Knime

Integração de Sistemas de Informação

13579 – Rui Costa

Licenciatura em Engenharia de Sistemas Informáticos

3ºano

Barcelos | Outubro, 2024

Lista de Abreviaturas e Siglas

ETL - Extração, Transformação e Carregamento de dados

IPCA – Instituto Politécnico do Cávado e do Ave

API – Application Programming Interface

1. Índice

2. Índice de Figuras

3. Enquadramento

Este projeto faz parte da disciplina de Integração de Sistemas de Informação, pertencente ao curso de Licenciatura em Engenharia de Sistemas Informáticos. A proposta central é aplicar as técnicas de ETL em cenários práticos que envolvem a manipulação e integração de dados provenientes de várias fontes, explorando o seu uso em grandes volumes de informação.

Para este projeto, foi utilizado um ficheiro JSON contendo dados diversos sobre filmes, como orçamento (Budget), receitas de bilheteira (Box Office), títulos, datas de lançamento e os atores principais. O objetivo foi criar um fluxo de técnicas ETL que fosse capaz de extrair essas informações, filtrar e normalizar os dados mais relevantes e organizá-los de forma clara. Os resultados foram apresentados em gráficos na aplicação e também com a possibilidade de serem extraídos em formato de imagem, permitindo visualizar comparações, como a o top do budget e a quantidade de filmes lançados por ano.

Este projeto ilustrou como o uso de ferramentas de ETL pode simplificar o processo de tratamento e análise de dados em larga escala, desde a sua extração e transformação até à representação visual, facilitando a interpretação e tornando o processo mais eficiente.

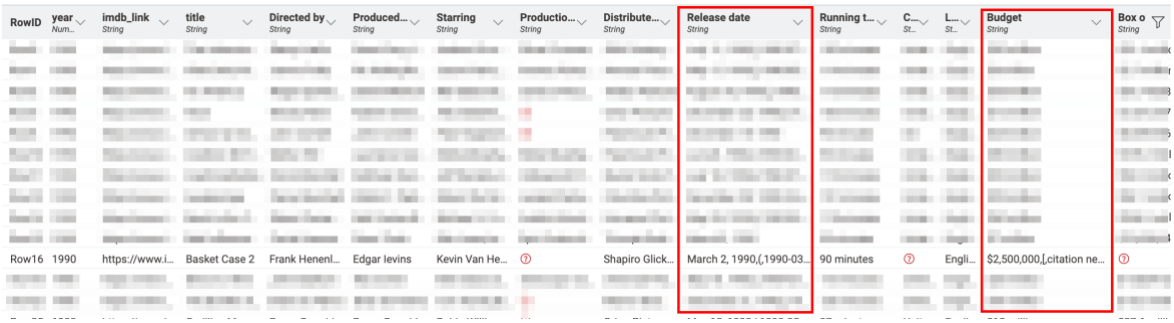
O trabalho foi desenvolvido utilizando a plataforma **KNIME**, uma ferramenta de ETL que permite a criação de fluxos de trabalho de forma intuitiva e visual. Com o uso do KNIME, foi possível automatizar a integração e processamento de dados complexos, transformando-os em outputs que foram analisados e apresentados por meio de gráficos e relatórios.

4. Problema

O desafio inicial enfrentado neste projeto foi a importação inicial de um ficheiro JSON que continha dados sobre filmes. Devido à falta de familiaridade com o formato e as ferramentas apropriadas, a primeira tentativa de trabalhar com o ficheiro resultou em dificuldades. Para contornar esta situação, o ficheiro foi convertido para o formato CSV, utilizando ferramentas online que acabou por ser o utilizado ao longo do projeto. Esta conversão facilitou o tratamento dos dados dentro do KNIME.

O que se pretende demonstrar com este trabalho é a experiência adquirida ao utilizar o KNIME para realizar operações de ETL. Durante a execução, encontrou-se um obstáculo adicional: o conjunto de dados apresentava uma grande quantidade de falhas em quase todas as colunas. Muitos dos dados não estavam normalizados, o que impedia a criação de relatórios e gráficos com eficiência. Assim, o foco do trabalho foi a normalização das colunas essenciais, de forma a torná-las consistentes para análise.

