Bootcamp: Analista de Dados

Desafio Prático

Módulo 5: Desafio Final

Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Bootcamp:

1. Realizar coleta e preparação de dados.
2. Realizar análise exploratória de dados.
3. Executar uma análise preditiva por meio da tarefa de regressão.
4. Construir um painel analítico.
5. Exercitar práticas utilizando a ferramenta Knime Analytics Platform.
6. Exercitar práticas utilizando a ferramenta Power Bi.

Enunciado

Uma empresa que trabalha com aluguel de veículos deseja estudar como seu negócio está funcionando e gostaria de entender como determinadas variáveis influenciam no preço do aluguel do veículo. Para confirmar essas suspeitas, foi reunido em um conjunto de dados um histórico contendo locações que foram exportadas aleatoriamente do banco de dados, ou seja, temos 121 observações. Os dados se encontram no arquivo "aluguel_veiculosOriginal.xlsx". A imagem abaixo apresenta um exemplo dos dados existentes no arquivo.

valor_total_locacao	qtde_portas	ar_condicionado	data_inicio_locacao	idade_locatario	genero	quilometragem	cotacao_dolar	Estado	cidade	qtde_diarias
368,38	2 portas	sem_ar_condicionado	01/06/2021	23	Masculino	957,44	4,41	Minas Gerais	Belo Horizonte	3
446,85	4 portas	com_ar_condicionado	12/05/2021	18	Feminino	829,53	5,63	Bahia	Feira de Santana	4
414,73	2 portas	com_ar_condicionado	14/04/2021	28	Feminino	923,30	8,81	Rio de Janeiro	Rio de Janeiro	2
434,29	4 portas	com_ar_condicionado	02/06/2021	21	Feminino	871,52	4,26	São Paulo	São Paulo	2

Foram selecionadas para este estudo as seguintes variáveis:

1. Valor total locação (variável contínua medida em reais) – Nos informa qual foi o valor total de uma determinada locação.

2. Quantidade portas (variável categórica com dois níveis) – Nos informa se o veículo alugado era de duas ou quatro portas.

3. Ar-Condicionado (variável categórica com dois níveis) – Nos informa se o veículo alugado tinha ou não tinha ar-condicionado.

4. Data (variável contínua expressa como data no formato DD/MM/YYYY) – Informa a data inicial daquela locação.

5. Idade do locatário (variável discreta medida em anos) – Nos informa qual a idade do indivíduo que realizou a locação.

6. Quilometragem (variável contínua medida em KM) – Nos informa quantos KM rodados o veículo tinha no ato da locação.

7. Dólar (variável contínua medida em dólares) – Nos informa qual a cotação do dólar no dia da locação.

8. Estado (variável categórica Nominal) - Nos informa em qual estado originou a locação.

9. Cidade (variável categórica Nominal) - Nos informa em qual cidade originou a locação.

10. Quantidade de dias locados (variável contínua medida em número de dias) - Nos informa quantos dias foi uma determinada locação.

Atividades

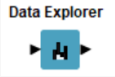
Os alunos deverão desempenhar as seguintes atividades:

Atividade 1:

Utilizando a ferramenta Knime Analytics Plataform, construa um workflow para realizar a coleta e análise exploratória de dados. Para executar esta atividade, seu workflow deve conter uma sequência de nós, conforme descrito abaixo.

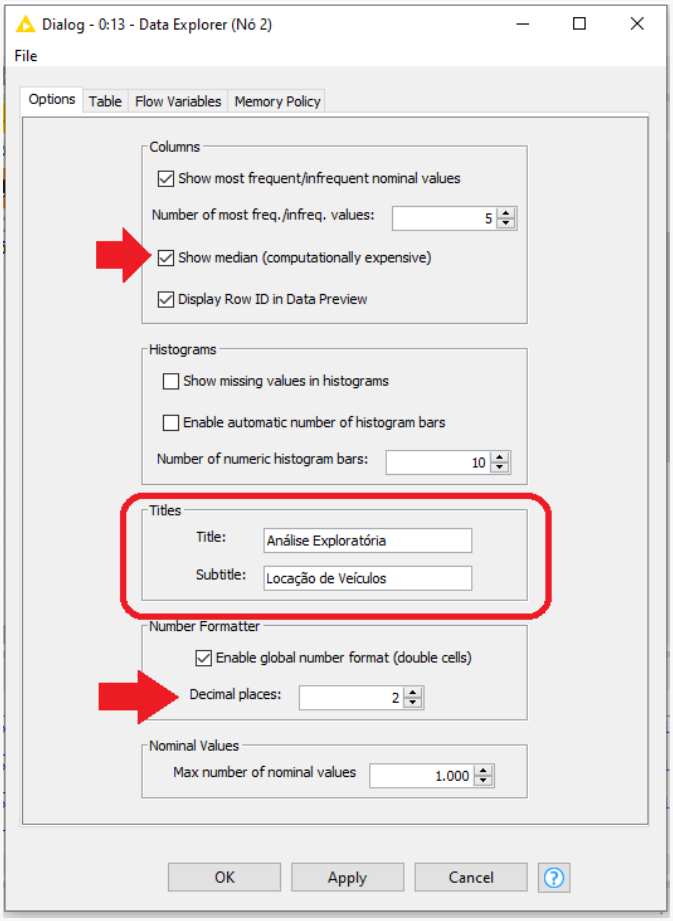
Passo 1. Inserir um nó para coleta/leitura do seu conjunto de dados. Abra o conjunto de dados que possui o histórico de locações, disponível no arquivo "aluguel_veiculosOriginal.xlsx".

