Modelação Dimensional
Etapa 2

Mestrado em Engenharia Informática
Tecnologias de Processamento de Dados
2017/2018

Grupo 2:
Gonçalo Lobo 44870
Liliana Ramos 44873
Pedro Gomes 44858

Índice

1. Introdução

Este trabalho foi realizado no âmbito da cadeira de Tecnologias de Processamento de Dados e envolve a modelação e construção de um *data warehouse*, que irá incorporar dados relativos às audiências televisivas do primeiro semestre do ano de 1996.

Na primeira etapa foi efetuada uma identificação das fontes de dados para o processo de negócio escolhido. Foram também analisados os dados e as ligações entre os dados das diferentes fontes e depois foram mostrados num diagrama.

Com este relatório pretende-se melhorar a etapa anterior utilizando assim, as fontes de dados disponibilizadas pelo docente.

Este relatório refere-se à segunda etapa em que será efetuada a criação de um modelo multidimensional adequado a um data warehouse. Como tal será determinado o grão da tabela de factos, serão modeladas as dimensões do negócio e, por fim, identificadas as medidas numéricas da tabela de factos. Ainda nesta etapa, foram também realizadas melhorias à etapa anterior.

2. Fontes de dados

Para a realização deste projeto foram disponibilizadas fontes de dados provenientes de 5 fontes distintas: os espetadores de televisão, os tipos de programas televisivos, os canais vistos pelos espetadores, a programação dos canais de televisão e as classes sociais.

A. Espetadores

O ficheiro *espetadores.csv* contém informação sobre os espetadores. Em baixo encontram-se descritos os 8 campos que compõem o ficheiro:

Tabela 1 - Detalhes dos dados do ficheiro espetadores.csv

Campo	Descrição	Tipo de Dados	Exemplo
ID	Identificador único de registo	Inteiro	6
Código	Identificador único de espetador	Inteiro	3001
Região	Região do país de residência do espetador	Texto	“Gr. Lisboa”
Sexo	Masculino ou Feminino	Texto	“Femin.”
DonaDeCasa	Se o espetador trabalha em casa ou não	Texto	“DDC”
EscalãoEtário	Escalão etário do espetador	Texto	“+64”
Classe	Classe social do espetador	Texto	“D”
Data	Data de criação do registo	Data	#1996-01-01#

B. Tipologia

O ficheiro *tipologia.csv* contém informação sobre as classificações dos vários tipos de programas televisivos. Em baixo estão descritos os 2 campos que compõem o ficheiro:

Tabela 2 - Detalhes dos dados do ficheiro tipologia.csv

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

C. Canais Vistos pelos espetadores

O ficheiro *audiencias.csv* contém informação sobre os tempos de seleção de canais por cada espetador. Em baixo estão descritos os 6 campos que compõem o ficheiro:

Tabela 3 - Detalhes dos dados do ficheiro audiencias.csv

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

D. Programação dos canais de televisão

Esta fonte de dados é constituída por 182 ficheiros com uma extensão do tipo .pet e contém informação sobre a programação do primeiro trimestre do ano de 1996.

Tabela 4 - Detalhes dos dados do ficheiro yyyymmdd.pet

Campo	Descrição	Tipo de Dados	Exemplo
Canal	Número do canal	Inteiro	1
HoraInício	Hora de início do programa, no formato hhmmss	Inteiro	20000
Duração	Duração do conteúdo televisivo em segundos	Inteiro	162
Zero	Sem significado	Inteiro	0
Nome1	Nome do conteúdo televisivo	Texto	"SESSAO DUPLA I"
Nome2	Segundo nome do conteúdo televisivo	Texto	"CLASSE"
Classificação	Classificação do conteúdo, detalhada a seguir	Texto	"P"
Tipo	Tipo do conteúdo, de acordo com a tipologia em cima	Texto	"aae"

ParteTodo	Se representa o conteúdo todo ou uma das suas partes	Inteiro	1

E. Classes sociais

O ficheiro *classes.tsv* descreve o significado das letras A, B, e outras, que identificam classes sociais de cada um dos espetadores.

Tabela 5 - Detalhes dos dados do ficheiro classes.csv

Campo	Descrição	Tipo de Dados	Exemplo
Classe	Representa a classe social	Texto	B
Estatuto	Estatuto social	Texto	Classe média
Ocupação	Ocupações representativas	Texto	Gestor, administrador, ou profissional intermédio

3. Diagrama de dados

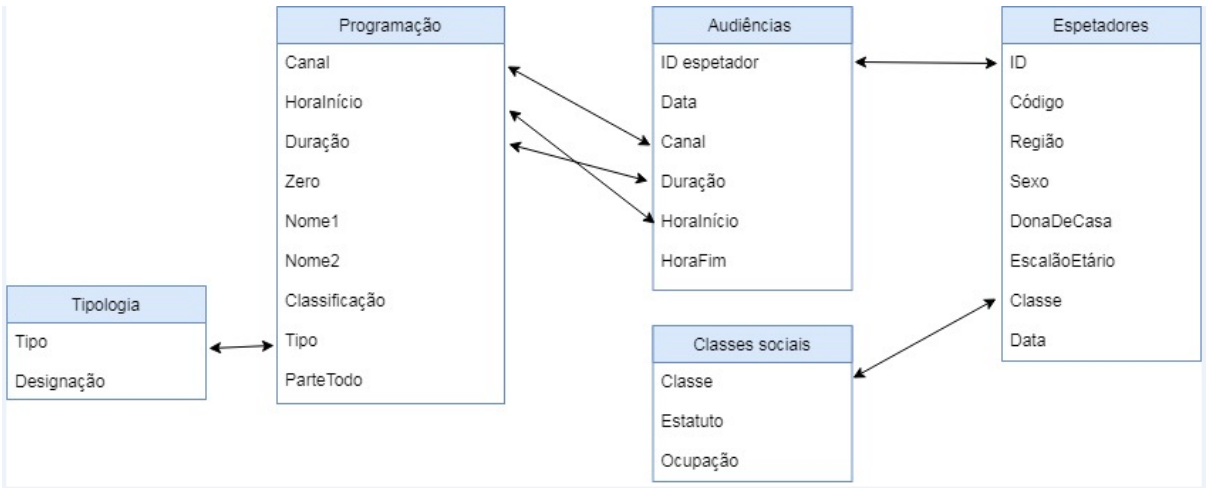


Figura 1 - Diagrama dos Dados

A partir do Diagrama de dados podemos reparar que na fonte de dados "Espetadores" o campo **classe** se relaciona com o campo **classe** da fonte de dados "Classes sociais" e que o **ID** se relaciona com o campo **ID espetador** das "Audiências". Com estas ligações conseguimos obter informação relativa ao espectador, à sua classe social, e o que um determinado espetador via na televisão no primeiro semestre de 1996.

