

Integração de Sistema de Informação

Relatório de Trabalho Prático 1

Júlia Dória Rodrigues - 24204

Engenharia de Sistemas Informáticos

Outubro de 2024

Afirmo por minha honra que não recebi qualquer apoio não autorizado na realização deste trabalho prático.
Afirmo igualmente que não copiei qualquer material de livro, artigo, documento web ou de qualquer outra fonte exceto onde a origem estiver expressamente citada.

Júlia Dória Rodrigues - 24204

Índice

1.	INTRODUÇÃO	5
2.	DESCRIÇÃO DO PROBLEMA	6
3.	CONFECÇÃO DE DADOS	9
	Dados em branco e modificações	9
	Alterações manuais	10
4.	PROCESSO ETL	11
	Extração	11
	Transformação	12
	Carga (Load)	19
5.	JOBs	21
	Cálculo de Bónus salarial	21
	Extração de campos vazios	22
	Estatísticas	23
6.	TABLEAU	24
7.	CONCLUSÃO E TRABALHOS FUTUROS	27
8.	REFERÊNCIAS	29

Lista de Figuras

Figura 1 - Criação de dados fictícios através da plataforma Mockaroo	9
Figura 2 - Dados alterados para posterior limpeza e formatação.....	10
Figura 3 - Número de linhas geradas no ficheiro CSV	10
Figura 4 - Número de linhas com duplicação em ficheiro CSV.....	10
Figura 5 - Visão geral do projeto Knime	11
Figura 6 - Caminho relativo ficheiro CSV.....	12
Figura 7 - Remoção de linhas duplicadas	12
Figura 8 - Regex para manipulação das datas	13
Figura 9 - Nó para calcular idade e tempo de contrato	14
Figura 10 - Expressão para construção de coluna com valor de porcentagem do bônus através da quantidade de dias de ausência.....	14
Figura 11 - Expressão para construção de coluna com o valor do bônus pelo tempo de contrato ...	15
Figura 12 - Cálculo de valor total do bónus.....	15
Figura 13 - Filtragem de funcionários com e sem bonus	15
Figura 14 – Filtragem de colunas.....	16
Figura 15 - Filtragem de colunas (2)	16
Figura 16 - Filtragem de colunas (3)	17
Figura 17 - Preenchimento de campos vazios.....	17
Figura 18 - Criação de faixas de idade com o nó Numeric Binner	18
Figura 19 - Nó Group By para contagem de gênero por departamento	18
Figura 20 - Configuração de nó Group By.....	19
Figura 21 - Configuração de nó excel writer	19
Figura 22 - Configuração do nó Tableau Writer	20
Figura 23 - Configuração do nó CSV Writer.....	20
Figura 24 - Processo de cálculo de bónus salarial.....	21
Figura 25 - Processo de extração de campos vazios	22
Figura 26 - Processo de obtenção de dados estatísticos	23
Figura 27 - Gráfico de média de ausência por faixa etária	25

Figura 28 - Gráfico de distribuição de gênero por departamento	25
Figura 29 - Gráfico de distribuição de média salarial por gênero	26

1. Introdução

Este artigo aborda a aplicação dos processos ETL (Extract, Transform, Load) no contexto da integração de sistemas de informação, com enfoque na análise e tratamento de dados empresariais. O principal objetivo é desenvolver a prática e a competência destes processos, dada a sua importância na preparação dos dados, quer para fornecer informação relevante, quer para ajudar a tomar decisões organizacionais.

Neste projeto, será simulado um cenário de uma empresa com um arquivo contendo os dados de todos os funcionários. O projeto visa inicialmente formatar os dados, para depois calcular os prémios dos seus colaboradores com base em métricas como o número de faltas e o tempo de contrato. Analisará também a distribuição dos salários por género em cada um dos departamentos da empresa e a relação entre o número de faltas e a idade de cada funcionário. Por fim, os dados recolhidos serão analisados com recurso à ferramenta Tableau.

O fluxo de trabalho será desenvolvido com recurso ao KNIME, uma plataforma de análise de dados e de aprendizagem automática de código aberto que fornece suporte para a criação de processos ETL com pouco ou nenhum código.

O trabalho inclui a geração de dados fictícios para simular a situação real utilizando o Mockaroo (<https://www.mockaroo.com>), aplicando transformações para limpar e enriquecer os dados, bem como cálculos baseados em regras comerciais. O resultado final será um conjunto de dados transformados que serão analisados de diferentes formas para fornecer informações à empresa sobre o seu contingente.

2. Descrição do Problema

A empresa fictícia XYZ é uma organização que emprega um grande número de funcionários em vários departamentos. Como parte do seu processo de otimização de recursos humanos e análise de desempenho, a empresa enfrenta o desafio de integrar, transformar e analisar os dados dos colaboradores para apoiar a tomada de decisões estratégicas. Estes dados incluem informações sobre os colaboradores, tais como:

- **Employee ID:** Identificador único para cada funcionário.
- **First Name:** Nome do funcionário.
- **Last Name:** Último nome do funcionário.
- **Date of Birth:** Data de nascimento do funcionário.
- **Hire Date:** Data de contratação do funcionário.
- **Job Title:** Cargo atual.
- **Department:** Departamento no qual o funcionário trabalha.
- **Annual Salary:** Salário anual do funcionário.
- **Absence Days:** Quantidade de dias de ausência no trabalho no último ano.
- **Gender:** Gênero com o qual o funcionário se identifica.
- **Marital Status:** Estado civil do funcionário.
- **City:** Cidade onde o funcionário reside.
- **Country:** País de residência do funcionário.

O primeiro problema a enfrentar é que o conjunto de dados apresentado no ficheiro *mock_data.csv* contém algumas inconsistências, tais como diferentes formatos de data (por exemplo, algumas datas estão no formato MM/DD/AAAA e outras no formato MM.DD.AAAA), e há também regtos com campos em branco, especialmente nos campos *Job Title* e *Gender*.

Além disso, a empresa pretende efetuar análises específicas para:

- Cálculo da idade e do tempo de serviço: Determinar a idade de cada empregado com base na sua data de nascimento e o seu tempo de serviço com base na sua data de contratação.

- Cálculo de bónus de desempenho: Após calcular o tempo de serviço de cada funcionário, atribuir bónus a eles com base no número de faltas e no tempo de contrato, utilizando as seguintes regras:

Primeiro em atenção ao tempo de contrato:

- 200 euros para os trabalhadores com 0 a 5 anos de contrato.
- 400 euros para os trabalhadores com 6 a 10 anos de contrato.
- 1000 euros para os trabalhadores com mais de 10 anos de contrato.

Em seguida, o montante do prémio é determinado em função da duração do contrato:

- Os trabalhadores com mais de 15 faltas não recebem qualquer bónus.
- Os colaboradores com 0 a 5 faltas recebem 100% do bónus.
- Os colaboradores com 6 a 10 faltas recebem 50% do bónus.
- Empregados com 11 a 15 faltas recebem 25% do bónus.
- Análise de Género por Departamento: Analisar a distribuição do género por departamento, considerando ainda que alguns colaboradores não preencheram o campo do género.
- Análise das disparidades salariais entre géneros: Comparar os salários médios por género e identificar se existem disparidades significativas.
- Análise do Número de Ausências por Faixa Etária: Avaliar como varia o número de ausências em função da idade dos colaboradores, utilizando faixas etárias pré-definidas.
- Identificação de Registros Incompletos: Gerar uma lista de nomes de colaboradores com campos em branco (como por exemplo, Função e Sexo) para que a empresa possa posteriormente corrigir e completar essa informação.

O objetivo é que, no final de toda a extração e transformação destes dados, a empresa obtenha novos ficheiros com informações que possam melhorar o seu desempenho em relação aos seus colaboradores.

Em primeiro lugar, para saber exatamente qual o valor do bónus que cada colaborador irá receber. Além disso, poder analisar a distribuição do género dos seus empregados por departamento, e também a distribuição salarial por género, para analisar se existe alguma disparidade tanto salarial como de oportunidades dentro de cada departamento.

Existe ainda a necessidade de preencher os dados em falta, uma vez que esta informação é crucial para a continuidade das análises da empresa. Por fim, a análise das ausências dos colaboradores em função da idade tem como objetivo prever qual a faixa etária é mais presente, para que se possa ponderar se esta deve ser tida em conta em futuras contratações.