Passo 2. Inserir um nó para executar a análise exploratória de dados nesse conjunto de dados. Sugere-se utilizar os seguintes nós:

, disponível na extensão: KNIME, Excel Support e 

, disponível na extensão: KNIME JavaScript Views (Labs).

Nas configurações do nó "Data Explorer", habilite a opção para calcular a mediana (*Show median*) e formate o número de casas decimais (*Decimal places*) para 2. Opcionalmente, atribua um título e subtítulo ao relatório que será gerado. A imagem a seguir demonstra um exemplo de configuração desse nó.



Após a execução do seu workflow de análise exploratória, analise o resultado para melhor conhecer seus dados. Note que na análise, a variável de data não é considerada. Para todas as variáveis (numéricas e nominais), observe se existe a ocorrência de valores ausentes (*No. missings*) e quantos são. Analise cada uma das variáveis numéricas buscando identificar os valores das Medidas de tendência central (Média aritmética, Mediana, Moda), Medidas de Posição (Máximo, Mínimo), Medidas de dispersão (Variância, Desvio Padrão, Amplitude). Lembre-se que a amplitude é determinada pela diferença entre o valor máximo e mínimo. Para as variáveis categóricas ou nominais, observe as ocorrências de cada um deles, quantas ocorrências distintas cada variável possui, como são escritas, tentando identificar inclusive se alguma ocorrência possui erro de ortografia.

Atividade 2

Após a análise exploratória, execute as atividades de preparação de dados, mais especificamente, tratamento de dados ausentes. Lembre-se de realizar a leitura de seu conjunto de dados novamente. Utilizando a ferramenta Knime Analytics Plataform, riar um workflow para fazer os seguintes tratamentos de dados:

Passo 1. Inserir um nó para coleta/leitura do seu conjunto de dados. Abra o conjunto de dados que possui o histórico de locações, disponível no arquivo "aluguel_veiculosOriginal.xlsx".

Por meio da análise exploratória de dados, observamos que existem valores ausentes em diversas variáveis. Vamos tratar cada caso de uma forma diferente.

Passo 2. Para a variável *idade_locatario,* vamos excluir as linhas que possuem valores ausentes nesta variável. Sugere-se utilizar o nó "Row Filter", disponível na extensão "KNIME Base nodes" instalada junto com a instalação do Knime

.

Passo 3. Para a variável ar_condicionado, vamos tratar as linhas que possuem valores ausentes, inserindo o valor "com_ar_condicionado" para as locações onde a variável "qtde_portas" for igual a "4 portas", e inserindo o valor "sem_ar_condicionado" para as locações onde a variável "qtde_portas" for igual a "2 portas". Sugere-se utilizar o seguinte nó

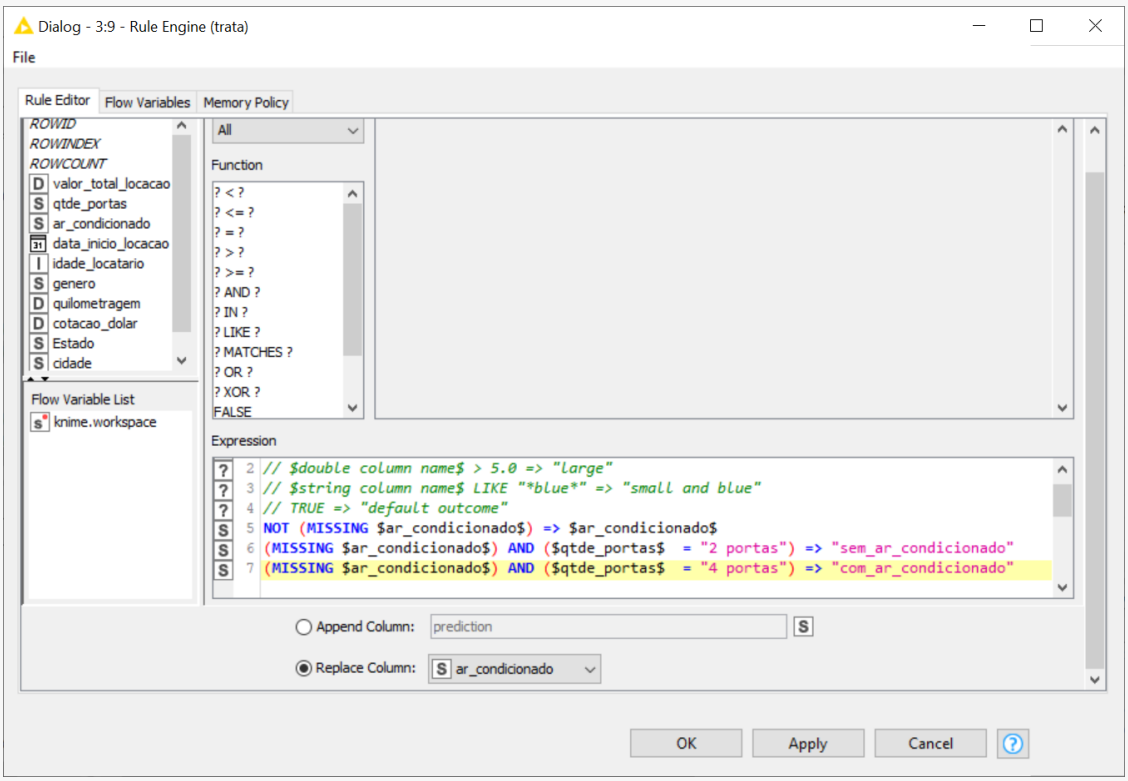
, disponível na extensão KNIME Javasnippet. Esse nó é usado para você criar regras baseada nas variáveis e valores de seus dados. Você precisa criar as seguintes regras:

- Se a variável ar_condicionado não possui valores ausentes, você mantem o valor que existe na variável.

