Trabalho prático 1: ETL

Rodrigo Carreiras Pilar

Nº 26536 – Regime Pós-laboral

Ano letivo 2024/2025

Licenciatura em Engenharia de Sistemas Informáticos

Escola Superior de Tecnologia

Instituto Politécnico do Cávado e do Ave

Identificação do Aluno

Rodrigo Carreiras Pilar

Aluno número 26536, regime pós-laboral

Licenciatura em Engenharia de Sistemas Informático

RESUMO

Neste projeto, o objetivo foi desenvolver um fluxo de análise de dados automóvel, utilizando uma base de dados contendo informações técnicas de veículos, como cilindrada, consumo de combustível, ano de fabrico, entre outros. O problema principal a resolver consistia em tratar e validar estes dados, eliminando valores incorretos ou inconsistentes, e aplicando transformações necessárias para tornar a informação útil para futuras análises.

A abordagem realizada incluiu a importação dos dados a partir de um ficheiro CSV, seguida de uma série de operações de limpeza e transformação. Valores nulos foram tratados, cilindros fora do intervalo permitido foram eliminados e as colunas numéricas, como o ano de fabrico, foram utilizadas para calcular a idade dos veículos. Adicionalmente, foi garantido que os dados pudessem ser exportados em diversos formatos, como Excel e JSON, adequando-se às diferentes necessidades de output.

Como resultado, obteve-se um conjunto de dados limpo e estruturado, pronto para ser utilizado em análises futuras. A aplicação de métodos automáticos para garantir a qualidade dos dados permitiu um processamento eficiente e com pouca intervenção manual, demonstrando a viabilidade e utilidade da solução desenvolvida.

ABSTRACT

In this project, the aim was to develop an automobile data analysis flow, using a database containing technical vehicle information, such as cylinder capacity, fuel consumption, year of manufacture, among others. The main problem to be solved was to process and validate this data, eliminating incorrect or inconsistent values and applying the necessary transformations to make the information useful for future analysis.

The approach taken included importing the data from a CSV file, followed by a series of cleaning and transformation operations. Null values were treated, cylinders outside the permitted range were eliminated and numerical columns, such as the year of manufacture, were used to calculate the age of the vehicles. In addition, it was ensured that the data could be exported in various formats, such as Excel and JSON, to suit different output needs.

As a result, a clean and structured data set was obtained, ready to be used in future analyses. The application of automatic methods to guarantee data quality enabled efficient processing with little manual intervention, demonstrating the viability and usefulness of the solution developed.

ÍNDICE

1. Introdução 1
1.1. Objetivos 2
1.2. Contexto 3
1.3. Estrutura do documento 4
2. Análise do problema 5
2.1. Problemas Identificados 5
2.2. Consequências da Falta de Tratamento dos Dados 6
2.3. Necessidade de Solução 6
3. Implementação 7
3.1. Leitura e Preparação dos dados 9
3.2. Organização e Integração de Dados Temporais 12
3.3. Filtragem de Dados para Cálculo da Idade e Organização das Colunas 13
3.4. Processamento e Exportação de Dados 15
4. Conclusão 17

ÍNDICE DE FIGURAS

Figura 1 - Visão Geral do Projeto 7
Figura 2 - Leitura e preparação dos dados 9
Figura 3 - Configuração do Colunm Expressions 10
Figura 4 - Organização e Integração de Dados Temporais 12
Figura 5 - Filtragem de Dados para Cálculo da Idade e Organização das Colunas 13

ÍNDICE DE TABELAS

Tabela 1 - Ajuste do Ano do Modelo 10
Tabela 2 - Extração da Marca 11
Tabela 3 - Extração do Modelo 11
Tabela 4 - Expressão do Cálculo da Idade 13
Tabela 5 - Expressão da verificação dos valores da coluna cylinders 14
Tabela 6 - Expressão para filtrar coluna mpg e horsepower 15

Siglas e Acrónimos

ETL - Extract, Transform, Load: Processo de extração, transformação e carregamento de dados, essencial em fluxos de integração de dados.

CSV - Comma-Separated Values: Formato de ficheiro para armazenamento de dados em tabelas de texto.