3. Confecção de Dados

Os dados utilizados para este projeto foram gerados através da ferramenta Mockaroo, que permite criar dados fictícios com base em critérios personalizados. Esta ferramenta foi escolhida para simular um cenário realista de uma empresa, permitindo a realização de várias análises e operações de transformação e limpeza de dados ao longo do projeto.

Como já relatado supra, os dados gerados consistem em informações desde uma id para cada funcionário, até a cidade e país em que residem (Figura1).

The screenshot shows the Mockaroo data generation interface with the following field configurations:

Field	Type	Configurations
employee_id	Row Number	blank: 0 %, Σ, X
First Name	First Name	blank: 0 %, Σ, X
Last Name	Last Name	blank: 0 %, Σ, X
Date of Birth	Datetime	10/16/2023 to 10/16/2024, format: m/d/yyyy, blank: 0 %, Σ, X
Hire Date	Datetime	10/16/2023 to 10/16/2024, format: m/d/yyyy, blank: 0 %, Σ, X
Job Title	Job Title	blank: 15 %, Σ, X
Department	Department (Corpor...)	blank: 0 %, Σ, X
Annual Salary	Number	min: 20000, max: 150000, decimals: 2, blank: 0 %, Σ, X
Absence Days	Number	min: 0, max: 30, decimals: 0, blank: 0 %, Σ, X
Gender	Gender	blank: 5 %, Σ, X
Marital Status	Custom List	Single, Married, Divorced, Widowed, random, blank: 0 %, Σ
city	City	blank: 0 %, Σ, X
Country	Country	Portugal, Brazil, Germany, Switzerland, blank: 0 %, Σ

Figura 1 - Criação de dados fictícios através da plataforma Mockaroo

Dados em branco e modificações

Alguns campos foram intencionalmente deixados em branco (vazios) para simular cenários comuns em bases de dados reais, onde pode haver informação incompleta. Em particular, os campos *Gender* e *Job Title* foram os mais afectados por esta falta de dados, o que nos permitiu aplicar posteriormente técnicas de limpeza e gestão de dados para identificar e colmatar estas lacunas.

Alterações manuais

Após a geração inicial dos dados, foram efectuados alguns ajustamentos manuais utilizando ferramentas de manipulação de texto. Especificamente:

Modificações nas datas: Algumas datas foram alteradas para uma formatação diferente, como a utilização de “.” (ponto) em vez de “/” (barra).

1,Cecilia,Hodding,30/12/1980,25/1/2018,,Project Manager,Services,20870.55,18,Gender Female,Married,Sorocaba,018211
4,Briant,Odell,10/6/1978,12/5/2019,Recruiter,Legal,93375.83,2,Male,Single,Imbituba,Brazil
5,Franklyn,Lanchberry,15/3/1993,1/10/2023,Software Test Engineer IV,Business Development,47840.02,21,,Widowed,Avelar,Portugal
6,Elora,Tetsall,17/6/2001,15.9.2020,IP Sales,Marketing,98263.81,4,Female,Widowed,Ipatinga,Brazil
7,Gilli,Moses,30.12.1990,5/8/2014,quality Engineer,Product Management,52306.66,24,Female,Married,Cortes,Portugal
8,Cristian,Giacomasso,20/9/1991,6/2/2017,,Legal,104350.35,15,Male,Single,Formosa,Brazil
9,Skipton,Bellie,15/12/1961,4.6.2021,clinical Specialist,Product Management,110164.46,0,Male,Married,Buriti Alegre,Brazil
10,Derron,Pither,7/3/1987,26/8/2023,Developer II,Accounting,121116.47,18,Male,Widowed,Andradina,Brazil
11,Olga,Ewen,9.3.1998,31/5/2023,Financial Advisor,Sales,130331.38,5,Non-binary,Married,Soure,Portugal
12,Kathi,Fellgate,30/4/1970,7/3/2019,Software Test Engineer II,Services,122323.63,18,Female,Widowed,Santa Luzia,Brazil
13,Brucie,Herbi,30/11/2000,1/10/2015,Health Coach IV,Training,128214.57,47,Male,Single,Hamburg Bramfeld,Germany
14,Val,Heselwood,3/4/1975,1/9/2023,,Legal,94619.63,20,Male,Divorced,Orleans,Brazil
15,James,Tolossi,14/1/1997,16/11/2016,,Services,64574.91,27,Male,Divorced,Piracuruca,Brazil

Figura 2 - Dados alterados para posterior limpeza e formatação

Inserção de linhas duplicadas: Algumas linhas duplicadas foram deliberadamente introduzidas na base de dados para representar erros comuns em bases de dados reais, tornando possível a aplicação de técnicas de deteção e remoção de duplicados durante o processo de limpeza de dados. Como é possível notar, através do *Mockaroo*, foi gerado um ficheiro CSV (Figura 3) com 1000 linhas, mas o ficheiro CSV *mock_data.csv* (Figura 4) possui 1018 linhas.

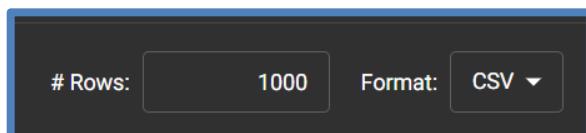


Figura 3 - Número de linhas geradas no ficheiro CSV

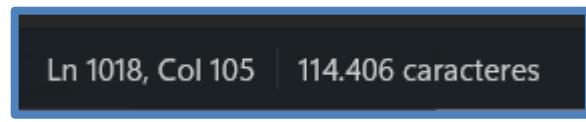


Figura 4 - Número de linhas com duplicação em ficheiro CSV

Este conjunto de dados foi concebido para proporcionar um cenário ao menos perto da realidade e permite a exploração de diversos aspectos como será exposto nas próximas etapas.

4. Processo ETL

O processo ETL (Extract, Transform, Load) é essencial para integrar dados de diferentes fontes, transformando-os em formatos padronizados adequados para análise. Garante que os dados estão prontos para uma análise eficiente, permitindo a extração de conhecimentos estratégicos e garantindo a sua precisão.

Através do ETL, podemos transformar grandes volumes de dados em informação utilizável, automatizando etapas que minimizam erros manuais e garantem a qualidade. Este processo é fundamental para que as empresas possam tomar decisões com base em dados fiáveis e históricos (Haider, 2023).

Na visão geral do projeto knime (Figura 5), é possível notar as etapas do processo ETL que serão apresentadas detalhadamente a seguir.

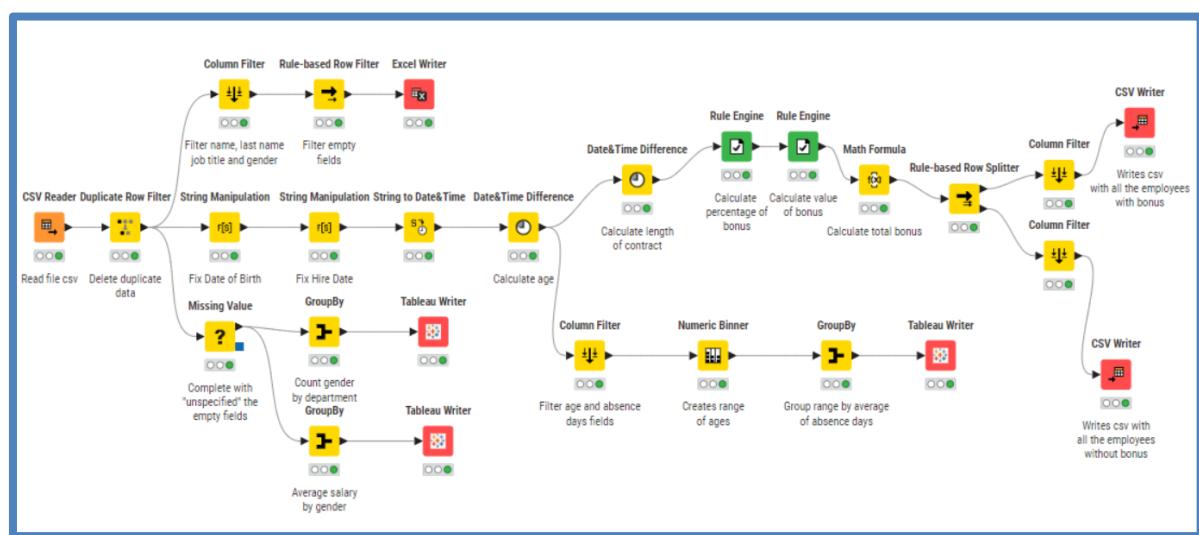


Figura 5 - Visão geral do projeto Knime

Extração

Na fase de extração, o primeiro passo foi definir um caminho relativo para o ficheiro CSV gerado no Mockaroo, como mostra a Figura 6. Este ficheiro contém informação sobre os trabalhadores, incluindo os campos employee_id, Nome próprio, Apelido, Data de

nascimento, Data de contratação, Função, Departamento, Salário anual, Dias de ausência, Sexo, Estado civil, Cidade e País.

