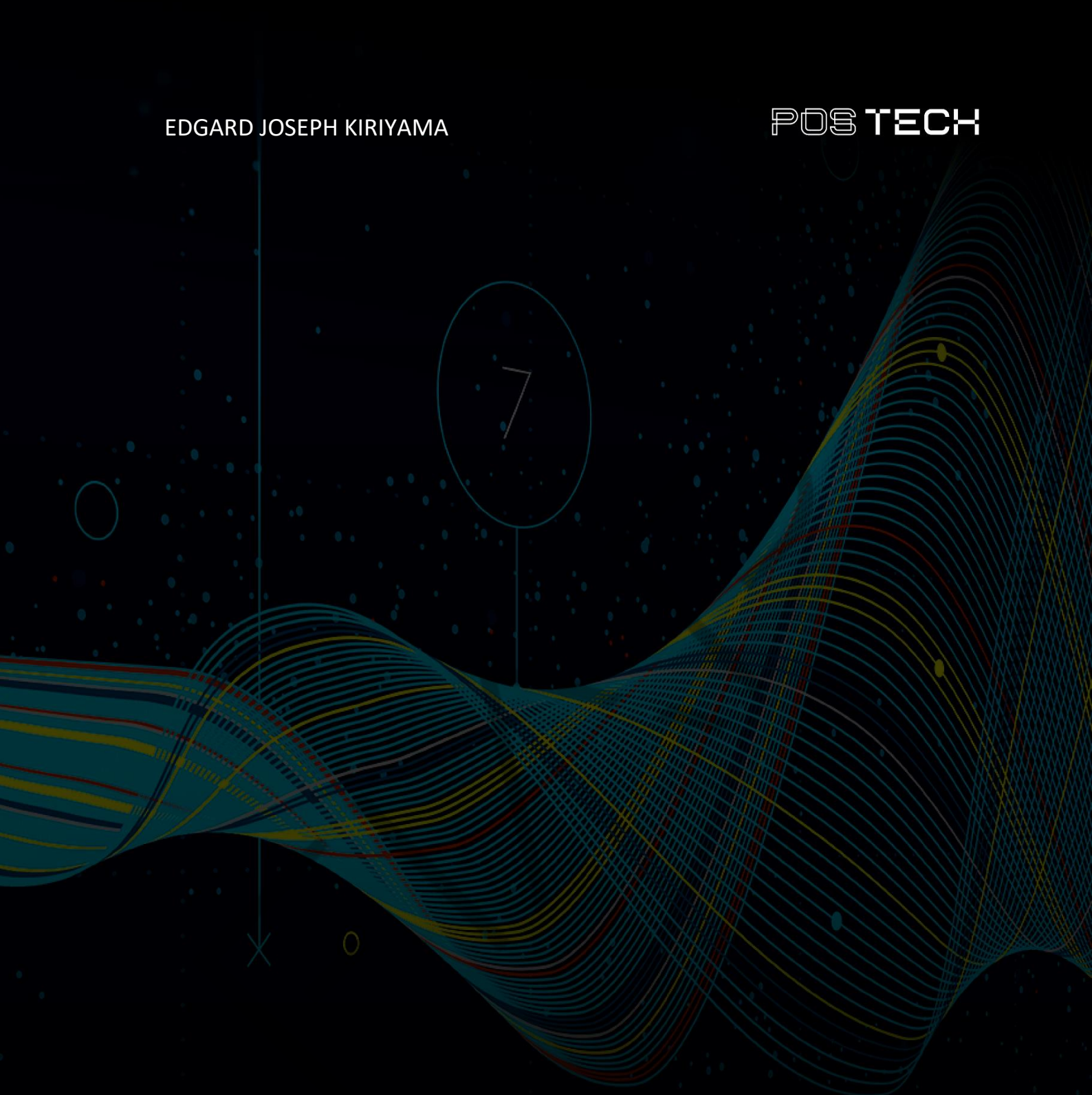
EDGARD JOSEPH KIRIYAMA

PÓS TECH



DATA ANALYTICS

MACHINE LEARNING COM PYTHON

AULA 01

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

SUMÁRIO

O QUE VEM POR AÍ? 3
CONHEÇA SOBRE O ASSUNTO 4
HANDS ON 22
O QUE VOCÊ VIU NESTA AULA? 23
REFERÊNCIAS.....24
PALAVRAS-CHAVE.....25

FIAP

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

O QUE VEM POR AÍ?

Olá, jovem analytic!

Você está na terceira disciplina do curso e aqui começa mais um ponto importantíssimo da jornada em que você está trilhando.

Nesta aula iremos trabalhar com a introdução de conceitos básicos em estatística e álgebra linear.

Este é o momento de aprimorarmos e elevarmos o nosso nível! Vamos te mostrar como a estatística e a álgebra linear são aplicadas na área de dados, a sua importância na avaliação no conjunto de dados e como essas ferramentas ajudam na construção de modelos de Machine Learning.

Para essa aula, estamos disponibilizando a base de dados no GitHub.

Vamos entender como este aprendizado pode ajudar no nosso dia a dia!

FIAP

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

CONHEÇA SOBRE O ASSUNTO

Estatística Básica

Estatística é uma área da matemática que se dedica a coletar, analisar e interpretar dados. Na análise de dados, a estatística é uma ferramenta fundamental para extrair informações valiosas a partir dos dados disponíveis.

Esta disciplina é usada para descrever e resumir dados, identificar padrões e tendências, testar hipóteses e ajudar a realizar previsões. As técnicas estatísticas incluem a análise descritiva, que envolve a apresentação e descrição dos dados em gráficos e tabelas, e a inferência estatística, que permite fazer generalizações sobre uma população com base em uma amostra de dados.

