Universidade do Minho
Escola de Engenharia

Catarina da Cunha Malheiro da Silva Pereira

Preços de Veículos Australianos

7 de janeiro de 2024

Preços de Veículos Australianos
Catarina Pereira
Uminho | 2024

Universidade do Minho
Escola de Engenharia

Catarina da Cunha Malheiro da Silva Pereira

Preços de Veículos Australianos

Trabalho de Individual
Engenharia em Telecomunicações e Informática
Inteligência Artificial para as Telecomunicações

Trabalho efetuado sob a orientação de:
Professora Doutora Dalila Durães

7 de janeiro de 2024

Índice

Índice de Figuras iii

Lista de Acrónimos v

Acrónimos v

Glossário v

1 Introdução 1
1.1 Enquadramento . . . 1

2 Revisão da Literatura 2

3 Abordagem Metodológica 3

4 Trabalho Realizado 5
4.1 Carregar o *dataset* descarregado . . . 5
4.2 Tratamento dos Dados . . . 5
4.2.1 Cell Splitter . . . 6
4.2.2 Collumn Filter . . . 7
4.2.3 String Manipulation . . . 8
4.2.4 String To Number . . . 8
4.2.5 Category To Number . . . 9
4.2.6 Numeric Outliers . . . 10
4.2.7 Missing Values . . . 10
4.2.8 Auto-binner . . . 11
4.3 Exploração dos Dados . . . 11
4.4 Filtragem . . . 13
4.5 Modelação dos dados . . . 14
4.5.1 Modelo Supervisionado de Regressão Linear . . . 15
4.5.2 Modelo Supervisionado de Árvores de Decisão . . . 18
4.5.3 Modelo Não Supervisionado de *K-Means Clustering* . . . 20

5 Conclusão 22

Referências Bibliográficas 23

Índice de Figuras

1 Previsão do preço de carros usando três modelos. . . . 5
2 Carregamento do *dataset* "Australian Vehicle Prices" no nó **CSV Reader**. . . . 5
3 Metadono "Tratamento dos Dados". . . . 6
4 Dentro do metadono da "Tratamento dos Dados". . . . 6
5 Coluna do atributo *Engine*. . . . 6
6 Nó **Cell Splitter**. . . . 6
7 Janela de configuração do **Cell Splitter**. . . . 7
8 Nó **Column Filter**. . . . 7
9 Janela de configuração do **Column Filter**. . . . 7
10 Tabela de saída do nó **Column Filter**. . . . 7
11 Nó **String Manipulation**. . . . 8
12 Janela de configuração de um "String Manipulation". . . . 8
13 Nó **String To Number**. . . . 8
14 Janela de configuração do "String To Number". . . . 8
15 Janela de configuração do nó **Category To Number**. . . . 9
16 Saída do nó **Category To Number**. . . . 9
17 Nó **Numeric Outliers**. . . . 10
18 Janela de configuração de "Numeric Outliers". . . . 10
19 Nó **Box Plot (JavaScript)**. . . . 10
20 Janela de configuração de "Box Plot (JavaScript)". . . . 10
21 Nó **Missing Values**. . . . 10
22 Tratamentos dos "Missing Values". . . . 10
23 Nó **Auto-binner**. . . . 11
24 Configuração do "Auto-binner". . . . 11
25 Metadono "Exploração de dados". . . . 12
26 Dentro do metadono da "Exploração de dados". . . . 12
27 Tabela Numérica do Nó **Data Explorer**. . . . 12
28 Estatísticas. . . . 13
29 Scatter Matrix. . . . 13
30 Correlação linear. . . . 13
31 Metadono "Filtragem". . . . 14
32 Dentro do metadono da "Filtragem". . . . 14
33 Tabela de saída do nó leitor **Column Filter**. . . . 14
34 Remover de consideração a linha de dados possivelmente incorreta. . . . 14
35 Partição dos dados 75% – 25%. . . . 15
36 Metanodo **Modelos de Aprendizagem de Regressão Linear**. . . . 15
37 Dentro do metadono da "Modelos de Aprendizagem de Regressão Linear". . . . 15
38 Configuração do nó **Linear Regression Learner**. . . . 16
39 Configuração do nó **Numeric Scorer**. . . . 16
40 Output do nó **Numeric Scorer**. . . . 16
41 Histograma do Erro Residual. . . . 17
42 Histograma do Preço. . . . 17
43 Histograma do Preço Previsto. . . . 17
44 Gráfico de Dispersão. . . . 18
45 Metanodo **Modelos de Aprendizagem de Árvores de Decisão**. . . . 18
46 Dentro do metadono da "Modelos de Aprendizagem de Árvores de Decisão". . . . 18
47 Configuração do nó **Decision Tree Learner**. . . . 19
48 Configuração do nó **Scorer (JavaScript)**. . . . 19
49 Output do nó **Numeric Scorer**. . . . 19
50 Metanodo **Modelo Não Supervisionado de K-Means Clustering**. . . . 20
51 Dentro do metadono da "Modelo Não Supervisionado de K-Means Clustering". . . . 20

52 Configuração do nó **K-means**. 20
53 Nós **Color Manager** e **Shape Manager**. 20
54 Configuração do nó **Color Manager**. 21
55 Configuração do nó **Shape Manager**. 21
56 Configuração do nó **Scatter Plot (Legacy)**. 21
57 *Output* do nó **Scatter Plot (Legacy)**. 21

Acrónimos

CRISPDM	Cross Industry Standard Process for Data Mining
KNIME	Konstanz Information Miner
UC	Unidade Curricular

1 Introdução

Este relatório faz parte da Unidade Curricular (UC) Inteligência Artificial para as Telecomunicações, do 1° semestre do 1° ano do Mestrado em Engenharia de Telecomunicações e Informática. Este projeto individual foi proposto pela docente Dalila Durães e tem como principal objetivo responder às várias tarefas propostas.

No dinâmico cenário automovel australiano, a compreensão dos fatores que influenciam os preços de veículos é crucial para consumidores, revendedores e analistas de mercado. O conjunto de dados "Australian Vehicle Prices" para o ano de 2023 surge como uma fonte abrangente de informações, oferecendo insights valiosos sobre a dinâmica do mercado de carros na Austrália. Com mais de 16000 registos, este conjunto de dados possui detalhes cruciais sobre uma variedade de marcas, modelos, tipos e características de veículos, refletindo a diversidade e complexidade do mercado automovel.

1.1 Enquadramento

A indústria automovel australiana é um ecossistema bastante animado, onde variáveis como marca, modelo, tipo de veículo, ano de fabrico e localização desempenham papéis fundamentais na determinação dos preços dos carros. Neste contexto, o conjunto de dados "Australian Vehicle Prices" oferece uma visão aprofundada, permitindo não apenas a previsão de preços com modelos de *machine learning*, mas também a análise do mercado e a identificação de características determinantes nos preços dos veículos.

Este estudo visa explorar as oportunidades que esse conjunto de dados oferece, desde a previsão de preços de carros até análises de mercado e identificação de características-chave. Ao longo desta análise, pretende-se fornecer uma visão abrangente dos fatores que impulsionam o mercado automovel australiano, promovendo uma compreensão mais profunda e informada para todos os interessados nesse setor dinâmico.

2 Revisão da Literatura

Este capítulo apresenta os conceitos teóricos necessários ao desenvolvimento deste projeto, assim como a respetiva revisão da literatura.

Este relatório fornece uma exploração aprofundada do Konstanz Information Miner (KNIME), projetada para capacitar utilizadores na área de ciência e análise de dados. KNIME oferece um ambiente versátil e fácil de usar, permitindo que indivíduos percorram perfeitamente o fluxo de trabalho de análise de dados, desde a integração de dados até análise avançada, modelação e implementação.

KNIME destaca-se como uma plataforma robusta que fornece uma interface intuitiva, com pouco código/sem código, tornando-a acessível a utilizadores em vários níveis de aptidão. Este artigo tem como objetivo elucidar os principais recursos e funcionalidades do KNIME e a sua aplicabilidade em cenários de análise de dados do mundo real.

Os recursos principais desta aplicação são as seguintes:

1. Integração de dados e automação de fluxo de trabalho;
2. Análise de dados e *machine learning*;
3. Visualização dos dados;
4. Implementação do *dataset*.

Concluindo, o KNIME surge como uma solução abrangente para ciência e análise de dados ponta a ponta. A sua interface amigável, recursos robustos e suporte ativo da comunidade o posicionam-no como um ativo valioso para indivíduos e organizações que procuram aproveitar o poder dos dados para uma tomada de decisão informada.

A metodologia utilizada neste projeto é o Cross Industry Standard Process for Data Mining (CRISPDM). Esta metodologia é a metodologia mais utilizada para projetos de verificação de dados [1].

Kaggle é uma plataforma global como o AirBnB para cientistas de dados, que oferece um espaço colaborativo para indivíduos de todo o mundo se envolverem na ciência de dados, *machine learning* e problemas de análise preditiva [2]. O impacto do Kaggle é multifacetado.O rápido crescimento da plataforma deve-se ao conteúdo de alta qualidade compartilhado pelo Kaggle, incluindo conjuntos de dados, posts em fóruns e kernels. As competições Kaggle são iniciadas por anfitriões que fornecem dados e descrevem um problema. Além de competições, Kaggle oferece conjuntos de dados abertos, cadernos de *machine learning* e tutoriais para ajudar os utilizaadores a aprender e praticar ciência de dados e técnicas de *machine learning*.

O projeto consiste em seis fases distintas.

1. Avaliar os dados;
2. Recolha de dados;
3. Preparação dos dados;
4. Escolha do modelo;
5. Treino dos dados;
6. Validação dos dados;
7. Apresentação dos dados.

3 Abordagem Metodológica

Este capítulo apresenta os conceitos práticos necessários ao desenvolvimento deste projeto, mais precisamente sobre o *dataset*.

Os atributos do *dataset* indicado:

- *Brand:* Nome do fabricante do carro
- *Year:* Ano de fabrico do carro
- *Model:* Nome ou código do modelo do carro
- *Car/Suv:* Tipo de carro (carro ou SUV)
- *Title:* Título ou descrição do carro
- *UsedOrNew:* Condição do carro (usado ou novo)
- *Transmission:* Tipo de transmissão (manual ou automática)
- *Engine:* Capacidade ou potência do motor (em litros ou quilowatts)
- *DriveType:* Tipo de tração (roda dianteira, traseira ou integral)
- *FuelType:* Tipo de combustível (gasolina, diesel, híbrido ou elétrico)
- *FuelConsumption:* Taxa de consumo de combustível (em litros por 100 km)
- *Kilometres:* Distância percorrida pelo carro (em quilometros)
- *ColourExtInt:* Cor do carro (exterior e interior)
- *Location:* Localização do carro (cidade e estado)
- *CylindersinEngine:* Número de cilindros no motor
- *BodyType:* Formato ou estilo da carroceria do carro (sedan, hatchback, cupê, etc.)
- *Doors:* Número de portas no carro
- *Seats:* Número de assentos no carro
- *Price:* Preço do carro (em dólares australianos)

Mediante o *dataset* disponível, pretende-se desenvolver um modelo de regressão linear, utilizando táticas adequadas para tratamento de valores discretos.