Na fonte de dados "Programação" é possível notar que existem vários campos associados à fonte de dados "Audiências", sendo assim possível relacionar o **Canal**, a **HoraInício** e a **Duração** de um determinado programa. Para além disto, na fonte de dados "Programação" existe uma associação, no campo do **tipo,** à fonte de dados "Tipologia", que contém o tipo de programa.

4. Análise das Fontes de Dados

Em cada análise apresentada de seguida incluímos o código necessário e usado no RStudio para obtermos os dados para análise. Caso tenha sido usado outro método é também explicado o processo.

A. espetadores.csv

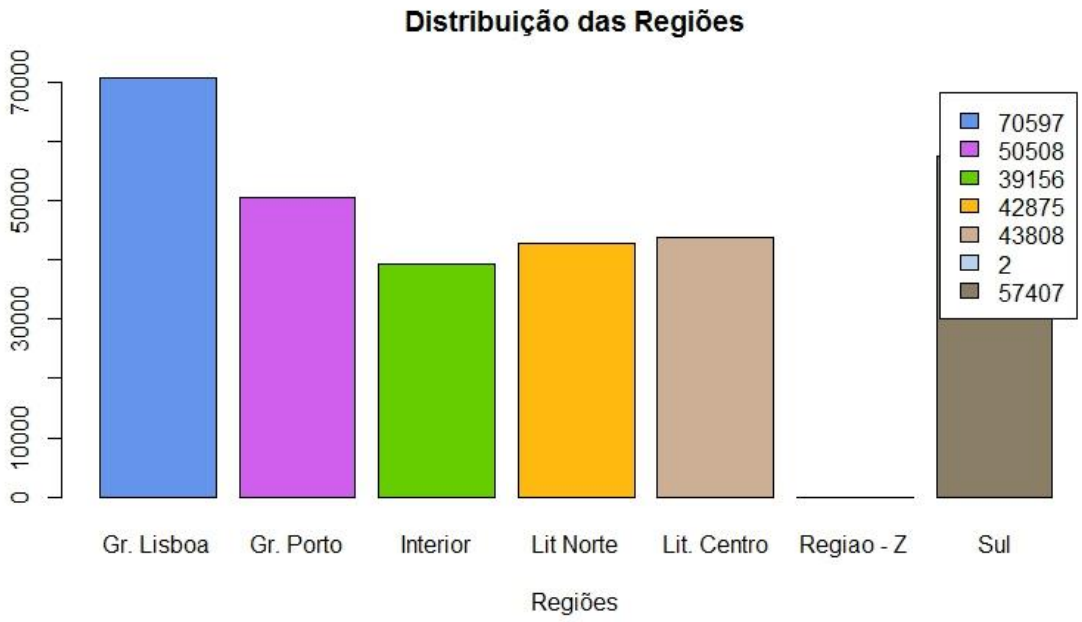


Figura 2 - Distribuição das Regiões

Região: É possível observar, a partir do gráfico, que existe uma maior afluência de espetadores (70597) na região da Grande Lisboa.

```
> count_regiao <- table(espetadores$Região)
> barplot(count_regiao, main="Distribuição das Regiões", xlab="Regiões", col = c("cornflowerblue", "mediumorchid2", "chartreuse3", "darkgoldenrod1", "peachpuff3", "slategray2", "wheat4"), legend.text = count_regiao)
```

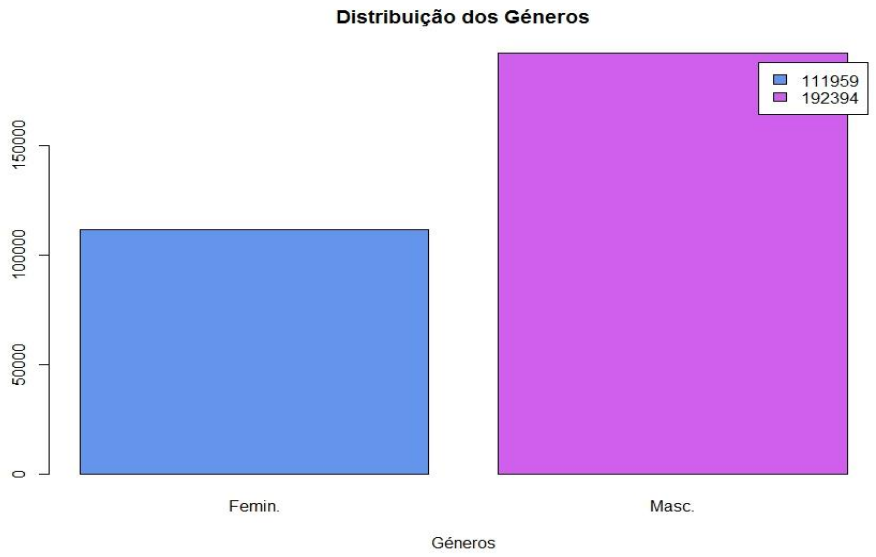
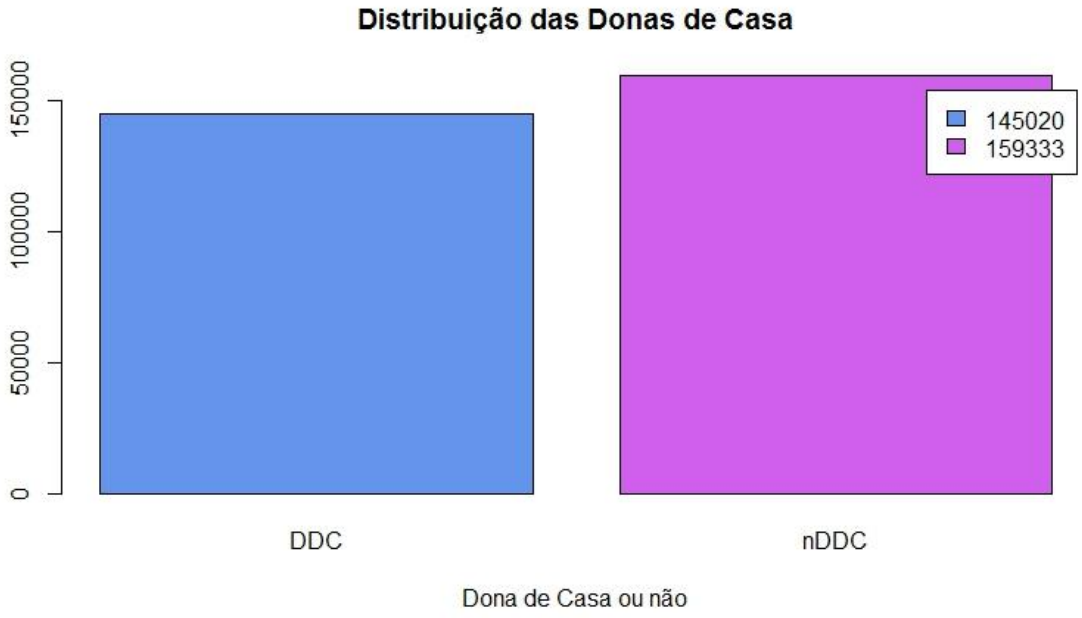