O objetivo final foi conseguir transformar estes dados desestruturados em informações úteis, aplicando técnicas de limpeza e normalização de dados no KNIME, de modo a gerar gráficos e relatórios visuais que pudessem refletir corretamente as tendências e relações contidas nos dados.



RowID	year	imdb_link	title	Directed by	Produced...	Starring	Productio...	Distribute...	Release date	Running t...	C...	L...	Budget	Box o
	Num...	String	String	String	String	String	String	String	String	String	St...	St...	String	String
Row16	1990	https://www.i...	Basket Case 2	Frank Henenl...	Edgar levins	Kevin Van He...	?	Shapiro Glick...	March 2, 1990,(,1990-03...	90 minutes	?	Engli...	$2,500,000,[,citation ne...	?

Figura 1 - Exemplo Dados Iniciais

Na *Fig. 1* é possível analisar as colunas que foram normalizadas neste trabalho prático numa fase inicial. Sendo que a grande maioria dos dados estava assim representado.

5. Estratégia Utilizada

A abordagem para resolver o problema focou-se na criação de um fluxo de trabalho estruturado no KNIME, utilizando operadores e processos que permitiram a limpeza, normalização e organização dos dados. Abaixo estão os passos principais e os nodes utilizados ao longo do processo.

1. **Leitura e Filtragem dos Dados**: O primeiro passo foi utilizar o node **CSV Reader** para importar o ficheiro CSV que continha os dados convertidos do JSON original. Em seguida, aplicou-se o node **Row Filter** para eliminar dados desnecessários, mantendo apenas as informações relevantes para a análise.

2. **Normalização de Dados com Expressões Regulares**: Para tratar inconsistências nos dados textuais, o node **String Replacer** foi utilizado. Este operador permitiu a substituição de determinados padrões utilizando expressões regulares, ajudando a uniformizar e corrigir os valores nas colunas selecionadas.

3. **Transformações Específicas**: Além do **String Replacer**, foi necessário recorrer ao node **Expression** para executar substituições de palavras e organizar os dados de acordo com critérios específicos.

4. **Junção de Dados**: Após a normalização, foi necessário combinar os dados tratados em diferentes colunas, o que foi feito com o node **Joiner**. Este operador permitiu unir colunas tratadas, consolidando a informação para a visualização final.

5. **Ordenação e Seleção Final**: Para concluir o fluxo de trabalho, aplicou-se o node **Sorter** para ordenar os dados conforme necessário e, finalmente, um **Row Filter** adicional para selecionar e apresentar apenas os dados essenciais nas visualizações criadas.

6. **Visualização:** De modo a apresentar os dados obtidos com uma fácil leitura foi utiizado o node **Generic Echart View** para criar gráficos dinâmicos e interativos dos dados. No entanto, este método não permitia a exportação direta dos gráficos, o que limitou a sua utilização.

7. **Exportação de Gráficos**: Esta exportação foi realizada com os nodes **Python View** e **Image Writer**: Para contornar esta limitação, recorreu-se ao Python View para criar os gráficos personalizados e exportá-los em formato .png conectando-o ao Image Writer. facilitando a partilha e armazenamento dos resultados. Finalmente, foi implementado o node **Send**

Email para automatizar o envio das imagens exportadas por email, permitindo uma distribuição prática e rápida dos gráficos criados.

6. Jobs/Transformações

Será feita uma explicação detalhada do processo com recurso a imagens dos diagramas, de modo a ilustrar cada etapa envolvida. Estes diagramas ajudam a compreender a sequência de operações realizadas, desde a importação dos dados até à criação e envio dos gráficos finais, permitindo uma visualização clara de todo o fluxo de trabalho.