JSON - JavaScript Object Notation: Formato leve de intercâmbio de dados.

API - Application Programming Interface: Interface que permite a interação entre diferentes sistemas de software.

AI - Artificial Intelligence: Inteligência Artificial, refere-se à simulação de processos de inteligência humana por sistemas computacionais.

KNIME - Konstanz Information Miner: Plataforma de análise de dados utilizada para criar fluxos de trabalho de data science.

Node - Elemento em KNIME que executa uma tarefa específica dentro do fluxo de trabalho, como um bloco de operações.

1. Introdução

Atualmente, o tratamento e análise de grandes volumes de dados tornaram-se indispensáveis em várias áreas de atuação, incluindo a indústria automóvel. Com o crescente desenvolvimento de tecnologias e a digitalização de processos, as organizações necessitam de ferramentas eficazes para gerir, processar e analisar dados de forma eficiente e precisa. Neste contexto, a área de Sistemas de Informação desempenha um papel fundamental, fornecendo soluções que permitem otimizar o uso de dados para a tomada de decisões mais informadas e estratégicas.

No setor automóvel, a análise de dados é particularmente relevante para avaliar o desempenho dos veículos, monitorizar padrões de consumo de combustível, e prever necessidades de manutenção. Além disso, a correta gestão de dados técnicos, como especificações de motores e cilindradas, permite às empresas do ramo automóvel obter informações cruciais para melhorar a eficiência dos seus processos e produtos. No entanto, o manuseio de dados automóveis também apresenta desafios, como a presença de informações incorretas ou incompletas, que podem comprometer a qualidade das análises realizadas.

O presente trabalho aborda o desenvolvimento de um fluxo de tratamento e transformação de dados automóveis, com o objetivo de garantir a qualidade e integridade da informação recolhida. Ao longo do projeto, foi realizada a importação de dados provenientes de uma base de dados automóvel, seguida de operações de limpeza, transformação e validação dos mesmos, para que possam ser utilizados em futuras análises de forma mais eficiente e confiável.

1.1. Objetivos

O principal objetivo deste trabalho foi desenvolver um fluxo eficiente de tratamento de dados automóveis, capaz de garantir a sua integridade e preparar a informação para futuras análises. Para atingir este objetivo geral, foram definidos os seguintes pontos principais:

- **Importação e Validação de Dados:** Assegurar a correta importação de dados automóveis a partir de um ficheiro CSV, garantindo que a estrutura dos dados é preservada e que possíveis erros ou omissões sejam identificados.
- **Limpeza e Transformação dos Dados:** Implementar processos de limpeza e transformação dos dados, de modo a eliminar informações inconsistentes ou inválidas, como valores fora de intervalos predefinidos ou a presença de valores nulos que comprometam a qualidade dos dados.
- **Conversão para Formatos de Dados Diferentes:** Automatizar a exportação dos dados limpos para diferentes formatos, como Excel, XML e JSON, de forma a garantir a compatibilidade com diferentes plataformas e ferramentas analíticas.
- **Melhoria da Qualidade dos Dados:** Garantir que os dados tratados possam ser utilizados em futuras análises de forma eficiente, ao aplicar boas práticas de manipulação de dados e assegurando a sua integridade ao longo do processo de transformação.

Estes objetivos foram pensados de modo a assegurar que, ao final do processo, os dados estejam prontos para serem utilizados em análises avançadas ou integrados em sistemas de monitorização e tomada de decisões dentro do contexto da indústria automóvel.

1.2. Contexto

O presente projeto foi desenvolvido no âmbito da unidade curricular de Integração de Sistemas de Informação do curso de Engenharia de Sistemas Informáticos, com o objetivo de proporcionar aos alunos uma experiência prática na aplicação de conceitos de tratamento e análise de dados. Através deste projeto, foi possível consolidar os conhecimentos teóricos adquiridos ao longo do curso, utilizando ferramentas avançadas para lidar com problemas reais de manipulação e transformação de dados.