O modelo tem três tipos de casos de uso potenciais e estes são:

- **Previsão de Preço:** Prever o preço de um carro com base nas características e localização usando modelos de *machine learning*.
- **Análise de Mercado:** Explorar as tendências do mercado e a procura de diferentes tipos de carros na Austrália usando estatísticas descritivas e técnicas de visualização.
- **Análise de Recursos:** Identificar os recursos mais importantes que afetam os preços dos carros e como eles variam entre diferentes marcas, modelos e locais usando análise de correlação e regressão linear.

O modelo escolhido para este trabalho será a previsão do preço dos carros na Austrália, segundo o que foi dado na aula teórica do dia 22 de novembro de 2023.

A especificção dos atributos mais importantes que afetam os preços dos carros:

- *Brand*;
- *Year*;
- *Car/Suv*;
- *UsedOrNew*;
- *Transmission*;
- *Kilometres*;
- *CylindersinEngine*;
- *Doors*;
- *Seats*;
- *Price*.

4 Trabalho Realizado

O desenvolvimento do *workflow* será dividido em várias tarefas distintas, cada uma para a contribuição da construção de uma análise sólida e abrangente. A seguir, apresenta-se uma breve visão geral de cada tarefa.

Com o propósito de entender o relacionamento dos dados e a sua qualidade, procedeu-se à sua exploração (*Data Understanding*) através da utilização dos componentes *Linear Correlation*, *Statistics*, *Crosstab* e *Data Explorer*, permitindo a extração de valores estatísticos para sustentar a tomada de decisão da fase seguinte.

Para modelar esta previsão com KNIME, seguiu-se as seguintes etapas, também demonstradas na Figura 1:

- Leitura de dados;
- Limpeza de dados;
- Categorização de dados;
- Filtragem;
- Técnicas de *encoding*;
- Tratamento de valores em falta;
- Tratamento de *outliers*;
- Partitioning de dados;
- Modelo Supervisionado de Regressão linear;
- Modelo Supervisionado de Árvores de Decisão;
- Modelo Não Supervisionado de *K-Means Clustering*;
- Visualização dos dados de cada modelo.

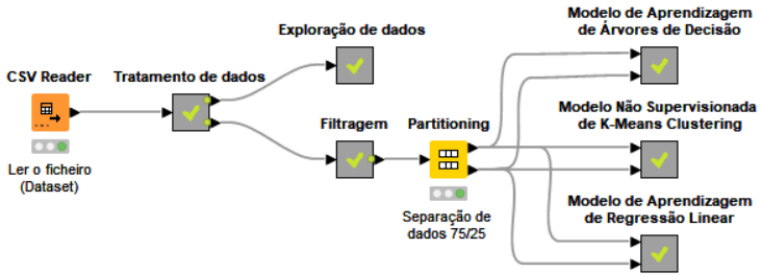


Figura 1: Previsão do preço de carros usando três modelos.

4.1 Carregar o *dataset* descarregado

O nó **CSV Reader** é usado para leitura dos dados de entrada. Como o ficheiro descarregado do Kaggle é .csv, no Knime coloca-se um **CSV Reader** para a importação e leitura do ficheiro, como está ilustrado na Figura 2.

Figura 2: Carregamento do *dataset* "Australian Vehicle Prices" no nó **CSV Reader**.

4.2 Tratamento dos Dados

A *Feature Engineering* envolve a extração e transformação de variáveis de dados brutos, como listas de preços, descrições de produtos e volumes de vendas, para que se possa usar recursos para treino e previsão.

A extração e transformação pode ser considerada a transformação de *strings* para números, separar valores da mesma coluna e remover colunas. No metanodo "Tratamento dos Dados", Figura 3 e Figura 4, contém as seguintes instruções:

- **Cell Splitter -** Divisão do atributo *Engine* em duas colunas;
- **Collumn Filter -** Remoção de uma coluna repetida;

- **String Manipulation -** Remoção das unidades;
- **String To Number -** Transformação de *strings* para números;
- **Category To Number -** Transformação de valores categóricos para numérico;
- **Numeric Outliers -** Detecta e trata os *outliers* de cada uma das colunas selecionadas individualmente por meio do intervalo interquartil;
- **Missing Value -** Ajuda a lidar com valores ausentes encontrados nas células da tabela de entrada;
- **Auto-binner -** Permite agrupar dados numéricos em intervalos - chamados *bins*.

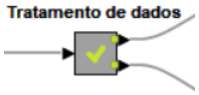


Figura 3: Metadono "Tratamento dos Dados".

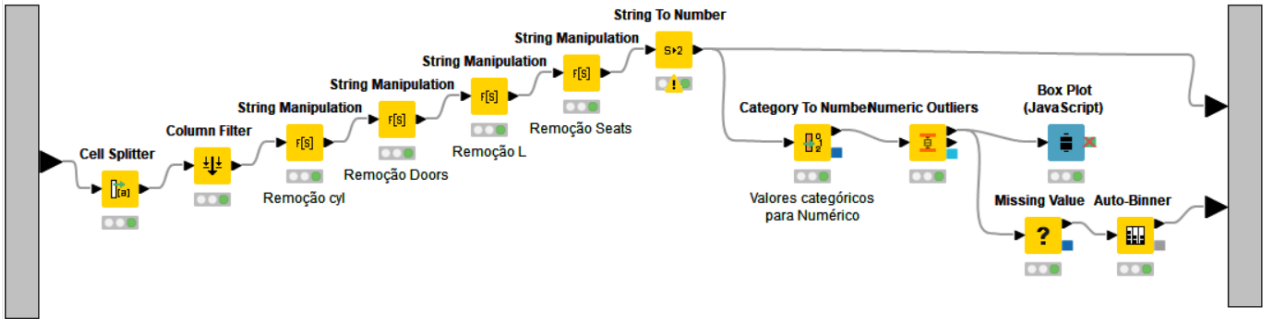


Figura 4: Dentro do metadono da "Tratamento dos Dados".

4.2.1 Cell Splitter

A *Engine* tem dois tipos de capacidade do motor, a cilindrada (*CylindersinEngine* e os litros), Figura 5. Para a remoção de informação repetida terá de se fazer primeiramente uma divisão de uma coluna em duas colunas utilizando um **Cell Splitter**, Figura 6. Posteriormente a remoção da coluna que está duplicada. A configuração do **Cell Splitter**, Figura 7, corresponderá ao que está no *dataset*. O delimatador entre os dois tipos de valores que se encontra no atributo *Engine* são delimitados com uma vírgula. Dado que não se quer que haja informação repetida a coluna inicial, *Engine*, Figura 5 será removida, ficando as duas novas colunas.

S Engine
4 cyl, 2.2 L
4 cyl, 1.5 L
4 cyl, 2 L
8 cyl, 5.5 L
4 cyl, 1.3 L
-
4 cyl, 2 L
4 cyl, 1.6 L
4 cyl, 2 L
4 cyl, 1.3 L
-
4 cyl, 2.7 L
-

Figura 5: Coluna do atributo *Engine*.

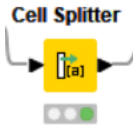


Figura 6: Nó **Cell Splitter**.

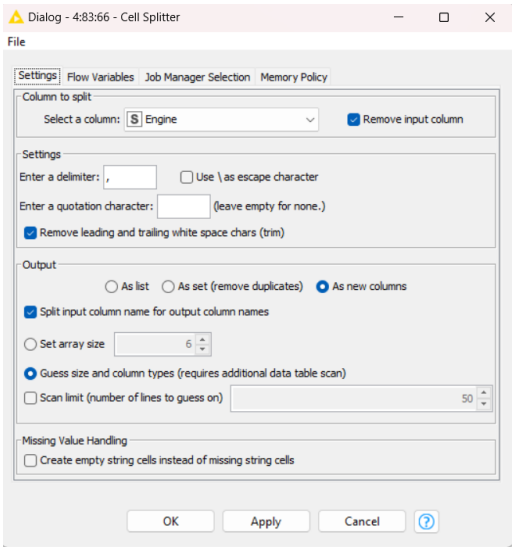


Figura 7: Janela de configuração do **Cell Splitter**.

4.2.2 Collumn Filter

Após a separação dos dois tipos de valores, cria-se um novo problema: colunas duplicadas. Os valores da cilindrada do *Engine* é igual aos valores da coluna *CylindersinEngine*. Como os valores da cilindrada do *Engine* estão agora numa coluna independente, esta será removida através de um **Column Filter**, Figura 8.

Pode se observar na Figura 9 que a coluna que é duplicada do atributo *CylindersinEngine* e anteriormente criada será removida.

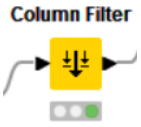


Figura 8: Nó **Column Filter**.

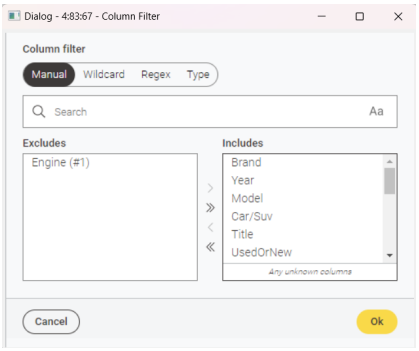
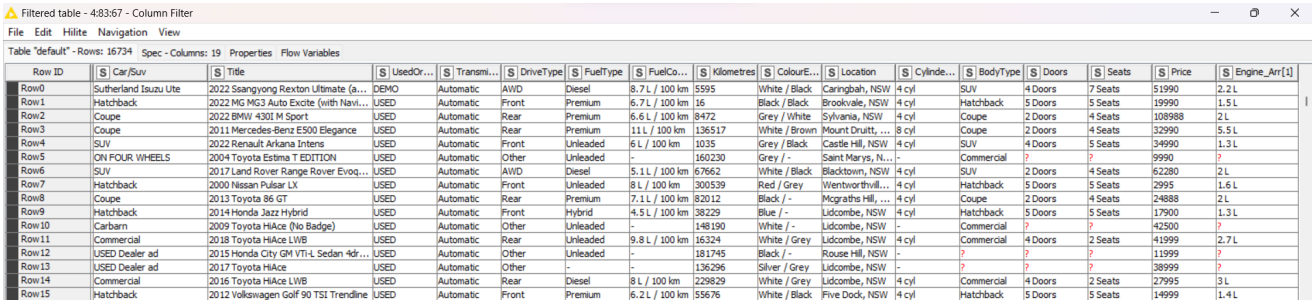


Figura 9: Janela de configuração do **Column Filter**.