- Se a variável ar_condicionado possui valores ausentes e a variável qtde_portas for igual a “2 portas”, altera o valor da variável para “sem_ar_condicionado”.

- Se a variável ar_condicionado possui valores ausentes e a variável qtde_portas for igual a “4 portas”, altera o valor da variável para “com_ar_condicionado”.

A imagem a seguir demonstra um exemplo de configurações desse nó.



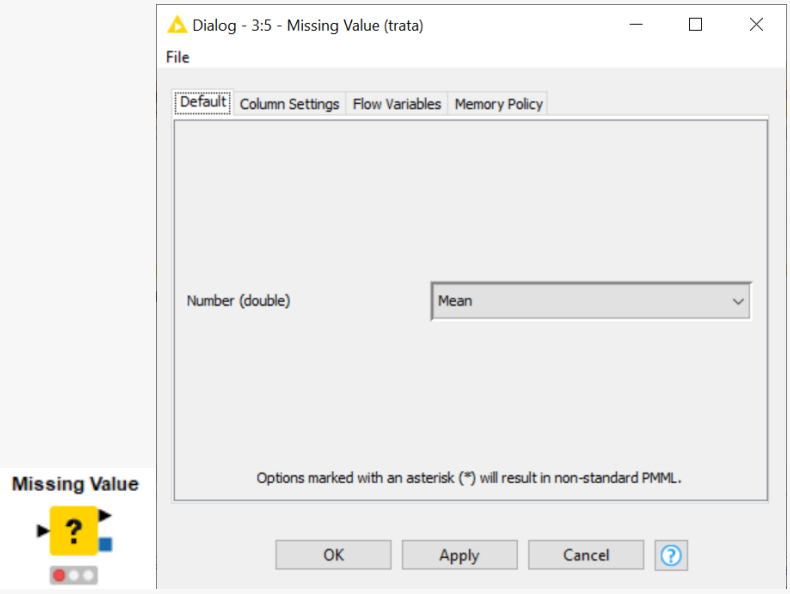
Passo 4. Para as variáveis numéricas cotacao_dolar e valor_total_locacao, vamos tratar as linhas que possuem valores ausentes, substituindo o valor ausente pelo valor da média da respectiva coluna. Sugere-se utilizar as seguintes etapas:

a. Separar (Split) o conjunto de dados em dois subconjuntos, onde um dos subconjuntos terá apenas as variáveis (colunas) cotação_dolar e valor_total_locacao, e o outro subconjunto terá as demais variáveis.

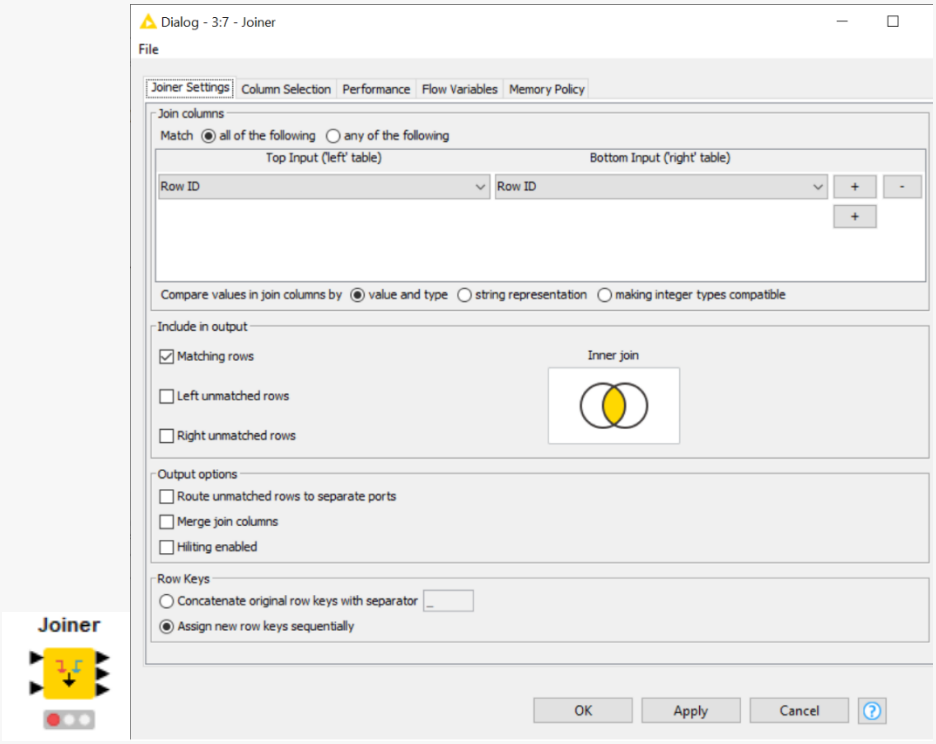
Sugere-se utilizar o nó

, disponível na extensão “KNIME Base nodes” instalada junto com a instalação do Knime.

