



Licenciatura em Ciência de Dados - 3º ano

FilmData Insights: Inteligência Analítica Aplicada ao Cinema

Sistemas de Informação Analíticos

25 de maio de 2025

Discente: João Dias nº 110305

2024/2025

ÍNDICE

1. Título	1
2. Objetivo	1
3. Descrição da Empresa e do Negócio	1
4. Definição de Objetivos e Requisitos	3
4.1 Objetivos de Negócio	3
4.2 Requisitos Funcionais	4
4.3 Requisitos Técnicos	4
4.4 Requisitos de Dados	4
5. Análise e Planeamento	5
5.1 Sistemas e fontes de informação	5
5.1.1 Entidade: Filme	5
5.1.2 Entidade: Avaliação	6
5.1.3 Entidade: Avaliador	7
5.2 Plano de projeto	7
6. Modelação de dados	10
7. Seleção e extração de features	12
7.1 Fontes de Dados Estruturadas do IMDB	12
7.2 Integração de Dados Não Estruturados: Scraping de Comentários do IMDB	13
8. Limpeza e Transformação e Carregamento de Dados (ETL)	16
8.1 ETL	16
8.2 Carregamento de Dados no Data Warehouse	19
9. Implementação de Ferramentas de BI e Analítica	19
10. Qualidade de dados	25
11. Implementação e Deployment	27
12. Manutenção e Evolução	27

ÍNDICE DE FÍGURAS

Figura 1 - Diagrama Entidade-Relação	5
Figura 2 - Model view da solução implementada no Power BI	11
Figura 3 - Vista 1 (Reviews e Avaliações) do dashboard em Power BI	21
Figura 4 - Vista 2 (Reviews e Avaliações) do dashboard em Power BI	22
Figura 5 - Vista 3 (Reviews e Avaliações) do dashboard em Power BI	23
Figura 6 - Vista 1 (Perfis dos Avaliadores) do dashboard em Power BI	24
Figura 7 - Vista 2 (Perfis dos Avaliadores) do dashboard em Power BI	25
Figura 8 - Análise de missing values e resumos estatísticos dos CSVs de oscares (à esquerda) e de filmes (à direita)	26
Figura 9 - Análise de missing values e resumos estatísticos dos CSVs de tipos de filme (esq.) e de avaliadores (dir.)	26
Figura 10 - Análise de missing values e resumos estatísticos do CSVs de avaliações de filmes	
	27

ÍNDICE DE TABELAS

Tabela 1 - Atributos da entidade de Filme	6
Tabela 2 - Atributos da entidade de Avaliação	6
Tabela 3 - Atributos da entidade de Avaliador	7
Tabela 4 - Número de reviews extraídas por cada filme	15

1. Título

FilmData Insights: Inteligência Analítica Aplicada ao Cinema

2. Objetivo

O objetivo deste projeto é desenvolver e implementar um Data Warehouse especializado na análise de dados do cinema, utilizando dados do IMDB, que incluem informação sobre **filmes, prémios Óscar e avaliações**. O sistema será projetado para armazenar, processar e explorar grandes volumes de dados, permitindo análises detalhadas e extração de insights relevantes, tais como:

- Identificação de **padrões de sucesso** (relações entre prémios Óscar e classificações dos filmes);
- Análise de **tendências cinematográficas** (evolução da popularidade de diferentes géneros e estilos ao longo do tempo);
- Relação entre **realizadores e sucesso** dos filmes (impacto dos realizadores no desempenho dos filmes em termos de avaliação e prémios);

Além das avaliações numéricas já disponíveis nos datasets do IMDB, o projeto incluirá **scraping de comentários, também do IMDB**, para recolher reviews textuais sobre os filmes. Através da aplicação de técnicas de Text Mining e Processamento de Linguagem Natural (NLP), será possível identificar **temas recorrentes, padrões de opinião e análise de polaridade e de sentimento**, enriquecendo a análise com uma perspetiva mais qualitativa. Estes insights permitirão compreender melhor as preferências do público e da crítica, auxiliando a indústria cinematográfica na tomada de decisões estratégicas sobre produção e distribuição de filmes.

O sistema integrará um **processo ETL** robusto, garantindo a qualidade, consistência e performance das consultas analíticas. Além disso, serão utilizadas ferramentas de **Business Intelligence (BI)** para criar **dashboards interativos** e relatórios dinâmicos que facilitem a visualização dos insights e a tomada de decisões estratégicas.

3. Descrição da Empresa e do Negócio

A FilmData Insights é uma empresa especializada em **análise avançada de dados do setor cinematográfico**, focando-se na identificação de padrões e tendências na indústria do entretenimento. Com recurso a Data Warehouses,

técnicas avançadas de Business Intelligence (BI) e Machine Learning, a FilmData Insights ajuda os seus clientes a compreender melhor as dinâmicas do mercado cinematográfico e a tomar decisões estratégicas fundamentadas em dados.

A FilmData Insights **opera no setor do entretenimento e análise de mercado**, fornecendo insights detalhados e personalizados para:

- **Estúdios de Cinema** → Otimização de estratégias de lançamento e marketing, identificando géneros e formatos mais atrativos para diferentes audiências.
- **Plataformas de Streaming** → Análise de padrões de consumo e recomendações personalizadas para maximizar a retenção de utilizadores e, consequentemente, maximizar o Lifetime Value dos clientes.
- **Críticos e Analistas de Mercado** → Ferramentas de BI e dashboards interativos para avaliar o impacto de filmes ao longo do tempo e prever tendências futuras.
- **Organizadores de Prémios e Festivais** → Modelos preditivos para identificar potenciais candidatos a prémios e fatores que influenciam o reconhecimento crítico.

Distingue-se pela sua capacidade de transformar grandes volumes de dados em insights açãoáveis, ajudando os seus clientes a adaptar-se à constante evolução do setor cinematográfico. A empresa utiliza dados estruturados do IMDB e dos Óscars, complementados por análises qualitativas de dados não estruturados de avaliações do público através de Text Mining. O seu portefólio de serviços inclui:

- **Análises preditivas de sucesso de filmes** → Modelos estatísticos para prever o impacto de um filme com base em padrões históricos de avaliação e prémios.
- **Identificação de tendências cinematográficas** → Evolução dos géneros, estilos de produção e receção crítica ao longo do tempo.
- **Análise da influência de realizadores e produtores** → Como a escolha de profissionais impacta a performance dos filmes em termos de audiência e reconhecimento crítico.
- **Correlação entre prémios e sucesso comercial** → Estudo detalhado sobre como os prémios influenciam a bilheteira e a popularidade dos filmes.
- **Sentiment Analysis de avaliações do público** → Aplicação de Processamento de Linguagem Natural (NLP) para extrair padrões emocionais e percepções qualitativas.

- **Benchmarking e estratégias de lançamento** → Comparação entre diferentes abordagens de distribuição, marketing e posicionamento de filmes.

A FilmData Insights posiciona-se como uma referência na análise de dados cinematográficos, permitindo aos seus clientes:

- **Antecipar tendências** → Compreender mudanças nas preferências do público e evolução dos formatos de produção.
- **Maximizar impacto de lançamentos** → Ajustar estratégias de marketing e distribuição para otimizar o sucesso dos filmes.
- **Apoiar decisões estratégicas** → Fornecer insights açãoáveis para estúdios, plataformas de streaming e críticos.

4. Definição de Objetivos e Requisitos

Antes da implementação do Data Warehouse da FilmData Insights, é essencial definir os **objetivos de negócio, requisitos funcionais, técnicos** e de **dados**. O foco principal é garantir que o sistema seja robusto, eficiente e capaz de suportar consultas analíticas complexas adaptadas aos diferentes stakeholders da indústria cinematográfica.

4.1 Objetivos de Negócio

O Data Warehouse da FilmData Insights tem como principais objetivos:

- **Melhorar a análise e previsão de tendências cinematográficas** → Identificação de padrões de sucesso com base em classificações, prémios e avaliação do público.
- **Compreender o impacto dos realizadores** → Avaliação do desempenho de filmes e identificação de tendências de sucesso associadas à direção de cada filme.
- **Estudar a relação entre prémios e desempenho comercial** → Correlacionar o impacto de distinções como os Óscares com a popularidade e bilheteira dos filmes.
- **Comparação entre diferentes fontes de avaliação** → Identificar discrepâncias entre a avaliação quantitativa e os comentários textuais do IMDB.
- **Enriquecimento da análise com dados textuais sobre os filmes** → Extrair padrões de opinião e sentimentos dos comentários do IMDB.

- **Relatórios e dashboards interativos** → Criar visualizações dinâmicas para apoiar a tomada de decisões de estúdios e plataformas de streaming.