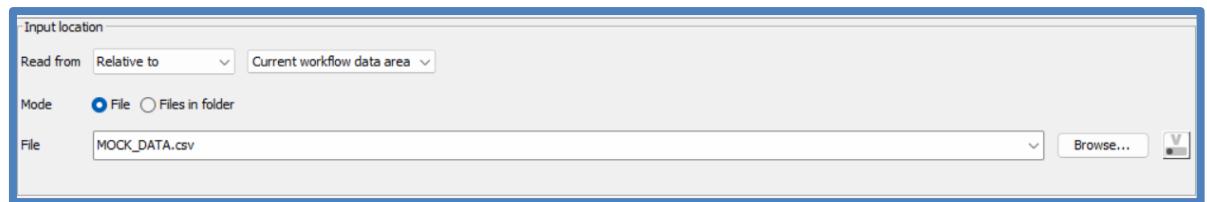


Figura 6 - Caminho relativo ficheiro CSV

O nó *CSV Reader* (Figura 5) foi utilizado para extrair os dados do ficheiro e integrá-los no processo ETL. Era importante garantir que todos os dados fossem carregados corretamente, tendo em conta que alguns campos foram propositalmente deixados em branco, simulando dados incompletos, e foram também inseridos valores duplicados para resolver problemas reais de qualidade dos dados durante as fases de transformação subsequentes.

Transformação

A primeira transformação realizada no ficheiro CSV foi através do nó *Duplicate Row Filter* para remover todas as linhas repetidas. Assim, como se pode depreender da figura 7, e tendo também em conta a figura 4, verifica-se que todas as linhas duplicadas foram removidas. A configuração do nó teve em conta todos os campos da tabela, embora apenas o *employee_id* pudesse ter sido utilizado. Além disso, foi utilizada a opção de manter o primeiro registo encontrado e descartar o segundo duplicado.

1: Filtered/Labeled Data													Flow Variables
		Rows: 1000 Columns: 13											
		RowID	employee_id	First Name	Last Name	Date of Birth	Hire Date	Job Title	Department	Annual Salary	Absence Count	Gender	
	1	Row0	1	Pollyanna	Papaminas	12/5/1969	19/3/2014	Environmental	Sales	24,817.24	30	Female	
	2	Row1	2	Minny	Grouer	25/7/1980	14/1/2022	Web Designer	Engineering	57,166.09	25	Bigende	

Figura 7 - Remoção de linhas duplicadas

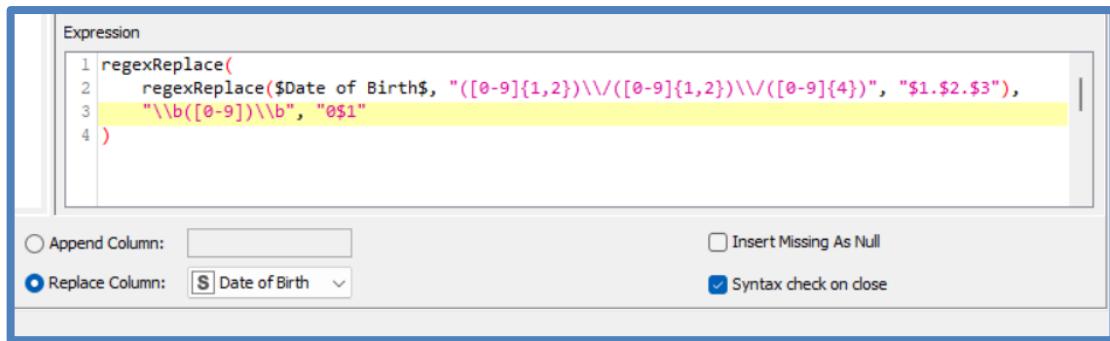
A partir desta filtragem dos dados, foram efectuadas outras transformações no ficheiro original. Como as datas estavam no formato d/m/yyyy, mas algumas também estavam no

formato d.m.yyyy. Foram utilizados dois nós de *string manipulation*, um para o campo da data de nascimento e o outro para a data de emprego.

O nó utilizado para formatar as datas exigia a aplicação de expressões regulares, pelo que foi aplicada a expressão apontada na figura 8. O código utiliza duas expressões regulares para transformar e padronizar a coluna *Date of Birth*.

A primeira expressão converte as datas do formato m/d/yyyy para m.d.yyyy, substituindo as barras por pontos. Em seguida, um segundo regex insere um zero inicial em dias ou meses representados por apenas um dígito, padronizando-os para dois dígitos.

Através deste nós de manipulação ainda é possível sobrescrever a coluna referente as datas alteradas, evitando que sejam apenas adicionados dados, aumentando o tamanho do ficheiro sem necessidade.



The screenshot shows a software interface for manipulating data. In the main window, there is a code editor with the following content:

```
Expression
1 regexReplace(
2     regexReplace($Date of Birth$, "(\\d{1,2})\\/(\\d{1,2})\\/(\\d{4})", "$1.$2.$3"),
3     "\\b([0-9])\\b", "0$1"
4 )
```

Below the code editor, there are several options:

- Append Column: [empty input field]
- Insert Missing As Null
- Replace Column:
- Syntax check on close

Figura 8 - Regex para manipulação das datas

As transformações seguintes incluem a adição de mais duas colunas ao ficheiro, a partir do cálculo da diferença de tempo, utilizada para calcular a idade de cada empregado e a duração do contrato. Para tal, foram utilizados dois nós *Date&time Difference* (Figura 9), que utilizaram o tempo de execução do programa atual para efetuar os cálculos.

Além disso, foi escolhida a granularidade do valor final em anos, e foram nomeadas as novas colunas que irão armazenar esta nova informação. Assim, a coluna que armazenará a idade foi nomeada de *Age*, e a coluna do tempo de contrato como *Length of Contract*.

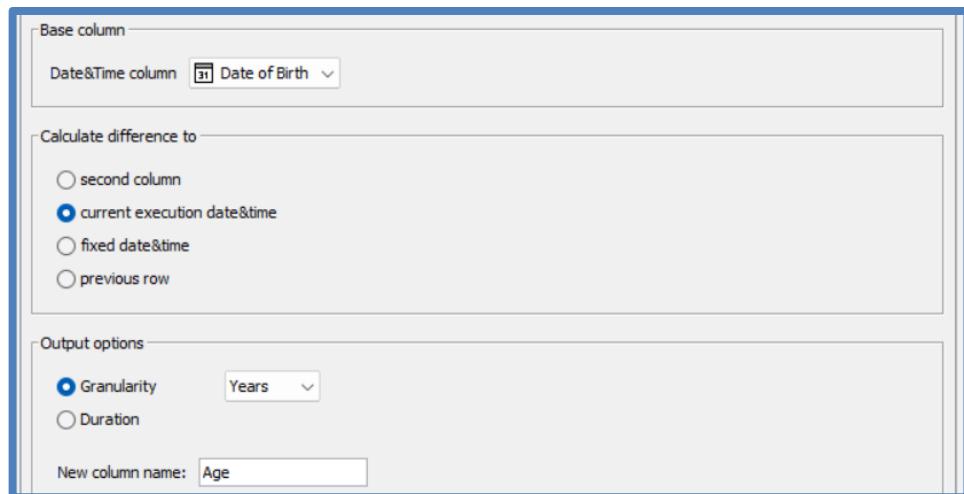


Figura 9 - Nó para calcular idade e tempo de contrato

De seguida, como já foi referido no tópico anterior, foram aplicadas algumas regras com base no tempo de serviço de cada colaborador na empresa, bem como nas suas ausências anuais. Para o efeito, foram utilizados dois nós do *Rule Engine*. Para que este nó pudesse aplicar a lógica necessária, era preciso inserir expressões e adicionar uma nova coluna ao final do ficheiro utilizando a opção *Append Column* (TosinLitcs, 2021). O processo foi repetido tanto para os dias em falta (Figura 10) como para o tempo de contrato (Figura 11), tendo em conta a lógica de cada um.

