Fundamentos de Mineração de Dados

Exercício teórico sobre análise de dados

1- Dado o conjunto de dados {1, 2, 3, 4, 5, 80}, calcular:

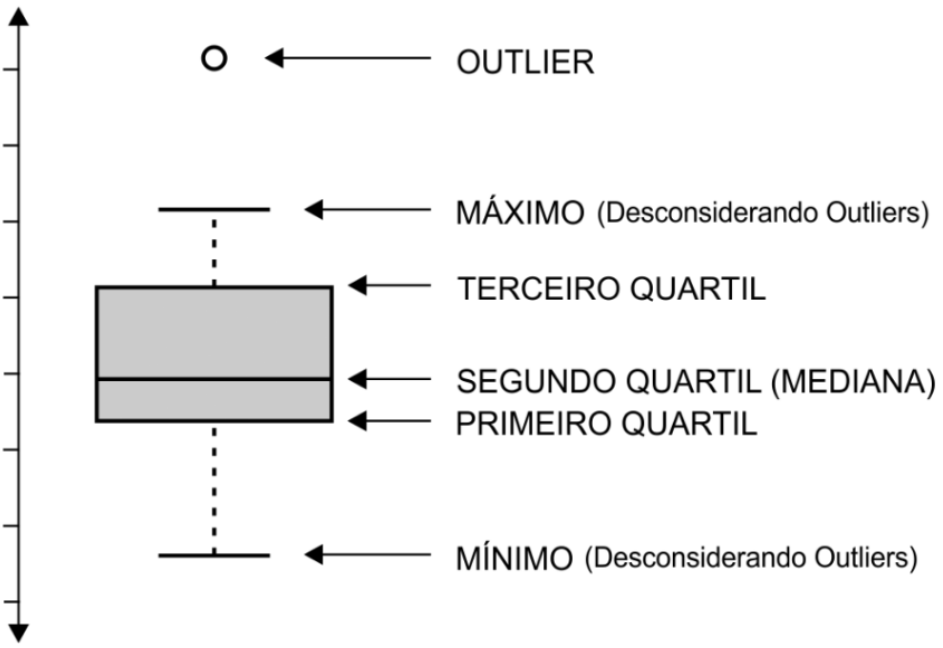
Média:

x_media = (1 + 2 + 3 +4 +5 +80)/6 = 15,83

Mediana:

x_mediana = (3+4)/2 = 3,5

2- Comente sobre o uso de um box plot para explorar um conjunto de dados com quatro atributos: idade, peso, altura e renda.



O boxplot ou diagrama de caixa é uma ferramenta gráfica que possibilita visualizar a distribuição e valores discrepantes (outliers) dos dados em questão, fornecendo assim um meio complementar para desenvolver uma perspectiva sobre a característica desses dados. Além disso, o boxplot também é uma disposição gráfica comparativa.

As medidas de estatísticas descritivas como mínimo, máximo, primeiro quartil, segundo quartil ou mediana e o terceiro quartil formam o boxplot.

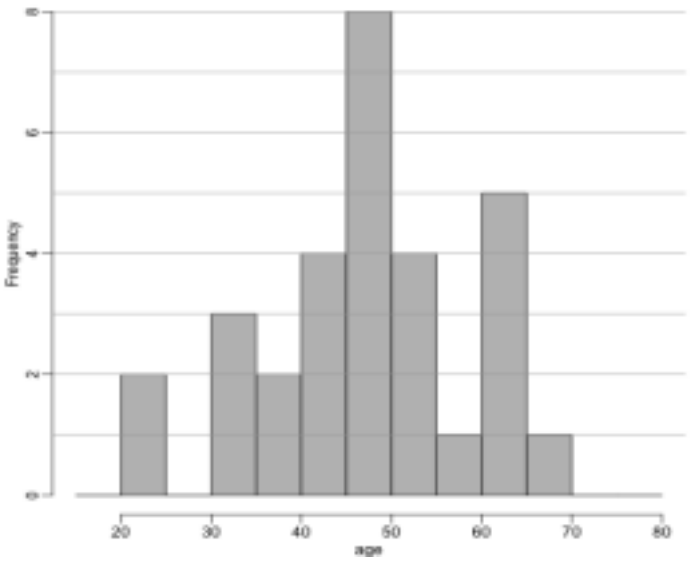
O boxplot nos fornece uma análise visual da posição, dispersão, simetria, caudas e valores discrepantes (outliers) do conjunto de dados.

