Análise Exploratória de Dados
Segunda chamada da avaliação presencial - Gabarito - 2023/01

Prof. Hugo Carvalho

04/07/2023

Questão 1:

a) A fórmula para calcular x_{MIQ} é uma média aritmética referente ao conjunto de dados x, porém não é uma média que leva em consideração o conjunto de dados como um todo: uma parcela do começo e uma parcela do final, após ordenação, são removidas. Tais parcelas são exatamente as menores 25% e as maiores 25% observações, mantendo somente os 50% centrais, entre o primeiro e o terceiro quartil. Comparando com a média $\overline{x}$, vê-se que x_{MIQ} tenta remover observações discrepantes para obter uma (possivelmente) melhor medida de centralidade do conjunto de dados x.

b) Uma possível vantagem de x_{MIQ} é que, devido à remoção de observações potencialmente discrepantes, dará uma melhor noção de centralidade do conjunto de dados x. Caso tal conjunto de dados não tenha observações discrepantes e seja aproximadamente simétrico, teremos $x_{\text{MIQ}} \approx \overline{x}$, sem muita perda de informação. Um ponto negativo dessa nova métrica é que ela descarta metade do conjunto de dados, sendo portanto, menos confiável, a não ser que estivéssemos em um cenário de um conjunto de dados muito grande. Além disso, a remoção de observações pode levar a conclusões errôneas da base de dados. Dessa forma, talvez não seja recomendável usar somente x_{MIQ} mas também $\overline{x}$, comparando ambas as métricas.

c) Veja se o conjunto de dados que a pessoa criou é algo que faz sentido e que justifica algo que ela falou no item b)! Não coloquei gabarito pois é algo beeeeeeem livre.

d) Pouco sensível a valores discrepantes, justamente por descartá-los em seu cálculo.

e) Nesse caso, pouca coisa pode ser dita além de que a metade central do conjunto de dados x (ou seja, aquela parte compreendida entre o primeiro e o terceiro quartil) têm média zero, e isso pode acontecer de diversas maneiras possíveis. Sobre a parte descartada, nada pode ser dito, pois essa parcela nem entra na conta.

Questão 2: Também não vou colocar gabarito por ser bem livre, mas alguns pontos importantes que a pessoa pode tocar:

- A figura é condizente com um quadro depressivo, dada a alta incidência de dias com sentimentos negativos e a menor incidência de dias com sentimentos positivos.
- Algumas variáveis apresentam muita variabilidade (ansioso p. ex.) enquanto que outras são bem mais concentradas (deprimido, p. ex.).
- Apesar do quadro geral, há picos de felicidade e satisfação, porém são outliers.
- O fato de "Feliz" e "Satisfeito" não ter outliers para baixo é preocupante pois indica que estar no nível mínimo nessas variáveis não é raro.
- O gráfico tem algumas coisas esquisitas, como outliers aparentemente coincidindo com o 3o quartil?!
- Em alguns casos ("Deprimido" e "Ansioso") não dá para identificar a mediana corretamente. Mais especificamente, não fica claro nem se ela coincide com o primeiro ou o terceiro quartil ou se houve um arredondamento!
- A falta de informações não-inteiras dá a entender que houve um arredondamento, potencialmente desnecessário.