b. Substituir o valor ausente das colunas (variáveis) cotacao_dolar e valor_total_locacao pelo valor da média respectiva coluna. Sugere-se utilizar o nó “Missing Value”, que da extensão “KNIME Base nodes” instalada junto com o Knime; configurá-lo conforme imagem abaixo:

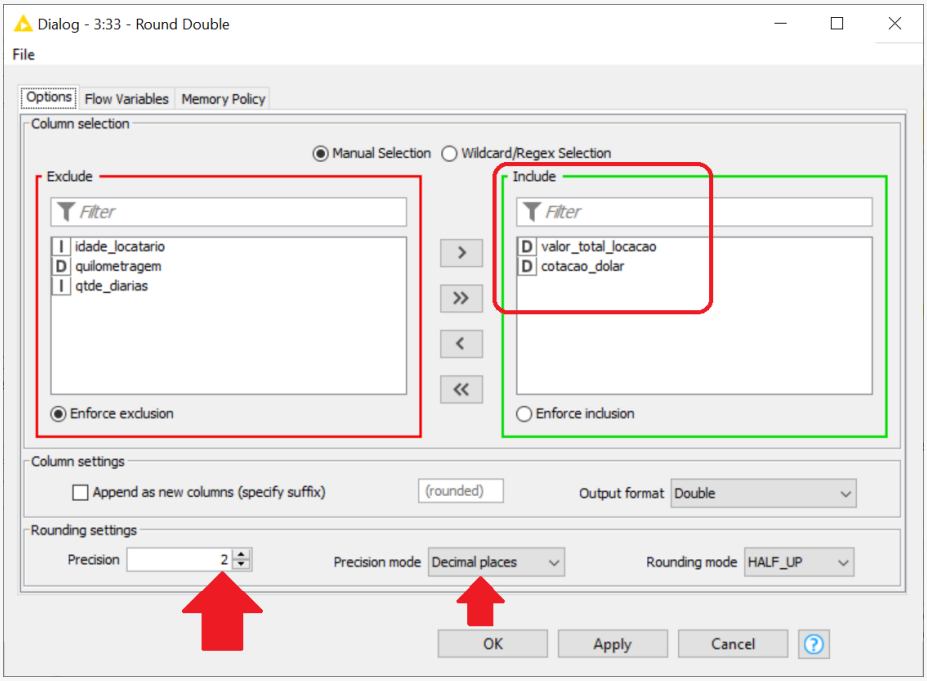


c. Fazer a junção dos dois subconjuntos de dados, o subconjunto que contém as colunas variáveis cotacao_dolar e valor_total_locacao e o subconjunto que possui as demais colunas. Sugere-se utilizar o seguinte nó “Joiner”, que da extensão “KNIME Base nodes” instalada junto com o Knime; configurá-lo conforme imagem abaixo. Note que a coluna usada na junção dos subconjuntos deve ser a coluna RowID.



d. Para garantir que os valores calculados possuam no máximo duas casas decimais, utilize o nó

da extensão “KNIME Base nodes”, e faça a configuração desse nó conforme imagem abaixo.



Passo 5. Para a variável estado, vamos tratar as linhas que possuem valores ausentes, considerando o valor da cidade e, consequentemente, substituindo o valor ausente pelo valor correspondente ao estado no qual a respectiva cidade pertence. Sugere-se utilizar as seguintes etapas:

a. Separar (Split) o conjunto de dados em dois subconjuntos, onde um dos subconjuntos terá apenas as variáveis (colunas) *estado* e *cidade*, e o outro subconjunto terá as demais variáveis. Sugere-se utilizar o nó "*Column Splitter*".

b. Substituir o valor ausente das colunas (variáveis) estado pelo valor correspondente ao estado no qual a respectiva cidade pertence. Para isso, você pode usar como modelo o tratamento realizado nos dados ausentes da variável "ar_condicionado", no Passo 3. Sugere-se criar a seguinte regra nó "*Rule Engine*":

- Se a variável estado não possui valores ausentes, você mantém o valor que existe na variável.

- Se a variável estado possui valores ausentes e a variável cidade a um dos seguintes valores ("Anápolis","Goiânia"), substitui o valor ausente da variável estado para "Goiás". Abaixo um exemplo de como escrever esta regra: `MISSING $Estado$ AND $cidade$ IN ("Anápolis","Goiânia") => "Goiás"`

- Se a variável estado possui valores ausentes e a variável cidade a um dos seguintes valores ("Belo Horizonte","Betim","Divinópolis","Uberlândia"), substitui o valor ausente da variável estado para "Minas Gerais".

- Se a variável estado possui valores ausentes e a variável cidade a um dos seguintes valores ("Feira de Santana","Salvador","Ilhéus"), substitui o valor ausente da variável estado para "Bahia".