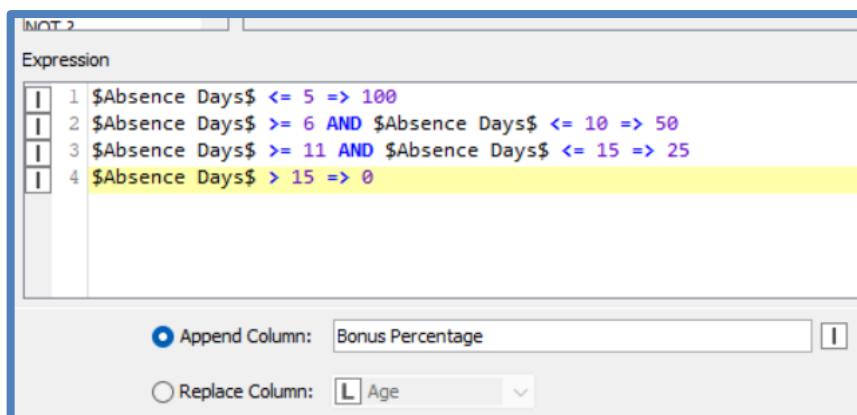


Figura 10 - Expressão para construção de coluna com valor de porcentagem do bônus através da quantidade de dias de ausência

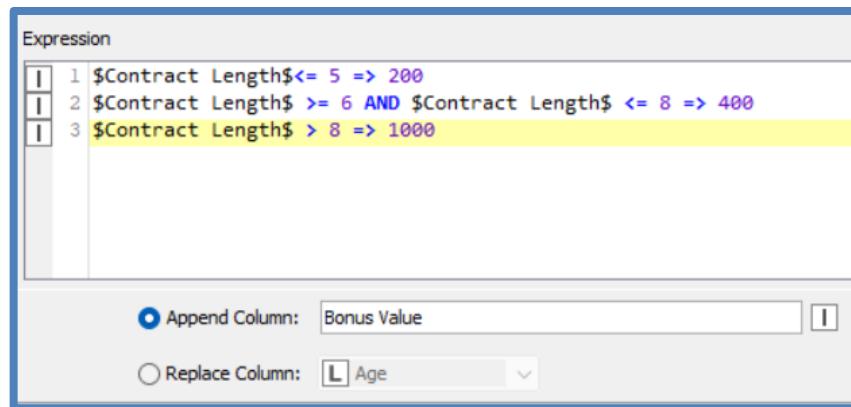


Figura 11 - Expressão para construção de coluna com o valor do bônus pelo tempo de contrato

As transformações realizadas pelos nós *Math Formula* e *Rule-based Row Splitter* consistem em utilizar as duas colunas criadas pelos nós *Rule Engine* para calcular o valor total de bônus para cada funcionário e acrescentar mais uma coluna ao ficheiro contendo estes valores (Figura 12), e em seguida dividir em sessões diferentes (Figura 13), uma com os funcionários que possuem direito ao bónus e outra com os demais.

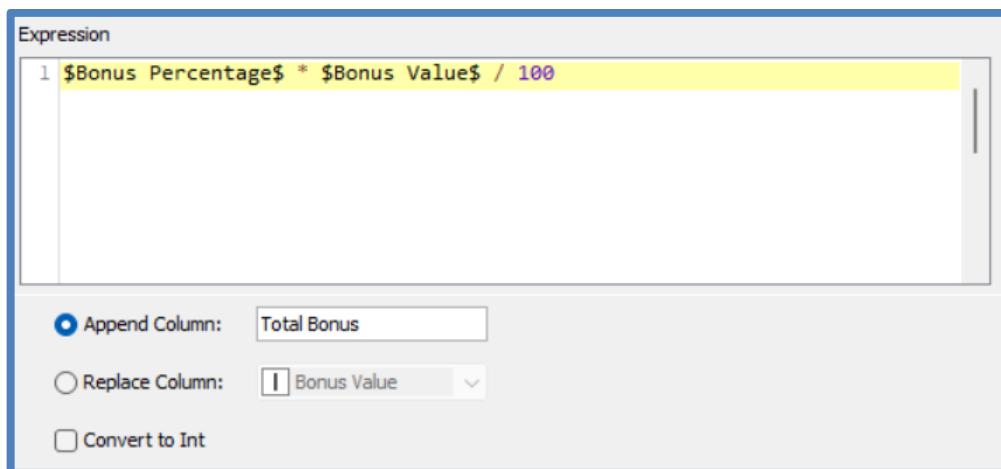


Figura 12 - Cálculo de valor total do bónus

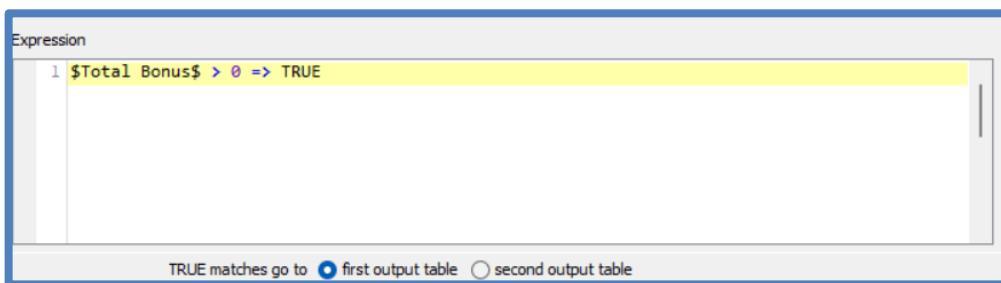


Figura 13 - Filtragem de funcionários com e sem bonus

O nó *Rule-based Row Filter* também foi utilizado para encontrar apenas as linhas do ficheiro em que havia dados vazios. A lógica a seguir foi usada para isso, pois já se sabia que apenas esses dois campos tinham dados que poderiam estar vazios.

MISSING \$Gender\$ => TRUE

MISSING \$Job Title\$ => TRUE

O nó *Column Filter* foi utilizado 4 vezes no processo. Este nó é responsável por filtrar os dados e utilizar apenas as colunas desejadas. Neste caso, foi utilizado para filtrar apenas as colunas *employee id*, *first name*, *last name* e *bonus value* (Figura 14) para os empregados que tinham ou não direito a bónus, e para separar *employee id*, *first name*, *last name*, *job title* e *gender* (Figura 15) para filtrar os campos vazios, como apresentado anteriormente.

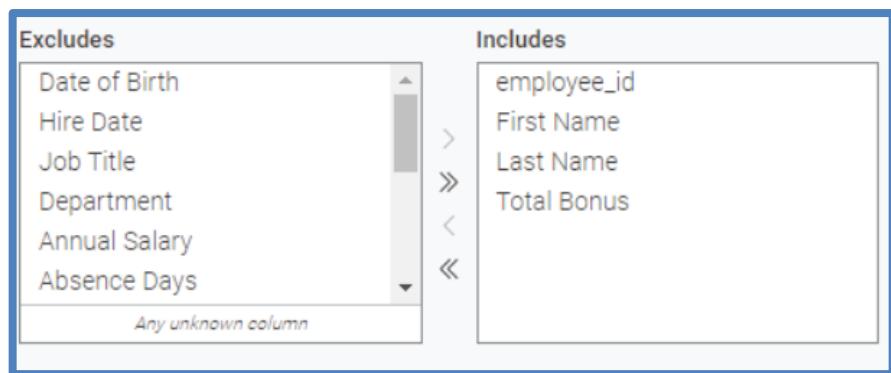


Figura 14 – Filtragem de colunas

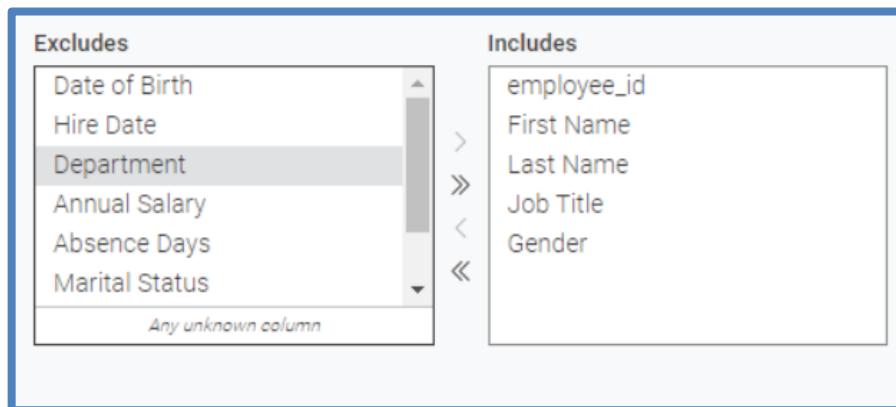


Figura 15 - Filtragem de colunas (2)

