PROGRAMA DE RESIDÊNCIA EM TECNOLOGIA DA INFORMAÇÃO
JUSTIÇA FEDERAL NO RIO GRANDE DO NORTE

EDITAL 004/2019 – PROVA DE CONHECIMENTOS ESPECÍFICOS
ÁREA DE CONCENTRAÇÃO 3: BUSINESS INTELLIGENCE E ANALYTICS
24 / 02 / 2019

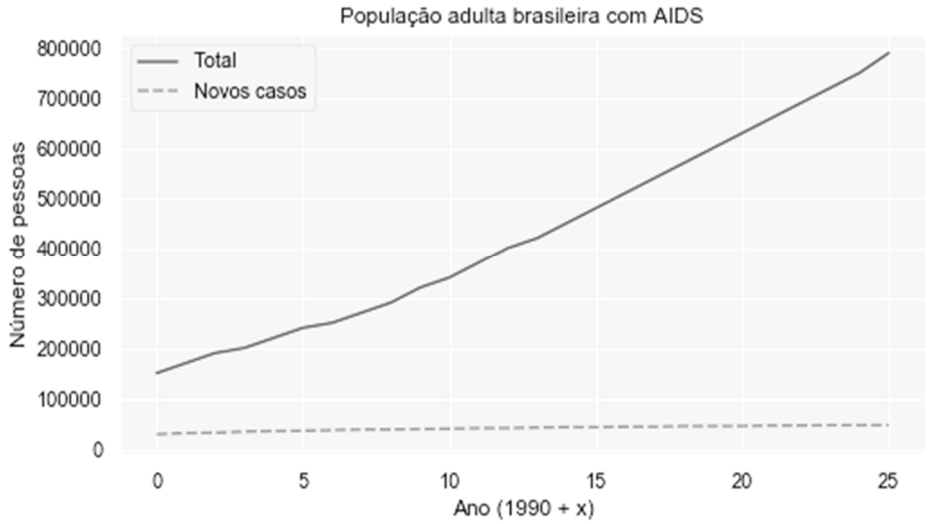
Identificação do Candidato	
Nome completo:	
CPF:	**Assinatura:**

Leia com atenção as seguintes instruções:

1. Aguarde a autorização do(s) fiscal(is) para poder iniciar a Prova.
2. Não esqueça de colocar seu **nome completo** (preferencialmente em letras maiúsculas) e de assinar o campo acima.
3. Este Caderno de Prova, com páginas numeradas de 1 a 12, é constituído de **30 (trinta) questões** de múltipla escolha, cada uma com quatro alternativas. Verifique se o Caderno de Prova está completo, sem falhas de impressão ou problemas que comprometam sua leitura. Caso necessário, solicite imediatamente ao(s) fiscal(is) a substituição do Caderno de Prova completo.
4. Confira se este Caderno de Prova corresponde à área de concentração para a qual foi inscrito. Caso haja alguma divergência, notifique imediatamente o(s) fiscal(is).
5. Leia com atenção o enunciado das questões antes de responde-las.
6. Cada questão possui **apenas uma** alternativa correta. Você deverá marcar a resposta que julgar correta usando caneta esferográfica de tinta na cor azul ou preta no local correspondente à respectiva questão. A interpretação das questões faz parte da Prova, de modo que não será permitida qualquer tipo de pergunta ou explicação ao(s) fiscal(is).
7. Não serão computadas questões sem marcação de resposta ou que contenham mais de uma marcação, marcação rasurada ou emendada.
8. Tenha cuidado ao manusear este Caderno de Prova, evitando rasuras, pois ele **não será substituído** por esse motivo. Também não é permitido destacar quaisquer das folhas que compõem este Caderno de Prova.
9. O tempo máximo para resolução desta Prova é de **2 (duas) horas**, para o qual **não haverá prorrogação**. Transcorrido esse tempo, o Caderno de Prova será recolhido pelo(s) fiscal(is).
10. Terminada a realização da Prova, este Caderno de Prova deverá ser **obrigatoriamente** entregue ao(s) fiscal(is) antes de se retirar da sala de realização.

QUESTÕES

1. Considere o gráfico abaixo, onde a linha contínua representa o número de adultos com HIV, enquanto a linha tracejada representa a quantidade de novos casos de adultos com HIV:



Avalie as afirmações a seguir e escolha a opção correta:

I. O número de casos entre os anos de 1991 e 1992 contraria a tendência geral do gráfico.
II. A proporção entre casos novos e antigos tem crescido ao longo dos anos.
III. A forma das curvas indica que dados mais recentes podem ter sido obtidos através de extrapolação dos dados.

