4.3 - Tratando valores discrepantes - outliers (parte 2)

00;00;00;05 - 00;00;23;17
O que fazemos com esses dados agora que a gente identificou quais são os pontos discrepantes? Uma abordagem simples é remover os valores discrepantes do conjunto de dados. Isso pode ser feito manualmente ou utilizando métodos automáticos baseados em critérios estatísticos. No entanto, deve-se ter cuidado ao remover os outliers, pois isso pode distorcer a distribuição dos dados e influenciar as conclusões.

00;00;24;02 - 00;00;47;06
Em vez de remover os outliers, outra abordagem é substituí-los por valores mais adequados. Isso pode envolver a substituição por estatísticas robustas, como a mediana, em vez da média (que é sensível a valores discrepantes) da mesma forma que a gente viu na aula de valores faltantes ou substituir pelo valor máximo, se fizer sentido olhando outras variáveis, como o cargo, a região e tudo mais.

00;00;47;14 - 00;01;09;19
Então, mais uma vez, é muito importante estudar o conceito, o contexto dos dados, beleza? Ainda há casos em que é necessário manter esses valores, pois apesar de serem eventos raros, são reais sim, e importantes para análise. Então, de novo, é importante considerar o contexto e a natureza dos dados ao lidar com os outliers. No nosso caso, nós temos uma coluna de faixa salarial.

00;01;10;15 - 00;01;36;07
Vamos ver se essas pessoas com salário acima do limite superior marcaram uma faixa salarial correspondente e a gente pode fazer o quê? Um filtro, que é assim: dados (que é a nossa tabela), colchetes dados. E aí, qual é a coluna que a gente está querendo ver? É a de salário. E coloca aquele símbolo matemático de “acima”, os salários acima do limite superior, porque a gente está querendo filtrar esses outliers e ver o quê?

00;01;36;07 - 00;02;08;23
A faixa salarial. Então, aqui a gente tem um filtro. Aqui a gente vai pegar todos os salários que estão acima do limite superior e agora a gente quer ver a faixa salarial deles. Ok. A gente coloca aqui um value counts no final, porque a gente consegue ver a quantidade por faixa salarial. Bom, então aqui a gente tem 19 outliers que estão acima de 40 mil e 3 outliers que estão entre 30 mil e 40 mil.

00;02;08;23 - 00;02;33;03
E a gente pode fazer que nem a gente fez com idade para poder ser mais assertivo na hora de substituir esses valores. Podemos fazer a média dos salários, que são acima de 40 mil, mas que não sejam outliers e substituir nos outliers e depois fazer a média dos valores entre 30 mil e 40 mil que não sejam outliers e substituir. Vou fazer um e o outro é para vocês.

00;02;33;05 - 00;02;57;26
Então vamos prestar atenção aqui. Bora lá! Primeiro a gente vai fazer o filtro da faixa salarial de 30 mil a 40 mil e acima de 40 mil fica para vocês. Dados (que é a nossa tabela), e a coluna faixa salarial, que já apareceu aqui para mim, igual...

00;03;01;12 - 00;03;25;02
Vou copiar exatamente como tá aqui, colocar entre aspas simples ou duplas. Ok. Esse é o primeiro filtro. Eu estou filtrando todas as pessoas que marcaram que são da faixa salarial de 30 mil a 40 mil. Só que eu quero desconsiderar os outliers. Então, vai ser um outro filtro. Então também já vou colocar esse primeiro filtro aqui entre parênteses. Selecionei tudo, cliquei aqui nos parênteses e tal.

00;03;26;19 - 00;03;52;06
Uma coisa "E" outra, então eu uso o &. Ok, meu próximo filtro entre parênteses que vai ser: dados... E o que a gente quer mesmo? A gente quer olhar a coluna de salário agora, então: salário... E a gente quer desconsiderar os outliers. A gente tem um limite superior. Então a gente quer todos os salários que estão abaixo desse limite superior, porque acima são outliers.

00;03;52;06 - 00;04;20;27
Então a gente quer tudo que esteja abaixo. Vamos usar aquele símbolo matemático bonito de tudo que está abaixo, de menor. Vamos copiar aqui igual o limite superior aqui, colar... Beleza. A gente fez dois filtros: faixa salarial de 30 a 40 e todos os salários que estão abaixo do limite superior. O que a gente quer, mesmo? A gente quer a média. Então já vamos colocar aqui: ponto min e os parênteses.

00;04;21;05 - 00;04;39;08
Executando aqui, deu um erro gigantesco falando que não tem como converter alguma coisa numérica. Vamos revisar tudo que a gente queria fazer. A gente queria o filtro de faixa salarial, a gente queria o salário, e a gente quer calcular a média de salário. Então a gente colocou o ponto min. A gente esqueceu de colocar o que a gente está calculando de média, que é a de salário.

00;04;39;25 - 00;05;09;25
Então, antes do ponto min, a gente tem que colocar qual é a coluna que a gente está calculando a média: salário. E agora a gente vai rodar e agora a gente tem um valor, olha, de 39 mil. Então a média dos valores de 30 mil a 40 mil que estão abaixo do limite superior é 39 mil. Ok, vamos colocar isso numa variável para poder facilitar. Média underline 30 underline 40.

