03

Agrupando e Sumarizando Informações

Transcrição

[00:00] Vamos continuar o nosso projeto de análise de dados, utilizando o R. Só lembrando, o nosso projeto, como a gente tem o interesse de avaliar os dados da nossa base de aluguel.

[00:15] Imagina que a gente recebeu uma nova demanda da área de planejamento, onde a gente precisa agora gerar um relatório de forma organizada, agrupando os dados por tipo de imóvel e os respectivos valores para cada tipo.

[00:30] Então, agora, para fazer esse tipo de análise, nós vamos utilizar um comando... a gente vai utilizar basicamente dois comandos aqui no R, então vamos lá, vamos colocar aqui, gerando novas informações de forma agrupada e sumarizada.

[01:01] A gente vai utilizar basicamente dois comandos aqui, a gente vai utilizar o comando chamado group_by e o comando summarise no R, só lembrando aqui, que também, a partir de a agora, a gente vai utilizar aquele comando, aquela funcionalidade do pipe.

[01:21] O que que é o pipe? Só lembrando, era aquele comando que era asterisco, maior, asterisco, que comando... que tudo o que você digitava, ele passava... o que você digitou, o comando recebido para a linha de baixo. Esse comando aqui também, ele pode ser utilizado, mais fácil utilizar o “Ctrl + Shift + M”, que é a tecla de atalho.

[01:52] Então, eu vou abrir um novo chunk aqui, que é o “Ctrl + Alt + I”, abriu. Então, a gente vai continuar ainda fazendo... utilizando a base de aluguel tratada, então agora eu vou colocar o “Ctrl + Shift + M”, ele já aparece o pipe ali. Então, o que que o pipe faz?

[02:15] Olha, o R entende que a partir de agora, eu estou utilizando essa base aluguel. Então, ele está passando a base aluguel para a linha de baixo. Então, o primeiro passo é agrupar a base, então a gente vai dar o comando group_by na variável tipo, então a gente vai sumarizar aqui, vai agrupar o tipo de imóvel.

[02:38] E na sequência, o que que a gente vai fazer? O comando summarise, a gente vai sumarizar, eu vou criar uma nova coluna aqui chamada quantidade, igual a n, abre e fecha parênteses. Então, o que que ele está fazendo?

[02:56] Ah, e aqui, a gente tem que colocar também o “Ctrl + Shift + M”, uma ou outra pipe ou seja, ele passar esse comando aqui do group_by para a linha de baixo e aí, então, fazer o summarise. Então, vou executar, olha, o que que ele fez? Para cada tipo de imóvel, ele trouxe todas as quantidades.

[03:14] Então, a gente sabe que na base de dados tem 18.854 apartamentos, tenho 74 box de garagem, 937 casas e assim por diante, casa comercial 260, casa de condomínio 973, uma chácara. Então, tenho... aqui eu posso visualizar todos os tipos de imóveis que estão disponíveis nessa base de dados.

[03:43] Eu posso agora, na sequência aqui, eu vou abrir um outro chunk, “Ctrl + Shift + I”, opa, “Ctrl + Shift + M”, opa, aqui... Eu vou abrir um novo chunk, “Ctrl + Alt + M”, opa... “Ctrl + Shift + I”. Não estou conseguindo, a gente vem aqui... ou não lembra, adiciona o chunk aqui que vai funcionar também.

[04:09] Então, agora, a gente pode... eu vou fazer essa agrupamento por mais de um campo. Então, a gente pode fazer esse agrupamento por mais de um campo também. Então, eu vou utilizar agora para agrupar o tipo e o bairro. Então, agora a gente vai agrupar a base de dados.

[04:33] Então, de novo, eu estou utilizando a base aluguel tratada, “Ctrl + Shift + M”, passo o chunk para baixo, vou subir aqui para ficar mais visual. Faço o group_by, agora eu faço por dois campos, tipo, virgula e bairro. “Ctrl + Shift + M”, dou outro chunk para passar para baixo, sumarizo de novo a base agora, mesma coisa.

