01

Verificando Duplicidades da base

Transcrição

[00:00] Tudo bem, galera? Vamos continuar aqui o nosso projeto de análise de dados utilizando o R. Com a base já carregada dentro do R, nós carregamos a base aqui, já está a base de aluguel com 32.960 registros. Lembrando, qual que é o nosso desafio aqui?

[00:19] Você como cientista de dados, recebeu essa base de aluguel com preço de imóveis, com algumas características. E aí, você tem o desafio de passar para a área de planejamento algumas informações para a tomada de decisão, então a gente está com a base aqui e a gente vai fazer essa análise de dados.

[00:39] Primeiramente, antes de a gente começar a fazer essa análise de dados, eu vou incluir um chunk aqui, então já está com o R Notebook aberto aqui na linha, a gente tem que sempre colocar o chunk, incluir aqui com a tecla de atalho, “Ctrl + Alt + Enter” ou então, a gente pode inserir por aqui, como eu já falei.

[01:00] A primeira coisa, antes de começar essa análise, é preciso nós verificarmos a estrutura da base, verificarmos se existe em duplicidade na base que a gente recebeu, se existem valores ausentes. Então, são algumas análises preliminares que vamos verificar para realmente, depois continuar fazendo as demandas que a gente recebeu da área de planejamento, os filtros e seleções que a gente vai fazer ao longo desse projeto.

[01:33] Uma biblioteca bastante interessante que a gente precisa instalar aqui, é a biblioteca chamada tidyverse, essa biblioteca... Então, vamos colocar aqui: “install.packages()”, aqui dentro do parênteses entre aspas, tidyverse, só lembrando que eu vou... A gente tem diversas maneiras de instalar esse pacote e eu estou instalando aqui pelo código, mas de novo, a gente também poderia estar instalando aqui pelo menu.

[02:18] Aqui, a hora que a gente coloca a biblioteca tidyverse, ele já aparece aqui, tadyverse. Eu vou dar um install, ele vai dar um start aqui, por isso que é interessante já instalar no começo. Instalou e depois, na sequência, a gente tem que sempre chamar a biblioteca, chamando aqui pela library, de novo tidyverse.

[02:48] Ele já aparece aqui em baixo, eu vou comentar aqui, só para não rodar de novo a instalação, comentei aqui, então ele não vai instalar de novo e eu vou executar aqui. Executando, ele vai carregar, isso aqui é interessante olhar, o que que acontece?

[03:06] Quando a gente instala essa biblioteca tidyverse, ele já engloba várias outras bibliotecas bastante interessantes, que a gente vai precisar. Então, a gente tem aqui ggplot para construção de gráficos; tibble, que é tabelas; readr, que a gente até usou para carregar a base aqui; o purrr; o dplyr; forcats, entre outras aqui.

[03:28] Então, ela já faz um conjunto de pacotes, então isso, até facilita a nossa análise para frente. Lembrando aqui, que a gente já está com a tabela carregada, um comando para olhar a estrutura... tem alguns comandos para a gente olhar a estrutura da base, que é o “str(aluguel)”, que é o nome da base.

[03:48] Então, esse comando aqui, eu vou dar “Ctrl + Enter”, ele vai mostrar a estrutura da base, então eu tenho aqui todas as... eu tenho 32.960 observações, tenho cada uma das colunas, tipo, bairro, quartos, vagas, suítes, área, valor, condomínio e o IPTU.

[04:14] Tem um outro comando similar também aqui, a... esse comando de estrutura, eu vou abrir um outro chunk aqui, “Ctrl + Alt + I”, comentar aqui: “olhando a estrutura da base de dados”. Tem um outro comando bem similar a esse comando estrutura, str, que é chamado de glimpse.

[04:44] Ele faz exatamente a mesma coisa, porém ele é... o output dele, a saída é um pouco diferente do str, então dá uma olhada aqui, ele parece até uma saída um pouco mais limpa. Então aqui, ele tem aqui, 32.960 observações, nove variáveis, a variável tipo é caractere, bairro é caractere, quartos é dbl, aí é numérica, então é a mesma coisa que numérica.

[05:13] Vagas; suíte também numérica; área, numérica; valor, condomínio, IPTU. Veja que, condomínio e IPTU, a gente já consegue ver uns NAs aqui, esse NA, ele é como se fosse valor ausente, como se fosse não, é um valor ausente no R. O R entende isso aqui como ausência de valor, ausência numérica.

[05:34] Depois a gente vai verificar como é que a gente trata esses casos. Na sequência, o que que a gente precisa verificar? Então, verificando se a base existe duplicidade, isso é muito importante verificar.

[05:58] Toda a base que a gente recebe, quando a gente vai fazer uma análise de dados, a primeira coisa, bem simples, verifica se a base não tem nenhuma duplicidade, porque se tiver duplicidade, os dados, o resultado, as conclusões que você pode estar tirando dessa análise, ela pode ser inconclusiva ou pode ser destorcida.

[06:20] Então, não custa nada verificar sempre, recebeu a base, dá uma olhada na estrutura, no tipo de variável, se existe valor ausente, se não existe e olhar também se tem duplicidade ou não. Para a gente verificar se tem duplicidade, tem um comando chamado no R: unique.

[06:40] Esse comando unique, ele vai verificar exatamente se a base está única ou não, ele já vai tratar essas duplicidades. Eu vou salvar essa saída da base tratada, eu vou chamar de “aluguel_t”, aluguel tratado, vamos dizer assim, ele vai receber o quê? A gente vai usar esse comando unique, ele aparece aqui, unique, aluguel, na base aluguel.

[07:06] Então, simples assim, bem tranquilo. Eu vou executar aqui. Aparentemente ele executou e aí, eu vou chamar de novo... dar um glimpse aqui, a base “aluguel_t” agora, tratada e vou só executar essa linha. Executando, veja bem, a gente tinha a base sem duplicidade... com duplicidade, sem tratamento, a gente tinha 32.960 registros.

[07:38] Agora, a gente tem uma base de 31.800, ou seja, a base tinha, sim, duplicidade e essas duplicidades foram tratadas. A gente pode até investigar depois, se a gente quisesse isolar essas duplicidades, a gente poderia dar uma olhada para... a gente pode ver, “duplicidades”, eu vou chamar isso aqui de “duplicidades”, recebe exatamente o que?

[08:05] Ou melhor, se a gente pegar e executar só esse pedaço aqui: “unique(aluguel)”, dá um “Ctrl + Enter”, o que que ele vai fazer? Ele vai trazer exatamente as observações que ele tratou aqui já como sendo sem duplicidade. Então, agora a gente tem essa base.

[08:26] Então, isso é muito importante, toda a vez que a gente for fazer qualquer tipo de análise de dados, precisa verificar realmente, se tem alguma duplicidade ou não. Então, essa base aqui, a gente... agora, a base que a gente vai trabalhar com o nosso projeto, vai ser essa base com 32.800 [31.800] registros aqui, que ela está salva com aluguel tratado.

[08:54] Eu vou executar de novo aqui, então 31.800 registros tratados. Ela mantém a mesma estrutura, ela só... realmente, elimina os valores em duplicidade. Depois a gente continua, mais uma outra etapa para verificar, para fazer a análise de dados aqui, utilizando o nosso projeto.