A estatística é amplamente utilizada em diversas áreas, como negócios, ciência, engenharia, saúde e muitas outras. Com a crescente disponibilidade de dados e o aumento da importância da análise de dados para a tomada de decisões, a estatística se tornou uma habilidade fundamental para profissionais em muitas áreas.

Na nossa área, de análise de dados, é importante ter conhecimentos sólidos em estatística para escolher as melhores técnicas e interpretar corretamente os resultados. Uma análise de dados mal feita pode levar a conclusões equivocadas e decisões erradas, enquanto uma análise bem feita pode fornecer informações valiosas para a solução de problemas e a tomada de decisões informadas.

Estatística Descritiva

A Estatística Descritiva é uma das principais áreas da Estatística e tem como objetivo descrever e resumir um conjunto de dados de maneira objetiva. Ela é utilizada para entender a distribuição dos dados, identificar padrões e tendências, assim como resumir as características mais importantes dos dados de interesse.

Na Estatística Descritiva, o primeiro passo é a coleta de dados. Em seguida, são realizados os cálculos das medidas de tendência central, que incluem a média, mediana e moda, e medidas de dispersão, como o desvio padrão e a variância. Essas medidas são utilizadas para entender como os dados estão distribuídos e identificar a presença de valores extremos.

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

Além das medidas de tendência central e dispersão, a Estatística Descritiva também utiliza gráficos para visualizar os dados. Os gráficos mais comuns são o histograma, o gráfico de barras e o diagrama de caixa (boxplot), que permitem identificar a forma da distribuição, a presença de valores atípicos e também comparar grupos de dados.

A Estatística Descritiva é amplamente utilizada em diversas áreas, tais como negócios, ciência, engenharia, TI e saúde. Ela permite que os profissionais analisem e entendam seus dados de maneira clara e objetiva, o que é fundamental para a tomada de decisões informadas.

É importante lembrar que a Estatística Descritiva se limita apenas a descrever e resumir os dados disponíveis, e não permite fazer generalizações ou inferências sobre uma população. Para isso, é necessário utilizar técnicas da Inferência Estatística.

População

A população, na estatística, é o conjunto completo de indivíduos, objetos, eventos ou medidas que se deseja estudar e sobre os quais se quer fazer inferências. A população é definida com base nos objetivos da pesquisa (ou estudo) e pode ser constituída por elementos físicos, como pessoas, animais, objetos ou por conceitos abstratos, como ideias ou valores.

A população é importante na Estatística, porque é a partir dela que se fazem inferências sobre características e propriedades dos elementos que a compõem. No entanto, é muitas vezes inviável ou impraticável estudar toda a população, especialmente quando esta é muito grande ou está espalhada por uma área geográfica muito ampla. Nesses casos, são selecionadas amostras da população para serem estudadas, utilizando técnicas de amostragem adequadas. As inferências estatísticas feitas com base em amostras podem ser extrapoladas para a população como um todo, desde que a amostra seja representativa da população em questão.

Escala de Mensuração

A Escala de Mensuração é uma classificação utilizada na Estatística para descrever os diferentes tipos de dados que podem ser coletados em uma pesquisa ou

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

estudo. Essa classificação é importante porque as técnicas estatísticas apropriadas para a análise dos dados dependem do tipo de escala de mensuração utilizada.

Existem quatro tipos principais de escala de mensuração: nominal, ordinal, intervalar e de razão. A escala nominal é a mais simples e é usada para classificar dados em categorias ou classes sem uma ordem específica, como cor dos olhos ou sexo dos indivíduos. A escala ordinal também é usada para classificar dados em categorias, mas nesse caso há uma ordem específica, como notas de uma prova ou classificação de equipes em um campeonato.

A escala intervalar é usada para medir dados em uma escala numérica, onde as diferenças entre os valores são iguais, mas não há um valor zero absoluto, como temperatura em graus Celsius. Já a escala de razão é semelhante à escala intervalar, mas inclui um valor zero absoluto, como altura e peso, e permite realizar cálculos, como multiplicação e divisão entre as medidas.

A escolha da escala de mensuração adequada é importante para garantir que os dados sejam coletados e analisados da maneira correta. Uma escala inadequada pode levar a análises incorretas e conclusões equivocadas. Por isso, é importante definir a escala de mensuração adequada desde o início da coleta de dados em uma pesquisa ou estudo.

Medidas Resumo de Tendência

As medidas resumo de tendência central são utilizadas na Estatística para resumir e descrever a localização central dos dados em um conjunto de observações. Elas fornecem um valor único que representa a tendência central dos dados e ajudam a entender a distribuição dos dados.

As medidas de tendência central mais comuns são a média, a mediana e a moda. A média é calculada somando-se todos os valores do conjunto de dados e dividindo pelo número total de observações. Ela é afetada por valores extremos ou discrepantes no conjunto de dados. A mediana é o valor que divide o conjunto de dados ao meio, onde metade dos valores são menores e metade são maiores. Ela é menos afetada por valores extremos do que a média. A moda é o valor que aparece com mais frequência no conjunto de dados.

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

Média

A média é definida como a soma de todos os valores em um conjunto de dados, dividida pelo número total de observações. Ela é uma medida que representa a tendência central dos dados, pois indica o valor que os dados tendem a se concentrar em torno. Ela é comumente utilizada em estudos estatísticos para resumir e descrever características de um conjunto de dados.