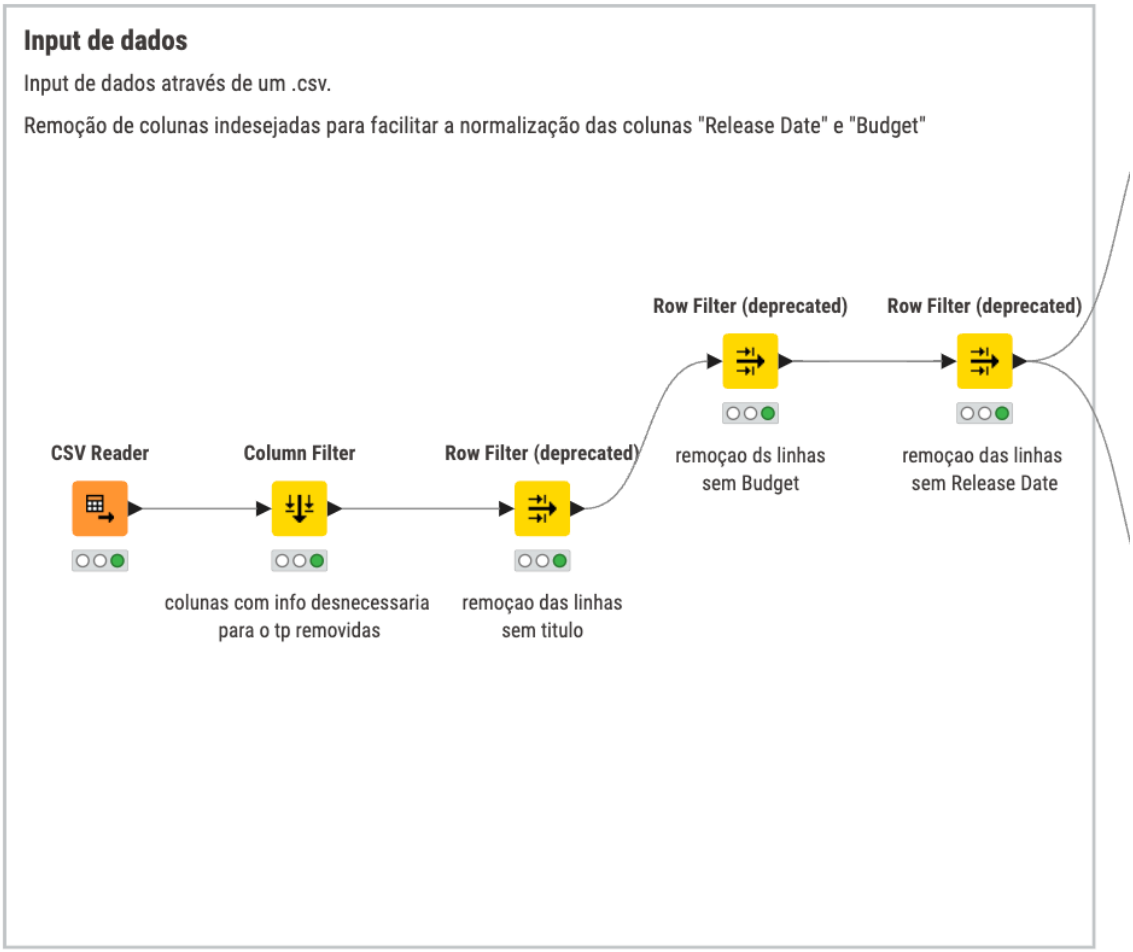


Figura 2 - Input Dados

Na imagem acima podemos verificar que o node “CSV Reader” irá ler o ficheiro, cujo tem um caminho relativo, e foram feitas filtragens de colunas de modo a preparar os dados para a sua

normalização, sendo que foram removidas colunas que não tinham interesse futuro. Nas colunas a serem tratadas foram removidas as linhas que não continham informação com objetivo de não atrapalhar o tratamento. No decorrer desta fase foi explorada a possibilidade de retirar essas linhas e exportá-las para um ficheiro para serem atualizadas futuramente, não avançou pois não teria impacto no resultado pretendido.