A filtragem de colunas também foi utilizada para que fossem separadas apenas as informações de idade e dias de ausência dos funcionários (Figura 16) para posterior apresentação dos dados.

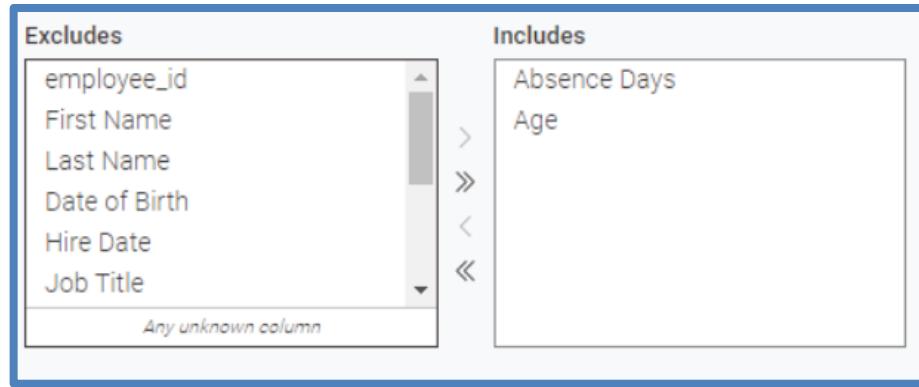


Figura 16 - Filtragem de colunas (3)

O nó *Missing Value* foi utilizado para preencher com *Unspecified* os campos em branco dos dados lidos, para que estes funcionários aparecessem em análises como a de gênero por departamento. Apesar deste dado estar em branco, eles entrarão na estatística como *unspecified*.

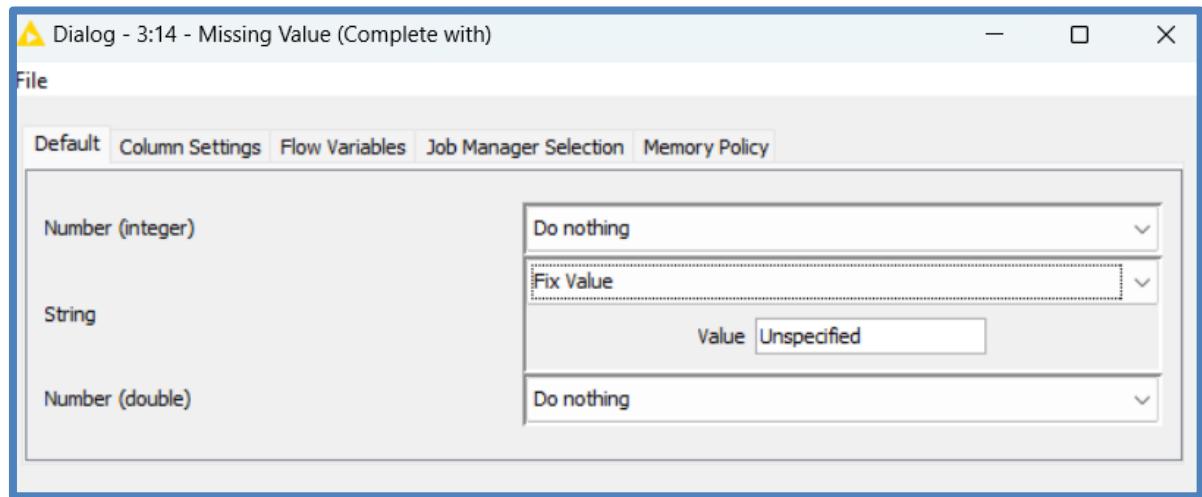


Figura 17 - Preenchimento de campos vazios

O nó *Numeric Binner*, utilizado após a separação das colunas *Age* e *Absence Days*, é apresentado para estabelecer faixa de idades. O que este nó faz é receber uma sequência de valores, um em cada linha, e assim os agrupa de acordo com um intervalo. Neste caso, os intervalos são para as idades conforme apresentado na figura 18.

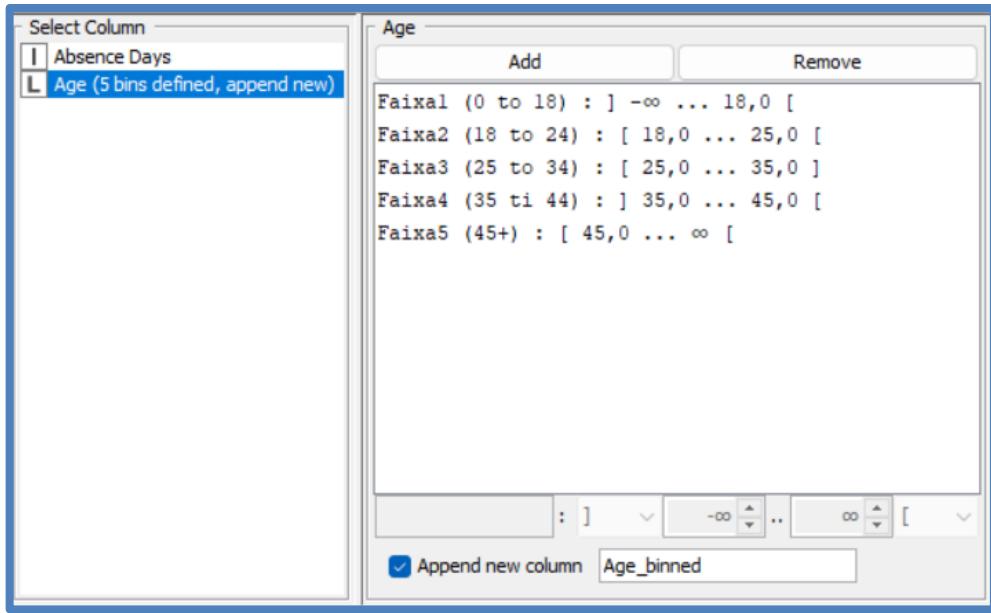


Figura 18 - Criação de faixas de idade com o nó Numeric Binner

Por fim, o nó *Group By* (Figura 19) foi utilizado para agrupar dados em categorias e efetuar operações de agregação sobre eles, tais como médias e contagem. No processo, o nó foi aplicado três vezes com objectivos diferentes:

1. **Grupo de Género por Departamento:** Para contabilizar quantos colaboradores de cada género existem em cada departamento.
2. **Calcular o Salário Médio por Género:** Neste caso, o nó foi configurado para calcular a média salarial, agrupando por género.
3. **Média de Ausências por Faixa Etária:** Por fim, o nó foi utilizado para calcular o número médio de ausências de acordo com os diferentes grupos etários.

