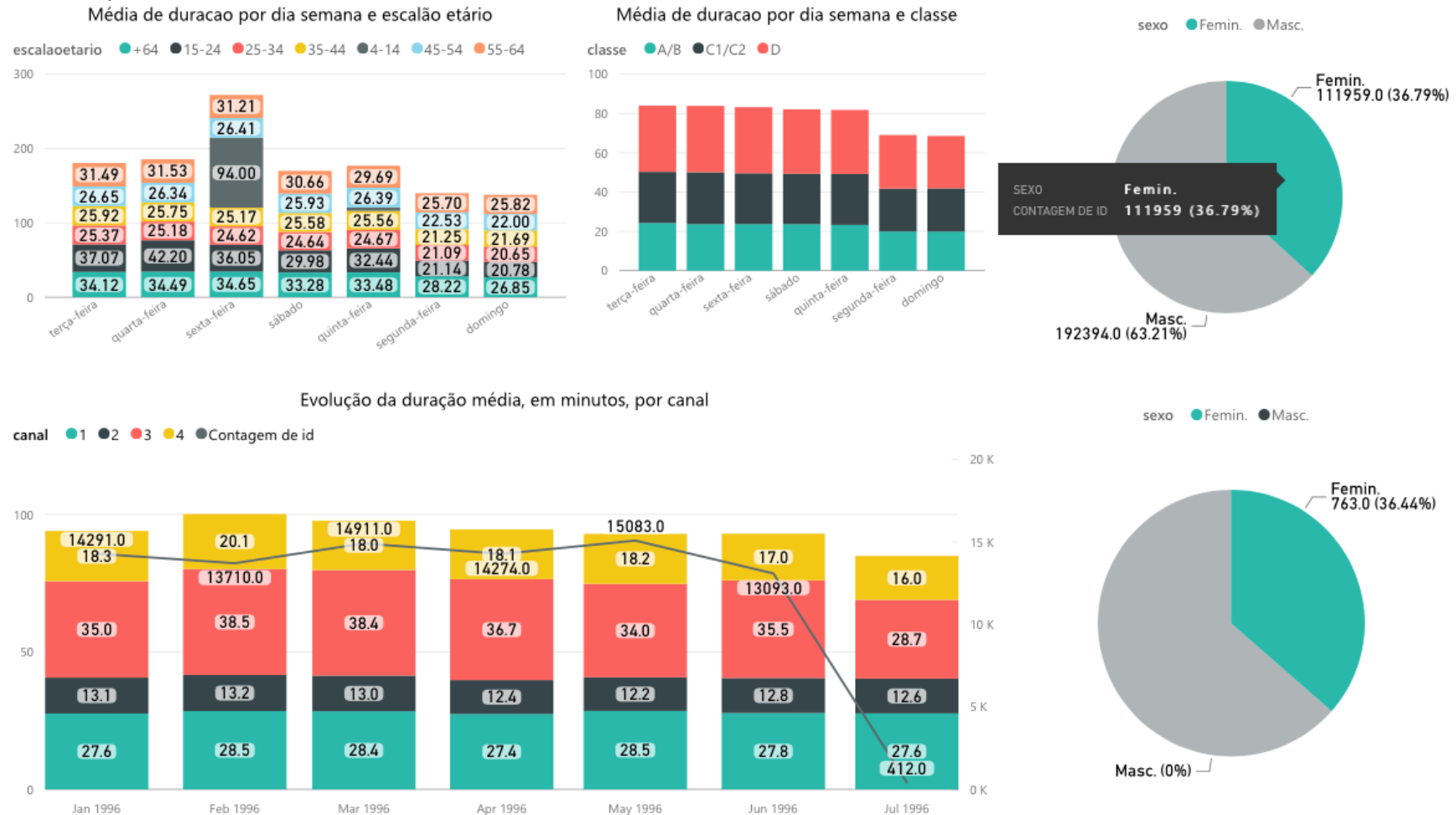
Ainda assim, com este trabalho foi possível obter uma mais clara noção da realidade da televisão portuguesa, atingindo o objetivo final deste projeto.