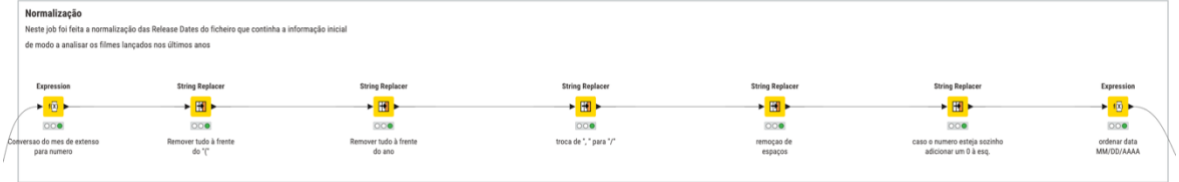


Figura 3 Normalização "Release Date"

Na *fig. 3* podemos verificar os passos seguidos. Na *fig. 1* podemos analisar o estado inicial dos dados.

Os passos para a normalização das datas foram os seguintes:

1. Criar uma expressão para substituir o nome dos meses que estavam escritos por extenso pelo seu respetivo número.
2. Usar o String Replacer para remover tudo na frente do "(", todas as utilizações deste node foram realizadas através de expressões regulares.
3. Remover tudo na frente do ano, detetando os primeiros 4 números seguidos e substituindo tudo na frente por " ".
4. Com os dados quase normalizados foi realizada a remoção dos espaços.
5. Como havia datas, por exemplo, 04/5/1990 foi adicionado um 0 à esquerda dos números isolados para padronizar a data.
6. Para finalizar foi utilizada uma expressão para organizar as datas por MM/DD/AAAA

Esta foi a metodologia utilizada para resolver o problema de apresentação dos dados na coluna "Release Date".

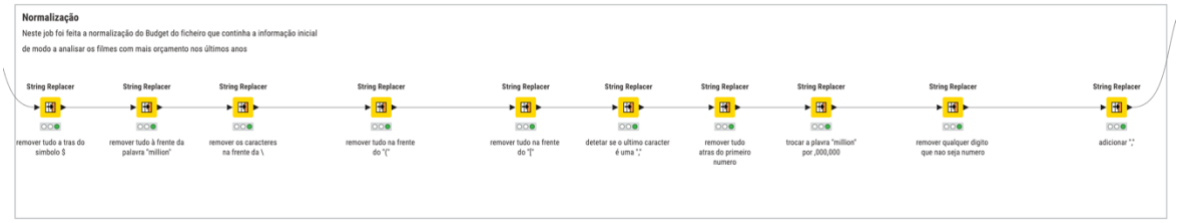


Figura 4 – Normalização "Budget"

Na *Fig. 4*, podemos observar o processo de normalização aplicado à coluna "Budget". Este fluxo foi construído para garantir que os dados financeiros fossem padronizados e ficassem consistentes para análise. Abaixo estão os passos seguidos durante o processo:

1. Utilizou-se o **String Replacer** para remover todos os caracteres antes do símbolo "$", de modo a eliminar qualquer informação desnecessária.

2. A seguir, foi removido tudo à frente da palavra "million", utilizando novamente o **String Replacer**, para manter apenas o valor numérico relevante.

3. Continuou-se a limpeza ao remover todos os caracteres após a barra invertida "\", eliminando partes do texto que não eram necessárias para a análise.

4. Com outro **String Replacer**, removeu-se tudo à frente do símbolo "(", seguindo-se o mesmo procedimento para o símbolo "[", mantendo apenas a informação relevante.

5. Verificou-se se o último carácter era uma vírgula, para remover eventuais separadores que pudessem prejudicar a normalização dos dados.

6. Foi aplicado um **String Replacer** para eliminar todos os caracteres que estivessem antes do primeiro número encontrado, garantindo que apenas o valor numérico fosse mantido.

7. A palavra "million" foi substituída por ",000,000" para padronizar os valores financeiros, convertendo-os em números completos.

8. Por fim, foi removido qualquer dígito que não fosse numérico e adicionou-se uma vírgula onde necessário para garantir a consistência na apresentação dos números.