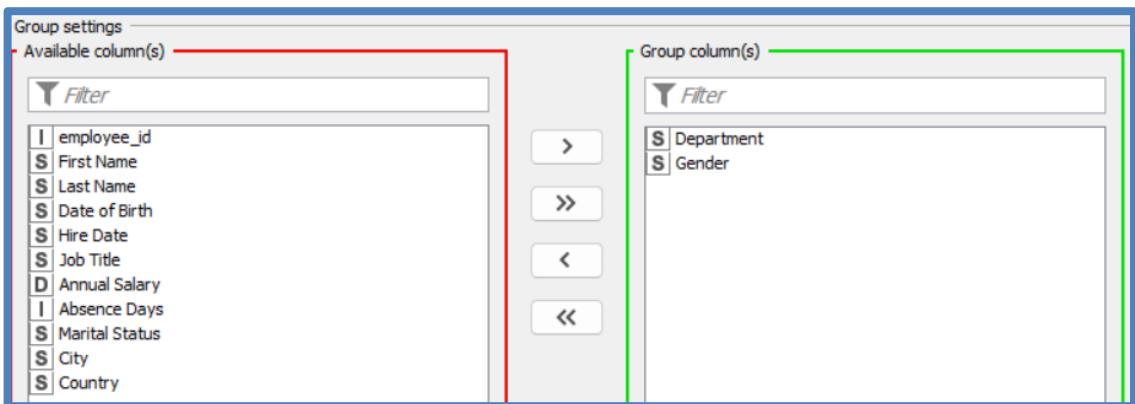


Figura 19 - Nó Group By para contagem de gênero por departamento

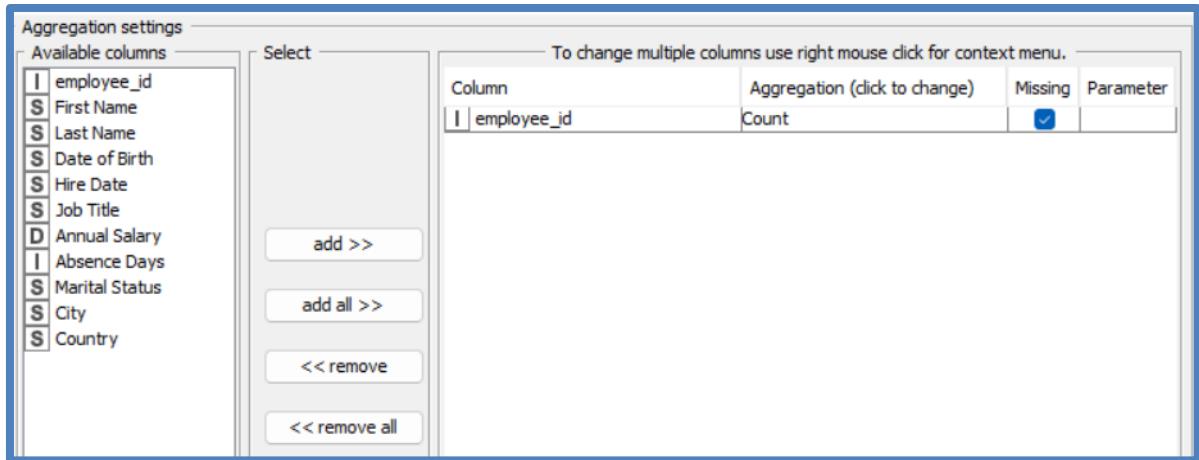


Figura 20 - Configuração de nó Group By

Carga (Load)

A última etapa do processo de ETL é o carregamento, onde os dados processados e transformados são organizados e enviados para armazenamento final ou integração com ferramentas de análise. Nesta fase, todos os dados ficam acessíveis à empresa para apoio às decisões estratégicas e operacionais.

Foram utilizados os seguintes nós de carga para esta fase:

Excel Writer: Este nó foi configurado para gerar um ficheiro Excel contendo os registo com campos em branco, como o género ou o cargo, facilitando o seu preenchimento posterior. Este formato é facilmente editável pelos RH, permitindo a correção e complemento de informação pendente. Para configuração foi utilizado um diretório local.

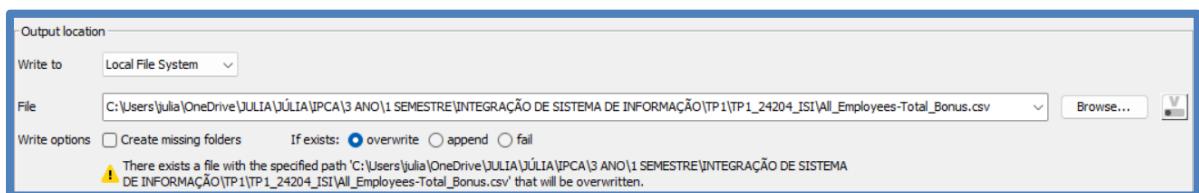


Figura 21 - Configuração de nó excel writer

Tableau Writer: Com três nós do *Tableau Writer*, foram criados ficheiros de extração específicos para o Tableau. Estes ficheiros organizam os dados em três categorias de análise: (1) distribuição do género por departamento, (2) salário médio por género e (3) média de

ausências por grupo etário. Estas folhas de cálculo foram concebidas para serem exploradas visualmente no Tableau, onde as análises são detalhadas no tópico 6. Todavia, os ficheiros também foram salvos localmente, mas uma futura alteração no projeto poderá trocar este nó por um Tableau Server, para que o processo se torne mais independente de diretório local.

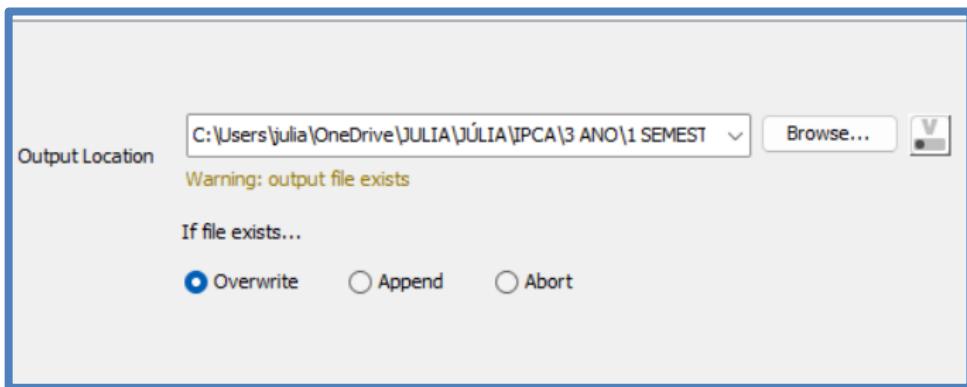


Figura 22 - Configuração do nó Tableau Writer

CSV Writer: Finalmente, foram utilizados dois nós *CSV Writer* para exportar dados sobre os empregados elegíveis e não elegíveis para receber bónus. Os ficheiros CSV foram criados para apoiar os RH no cálculo e pagamento dos bónus e para comunicar diretamente aos empregados que não os receberam, encorajando-os a trabalhar mais e a reduzir as suas ausências, maximizando assim a sua oportunidade de receber um bónus futuro.