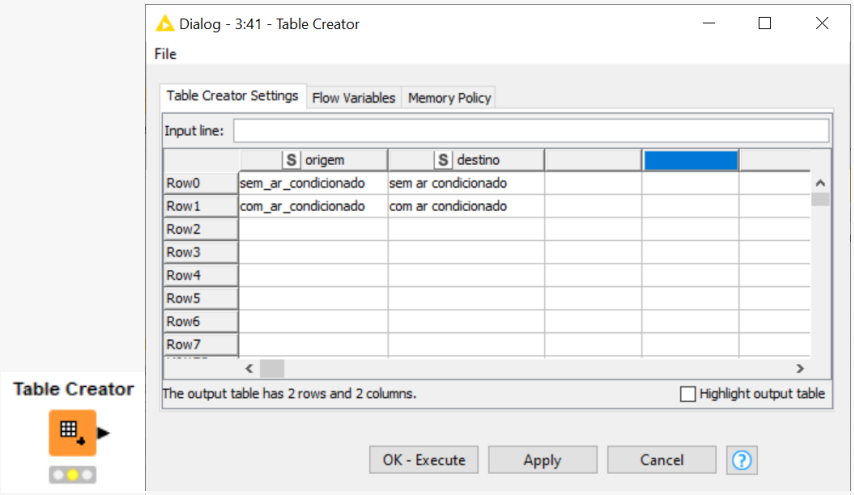
- Se a variável estado possui valores ausentes e a variável cidade a um dos seguintes valores (“São Paulo”,”Ribeirão Preto”,”Marília”), substitui o valor ausente da variável estado para "São Paulo".

- Se a variável estado possui valores ausentes e a variável cidade a um dos seguintes valores ("Rio de Janeiro","Niterói","Teresópolis"), substitui o valor ausente da variável estado para "Rio de Janeiro".

c. Fazer a junção dos dois subconjuntos de dados, o subconjunto que contém as colunas (variáveis) estado e cidade, e o subconjunto que possui as demais colunas. Sugere-se utilizar o nó Joiner, seguindo a configuração já apresentada no Passo 4.

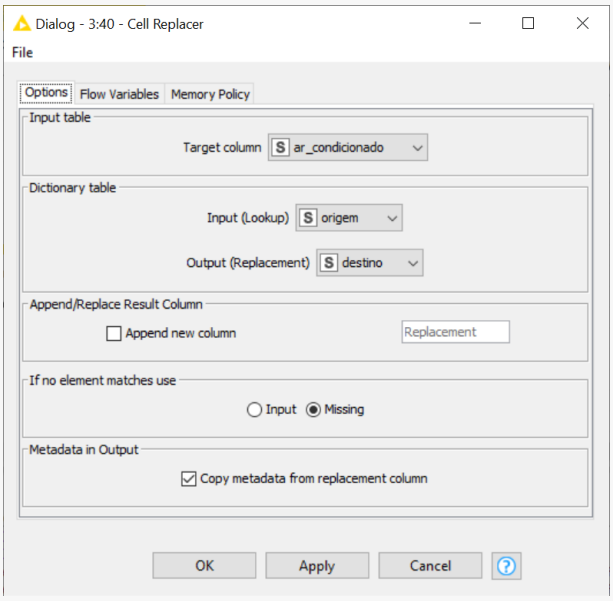
Passo 6. Agora você deve tratar os valores da variável “ar_condicionado” de modo a torná-los mais inteligíveis. Para isso, você deve substituir o valor “com_ar_condicionado” por “com ar condicionado” e “sem_ar_condicionado” por “sem ar condicionado”. Sugere-se utilizar as seguintes etapas:

1. Criar uma tabela de referência com o valor original (input) e o valor destino (output). Para isso, sugere-se utilizar o nó “Table Creator”, que da extensão ‘KNIME Base nodes” instalada junto com o Knime; configurá-lo conforme imagem abaixo:



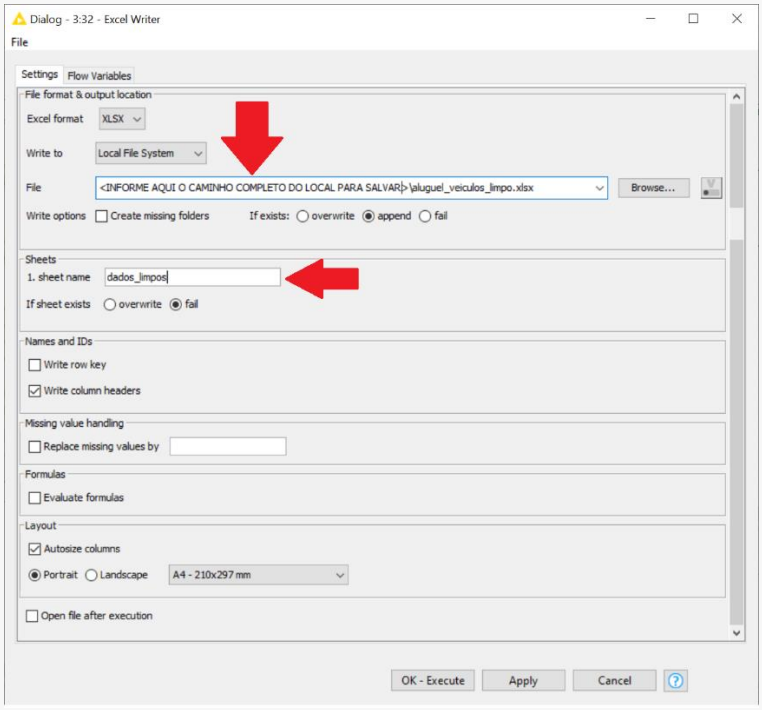
2. Substituir os valores conforme a tabela criada, onde a coluna alvo (*Target column*) é “ar_condicionado”, e temos as colunas origem e

destino respectivamente como Input e Output. Para isso, sugere-se utilizar o nó "*Cell Replacer*", que da extensão "KNIME Base nodes" instalada junto com o Knime; configurá-lo conforme imagem abaixo:



Passo 7. Crie um novo arquivo no Excel com o conjunto de dados tratado. Utilize o nó

e nomeie o arquivo para "aluguel_veiculos_limpo.xlsx". O exemplo de configuração deste nó na imagem abaixo. No campo File informe o caminho onde vocês deseja salvar este arquivo.



Atividade 3

Considerando o conjunto de dados tratado na atividade anterior e salvo no arquivo “aluguel_veiculos_limpo.xlsx”, aplique o que você aprendeu no módulo 2 para calcular:

- O coeficiente de correlação linear de Pearson entre as variáveis valor_total_locacao e cotacao_dolar.
- O R^2 da regressão linear entre as variáveis valor_total_locacao e idade_locatario.
- As Medidas de Centralidade (mediana) e Dispersão (quartis) da variável idade_locatario.

Atividade 4

Nesta atividade, vamos prever o valor total da locação para o conjunto de dados disponível no arquivo “aluguel_preverValorTotal.xls”, com os dados conforme figura a seguir.

qtde_portas	ar_condicionado	data_inicio_locacao	idade_locatario	genero	quilometragem	cotacao_dolar	Estado	cidade	qtde_diarias
4 portas	com ar condicionado	2021-05-13	18	Masculino	930,7	6,93	Minas Gerais	Divinópolis	2
2 portas	sem ar condicionado	2021-06-01	23	Feminino	957,44	4,41	Minas Gerais	Belo Horizonte	3
2 portas	com ar condicionado	2021-04-14	28	Feminino	923,3	8,81	Rio de Janeiro	Niterói	1
2 portas	com ar condicionado	2021-05-14	20	Feminino	665,44	4,67	Goiás	Goiás	4
2 portas	com ar condicionado	2021-04-18	18	Feminino	930,7	5,93	Minas Gerais	Betim	2
4 portas	com ar condicionado	2021-06-02	21	Feminino	871,52	4,26	São Paulo	São Paulo	5
2 portas	sem ar condicionado	2021-06-01	23	Masculino	957,44	4,41	Minas Gerais	Divinópolis	4
2 portas	com ar condicionado	2021-04-18	18	Feminino	930,7	5,93	Minas Gerais	Belo Horizonte	4
4 portas	com ar condicionado	2021-05-17	18	Masculino	501,94	6,94	Rio de Janeiro	Rio de Janeiro	3
2 portas	com ar condicionado	2021-04-19	20	Masculino	447,87	4,74	Rio de Janeiro	Niterói	2

Para realizar esta análise preditiva, vamos utilizar Regressão Linear Simples, e a base de treinamento será o arquivo “aluguel_veiculos_limpo.xlsx” criado no passo 7 da atividade 2. Para isso, utilizando a ferramenta *Knime Analytics Platform*, criar um workflow que leia os dois arquivos, execute o treinamento de Regressão Linear no arquivo “aluguel_veiculos_limpo.xlsx”, e aplique o modelo na base do arquivo “aluguel_preverValorTotal.xls”. Segue abaixo o conjunto de passos que devem ser seguidos para executar esta atividade.

Passo 1. Inserir um nó do tipo

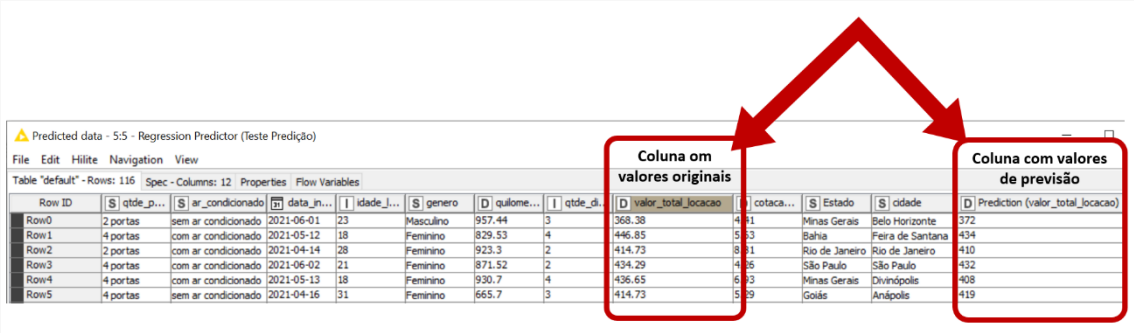
para coleta/leitura dos dois conjuntos de dados disponíveis nos arquivos “aluguel_veiculos_limpo.xlsx” e “aluguel_preverValorTotal.xls”. Lembre-se que deve ser um nó para cada arquivo.

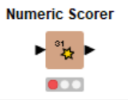
Passo 2. Para o conjunto de dados “aluguel_veiculos_limpo.xlsx”, utilize o nó

para treinar o algoritmo de análise e gerar o modelo de predição.