00;05;10;09 - 00;05;41;18
Ok, vamos executar? Ótimo. E aí a gente vai querer substituir naqueles valores outliers. E a gente pode o quê? Localizar eles primeiro. Para localizar, a gente pode usar o quê? Ponto loc. Então vamos lá: dados (que é a tabela) ponto loc, a gente abre o nosso colchete e o que a gente quer localizar? A gente quer localizar todas as pessoas que marcaram que são de 30 mil a 40 mil na faixa salarial.

00;05;42;11 - 00;06;12;08
Então, primeiro filtro... vamos colocar o primeiro filtro devagar. A gente pode até copiar esse filtro do jeito que está aqui para facilitar. Aqui copiamos, colocamos embaixo. Então, o primeiro filtro foi faixa salarial de 30 a 40 mil. "E", ou seja, &, o segundo filtro que é: agora a gente está querendo pegar os outliers, a gente está querendo identificar os outliers. Então vamos colocar todos os salários que estão acima do limite superior.

00;06;12;22 - 00;06;40;22
Ok, então vou copiar do jeito que está aqui para facilitar. Vou colar e vou trocar esse símbolo matemático porque agora eu não quero o que está abaixo, eu quero o que está acima. Vamos colocar aqui: se a gente executar, a gente vai ter como retorno a tabela, só que a gente não quer a tabela, a gente quer os valores numéricos ali da coluna de salário, que são esses outliers.

00;06;40;22 - 00;07;11;03
Então a gente coloca a vírgula e coloca salário, que é a coluna que a gente está querendo. Então executando, a gente vai ter aqui os outliers. Olha que legal: aqui em cima a gente tem que foram 3 outliers identificados, 3 valores dessa faixa salarial que estão acima do limite superior. E quando a gente fez esse loc, de localizar eles, foram exatamente três valores, que são valores de 470 mil ou 358 mil.

00;07;11;19 - 00;07;40;11
Engraçado, né, gente, que parece até mesmo um erro de digitação. Às vezes a pessoa colocou um zero a mais ali. Às vezes, o salário era 47 mil ou 35 mil. Enfim, só teorizando aqui. A gente encontrou esses valores e agora a gente quer atribuir a eles a média 30 a 40. Então a gente coloca o igual, copia lá a média do 30 a 40 e coloca neles. E executamos a célula. E assim a gente substituiu.

00;07;40;25 - 00;08;06;07
Se a gente copiar essa célula ali em cima que a gente está vendo, o value counts da quantidade dos valores de salário acima do limite superior, e executar, esses três números aqui têm que sumir, tem que ficar só esse 19 do acima de 40 mil. Vamos executar. Perfeito, ficaram os 19 acima do 40 mil. E é o seguinte: fiz um, o outro é por conta de vocês aí.

00;08;06;08 - 00;08;35;05
Vou estar aqui esperando vocês voltarem, tá? Ok, voltando aqui, vocês fizeram aí a média de 40, salários acima de 40, que deu 53.027. Deu isso, né? Aí vocês com certeza atribuíram os valores aqui bonitinho. E aí, quando vão lá fazer o value counts, eu tenho certeza que deu isso aqui vazio porque a gente não tem mais outlier, correto?

00;08;36;00 - 00;09;02;10
Ótimo. Arrasaram. Vamos agora fazer o boxplot de novo pra ver como é que ele fica. Então vamos lá: plt ponto boxplot e aqui dentro nossa tabela dados e a nossa coluna de salário. E aí a gente executa. E olha só como ficou. Dá pra ver a caixinha um pouquinho melhor e agora ela está maior e tal.

00;09;02;18 - 00;09;25;23
Lembrando que a caixa é onde se concentra a maior quantidade de valores e a linha laranja aqui é a mediana dos valores. A gente pode ver uns pontinhos aqui acima e tudo mais, porém esses pontinhos estão abaixo do nosso limite superior, na verdade. Se a gente olhar aqui: 60 mil 60 e pouco, a gente já calculou que estão abaixo do nosso limite superior.

00;09;25;23 - 00;09;45;27
Tudo certo. E pela faixa salarial também, que a gente já deu uma olhada, a gente sabe que os salários podem ir até um pouco mais de 40 mil. Então os pontos, nesse caso, são valores que distoam um pouco dos valores gerais que estão na caixinha, mas entendendo o contexto, nós podemos concluir que não é um valor discrepante. Viram de novo a importância de entender o contexto?

00;09;46;16 - 00;10;14;06
Beleza então. Lembre-se sempre de que a abordagem para lidar com os outliers deve ser escolhida com base na natureza dos dados, nos objetivos da análise e no contexto específico do problema. Não há uma solução única que seja adequada para todos os casos e é importante avaliar as diferentes opções com cuidado. Bom, é isso, pessoal. Nós discutimos sobre valores discrepantes nessa aula e tratamos os valores discrepantes da coluna de salário. É isso e até a próxima aula.