Para o desenvolvimento do projeto, foi utilizado o KNIME, uma plataforma de análise de dados que oferece uma grande variedade de ferramentas e nodes para manipulação, transformação e visualização de dados. O KNIME permitiu a integração de diferentes fontes de dados, facilitando o processo de extração, transformação e carregamento (ETL) dos dados. Além disso, possibilitou a implementação de filtros, transformações e validações automáticas, essenciais para garantir a qualidade e integridade dos dados processados.

Este ambiente proporcionou um espaço intuitivo e eficiente para desenvolver o fluxo de trabalho necessário para o tratamento dos dados, utilizando as suas funcionalidades visuais e orientadas a nós, que permitem uma construção clara e modular do processo de análise.

1.3. Estrutura do documento

Este relatório está organizado em cinco capítulos, além da introdução, e cada um deles foca numa etapa distinta do desenvolvimento do projeto.

Capítulo 2 – Análise do problema: Neste capítulo, é apresentada uma análise detalhada do problema abordado pelo projeto. Serão discutidas as necessidades de tratamento e análise de dados automóveis e a importância de garantir a qualidade da informação para suportar decisões estratégicas. Também inclui uma breve análise sobre o estado da arte e as soluções existentes no mercado para este tipo de problema.

Capítulo 3 – Implementação: Este capítulo aborda a implementação do projeto, incluindo uma descrição detalhada das tecnologias utilizadas, com especial destaque para o uso do KNIME e os nodes que suportaram as diferentes transformações de dados. Serão discutidas as decisões técnicas tomadas e os principais desafios encontrados ao longo do desenvolvimento.

Capítulo 4 – Análise de resultados e testes: Aqui, serão analisados os resultados obtidos após a implementação do sistema de tratamento de dados. Inclui uma avaliação da eficácia das transformações aplicadas e dos outputs gerados, bem como uma análise dos testes realizados, com foco na validação dos dados e no cumprimento dos objetivos estabelecidos.

Capítulo 5 – Conclusão: O último capítulo apresenta as conclusões finais do trabalho desenvolvido. São discutidos os principais aprendizados e desafios, além de potenciais trabalhos futuros que poderiam dar continuidade ao projeto

2. Análise do problema

O volume crescente de dados em diversas áreas da indústria exige a adoção de ferramentas robustas e metodologias eficazes para tratar e analisar a informação disponível. No contexto do setor automóvel, dados como consumo de combustível, cilindrada, ano de fabrico, e outras especificações técnicas dos veículos são fundamentais para várias finalidades, desde o estudo de desempenho e eficiência até à análise de padrões de produção e vendas. No entanto, a qualidade dos dados disponíveis nem sempre está garantida, devido a valores incorretos, dados incompletos ou informações inconsistentes, que podem comprometer a análise e prejudicar a tomada de decisões.

2.1. Problemas Identificados

Ao lidar com dados automóveis, surgem diversos problemas relacionados com a integridade e validade da informação. Alguns dos desafios mais comuns incluem:

Dados incompletos: A ausência de informações essenciais, como o consumo de combustível (mpg), a cilindrada (cylinders) ou o ano de fabrico, afeta a capacidade de realizar análises precisas. Dados incompletos podem levar a conclusões incorretas ou enviesadas.

Valores fora do intervalo: Em muitas bases de dados, é comum encontrar valores fora do intervalo esperado, como no caso de cilindradas que não correspondem aos padrões conhecidos (ex: cilindros abaixo de 3 ou acima de 12, o que não é comum em veículos). Esses valores podem ser resultado de erros de inserção ou falta de normalização na recolha de dados.

Valores inválidos: Alguns valores pareciam fora do padrão esperado, exigindo verificações e ajustes automáticos, como a substituição de valores nulos por etiquetas como "ND" (Não Definido), de modo a manter a integridade do conjunto de dados.

2.2. Consequências da Falta de Tratamento dos Dados

Sem um tratamento adequado, a análise destes dados pode levar a conclusões erradas. Problemas como dados incompletos ou valores fora do intervalo esperado podem resultar em:

Conclusões enviesadas: Dados incorretos podem levar a uma análise distorcida, afetando métricas importantes como o desempenho dos veículos ou o consumo de combustível.