Este processo de normalização do "Budget" foi essencial para assegurar que os dados estivessem formatados corretamente, permitindo uma análise precisa dos filmes com maior orçamento ao longo dos últimos anos.

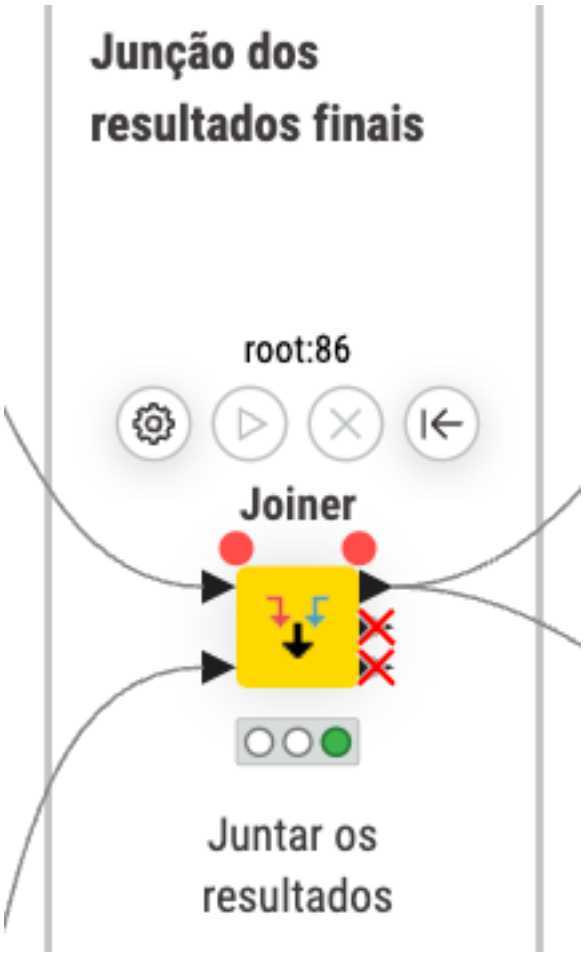


Figura 5 – Joiner

O node Joiner foi utilizado para realizar a junção dos resultados obtidos após o tratamento das colunas. Neste processo, as colunas que foram previamente normalizadas foram combinadas, permitindo a criação de um conjunto de dados consolidado.

Esta junção foi essencial para unir a informação tratada, garantindo que os diferentes aspetos dos dados, como o "Budget" e a "Release Date", pudessem ser analisados em conjunto, proporcionando uma visão mais ampla e completa do conjunto de dados final.