Figura 3 - Distribuição dos Géneros

Sexo: A partir deste campo conseguimos perceber que existem 111959 espetadores do sexo feminino e 192394 espetadores do sexo masculino.

```
> count_gender <- table(espetadores$Sexo)
> barplot(count_gender, main="Distribuição dos Géneros", xlab="Géneros", col =
```




```
c("cornflowerblue", "mediumorchid2"), legend.text = count_gender)
```

Figura 4 - Distribuição das Donas de Casa

DonaDeCasa: Através da análise deste gráfico é possível concluir que a maioria dos espetadores não são Donas de Casa (159333).

```
> count_ddc <- table(espetadores$DonaDeCasa)
> barplot(count_ddc, main="Distribuição das Donas de Casa", xlab="Dona de Casa
ou não", col = c("cornflowerblue", "mediumorchid2"), legend.text = count_ddc)
```

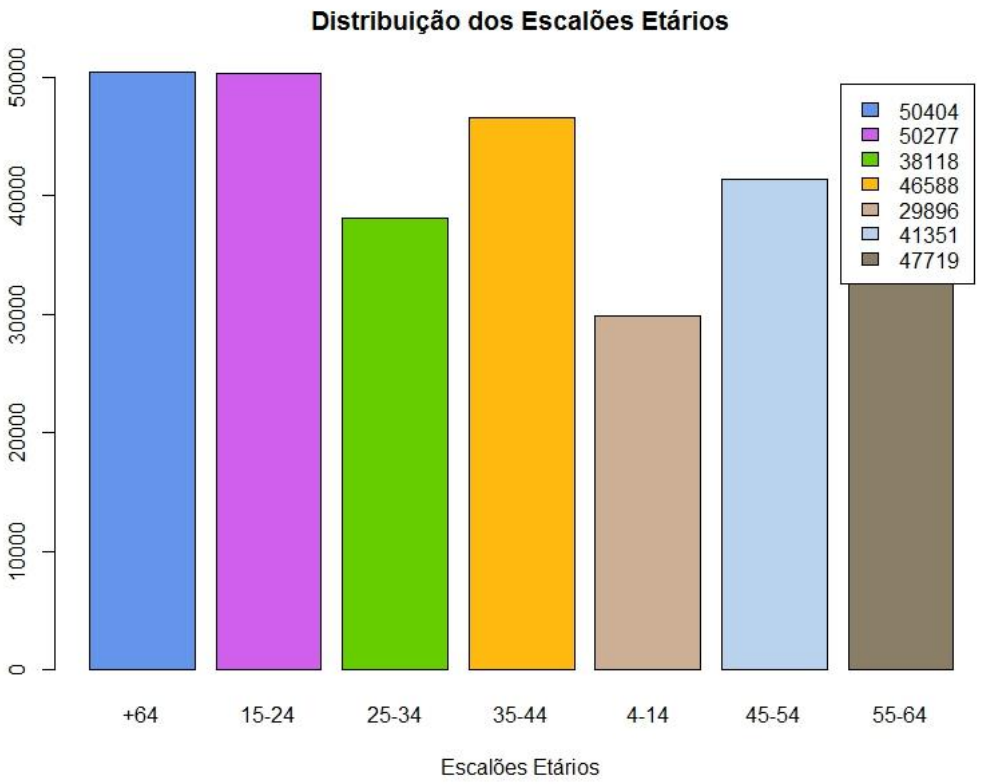


Figura 5 - Distribuição dos Escalões Etários

EscalãoEtário: É possível observar que o escalão etário mais influente nos dados é o escalão de "+64" apesar de o escalão "15-24" ter um número de espetadores muito próximo.

```
> count_escalao <- table(espetadores$EscalãoEtário)
> barplot(count_escalao, main="Distribuição dos Escalões Etários", xlab="Escalões
Etários", col = c("cornflowerblue", "mediumorchid2", "chartreuse3", "darkgoldenrod1",
"peachpuff3", "slategray2", "wheat4"), legend.text = count_escalao)
```