Incompatibilidade com outras ferramentas: Dados mal formatados ou inválidos podem dificultar a exportação para diferentes formatos e plataformas de análise, como Excel ou JSON, limitando a sua usabilidade.

2.3. Necessidade de Solução

A necessidade de uma solução para garantir a qualidade dos dados tornou-se clara. Foi importante desenvolver um fluxo de trabalho para a limpeza e validação dos dados, utilizando a plataforma KNIME. O KNIME, através dos seus diversos nós e ferramentas, permitiu realizar operações automáticas de transformação e validação dos dados, garantindo que o conjunto de dados final estivesse limpo, consistente e pronto para ser utilizado em análises futuras.

O principal objetivo foi preparar os dados automóveis, inicialmente disponíveis em formato CSV, e garantir que eles pudessem ser exportados em formatos adequados (como Excel e JSON), mantendo a integridade e utilidade da informação processada.

3. Implementação

Neste capítulo, será descrito detalhadamente o processo de implementação do projeto, focando nos diferentes componentes e operações realizados no fluxo de trabalho. O desenvolvimento foi feito utilizando a plataforma KNIME, que permitiu criar um fluxo modular e eficiente para o tratamento dos dados automóveis. Além disso, serão incluídos prints dos fluxos implementados, juntamente com explicações das funcionalidades dos diferentes nós utilizados.

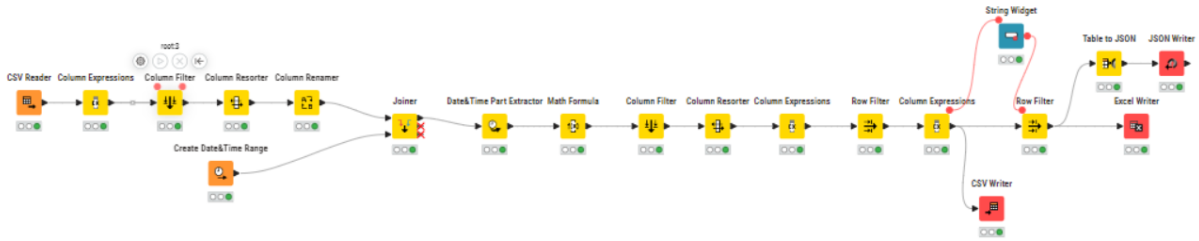


Figura 1 - Visão Geral do Projeto

A imagem apresenta um fluxo que inclui etapas de limpeza e manipulação de dados, como a aplicação de expressões em colunas, filtragem e reordenamento de colunas. Adicionalmente, um nó de criação de intervalos de datas foi integrado ao processo. Os dados resultantes são exportados em três formatos diferentes: JSON, Excel e CSV, utilizando os respetivos "writers" do KNIME para garantir a flexibilidade do formato de saída.

3.1. Leitura e Preparação dos dados

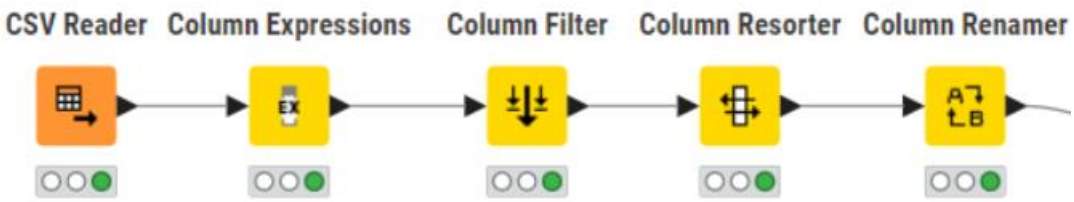


Figura 2 - Leitura e preparação dos dados

Nesta primeira fase do fluxo, o **CSV Reader** carrega o ficheiro CSV com os dados dos automóveis. Em seguida, o node **Column Expressions** cria duas colunas com base na coluna “name” separando o nome da marca do modelo através de uma expressão personalizada. O **Column Filter** filtra as colunas, removendo aquelas que não são necessárias para o processo, neste caso, remove a coluna “name”. Enquanto o **Column Resorter** reorganiza a ordem das colunas para preparar os dados. Finalmente, o **Column Renamer** ajusta os nomes das colunas para garantir consistência e clareza. Essas operações são essenciais para preparar os dados brutos para as etapas seguintes.