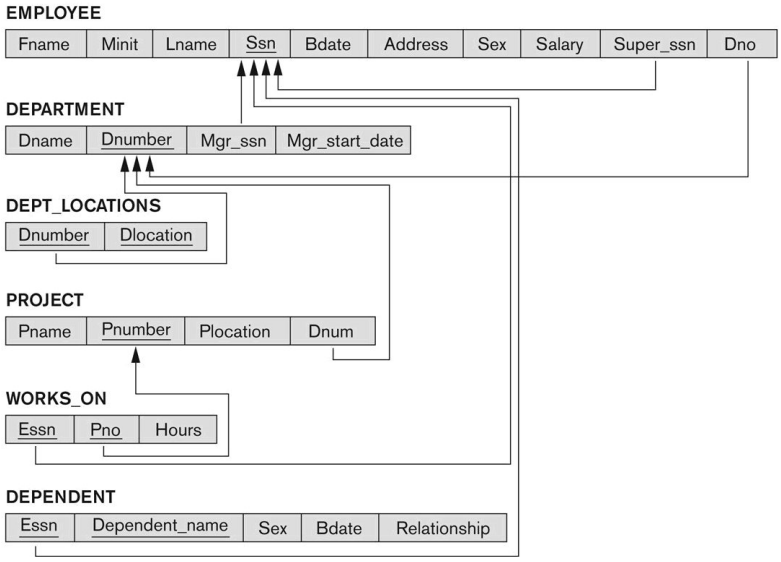
a) Apenas a afirmativa I está correta
b) Apenas a afirmativa II está correta
c) Apenas a afirmativa III está correta
d) Apenas as afirmativas I e III estão corretas

2. Sobre a análise de agrupamentos, é possível afirmar que:
a) o *k-Nearest Neighbor* (*k*-NN) é um popular algoritmo utilizado para realizar a tarefa de agrupamento
b) é uma técnica que tem como principal objetivo realizar a classificação de novos dados a partir de um modelo construído a partir de uma base com dados rotulados
c) é uma técnica que identifica elementos semelhantes e os separa para uma análise exploratória mais profunda
d) é uma técnica cujo uso requer que os dados estejam devidamente rotulados com as respectivas classes

3. O processo de ETL (extração, transformação e carga) é utilizado para realizar carga de dados em bancos de dados do tipo *data warehouse*. Acerca desse processo, assinale a alternativa verdadeira:

a) A carga incremental é usualmente mais fácil de ser implementada do que a carga completa, uma vez que trabalha com uma menor quantidade de dados por vez.
b) Não é uma prática recomendada criar uma base intermediária para integração de dados antes da carga no *data warehoue*, pois gera duplicação de dados.
c) Primeiramente, deve ser realizada a carga das tabelas Dimensão, que podem ser realizadas em paralelo, para depois se realizar a carga nas tabelas Fato.
d) As chaves primárias no *data warehouse* devem refletir as chaves primárias dos sistemas de origem, visando garantir a rastreabilidade dos dados.

Para responder as questões 4 e 5, considere o seguinte modelo relacional:



4. Qual das opções abaixo apresenta a consulta que lista os primeiros nomes de todos os gerentes que não possuem dependentes?

a)
```
SELECT fname FROM EMPLOYEE WHERE
  (ssn IN (SELECT mgr_ssn FROM DEPARTMENT)) AND
  (ssn NOT IN (SELECT essn FROM DEPENDENT));
```
b)
```
SELECT fname FROM EMPLOYEE WHERE
  (ssn IN (SELECT mgr_ssn FROM DEPARTMENT)) OR
  (ssn NOT IN (SELECT essn FROM DEPENDENT));
```
c)
```
SELECT fname FROM EMPLOYEE WHERE
  (ssn IN (SELECT mgr_ssn FROM DEPARTMENT)) OR
  (ssn IN (NOT SELECT essn FROM DEPENDENT));
```
d)
```
SELECT fname FROM EMPLOYEE WHERE
  (ssn NOT IN (SELECT mgr_ssn FROM DEPARTMENT)) AND
  (ssn NOT IN (SELECT essn FROM DEPENDENT));
```

5. Qual das opções abaixo apresenta a consulta que lista os primeiros nomes (`fname`) e endereços (`address`) de todos os empregados que trabalham em pelo menos um projeto localizado em `'Natal'` (`plocation`) mas cujo departamento do empregado (`dno`) não é localizado em `'Natal'` (`dlocation`)?