A visualização da saída do nó **Column Filter** com a tabela lida na Figura 10.



Row ID	Car/Suv	Title	UsedOr...	Transmi...	DriveType	FuelType	FuelCo...	Kilometres	ColourE...	Location	Cylinde...	BodyType	Doors	Seats	Price	Engine_Arr[1]
Row0	Sutherland Isuzu Ute	2022 Ssangyong Rexton Ultimate (a...	DEMO	Automatic	AWD	Diesel	8.7 L / 100 km	5595	White / Black	Caringbah, NSW	4 cyl	SUV	4 Doors	7 Seats	51990	2.2 L
Row1	Hatchback	2022 MG MG3 Auto Excite (with Navi...	USED	Automatic	Front	Premium	6.7 L / 100 km	16	Black / Black	Brookvale, NSW	4 cyl	Hatchback	5 Doors	5 Seats	19990	1.5 L
Row2	Coupe	2022 BMW 430I M Sport	USED	Automatic	Rear	Premium	6.6 L / 100 km	8472	Grey / White	Sylvania, NSW	4 cyl	Coupe	2 Doors	4 Seats	108988	2 L
Row3	Coupe	2011 Mercedes-Benz E500 Elegance	USED	Automatic	Rear	Premium	11 L / 100 km	136517	White / Brown	Mount Druitt, ...	8 cyl	Coupe	2 Doors	4 Seats	32990	5.5 L
Row4	SUV	2022 Renault Arkana Intens	USED	Automatic	Front	Unleaded	6 L / 100 km	1035	Grey / Black	Castle Hill, NSW	4 cyl	SUV	4 Doors	5 Seats	34990	1.3 L
Row5	ON FOUR WHEELS	2004 Toyota Estima T EDITION	USED	Automatic	Other	Unleaded	-	160230	Grey / -	Saint Marys, N...	-	Commercial	?	?	9990	?
Row6	SUV	2017 Land Rover Range Rover Evoq...	USED	Automatic	AWD	Diesel	5.1 L / 100 km	67662	White / Black	Blacktown, NSW	4 cyl	SUV	2 Doors	4 Seats	62280	2 L
Row7	Hatchback	2000 Nissan Pulsar LX	USED	Automatic	Front	Unleaded	8 L / 100 km	300539	Red / Grey	Wentworthvill...	4 cyl	Hatchback	5 Doors	5 Seats	2995	1.6 L
Row8	Coupe	2013 Toyota 86 GT	USED	Automatic	Rear	Premium	7.1 L / 100 km	82012	Black / -	Mcgraths Hill, ...	4 cyl	Coupe	2 Doors	4 Seats	24888	2 L
Row9	Hatchback	2014 Honda Jazz Hybrid	USED	Automatic	Front	Hybrid	4.5 L / 100 km	38229	Blue / -	Lidcombe, NSW	4 cyl	Hatchback	5 Doors	5 Seats	17900	1.3 L
Row10	Carbarn	2009 Toyota HiAce (No Badge)	USED	Automatic	Other	Unleaded	-	148190	White / -	Lidcombe, NSW	-	Commercial	?	?	42500	?
Row11	Commercial	2018 Toyota HiAce LWB	USED	Automatic	Rear	Unleaded	9.8 L / 100 km	16324	White / Grey	Lidcombe, NSW	4 cyl	Commercial	4 Doors	2 Seats	41999	2.7 L
Row12	USED Dealer ad	2015 Honda City GM VTi-L Sedan 4dr...	USED	Automatic	Other	Unleaded	-	181745	Black / -	Rouse Hill, NSW	-	?	?	?	11999	?
Row13	USED Dealer ad	2017 Toyota HiAce	USED	Automatic	Other	-	-	136296	Silver / Grey	Lidcombe, NSW	-	?	?	?	38999	?
Row14	Commercial	2016 Toyota HiAce LWB	USED	Automatic	Rear	Diesel	8 L / 100 km	229829	White / Grey	Lidcombe, NSW	4 cyl	Commercial	4 Doors	2 Seats	27995	3 L
Row15	Hatchback	2012 Volkswagen Golf 90 TSI Trendline	USED	Automatic	Front	Premium	6.2 L / 100 km	55676	White / Black	Five Dock, NSW	4 cyl	Hatchback	5 Doors	5 Seats	14999	1.4 L

Figura 10: Tabela de saída do nó **Column Filter**.

4.2.3 String Manipulation

Este passo consiste em quatro nós **String Manipulation**, Figura 11, cada um dos quais é usado para uma tarefa específica. Observando a *filtered table* do nó **Column Filter**, Figura 10, observa-se que a coluna "Cylinder" possui as unidades "cyl". A coluna "Doors" possui unidades de "Doors", a coluna "Seats" possui unidades de "Seats" e a coluna "Engine" possui unidades de "L". As unidades precisam ser removidos para que os dados sejam considerados um número em vez de uma *string*. Precisa-se de quatro nós **String Manipulation** para realizar isto, em vez de um, pois a saída do nó **String Manipulation**, substitui a coluna ou anexa uma nova coluna. O nó não tem a capacidade de substituir caracteres diferentes de colunas diversas de uma só vez e, em seguida, substituí-los cada um na coluna correspondente (ou anexar cada um deles como uma nova coluna). Há um nó **String Manipulation (Multi-Column)** disponível no KNIME. No entanto, isso também não é adequado às necessidades porque esse nó permite a manipulação de múltiplas colunas - mas para a mesma *string*/caracteres. Neste caso, precisa-se de manipular strings diferentes para várias colunas. Portanto, usar-se-á quatro nós **String Manipulation**.
Para configurar cada atributo mencionado anteriormente, utiliza-se a seguinte fórmula: `replace($Atributo$, "Unidade","")`, conforme ilustrado na Figura 12. Além disso, caso seja necessário remover as unidades dessa coluna, é preciso executar a operação "Replace Column" para substituir a coluna com as unidades e marcar a opção "Insert Missing As Null".

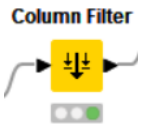


Figura 11: Nó **String Manipulation**.

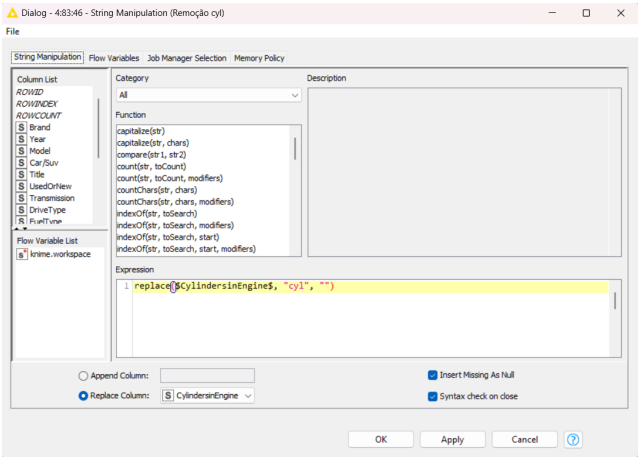


Figura 12: Janela de configuração de um "String Manipulation".

4.2.4 String To Number

Quando o ficheiro .csv é lido pela primeira vez os atributos são considerados todos *string*. O nó **String To Number**, Figura 13, é utilizado para que as colunas que só tenham números sejam considerados uma variável numérica e não *string*. A configuração do nó **String To Number** está representado na Figura 14.

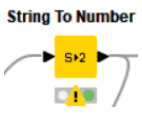


Figura 13: Nó **String To Number**.

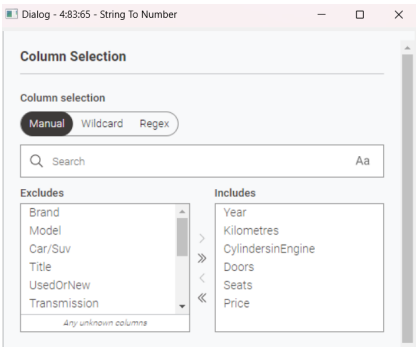


Figura 14: Janela de configuração do "String To Number".

4.2.5 Category To Number

Os dados que se lêem no CSV têm todas as variáveis categóricas diferentes. Na janela de configuração do nó **Category To Number**, Figura 15 foi preciso alterar o `Max. Categories` de 100 (valor default) para 16800, esta mudança teve de ser feita dado que existem 16733 linhas de valores.

O nó **Category To Number** recolhe os dados categorizados e atribui números às diferentes categorias. Além disso, cria novas colunas com a categoria como título para que as linhas dessas colunas tenham o número correspondente associado a essa categoria, Figura 16. Em cada categoria o número correspondente a uma nova *string* corresponde ao apareciemento pela primeira vez dessa mesma *string*. Por exemplo, se tiver um carro com a **Brand** 5, o **Year** 7 e o **Car/SUV** 5. Então haverá três colunas com informações 5, 7 e 5 que diria ao modelo que os dados associados a estas três colunas referem-se a um carro de marca Toyota de 2009 e o modelo Carbarn, Figura 16.

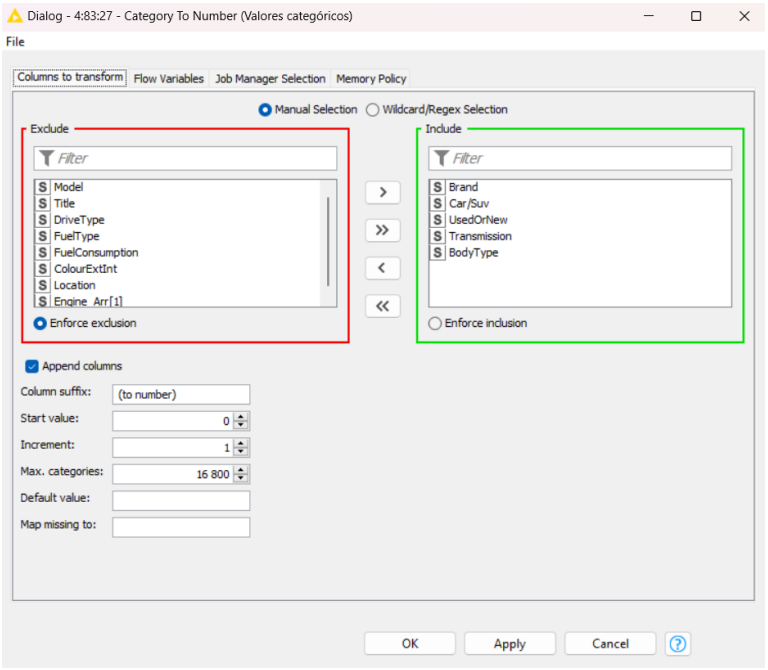


Figura 15: Janela de configuração do nó **Category To Number**.