4.2 Requisitos Funcionais

Os requisitos funcionais estabelecem as capacidades essenciais do sistema:

- **Extração e integração de dados** → Recolher e consolidar dados estruturados do IMDB e prémios Óscar, além de dados não estruturados via scraping de comentários, com recurso aos principais métodos de Text Mining e Processamento de Linguagem Natural.
- **Processamento ETL (Extração, Transformação e Carregamento)** → Limpeza, normalização e enriquecimento dos dados para garantir coerência e qualidade.
- **Armazenamento e modelação de dados** → Implementação de um esquema dimensional otimizado para consultas analíticas.
- **Disponibilização de dados para análise e visualização** → Integração com ferramentas de BI para construção de dashboards e relatórios interativos.

4.3 Requisitos Técnicos

A arquitetura do Data Warehouse será projetada para garantir performance, escalabilidade e integração eficiente.

- Tecnologia de Armazenamento → Implementação de um Data Warehouse dimensional, garantindo rapidez nas consultas analíticas.
- Pipeline ETL → Desenvolvimento de um processo automatizado para extrair, transformar e carregar dados estruturados e não estruturados.
- Infraestrutura e Ferramentas:
 - Business Intelligence: Power BI para criação de relatórios interativos.
 - Machine Learning & NLP: Utilização de bibliotecas como SpaCy, NLTK e modelos da Hugging Face para processamento de linguagem natural.

4.4 Requisitos de Dados

O Data Warehouse será construído a partir de diferentes fontes de dados, garantindo a diversidade e riqueza da análise.

- IMDB (FILMES.csv, FILMES_AVALIACAO.csv, TIPO_FILME.csv, AVALIADOR.csv) → Dados estruturados sobre filmes, classificações, géneros e avaliações.

- Óscars (OSCAR.csv) → Informação com a descrição dos prémios Óscar.
- Reviews Textuais (Scraping) → Recolha de comentários e avaliações textuais para análise.

Cada conjunto de dados será submetido a processos de limpeza, transformação e normalização antes de ser integrado no Data Warehouse.

5. Análise e Planeamento

5.1 Sistemas e fontes de informação

A Figura 1 mostra o diagrama entidade-relação através que vai modelar as relações entre filmes, avaliadores e avaliações, permitindo a estruturação de dados para análise detalhada da receção crítica e popular dos filmes. A estrutura do modelo é composta pelas seguintes entidades Filme, Avaliador e Avaliação.

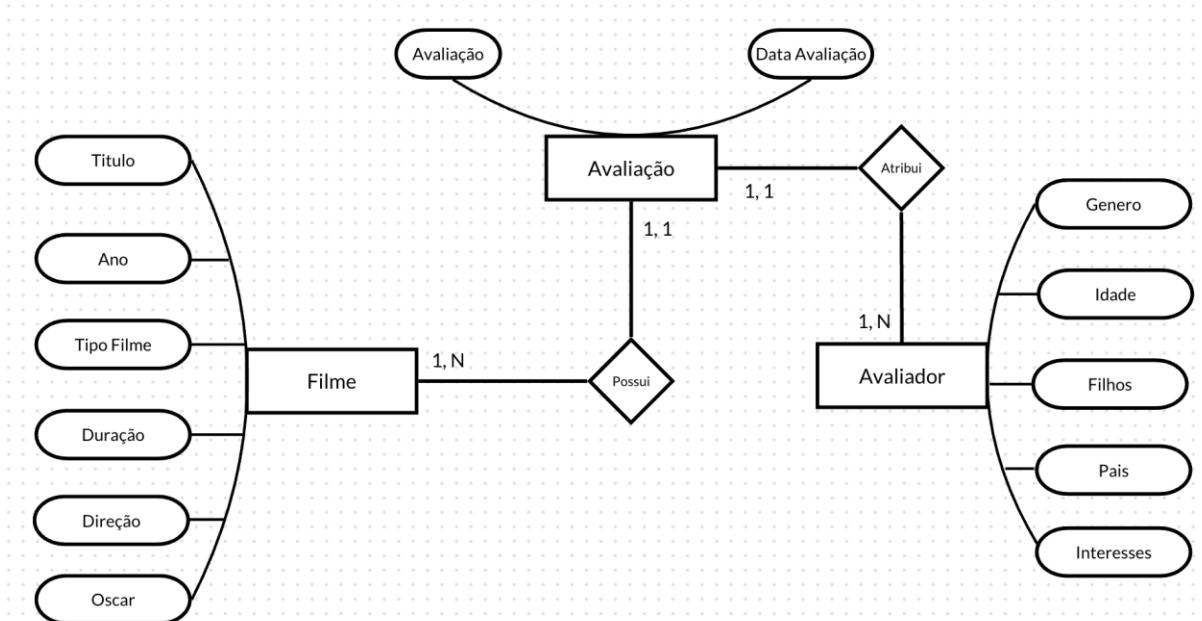


Figura 1 - Diagrama Entidade-Relação

5.1.1 Entidade: Filme

A entidade Filme representa cada obra cinematográfica contida no sistema e os seus atributos estão representados na Tabela 1.

Atributo	Descrição
Título	Nome do filme
Ano	Ano de lançamento
Tipo de Filme	Categoria ou género associado ao filme
Duração	Tempo de exibição em minutos
Direção	Nome do realizador do filme
Óscar	Indicação ou vitória em prémios da Academia

Tabela 1 - Atributos da entidade de Filme

Cada filme pode ter múltiplas avaliações, estabelecendo uma relação um para muitos (1, N) com a entidade Avaliação, descrita na secção seguinte.

5.1.2 Entidade: Avaliação

A entidade Avaliação representa as notas quantitativas numa escala de 1 a 5, atribuídas pelos avaliadores a cada filme. Os seus atributos encontram-se descritos na Tabela 2.

Atributo	Descrição
Avaliação	Classificação atribuída ao filme
Data da avaliação	Momento em que a avaliação foi registada

Tabela 2 - Atributos da entidade de Avaliação

Cada Avaliação está sempre associada a um único filme e um único avaliador, estabelecendo relações um para um (1, 1) com ambas as entidades. É importante reforçar que como se está a trabalhar com dados do IMDB e a plataforma apresenta no seu website, atualmente, as avaliações quantitativas numa escala numérica compreendida entre 1 e 10, será necessário fazer o rescaling das classificações presentes na base de dados para esta nova escala.

5.1.3 Entidade: Avaliador

A entidade Avaliador representa os utilizadores que atribuem classificações aos filmes e os seus atributos encontram-se descritos na Tabela 3.

Atributo	Descrição
Género	Identificação de género do avaliador
Idade	Faixa etária do utilizador
Filhos	Indica se o avaliador tem filhos, um fator que pode influenciar as preferências cinematográficas
País	Nacionalidade do avaliador
Interesses	Tipos de filmes preferidos pelo utilizador

Tabela 3 - Atributos da entidade de Avaliador

Cada avaliador pode realizar múltiplas avaliações, estabelecendo uma relação um para muitos (1:N) com a entidade Avaliação.

5.2 Plano de projeto

Para garantir uma implementação eficiente, foi delineado um plano de projeto estruturado constituído por várias etapas. Este plano abrange desde a recolha e preparação dos dados até a construção do Data Warehouse e o desenvolvimento de dashboards analíticos. O objetivo é criar um sistema robusto que permita a extração de insights relevantes sobre a indústria cinematográfica, facilitando a análise de tendências e a previsão de sucesso de filmes.

1. Recolha de Dados:

- Integração de dados com múltiplas estruturas, incluindo:
 - **Bases de dados estruturadas** do IMDB.
 - **Scraping de Reviews** do IMDB para recolha de dados textuais sobre as opiniões do público.

- Uniformização e padronização dos diferentes formatos de dados para garantir coerência na análise.

2. Processamento e Transformação de Dados:

- Implementação de uma pipeline **ETL** para:
 - **Remover valores nulos, duplicados e erros de formatação.**
 - **Normalizar variáveis** como categorias de filmes e escalas de avaliação.
 - **Converter texto de comentários em métricas estruturadas** através de **Processamento de Linguagem Natural (NLP).**
- Validação da **qualidade dos dados** antes da carga no Data Warehouse.

3. Construção do Data Warehouse:

- Criação de uma base de dados com estrutura otimizada para **consultas analíticas eficientes.**
- Implementação de **procedimentos de indexação e partição de dados** para otimizar o desempenho.

4. Desenvolvimento de Dashboards e Relatórios Analíticos:

- Construção de **dashboards interativos em Power BI** com métricas-chave:
 - Evolução da popularidade dos filmes ao longo do tempo.
 - Relação entre prémios e sucesso comercial.
 - Comparação entre avaliações do público e da crítica especializada.