9. 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

10. Anexos

10.1. Esquema de Gráficos



Esquema 2 - Conjunto de gráficos relativos a todas as fontes de dados, onde é destacado o conjunto de dados relativo aos registos do sexo feminino

10.2. Código SQL para Criação das Tabelas

```
CREATE TABLE dimHorario (  
    id NUMERIC(6,0),  
    [horario completo] NVARCHAR(MAX) NOT NULL,  
    [periodo do dia] NVARCHAR(MAX) NOT NULL,  
    hora NUMERIC(2,0) NOT NULL,  
    minutos NUMERIC(2,0) NOT NULL,  
    segundos NUMERIC(2,0) NOT NULL,  
    constraint pk_dimHorario  
        primary key (id),  
    constraint ck_dimHorario  
        check (id>=0)  
);
```

Captura de Ecrã 1 - Comando SQL para criar a tabela referente à dimensão Horário

```
CREATE TABLE dimPrograma (  
    id NUMERIC(9,0),  
    [nome geral] NVARCHAR(MAX) NOT NULL,  
    [nome especifico] NVARCHAR(MAX) NOT NULL,  
    canal NUMERIC(2,0) NOT NULL,  
    tipo NVARCHAR(MAX) NOT NULL,  
    categoria NVARCHAR(MAX) NOT NULL,  
    genero NVARCHAR(MAX) NOT NULL,  
    constraint pk_dimPrograma  
        primary key (id),  
    constraint ck_dimPrograma_id  
        check (id>0)  
);
```

Captura de Ecrã 2 - Comando SQL para criar a tabela referente à dimensão Programa

```
CREATE TABLE dimEspetador (  
    id NUMERIC(9,0),  
    [chave supernatural espetador] NUMERIC (9,0),  
    codigo NUMERIC(9,0) NOT NULL,  
    genero NVARCHAR(MAX) NOT NULL,  
    [escalao etario] NVARCHAR(MAX) NOT NULL,  
    regiao NVARCHAR(MAX) NOT NULL,  
    estatuto1 NVARCHAR(MAX) NOT NULL,  
    ocupacao1 NVARCHAR(MAX) NOT NULL,  
    estatuto2 NVARCHAR(MAX) NOT NULL,  
    ocupacao2 NVARCHAR(MAX) NOT NULL,  
    [dona de casa] NVARCHAR(MAX) NOT NULL,  
    [data inicio] DATE,  
    [data fim] DATE,  
    [em vigor] NVARCHAR(MAX),  
    constraint pk_dimEspetador  
        primary key (id),  
    constraint ck_dimEspetador_id  
        check (id>0)  
);
```

Captura de Ecrã 3 - Comando SQL para criar a tabela referente à dimensão Espetador

```

CREATE TABLE dimData (
    id NUMERIC(8,0),
    [data completa] DATE NOT NULL,
    [dia do mes] NUMERIC(2,0) NOT NULL,
    mes NUMERIC(2,0) NOT NULL,
    [nome do mes] NVARCHAR(MAX) NOT NULL,
    ano NUMERIC(4,0) NOT NULL,
    [dia da semana] NVARCHAR(MAX) NOT NULL,
    [semana do ano] NUMERIC(2,0) NOT NULL,
    [fim de semana] NVARCHAR(MAX) NOT NULL,
    [indicador feriado] NVARCHAR(MAX) NOT NULL,
    [indicador data comemorativa] NVARCHAR(MAX) NOT NULL,
    [nome data comemorativa] NVARCHAR(MAX) NOT NULL,
    constraint pk_dimData
        primary key (id),
    constraint ck_dimData
        check (id>0)
);

```

Captura de Ecrã 4 - Comando SQL para criar a tabela referente à dimensão Data

```

CREATE TABLE factAudiencias (
    espetador NUMERIC(9,0),
    programa NUMERIC(9,0),
    [data inicio] NUMERIC(8,0),
    [hora inicio] NUMERIC(6,0),
    duracao NUMERIC(10,0) CONSTRAINT nn_factVenda_duracao NOT NULL,

    CONSTRAINT pk_factAudiencias
        PRIMARY KEY (espetador, programa, [data inicio], [hora inicio]),

    CONSTRAINT fk_factAudiencias_espetador
        FOREIGN KEY (espetador)
        REFERENCES dimEspetador(id),

    CONSTRAINT fk_factAudiencias_programa
        FOREIGN KEY (programa)
        REFERENCES dimPrograma(id),

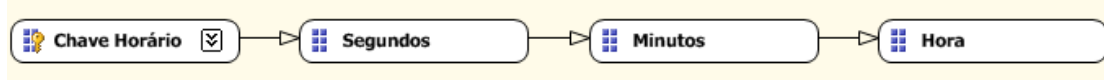
    CONSTRAINT fk_factAudiencias_data
        FOREIGN KEY ([data inicio])
        REFERENCES dimData(id),

    CONSTRAINT fk_factAudiencias_horario
        FOREIGN KEY ([hora inicio])
        REFERENCES dimHorario(id)
);

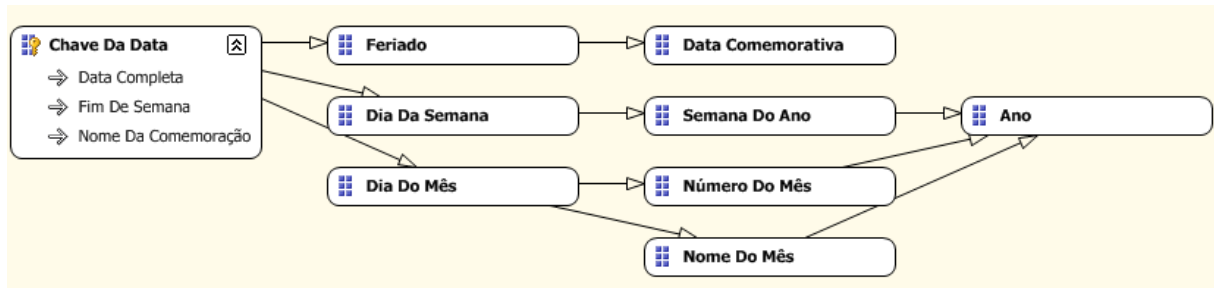
```