Row ID	Brand	Brand (...	Year	Model	Car/Suv	Car/Su...	Title	Used...	UsedOr...	Tran...	Transmi...	Drive...	FuelType	FuelCo...	Kilometres	ColourE...	Location	Cylinde...	BodyType	BodyTy...
Row0	Ssangyong	0	2,022	Rexton	Sutherland ...	0	2022 Ss...	DEMO	0	Automatic	0	AWD	Diesel	8.7 L / 100 km	5,595	White / Black	Caringbah, NSW	4	SUV	0
Row1	MG	1	2,022	MG3	Hatchback	1	2022 M...	USED	1	Automatic	0	Front	Premium	6.7 L / 100 km	16	Black / Black	Brookvale, NSW	4	Hatchback	1
Row2	BMW	2	2,022	430I	Coupe	2	2022 B...	USED	1	Automatic	0	Rear	Premium	6.6 L / 100 km	8,472	Grey / White	Sylvania, NSW	4	Coupe	2
Row3	Mercedes-B...	3	2,011	E500	Coupe	2	2011 M...	USED	1	Automatic	0	Rear	Premium	11 L / 100 km	136,517	White / Brown	Mount Druitt, ...	8	Coupe	2
Row4	Renault	4	2,022	Arkana	SUV	3	2022 R...	USED	1	Automatic	0	Front	Unleaded	6 L / 100 km	1,035	Grey / Black	Castle Hill, NSW	4	SUV	0
Row5	Toyota	5	2,004	Estima	ON FOUR ...	4	2004 T...	USED	1	Automatic	0	Other	Unleaded	-	160,230	Grey / -	Saint Marys, N...	?	Commercial	3
Row6	Land	6	2,017	Rover	SUV	3	2017 La...	USED	1	Automatic	0	AWD	Diesel	5.1 L / 100 km	67,662	White / Black	Blacktown, NSW	4	SUV	0
Row7	Nissan	7	2,000	Pulsar	Hatchback	1	2000 Ni...	USED	1	Automatic	0	Front	Unleaded	8 L / 100 km	300,539	Red / Grey	Wentworthvill...	4	Hatchback	1
Row8	Toyota	5	2,013	86	Coupe	2	2013 T...	USED	1	Automatic	0	Rear	Premium	7.1 L / 100 km	82,012	Black / -	Mcgraths Hill, ...	4	Coupe	2
Row9	Honda	8	2,014	Jazz	Hatchback	1	2014 H...	USED	1	Automatic	0	Front	Hybrid	4.5 L / 100 km	38,229	Blue / -	Lidcombe, NSW	4	Hatchback	1
Row10	Toyota	5	2,009	HiAce	Carbarn	5	2009 T...	USED	1	Automatic	0	Other	Unleaded	-	148,190	White / -	Lidcombe, NSW	?	Commercial	3
Row11	Toyota	5	2,018	HiAce	Commercial	6	2018 T...	USED	1	Automatic	0	Rear	Unleaded	9.8 L / 100 km	16,324	White / Grey	Lidcombe, NSW	4	Commercial	3

Figura 16: Saída do nó **Category To Number**.

4.2.6 Numeric Outliers

O nó **Numeric Outliers** é aplicado para identificar valores fora do intervalo e considerados raros, Figura 17 e Figura 18.

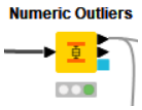


Figura 17: Nó **Numeric Outliers**.

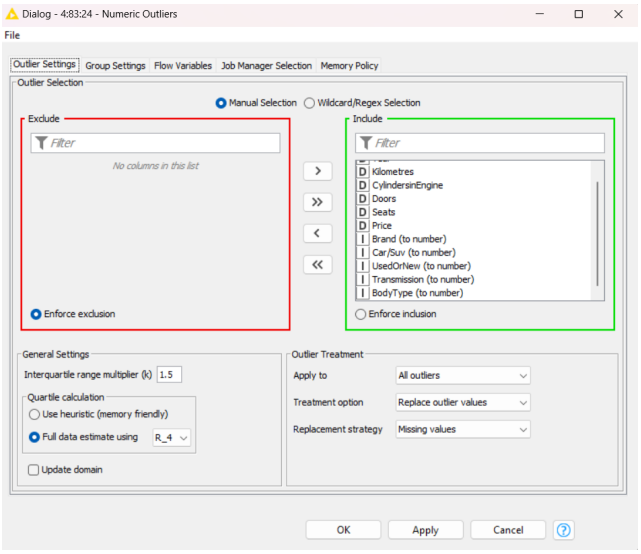


Figura 18: Janela de configuração de “Numeric Outliers”.

A visualização do Box Plot (Figura 19) gerada por este nó facilita a detecção destes *outliers*, nó **Box Plot (JavaScript)**, Figura 20.

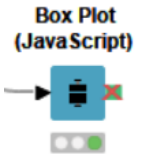


Figura 19: Nó **Box Plot (JavaScript)**.

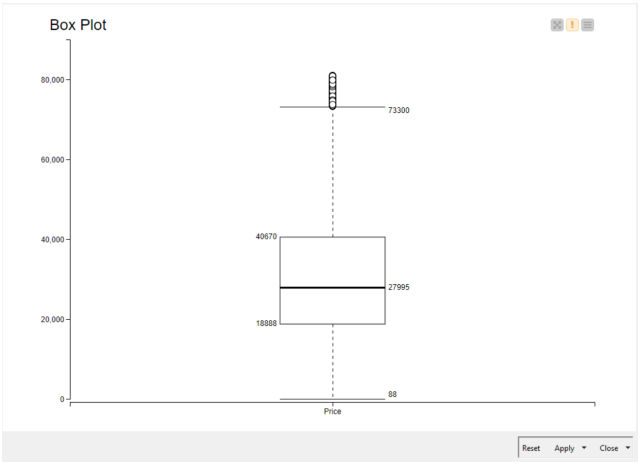


Figura 20: Janela de configuração de “Box Plot (JavaScript)”.

4.2.7 Missing Values

O tratamento de valores ausentes é realizado por meio do nó **Missing Values**, Figura 21, conforme apresentado na Figura 22.

Figura 21: Nó **Missing Values**.

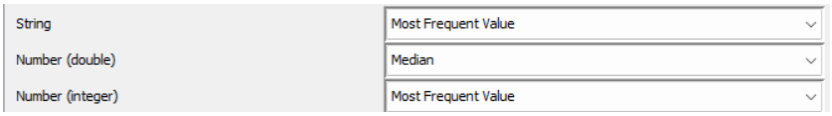


Figura 22: Tratamentos dos “Missing Values”.

Para substituir os valores desconhecidos, podem ser adotadas as seguintes estratégias:

- **Variáveis do Tipo String -** Valor Mais Frequente:
  Substituir valores ausentes em colunas do tipo *string* pelo valor mais frequente é uma escolha que preserva a

tendência central dos dados. A proximidade com os valores do Box Plot garante uma imputação alinhada com a distribuição predominante.

- **Variáveis do Tipo Number (double) -** Mediana:
  A mediana é usada para preencher valores numéricos do tipo *double*, sendo uma opção em situações com possíveis *outliers*. Menos sensível a valores extremos do que a média, a mediana mantém a centralidade da distribuição.

- **Variáveis do Tipo Number (Integer) -** Valor Mais Frequente.
  Para variáveis numéricas do tipo inteiro, o valor mais frequente preserva a moda da distribuição. Esta abordagem é eficaz em variáveis discretas, mantendo a consistência e representatividade da tendência central.

Estas escolhas visam preservar a integridade e a representatividade dos dados ausentes durante a análise.

4.2.8 Auto-binner

O nó **Auto-binner**, Figura 23, permite agrupar dados numéricos em intervalos - chamados *bins*. Existem duas opções de nomenclatura para os compartimentos e dois métodos que definem o número e o intervalo de valores que caem mum compartimento. A especificação da configuração do "Auto-binner" está representada na Figura 24.

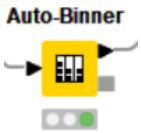


Figura 23: Nó **Auto-binner**.

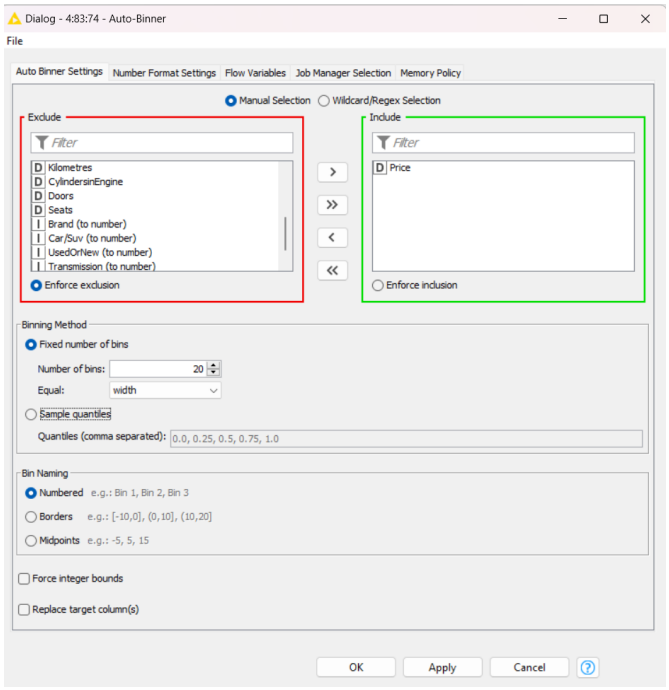


Figura 24: Configuração do "Auto-binner".

4.3 Exploração dos Dados

A Figura 25 e a Figura 26 representam o metadono "Exploração de dados". A Figura 25 representa o exterior e a Figura 26 o seu interior.

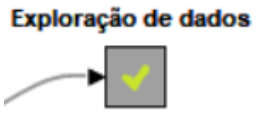


Figura 25: Metadono "Exploração de dados".

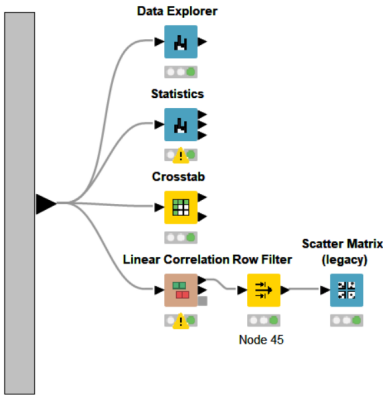


Figura 26: Dentro do metadono da "Exploração de dados".

A Figura 27 apresenta a tabela numérica gerada como resultado da execução do nó **Data Explorer**. Esta Figura 27 contém informações essenciais para a compreensão do relacionamento e qualidade dos dados explorados, como o máximo, o minimo e a média.