Vale ressaltar que a média é uma medida sensível a valores extremos (outliers), ou seja, valores que são muito diferentes dos demais no conjunto de dados podem influenciar bastante a média. Por essa razão, é importante avaliar se o conjunto de dados tem valores extremos antes de usar a média como medida de tendência central.

Mediana

A mediana representa o valor central de um conjunto de dados, ou seja, o valor que divide os dados em duas partes iguais. Para calculá-la, é necessário ordenar os dados do menor para o maior e, em seguida, selecionar o valor que está no meio da lista. Se o conjunto de dados tiver um número par de elementos, a mediana é a média aritmética dos dois valores centrais.

Ela é uma medida robusta para descrever a tendência central de um conjunto de dados, pois ela não é afetada por valores extremos (outliers).

A mediana é frequentemente utilizada em distribuições com assimetria, onde a média não seria uma medida adequada para descrever a tendência central dos dados.

Moda

A moda representa o valor que ocorre com maior frequência em um conjunto de dados. Em outras palavras, é o valor mais comum em uma distribuição. Para calcular a moda, é necessário identificar qual valor ocorre com mais frequência no conjunto de dados. Pode haver mais de uma moda em uma distribuição, por exemplo, se dois valores diferentes ocorrem com a mesma frequência máxima.

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

A moda é uma medida menos comum do que a média e a mediana, mas pode ser muito útil para descrever características de um conjunto de dados, principalmente quando há valores repetidos. A moda é uma medida robusta em relação a valores extremos, pois ela não é influenciada por valores muito diferentes dos demais.

Medidas Resumo de Variabilidade

As medidas resumo de variabilidade são utilizadas na Estatística para descrever a dispersão dos dados em torno de uma medida de tendência central, como a média, mediana ou moda. Essas medidas ajudam a entender a variação dos dados, quão próximos ou distantes eles estão uns dos outros.

As principais medidas resumo de variabilidade são:

- Amplitude: é a diferença entre o maior e o menor valor em um conjunto de dados. Ela indica a extensão total dos dados, mas é uma medida pouco precisa de variabilidade, pois é sensível a valores extremos.
- Desvio padrão: é uma medida que indica a dispersão dos dados em relação à média. Quanto maior o desvio padrão, maior é a variação dos dados em relação à média. Ele é muito utilizado em inferência estatística para avaliar a precisão de uma estimativa.
- Variância: é uma medida que indica o quão distantes os valores estão da média, ou seja, a variação dos dados. É calculada como a média dos quadrados das diferenças entre cada valor e a média. A variância é uma medida útil para avaliar a dispersão dos dados em torno da média.
- Quartis: são valores que dividem os dados em quatro partes iguais, cada uma contendo 25% dos dados. O primeiro quartil (Q1) é o valor que divide os 25% menores dados, o segundo quartil (Q2) é a mediana e o terceiro quartil (Q3) é o valor que divide os 25% maiores dados.

As medidas resumo de variabilidade são importantes para entender a dispersão dos dados e avaliar a precisão das estimativas. Elas são amplamente utilizadas em diversas áreas, como finanças, saúde, pesquisa de mercado, entre outras.

FIAP

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

Forma de Distribuição

A forma da distribuição é uma característica da distribuição de dados em um conjunto de observações. Ela descreve a aparência geral da distribuição, incluindo a simetria, o achatamento e a presença de valores extremos (outliers).

Existem diferentes tipos de formas de distribuição, sendo as principais:

- Simétrica: a distribuição é simétrica quando a metade esquerda da curva é igual à metade direita. Um exemplo de distribuição simétrica é a distribuição normal, também conhecida como curva em forma de sino.
- Assimétrica à direita: a distribuição é assimétrica à direita quando a cauda da curva se estende mais para o lado direito. Isso ocorre quando há valores extremamente altos que puxam a média para cima, enquanto a maioria dos dados está concentrada no lado esquerdo da distribuição.
- Assimétrica à esquerda: a distribuição é assimétrica à esquerda quando a cauda da curva se estende mais para o lado esquerdo. Isso ocorre quando há valores extremamente baixos que puxam a média para baixo, enquanto a maioria dos dados está concentrada no lado direito da distribuição.
- Bimodal: a distribuição é bimodal quando existem dois picos na curva, indicando que existem dois grupos diferentes de observações.

A forma da distribuição é uma importante característica dos dados, pois influencia a escolha das medidas estatísticas adequadas para descrever e analisar os dados. Além disso, a forma da distribuição pode indicar a presença de padrões e relações entre variáveis.

Detecção de Outliers

A detecção de outliers, ou valores extremos, é um processo importante na análise estatística para identificar valores que se afastam significativamente do padrão dos demais valores em um conjunto de dados. Os outliers podem ter um impacto significativo nas estatísticas descritivas, como média e desvio padrão, levando a resultados incorretos ou imprecisos.

Existem várias abordagens para detectar outliers, como:

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

- Análise gráfica: a visualização dos dados em gráficos, como histogramas, boxplots ou scatterplots, pode ajudar a identificar valores discrepantes que se afastam significativamente do padrão dos demais valores.
- Testes estatísticos: existem testes estatísticos, como o teste Z-score ou o teste de Dixon, que podem ser usados para identificar valores extremos com base em desvios da média ou em comparações entre valores.
- Métodos de modelagem: alguns modelos estatísticos, como a regressão linear, são sensíveis a outliers. A detecção e remoção de outliers pode melhorar a precisão do modelo e as inferências obtidas a partir dele.

