|  |  |
| --- | --- |
| **PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ (PUCPR)****PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA (PPGIa)****CIÊNCIA DE DADOS** **PROF. JEAN PAUL BARDDAL** | **Logo, company name  Description automatically generated** |

**PROVA**

Nome completo: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Data: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Instruções:**

* Leia a prova em sua completude e com atenção;
* Você deve realizar a entrega da prova via o link a seguir até as **12:00 (meio-dia) do dia 17 de abril 2024**. Link de entrega: <https://www.dropbox.com/request/A4NShoynUIxnPF8cX2TZ>
* O arquivo de entrega **deve** estar em formato **ZIP** contendo internamente o **DOCX** deste documento com as respostas, o arquivo da base de dados da parte prática e o jupyter notebook em formato **IPYNB**;
* Nesta prova, as primeiras 7 questões devem ser respondidas e uma justificativa explicitando o motivo das afirmações FALSAS estarem INCORRETAS também deve ser fornecida. Por exemplo, se uma questão possui 3 (três) afirmativas falsas, você deverá fornecer 3 (três) justificativas. Uma exceção é a questão 8, de caráter discursivo;
* As questões 9 em diante são práticas e cada estudante deverá responder de acordo com a sua base de dados. O código-fonte da análise de dados e a base de dados devem ser entregues juntamente do arquivo DOCX com as respostas da prova.
* Respostas **sem** as requisitadas justificativas receberão **nota zero**;
* O enunciado de cada questão **NÃO** pode ser editado. Caso isso ocorra, será atribuída a **nota zero** para cada questão em que isso ocorrer; e
* Não há consulta ao professor durante a prova. Caso um enunciado esteja vago, ambíguo ou incorreto, forneça um detalhamento e justificativa no campo de resposta.

|  |
| --- |
| **Questão 1 - (1 ponto)** |
| A análise exploratória de dados combina estatística, visualização de dados e raciocínio crítico. Focando em componentes estatísticas, média, mediana, moda, desvio padrão e quartis são conceitos amplamente utilizados. Sobre estes descritores, analise as afirmativas a seguir:1. A média é pouco impactada por outliers.
2. A mediana pode ser calculada ao dividir a soma de todos os valores de uma amostra de dados pelo tamanho desta amostra.
3. A moda é calculável apenas para variáveis categóricas.
4. O intervalo interquartil é uma alternativa ao desvio padrão pois também quantifica quão disperso um conjunto de dados é.

Quais das alternativas acima estão corretas? |
| **Resposta:****Exemplo: Afirmativas ii e iv são verdadeiras.** |
| **Justificativa(s):****Exemplo:** Afirmativa I é falsa pois...**Exemplo:** Afirmativa III é falsa pois... |

|  |
| --- |
| **Questão 2 - (1 ponto)** |
| Histogramas são gráficos univariados que permitem analisar e entender a distribuição de variáveis. Sobre histogramas, analise as afirmativas abaixo:1. O número de partições (colunas) em um histograma não afeta o entendimento do comportamento de uma variável.
2. Histogramas são gráficos de frequência aplicáveis, estritamente, a variáveis categóricas.
3. Histogramas permitem análise de curtose e assimetria de forma visual.
4. Quando o número de partições (colunas) e a quantidade de dados de um histograma são grandes, histogramas se tornam idênticos a gráficos de densidade.

Quais das afirmativas acima estão corretas? |
| **Resposta:** |
| **Justificativa:** |

|  |
| --- |
| **Questão 3 - (1 ponto)** |
| Um analista de dados está conduzindo uma análise univariada sobre um dataset que contém as gorjetas recebidas por um restaurante. O analista computou as seguintes estatísticas sobre as gorjetas obtidas durante um mês:- Média: 2.481- Mediana: 2.0, e- Desvio padrão: 1.59Além disso, o analista construiu o seguinte box-plot para esta distribuição:A picture containing door, clock  Description automatically generatedDe acordo com as estatísticas e box-plots, analise as afirmativas abaixo:1. O box-plot não representa a mesma variável descrita pelas estatísticas.
2. O motivo é que podemos verificar que a média calculada algebricamente não condiz com a média apresentada no box-plot.
3. O box-plot indica a existência de outliers.
4. A variável apresentada no box-plot tem comportamento simétrico, isto é, similar a uma distribuição gaussiana.
5. Na amostra de dados analisada, não existem valores negativos.

Quais das afirmativas acima estão corretas? |
| **Resposta:** |
| **Justificativa:** |

|  |
| --- |
| **Questão 4 - (1 ponto)** |
| A análise de correlações nos permite verificar o comportamento de duas variáveis e como elas se relacionam entre si. Para conduzir uma análise de correlações, os coeficientes de Pearson e Spearman são amplamente utilizados. Sobre o coeficiente de Spearman, analise as afirmativas abaixo:1. O coeficiente de Pearson indica tendência e força de uma correlação entre duas variáveis;
2. A análise de Pearson quantifica a relação linear entre um par de variáveis;
3. Um coeficiente de Pearson alto para um par de variáveis indica causalidade entre estas variáveis;