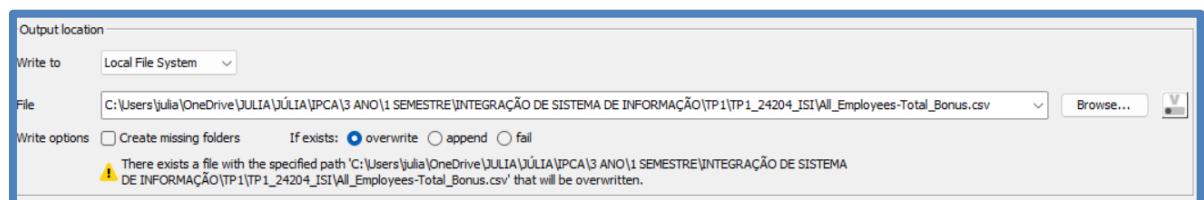


Figura 23 - Configuração do nó CSV Writer

5. Jobs

Cálculo de Bónus salarial

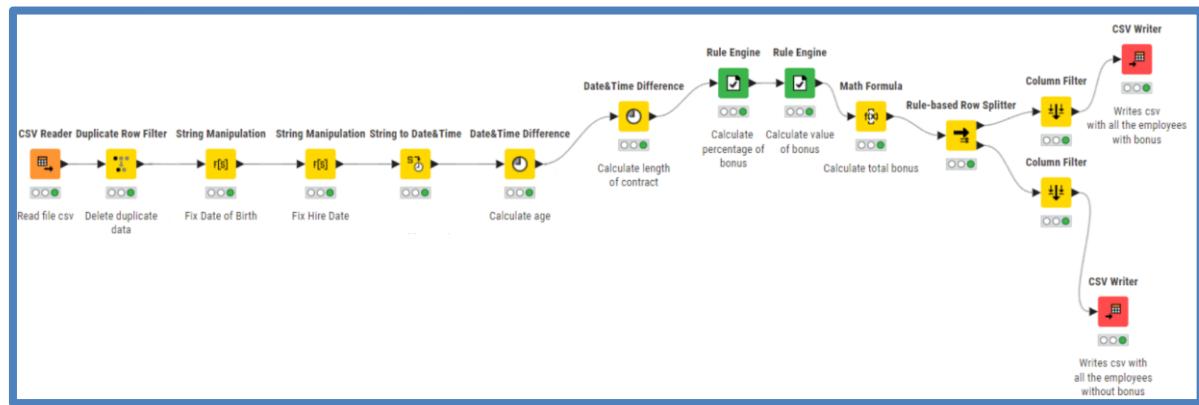


Figura 24 - Processo de cálculo de bónus salarial

O primeiro *job*, centrado no cálculo do abono salarial, começa com a uniformização das datas de nascimento e de admissão dos trabalhadores. Este passo é essencial para garantir que os dados estão uniformes e preparados para o cálculo da idade e do tempo de contrato, factores fundamentais para a atribuição de prémios. Com esta informação temporal em mãos, calcula-se a idade e o tempo de contrato do colaborador.

Estes dados alimentam um conjunto de regras que atribuem percentagens de prémios em função da experiência e da idade, reflectindo o valor da antiguidade e da maturidade na empresa. Depois de definidas as percentagens de bónus, são aplicadas regras específicas para calcular o valor monetário correspondente, ajustando-o a um valor final preciso.

Posteriormente, os colaboradores são divididos entre os que recebem o bónus e os que não recebem, sendo os dados exportados para dois ficheiros distintos. Estes ficheiros servem de referência para o departamento de recursos humanos, tanto para implementar o pagamento como para comunicar os critérios de incentivo aos que ficaram de fora da lista de prémios.

Extração de campos vazios

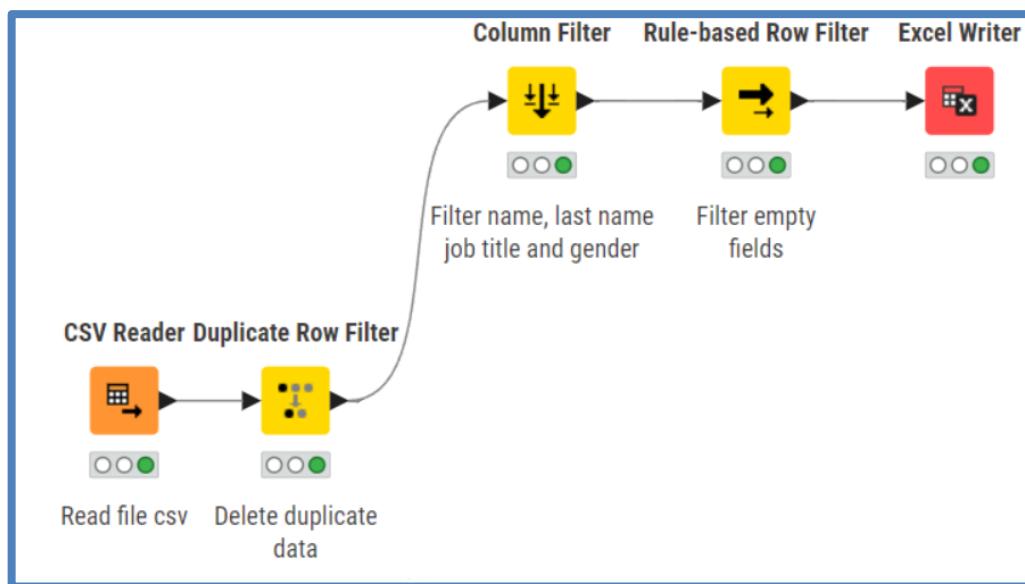


Figura 25 - Processo de extração de campos vazios

A segunda tarefa tem como objetivo identificar os campos em branco que, se não forem tratados, podem comprometer a análise. Para isso, inicialmente são filtradas apenas as colunas que podem conter dados incompletos, como *First Name*, *Last Name*, *Job Title* e *Gender*, de forma a centrar a análise nos campos mais críticos. Em seguida, utilizando um conjunto de regras, os registos com campos em branco são identificados e armazenados num ficheiro Excel. Este ficheiro fornece uma lista útil para o departamento de RH preencher as informações em falta, garantindo a integridade e a exatidão dos dados.

Estatísticas

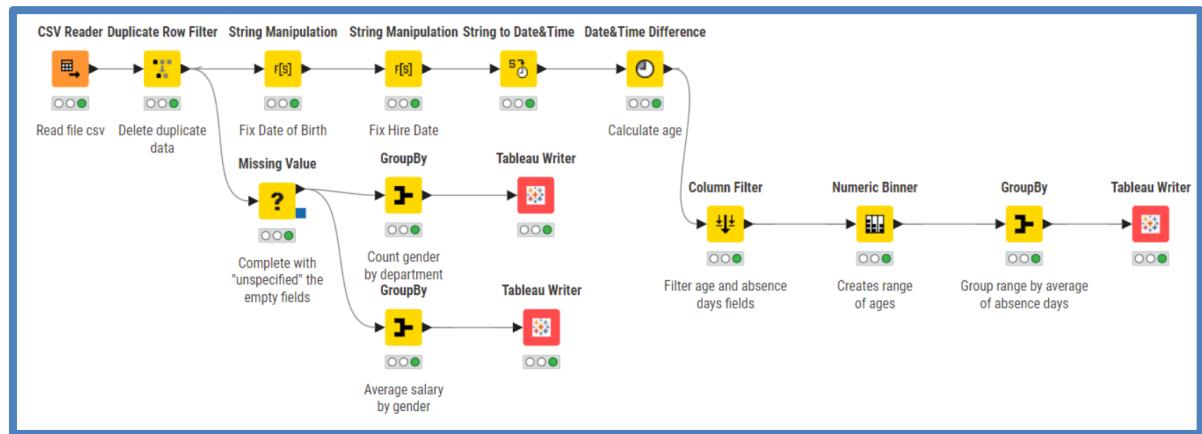


Figura 26 - Processo de obtenção de dados estatísticos

O terceiro e último *job* centra-se na análise estatística e subdivide-se em duas frentes principais: ausência por grupo etário e salário médio por género. Na primeira frente, a análise incide sobre os padrões de ausência entre diferentes grupos etários. Após a uniformização das datas e o cálculo da idade, os colaboradores são agrupados em categorias etárias específicas, permitindo avaliar o comportamento das ausências em função da idade. O número médio de dias de ausência em cada grupo etário é então calculado, revelando padrões de ausência que podem influenciar as estratégias de contratação e retenção. Esta informação é depois exportada para o Tableau, onde gráficos interactivos ajudam a visualizar as taxas de absentismo de acordo com a idade dos colaboradores.

Na segunda frente, centrada na distribuição por género e no salário médio, o trabalho utiliza os dados já padronizados para preencher os campos em branco de Género e Cargo. Depois, com a certeza de que os dados estão completos, a informação é agrupada para analisar a proporção de género em cada departamento e calcular o salário médio por género. Estes dados são também enviados para o Tableau, facilitando a análise visual da distribuição de géneros nos departamentos e possíveis diferenças salariais entre géneros.

6. Tableau

No Tableau, o processo foi desenvolvido de forma a que cada ficheiro gerado pelo KNIME seja automaticamente lido e atualizado em cada execução do livro de trabalho. Foram criadas três planilhas, uma para cada fonte de dados processada no KNIME, proporcionando uma visualização focada nas necessidades da análise.

Em cada folha de cálculo, os dados foram organizados e transformados em diferentes tipos de gráficos para facilitar uma interpretação dinâmica e visualmente apelativa. Os gráficos foram ajustados para destacar tendências e comparações, permitindo à empresa observar rapidamente a distribuição de géneros por departamento, os salários médios por género e as taxas de ausência por grupo etário.

Os gráficos desenvolvidos no Tableau revelam informações valiosas sobre a relação entre a idade e o número de ausências, a distribuição do género por departamento e o salário médio anual por género.