Numeric | Nominal | Data Preview

Search:

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness	Kurtosis	Overall Sum	No. zeros	No. missings	No. NaN	No. +∞
Year	☐	1940	2023	2016.229	5.248	27.538	-1.371	5.746	33737564	0	1	0	0
Kilometres	☐	1	533849	100096.110	78213.115	6117291293.727	0.968	1.029	1615751403	0	592	0	0
CylindersinEngine	☐	2	12	4.455	1.061	1.125	2.240	5.688	66613	0	1782	0	0
Doors	☐	2	5	4.005	0.693	0.480	-1.350	2.997	60312	0	1675	0	0
Seats	☐	2	22	5.101	1.119	1.253	0.874	12.536	76669	0	1705	0	0
Price	☐	88	1500000	37303.335	37177.867	1382193794.309	8.662	190.413	622256925	0	53	0	0

Showing 1 to 6 of 6 entries

Figura 27: Tabela Numérica do Nó **Data Explorer**.

A coluna "Year" representa o ano de fabrico dos veículos, com um intervalo abrangente entre 1940 e 2023. A média de 2016,229 sugere uma concentração de carros fabricados recentemente, enquanto a assimetria negativa indica uma distribuição levemente inclinada para carros mais antigos.

A coluna "Kilometres" representa a distância percorrida pelos veículos em quilometros. A média de 100096,110 indica uma distribuição provavelmente centrada em torno deste valor, enquanto a assimetria positiva sugere uma inclinação para a direita, indicando a presença de alguns veículos com uma quilometragem muito alta.

A coluna "CylindersinEngine" representa o número de cilindros no motor dos veículos. A média de 4,455 indica uma distribuição central em torno deste valor, enquanto a assimetria positiva sugere uma inclinação para veículos com mais cilindros.

A coluna "Doors" representa o número de portas nos veículos. A média de 4,005 indica que a maioria dos veículos possui aproximadamente quatro portas, enquanto a assimetria negativa sugere uma inclinação para veículos com menos portas.

A coluna "Seats" representa o número de assentos nos veículos. A média de 5.101 indica uma distribuição centrada em torno desse valor, enquanto a assimetria positiva sugere uma inclinação para veículos com mais assentos.

A coluna "Price" representa o preço dos veículos em dólares australianos. A média de 37303,335 indica uma distribuição centrada em torno deste valor, enquanto a assimetria positiva significativa sugere uma forte inclinação para veículos com preços mais elevados.

Estas estatísticas, também apresentadas na Figura 28, fornecem uma visão aprofundada da distribuição e variabilidade dos atributos-chave, permitindo análises mais avançadas e a construção de modelos preditivos.

Ao visualizar a Figura 29 que o nó **Scatter Matrix (Legacy)** criou pode-se observar que existe correlação forte positiva na diagonal da esquerda para a direita. Enquanto os outros valores não demonstram que existe uma boa correlação. A Figura 30 demonstra a correlação.

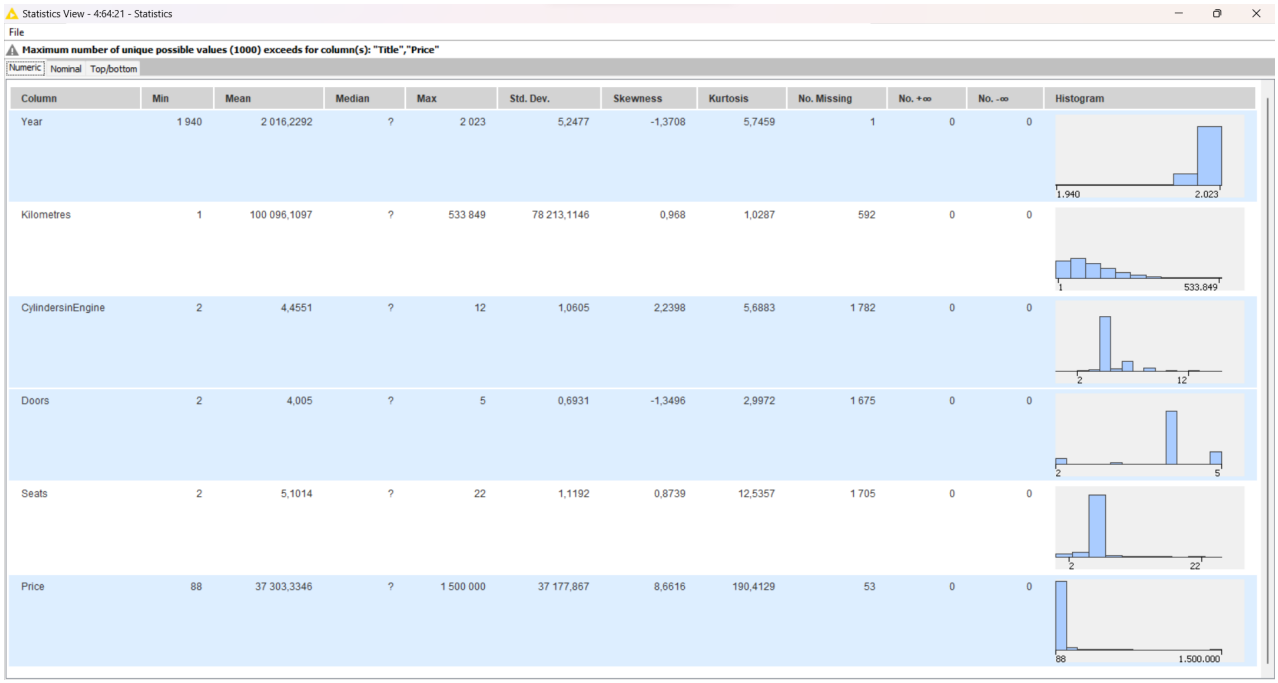


Figura 28: Estatísticas.

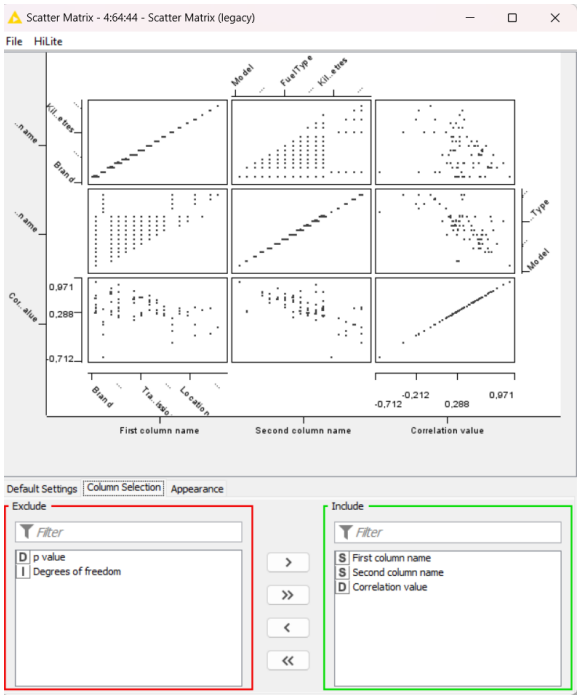


Figura 29: Scatter Matrix.

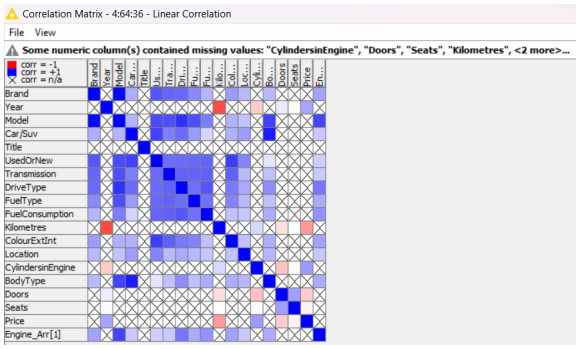


Figura 30: Correlação linear.

4.4 Filtragem

Na etapa de filtragem, realizada pelo metanodo “Filtragem”, Figura 31 e Figura 32, ocorre uma seleção criteriosa de colunas para remover redundâncias e concentrar o modelo em atributos relevantes. As colunas removidas incluem “Car/Suv”, “Title”, “Engine”, “DriveType”, “FuelType”, “FuelConsumption”, “ColourExtInt”, e “Location”. A Figura 33 exibe a tabela resultante após o filtro de colunas, evidenciando a presença de valores ausentes em algumas linhas.

Como se pode ver na Figura 33 a saída do nó **Column Filter** possui alguns valores ausentes de algumas linhas. As colunas “Transmisson”, “Kilometers”, “Cylinder”, “BodyType”, “Doors”, “Seats” e “Price” estão em formato de *string* com as unidades de quilometragem e cilindrada do motor.

Figura 31: Metadono “Filtragem”.

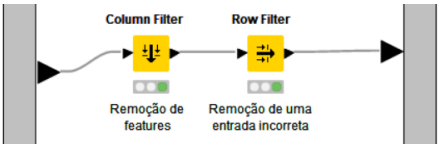


Figura 32: Dentro do metadono da “Filtragem”.

Filtered table - 4:56:26 - Column Filter (Remoção de)

File Edit Hilite Navigation View

Table "default" - Rows: 16734 Spec - Columns: 12 Properties Flow Variables

Row ID	Brand	Year	Car/Suv	UsedOr...	Transmi...	Kilometres	Cylinde...	BodyType	Doors	Seats	Price	Engine...
Row2391	?	?	nan	nan	?	?	?	?	?	?	?	?
Row51	Abarth	2,021	Hatchback	USED	Manual	13,200	4	Hatchback	3	4	36,990	1.4
Row6949	Abarth	2,022	02 **** ****	DEMO	-	17	?	Hatchback	?	?	44,824	?
Row8479	Abarth	2,021	Hatchback	USED	Automatic	14,901	4	Hatchback	3	4	37,990	1.4
Row751	Alfa	2,013	Awsum Autos - Car...	USED	Manual	115,050	?	Hatchback	?	?	10,995	?
Row945	Alfa	2,015	Hatchback	USED	Automatic	52,269	4	Hatchback	5	5	22,880	1.7
Row1046	Alfa	2,014	Hatchback	USED	Manual	136,311	4	Hatchback	5	5	12,995	1.4
Row1717	Alfa	2,015	Hatchback	USED	Automatic	98,446	4	Hatchback	5	5	24,995	1.7
Row2429	Alfa	2,015	Hatchback	USED	Automatic	132,907	4	Hatchback	5	5	15,800	1.7
Row6611	Alfa	2,013	Hatchback	USED	Manual	154,400	4	Hatchback	5	5	8,999	1.4
Row9729	Alfa	2,019	SUV	USED	Automatic	63,319	6	SUV	4	5	91,990	2.9
Row9795	Alfa	2,014	Hatchback	USED	Manual	92,500	4	Hatchback	3	5	10,692	1.4
Row14581	Alfa	2,017	Hatchback	USED	Automatic	52,820	4	Hatchback	5	5	28,999	1.7
Row15831	Alfa	2,006	Sedan	USED	Manual	90,000	4	Sedan	4	5	88	2
Row16119	Alfa	2,021	SUV	USED	Automatic	2,100	4	SUV	4	5	76,830	2
Row16238	Alfa	2,012	Wagon	USED	Automatic	199,379	5	Wagon	4	5	12,990	2.4
Row16264	Alfa	2,008	Sedan	USED	Automatic	104,776	4	Sedan	4	5	12,890	2.2
Row16475	Alfa	2,014	Hatchback	USED	Automatic	215,977	4	Hatchback	3	5	8,999	1.4

Figura 33: Tabela de saída do nó leitor **Column Filter**.