Passo 3. Submeta o modelo gerado ao próprio conjunto de dados usado no treinamento para depois avaliar a acurácia do modelo. Para isso, utilize o nó

para gerar uma nova coluna com a predição da coluna de valor_total_locacao. Avalie o resultado da previsão comparando a coluna com os valores originais (valor_total_locacao) e a coluna com os valores de predição, conforme exemplificado na imagem abaixo.

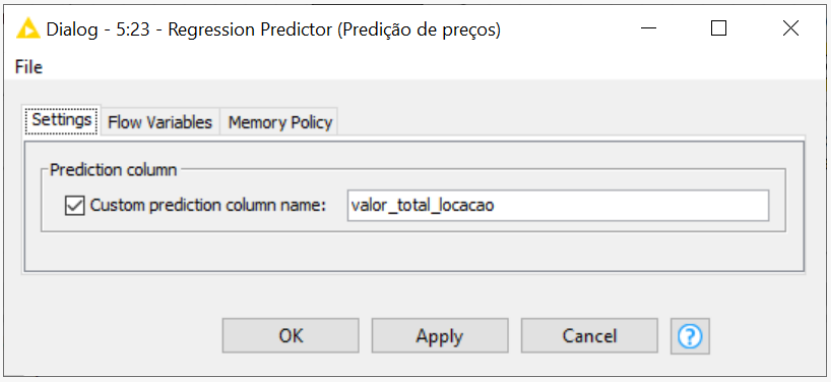


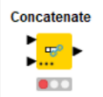
Passo 4. Utilize o nó 

para medir a acurácia do modelo. Analise os resultados, qual o valor do coeficiente de determinação (r^2).

Passo 5. Submeta o modelo de predição gerado no passo 3 ao conjunto de da onível no arquivo “aluguel_preverValorTotal.xls”. Para isso, use o

nó . Conecte o modelo do passo 3 à sua respectiva porta de entrada neste nó (quadrado azul) e o conjunto de dados à outra porta, e execute o nó. Nas configurações do nó, atribua o nome “valor_total_locacao” ao campo que será gravada a predição, conforme demonstrado na imagem abaixo.



Passo 6. Unifique os dois conjuntos de dados, para isso, utilize o nó 

. Conecte a saída do nó

em uma das portas de entrada do nó “*Concatenate*” e a saída do nó

do passo 5 na outra porta. Execute o nó e analise o resultado. Note que o resultado deve conter 126 linhas, sendo 10 do nó

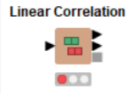
e 116 do nó

. Observe ainda que algumas linhas ficaram o RowID duplicado (RowID_dup). Podemos refazer a numeração dos RowIDs utilizando o nó

. Observe o resultado.

Passo 7. Para gerar um maior nível de detalhes no nosso conjunto de dados antes de construir o painel de visualização, vamos criar uma nova variável (coluna), chamada valor_diaria, para guardar o valor da diária de cada locação. Ou seja, o valor da diária é calculado pela divisão da variável valor_total_locacao pela variável qtde_diaria. Formate seu resultado para conter até 2 casas decimais. Para isso, utilize a seguinte sequência de nós:

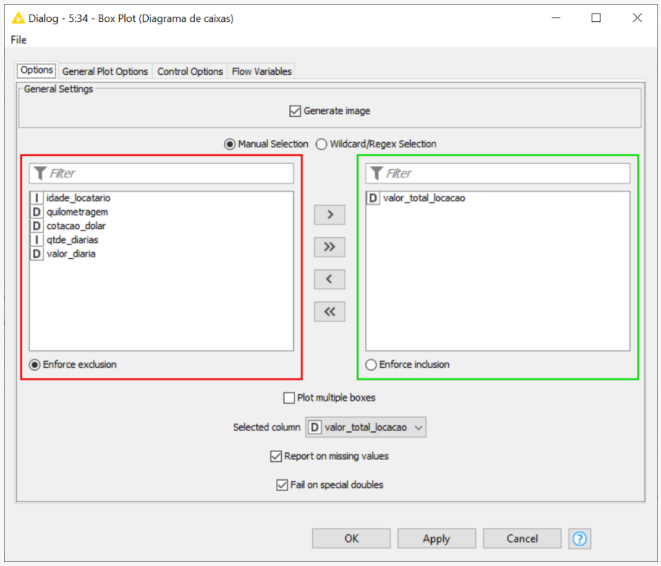
. O primeiro nó deve ser usado para criar a nova coluna com a fórmula de cálculo dos valores, e o segundo para arredondar o valor a 2 casas decimais.

Passo 8. Submeta seu conjunto de dados a uma correlação linear. Vamos utilizar o nó 

. Analise seus resultados observando a matriz e as medidas de correlação entre as variáveis.