a)
```
SELECT fname, address FROM EMPLOYEE
```

```
  WHERE ((ssn IN (SELECT essn FROM WORKS_ON JOIN PROJECT ON pnumber=pno
         WHERE plocation='Natal'))
  AND
  (ssn NOT IN
       (SELECT ssn FROM EMPLOYEE WHERE dno IN
             (SELECT dnumber FROM (DEPARTMENT NATURAL JOIN DEPT_LOCATIONS)
             WHERE dlocation='Natal'))));
```

b)
```
  SELECT fname, address FROM EMPLOYEE
  WHERE ((ssn NOT IN (SELECT essn FROM WORKS_ON JOIN PROJECT ON pnumber=pno
  WHERE plocation='Natal'))
  AND
  (ssn NOT IN
       (SELECT ssn FROM EMPLOYEE WHERE dno IN
       (SELECT dnumber FROM (DEPARTMENT NATURAL JOIN DEPT_LOCATIONS)
                     WHERE dlocation='Natal'))));
```

c)
```
  SELECT fname, address FROM EMPLOYEE
  WHERE((ssn IN (SELECT essn FROM WORKS_ON JOIN PROJECT ON pnumber=pno
  WHERE plocation='Natal'))
  AND
  (ssn NOT IN
       (SELECT ssn FROM EMPLOYEE WHERE dno NOT IN
       (SELECT dnumber FROM (DEPARTMENT NATURAL JOIN DEPT_LOCATIONS)
                     WHERE dlocation='Natal'))));
```

d)
```
  SELECT fname, address FROM EMPLOYEE
  WHERE((ssn IN (SELECT essn FROM WORKS_ON JOIN PROJECT ON pnumber=pno
  WHERE plocation='Natal'))
  AND
  (ssn IN (SELECT ssn FROM EMPLOYEE WHERE dno IN
          (SELECT dnumber FROM (DEPARTMENT NATURAL JOIN DEPT_LOCATIONS)
                        WHERE dlocation='Natal'))));
```

6. O *QlikView* é uma plataforma para implementação de serviços de *business intelligence* (BI). Acerca dessa plataforma, assinale a afirmativa correta:

a) Um conjunto de funções é disponibilizado pelo *QlikView* para que se implemente *scripts*. São exemplos disso as funções lógicas e as funções de agregação.

b) O *QlikView* é voltado para visualização e análise de dados, não sendo utilizado para operações de carregamento de dados.

c) Para facilitar a interação com o usuário de negócio, a plataforma oferece relatórios previamente elaborados e não mecanismos avançados de interação com dados.

d) A plataforma trabalha com o conceito de pastas, as quais refletem as pastas no sistema operacional e nas quais são salvos os relatórios gerados.

7. São exemplos de abordagens de mineração de dados mais apropriados para se criar um modelo de classificação:

a) algoritmo genético e árvores B

b) árvores de decisão e florestas aleatórias (*Random Forest*)

c) regressão linear simples e regressão multivariada

d) *k-means* e árvores rubro-negras

8. Ao se deparar com uma base de dados para analisar, é comum que tal base possua uma quantidade de dados faltosos ou indisponíveis. A falta de alguns valores nas bases de dados se deve, muitas vezes, por falta de preenchimento nos cadastros, falha em sensores, entre outros motivos. Para que técnicas de aprendizado de máquina possam utilizar corretamente os dados, é geralmente necessário corrigir esse tipo de problema. Dentre as abordagens utilizadas para tratar a dados faltosos, analise as seguintes afirmativas:

I. A retirada dos exemplos que possuem dados faltosos é obrigatória para dar seguimento à análise de dados.
II. Pode-se substituir o campo faltoso por valores retirados a partir dos próprios dados, como a média ou a moda.
III. É possível utilizar um modelo de classificação para preencher automaticamente os dados.