Resultados:

1. **Plataforma unificada** que centraliza e organiza dados do setor cinematográfico, permitindo **visualizações detalhadas e exploração de tendências.**
2. **Capacidade melhorada de análise**, oferecendo insights estratégicos para **estúdios de cinema, plataformas de streaming e críticos.**
3. **Dashboards interativos e relatórios dinâmicos**, possibilitando uma **tomada de decisão informada baseada em dados concretos.**

Este projeto consolida o trabalho desenvolvido pela **FilmData Insights** como uma ferramenta essencial para análise e previsão de tendências no setor do entretenimento, ajudando a indústria a **maximizar o impacto dos lançamentos e compreender melhor as preferências do público.**

6. Modelação de dados

A modelação de dados constitui uma fase crítica do projeto, já que é nesta etapa que se define o **esquema do Data Warehouse**, garantindo uma estrutura robusta e eficiente que permite análises rápidas e de qualidade. Para este projeto, a modelação de dados segue a **abordagem dimensional**, optando-se por um **esquema estrela** devido à sua **simplicidade** e ao seu **desempenho otimizado** para **consultas analíticas**.

O **Data Warehouse** foi implementado em **MySQL (com recurso à interface do Workbench)**, permitindo um controlo completo sobre a **estruturação, integridade e carregamento** dos dados. Por sua vez, a camada de exploração e visualização foi desenvolvida em Power BI, tirando partido da ligação direta ao Data Warehouse para construir dashboards interativos e relatórios dinâmicos.

O modelo de dados centra-se numa **tabela de factos**, que regista as **avaliações numéricas** atribuídas aos **filmes**. Esta tabela inclui identificadores únicos para cada avaliação, **chaves estrangeiras** para ligação às **tabelas dimensionais**, a nota atribuída e a data da avaliação. A sua estrutura foi cuidadosamente desenhada para permitir análises detalhadas, tais como a **evolução da receção crítica** ao longo do tempo, a **comparação entre diferentes categorias de filmes** e a identificação de **padrões de avaliação** por parte dos utilizadores.

As **tabelas dimensionais** são responsáveis por descrever o contexto em que cada avaliação foi realizada. Destacam-se as seguintes:

- A **tabela filmes** contém atributos descritivos sobre os filmes, incluindo título, ano de lançamento, duração (em minutos), nome do(s) realizador(es), presença ou não de indicação ao Óscar, e o tipo de filme (género), que é referenciado através de uma chave estrangeira.
- A **tabela avaliador** descreve os utilizadores que efetuam as avaliações, com atributos como género, idade, indicação de se têm filhos, cidade e interesses, permitindo análises mais segmentadas e personalizadas.
- A **tabela tipo_filme** armazena a descrição dos géneros dos filmes, proporcionando uma categorização clara e simples para análise.
- A **tabela imdb_comments** recolhe comentários textuais sobre os filmes, com a análise de sentimento associada. Esta tabela, embora não esteja diretamente

ligada à tabela de factos filmes_avaliacao, liga-se à tabela filmes, permitindo uma integração complementar de dados qualitativos e quantitativos.

Este modelo permite consultas analíticas rápidas e eficientes, suportando uma variedade de análises. Além disso, foram implementadas boas práticas na modelação, incluindo a **definição explícita de chaves primárias e estrangeiras**, garantindo a **integridade referencial** entre as tabelas e facilitando a **escrita de queries SQL otimizadas**. O diagrama estrela apresentado na Figura 2, ilustrando as **relações entre as tabelas** e destacando as ligações fundamentais do sistema.

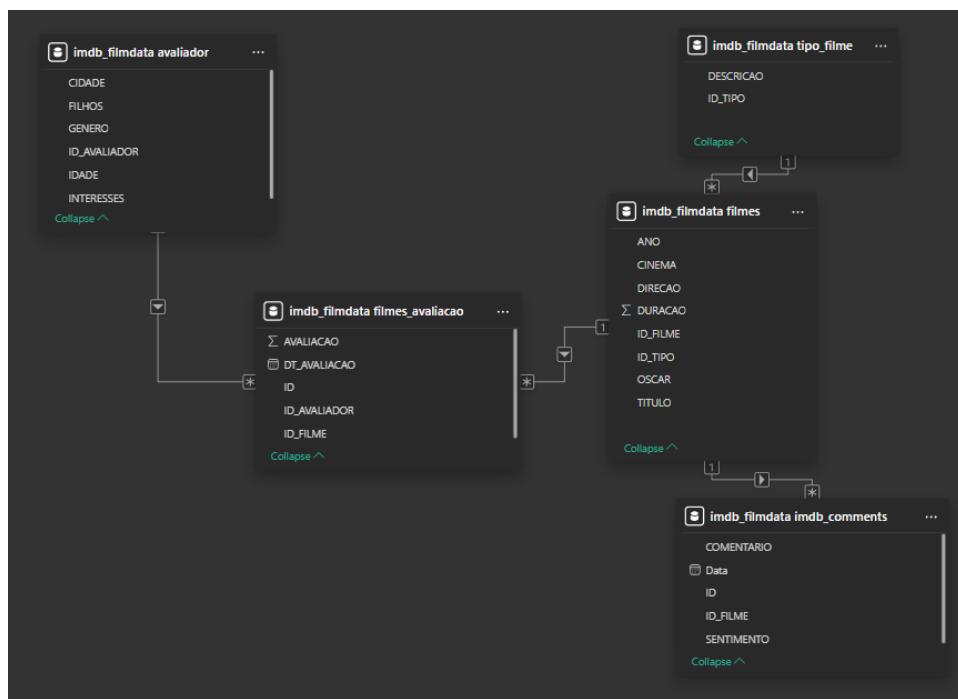


Figura 2 - Model view da solução implementada no Power BI

Para assegurar a consistência entre o modelo lógico e a implementação na base de dados, foram definidos os seguintes **Data Definition Language (DDL)** no **sistema MySQL**:

- Criação de **tabelas** com **tipos de dados apropriados** para cada **atributo**, garantindo a qualidade e coerência da informação.
- Definição de **restrições** como **chaves primárias** e **chaves estrangeiras**, essenciais para preservar a **integridade do modelo**.
- **Estruturação** clara e **escalável**, preparada para **suportar** a **introdução** de **novos dados** e a realização de futuras análises.

Este **modelo dimensional** não só garante a **integridade** e a **performance** necessárias para suportar o **volume** e a **complexidade** dos dados do projeto, como também se integra facilmente com ferramentas analíticas como o **Power BI**, possibilitando uma **exploração completa** e detalhada do **universo cinematográfico** que se pretende estudar.

7. Seleção e extração de features

7.1 Fontes de Dados Estruturadas do IMDB

A base deste projeto assenta em **dados estruturados** extraídos de **conjuntos** provenientes do **IMDB**, permitindo a criação de um **modelo dimensional** sólido e coerente. Estes dados foram obtidos a partir de **ficheiros CSV** contendo informações detalhadas sobre **filmes, óscares, avaliações, categorias e perfis de avaliadores**, todos fundamentais para o desenvolvimento do Data Warehouse.

As **fontes estruturadas** utilizadas baseiam-se, essencialmente, no processamento de cinco ficheiros CSV, entre os quais:

- **FILMES.csv**: Contém **10 filmes** e **detalhes descritivos sobre cada um**, incluindo título, ano de lançamento, duração, realizador, tipo (género) e presença de indicação ou prémio Óscar. Este ficheiro serviu como base para a construção da dimensão filmes no modelo, detalhada na secção 6.
- **FILMES_AVALIACAO.csv**: Apresenta **9.999 observações** e regista as **avaliações numéricas** atribuídas a cada filme por parte dos avaliadores, com atributos como nota e data de avaliação, constituindo a tabela de factos principal do modelo.
- **TIPO_FILME.csv**: Fornece a descrição de **7 categorias ou géneros cinematográficos**, permitindo a categorização dos filmes e facilitando análises por tipo.
- **AVALIADOR.csv**: Descreve o **perfil dos 1.033 utilizadores** que realizaram as avaliações, com atributos como género, idade, filhos, cidade e interesses, possibilitando análises segmentadas e personalizadas.
- **OSCAR.csv**: Contém informações sobre **prémios e nomeações de filmes**. Esta tabela identifica cada distinção, abrangendo categorias como "Melhor Filme", "Melhor Ator(a)", "Melhor Realização", entre outras.