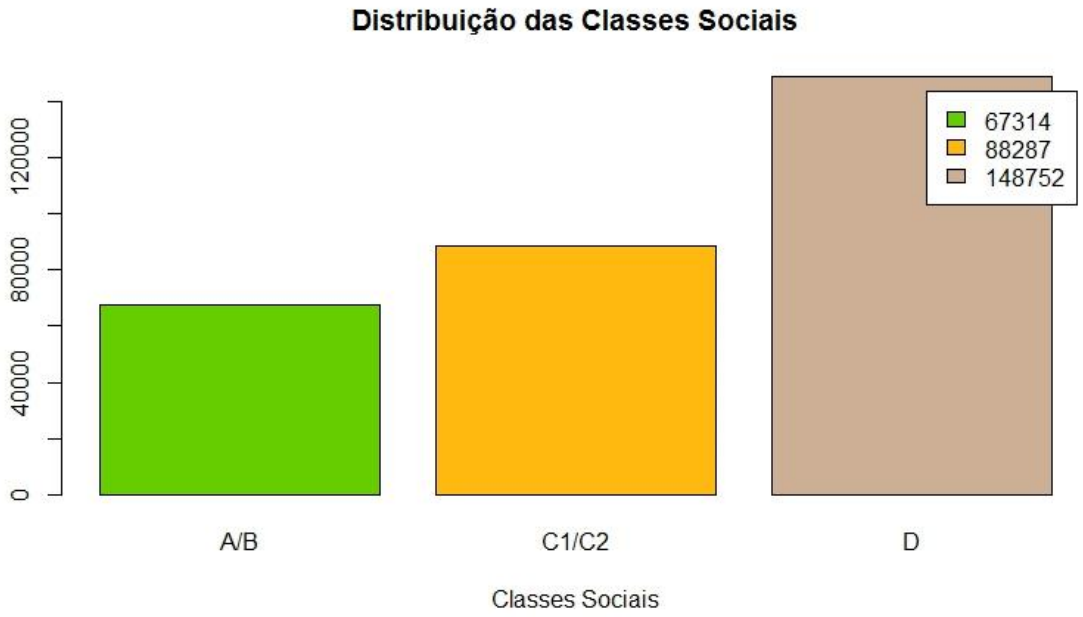


Figura 6 - Distribuição das Classes Sociais

Classe: Analisando o gráfico observamos que a classe social D, correspondente à "Classe Trabalhadora", é a classe com maior representação nos espetadores.

```
> count_classe <- table(espetadores$Classe)
> barplot(count_classe, main="Distribuição das Classes Sociais", xlab="Classes Sociais", col = c("chartreuse3", "darkgoldenrod1", "peachpuff3"), legend.text = count_classe)
```

```
> summary(espetadores$Data)
        Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
"1996-01-01" "1996-02-15" "1996-03-31" "1996-03-31" "1996-05-15" "1996-06-30"
```

Data: É possível observar que o primeiro dia das audiências é dia 01-01-1996 e o último dia é o dia 30-06-1996 (correspondente ao final do primeiro semestre).

B. tipologia.tsv

```
> summary(tipologia)
    Tipo            Designacao
Length:238         Length:238
Class :character   Class :character
Mode  :character   Mode  :character
```

Figura 7 - Tipologia

É possível observar que existem 238 tipos de programas diferentes. Outro aspeto verificado é que os tipos de programas encontram-se definidos em forma de hierarquia.

C. audiencias.csv

```
> summary(audiencias$ID)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     57   71030  150800  149900  226300  304400
```

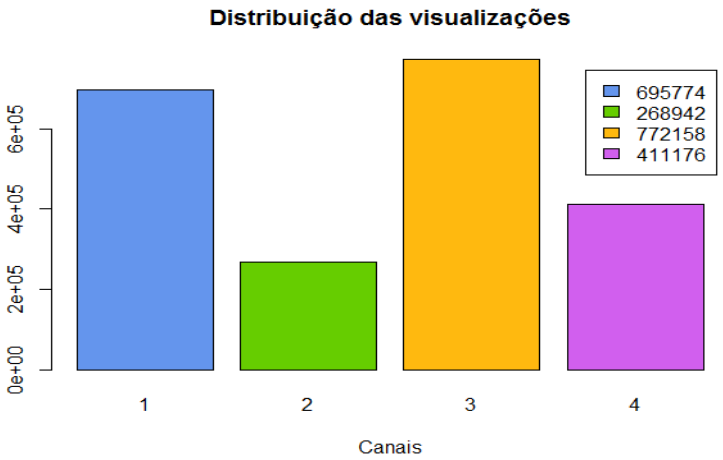
Figura 8 - Sumário das audiências

ID: É possível observar que existem 304400 entradas na fonte de dados *audiências.csv*. Foi utilizado novamente o comando *"summary"* no RStudio.

```
> summary(audiencias$Data)
        Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
"1996-01-01" "1996-02-12" "1996-03-30" "1996-03-29" "1996-05-14" "1996-06-30"
```

Figura 9 - Audiências

Data: É possível observar que o primeiro dia das audiências é dia 01-01-1996 e o último dia do semestre é o dia 30-06-1996 (como já concluído anteriormente). A partir deste campo podemos retirar que o mês com mais



visualizações foi em Janeiro, como demonstrado no gráfico seguinte:

Figura 10 - Visualizações

Canal: Podemos observar no gráfico anterior que existem 4 canais e que o canal com maior número de visualizações é o 3 com 772158 visualizações. Por outro lado temos o canal 2 com o menor número de visualizações.

```
> counts = table(audiencias$Canal)
> barplot(counts, main="visualizações", xlab = "canais", legend.text = counts, col = c("cornflowerblue","chartreuse3" , "darkgoldenrod1"  , "mediumorchid2")
```

```
> summary(audiencias$Duração)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    2.00    8.00   24.93   29.00 1055.00
```

Figura 11 - Summary Audiências

Duração: A partir da análise do sumário concluímos que a duração mínima de visualização de um canal é de 2 minutos e a maior duração é 1055 minutos. Podemos também concluir que a média da duração de visualização é de 24.93

```
  HoraInicio                    HoraFim
Min.   :1996-01-01 02:00:00   Min.   :1996-01-01 02:01:00
1st Qu.:1996-02-12 21:56:00   1st Qu.:1996-02-12 22:16:00
Median :1996-03-31 00:35:00   Median :1996-03-31 00:50:00
Mean   :1996-03-30 14:39:29   Mean   :1996-03-30 15:04:59
3rd Qu.:1996-05-14 21:02:00   3rd Qu.:1996-05-14 21:39:00
Max.   :1996-07-01 01:54:00   Max.   :1996-07-01 01:55:00
NA's   :3221                  NA's   :3858
```

minutos.

Figura 12 - Summary Horas

HoraInicio: A partir da análise do sumário, observamos que os registos das audiências começaram no dia 01-01-1996, com início às 2 horas da manhã.

HoraFim: A partir da análise do sumário, observamos que os registos das audiências terminaram no dia 01-07-1996, com fim à 01:55 horas da manhã.