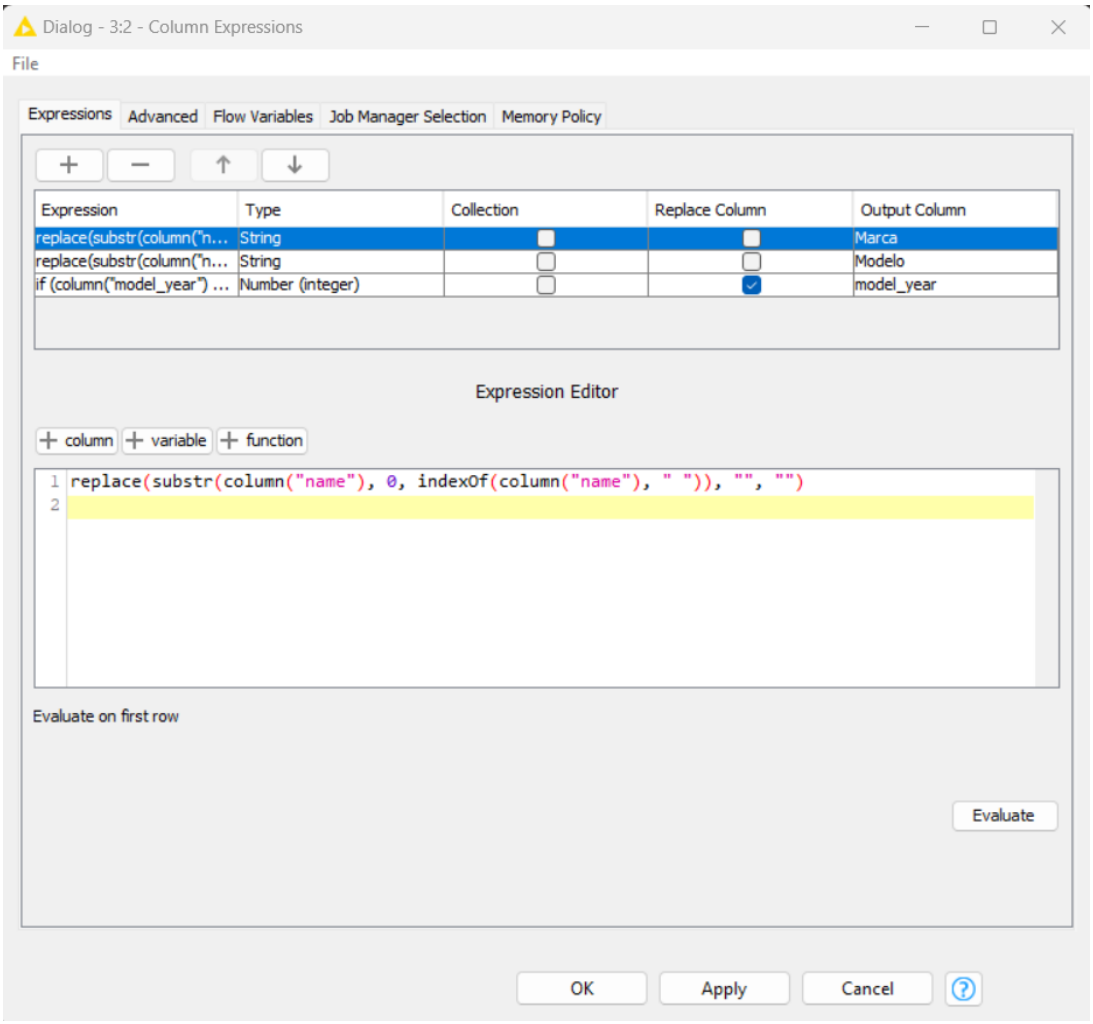


Figura 3 - Configuração do Colunm Expressions

Para o node **Column Expressions**, foram implementadas três expressões principais:

- **Ajuste do Ano do Modelo (model_year):** A primeira função corrige o ano do modelo do veículo. Se o valor de **model_year** for inferior ou igual a 24, significa que o veículo é do século XXI (anos 2000-2024). Portanto, 2000 é somado ao valor de **model_year**. Caso contrário, adiciona-se 1900 para corrigir os anos anteriores (1925-1999).

```
if (column("model_year") <= 24) {
    2000 + column("model_year")
} else {
    1900 + column("model_year")
}
```

Tabela 1 - Ajuste do Ano do Modelo

- **Extração da Marca:** A segunda função usa a coluna **name** para extrair a marca do veículo, removendo o nome do modelo e mantendo apenas a parte correspondente à marca.

```
replace(substr(column("name"), 0, indexOf(column("name"), " ")), "", "")
```

Tabela 2 - Extração da Marca

- **Extração do Modelo:** A terceira função, semelhante à segunda, utiliza a mesma coluna **name**, mas extrai o nome do modelo do carro, removendo a parte correspondente à marca.

```
replace(substr(column("name"), indexOf(column("name"), " ") + 1, length(column("name"))), "", "")
```

Tabela 3 - Extração do Modelo

As expressões anteriormente referidas estão escritas em **JavaScript**, que é a linguagem suportada pelo node **Colunm Expressions** para a manipulação de dados nas colunas. Estas expressões ajudam a padronizar e limpar os dados antes de prosseguir para as próximas etapas do fluxo.

3.2. Organização e Integração de Dados Temporais

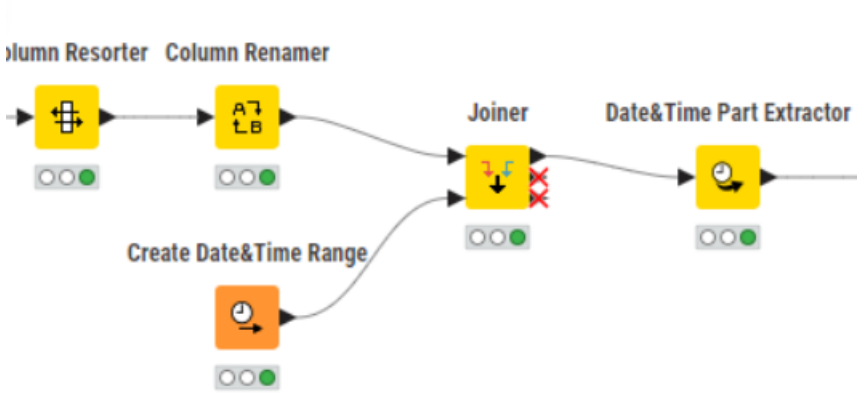


Figura 4 - Organização e Integração de Dados Temporais

No âmbito deste projeto, o node **Create Date&Time Range** foi utilizado com o objetivo específico de gerar uma coluna que contém a data atual de forma dinâmica. Esta abordagem permite que o cálculo da idade dos automóveis seja feito automaticamente, sempre com base na data em que o fluxo é executado. Assim, evita-se a necessidade de atualizar manualmente a data de referência, garantindo que a idade dos veículos seja calculada com precisão a partir da data atual.

Para simplificar o cálculo da idade, utilizou-se o node **Date&Time Part Extractor**. Este node foi configurado para extrair apenas o ano da data atual gerada pelo **Create Date&Time Range**, uma vez que, para o cálculo da idade, apenas esta parte específica da data é relevante. Ao extrair apenas o ano, torna-se possível realizar a subtração direta entre o ano atual e o ano de fabricação presente no dataset, obtendo assim a idade exata dos automóveis.

Adicionalmente, o node **Joiner** foi utilizado para combinar dados provenientes de diferentes tabelas dentro do fluxo, permitindo que as informações de datas e os restantes atributos dos automóveis fossem integrados e transformados em uma só tabela.