A **escolha das features** contidas nestes ficheiros foi guiada pela sua **relevância** para o **modelo analítico** e para os **objetivos de negócio** definidos no projeto, nas secções iniciais deste relatório. A combinação das variáveis extraídas destas fontes permite construir uma **visão integrada e multidimensional da indústria cinematográfica**, explorando desde as **características** intrínsecas de cada **filme**, passando pelas **preferências e padrões de comportamento** dos **avaliadores**, até ao **impacto** de **distinções** e **prémios de prestígio**.

Esta seleção assegura que o modelo é capaz de responder a perguntas de negócio complexas, como a **identificação de géneros mais populares**, a avaliação do **impacto dos prémios no sucesso dos filmes** ou a **deteção de relações entre perfis de utilizadores e avaliações** atribuídas. Além disso, a diversidade das features selecionadas permite ao modelo captar tanto a dimensão objetiva (dados quantitativos e estruturados) como a dimensão descritiva (categorias e perfis), proporcionando uma base sólida para a exploração de insights e a construção de relatórios interativos no Power BI.

7.2 Integração de Dados Não Estruturados: Scraping de Comentários do IMDB

Para **complementar as fontes de dados estruturadas**, descritas na secção 7.1, e enriquecer o modelo analítico com uma **perspetiva qualitativa**, foi realizada a **extração de dados não estruturados** provenientes dos **comentários** publicados pelos utilizadores na plataforma **IMDB**. Esta abordagem teve como objetivo ampliar a análise, **capturando as percepções reais do público** e permitindo a aplicação de técnicas de **text mining** e **análise de sentimento** sobre os filmes considerados no projeto.

A **origem** desta informação foi exclusivamente o **IMDB**, através das páginas públicas de reviews, correspondentes à lista de filmes previamente definida e que se encontra contemplada na base de dados (na tabela de filmes). Foram selecionadas páginas específicas para cada filme, identificadas pelas respetivas URLs, incluindo títulos como The Father, Soul, The Trial of the Chicago 7, Bad Boys for Life, Enola Holmes, Homem-Aranha: Sem Volta para Casa, Cruella, Shang-Chi and the Legend of the Ten Rings, A Quiet Place Part II e The Black Phone. Estas páginas disponibilizam um conjunto significativo de comentários de utilizadores, constituindo uma fonte rica e direta de dados não estruturados relacionados com os filmes analisados.

A implementação do **processo de extração** foi realizada em **Python**, utilizando a biblioteca **selenium**, responsável por **automatizar a navegação** nas páginas e permitir a **recolha completa dos comentários**. A estrutura do processo seguiu os seguintes passos:

1. **Configuração do ambiente de scraping:** O **scraper** começa por **inicializar um navegador Chrome**, configurado com opções para garantir estabilidade e desempenho, preparado para **navegar automaticamente** sem intervenção manual.
2. **Acesso às páginas de cada filme:** Com as URLs da lista acima referenciada, o navegador foi dirigido automaticamente à página de reviews correspondente a cada filme. Esta ligação garantiu o acesso a todas as reviews públicas disponibilizadas pelo IMDB para todos os filmes contemplados na Base de Dados.
3. **Interação automatizada com a página:** O **scraper** foi programado para **identificar e clicar no botão “All”** (quando disponível), expandindo a visualização para incluir todos os comentários disponíveis. Esta ação foi essencial para garantir a **extração completa** dos dados, mesmo em casos em que o conteúdo estava inicialmente truncado ou limitado a um subconjunto de reviews.
4. **Scroll automático até ao fundo da página:** Para carregar a totalidade do conteúdo, o script executa **scrolls automáticos** até ao final da página. Esta ação repetitiva é controlada através da **monitorização da altura da página**, assegurando que o carregamento está concluído antes de avançar para a recolha dos dados.
5. **Extração dos comentários e metadados:** Após o carregamento completo da página, o script procedeu à **identificação e recolha** dos **elementos HTML** correspondentes aos comentários. A prioridade foi dada à extração do corpo completo do comentário, complementado pelo título sempre que disponível, assegurando a captura do conteúdo mais representativo e informativo. Nos casos em que o IMDB oculta o corpo do comentário devido à **presença de spoilers**, foi realizada apenas a recolha do **título da review** como **alternativa**, garantindo que, mesmo nesses casos, o **feedback do utilizador não era desconsiderado**. A **data da publicação** de cada review foi igualmente extraída, permitindo associar o comentário ao respetivo filme e incluir uma **perspetiva temporal** na análise.

6. **Encerramento do navegador e consolidação dos dados:** Após a extração de todas as reviews para todos os filmes, o **navegador** foi **encerrado** automaticamente e os dados foram **estruturados** num **formato tabular**, pronto para integração com o modelo analítico e posterior aplicação de **técnicas** de **text mining** e **análise de sentimento**.

A escolha destas features e a implementação deste processo permitiram **enriquecer** significativamente o **modelo** com uma **camada qualitativa**, ampliando a capacidade analítica para além das métricas quantitativas. Esta estratégia posiciona o sistema como uma solução abrangente e enriquecida, capaz de fornecer uma compreensão profunda e detalhada da receção crítica e popular dos filmes analisados.

Após a implementação do processo de scraping, foram extraídas **20.851 reviews** associadas aos filmes analisados. Cada filme teve a sua respetiva página de reviews no IMDB processada, permitindo recolher o conteúdo textual das opiniões do público e associá-lo ao filme correspondente. A distribuição do número de reviews extraídas por filme encontra-se resumida na Tabela 4.

Filme	Número de reviews
A Quiet Place Part II	1.840
Bad Boys for Life	1.281
Cruella	2.050
Enola Holmes	1.759
Homem-Aranha: Sem Volta para Casa	6.172
Shang-Chi and the Legend of the Ten Rings	2.546
Soul	1.850
The Black Phone	1.561
The Father	1.113
The Trial of the Chicago 7	679

Tabela 4 - Número de reviews extraídas por cada filme

8. Limpeza e Transformação e Carregamento de Dados (ETL)

8.1 ETL

Para garantir uma leitura robusta das bases de dados do IMDB, foi definida uma **função de leitura** dos ficheiros providenciados pelo IMDB, que tenta inicialmente **ler os arquivos com codificação UTF-8** e, quando não é possível, recorre a **Latin-1**. Esta abordagem assegura a compatibilidade com diferentes formatos de ficheiros e evita perdas de dados devido a erros de codificação. Cada ficheiro foi **carregado para um DataFrame** de pandas, **armazenado num dicionário** para facilitar a manipulação subsequente.

A fase de transformação consistiu numa série de **operações** realizadas sobre os dados extraídos, com o objetivo de **garantir consistência, coerência e qualidade** antes do carregamento no Data Warehouse. Esta etapa foi implementada em **Python**, com o uso extensivo das bibliotecas **pandas** e a pipeline dos **transformers**. As principais transformações aplicadas foram:

1. Conversão e normalização de datas

- As **colunas** de data foram convertidas para o **formato datetime** de pandas. Foi aplicado um tratamento específico para garantir que as **datas inválidas** ou em **formatos inconsistentes** fossem **detetadas e tratadas**.
- Esta normalização assegura que as análises temporais subsequentes sejam consistentes e precisas e a assegura o correto carregamento das datas no Data Warehouse.

2. Pipeline de análise de sentimento dos comentários

- A análise de sentimento foi implementada para enriquecer o conjunto de dados das reviews textuais, que contém os comentários recolhidos via scraping das páginas do IMDB. O objetivo foi **identificar a polaridade dos comentários** e classificá-los como **Positivo, Neutro ou Negativo**, fornecendo uma **camada qualitativa** essencial ao modelo analítico.
- A implementação foi realizada com recurso à pipeline de **sentiment-analysis** do modelo **cardiffnlp/twitter-roberta-base-sentiment**, especializado na análise de sentimentos em texto, desenvolvido pelo **Research Group in Natural Language Processing** da **Cardiff University**, um modelo **RoBERTa-base** treinado em **~58M tweets** e

afinado para **classificação de sentimento com 3 classes** (Negativo, Neutro, Positivo) através do **TweetEval benchmark**¹. O processo seguiu estes passos:

- A pipeline foi inicializada com **return_all_scores=True**, permitindo a **obtenção** não apenas do sentimento principal, mas também dos **scores de probabilidade para cada classe de sentimento** (Negativo, Neutro, Positivo).
- Foi criado um **dicionário de mapeamento** entre os labels originais do modelo (LABEL_0, LABEL_1, LABEL_2) e os sentimentos reais correspondentes: Negativo, Neutro e Positivo.
- Para cada comentário no campo COMENTARIO, o modelo **analisou** até aos **512 caracteres iniciais** (por limitação do modelo) e produziu uma **lista de scores para cada label**. O código identificou a **label com a maior pontuação** utilizando a função `max(resultado, key=lambda x: x['score'])`, garantindo que o sentimento principal era determinado de forma objetiva.
- Para cada comentário, além de determinar o sentimento principal, o **código gerou um log detalhado**, que incluiu:
 - a) O texto inicial do comentário.
 - b) O sentimento principal identificado (Negativo, Neutro ou Positivo).
 - c) Os scores individuais de cada label, apresentados como percentagens (ex. Negativo: 85,5%, Neutro: 10,2%, Positivo: 4,3%).
 - d) Em caso de erro, o log registava a mensagem de exceção e a identificação do comentário.
- O **progresso da classificação** foi visualizado através da biblioteca **tqdm**, que apresentou uma barra de progresso enquanto os comentários eram processados, essencial para **acompanhar o estado do processamento** dado o volume elevado de dados.
- O resultado final foi a criação de um **novo atributo de Sentimento** no DataFrame dos comentários, atribuindo a cada comentário a classificação resultante da análise. Esta nova feature permitiu, posteriormente, a **análise cruzada entre percepção do público**.