No gráfico das ausências por grupo etário, verifica-se uma tendência decrescente do número de ausências à medida que a idade dos trabalhadores aumenta. Este padrão indica que os colaboradores mais jovens têm uma taxa de absentismo mais elevada, o que sugere a eventual necessidade de estratégias específicas para este grupo, como programas de estágio ou de trainees, para garantir um maior empenho.



Figura 27 - Gráfico de média de ausência por faixa etária

O gráfico da distribuição dos géneros por departamento revela uma proporcionalidade equilibrada, com uma representação relativamente uniforme de cada género nas diferentes áreas.

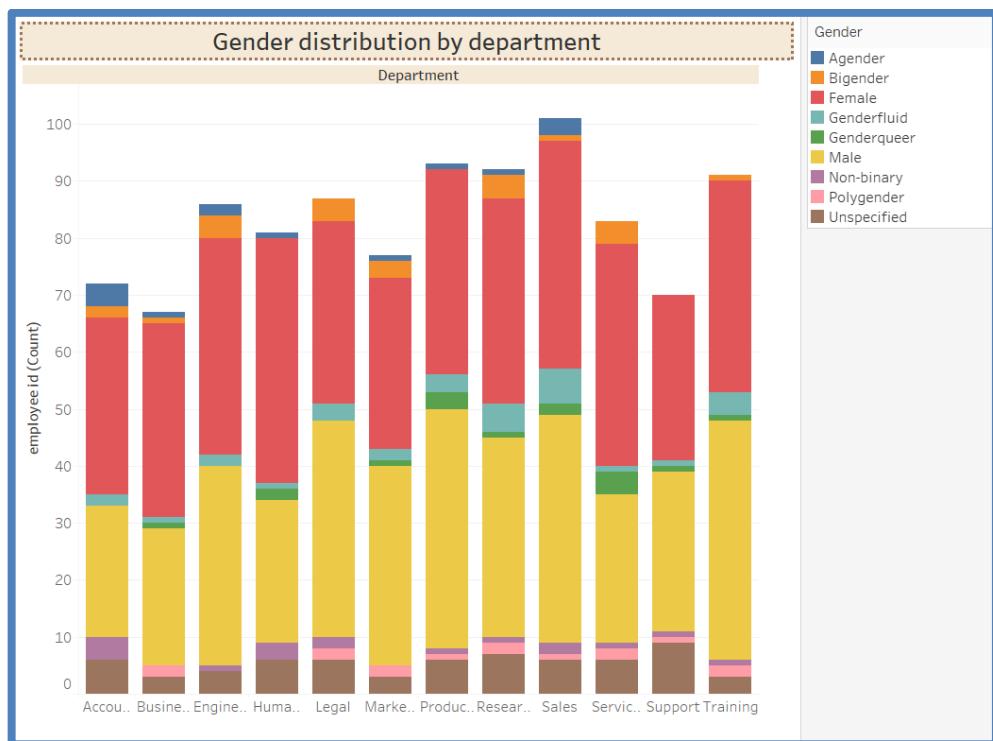


Figura 28 - Gráfico de distribuição de gênero por departamento

Por último, o gráfico dos salários médios anuais por género mostra que o género masculino tem um salário médio mais elevado, seguido do género feminino. No entanto, é o género bigénero que apresenta o salário médio mais baixo, o que merece uma análise mais aprofundada para identificar as causas desta disparidade.

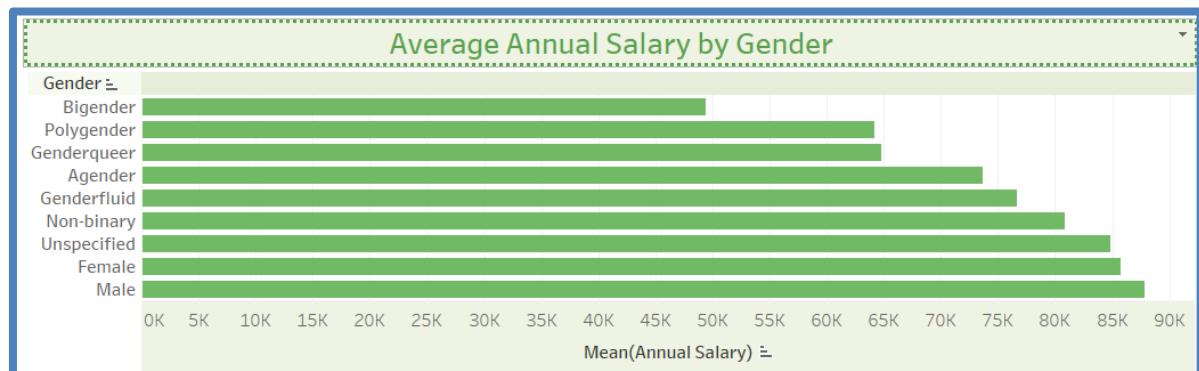


Figura 29 - Gráfico de distribuição de média salarial por gênero

A ferramenta Tableau é extremamente versátil, permitindo à empresa explorar e manipular estes dados de várias formas. No futuro, pode ser possível alargar a utilização do Tableau, adicionando novas fontes de dados ou simplesmente actualizando os ficheiros CSV originais e reexecutando o processo no KNIME para uma atualização automática das análises. Esta abordagem garante flexibilidade e escalabilidade para a empresa, facilitando a integração contínua de dados nas suas operações analíticas.

7. Conclusão e Trabalhos Futuros

A análise dos dados da empresa revelou uma série de informações valiosas que não só permitiram identificar lacunas na informação, como também forneceram uma base para futuras decisões estratégicas. Durante o processo, verificou-se que alguns campos estavam em branco, como o género e o cargo de alguns colaboradores. Para resolver esta questão, foi gerado um ficheiro Excel com a lista destas inconsistências. No futuro, será necessário formatar este Excel e enviá-lo por correio eletrónico para o departamento de Recursos Humanos, para que este possa preencher a informação em falta.

Em termos de análise da distribuição de género por departamento, os dados apresentados no Tableau indicam que a distribuição é geralmente equilibrada. No entanto, a diferença salarial entre géneros é alarmante e requer uma investigação mais aprofundada. É fundamental que a empresa compreenda as razões subjacentes a esta diferença salarial e analise se existem outros factores correlacionados que possam contribuir para este cenário.

Para além disso, verificou-se que os trabalhadores mais jovens têm uma taxa de absentismo significativamente mais elevada. Este facto deve alertar a empresa para ponderar a implementação de contratos de estágio ou de trainee para estas faixas etárias. Tal abordagem permitiria estabelecer um período de avaliação do empenho e motivação dos novos colaboradores, mitigando a possível contratação de indivíduos que não demonstrem interesse em permanecer a longo prazo.

No que diz respeito à manipulação dos dados para análise no Tableau, foi necessário recorrer a ficheiros locais, o que, apesar de funcional, traz limitações em termos de automatização e atualização. No futuro, pretende-se estabelecer uma ligação ao servidor, o que permitirá um processo mais automatizado, reduzindo a dependência de regístos locais e garantindo que a análise é sempre efectuada com dados actualizados.

Por fim, a empresa obteve ainda os valores dos bónus atribuídos aos seus colaboradores, bem como a lista dos que não os receberam. Esta informação pode ser utilizada para o envio de mensagens de incentivo, com o objetivo de motivar os colaboradores a melhorar o seu desempenho e empenho nas actividades diárias.

Desta forma, este trabalho serviu não só para analisar a situação atual da empresa, mas também para delinear um conjunto de acções futuras que visam a melhoria contínua do ambiente de trabalho e a promoção de uma cultura organizacional mais inclusiva e produtiva.

8. Referências

Haider, K. (2023, September 28). *What is ETL? - Extract, Transform, Load Explained.* Astera. <https://www.astera.com/type/blog/etl/>

Zurich, 2024

Zurich, A. (2024). *KNIME Tableau Integration User Guide.* Knime.com.
https://docs.knime.com/202406/tableau_integration_user_guide/index.html#_tableau_online_setup

TosinLitics. (2021, November 10). *Day 14 - Rule Engine Node - 30 Days of KNIME.* YouTube. <https://www.youtube.com/watch?v=LhwpEuovKpo>