04

Seleção dos dados

Transcrição

[0:01] Bom, agora nós temos o objeto com todos os registros e colunas. Porém, você não vai utilizar todas essas colunas para fazer as análises, porque de acordo com os pedidos de análise da instituição você pode analisar, selecionar as colunas de interesse. Consequentemente, o objeto vai ficar mais leve. Eu vou mostrar já para você a importância dessa etapa de seleção.

[0:25] Primeiro, você vai criar um vetor com o nome das colunas. O vetor já está virado aqui ó. São essas colunas que nós vamos utilizar durante todo o curso.

[0:34] Você vai salvar num objeto chamado colunas, vai atribuir, vamos alinhar aqui para ficar organizado. Pronto. Criei aqui um vetor chamado colunas. Agora, você vai utilizar uma função do dplyr para fazer a seleção apenas dessas colunas com base nesse objeto.

[1:01] Primeiro você vai colocar aqui o objeto que você deseja, original, que é o merge Enem, correto? Vai utilizar aqui o concatenador de função do dplyr que foi explicado no curso anterior, vai chamar a função select_, vai passar um parâmetro chamado dots, que ele já indica para você, e vai passar o objeto colunas como valor desse parâmetro.

[1:36] E esse resultado aqui dessa execução, você vai salvar num objeto chamado Enem. Vai salvar nele, vamos executar. Pronto, executamos, olhamos aqui do lado direito, na aba enviroments, temos os dois objetos: um chamado Enem e o merge Enem.

[2:01] Um com 23 colunas, que é o que fizemos agora, o Enem, e o outro com 37 colunas. Vamos mudar nossa visualização aqui para grid e eu vou mostrar uma coisa bem interessante. Vamos aumentar aqui o espaçamento.

[2:14] Olha só a diferença: o nosso objeto Enem tem 338 megas e o nosso objeto merge Enem tem 566. Então é muito importante você se atentar à seleção das colunas porque isso pode desocupar muita memória da sua máquina de trabalho, da sua estação de trabalho, assim, evitando prováveis problemas no futuro por falta de memória, principalmente na hora de fazer, elaborar os gráficos, que consome bastante memória.

[2:47] Agora você pode simplesmente apagar o merge Enem. Pronto. Você tem todos os dados em um objeto bem menor com todas as colunas desejadas anteriormente, aqui na nossa variável colunas, que serão utilizadas durante todas as análises até o final do curso.

[3:21] Depois de obtermos todos os dados necessários, ter feito a seleção agora anteriormente apenas as colunas que nós desejamos para fazer as análises, futuros gráficos, agora nós precisamos conhecer um pouco nossa base de dados, por exemplo, o que cada coluna indica, o que cada coluna armazena. Então, você pode fazer isso aqui com a função str enem, passando os dados, executa, pronto.

[3:55] Aqui você tem um resumo e vamos fazer uma exploração bem rápida de cada coluna da nossa base de dados para saber o que é, por exemplo, o que é uma coluna numérica, o que é uma coluna categórica e assim vai.

[4:19] No total, nós temos 23 colunas, correto? Com o número da inscrição, o ano que foi realizado a prova, essa coluna é numérica, né? O código do município, o nome do município, a UF de residência do candidato, da candidata, a UF da escola onde foi realizada, a idade do estudante, o sexo do mesmo, F, M, F, M; situação de conclusão, se ele precisa da prova em Braile ou não, município da prova, UF da prova, o tipo de presença em cada matéria, que é do tipo numérico, a nota de cada matéria também era para ser numérico, você pode observar aqui que está em chr, mas mais para frente nós vamos fazer uma exploração mais detalhada para saber porque está em chr, o tipo de língua, que significa o idioma da prova, da parte de

idiomas, se é inglês ou espanhol, o status da redação, se ele está presente, se ele fez ou não, e a nota da redação também que era para ser numérico, que era para ser uma nota, ou seja, uma nota decimal, mas está armazenado também como chr, ou seja, uma variável categórica.

[5:47] Essas variáveis que eram para ser numéricas e estão como categóricas, deve ser algum problema de inserção de dados, deve ter algum caractere especial, por exemplo, alguma letra que não era para estar, alguma pontuação que não era para estar e ela foi, o próprio R reconheceu essa coluna como categórica, como textual, como chr, porém, mais para frente nós vamos corrigir esse problema para gerar as análises e também os gráficos corretamente.