¹ Para mais informações sobre o modelo consultar: <https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment>

(qualitativa) e métricas quantitativas, como o volume de comentários e as notas atribuídas aos filmes.

3. Padronização e Renomeação de Colunas:

- Foi realizada uma junção entre a tabela de **filmes** e a tabela de **óscares**, com o objetivo de **integrar a descrição dos óscars diretamente na tabela de filmes**.
- As **variáveis** foram **renomeadas** nas várias tabelas de forma a facilitar a interpretação e o **alinhamento na nomenclatura de atributos presentes em várias tabelas**.

4. Rescaling da Avaliação:

- A coluna que contém as avaliações dos filmes, originalmente expressa numa escala de 1 a 5, foi **ajustada para uma nova escala mais granular e detalhada**, com valores compreendidos entre 1 e 10, que reflete a **amplitude de valores utilizados pelo IMDB nas suas plataformas**. O rescaling foi realizado através de uma fórmula que amplia a amplitude das avaliações e que se pode ver na Equação 1. Além disso, esta transformação procurou expandir as diferenças entre avaliações próximas, destacando variações nas opiniões do público e facilitando comparações mais precisas entre filmes. Este ajuste contribuiu para uma melhor visualização e interpretação dos dados no sistema analítico, tornando possível a **identificação mais clara de padrões de receção crítica entre os diferentes filmes e géneros**.

$$Avaliação (A) = A(nova)_{min} + (A(original) - A(original)_{min}) \times \frac{A(nova)_{max} - A(nova)_{min}}{A(original)_{max} - A(original)_{min}}$$

onde:

$A(original)$ é a avaliação original (escala inicial, 1 a 5)

$A(original)_{min}$ e $A(original)_{max}$ são os valores mínimo e máximo da escala original (1 e 5 respectivamente)

$A(nova)_{min}$ e $A(nova)_{max}$ são os valores mínimo e máximo da nova escala (1 e 10 respectivamente)

Equação 1 - Rescaling das avaliações

8.2 Carregamento de Dados no Data Warehouse

Após a fase de extração e transformação, os dados preparados foram **carregados** no **Data Warehouse**, implementado numa **base de dados MySQL**. O carregamento foi efetuado com o objetivo de garantir a **integridade, coerência e disponibilidade** dos dados para análise posterior em ferramentas de Business Intelligence, como o **Power BI**. O processo de carregamento foi **implementado** em **Python**, utilizando a biblioteca **sqlalchemy** para estabelecer uma **ligação segura e eficiente** com o **MySQL**.

Com a ligação estabelecida, foi implementada a **criação automática das tabelas** no **MySQL**, utilizando instruções CREATE TABLE IF NOT EXISTS com a **definição explícita** dos **tipos de dados** para cada coluna. Este passo garantiu que a estrutura do Data Warehouse refletia o modelo analítico dimensional concebido na fase anterior.

Posteriormente, os dados foram **carregados** para as **respetivas tabelas** através do método **to_sql**, especificando o modo if_exists='append' para **adicionar os dados sem substituir os registos já existentes**. Esta estratégia assegurou a preservação da **integridade dos dados** e permitiu a **atualização incremental das tabelas**, suportando volumes de dados elevados. O processo de carregamento incluiu:

1. **Escrita sequencial** dos **dados transformados** para cada tabela.
2. **Verificação do sucesso** de cada operação de carregamento, com registo de mensagens de confirmação para cada tabela.
3. **Gestão automática** de exceções para deteção e correção de potenciais erros durante a execução, assegurando a robustez do processo.

Este processo garantiu que os dados limpos e enriquecidos foram disponibilizados no Data Warehouse de forma completa e consistente. A utilização da biblioteca **sqlalchemy** proporcionou uma **integração segura e eficiente** entre o **ambiente de desenvolvimento Python** e o **sistema de base de dados MySQL**, facilitando a **manutenção e atualização futuras** dos dados.

9. Implementação de Ferramentas de BI e Analítica

Com o Data Warehouse em plena operação, foi implementada uma camada de Business Intelligence (BI) e analítica, concebida para **transformar os dados**

estruturados e enriquecidos em **insights acionáveis**. Esta camada foi desenvolvida utilizando o **Power BI**, uma ferramenta robusta e versátil que permite a criação de relatórios dinâmicos e dashboards interativos.

A configuração do ambiente de BI teve como base a **ligação direta ao Data Warehouse em MySQL**, estabelecendo uma **conexão segura e eficiente** para a **extração de dados em tempo real**. Esta ligação permitiu o **acesso completo às tabelas dimensionais e de factos criadas na fase de ETL**, garantindo que os utilizadores podem consultar e explorar os dados de forma intuitiva e sem necessidade de replicações manuais.

No Power BI foram desenvolvidos **dashboards e relatórios com visualizações interativas**, incluindo:

- Análises detalhadas por tipologia de filme, com gráficos que ilustram a distribuição das avaliações e a evolução da receção crítica ao longo do tempo.
- Comparação entre géneros cinematográficos e padrões de avaliação, permitindo identificar tendências e preferências do público.
- Integração da análise de sentimentos dos comentários, com visualizações que destacam a polaridade das opiniões expressas pelo público, complementando as métricas quantitativas.
- Exploração do impacto das óscares, relacionando a obtenção de prémios com a avaliação média e o volume de comentários.

A implementação foi pensada para permitir aos utilizadores finais explorar livremente os dados, com possibilidade de **criação de filtros, drill-downs e visualizações personalizadas adaptadas** a diferentes perspetivas de negócio. Além disso, a integração do Power BI com o MySQL assegura que os dashboards estão sempre atualizados com os dados mais recentes, suportando a tomada de decisões informadas e a identificação de insights estratégicos.

Este ambiente de BI e analítica, baseado na arquitetura desenvolvida, permite à organização obter uma compreensão profunda do desempenho dos filmes e das tendências de comportamento do público, transformando dados brutos em informação de valor para suportar decisões de negócio.

A Figura 3 ilustra a vista 1 da secção de “Reviews e Avaliações” do dashboard desenvolvido no Power BI. Esta visualização destaca os **indicadores chave de**

desempenho (número de filmes, duração média, número total de reviews e avaliação média), permitindo uma **análise agregada e imediata**. Em complemento, a **nuvem de palavras** apresenta uma **representação visual das palavras mais frequentes** nos comentários extraídos do IMDB, oferecendo uma perspetiva qualitativa sobre a percepção do público.