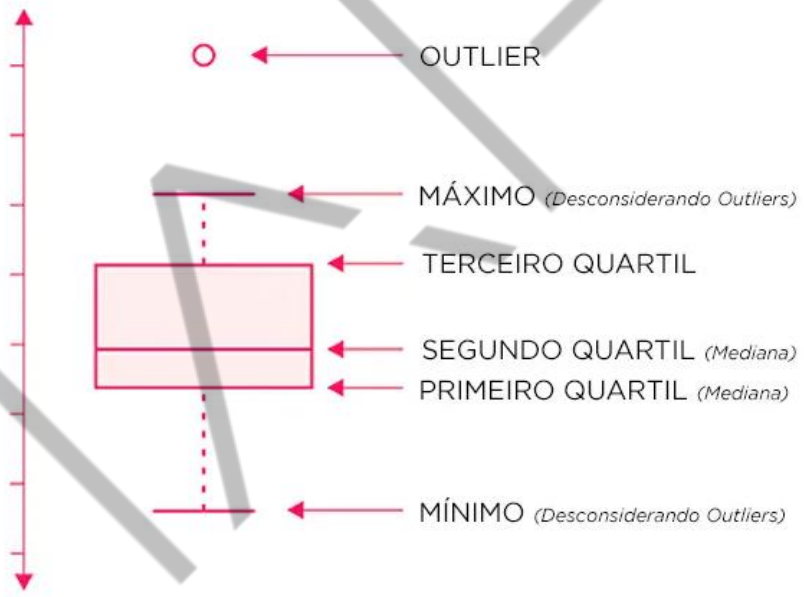


Figura 1 – Representação de um Box-plot (gráfico de caixa)
Fonte: Estat site (2018), adaptado por FIAP (2023)

É importante lembrar que a detecção de outliers não é uma tarefa fácil e depende do contexto em que os dados são coletados. Valores que parecem extremos em um conjunto de dados podem ser completamente justificáveis em outro. Por isso, é fundamental ter conhecimento do domínio do problema e avaliar cuidadosamente os resultados obtidos ao lidar com valores extremos.

Probabilidade

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

Probabilidade é uma medida numérica que indica a chance ou a probabilidade de um evento ocorrer. Na estatística, a probabilidade é usada para descrever e quantificar a incerteza associada a um evento aleatório.

Um evento aleatório é um evento cujo resultado não pode ser previsto com certeza, como o lançamento de um dado ou a escolha de uma carta de um baralho. A probabilidade de um evento é expressa como um número entre 0 e 1, sendo que 0 indica a impossibilidade de ocorrer e 1 indica a certeza de que ocorrerá.

Pode ser calculada usando diferentes abordagens, como a teoria clássica, a teoria da frequência relativa ou a teoria das probabilidades condicionais. Essas abordagens têm suas próprias definições e formulações matemáticas, mas todas elas se baseiam em contar o número de resultados possíveis e dividir pelo número total de resultados. Ela tem aplicações em diversas áreas da estatística, incluindo o cálculo de estimativas, testes de hipóteses, análise de séries temporais e modelagem estatística. É uma ferramenta importante para a tomada de decisões em ambientes incertos e para a análise de riscos em diversos contextos.

Distribuição de Poisson

A distribuição de Poisson é uma distribuição de probabilidade discreta que descreve a probabilidade de um número de eventos ocorrer em um intervalo de tempo ou espaço específico. Ela é amplamente usada em situações em que o evento ocorre aleatoriamente em uma taxa constante e independente de qualquer outro evento.

A distribuição de Poisson é caracterizada por um único parâmetro, lambda (λ), que representa a taxa média de ocorrência do evento. A partir desse parâmetro, é possível calcular a probabilidade de um número específico de eventos ocorrer em

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

um determinado intervalo de tempo ou espaço.

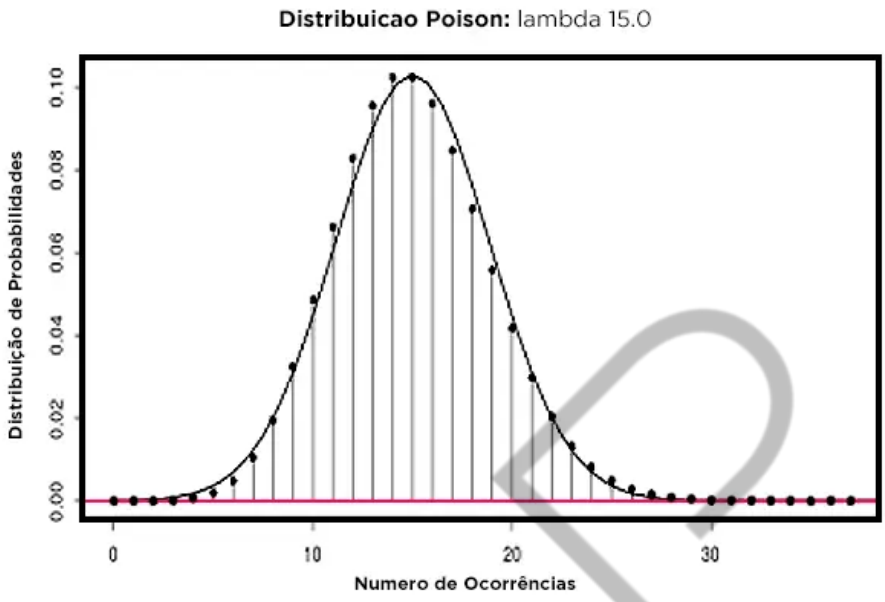


Figura 2 – Representação da distribuição de poisson
Fonte: Lampada UERJ (2022), adaptado por FIAP (2023)

Distribuição de Normal

A distribuição normal, também conhecida como distribuição gaussiana, é uma das distribuições de probabilidade mais importantes e amplamente utilizadas na estatística. Ela é caracterizada por uma curva simétrica em forma de sino, onde a maior parte dos dados está concentrada em torno da média, com a frequência de ocorrência diminuindo à medida que se afasta da média.