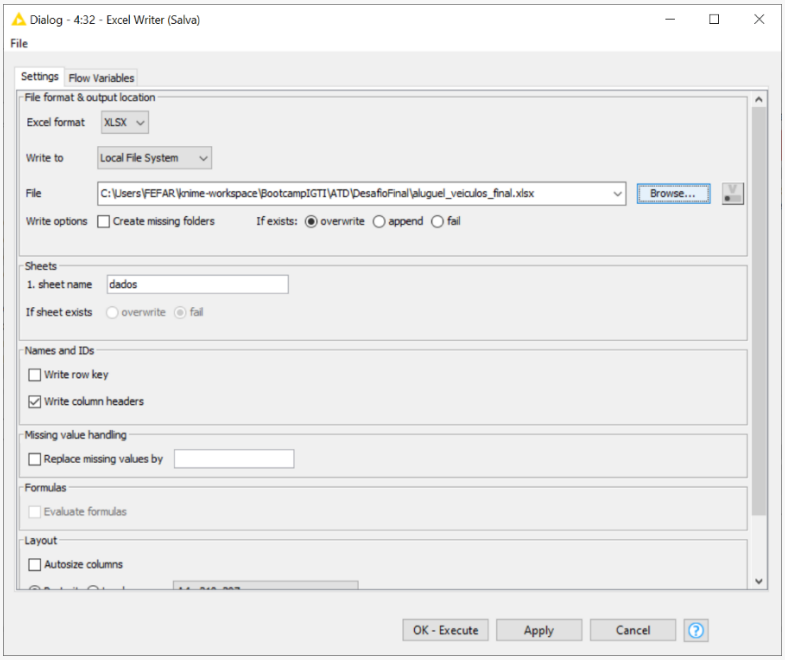
Passo 9. Gere o gráfico de caixa para a variável valor_total_locacao. Utilize o nó

. Configure seu nó conforme imagem abaixo. Execute o nó e analise seus resultados observando se existe algum *outlier*, os valores dos quartis, valor máximo e mínimo.



Passo 10: Salve os conjuntos de dados em um novo arquivo no Excel. Utilize

o nó e nomeie o arquivo para “aluguel_veiculos_final.xlsx”. Esse arquivo será usado para construção do nosso painel no power BI.



Atividade 5

Nesta atividade, vamos construir nosso painel visual (dashboard) a partir dos dados tratados nas atividades anteriores. Crie seu painel no Power BI utilizando o conjunto de dados do arquivo “aluguel_veiculos_final.xlsx” gerado no passo 10 da atividade anterior. Seu dashboard deve responder as seguintes perguntas:

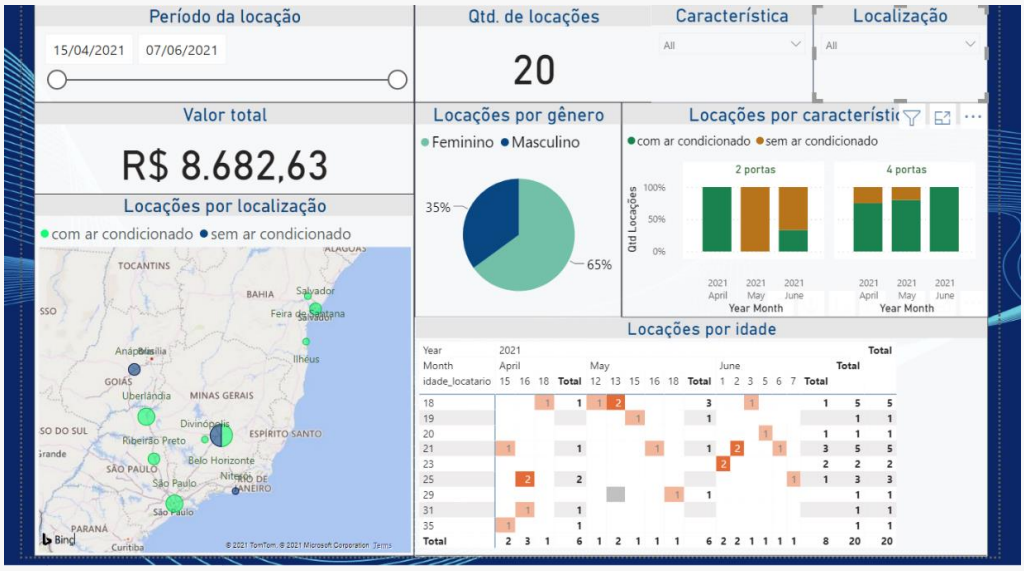
- Considerando a idade dos locatários, qual a idade que mais contribuiu financeiramente com o faturamento da empresa? A idade de maior contribuição varia conforme o gênero?

- Para os veículos que possuem Ar-condicionado, qual o mês que teve o maior faturamento (R$) dentre todas as locações realizadas entre os meses de abril e junho?

- Qual a cidade que possui o maior número de locações dos veículos de 4 portas? Para responder esta pergunta, é necessário criar uma nova medida no seu conjunto de dados para calcular a quantidade de locação. A fórmula usada pode ser: `Qtd Locações = COUNTROWS(dados)`

- Quais características (ar_condicionado e qtde_portas) possuem o maior volume de locações?

Para realizar esta atividade, você deve construir visualizações que permita identificar o faturamento e a quantidade de locações pela idade do locatário, pelo gênero, por características do veículo (ar condicionado e quantidade de portas), por período de locação, por cidade. A imagem a seguir apresenta um exemplo de dashboard criado com dados amostrais.



Correlação dos módulos com as atividades e questões:

Módulo	Atividade	Questão
Módulo 1	Atividade 1	Questão 1
Módulo 1	Atividade 2	Questões 2 e 3
Módulo 2	Atividade 3	Questões de 4 a 7
Módulo 3	Atividade 4	Questões de 8 a 11
Módulo 4	Atividade 4	Questões de 12 a 15