3.3. Filtragem de Dados para Cálculo da Idade e Organização das Colunas

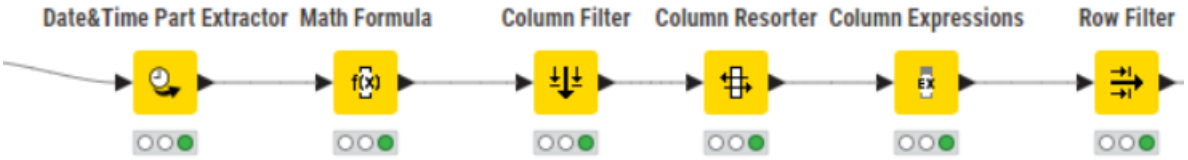


Figura 5 - Filtragem de Dados para Cálculo da Idade e Organização das Colunas

Neste conjunto de nodes, o **Math Formula** e o **Column Expressions** são utilizados para realizar cálculos e aplicar condições aos dados de forma a preparar a informação para análise. O **Row Filter**, por sua vez, desempenha um papel essencial ao garantir que apenas as linhas com dados relevantes sejam mantidas. O **Colunm Resorter** é utilizado novamente apenas para modificar a ordem das colunas. Abaixo está a explicação mais detalhada de alguns destes nodes e o papel que desempenham no fluxo.

O **Math Formula** é usado para calcular a idade dos veículos. A fórmula subtrai o ano de fabricação do veículo ao ano atual, resultando assim na idade do veículo. Este node cria uma nova coluna com a idade de cada veículo.

$Ano Atual$-$Year$

Tabela 4 - Expressão do Cálculo da Idade

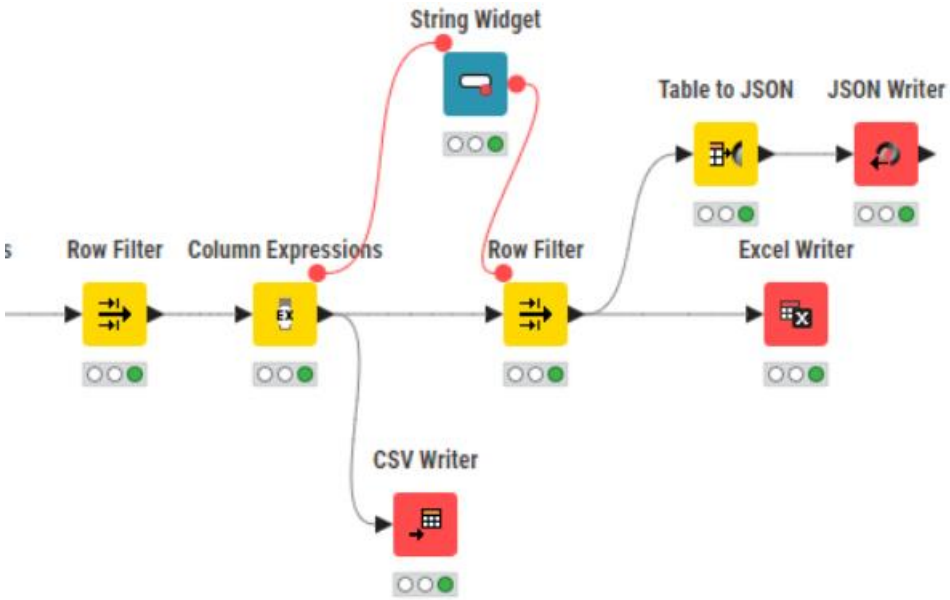
O **Column Expressions** permite verificar e ajustar os valores da coluna **cylinders** com base em condições específicas. No código apresentado, o node verifica se o valor na coluna **cylinders** está dentro de um intervalo aceitável, de 3 a 12 cilindros. Caso o valor esteja dentro deste intervalo, o próprio valor é mantido; caso contrário, é atribuído o valor **null** para indicar que o dado não atende aos critérios de validação. Este filtro assegura que apenas veículos com um número de cilindros razoável (entre 3 e 12) sejam considerados.

```
if (column("cylinders") >= 3 && column("cylinders") <= 12)
{
    column("cylinders")
}
else
{
    null
}
```

Tabela 5 - Expressão da verificação dos valores da coluna cylinders

O **Row Filter** é configurado para eliminar todas as linhas onde o valor da coluna **cylinders** está em falta (missing). Esta etapa é importante para remover dados incompletos e garantir a qualidade dos dados analisados. No caso específico, o filtro é configurado com a condição **cylinders is not missing** para permitir que apenas as linhas que possuem dados válidos na coluna **cylinders** avancem para as etapas seguintes do fluxo.

