4.3 - Tratando valores discrepantes - outliers (parte 1)

00;00;00;06 - 00;00;22;24
Olá pessoal! Bem-vindes a mais uma aula. Nessa aula nós vamos falar sobre valores discrepantes, também conhecidos pelo termo em inglês “outliers”. O que são esses valores discrepantes? A gente já falou um pouquinho deles quando comentamos sobre média, mas vale relembrar. Lembram da aula em que aprendemos sobre média? Nós começamos usando uma lista de idades com menos números para entender melhor o que era a média.

00;00;23;10 - 00;00;42;27
A lista era essa daqui. Aí nós executamos a média. Vamos executar aqui para a gente poder ver. Antes de tudo, vamos fazer o quê? Criar uma célula de texto, colocar nossos três joguinhos da velha aqui e colocar “valores discrepantes” outliers aqui entre parênteses.

00;00;44;29 - 00;01;24;10
Para a gente saber que começou uma nova aula, um novo tema. Ok, agora eu vou copiar aqui minha lista e vou colocar aqui... a gente tinha essa lista na aula de média. E aí a gente calculou a média, né? NP, numpy, min, lista idades. E executamos. A gente chegou à média de 29,8. Bom, agora imagina que a pessoa ali na hora de preencher a lista deu um escorregada aqui no zero, e ao invés de 40 digitou 400 nomes.

00;01;24;10 - 00;01;48;01
Vamos executar aqui a lista e vamos executar de novo e calcular de novo a média. Então a gente tinha uma média de 29.8 e com essa escorregada ali no zero a média foi para 62. Imagina, gente, a gente foi de uma média 29 anos para 62 anos. Muito diferente... Neste conjunto de dados, a maioria das idades parece normal: 23, 43 anos, enfim.

00;01;48;07 - 00;02;13;24
No entanto, tem uma ali que você destaca absurdamente, que é esse 400. Esse valor 400 é um exemplo de um valor discrepante ou outlier, é muito maior do que os outros valores e parece ser uma entrada incorreta. Nesse caso, alguém pode mesmo ter digitado esse zero a mais. Tratar os valores discrepantes é uma parte importante da análise de dados para garantir que nossas conclusões sejam mais precisas e robustas.

00;02;14;08 - 00;02;36;10
Nesse exemplo mesmo, imagina se quiséssemos preparar um conteúdo especial para esse grupo de pessoas aí achando que a idade média era 62 anos e gente fizesse algo aí para idosos, por exemplo. Não ia ser legal, porque a média real é 29 anos. Antes de lidar com os outliers, é crucial a gente identificar. Isso pode ser feito utilizando métodos estatísticos, como a regra do desvio padrão.

00;02;36;23 - 00;03;11;22
Lembram dela? Quanto mais próximo de zero, quer dizer que nossos dados estão mais homogêneos. Vamos calcular aqui qual o desvio padrão dessa lista de idades aqui com esse 400. NP (de numpy), std (do desvio padrão) e aí a lista de idades. Veja bem, deu 106, é um valor bem alto, ainda mais considerando que a nossa lista de idade tem poucos valores. Portanto, a chance de termos outliers é gigante. Existem cálculos para confirmar isso.

00;03;11;23 - 00;03;51;11
Um deles é usando a média de desvio padrão. É bem simples: a gente pega a média. Vamos atribuir esse valor aqui a uma média... Média, para ficar mais fácil... Então aqui "média", só pra gente poder continuar tendo resultado e esse de baixo aqui para desvio. Aí a gente coloca aqui embaixo também, só pra gente poder visualizar. Beleza. A gente pega a média, somando três vezes o desvio padrão. Então: média mais três vezes desvio. "Mais" de soma mesmo e essa estrelinha, que é de multiplicação.

00;03;51;27 - 00;04;17;05
Ok? Bom, então média mais três vezes o desvio. A gente tem 383 e aí é o nosso limite superior. Acima de 383, é tudo outlier. E a gente precisa achar o nosso limite inferior. Aqui a gente pode copiar aqui, colocar numa célula de código embaixo e, ao invés de somar, a gente vai subtrair. E aqui o nosso limite inferior é -258.

00;04;17;05 - 00;04;42;08
Então tudo abaixo disso é outlier. A gente tem o limite superior e inferior. É bem importante entendermos o contexto para saber se os limites calculados fazem sentido, então precisamos de algo mais conservador. Para isso, vamos usar a mediana e aprender um conceito novo, chamado de quartis. Como o próprio nome mesmo diz, a gente divide o dado em quatro partes, usando três valores.

00;04;42;13 - 00;05;03;25
É bem simples de fazer, muito parecido com o que fizemos com a mediana, que, aliás, é o nosso quartil do meio, está pegando o meio dos dados. Então, o primeiro quartil é a metade de baixo da mediana e o terceiro quartil é a metade de cima. Existe um modo visual da gente enxergar isso chamado boxplot. Vamos visualizar pra gente poder entender melhor?