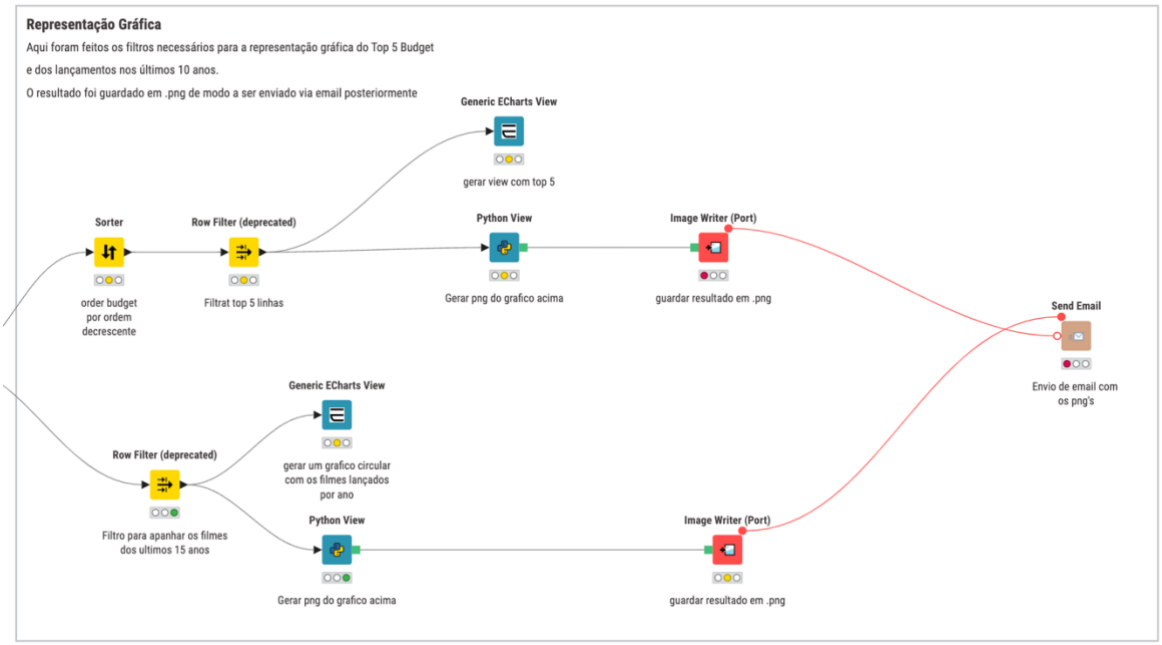


Figura 6 - Representação Gráfica

Para apresentar os dados de forma clara e compreensível, foi utilizada uma combinação de ferramentas que permitiu tanto a visualização interativa quanto a exportação dos resultados.

Inicialmente foram organizados os dados, e, de seguida, recorreu-se ao **Generic Echart View**, que possibilitou a criação de gráficos dinâmicos e interativos diretamente no KNIME, oferecendo uma visão inicial do comportamento dos dados. No entanto, como esta abordagem não permitia a exportação dos gráficos para outros formatos, optou-se por complementar o processo com o **Python View**. Este node foi utilizado para gerar gráficos personalizados e exportá-los em formato de imagem (.png), garantindo a flexibilidade necessária para partilhar e armazenar os resultados. Depois de criados, os gráficos foram guardados com o node **Image Writer**, permitindo a exportação direta das visualizações. Finalmente, utilizou-se o node **Send Email** para automatizar o envio das imagens por email, assegurando que os resultados fossem enviados de forma prática.