[05:07] Vou sumarizar retornando a quantidade, igual a n, abre e fecha, ele vai fazer a contagem de todos os imóveis agora por tipo e bairro. Vou executar aqui. Então, olha, agora, o que que ele fez? Para cada um do tipo, ele disponibilizou todos os bairros possíveis, apartamento aqui, ele colocou apartamento no bairro da Abolição tem 10 imóveis, no bairro da Água Santa tem oito e assim por diante.

[05:38] A gente pode andar aqui também dando o preview aqui na base, a gente vai ter toda a base aqui, 97 páginas aqui, aonde ele abriu todos esses tipos e imóveis e para o bairro. Então, eu posso abrir, fazer esse group_by, essa sumarização por mais de um campo.

[06:00] O comando group_by, o comando summarise, ele aceite não só somar outras quantidades, a gente pode ter outros comandos para resumir a base de dados. Então, vamos utilizar agora... utilizando outros comandos no summarise para resumir a base de dados.

[06:35] Então, eu posso utilizar, quais são esses outros comandos? Eu posso utilizar o máximo, que é o max, o mínimo, posso utilizar a média e assim por diante, tem outros comandos que a gente pode estar utilizando. Então, vou mostrar aqui para vocês, como é que a gente vai fazer isso.

[06:55] Vamos abrir um novo chunk aqui. Então, eu vou fazer agora um agrupamento para retornar o valor médio. Então, eu vou agrupar os tipos de imóveis, vou fazer uma sumarização por tipo de imóvel, retornando o valor médio de cada tipo. Como é que a gente faz isso?

[07:18] De novo, aponta aqui a base aluguel tratada, coloca um chunk, faço o group_by, da mesma forma que a gente fez, vou fazer agora de novo por um capo só, que é o tipo, tipo de imóvel e aqui no summarise, agora, a gente vai trocar. Agora, eu quero o quê?

[07:38] Eu quero a... Eu vou retornar... eu posso retornar mais de um indicador, mais de um capo aqui também. Então, eu vou retornar o mesmo campo que a gente estava retornando aqui da quantidade, que expressa por esse comando aqui.

[07:45] Eu vou colocar uma vírgula, aonde eu vou trazer agora a minha... eu vou aguardar na minha variável média, chamar uma nova média, num novo campo chamado média, aonde ele vai trazer o comando mean, que é de média. E aí, ele vai trazer média de quem? Eu vou trazer a média do valor.

[08:15] Só lembrando, o valor é o campo aonde tem a informação de valor do aluguel. Ah, esqueci aqui do “Ctrl + Shift + M” no chunk, não pode esquecer, sempre... na última linha, ele não precisa do chunk, sempre lembrando, chunk... Chunk, não, o pipe aqui, ele sempre vai servir para passar o comando que você está fazendo na linha, para a linha seguinte.

[08:42] Vou executar. Então, aqui a gente retornou o valor médio de aluguel de cada tipo de imóvel. Então, o valor médio do tipo de aluguel de apartamento é... eu tenho 18.854 alugueis. Ah, ele não retornou o valor médio aqui para apartamento, por quê? Porque ele encontrou o NA, o que que é o NA?

[09:07] O NA é o valor ausente, a gente vai... depois eu vou mostrar para vocês como é que mostra, trata esse valor ausente. Quando existe esse valor ausente, eu não consigo fazer nenhum tipo de transformação aritmética, eu não consigo fazer nenhum tipo de operação aritmética.

[09:25] Então, como eu pedi a média aqui, ele não fez, já de box de garagem, ele fez, não tem nenhum valor ausente. Então, eu tenho aqui 74 imóveis considerados tipo de box, garagem, valor médio é dois mil e setenta e oito, de casa

também existe valor ausente, casa comercial, ele trouxe aqui 260 imóveis disponíveis, com valor médio de 14.439.

[09:50] E assim por diante, Casa de Vila, com 245, valor médio, 1.582. E aí, a gente pode paginar aqui, olhar também... Então, esse group_by, junto com o summarise, trazendo outros valores também é bem interessante para a gente observar onde tem valor ausente, onde não tem, então facilita bastante.

