02

Importação e merge dos dados

Transcrição

[0:00] Primeiramente, você vai definir o diretório que você quer trabalhar.

[0:06] Para descobrir o diretório onde você está trabalhando, você vai utilizar a função getwd, executar e vai passar aqui no console o caminho.

[0:14] Esse é o diretório atual. Se você deseja mudar, você vai utilizar a função setwd, passar o caminho, por exemplo, o caminho que eu quero trabalhar é esse daqui, o caminho completo se for necessário, executar essa função e vamos verificar novamente... pronto, agora nós estamos trabalhando no diretório desejado.

[0:41] Em seguida, você vai importar as bibliotecas que serão trabalhadas utilizando a função library, vai chamar o pacote que você deseja, por exemplo, o primeiro aqui é data table, e executar.

[0:55] Pronto, habilitei o pacote data table.

[1:02] Vamos também trabalhar com o pacote dplyr, habilitei.

[1:05] E, por fim, o pacote de gplot2. Se você já fez o curso pré-requisito, o curso anterior chamado Dada Vis com 1 variável, você já vai ter instalado esses três pacotes, você só vai precisar habilitar.

[1:21] Caso contrário, você vai utilizar a função install.packages e passar o nome do pacote, por exemplo o gplot2, e vai executar essa linha de comando e, posteriormente, habilitar.

[1:38] Eu não vou instalar aqui porque já foi feito no curso anterior e eu já tenho esse pacote instalado.

[1:43] Agora que você já instalou e já importou os pacotes e definiu o diretório que você deseja trabalhar, vamos importar todos os dados que serão utilizados.

[1:53] Todos os dados estão no formato csv e estão disponíveis para você no curso.

[1:59] Primeiro, você vai importar cada registro, cada base de dados, que eles estão separados por ano, por exemplo, o ano de 2010 está em uma base de 2010, o ano de 2011, na de 2011 e assim por diante.

[2:13] Você vai utilizar a função fread, vai passar o nome do arquivo, enem 2010.csv, vai passar o parâmetro encoding, para definir a questão de acentuação e caracteres especiais na base de dados em português, que vai ser definido por UTF-8.

[2:36] Você vai salvar tudo isso num objeto chamado enem 2010.

[2:43] Colocar no atribuidor, vai executar e pronto.

[2:49] Aqui do lado direito, já podemos ver que os dados foram executados, ali enem 2010, havendo mais de 230 mil registros com 23 colunas.

[3:02] Agora você vai fazer isso para a base de 2011, apenas mudando o nome para a de 2011, vai carregar, pronto, carregamos aqui mais um, enem 2011, mais de 260 mil registros e 27 colunas.

[3:24] Eu vou cortar agora o vídeo, você vai fazer isso até 2017 e eu vou mostrar o resultado final do código para você.

[3:29] Pronto, aqui está o código. Você vai carregar todos os dados, nada vai mudar, só vai mudar apenas aqui o nome do arquivo, 2012, 2013, até 2017.

[3:40] Do lado direito, você pode observar aqui que temos todos os objetos, todos os dados já carregados.

[3:43] Após carregar todos os dados, agora é necessário, ao invés de trabalhar com os muitos objetos que nós temos aqui de 2010 até 2017, correto?

[3:55] Vamos trabalhar apenas com 1 objeto só, é mais fácil fazer uma junção em todos esses dados em um único objeto.

[4:06] E isso é possível utilizando a função rbind, e você precisa acrescentar como parâmetro o nome dos objetos que você deseja fazer o merge.

[4:22] Então vai ser de enem 2010 até o 2017. 2010, 11, 12, 13, 14, 15, 16, e por fim, 2017.

[4:40] Você vai salvar isso tudo em um objeto chamado merge enem, que vamos executar e, olha só: ao olhar aqui no console, aconteceu um erro.

[4:56] O que esse erro indica? Esse erro indica que os objetos para fazer o merge, esses objetos aqui 2010 até 2017 não são exatamente iguais.

[5:10] Nesse caso aqui, por exemplo, ele está indicando que tem objetos com 23 colunas e os outros não tem. E a gente olhando aqui o primeiro objeto tem 23 colunas, 27, 27, 31, 28, 28; e para utilizar essa função todos os objetos têm que ser idênticos na quantidade de colunas, por que ele faz o merge por linhas.

[5:33] Então, para resolver esse problema, faremos o merge das colunas em comum, adicionando o valor em a, ou seja, não definindo as colunas que não são em comum entre as bases.

[5:45] Essa solução pode ser feita aqui utilizando o parâmetro como o próprio erro já indica fio, o argumento fio, você vai adicionar aqui no final da linha "fio igual a TRUE" , vai executar novamente, e pronto, o merge está sendo executado.

[6:07] E no final teremos apenas 1 objeto com todos os dados.

[6:12] Pronto. Fizemos o merge de todas as bases em uma única base chamada merge enem.

[6:17] Se você olhar aqui do lado direito nessa janela, merge enem há mais de 2 milhões de registros com 37 colunas. Ou seja, ele fez um merge, uma combinação das colunas aqui, as colunas não comum ele inseriu o valor não definido, ou seja, na.

[6:34] Agora, nós podemos descartar de todos esses registros aqui, essas bases que não serão utilizadas mais, por quê?

[6:41] Nós vamos utilizar agora somente esse objeto chamado merge enem.

[6:46] Para eliminar esses outros registros, você vai utilizar a função rm, passando o nome dos objetos, pode dar um Ctrl C aqui, porque a gente já utilizou o nome dos objetos, e vamos apagar.

[7:01] Pronto, agora nós temos somente o objeto merge enem.

[7:08] Esse procedimento é muito importante e muito recomendado também porque a cada objeto armazenado e criado dentro do RStudio, ou se você estiver trabalhando em outra DE, ocupa a memória do computador, principalmente nessas nossas bases de dados que são muito grandes.

[7:24] Então, é recomendado que vocês sempre eliminem, apaguem os conjuntos de dados que vocês têm certeza que não vai ser utilizado mais.

[7:32] Nesse caso nós temos certeza porque todos esses dados aqui agora estão dentro do objeto merge enem, ok?