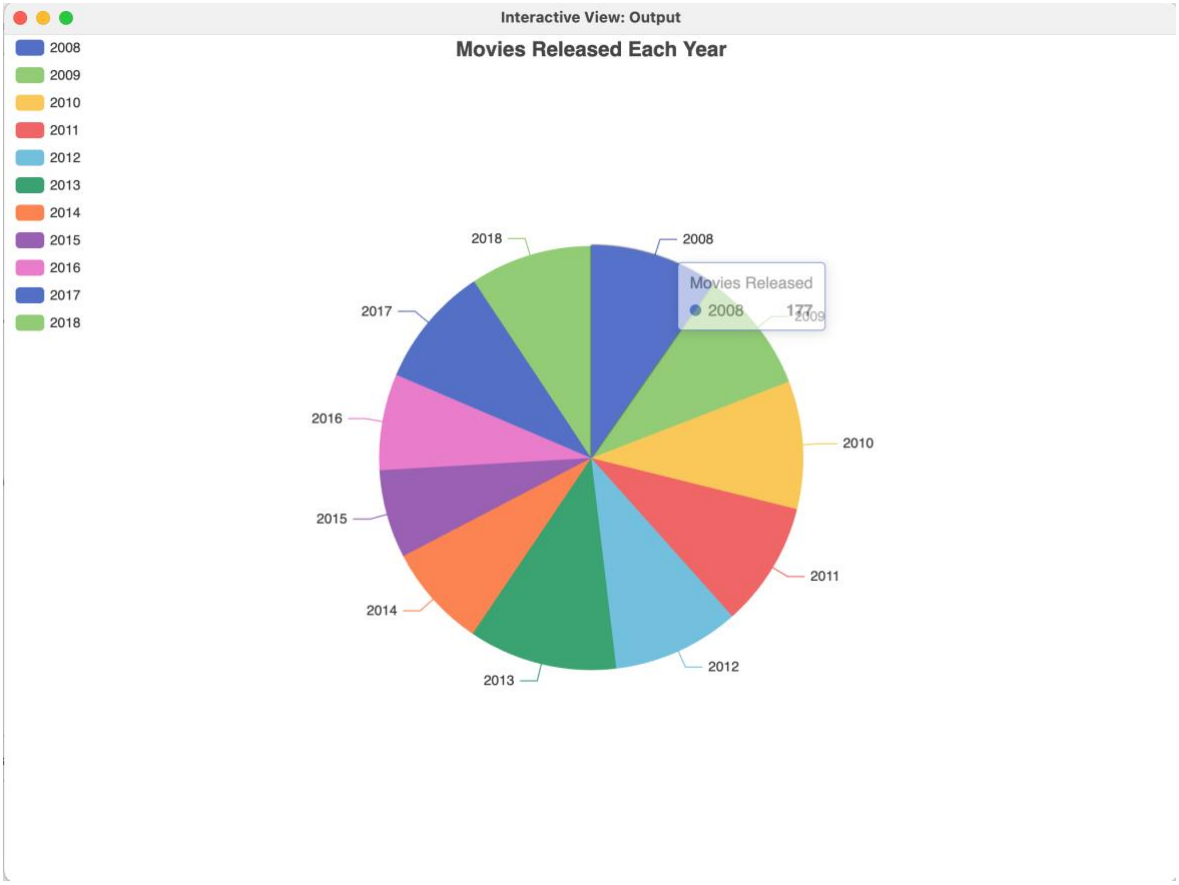


Figura 7 - Gráfico "Released Movies"

O gráfico circular que representa a quantidade dos filmes lançados entre 2008 e 2018 e mostra a distribuição dos lançamentos ao longo dos anos, permitindo observar quais anos tiveram mais ou menos produções.

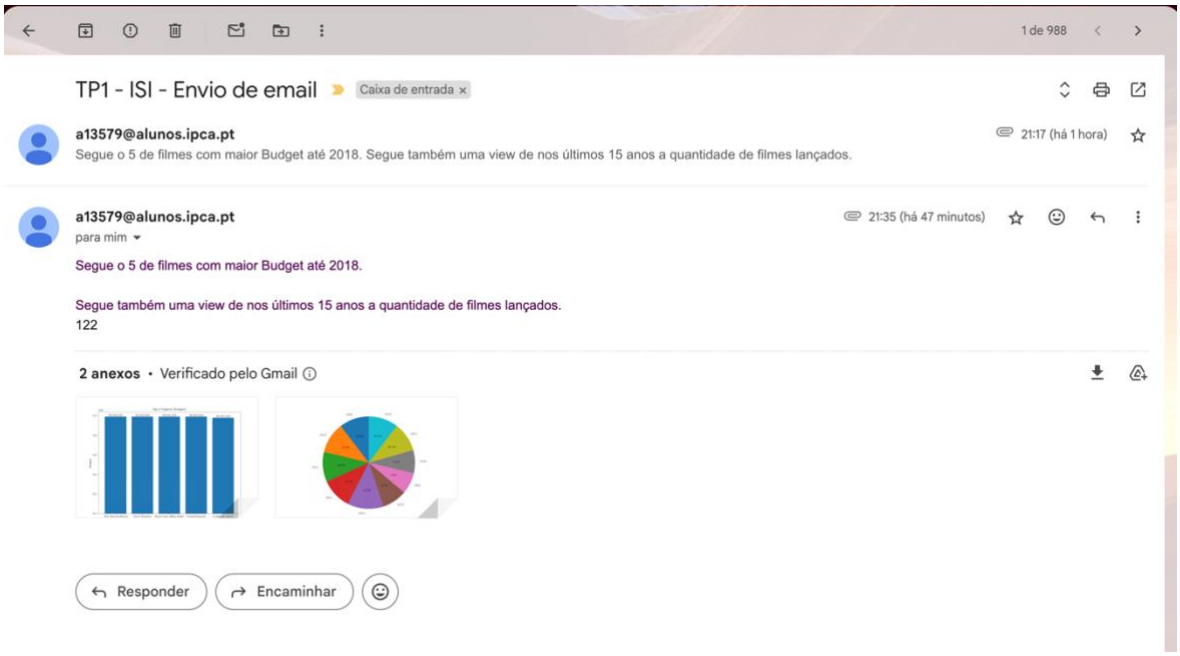


Figura 8 - Envio de email

Na *Fig. 8* podemos verificar o resultado obtido após gerar os gráficos via Python View e exportá-los para ".png". Esses gráficos utilizando o node "Send Email" foram enviados a partir do email escolar do *IPCA* para a minha conta de email pessoal com sucesso.