A distribuição normal é definida por dois parâmetros: a média (μ) e o desvio padrão (σ). A média é o valor central da distribuição e o desvio padrão é uma medida de dispersão que indica o quão longe os valores estão da média.

A distribuição normal é útil para modelar dados contínuos que são afetados por múltiplos fatores aleatórios. Ela é frequentemente usada em testes estatísticos, como testes de hipóteses e intervalos de confiança, e em análises de regressão, entre outras aplicações.

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

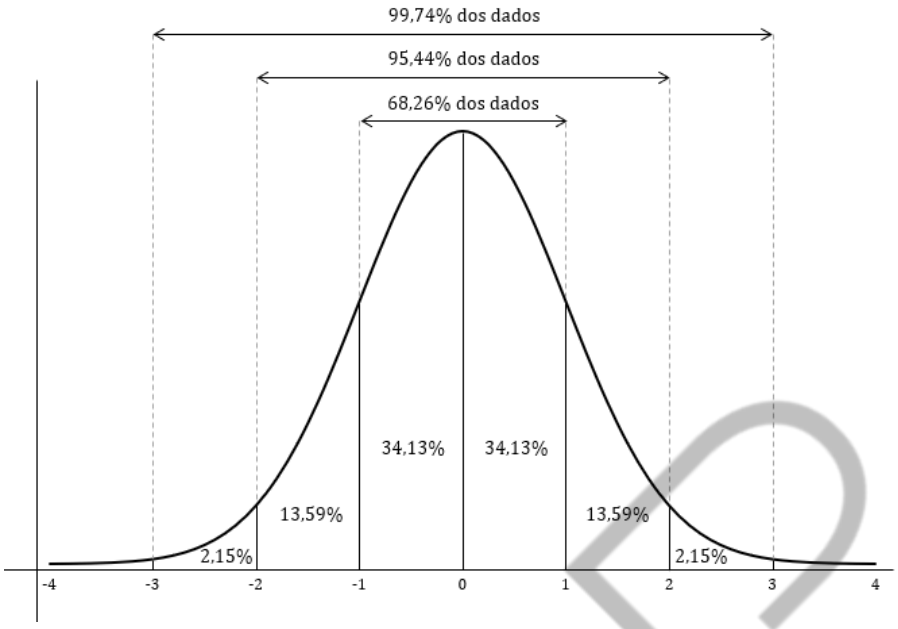


Figura 3 – Representação da Distribuição Normal
Fonte: Pro Educacional (s.d.)

Probabilidade Condicional

A probabilidade condicional é uma medida de probabilidade que leva em consideração a ocorrência de um evento em relação a outro evento que já ocorreu ou que se sabe que ocorreu. Ela é expressa pela probabilidade de um evento A ocorrer, dado que um evento B já ocorreu ou é conhecido.

A probabilidade condicional é representada matematicamente pela fórmula P(A|B), onde P é a probabilidade, A é o evento em questão e B é o evento conhecido. Essa fórmula é lida como "a probabilidade de A dado que B ocorreu".

Inferência Estatística

Inferência estatística é o processo de usar amostras de dados para fazer afirmações ou inferências sobre uma população maior a partir da qual os dados foram coletados. Ela envolve a análise de dados amostrais para estimar parâmetros desconhecidos da população e testar hipóteses sobre a população a partir da amostra.

A inferência estatística é importante, porque muitas vezes é impraticável ou impossível coletar dados de uma população inteira, especialmente quando a população é grande. Em vez disso, é mais viável coletar amostras dos dados e usar inferência estatística para extrair informações sobre a população a partir da amostra.

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

Os métodos de inferência estatística incluem a estimativa de parâmetros, testes de hipóteses e intervalos de confiança. A estimação de parâmetros envolve a estimativa de valores desconhecidos, como a média e a variância da população, com base na amostra. Os testes de hipóteses envolvem a avaliação da evidência da amostra em relação a uma hipótese sobre a população. Os intervalos de confiança fornecem uma estimativa de um intervalo dentro do qual o parâmetro da população pode estar com um determinado nível de confiança.

Intervalo de Confiança

Intervalo de confiança é uma medida estatística que fornece um intervalo de valores plausíveis para um parâmetro desconhecido da população com um determinado nível de confiança. É um intervalo estimado a partir de uma amostra de dados que provavelmente inclui o valor verdadeiro do parâmetro.

O intervalo de confiança é expresso como uma faixa de valores com um limite inferior e um limite superior. A estimativa pontual do parâmetro é geralmente o valor central do intervalo. O nível de confiança indica a probabilidade de que o intervalo de confiança inclua o valor verdadeiro do parâmetro.

Teste de Hipótese

O teste de hipótese é uma técnica estatística que permite avaliar se uma afirmação sobre uma população é suportada ou não pelos dados amostrais disponíveis. Em outras palavras, o teste de hipótese é um procedimento que permite tirar conclusões sobre uma população com base em informações obtidas de uma amostra.

O processo de teste de hipótese envolve a formulação de uma hipótese nula (H0), que representa a afirmação a ser testada e uma hipótese alternativa (H1), que representa a negação da hipótese nula. O objetivo do teste é avaliar se há evidências suficientes para rejeitar a hipótese nula em favor da hipótese alternativa.