00;05;04;21 - 00;05;39;29
Bom, pra isso a gente precisa de mais uma biblioteca, que é o Matplotlib, que é uma biblioteca com vários recursos de visualização de dados para plotagem de gráficos 2D e 3D. Nós conseguimos importar essa biblioteca assim: import matplotlib ponto pyplot. Essa biblioteca é mais longuinha, é mais chatinha de fazer o import. E aí a gente vai dar um apelido pra ela também, porque imagina toda vez que a gente for usar ter que colocar "matplotlib"? Então vai ser "as plt", ok?

00;05;40;07 - 00;06;11;21
Então toda vez que a gente for usar ela, a gente vai só colocar ali o “plt”. Vamos executar. Vamos plotar o nosso boxplot para visualizar os nossos quartis. A gente plota dessa forma: plt ponto boxplot. Opa, escrevi tudo errado, é: plt ponto boxplot, parênteses, ok? E aí dentro a gente vai colocar a nossa lista idades e vamos rodar.

00;06;13;14 - 00;06;39;04
Bom, aqui a gente tem o nosso gráfico. Deixa eu diminuir um pouquinho pra gente poder conseguir ver o gráfico melhor. Bom, a linha laranja no meio dessa caixinha aqui embaixo é a nossa mediana, as bordas da nossa caixinha (que está dando pra ver muito bem), mas as bordas é o nosso quartil Q1 é o Q3 e as linhas abaixo (que é essa linha aqui embaixo) é que define o nosso limite inferior.

00;06;39;04 - 00;07;06;21
E essa linha aqui de cima é o que define o nosso limite superior. Tudo, além dessas linhas, são os nossos outliers. Bem aqui em cima a gente vê essa bolinha aqui bem longe da nossa linha de limite superior. Ou seja, é um outlier, é o nosso outlier de 400, inclusive, que é bem nítido. Legal, né? Eu acho que é bem fácil de a gente visualizar os outliers com esse boxsplot. Vamos dar uma olhada se existem valores discrepantes na coluna de salário lá da nossa tabela?

00;07;07;05 - 00;07;35;15
Bom, vamos lá. Vou colocar assim: plt ponto boxplot e a gente abre parênteses. E aí a gente vai colocar: dados (que é a nossa tabela), abre colchetes e aí colocamos qual é a nossa coluna, que é de salário que a gente está querendo visualizar. Vamos executar. E olha só o que apareceu aqui: com esse gráfico, nós conseguimos ter uma noção rápida da dispersão dos salários.

00;07;35;28 - 00;08;03;04
Aqui no eixo Y a gente tem os valor de salários aqui e conseguimos ver que a caixa está abaixo dos 50 mil mais ou menos. Imagine que é 100 mil, a metade é 50 mil. A caixinha está abaixo disso. E aí a gente tem alguns valores que estão lá em cima, indicando, sei lá, 400 mil, 350 mil, vários outliers. A gente até queria ser um desses outliers, ganhando 400 mil.

00;08;04;06 - 00;08;30;02
Bom, estamos vendo visualmente: parece que os nossos outliers estão todos aí, acima de 100 mil, mas vamos aprender agora a calcular esses limites certinho: o limite superior e inferior para a nossa coluna de salário. Pra isso, vamos usar uma função do Pandas chamada de quartile, em inglês quartil, que nos retornará os nossos quartis. O primeiro quartil pega o valor que separa 1/4 dos dados.

00;08;30;16 - 00;09;08;02
Então chamamos ele passando o parâmetro de 0.25, que é 1/4 de 100, de 1. Então vamos colocar mais uma célula de código e aqui vamos colocar Q1 é igual a dados. A coluna é salário ponto e a função é quantile. Assim, ok? E aí a gente abre parênteses e coloca lá os 0.25.

00;09;09;06 - 00;09;40;03
Uma coisa importante aqui é que no Python, para ter valores decimais ou com ponto flutuante, a gente usa o ponto mesmo, não a vírgula, tá gente? Então, a gente coloca 0.25, vamos colocar o Q1 aqui embaixo para gente poder visualizar? E vamos executar. O nosso primeiro quartil é 400751. Agora vamos calcular o terceiro quartil. O terceiro quartil separa 3/4 dos dados.

00;09;40;13 - 00;10;21;09
Então chamamos ele passando o parâmetro de 0.75: 0.25 o primeiro quartil, 0.75 o terceiro quartil. Ok, vamos copiar aqui em cima e colar embaixo para facilitar pra gente? E agora a gente vai calcular o Q3, o terceiro quartil. Aqui embaixo o Q3, que não é 0.25, é 0.75. Vamos executar e temos aqui 11794. Aqui vamos calcular um valor chamado de interquartil, que define o valor que tem entre o primeiro e o terceiro quartil.