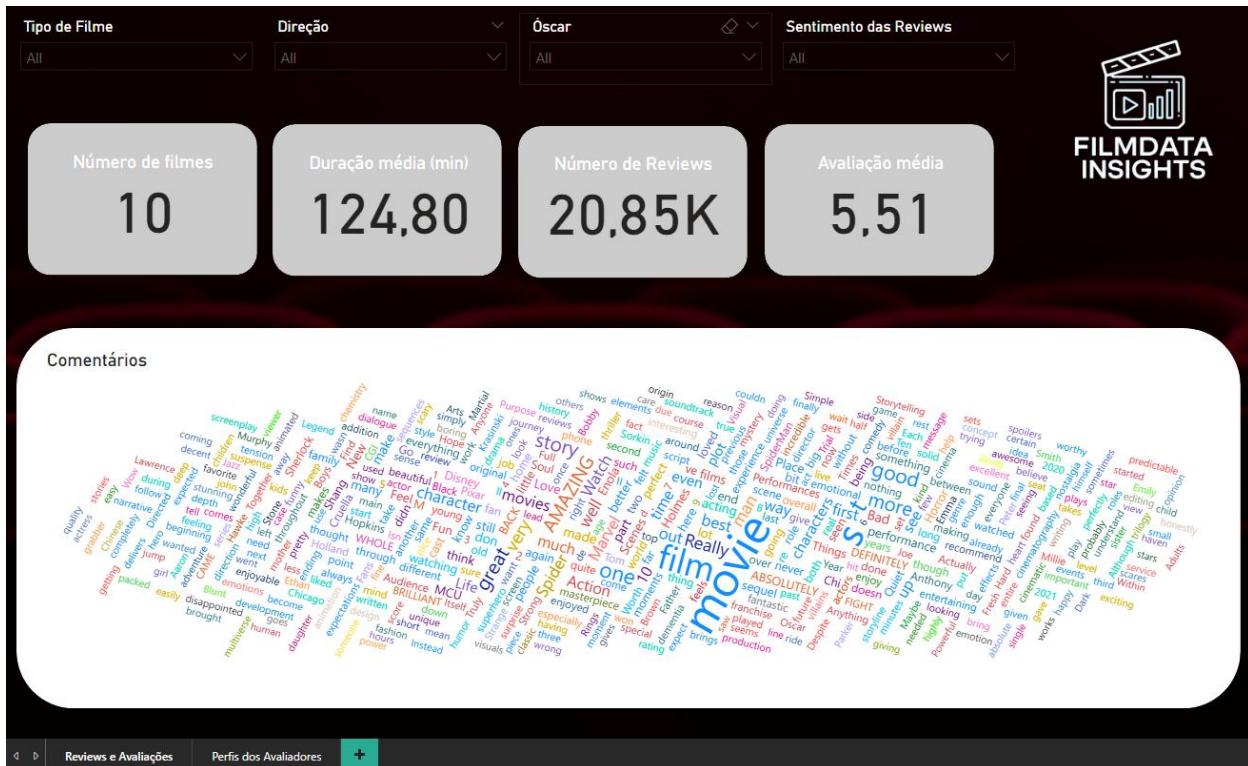


Figura 3 - Vista 1 (Reviews e Avaliações) do dashboard em Power BI

A Figura 4 apresenta uma segunda vista da secção “Reviews e Avaliações” do dashboard desenvolvido no Power BI, destacando a **evolução do número de reviews segmentadas por sentimento ao longo do tempo**, a **distribuição de sentimento** por tipologia de filme e a **relação entre Óscars** e a **percepção** do público. É importante destacar que há um **alinhamento entre as reviews positivas e os óscars**, e que os **filmes que não têm óscars são também os que apresentam o maior volume de reviews negativas**.

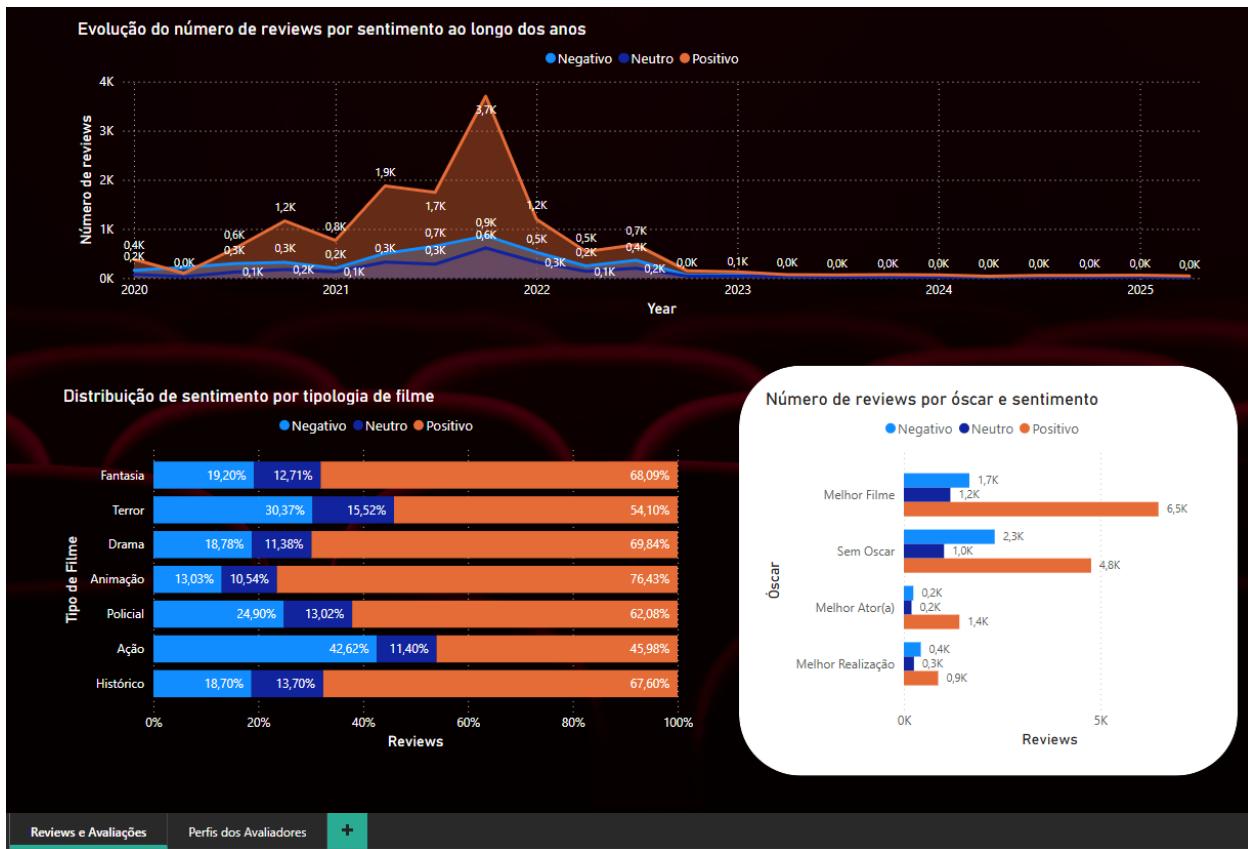


Figura 4 - Vista 2 (Reviews e Avaliações) do dashboard em Power BI

A Figura 5 apresenta a terceira vista do dashboard desenvolvido no Power BI, evidenciando a **relação** entre o **total de reviews** e a **avaliação média por tipologia de filme**, assim como uma **listagem detalhada das avaliações médias, número de reviews e duração** dos **filmes analisados**. Esta visualização **combina análise quantitativa e textual**, com destaque para a **inclusão de exemplos reais de uma tabela com os comentários extraídos**, permitindo compreender as razões por detrás das classificações atribuídas. A distribuição dos pontos no gráfico revela tendências por género, destacando, por exemplo, a concentração de reviews em géneros como Fantasia e Terror, e permitindo observar variações na popularidade e receção crítica dos filmes.