[10:11] Do mesmo modo, eu poderia mudar aqui, trazendo, por exemplo, o valor máximo ou o valor mínimo, vou executar, troquei aqui pelo máximo, vou executar de novo, ele vai trazer... Aí, aqui, a gente poderia mudar também, então max. Vou executar de novo, então, valor máximo aqui e outra coisa... mesmo trocando por mínimo.

[10:34] Então, tem os principais... operações matemáticas aqui, que a gente pode trazer. Existe um outro comando que facilita um pouco... retorna, na verdade, essa sumarização, esse agrupamento, que é o comando summary.

[10:53] Então, eu vou mostrar um outro comando bastante utilizando para sumarizar, para resumir a base de dados, fazer de forma bem rápida e todas as variáveis, que é outro comando, que é o... bem parecido com o summarise, mas ele chama summary. Esse comando summary, como é que ele vai funcionar?

[11:24] Ele vai funcionar da seguinte maneira, ele é bem simples até, a gente só colocar, summary, agora, não mais summarise, mas summary mesmo e o nome da base, aluguel tratada. Então, assim, o que que eu vou fazer?

[11:42] Ele vai sumarizar toda a base aqui para todos os campos, ele vai fazer um resumo dos dados, uma estatística descritiva de alguns indicadores, para a gente visualizar a base, como é que está. Então, vou executar aqui e vamos analisar. Então, o que que ele trouxe?

[11:57] Ele trouxe aqui para variáveis que são caracteres, ele não vai fazer nenhum tipo de operação. Por exemplo, tipo, ele não fez nada, ele só contou aqui, ele só fez uma contagem. Então, a variável tipo, aqui, eu tenho 31.800, tipos diferentes, bairro também é um campo categórico, caractere.

[12:21] Agora, nas variáveis que são retornados valores, então ele faz um resumo aqui dos dados, quartos, eu tenho o mínimo de zero quartos, esse 1SQS é o primeiro (quartil), a mediana, a mediada são dois quartos, a média é 1.77 e o terceiro quartil, que é 75% dos dados ordenados, três quartos e máximo, apareceu um com 100 quartos aqui.

[12:48] Então, esse tipo de comando aqui, ele serve justamente também para a gente verificar se existem valores discrepantes na base, por exemplo, apareceu um dado aqui, que está falando que tem um... dados de imóveis, um imóvel com 100 quartos, então tem que checar mesmo, ver se é isso, se não é.

[13:06] Aí, para todos os campos da nossa base, quartos, vagas, suítes, aquele valor de ausente, que eu falei para vocês, de valo, que a gente tinha feito, então eu tenho aqui o número mínimo do valor disponível para imóveis, é 75, um valor bem baixo também, depois teria que checar se é aquilo mesmo ou não.

[13:27] O primeiro (quartil) aqui da base é 1.500, a mediana é 2.800, a média de aluguéis é 13.125 e aqui, ele aparece, aqui, número máximo, também aparece um número bem absurdo. Então, isso aqui serve para a gente fazer um tratamento também da base de dados, nesse valor aqui máximo, mostrando aqui, provavelmente deve ser um erro de digitação.

[13:50] Então, também serve para a gente pegar e fazer algum tratamento de (outlier), talvez desconsiderar esse valor e aqui, são os NA's, os NA's são os valores ausentes que eu comentei, então ele aparece aqui também. Então, para valor, tem 17 valores ausentes, para condomínio tenho 3.949 valores ausentes, IPTU, eu tenho 9.886 valores ausentes.

[14:17] Então, esses valores ausentes, a gente precisa também fazer um tratamento na base e depois a gente vai... eu vou mostrar para vocês como é que faz, existem vários tipos de tratamento, eu vou mostrar alguns tipos de tratamentos de valores ausentes.

[14:32] Essa parte aí, era isso que eu queria mostrar para vocês. Então, até mais.