Para realizar um teste de hipótese, é necessário escolher um nível de significância, que representa a probabilidade máxima de cometer um erro do tipo I (rejeitar erroneamente a hipótese nula). Em seguida, é necessário calcular uma

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

estatística de teste apropriada para o problema em questão e compará-la com um valor crítico ou p-valor determinado pelo nível de significância escolhido.

Se a estatística de teste calcular for maior do que o valor crítico ou se o p-valor for menor do que o nível de significância, a hipótese nula é rejeitada em favor da hipótese alternativa. Caso contrário, não há evidências suficientes para rejeitar a hipótese nula.

Análise de Associação

Análise de associação é uma técnica estatística que busca avaliar a relação entre duas variáveis. Existem vários métodos de análise de associação, mas dois dos mais comuns são o teste Qui-quadrado e a correlação de Pearson.

O teste Qui-quadrado é usado para avaliar a associação entre duas variáveis categóricas. Ele compara as frequências observadas de cada categoria em uma tabela de contingência com as frequências esperadas sob a hipótese nula de que não há associação entre as variáveis. Se as frequências observadas diferem significativamente das esperadas, então a hipótese nula é rejeitada em favor da hipótese alternativa de que há associação entre as variáveis.

Já a correlação de Pearson é usada para avaliar a associação linear entre duas variáveis quantitativas. Ela mede o grau e a direção da relação entre as variáveis, variando de -1 (correlação perfeita negativa) a +1 (correlação perfeita positiva), com 0 indicando ausência de correlação. Uma correlação positiva significa que as variáveis tendem a aumentar ou diminuir juntas, enquanto uma correlação negativa indica que elas tendem a variar em direções opostas. A correlação de Pearson assume que as variáveis seguem uma distribuição normal e que a relação entre elas é linear.

Ambos os métodos podem ser úteis em diferentes contextos de análise de dados. O teste Qui-quadrado é útil para investigar associações entre variáveis categóricas, como a relação entre gênero e preferência de um produto. Já a correlação de Pearson é útil para avaliar associações entre variáveis quantitativas, como a relação entre idade e renda. No entanto, é importante lembrar que a presença de uma associação entre duas variáveis não implica necessariamente uma relação causal entre elas, e outros fatores podem estar influenciando os resultados observados.

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

Que tal dar uma lida nesse artigo do blog do medium? O autor Lucas Ribeiro traz uma boa visão e exemplos na aplicação da estatística para análise de dados.

Álgebra Linear

Álgebra linear é um ramo da matemática que se dedica ao estudo de vetores, espaços vetoriais, transformações lineares, sistemas de equações lineares e matrizes. Ela permite a representação e manipulação de dados em várias dimensões, o que a torna uma ferramenta poderosa para resolver problemas complexos. Algumas das principais técnicas utilizadas em álgebra linear incluem diagonalização de matrizes, decomposição de valores singulares, e resolução de sistemas de equações lineares.

Na análise de dados, a álgebra linear é uma ferramenta fundamental para lidar com conjuntos de dados que podem ser representados como matrizes e vetores, no qual auxilia (e muito) na resolução de problemas de otimização, como a minimização de erros ou a maximização de funções de custo, por exemplo. Além disso, é utilizada para realizar transformações de dados, como rotações, escalas e projeções, que podem ser usadas para melhorar a visualização e interpretação de dados.

Algumas das principais técnicas de álgebra linear utilizadas na análise de dados incluem a decomposição de matrizes, como a decomposição em valores singulares (SVD) e a decomposição de autovalores e autovetores (PCA), além de técnicas de resolução de sistemas lineares, como a eliminação de Gauss e a decomposição LU.

Teoria dos Conjuntos

A teoria dos conjuntos é uma área da matemática que estuda a propriedade e a estrutura dos conjuntos, que são coleções de objetos bem definidos. Na análise de dados, a teoria dos conjuntos é uma ferramenta importante para lidar com conjuntos de dados e realizar operações entre eles.

Os conjuntos são utilizados na análise de dados para representar grupos de elementos com características semelhantes, como por exemplo, um conjunto de

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

clientes de uma empresa ou um conjunto de observações de uma variável específica em um estudo estatístico.

Algumas das operações de conjuntos mais comuns utilizadas na análise de dados incluem a união, interseção e diferença entre conjuntos. Essas operações podem ser usadas para identificar padrões, comparar grupos e filtrar dados.

Além disso, a teoria dos conjuntos é utilizada em técnicas de Machine Learning, como a classificação de dados em grupos ou categorias, baseada na semelhança entre conjuntos de dados.

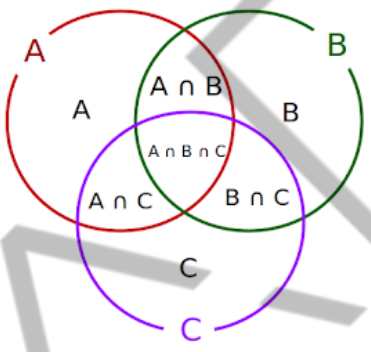


Figura 4 – Exemplo ilustrativo sobre a união entre os conjuntos
Fonte: Medium – Teoria dos Conjuntos – João Renato (2020)

Sistemas Lineares

Sistemas lineares são um conjunto de equações lineares que descrevem as relações entre várias variáveis em um sistema. Na análise de dados, sistemas lineares são frequentemente utilizados para modelar problemas que envolvem múltiplas variáveis e suas interações.