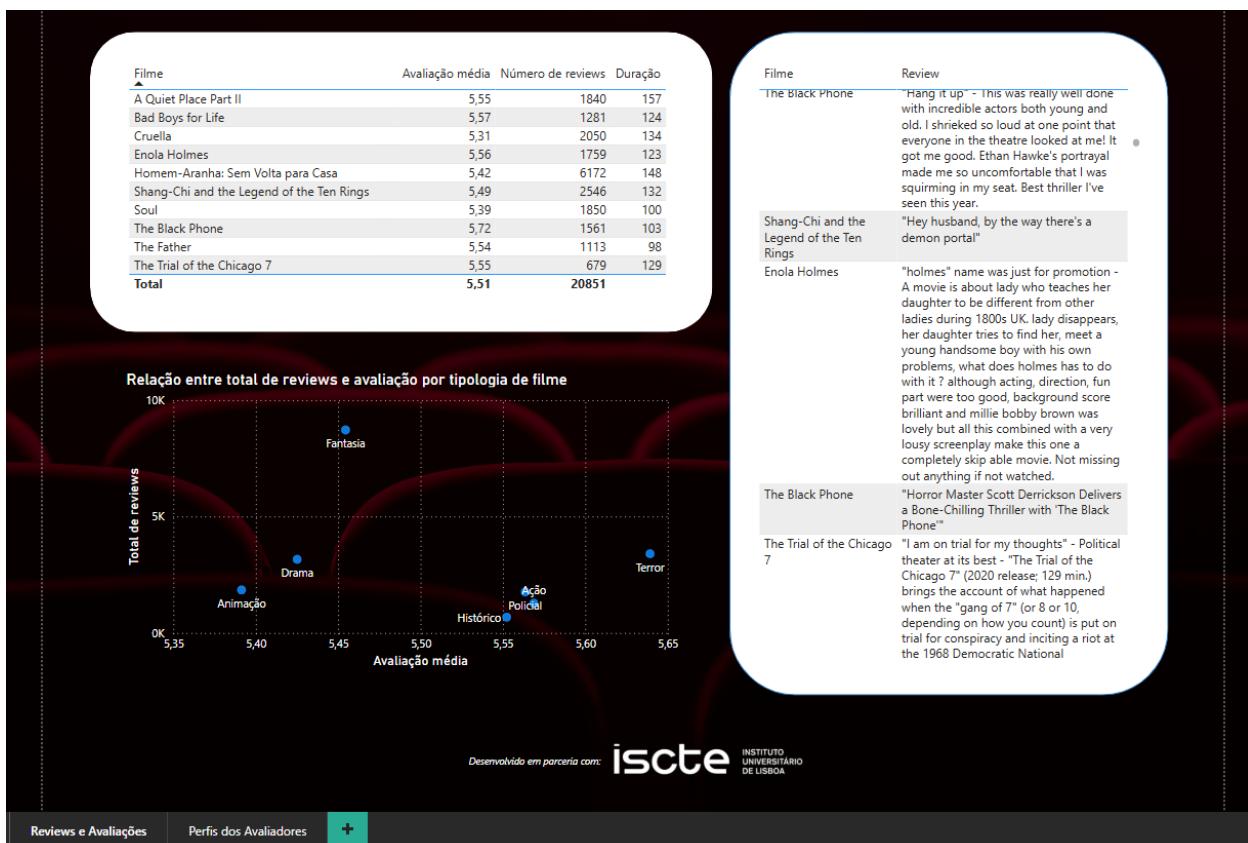


Figura 5 - Vista 3 (Reviews e Avaliações) do dashboard em Power BI

A Figura 6 apresenta a vista 1 da secção de “Perfis dos Avaliadores” do dashboard desenvolvido no Power BI, destacando a **distribuição dos avaliadores por cidade, a média global das avaliações, e a segmentação por género e existência de filhos**. O gráfico de barras exibe de forma clara a **concentração de avaliadores em determinadas cidades**, enquanto os gráficos circulares ilustram a distribuição entre **avaliadores com filhos e sem filhos**, e entre os **géneros masculino e feminino**. Esta visualização permite compreender a **composição demográfica e comportamental dos utilizadores** que realizaram avaliações, fornecendo uma camada adicional de contexto para as análises exploratórias e permitindo cruzar padrões de avaliação com perfis específicos de avaliadores.

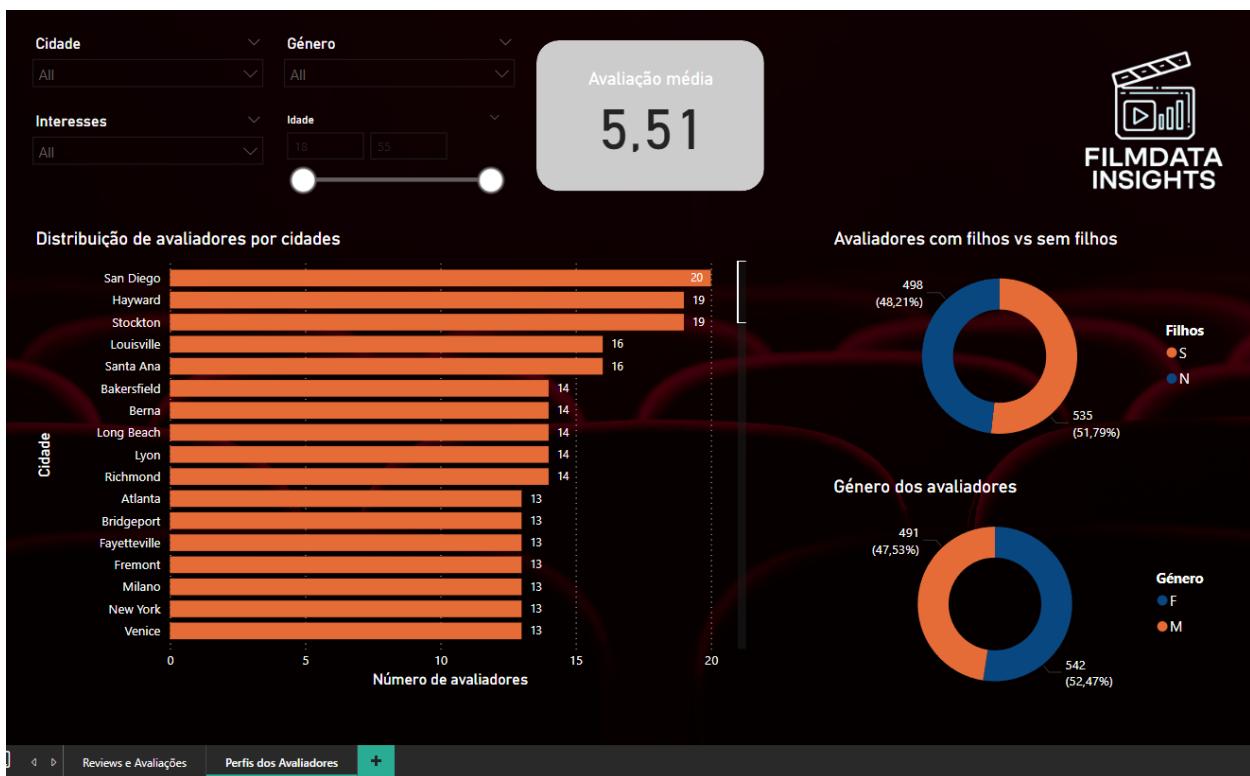


Figura 6 - Vista 1 (Perfis dos Avaliadores) do dashboard em Power BI

A Figura 7 revela uma evolução da avaliação média dos filmes ao longo do tempo, onde é possível identificar alguma **sazonalidade**. Os gráficos de barras mostram que **utilizadores sem filhos tendem a dar avaliações mais baixas**, e que interesses como **Ação-Policial** recebem as **classificações mais elevadas (5,7)** no **universo feminino**, em contraste com categorias como **Terror-Fantasia**, com **médias mais baixas (5,4)**. A tabela lateral destaca as cidades que atribuem as **avaliações mais altas**, como **Detroit (6,31)** e **New Orleans (6,16)**, complementando a análise com uma **perspetiva geográfica à receção dos filmes**.

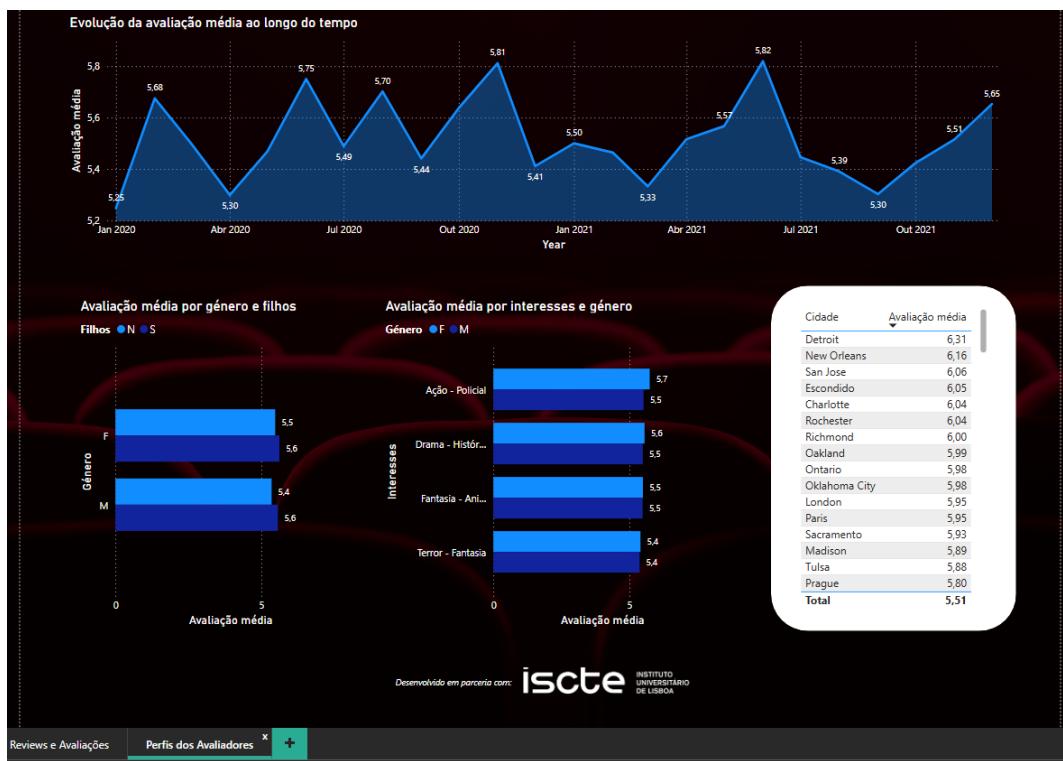


Figura 7 - Vista 2 (Perfis dos Avaliadores) do dashboard em Power BI

10. Qualidade de dados

Antes de se proceder às transformações, à construção do ETL e, naturalmente, à colocação do Data Warehouse em produção, foram realizados **testes sistemáticos de qualidade de dados** para garantir a **integridade, coerência e completude** dos dados extraídos. Esta etapa preliminar foi essencial para identificar potenciais problemas nos dados de origem e orientar o processo subsequente de **preparação e carregamento**. As verificações incluíram:

- Análise de valores nulos (missing values):** Para cada tabela de dados estruturados (detalhadas na secção 7.1), foi calculada a contagem de valores ausentes em cada variável, permitindo identificar possíveis lacunas. Depois de executados estes testes, foi possível concluir que **não existem valores em falta na base de dados estruturada**.
- Resumos estatísticos descritivos:** Foram geradas estatísticas como média, desvio padrão, valores mínimo e máximo, quartis e contagem de valores únicos para todas as variáveis numéricas e categóricas. Esta análise permitiu garantir

que **não existem outliers** e verificar a **coerência das distribuições iniciais**, antes da realização de quaisquer transformações.