Quais das afirmativas acima estão corretas? |
| **Resposta:** |
| **Justificativa:** |

|  |
| --- |
| **Questão 5 - (1 ponto)** |
| A identificação e remoção de outliers é de extrema importância durante as etapas de análise e preparação de dados. Uma das abordagens mais comuns para identificar outliers é o método de Tukey. Focando neste método, analise as afirmativas abaixo:1. O método de Tukey é baseado no intervalo inter-quartis das variáveis. O intervalo inter-quartil de uma variável abrange 50% dos dados analisados.
2. A definição de valores máximos e mínimos (whiskers) segue o formato [$Q\_{1}-k(Q\_{3}-Q\_{1}), Q\_{3}+k(Q\_{3}-Q\_{1})$].
3. Em uma análise multivariada de Tukey, os valores anômalos detectados em ao menos duas variáveis devem ser automaticamente removidos do dataset.
4. Uma característica relevante do método de Tukey é que ele pode ser aplicado a datasets com assimetria forte pois seus cálculos são baseados em média e desvio padrão.

Quais das afirmativas acima estão corretas? |
| **Resposta:** |
| **Justificativa:** |

|  |
| --- |
| **Questão 6 - (1,5 pontos)** |
| Visualização efetiva de dados objetiva mandar uma mensagem clara e precisa ao leitor. Desta forma, é de extrema importância que visualizações sejam projetadas de forma adequada e aderente aos tipos de dados sendo apresentados, incluindo preocupações com componentes visuais de tamanho, forma e cores. Abaixo temos um gráfico de pizza que representa o número de mortes por COVID-19 no Brasil.Mortes por Covid-19 no Brasil com doenças associadas ate 13 de abril — Foto: Carolina Dantas/G1Fonte: https://g1.globo.com/bemestar/coronavirus/noticia/2020/04/15/pacientes-com-cancer-e-cardiacos-deixam-de-buscar-atendimento-por-medo-da-covid-19-alertam-medicos.ghtml - Acesso em: 6 de maio de 2020.Dada a visualização acima, analise as afirmativas abaixo:1. Uma possível melhoria para esta visualização seria adotar uma escala de cores quantitativa, uma vez que não há ordem ou relação entre as comorbidades apresentadas.
2. Uma possível melhoria para esta visualização seria rotular cada segmento do gráfico com o número de casos ou percentual associado.
3. Uma possível melhoria para esta visualização seria mudar o seu tipo para um gráfico de barras.
4. Uma possível melhoria para esta visualização seria apresentar no gráfico o número de mortes em que nenhuma comorbidade estava presente. Esta mudança acarretaria uma grande mudança nos segmentos e suas proporções visuais.

Quais das afirmativas acima estão corretas? |
| **Resposta:** |
| **Justificativa:** |

|  |
| --- |
| **Questão 7 - (1,5 pontos)** |
| Um método popular na Ciência de Dados é a Análise de Componentes Principais, ou Principal Component Analysis (PCA). Sobre o PCA, analise as afirmativas abaixo:1. PCA é um método não-supervisionado, isto é, não considera as classes de um problema em consideração.
2. PCA estabelece componentes principais baseando-se na covariância entre as variáveis de um dataset.
3. O número máximo de componentes principais gerado pelo PCA é maior ou igual ao número de variáveis originais em um dataset.
4. PCA é não-paramétrico, i.e., não pode ser usado para aprender componentes em um conjunto de treinamento e posteriormente aplicado em um conjunto de teste.

Quais das afirmativas acima estão corretas? |
| **Resposta:** |
| **Justificativa:** |

|  |
| --- |
| **Questão 8 - (2 pontos)** |
| Você recebeu um conjunto de dados (dataset) tabular com milhares de colunas. Seu objetivo é reduzir a dimensionalidade deste dataset de forma que instâncias similares no espaço original de atributos se mantenham próximas, (isto é, sejam projetadas na mesma vizinhança) em um espaço de dimensionalidade reduzido. Neste caso, qual algoritmo você deve usar e por quê? |
| **Resposta e justificativa:** |

**PARTE PRÁTICA**

|  |
| --- |
| **Questão 9 - (2 pontos)** |
| Dada a base de dados, apresente uma análise geral da base de dados no que tange número de linhas, número de atributos e valores faltantes. Apresente prints de tela legíveis e discorra sobre os prints apresentados. |
| **Resposta:** |

|  |
| --- |
| **Questão 10 - (2 pontos)** |
| Dada a base de dados, apresente uma análise univariada de ao menos 3 variáveis que você julga mais relevante. Faça uso de estatísticas e visualizações e aponte os principais comportamentos observados e o significado de cada um deles no contexto da base de dados. |
| **Resposta:** |

|  |
| --- |
| **Questão 11 - (2 pontos)** |
| Dada a base de dados, apresente uma análise multivariada de ao menos 3 subconjuntos de variáveis que você julga mais relevante. Faça uso de estatísticas e visualizações e aponte os principais comportamentos observados e o significado de cada um deles no contexto da base de dados. |
| **Resposta:** |

|  |
| --- |
| **Questão 12 - (2 pontos)** |
| Dadas as visualizações multivariadas da questão anterior, escolha uma delas e realize aprimoramentos pensando em uma análise explanatória de dados, isto é, que eventualmente possa ser usada em uma apresentação, relatório ou documento formal ou científico. Forneça abaixo um print da visualização preparada e justifique as mudanças realizadas em relação a visualização criada inicialmente. |
| **Resposta:** |