7. Extra

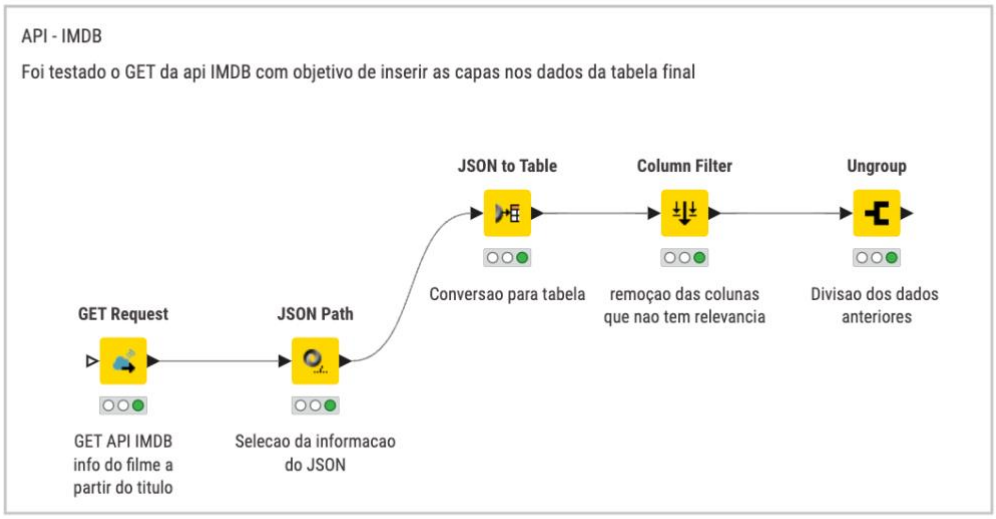


Figura 9 - API IMDB

Na *Fig. 9*, já com o objetivo do trabalho prático concluído, na procura de mais funcionalidades a serem utilizadas através da ferramenta KNIME, foi feito um GET Request à API do IMDB, o processo de aprendizagem foi relativamente simples e foi analisada a possibilidade de integrar uma imagem da capa para cada respetivo filme nos dados iniciais. Infelizmente, não foi possível realizar este processo, no entanto o conhecimento obtido será aplicado num outro projeto, seja a nível académico, pessoal e/ou profissional.

8. Conclusão

Em suma, o processo de limpeza e normalização das colunas essenciais revelou-se fundamental para garantir a qualidade dos dados, permitindo a criação de relatórios e gráficos eficazes.

Ao longo do projeto, foi possível transformar dados desestruturados em informações valiosas, demonstrando a relevância das operações de ETL na análise de dados. O trabalho realizado não só possibilitou a visualização das tendências e relações subjacentes nos dados, como também proporcionou uma experiência prática significativa na utilização do KNIME. Esta experiência é um passo importante para a continuidade do desenvolvimento de competências na área da Engenharia de Sistemas Informáticos e na aplicação de ferramentas analíticas para resolver problemas complexos.

9. Apresentação em Vídeo

Abaixo podemos verificar o QR Code que remete para a apresentação do trabalho desenvolvido num breve vídeo.

Figura 10 - QRCode

10. Bibliografia

Collect API –

https://collectapi.com/api/imdb/imdb-api

IPCA Moodle –

https://elearning.ipca.pt/2425/my/

KNIME Community Hub –

https://hub.knime.com/

Json to CSV Converter –

https://www.convertcsv.com/json-to-csv.htm