3. **Deteção de anomalias e inconsistências:** Com base nas estatísticas geradas, foram identificadas potenciais inconsistências como o **desalinhamento entre as escalas das avaliações contidas na Base de Dados e os valores apresentados, por norma, na plataforma web do IMDB**.
4. **Verificação de conformidade dos formatos e tipos de dados:** Foram ainda **validados os tipos e formatos de todas as colunas**, para garantir que a leitura inicial estava alinhada com o modelo dimensional a implementar.

As figuras 8, 9 e 10 ilustram os resumos estatísticos e as análises iniciais realizadas, suportando assim os pontos anteriores supracitados e serviram como base para identificar e corrigir potenciais anomalias antes da execução das transformações, da construção do processo ETL e da colocação do Data Warehouse em produção.

Missing Values:													
ID	0												
TITULO	0												
ANO	0												
ID_TIPO	0												
DURACAO	0												
DIRECAO	0												
CINEMA	0												
ID OSCAR	0												
dtype: int64													
Resumo estatístico:													
	count	unique	top	freq	mean	std	min	25%	50%	75%	50%	75%	max
ID OSCAR	6.0	NaN	Nan	NaN	2.5	1.870829	0.0	1.25	2.5	3.75	5.0		
DESCRICAO	6	6	Sem Oscar	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	3.0

Missing Values:													
ID OSCAR	0												
DESCRICAO	0												
dtype: int64													
Resumo estatístico:													
	count	unique	top	freq	mean	std	min	25%	50%	75%	50%	75%	max
ID OSCAR	6.0	NaN	Nan	NaN	2.5	1.870829	0.0	1.25	2.5	3.75	5.0		
DESCRICAO	6	6	Sem Oscar	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	3.0

Figura 8 - Análise de missing values e resumos estatísticos dos CSVs de oscares (à esquerda) e de filmes (à direita)

Missing Values:													
ID	0												
GENERO	0												
IDADE	0												
FILHOS	0												
PAIS	0												
INTERESSES	0												
dtype: int64													
Resumo estatístico:													
	count	unique	top	freq	mean	std	min	25%	50%	75%	50%	75%	max
ID	1033.0	NaN	Nan	NaN	717.0	298.345717	201.0	459.0	717.0	975.0	1233.0		
GENERO	1033	2		F	542	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
IDADE	1033.0	NaN	Nan	NaN	37.31365	10.916188	18.0	28.0	38.0	47.0	55.0		
FILHOS	1033	2		S	535	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
PAIS	1033	107		San Diego	20	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
INTERESSES	1033	4	Fantasia - Animacão	268	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

Figura 9 - Análise de missing values e resumos estatísticos dos CSVs de tipos de filme (esq.) e de avaliadores (dir.)

Missing Values:											
ID_FILME	0										
ID_AVALIADOR	0										
AVALIACAO	0										
DT_AVALIACAO	0										
dtype:	int64										
Resumo estatístico:											
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
ID_FILME	9999.0	NaN	NaN	NaN	5.49615	2.875902	1.0	3.0	5.0	8.0	10.0
ID_AVALIADOR	9999.0	NaN	NaN	NaN	700.627863	288.970747	201.0	450.5	700.0	950.0	1233.0
AVALIACAO	9999.0	NaN	NaN	NaN	3.005301	1.407149	1.0	2.0	3.0	4.0	5.0
DT_AVALIACAO	9999	731	15/09/2020	26	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figura 10 - Análise de missing values e resumos estatísticos do CSVs de avaliações de filmes

11. Implementação e Deployment

Após a **conclusão dos testes de qualidade** e a **validação do processo ETL**, o sistema foi preparado para ser **implementado** e **colocado em produção**. Esta fase foi essencial para garantir a transição do ambiente de desenvolvimento para um ambiente operacional, assegurando que todas as funcionalidades estivessem prontas para utilização pelos utilizadores finais.

A implementação envolveu a **migração dos dados transformados e validados para o Data Warehouse**, localizado numa **base de dados MySQL**, garantindo que todas as tabelas e relações estivessem corretamente configuradas e que os dados estivessem acessíveis para exploração e análise.

Em paralelo, foram **configurados os dashboards e relatórios no Power BI, com ligações diretas ao Data Warehouse**, permitindo a atualização automática dos dados e a disponibilização de insights em tempo real. Esta configuração assegurou a integração perfeita entre o backend (MySQL) e a camada de apresentação (Power BI), permitindo que os utilizadores pudessem aceder a análises detalhadas e interativas de forma intuitiva.

Adicionalmente, procedeu-se à **compilação deste relatório técnico**, que **documenta detalhadamente todos os procedimentos, decisões e resultados** obtidos durante o projeto, servindo como **referência para futuras atualizações e manutenções** do sistema.

12. Manutenção e Evolução

A utilidade e relevância de um projeto de Data Warehousing dependem de um ciclo contínuo de manutenção e evolução, que assegure que o sistema se mantém

alinhado com as necessidades do negócio, com a integridade dos dados e com o desempenho exigido pelos utilizadores finais.

No contexto específico deste projeto, a manutenção do Data Warehouse implementado em MySQL envolverá a **atualização regular das tabelas com novas extrações de dados do IMDB e de reviews**, garantindo que o sistema reflete a **evolução do mercado cinematográfico e a percepção do público ao longo do tempo**. Isto incluirá a **reexecução periódica do processo de ETL**, com **verificações de qualidade adicionais** para assegurar que não existem incoerências ou valores anormais nos dados mais recentes. A **monitorização da performance do ambiente Power BI** será igualmente essencial, para garantir que a experiência do utilizador se mantém fluida mesmo com o **crescimento do volume de dados**.

A evolução do sistema poderá incluir a **integração de novas fontes de dados** externas, como **outras bases cinematográficas, plataformas de streaming**, ou até **fontes complementares de sentimento** (por exemplo, **redes sociais e portais de crítica**), para enriquecer ainda mais as análises disponíveis. Poderá também passar pela adaptação do modelo dimensional, **ajustando a granularidade das tabelas de factos** ou **incluindo novas dimensões** que permitam análises mais profundas e segmentadas, como por exemplo detalhes adicionais sobre **categorias de filmes, tendências de bilheteira**, ou ainda mais **preferências demográficas**.

Do ponto de vista técnico, a evolução incluirá ainda a **atualização de bibliotecas e pipelines** utilizados no **processo ETL**, garantindo compatibilidade com **novas versões de Python, sqlalchemy, pandas** ou **transformers**. Será também relevante avaliar a necessidade de **migrar** para uma **infraestrutura mais escalável**, como um **serviço cloud** de base de dados, caso o **volume e complexidade dos dados aumentem consideravelmente**.

Por fim, a **formação contínua dos utilizadores** e a **revisão dos dashboards** no Power BI terão um papel fundamental na manutenção e evolução do projeto. À medida que **novas funcionalidades** sejam introduzidas, será necessário **atualizar os dashboards** para refletir as alterações, garantindo que os **relatórios se mantêm relevantes e úteis** para a tomada de decisões. A **documentação técnica**, como este relatório, deve ser também **continuamente atualizada** para refletir as **alterações e adaptações** realizadas no **sistema**.

Desta forma, a manutenção e evolução do Data Warehouse e do ambiente analítico desenvolvido não só asseguram a longevidade do projeto, como também garantem a sua capacidade de adaptação às dinâmicas do setor cinematográfico e às necessidades de negócio identificadas, consolidando a solução como uma ferramenta essencial para a análise e compreensão do universo de filmes e avaliações.