Antes de iniciar a modelação e limpeza de dados, a primeira etapa é fazer uma verificação de integridade dos dados. Observando a primeira linha, na Figura 34, depara-se que há uma linha onde não existe nenhum tipo de dados sobre essa entrada. Será removida esta linha porque não se quer que isto afete o resultado do modelo, Figura 34. Para fazer isso, utilizo o nó **Row Filter** e usando a opção “Excluir linhas por ID de linha”, insire-se o ID da linha (Row2391) e clica-se em Aplicar. A tabela de saída terá 16733 linhas em vez de 16734 linhas, como inicialmente.

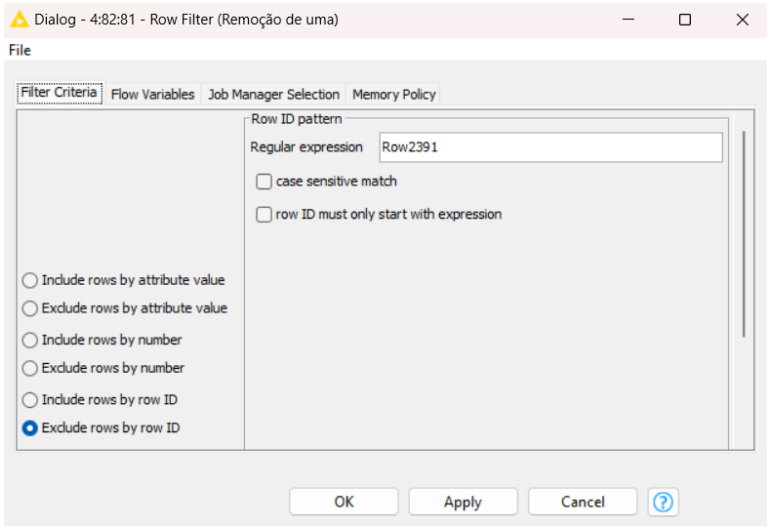


Figura 34: Remover de consideração a linha de dados possivelmente incorreta.

4.5 Modelação dos dados

Agora que os dados foram recolhidos, lidos e limpos, a próxima etapa é modelar os dados usando os Modelos de Supervisionada de Árvore de Decisão e de Regressão Linear e o modelo de aprendizagem não supervisionada *K-Means Clustering*.

A primeira etapa para fazer isso é particionar os dados usando o nó **Partitioning**. Os nós de **Partitioning** são usados para gerar um campo de partição que divide os dados em subconjuntos ou amostras separadas para os estágios

de treino, teste e validação da construção do modelo. A escolha da partição padrão foi de 75% – 25%, Figura 35 entre os dados de treino e os dados de teste. Marcou-se a caixa "Use Random seed". Isto garantirá que uma semente fixa seja desenhada para cada execução para obter resultados reproduzíveis. Este nó possui duas portas de saída – uma para os dados de treino e outra para os dados de teste.

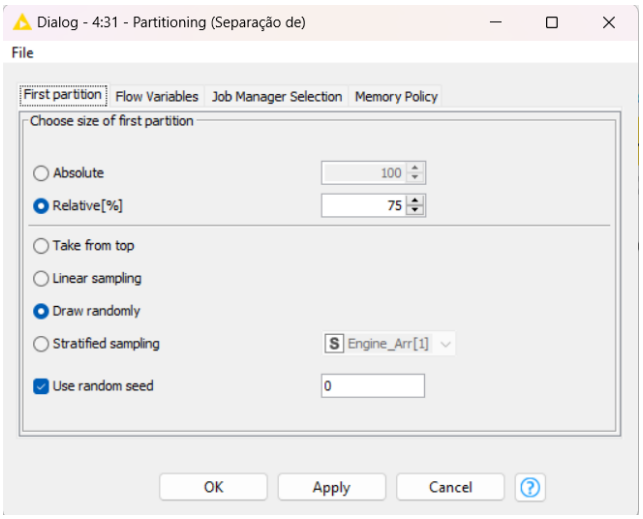


Figura 35: Partição dos dados 75% – 25%.

4.5.1 Modelo Supervisionado de Regressão Linear

A Regressão Linear está configurada num metanodo "Modelos de Aprendizagem de Regressão Linear", Figura 36 e Figura 37.

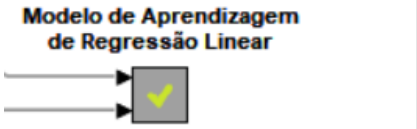


Figura 36: Metanodo **Modelos de Aprendizagem de Regressão Linear**.

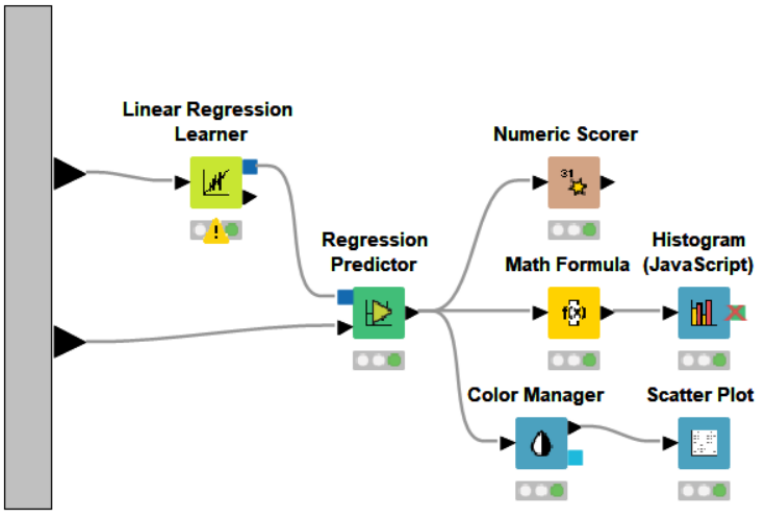


Figura 37: Dentro do metadono da "Modelos de Aprendizagem de Regressão Linear".

A saída de train do nó **Partitioning** está conectada ao nó **Linear Regression Learner**. Na janela de configuração, Figura 38, escolheu-se a variável alvo ("Price") e garantiu-se que todas as colunas necessárias estão incluídas.

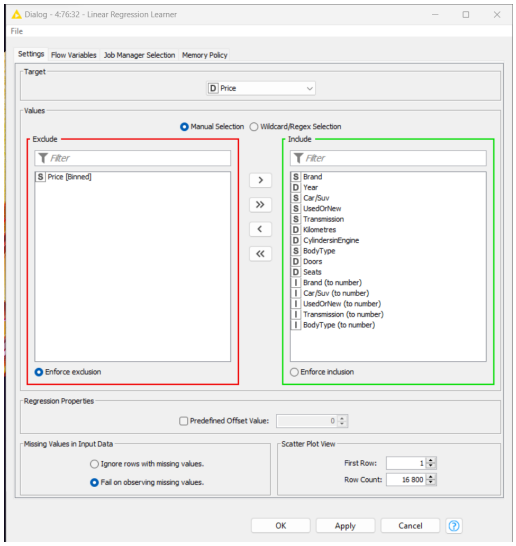


Figura 38: Configuração do nó **Linear Regression Learner**.

A saída do nó **Linear Regression Learner** consiste num modelo para o preditor e uma porta com coeficientes e estatísticas. Conecta-se o modelo ao nó **Regression Predictor** e conecta-se a saída de testes do nó **Partitioning** ao nó **Regression Predictor**. A porta de saída do nó **Regression Predictor** está conectada aos nós **Numeric Scorer** para poder avaliar o desempenho do modelo em relação aos dados de teste, o nó **Math Formula**,que está conectado ao nó **Histogram (JavaScript)**, para poder avaliar o erro residual e o nó **Color Manager**, que por sua vez está ligado ao nó **Scatter Plot**, para por avaliar a relacão "Price" - "Prediction (Price)". A configuração do **Numeric Scorer**, Figura 39 é a seguinte "Reference column" é a coluna com os dados reais e a "Predicted column" contém os dados previstos.

O *output* do modelo de regressão linear é mostrada na Figura 40. A estatística R^2 (medida de qualidade de ajuste) é 0,451, o que indica que há um ajuste de cerca de 45.1% para os dados no modelo de regressão. A Figura 40 mostra os coeficientes do nó **Linear Regression Learner**. Os coeficientes mostram como afetam o resultado final do modelo de regressão linear.

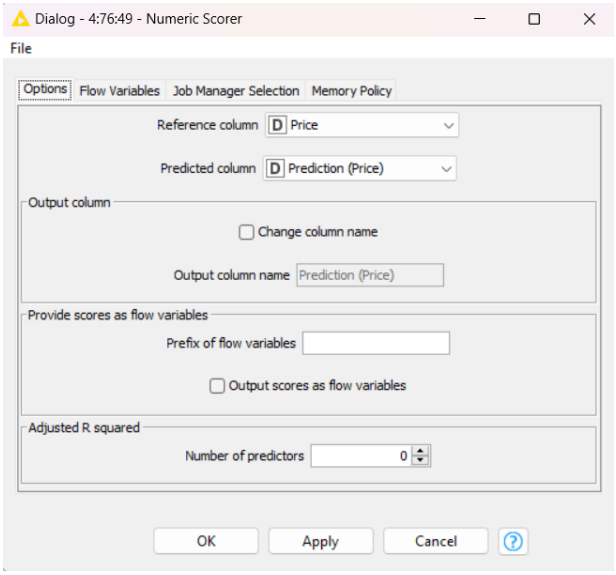


Figura 39: Configuração do nó **Numeric Scorer**.

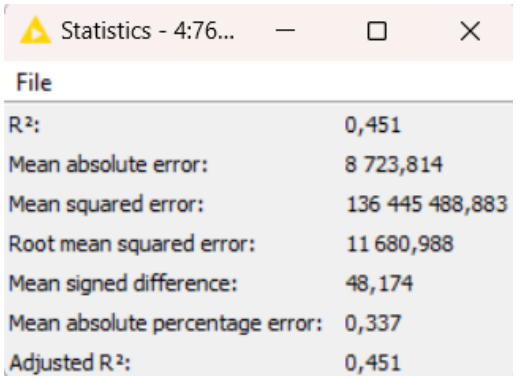


Figura 40: Output do nó **Numeric Scorer**.

Foram criados três outputs de histogramas: histograma erro residual, Figura 41que é a diferença entre o preço dado no *dataset* e o previsão do preço (Fórmula: `abs($Price$-$Prediction (Price)$)`), histograma do preço dado no *dataset*, Figura 42, e histograma do preço precisto, Figura 43.