- Posição – Em relação à posição dos dados, observa-se a linha central do retângulo (a mediana ou segundo quartil).
- Dispersão – A dispersão dos dados pode ser representada pelo intervalo interquartílico que é a diferença entre o terceiro quartil e o primeiro quartil (tamanho da caixa), ou ainda pela amplitude que é calculada da seguinte maneira: valor máximo – valor mínimo. Embora a amplitude seja de fácil entendimento, o intervalo interquartílico é uma estatística mais robusta para medir variabilidade uma vez que não sofre influência de outliers.
- Simetria – Um conjunto de dados que tem uma distribuição simétrica, terá a linha da mediana no centro do retângulo. Quando a linha da mediana está próxima ao primeiro quartil, os dados são assimétricos positivos e quando a posição da linha da mediana é próxima ao terceiro quartil, os dados são assimétricos negativos. Vale ressaltar que a mediana é a medida de tendência central mais indicada quando os dados possuem distribuição assimétrica, uma vez que a média aritmética é influenciada pelos valores extremos.
- Caudas – As linhas que vão do retângulo até aos outliers podem fornecer o comprimento das caudas da distribuição.
- Outliers – Já os outliers indicam possíveis valores discrepantes. No boxplot, as observações são consideradas outliers quando estão abaixo ou acima do limite de detecção de outliers.

Sendo assim, é possível analisar as características dos atributos: idade, peso, altura e renda. Comparando a dispersão dessas variáveis, as principais características desses conjuntos de dados, bem como a qualidade desse conjunto de dados. Além de comparar

a equivalência entre grupos amostrais, tanto para estudo transversal como longitudinal.

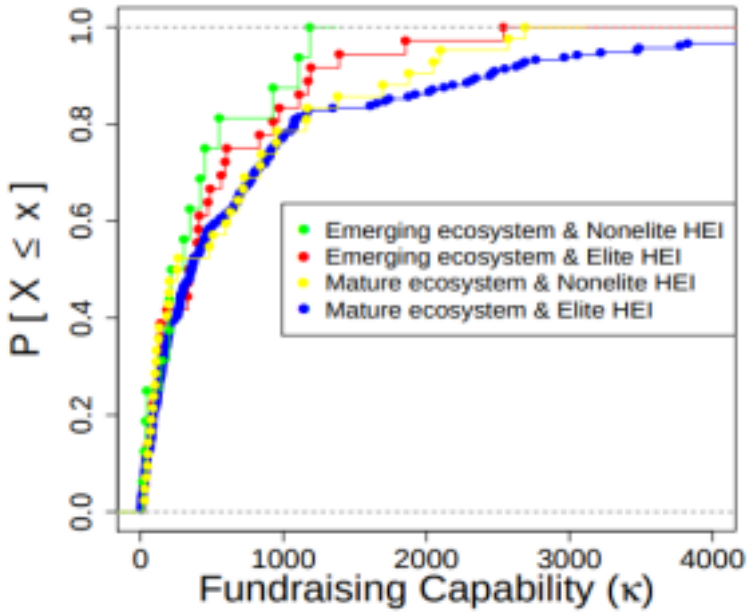
3- Considere o histograma abaixo, e responda quantas pessoas possuem entre 35 e 45 anos.



num_total = área do gráfico:

[(40-35)*0,2 + (45-40)*0,4]/10 = (1 + 2)/10 = 3/10 = 0,3 = 30%

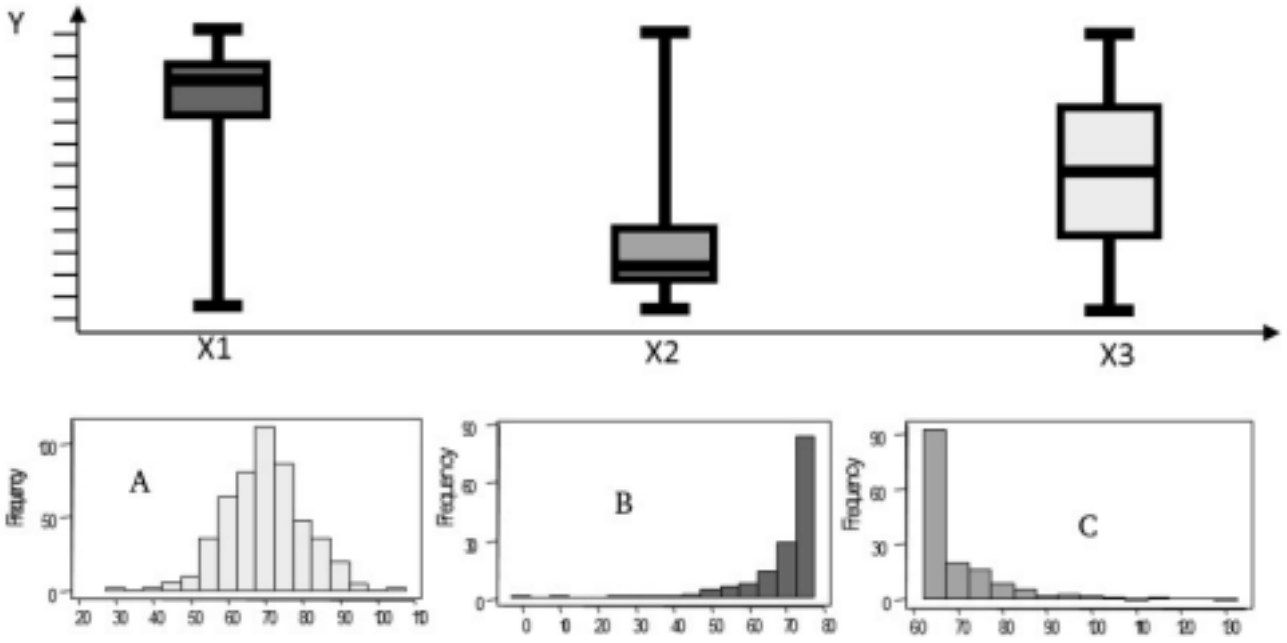
4- A CDF (cumulative distribution function) abaixo, refere-se à capacidade de arrecadação de fundos (κ) de startups do Brasil de acordo com a sua classe. Qual a probabilidade aproximada de startups do grupo "Emerging ecosystem & Elite HEI" levantarem κ mais do que 2000? E as do grupo "Mature ecosystem & Elite HEI"? (Mais informações em: https://arxiv.org/pdf/1904.12026.pdf)



A probabilidade aproximada de startups do grupo "Emerging ecosystem & Elite HEI" levantarem κ mais do que 2000 é de 5% ou menos.

Enquanto que a do grupo "Mature ecosystem & Elite HEI" é de menos do que 20%.

5- Associe cada um dos boxplots abaixo (X1, X2, X3) com o histograma mais provável (A, B, C).



A = X3, porque temos uma distribuição gaussiana, ou normal, em que que os valores se encontram distribuídos homogeneamente em torno da média

B = x1, porque tem-se uma distribuição muito grande de valores abaixo da mediana, com isso fazendo com que o menor valor de Y seja bem abaixo. Esse é um tipo de distribuição de probabilidade enviesada negativamente (negative skewed)

C = x2, porque tem-se uma distribuição muito grande de valores acima da mediana, com isso fazendo com que o maior valor de Y seja bem alto. Esse é um tipo de distribuição de probabilidade enviesada positivamente (positive skewed)