Captura de Ecrã 5 - Comando SQL para criar a tabela de factos

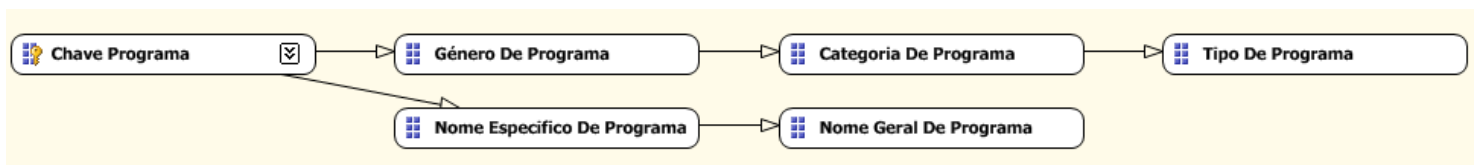
10.3. Hierarquias de Atributos



Esquema 3 - Hierarquia de atributos da dimensão Horário



Esquema 6 - Hierarquia de atributos da dimensão Data



Esquema 7 - Hierarquia de atributos da dimensão Programa

10.4 Código SQL para Prospecção de Informação

```
use IPAI07BD;
```

```
DROP TABLE #categoria_vista;
DROP TABLE #espetador_classificacao;
DROP TABLE data_mining;
```

```
SELECT f.espetador, p.categoria, SUM(f.duracao) AS soma
INTO #categoria_vista
FROM factAudiencias AS f, dimPrograma AS p, dimData AS d
WHERE f.programa = p.id
AND f.[data inicio] = d.id AND d.[data completa] BETWEEN '1996-01-01' AND '1996-01-07'
GROUP BY f.espetador, p.categoria
ORDER BY f.espetador ASC, soma DESC;
```

```
SELECT cl.espetador, cl.categoria AS classificacao
INTO #espetador_classificacao
FROM #categoria_vista AS cl
WHERE (cl.categoria = 'TELENOVELA' OR cl.categoria = 'NOTICIARIO' OR cl.categoria = 'PROG. INF. JUVENIL')
AND cl.soma >= ALL (SELECT c2.soma
FROM #categoria_vista AS c2
WHERE cl.espetador = c2.espetador);
```

```
SELECT f.espetador, e.genero, e.[escala etario], e.regiao, e.estatuto1, e.ocupacao1, e.estatuto2, e.ocupacao2, e.[dona de casa],
f.programa, p.canal, p.categoria, d.[data completa], h.[horario completo], h.[periodo do dia], f.duracao, c.classificacao
INTO data_mining
FROM factAudiencias AS f, dimEspetador AS e, dimPrograma AS p, dimData AS d, dimHorario AS h, #espetador_classificacao AS c
WHERE f.espetador = e.id AND f.programa = p.id AND f.[data inicio] = d.id AND c.espetador = f.espetador AND f.[hora inicio] = h.id AND
f.espetador IN (SELECT espetador FROM #espetador_classificacao) AND f.programa = p.id AND
(p.categoria = 'TELENOVELA' OR p.categoria = 'NOTICIARIO' OR p.categoria = 'PROG. INF. JUVENIL') AND
d.[data completa] BETWEEN '1996-01-01' AND '1996-01-07'
ORDER BY f.espetador, d.[data completa], h.[horario completo];
```

Captura de Ecrã 6 – Query SQL para gerar a tabela com dados que irão ser usadas para prospecção da informação