3.4. Processamento e Exportação de Dados



Neste projeto, cada node deste fluxo de trabalho desempenha um papel crucial para transformar, filtrar e exportar os dados, facilitando a análise e processamento específicos dos dados dos veículos.

Primeiramente, o node **Column Expressions** é configurado para verificar valores ausentes nas colunas ***mpg*** e ***horsepower***. Abaixo explico detalhadamente o funcionamento dessa configuração no contexto do projeto.

```
if
(isMissing(column("mpg"))) {
    "ND"
} else {
    column("mpg")
}
```

```
if
(isMissing(column("horsepower")))
{
    "ND"
} else {
    column("horsepower")
}
```

Tabela 6 - Expressão para filtrar coluna mpg e horsepower

Após este mesmo node, o fluxo divide-se em dois caminhos, um para o node **CSV Writer**, que exporta o conjunto de dados em formato CSV, um formato amplamente aceite e compatível com diversos sistemas de análise de dados. No outro caminho do fluxo, o **String Widget** é configurado para armazenar uma **string** com o valor "**Ford**". Esse valor é usado como um parâmetro dinâmico, permitindo flexibilidade no processo de filtragem dos dados. Em seguida, o **Row Filter** utiliza essa **string** para filtrar as linhas da tabela, mantendo apenas aquelas onde a coluna **name** contém o valor **"Ford"**. Isso significa que, ao alterar o valor no **String Widget**, o filtro pode ser atualizado automaticamente para refletir a nova entrada.

Para assegurar compatibilidade com outras ferramentas de análise, o **Excel Writer** exporta os dados para um arquivo Excel. Esse formato é ideal para relatórios em ferramentas de escritório, como o Microsoft Excel, facilitando a visualização e manipulação dos dados processados.

Depois, o **Table to JSON** converte a tabela de dados processada para o formato JSON, um formato amplamente utilizado em aplicações web e APIs. Esta conversão é essencial para o projeto, pois possibilita integrar os dados com sistemas ou plataformas que aceitam JSON como entrada, permitindo uma fácil reutilização dos dados. O **JSON Writer**, por sua vez, guarda esses dados em um arquivo externo no formato JSON, permitindo o armazenamento e a disponibilização dos dados de maneira estruturada e acessível.

4. Conclusão

Este projeto permitiu desenvolver e consolidar competências em manipulação e análise de dados, utilizando uma série de ferramentas de processamento e filtragem que tornaram o fluxo de trabalho mais eficiente e dinâmico. Ao longo das etapas, foi possível aplicar técnicas para limpar, organizar e transformar os dados de forma a extrair insights relevantes sobre os veículos, como a sua idade e características específicas.

A criação de funcionalidades dinâmicas, como o uso do **String Widget** para filtrar marcas específicas e a geração automática da idade dos veículos com base na data atual, trouxe flexibilidade e adaptabilidade ao processo. Ferramentas como o **Row Filter**, **Column Expressions**, e **Math Formula** foram essenciais para manipular e validar os dados de forma precisa, garantindo que a informação final estivesse limpa e adequada para exportação e análise.

Em conclusão, o projeto atingiu os objetivos propostos, demonstrando como a utilização de ferramentas de análise de dados pode otimizar processos de preparação e exploração de dados. A capacidade de manipular grandes volumes de dados com eficiência torna-se uma competência fundamental, e este trabalho reflete a importância de um pipeline bem estruturado para suportar decisões informadas e sustentadas por dados.

Bibliografia

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques. Elsevier.

Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media.

Wickham, H., & Grolemund, G. (2017). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media.

Knime AG. (2021). Data Science with KNIME: Using KNIME Analytics Platform for ETL, Machine Learning, and Reporting. Disponível em: https://www.knime.com/knimepress/data-science-with-knime

KNIME Documentation. KNIME Analytics Platform Guide. Disponível em: https://docs.knime.com/