Os histogramas do erro residual, Figura 41, e do preço, Figura 42, exibem assimetria para a direita, indicando uma tendência de subestimação dos preços. Por outro lado, o histograma do preço previsto, Figura 43, é simétrico, sugerindo uma distribuição mais equilibrada das estimativas de preço. Estas análises complementares fornecem uma visão mais holística do desempenho e comportamento do modelo de regressão linear.

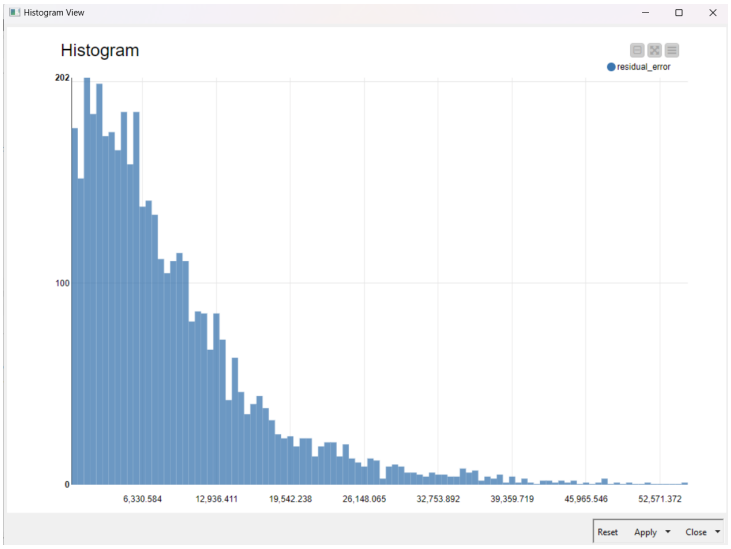


Figura 41: Histograma do Erro Residual.

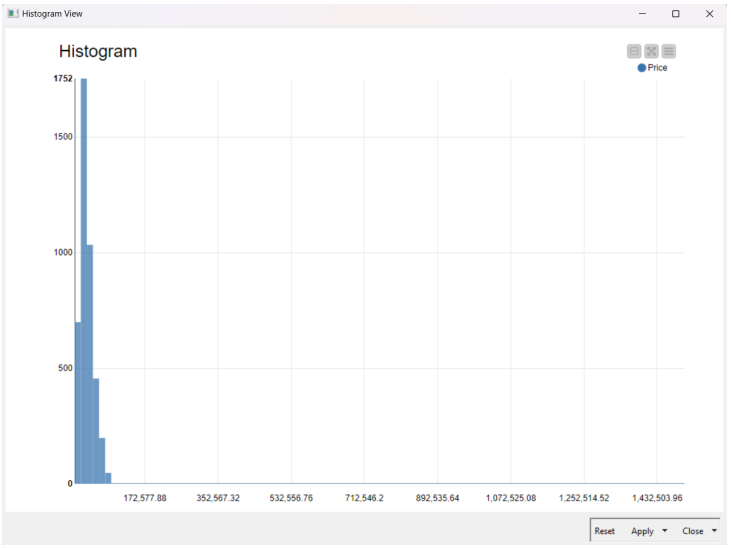


Figura 42: Histograma do Preço.

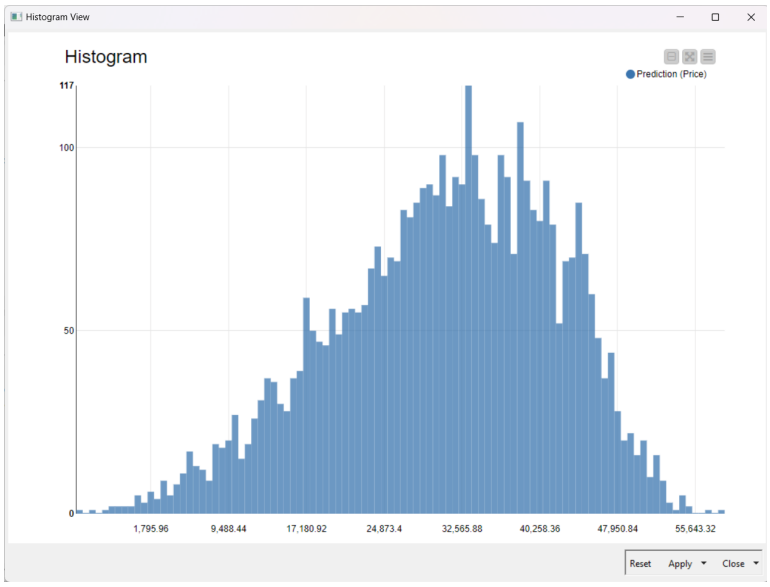


Figura 43: Histograma do Preço Previsto.

A Figura 44 representa um gráfico de dispersão (*Scatter Plot*) usa pontos para representar valores para duas variáveis numéricas diferentes. A posição de cada ponto nos eixos horizontal e vertical indica valores para um ponto de dados individual. Gráficos de dispersão são usados para observar relações entre variáveis. O eixo horizontal representa o *Predicted (Price)* e o o eixo vertical representa o eixo *Price*. Se criá-se uma reta nesta figura esta seria uma curva crescente virada para cima.

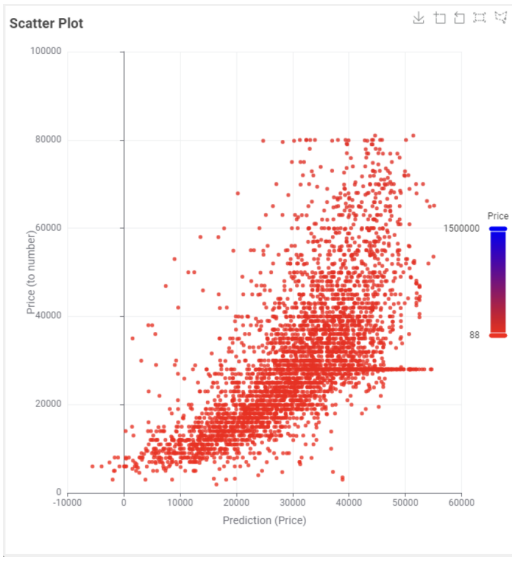


Figura 44: Gráfico de Dispersão.

4.5.2 Modelo Supervisionado de Árvores de Decisão

A Árvore de Decisão está configurada num metanodo "Modelos de Aprendizagem de Árvores de Decisão", Figura 45 e Figura 46.

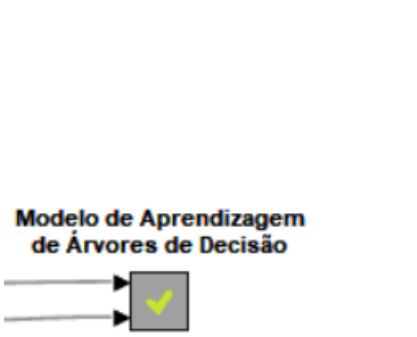


Figura 45: Metanodo **Modelos de Aprendizagem de Árvores de Decisão**.

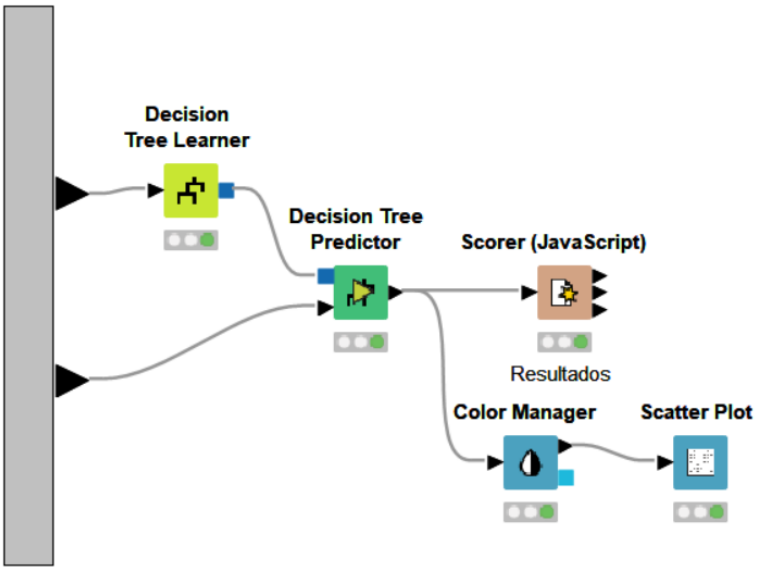


Figura 46: Dentro do metadono da "Modelos de Aprendizagem de Árvores de Decisão".

A saída de train do nó **Partitioning** está conectada ao nó **Decision Tree Learner**. Na janela de configuração, Figura 47, escolheu-se a *class column* ("Price [Binned]").

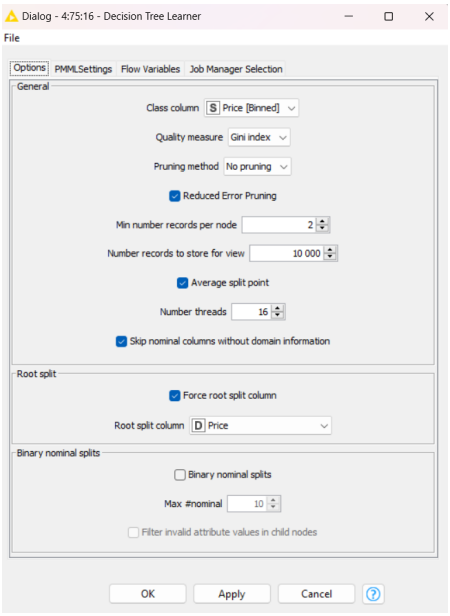


Figura 47: Configuração do nó **Decision Tree Learner**.

A saída do nó **Decision Tree Learner** consiste num modelo para o preditor e uma porta com coeficientes e estatísticas. Conecta-se o modelo ao nó **Decision Tree Predictor** e conecta-se a saída de testes do nó **Partitioning** ao nó **Decision Tree Predictor**. A porta de saída do nó **Decision Tree Predictor** está conectada aos nós **Scorer (JavaScript)** para poder avaliar o desempenho do modelo em relação aos dados de teste, o nó **Color Manager**, que por sua vez está ligado ao nó **Scatter Plot**, para por avaliar a relacão "Price (to Number)" - "Prediction (Price)". A configuração do **Numeric Scorer**, Figura 48 é a seguinte "Actual column" é a coluna com os dados reais (*Price [Binned]*) e a "Predicted column" contém os dados previstos, (*Prediction(Price [Binned])*).

O *output* do modelo de árvore de decisão é mostrada na Figura 40. O *Overall Error* indica que há um ajuste de cerca de 100% para os dados no modelo de árvore de decisão. A Figura 49 mostra os coeficientes do nó **Decision Tree Learner**. Os coeficientes mostram como afetam o resultado final do modelo de árvore de decisão.