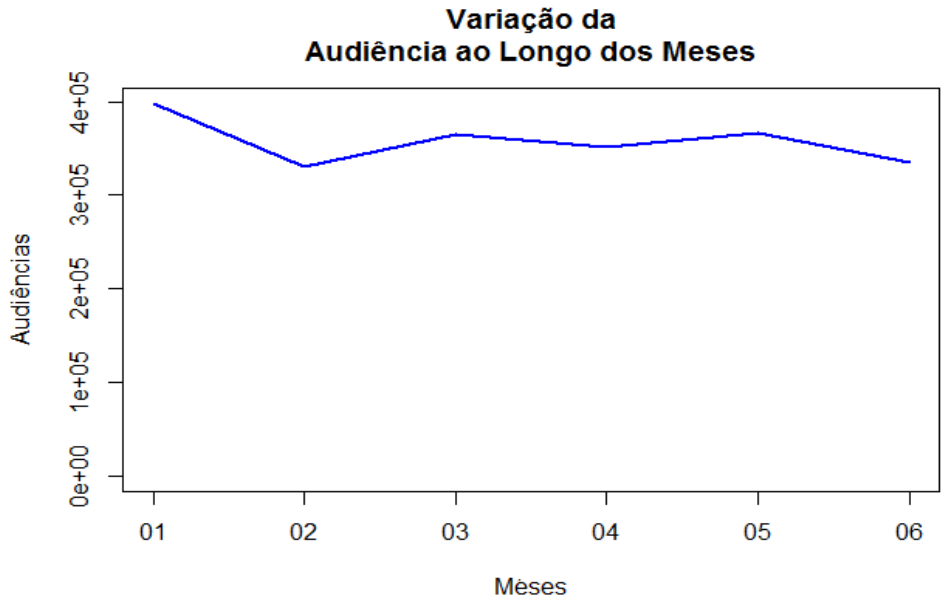


Figura 13 - Variação Audiência

Para além da análise das várias variáveis, fizemos também uma análise da variação das audiências ao longo dos meses. É possível observar que no início do mês 1 foi quando se teve o pico de audiências e o menor número de audiências

verificou-se no segundo mês. Entre o 3º mês e o 5º as audiências estiveram aproximadamente constantes tendo voltado a diminuir novamente no 6º mês.

```
> summary(audiencias)
> meses = substr(audiencias$Data, 6, 7)
> plot(table(mes), xlab="Meses", ylab="Audiências", type="l", main="Variação da Audiência ao Longo dos Meses", col="blue")
```

D. Ficheiros Pet

Para poder ser feita uma análise geral de todos os ficheiros pet disponibilizados foi necessário juntar todos num só (visto terem todos o mesmo número e tipo de variáveis). Para isso, foi executado o seguinte comando na linha de comandos:

```
> copy *.pet programacao.pet
```

```
> summary(programacao$Duracao)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0    21.0    64.0   602.6   522.5 12270.0
```

Figura 14 - Summary Duração Programação

Através de um sumário da duração dos conteúdos televisivos conseguimos perceber que o mínimo da duração é 2 segundos e o máximo é 12270 segundos. Como média de duração temos 602.6 segundos.

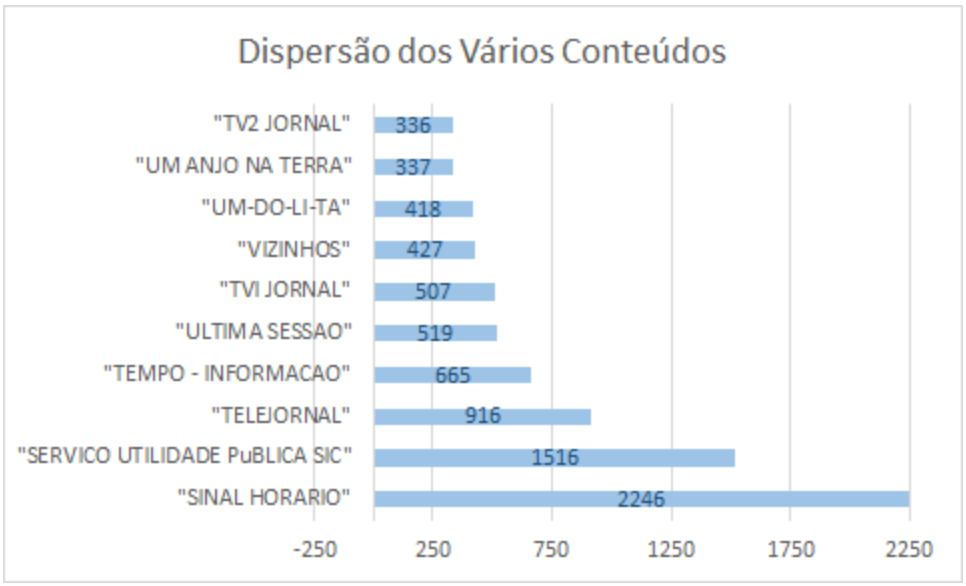
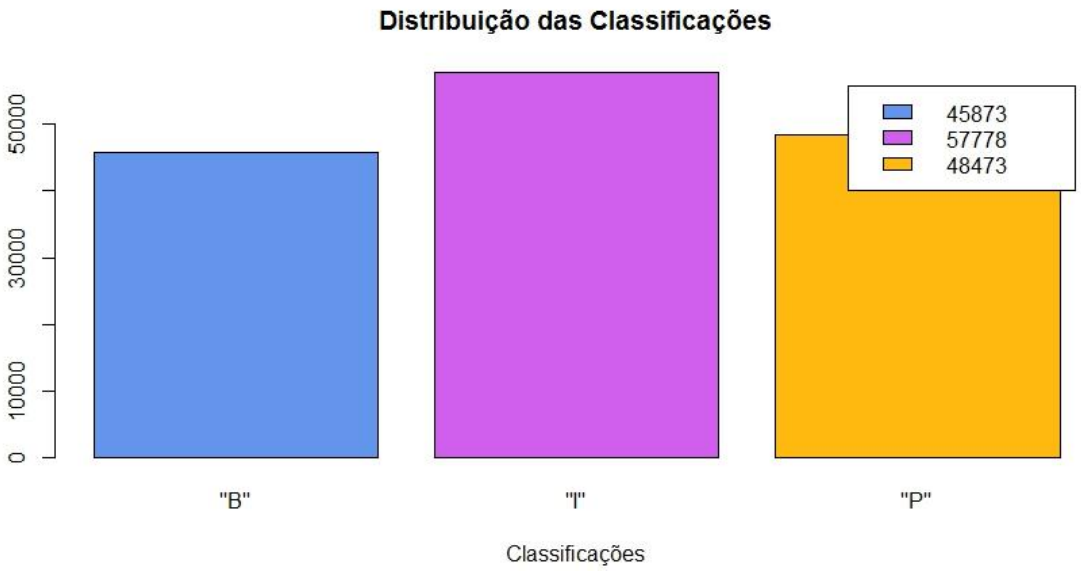


Figura 15 - Dispersão Conteúdos

Nome1: É possível observar que o programa "Sinal Horário" foi o mais visualizado. Em seguida tem-se o "Serviço Utilidade Pública SIC" e em 3º lugar o "Telejornal".