Um sistema linear é composto por um conjunto de equações lineares, onde cada equação é formada por uma combinação linear das variáveis envolvidas. Por exemplo, um sistema linear com duas variáveis x e y pode ser representado por duas equações:

$2x + 3y = 7$

$x - y = 2$

O objetivo da resolução de um sistema linear é encontrar os valores das variáveis que satisfazem todas as equações simultaneamente. Existem várias

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

técnicas para resolver sistemas lineares, incluindo a eliminação de Gauss, a decomposição LU e a utilização de matrizes.

Na análise de dados, os sistemas lineares são usados em diversas aplicações, como por exemplo, em modelos de regressão linear, onde as variáveis independentes são combinadas linearmente para prever a variável dependente. Os sistemas lineares também podem ser usados para resolver problemas de otimização, como encontrar os valores das variáveis que maximizam ou minimizam uma função objetivo sujeita a restrições lineares.

Vetores e Matrizes

Um vetor é uma sequência ordenada de números, representados como uma coluna ou uma linha de valores. Na análise de dados, os vetores são frequentemente usados para representar observações individuais ou amostras de dados, onde cada elemento do vetor representa uma variável diferente.

Por exemplo, um vetor de dados pode representar as idades de um grupo de pessoas:

[23, 29, 36, 42, 27]

Já uma matriz é uma coleção de números organizados em linhas e colunas. Na análise de dados, as matrizes são usadas para representar conjuntos de dados multidimensionais, onde cada linha da matriz representa uma observação e cada coluna representa uma variável.

Por exemplo, uma matriz de dados pode representar a altura, peso e idade de um grupo de pessoas:

| 1.70 | 70 | 23 |

| 1.60 | 60 | 29 |

| 1.80 | 80 | 36 |

| 1.65 | 65 | 42 |

| 1.75 | 75 | 27 |

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

As operações com vetores e matrizes são amplamente utilizadas na análise de dados, como por exemplo, a adição e subtração de vetores ou matrizes, a multiplicação de vetores e matrizes, a transposição de matrizes e a inversão de matrizes.

Espaço Vetorial

Um espaço vetorial é um conjunto de vetores, que podem ser somados e multiplicados por escalares, obedecendo a certas regras e propriedades matemáticas. Na análise de dados, os espaços vetoriais são amplamente utilizados para representar e manipular conjuntos de dados multidimensionais. Ele possui as seguintes propriedades:

- Adição: a adição de dois vetores em um espaço vetorial resulta em outro vetor no mesmo espaço vetorial.
- Multiplicação por escalar: a multiplicação de um vetor por um escalar em um espaço vetorial resulta em outro vetor no mesmo espaço vetorial.
- Comutatividade e associatividade: as operações de adição e multiplicação por escalar em um espaço vetorial são comutativas e associativas.
- Elemento neutro: existe um vetor no espaço vetorial, chamado de vetor nulo, que não altera o resultado quando adicionado a qualquer outro vetor.
- Elemento inverso: para cada vetor em um espaço vetorial, existe um vetor oposto que, quando adicionado ao vetor original, resulta no vetor nulo.
- Distributividade: a multiplicação por escalar distribui sobre a adição de vetores.

Na análise de dados, os espaços vetoriais são usados para representar e manipular conjuntos de dados multidimensionais, como por exemplo, vetores de características de imagens, conjuntos de dados em problemas de Machine Learning e séries temporais. As propriedades dos espaços vetoriais permitem realizar operações matemáticas e estatísticas eficientes, como a projeção de vetores, a identificação de subespaços e a aplicação de transformações lineares.

FIAP

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

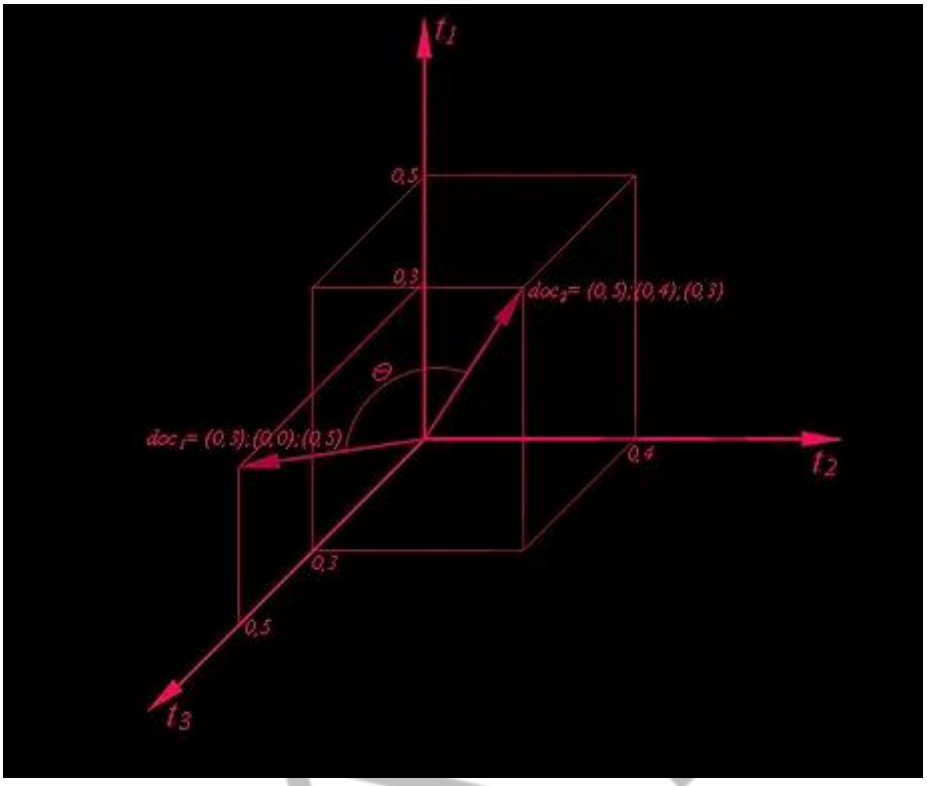


Figura 5 – Representação gráfica de vetores
Fonte: https://www.researchgate.net/figure/Figura-2-Representacao-do-modelo-de-espaco-vetorial_fig2_216835685 (2010), adaptado por FIAP (2023)

Distância Euclidiana

A distância euclidiana é uma medida de distância entre dois pontos em um espaço euclidiano. Na análise de dados, a distância euclidiana é frequentemente usada para calcular a distância entre pontos em um espaço vetorial n-dimensional.