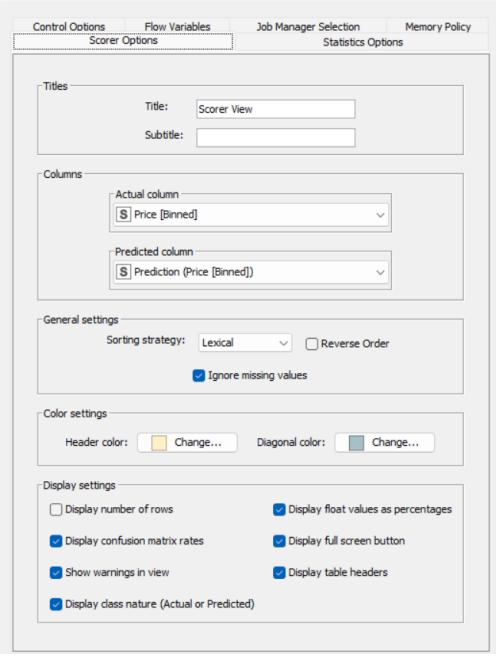


Figura 48: Configuração do nó **Scorer (JavaScript)**.

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (δ□□')	Correctly Classified	Incorrectly Classified
100.00%	0.00%	1.000	4184	0

Figura 49: Output do nó **Numeric Scorer**.

4.5.3 Modelo Não Supervisionado de *K-Means Clustering*

O *K-Means Clustering* está configurada num metanodo "Modelo Não Supervisionado de K-Means Clustering", Figura 50 e Figura 51.

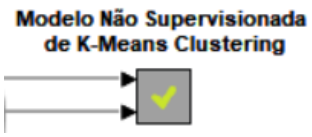


Figura 50: Metanodo **Modelo Não Supervisionado de K-Means Clustering**.

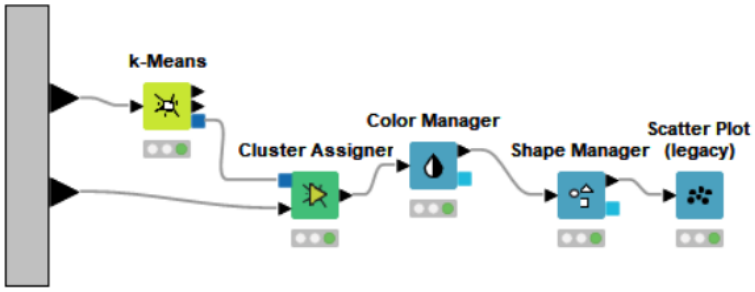


Figura 51: Dentro do metadono da "Modelo Não Supervisionado de K-Means Clustering".

A saída de train do nó **Partitioning** está conectada ao nó **K-Means**. Na janela de configuração, Figura 52, escolheu-se a 3 *clusters* e com *uma random seed* de 0 (zero).

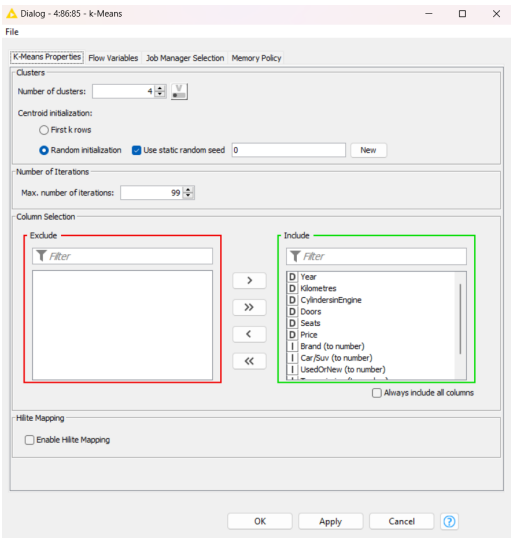


Figura 52: Configuração do nó **K-means**.

A saída do nó **K-Means** consiste num modelo para o nó **Cluster Assigner**. Conecta-se o modelo ao nó **Cluster Assigner** e conecta-se a saída de testes do nó **Partitioning** ao nó **Cluster Assigner**. A porta de saída do nó **Cluster Assigner** está conectada ao nó **Color Manager**, que por sua vez está ligado ao nó **Shape Manager** e por último conectado ao **Scatter Plot (Legacy)**, para por avaliar as diferentes relações entre as variáveis.

Os nós **Color Manager** e **Shape Manager**, Figura 53, Figura 54 e Figura 55, definem como os valores são representados no **Scatter Plot (Legacy)**, Figura 56 e Figura 57.

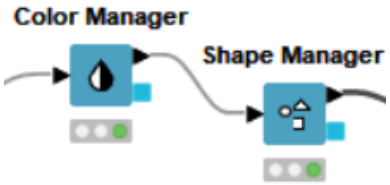


Figura 53: Nós **Color Manager** e **Shape Manager**.

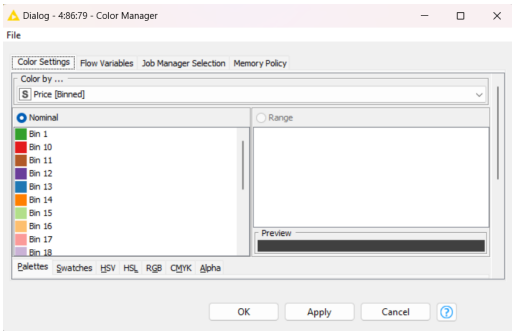


Figura 54: Configuração do nó **Color Manager**.

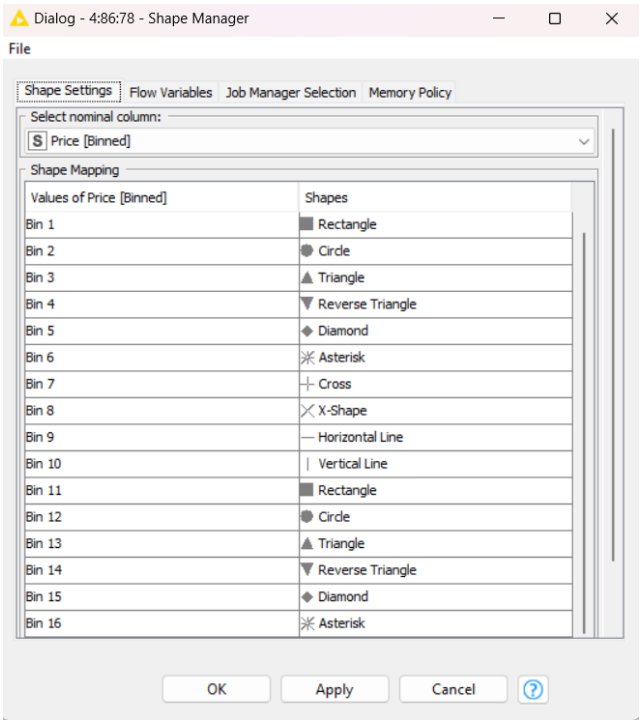


Figura 55: Configuração do nó **Shape Manager**.

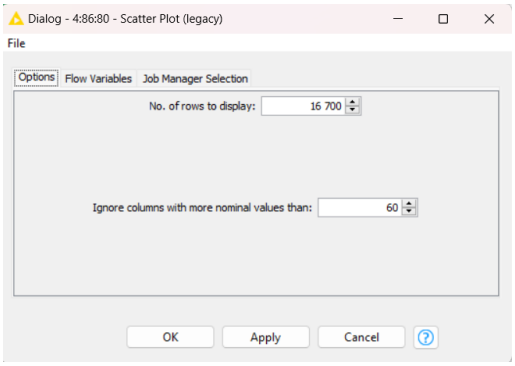


Figura 56: Configuração do nó **Scatter Plot (Legacy)**.

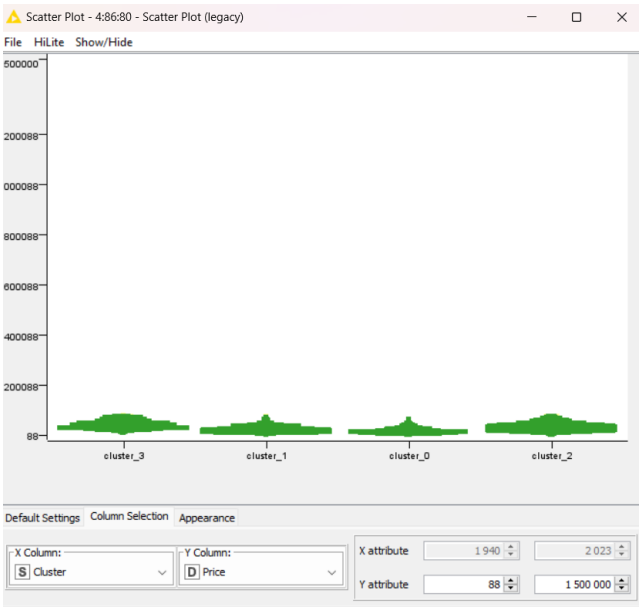


Figura 57: *Output* do nó **Scatter Plot (Legacy)**.

5 Conclusão

Em conclusão, este projeto teve como objetivo principal explorar e modelar o conjunto de dados "Australian Vehicle Prices" para o ano de 2023, fornecendo *insights* valiosos sobre o mercado automóvel na Austrália. A análise abrangeu desde a compreensão do cenário automóvel australiano até a aplicação de técnicas de limpeza, transformação e modelação de dados.

A abordagem metodológica adotada envolveu a utilização do KNIME, uma ferramenta versátil para ciência e análise de dados.

Os resultados da modelação incluíram a criação de um modelo de regressão linear para prever os preços dos carros com base em diversas características. A análise exploratória revelou *insights* sobre a distribuição e correlação entre as variáveis, enquanto os histogramas e gráficos de dispersão proporcionaram uma compreensão mais aprofundada do desempenho do modelo.

Em termos de aplicabilidade prática, o projeto oferece benefícios para consumidores, revendedores e analistas de mercado, fornecendo uma visão abrangente dos fatores que influenciam os preços dos veículos na Austrália. Além disso, destaca a importância de ferramentas como o KNIME e plataformas colaborativas como Kaggle na análise de dados do mundo real.

Em suma, a análise realizada contribui para uma compreensão mais profunda do mercado automóvel australiano, promovendo a tomada de decisões informadas e destacando o potencial das técnicas de *machine learning* na previsão de preços de carros e na análise de tendências de mercado.

Referências Bibliográficas

[1] Gonzalo Mariscal, Óscar Marbán e Covadonga Fernández. "A survey of data mining and knowledge discovery process models and methodologies". Em: *The Knowledge Engineering Review* 25.2 (2010), pp. 137–166. DOI: 10.1017/S0269888910000032.

[2] ZEESHAN-UL-HASSAN USMANI. *What is Kaggle, Why I Participate, What is the Impact?* https://www.kaggle.com/discussions/getting-started/44916. Acedido em 20 de dezembro de 2023. 2017.