Para poder chegar-se a estes dados para serem analisados foi feito o seguinte processo:

1. No RStudio:
   a. count_conteudos <- table(programacao$Nome1)
   b. View(count_conteudos)
2. Copiaram-se todos os dados da view para um ficheiro Excel.
3. Organizaram-se os dados por ordem decrescente pelo número de visualizações dos programas.
4. Com os primeiros 10 valores, criou-se um gráfico de barras horizontal para



podermos observar graficamente a dispersão dos vários conteúdos.

Figura 16 - Distribuição Classificações

Classificação: Através do gráfico anterior é possível concluir que a classificação que tem o maior número de conteúdos associado é a "I" para Publicidade ao Próprio Canal. De seguida encontra-se a classificação "P" para Programa e por fim a "B" para Intervalo Comercial.

```
> count_classificacao <- table(programacao$Classificação)
> barplot(count_classificacao, main="Distribuição das Classificações",
xlab="Classificações", col = c("cornflowerblue", "mediumorchid2", "darkgoldenrod1"),
legend.text = count_classificacao)
```

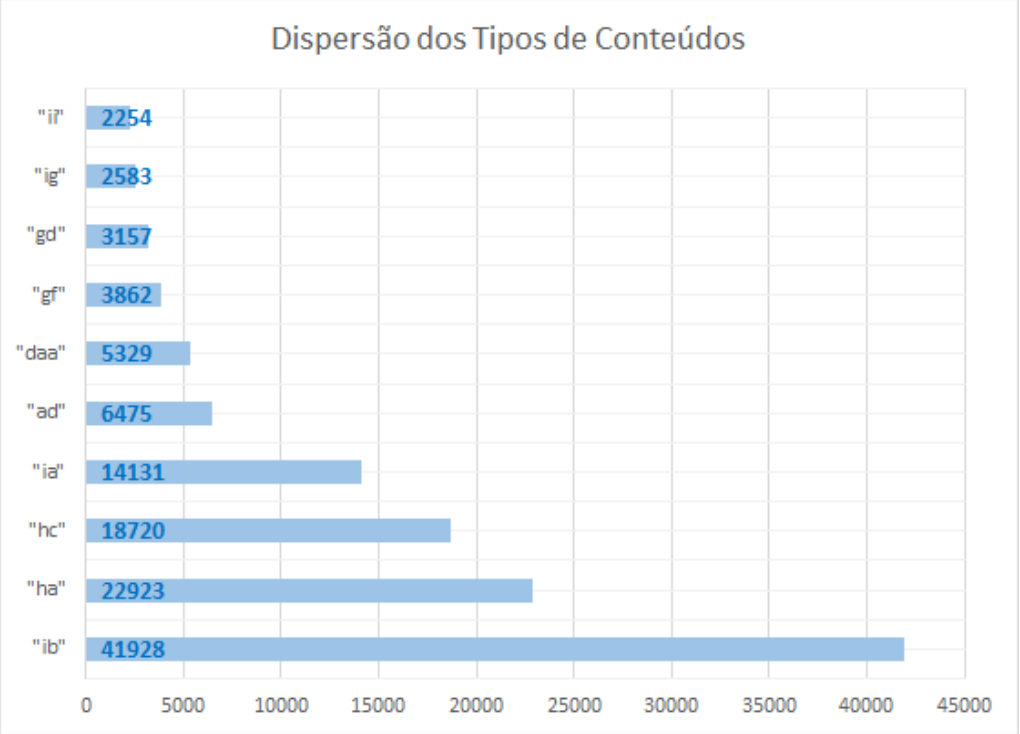


Figura 17 - Dispersão Conteúdos

Tipo: Neste gráfico é possível observar quais os tipos de conteúdos televisivos mais vistos que, neste caso, são o tipo "ib" para "INTERVALO AP.PROGRAM" e o tipo "ha" que corresponde a "ECRANS PUBLICITARIOS".

Para poder chegar-se a estes dados para serem analisados foi feito o seguinte processo:

1. No RStudio:
   a. count_tipos <- table(programacao$Tipo)
   b. View(count_tipos)
2. Copiaram-se todos os dados da view para um ficheiro Excel.
3. Organizaram-se os dados por ordem decrescente pelo número de programas com aquele tipo específico.
4. Com os primeiros 10 valores, criou-se um gráfico de barras horizontal para podermos observar graficamente a dispersão dos tipos de conteúdos mais frequentes.

E. classes.tsv

Neste ficheiro é possível verificar que existem 6 classes sociais diferentes: Classe média/alta, Classe média, Classe média/baixa, Classe trabalhadora qualificada, Classe trabalhadora e Aqueles com menor nível de subsistência.

5. Erros encontrados e correções das fontes de dados

A. Espetadores

Erro 1: Existem dois identificadores únicos dos espetadores repetidos sendo

```
> anyDuplicated(espetadores$v2, incomparables = FALSE)
[1] 275
```

274	274	12092204	Lit Norte	Masc.	DDC	4-14	A/B	#1996-01-01#
275	275	12092204	Interior	Masc.	DDC	55-64	A/B	#1996-01-01#

que têm regiões e idades diferentes (linhas 274 e 275).

Figura 18 - Demonstração Erro Espetadores

Resolução: Foi eliminada a primeira entrada encontrada, isto é, a 274.

B. Audiências

Erro 1: Existem durações com valor de 0, o que não faz sentido pois um programa não pode ter duração 0.

Resolução: Utilizamos o excel para remover essas mesmas entradas (linhas 696035, 1189667 e 1695440).

Erro 2: Reparámos também que existem datas de início e de fim incompletas (consideradas como NA no RStudio). Conseguimos concluir também que isto acontecia sempre que a data de início ou de fim tinha a hora de 00:00:00.

Resolução: Foi criado um script em Java (anexado) para corrigir estes dados.

	ID	Data	Canal	Duração	HoraInicio	HoraFim
794	183	1996-01-01	3	16	1996-01-01 23:10:00	1996-01-01 23:26:00
795	183	1996-01-01	3	11	1996-01-01 23:27:00	1996-01-01 23:38:00
796	183	1996-01-01	3	17	1996-01-01 23:43:00	*NA*

Figura 19 - Demonstração Erro Audiências

	ID	Data	Canal	Duração	HoraInicio	HoraFim
794	183	1996-01-01	3	16	1996-01-01 23:10:00	1996-01-01 23:26:00
795	183	1996-01-01	3	11	1996-01-01 23:27:00	1996-01-01 23:38:00
796	183	1996-01-01	3	17	1996-01-01 23:43:00	1996-01-02 00:00:00

Figura 20 - Demonstração Erro Audiências

Erro 3: Notámos que o ficheiro audiências apenas contém espetadores cujos id's correspondem a valores acima de 56.