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

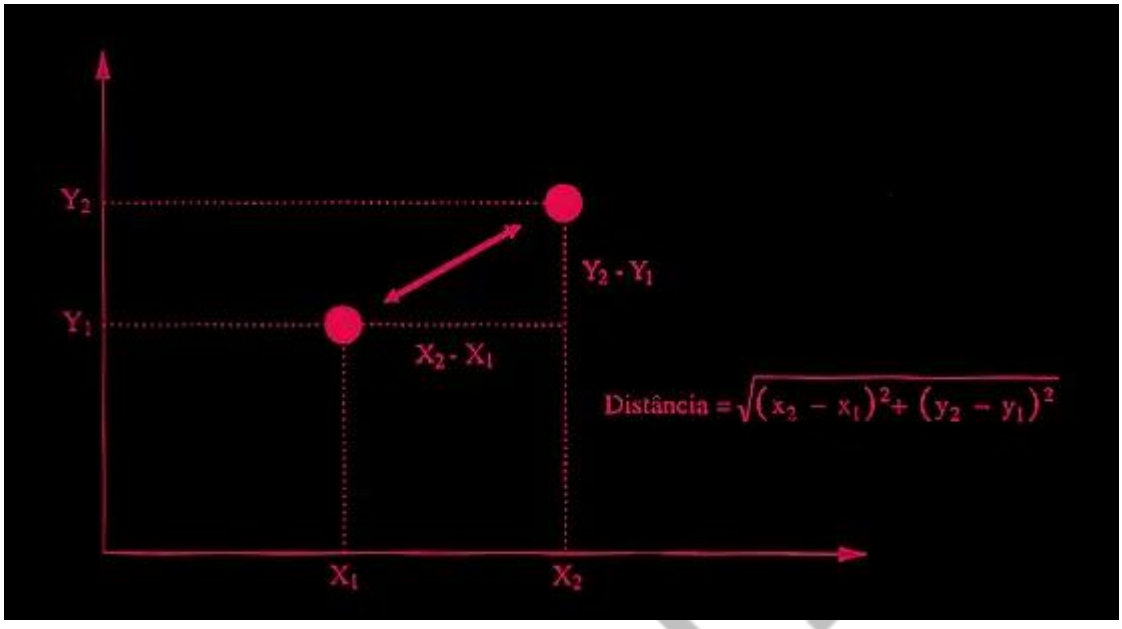


Figura 6 – Representação gráfica sobre a Distancia Euclidiana
Fonte: https://smolski.github.io/livroavancado/analise-de-clusters.html (2004), adaptado por FIAP (2023)

A distância euclidiana é uma medida importante na análise de dados, pois permite calcular a distância entre pontos em espaços n-dimensionais, o que pode ser útil em problemas de clustering, classificação, análise de séries temporais e outras aplicações. Além disso, a distância euclidiana pode ser usada como uma métrica para avaliar a qualidade de modelos de aprendizado de máquina, como a regressão linear e a classificação.

Que tal dar uma lida neste artigo no blog do medium? Nele, o autor Arthur Lamblet traz ótimos exemplos sobre álgebra linear.

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

HANDS ON

Agora, chegou o momento de ver na prática como começar a importar os nossos dados e trabalhar com eles via programação. A ideia é não se limitar apenas ao código explícito no hands on, então recomendamos que procure a documentação das bibliotecas, explore novas funcionalidades e muito mais!

Disponibilizamos o notebook dessa aula para que você possa aprender ainda mais. Que tal verificar?

Notebook - aula 1.

FIAP

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

O QUE VOCÊ VIU NESTA AULA?

Introdução a Modelos de Classificação; Machine Learning; Aplicação de Machine Learning.

Daqui em diante, é importante que você replique os conhecimentos adquiridos para fortalecer mais suas bases e conhecimentos.

IMPORTANTE: não esqueça de praticar com o desafio da disciplina, para que assim você possa aprimorar os seus conhecimentos!

Você não está sozinho(a) nesta jornada! Te esperamos no Discord e nas *lives* com os nossos especialistas, onde você poderá tirar dúvidas, compartilhar conhecimentos e estabelecer conexões!

FIAP

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

REFERÊNCIAS

FÁVERO, Luiz Paulo; BELFIORE, Patricia. **Manual de Análise de Dados - Estatística e Modelagem Multivariada com Excel®, SPSS® e Stata®**. Rio de Janeiro: Elsevier, 2017.

MORETIM, Pedro A.; BUSSAB, Wilson de O.. **Estatistica Básica**. 9. ed. São Paulo: Saraiva, 2017.

FIAP

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

PALAVRAS-CHAVE

Palavras-chave: Python. Pandas. EDA. Seaborn.

FIAP

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com

PDF exclusivo para Marcelo Cardoso da Silva - rm349622
mcardoso.dev@gmail.com