00;10;21;10 - 00;10;54;25
Interquartil vamos chamar de IQR. E aí, como é entre um quartil e outro, a gente pode fazer um menos outro. Então é: Q3 menos Q1. Vamos colocar aqui embaixo o IQR pra gente poder saber o valor: 7043. E por fim, para pegar os nossos limites, que é isso que a gente está buscando, a gente vai definir que o limite superior é três Q, três quartis, mais 1.5 vezes o interquartil.

00;10;54;28 - 00;11;39;07
E o inferior é um Q, terceiro quartil, menos 1.5 vezes o interquartil. E são esses limites que os tracinhos do boxplot mostram. Vamos calcular isso. Limite superior, vamos colocar assim: lim superior é terceiro quartil, que é o Q3 mais... A gente abre parênteses porque a multiplicação tem que ser feita antes da soma, 1.5 vezes (que é aquela estrelinha) o interquartil, que é IQR. Esse é o nosso limite superior, que é aquela barrinha lá em cima do boxplot, que é 22359.

00;11;39;23 - 00;12;19;22
E o limite inferior, vou colocar lim underline inferior é igual ao Q1 menos, abre parênteses, 1.5 vezes IQR. E vamos colocar aqui embaixo o lim inferior para a gente poder saber quanto é: -5813. Pronto! Agora sabemos que todos que estão com salário acima de 22359 estão acima do esperado. Como estamos lidando com a coluna de salário, podemos considerar o limite inferior como o máximo entre o limite inferior e zero, porque não tem salário negativo, né?

00;12;19;28 - 00;12;41;21
Pelo menos eu espero que não tenha. Ninguém está recebendo salário negativo. Bom, se a gente olhar na coluna de faixa salarial, a gente vê que tem salários que vão aí até os 40 mil. Vamos dar uma olhada: na tabela de dados, a coluna faixa salarial, ponto value counts, que assim a gente vai conseguir saber a quantidade de pessoas por cada faixa salarial.

00;12;43;00 - 00;13;17;08
Parênteses vazios. Opa, não é faixa salário, é faixa salarial. E olha, a gente tem aí as faixas salariais de 8 a 12 mil. Mas a gente também tem pessoas que estão ganhando de 25 a 30 mil, de 30 mil a 40 mil. Porém, pelo gráfico que a gente gerou, conseguimos visualizar que esses salários aqui no gráfico de 300 mil são realmente fora do contexto que a gente está analisando.

00;13;17;27 - 00;13;40;25
E o nosso limite calculado usando esses métodos de quartil deu um valor de 22 mil, que, olhando contexto, é um valor baixo, porque a gente tem as pessoas de 30 a 35. Nesse caso, não tem como a gente dizer que todos os salários acima de 22 mil são outliers. Existe um outro método de cálculo desse limite que considera que todo valor que está a uma certa distância da média é outlier.

00;13;42;14 - 00;14;14;14
Mas qual é essa distância? Essa distância é calculada considerando o desvio padrão. Lembra que o desvio padrão é uma medida de dispersão dos dados? Então faz sentido, já que o outlier é um dado completamente disperso. Vamos primeiro calcular aqui a média e o desvio padrão da coluna de salário. Vamos lá na nossa célula de código: média salário igual a dados (que é nossa tabela), nossa coluna salário ponto min.

00;14;15;03 - 00;14;55;27
Ok. Média salário só para a gente poder visualizar. Vamos executar. A média de salário está 10517. Ok. E vamos calcular o desvio padrão que é dado... Vamos colocar um nome: desvio salário igual a dados, salário (coluna salário) ponto std e os parênteses vazios. Vamos colocar aqui embaixo também desvio salário para a gente poder visualizar. E aí a gente tem um desvio padrão de salário de 18096.

00;14;56;26 - 00;15;20;23
Bom, pra calcular o limite superior nesse caso nós usamos uma fórmula que é a média mais um certo número de desvios padrões. Esse número pode variar dependendo da sensibilidade da análise e da distribuição dos dados. Pode ser média mais um desvio padrão, pode ser média mais dois desvios padrão e assim vai. O que determina isso é a nossa análise de contexto.

00;15;21;04 - 00;15;58;09
Como aqui nós vimos que os outliers estão bem acima dos valores comuns de salário, nós podemos tentar usar a média mais três desvios padrão. Assim: limite superior (para ficar diferente do que a gente tinha calculado de lim superior), limite underline superior igual a média salário, mais, abre parênteses (porque a gente vai fazer uma multiplicação agora) três vezes o desvio do salário, que é o desvio padrão de salário. E vamos dar uma visualizada aqui também nesse valor.

00;15;59;21 - 00;16;26;25
Ok: 64806. E olhando o nosso gráfico do boxplot, parece ser um valor muito bom para o limite superior. Então, todos os valores acima de 64806 são considerados outliers. Realmente temos muito valores altos de salário. Será que faz sentido essas pessoas terem esse salário? Ou foi erro de preenchimento de formulário?