Resolução: Para resolver este erro, no ficheiro espetadores.csv eliminaram-se todas as linhas dos espetadores com o id <= 56.

C. Ficheiros .pet

Erro 1: Existem linhas com o erro de estar a faltar uma vírgula entre dois campos, que tem a função de os separar, como no exemplo abaixo:

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

Resolução: Procurou-se todas as linhas que tinham este erro e entre esses campos foi colocada uma vírgula.

Erro 2: Existiam linhas com o erro de ter um dos campos, geralmente o campo ***Nome2***, apenas preenchido com aspas (“) quando o campo preenchido de forma correta deve estar com 2 aspas, como no exemplo abaixo:

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

Resolução: Colocou-se a restante “ em todas as linhas onde este erro ocorria.

6. Processo de Negócio

O processo de negócio da indústria televisiva tem como base o estudo das audiências de modo a entender quais as preferências dos espetadores em termos de conteúdos mais vistos e a duração dessa mesma visualização. Essa mesma indústria serve dois tipos de clientes: os clientes que assistem à programação sem qualquer tipo de pagamento, ou seja, os espetadores, e os clientes que compram o acesso a tempo publicitário para promover os seus produtos. Para isso é importante existir uma máxima gestão de recursos televisivos tentando otimizar a programação baseando-se nas audiências e consoante o tipo de espetadores. Assim o nosso processo de negócio recai na análise de audiências consoante a programação e os espetadores, de modo a conseguir perceber quais os programas com uma maior valorização televisiva, e consequentemente o tempo dedicado à publicidade.

Seguem-se então 5 questões analíticas que o grupo considerou relevantes para este processo de negócio em questão.

1. Quais os tipos de programas mais vistos pelos diferentes sexos e escalões etários?
2. Qual é o canal televisivo mais visto por região ao longo do semestre? E por escalão etário?
3. Qual é o dia em que um determinado canal registou maior audiência?
4. Qual a classe social que assiste mais a um intervalo comercial?
5. Qual o período do dia que um determinado canal registou maior audiência?

7. Grão e tipo da tabela de factos

O grão determina o nível máximo de detalhe. Este detalhe deve ser o maior possível para os recursos disponíveis. Por exemplo, tabelas com um grão mais fino são maiores mas também são mais expressivas. Para além disto, o grão identifica as dimensões e o detalhe a guardar nas mesmas. O número de dimensões tende a ser maior quanto mais fino for o grão.

A tabela de factos é composta por chaves estrangeiras que fazem referência às chaves primárias das tabelas de dimensões, em que cada valor da chave primária da tabela identifica de forma unívoca um facto. Sendo assim, considerámos que a nossa tabela de factos será composta pelas seguintes dimensões: Programa, Data, Espetador e Hora.

Então concluímos que cada linha da tabela corresponderá a um **programa**, emitido numa determinada **data** visto por um determinado **espetador** com início numa **horaInicio** e com uma **duração**.

Tabela 1 - Atributos da tabela de factos

Campo	Descrição	Tipo de Dados	Exemplo
ID Programa	Identificador único que identifica um Programa	Inteiro	15
ID Data	Identificador único que identifica uma Data	Inteiro	30
ID Espetador	Identificador único que identifica um espetador	Inteiro	58
ID horaInicio	Identificador único que identifica a hora de início	Inteiro	10

8. Modelar dimensões do negócio

8.1 Dimensão Programa

A tabela da Dimensão *Programa* contém informações sobre os programas que foram exibidos no primeiro semestre de 1996.

Hierarquia: Tipo > Subtipo > Subsubtipo

Tabela 2 - Atributos da Tabela de Dimensão Programa

Campo	Descrição	Tipo de Dados	Exemplo
ID Programa	Identificador do programa	Inteiro	1
Canal	Número do canal	Inteiro	1
Duração	Duração do conteúdo televisivo em segundos	Inteiro	162
Nome1	Nome do conteúdo televisivo	Texto	“SESSAO DUPLA I”
Nome2	Segundo nome do conteúdo televisivo	Texto	“CLASSE”
Classificação	Classificação do conteúdo, detalhada a seguir	Texto	“Programa”
Tipo	Tipo do conteúdo	Texto	“Ficção”
Subtipo	Subtipo do conteúdo	Texto	“Filme”
Subsubtipo	Subtipo do subtipo do conteúdo	Texto	“Comédia”
ParteTodo	Se representa o conteúdo todo ou uma das suas partes	Inteiro	1

8.2 Dimensão Espetador

A Tabela da Dimensão *Espetador* contém informações sobre os espetadores que viram televisão no primeiro semestre de 1996. Para a elaboração desta dimensão foram utilizados as fonte de dados do espetador e da classe.

Tabela 3 - Atributos da Tabela da Dimensão Espetador

Campo	Descrição	Tipo de Dados	Exemplo
ID Espetador	Identificador único de registo	Inteiro	6
Código	Identificador único de espetador	Inteiro	3001
Região	Região do país de residência do espetador	Texto	“Gr. Lisboa”
Sexo	Masculino ou Feminino	Texto	“Femin.”

DonaDeCasa	Se o espetador trabalha em casa ou não	Texto	“DDC”
EscalãoEtário	Escalão etário do espetador	Texto	“+64”
Estatuto	Estatuto social do espetador	Texto	“Classe trabalhadora”
Ocupação	Ocupação do espetador	Texto	“Trabalhador manual pouco ou não qualificado”

8.3 Dimensão Data

A tabela da Dimensão Data contém informações relativamente à data de emissão do programa televisivo.

Hierarquia: Ano > Mês > Dia

Tabela 4 - Atributos da Tabela da Dimensão Data

Campo	Descrição	Tipo de Dados	Exemplo
ID Data	Identificador único de registo	Inteiro	1
Data Completa	Data completa da criação do registo	Data	“1996-01-01”
Ano	Ano	Inteiro	1996
Mês	Mês	Texto	“Janeiro”
Dia	Dia	Inteiro	1
Dia da semana	Dia da semana	Texto	“Segunda”

8.4 Dimensão Hora

A tabela da Dimensão Hora contém informações relativamente à hora de emissão do programa televisivo.

