

Introdução

Transcrição

[0:00] Olá, seja bem-vindo a mais um curso da Alura, o curso Data Visualization - Gráfico multivariáveis. Meu nome é Gabriel e eu serei o instrutor durante todo o curso.

[0:13] Aqui, você irá aprender a criar códigos em R para fazer análises de dados por meio de gráficos, o que é muito mais fácil de se fazer do que analisar uma tabela com centenas de linhas ou até milhares de linhas, não é mesmo?

[0:24] Ao se trabalhar com análise de dados, visualização de dados, você precisa ter um problema em mãos para ser resolvido.

[0:32] O nosso problema a ser resolvido será atender a algumas solicitações de um cursinho preparatório para o exame do ENEM. Ou seja, nós vamos trabalhar com a análise dos microdados do ENEM entre 2010 e 2017.

[0:45] Esse cursinho tem filiais no Ceará, Minas Gerais, Rio Grande do Sul e DF, porém nós vamos fazer uma análise nacional e às vezes dando um foco maior nas filiais.

[1:01] O cursinho deseja obter informações como: identificar os tipos de idiomas que estão sendo mais escolhidos, o inglês ou o espanhol; a situação escolar das pessoas que estão fazendo a prova; a média de idade dessas pessoas; comparação das médias em ciências humanas e matemática por idade; média de cada matéria entre os anos 2010 até 2017; comparação das médias entre matemática, ciências humanas, redação por estados, inclusive, dando um destaque maior às filiais onde o cursinho tem; também deseja obter informações de valores estatísticos, como cálculos quartis, mediano, por estado e se há algum valor discrepante em um desses estados. São informações um pouco complexas de se obter, porém trabalhando com análise de dados fica bem mais fácil analisar essas informações.

[1:54] Também deseja obter informações sobre a média da redação por estado, comparando com a média nacional.

[2:03] A base de dados que nós iremos trabalhar, ela está disponível aqui no site do INEP. Esse site do INEP possui diferentes microdados, inclusive do ENEM.

[2:18] Aqui você pode ver ENAD, Encceja, entre outras bases de dados e tem aqui do ENEM. Nós vamos trabalhar aqui do ENEM de 2010 até 2017, está atualizado recentemente, em março de 2019.

[2:31] Mas esses dados também vão estar disponíveis no curso para você fazer o download porque os dados já passaram por uma etapa de pré-lançamento, um tratamento especial porque aqui os dados estão em diferentes formatos e colocamos tudo em um formato único, que é no formato CSV.

[2:50] Então você pode baixar os dados no curso, vai estar disponível e, posteriormente, se você tiver interesse, pode acessar os outros dados aqui no site do portal do INEP, o link também vai estar disponível no material do curso.

[3:04] Neste curso, você irá aprender a fazer algumas manipulações e transformações nos dados que são necessárias para gerar os dados.

[3:13] Por exemplo, aqui, você vai substituir o valor 1 e f pela palavra completa feminino, você vai aprender a explorar algumas colunas, por exemplo esse tipo de linguagem, você vai aprender a substituir algumas outras flags, por exemplo aqui do 0 que significa inglês, então o 0 vai se converter em inglês. Aqui na situação escolar que também são armazenados em

flag, você vai trocar os números pela respectiva situação, neste caso aqui, 1 significa concluído. Você também vai aprender a trabalhar, a converter strings em números quando será possível.

[3:50] Isso para que? Para você normalizar as colunas onde armazenam as notas, por exemplo aqui as notas de ciências da natureza, linguagens e códigos.

[4:00] E também vai aprender a fazer diferentes filtros, a trabalhar no centro com pacote dplyr, por exemplo aqui "filter", vamos fazer alguns filtros e selecionando alguns dados importantes, alguns dados necessários para aquele gráfico.

[4:15] Então não vamos trabalhar às vezes com todo o banco de dados e sim com uma pequena amostra que será o suficiente, ok?

[4:23] Então vamos lá, você vai fazer tudo isso, os gráficos, você vai trabalhar aqui, por exemplo, fazer esse gráfico de barras, você vai aprender a fazer esse outro gráfico de barras dividido por camadas, também vai aprender a fazer esse gráfico de pirâmide, o outro gráfico de pontos, gráfico de linhas, trabalhando sempre com muitas variáveis, o gráfico de bolhas que tem 4 variáveis num mesmo gráfico, gráfico de boxplot, e por fim vai aprender a trabalhar aqui a montar essa página que armazena diferentes gráficos numa mesma página.

[5:07] Porém, vamos aprender aqui a montar um gráfico por vez e até chegar nesse resultado final, ok? Vamos lá? Dar início aos nossos estudos?