Hierarquia: Hora > Minutos > Segundos

Tabela 5 - Atributos da Tabela da Dimensão Hora

Campo	Descrição	Tipo de Dados	Exemplo
ID Hora	Identificador único de registo	Inteiro	1
Hora de Início	Hora de início de visualização completa	Data	“14:47:00”
Hora	Hora	Inteiro	14

Minutos	Minutos	Inteiro	47
Segundos	Segundos	Inteiro	00
Período do dia	Período do dia relativo a hora de início	Texto	“Tarde”

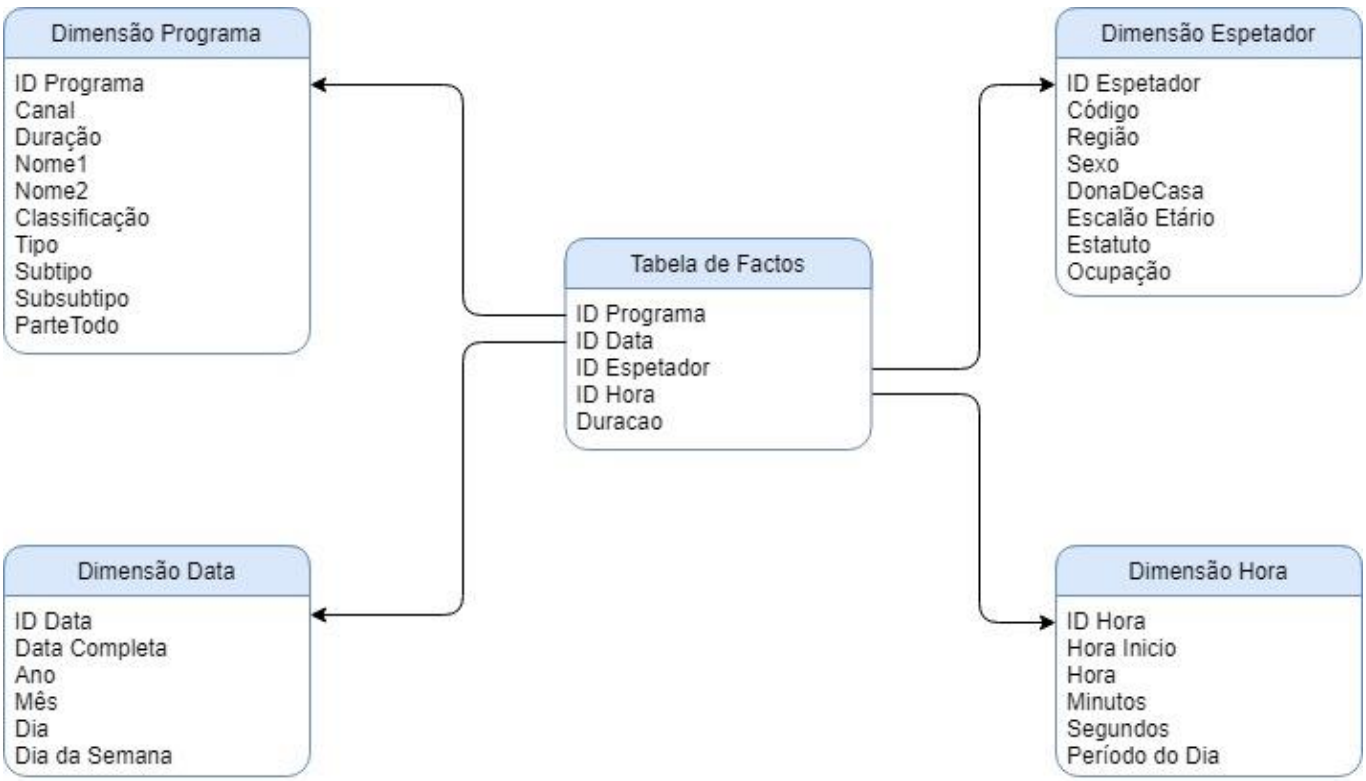
8.5 Medida numérica da tabela de factos

A tabela de factos vai conter medidas numéricas que permite avaliar um determinado processo de negócio. Anteriormente especificamos o processo de negócio, e com base nisso, consideramos importante a existência de uma medida aditiva correspondente a duração de visualização, pois é possível contabilizar os minutos correspondente a visualização de um programa.

Tabela 6 - Medida numérica da tabela de factos

Campo	Descrição	Tipo de Dados	Exemplo
Duração de visualização	Duração de visualização de um programa por um dado espectador em minutos	Inteiro	10

9. Diagrama em Estrela do *Data Warehouse*



10. Conclusão

Com a primeira etapa do projeto foi possível tratar e analisar e tratar as fontes de dados abertos recolhidas, que continham dados pouco perceptíveis e até alguns erros. Foi ainda possível elaborar um processo de negócio, bem como a elaboração das perguntas analíticas sobre o projeto que serão respondidas nas próximas etapas. Esta primeira etapa é uma melhoria em relação à etapa entregue anteriormente.

Na segunda etapa foi possível entender melhor como funciona a modelação dimensional de um *data warehouse*. Foi feita uma modelação dimensional para o processo de negócio que foi criado na etapa anterior, através da determinação do grão da tabela de factos, da modelação de dimensões adequadas e da determinação de medidas numéricas da tabela de factos.

Anexos:

A. Script Java para correção do ficheiro *audiencias.csv*:

```
import java.io.BufferedReader;
import java.io.File;
import java.io.FileReader;
import java.io.FileWriter;
import java.io.IOException;
import java.nio.file.Files;
public class Main {
    public static void main(String[] args) throws IOException {
            File audiencias = new
File("C:/Users/Gonçalo/Downloads/audiencias.csv");
            BufferedReader br = new BufferedReader(new
FileReader(audiencias));
            File fAux = new File("C:/Users/Gonçalo/Downloads/tmp.csv");
            FileWriter aux = new FileWriter(fAux, true);
            String linha;
            int i = 0;
            while ((linha = br.readLine()) != null) {
                    String[] param = linha.split(",");
                    String dataInicio = param[4];
                    String dataFim = param[5];
                    if (dataInicio.substring(1, dataInicio.length()-1).length() < 19) {
                            dataInicio = dataInicio.substring(0, dataInicio.length()-1) +
" 00:00:00#";
                    }if(dataFim.substring(1, dataFim.length()-1).length() < 19){
                            dataFim = dataFim.substring(0, dataFim.length()-1) + "
00:00:00#";
                    }
                    aux.write(param[0] + "," + param[1] + "," + param[2] + "," +
param[3] + "," + dataInicio + "," + dataFim + "\n");
                    i++;
                    System.out.println("i: " + i);
            }
            br.close();
            aux.close();
            Files.delete(audiencias.toPath());
            fAux.renameTo(new File("audiencias.csv"));
    }
}
```