Sobre as afirmativas anteriores, é correto dizer que:
a) somente II está correta
b) todas estão corretas
c) nenhuma está correta
d) somente I está correta

9. Considere que uma coleção MongoDB contém documentos com estrutura similar ao do documento a seguir:

```
{"title":"Once Upon a Time in the West",
 "year":1968,"rated":"PG-13",
 "runtime":175,
 "countries":["Italy","USA","Spain"],
 "genres":["Western"],
 "director":"Sergio Leone",
 "awards":{"wins":4,"nominations":5}}
```

Qual opção apresenta o filtro a ser usado no Mongo COMPASS a fim de retornar os filmes ranqueados com PG-13 e que têm exatamente dez nomeações de prêmios?
a) `{rated: "PG-13", "awards.nominations":10}`
b) `{rated="PG-13" & "awards.nominations"=10}`
c) `{rated:"PG-13", "awards":{nominations:10}}`
d) `{rated="PG-13", "awards.nominations"=10}`

10. Quais das seguintes técnicas não é mais adequada para realizar classificação de dados?
a) Máquinas de Vetores de Suporte (SVM)
b) *k-means*
c) *Perceptron* de Múltiplas Camadas (MLP)
d) *Random Forest*

11. Acerca dos tipos de gráficos existentes, assinale a alternativa correta:
a) Gráficos de barras ilustram tendências ao longo do tempo.
b) Gráficos de pizza são muito bons para representar correlações entre os dados.

c) Gráficos de caixa (*boxplots*) representam bem a dispersão dos dados.
d) Gráficos de linhas são úteis para se ilustrar dados percentuais.

12. Sobre os componentes de um *data warehouse*, considere as seguintes características:

I. informações identificadas por assuntos ou departamentos específicos.
II. capacidade de encontrar padrões nos dados.
III. capacidade de analisar informações em múltiplas perspectivas.
IV. processo de extração, tratamento e limpeza dos dados.

As características de I a IV são, respectivamente:
a) *Staging Area*, *Data Mining*, OLAP, ETL
b) *Data Mart*, *Data Mining*, OLAP, ETL
c) *Drill Through*, OLTP, *Drill Across*, *Staging Area*
d) Cubo de dados, OLTP, *Data Mining*, operações *Drill*

13. Os histogramas são representações gráficas de:
a) distribuição dos dados
b) correlação entre variáveis
c) acurácia de classificação
d) predição de classes

14. São exemplos de abordagens para treinamento e validação de modelos em um processo de mineração de dados:

I. *Weka*
II. *Bootstrap*
III. Validação cruzada (*cross-validation*).
IV. *Leave-one-out*.

Estão corretas as afirmativas:
a) I e III
b) I, III e IV
c) II, III e IV
d) I e III

15. Sobre o processo de amostragem de dados, analise as seguintes afirmativas:

I. Para ser considerada representativa, uma amostra deve possuir aproximadamente as mesmas propriedades dos dados originais.
II. O uso de uma amostra para construir um modelo de classificação tende a produzir um menor custo computacional.
III. Uma amostragem com reposição tem sempre o mesmo efeito que uma amostragem sem reposição.
IV. Para conseguir uma amostra representativa, seu tamanho deve ser sempre superior a 20% do total dos dados.

Estão corretas as afirmativas:
a) I e II
b) III e IV
c) II e III
d) II e IV

Considerando uma modelagem multidimensional através do esquema estrela sobre os julgamentos em um tribunal, onde deseja-se analisar quem julgou, o que se julgou, em que data, o valor financeiro especificamente solicitado e o valor liberado após julgamento (em termos de faixas de valores), analise as questões 16 a 19:

16. A realização do julgamento é representada como:
a) tabela Fato
b) tabela Dimensão
c) atributo de tabela Fato
d) atributo de tabela Dimensão

17. O mês do julgamento é representado como:
a) tabela Fato
b) tabela Dimensão
c) atributo de tabela Fato
d) atributo de tabela Dimensão

18. O juiz que julgou é representado como:
a) tabela Fato
b) tabela Dimensão
c) atributo de tabela Fato
d) atributo de tabela Dimensão

19. O valor financeiro solicitado no processo é representado como:
a) tabela Fato
b) tabela Dimensão
c) atributo de tabela Fato
d) atributo de tabela Dimensão.

20. Com relação aos tipos de bancos NoSQL, analise as afirmativas a seguir:

I. Bancos de dados de documentos armazenam dados como documentos (JSON, XML, etc.). Um exemplo de banco deste tipo é o MongoDB.
II. O banco de dados Neo4J é de um tipo de banco que possuem vértices e arestas representando as relações entre esses vértices.
III. Banco de dados colunares guardam colunas juntas, ao invés de linhas, sendo o tipo de banco do Neo4J.

Estão corretas as afirmativas:

a) I e III
b) I e II
c) II e III
d) I, II e III

Para responder as questões 21 e 22, considere as seguintes tabelas de uma base de dados relacional de um sistema utilizado por comerciantes de equipamentos:

- *Equipamento*: equipamentos comercializados, com o número de série e sua classificação em diferentes categorias, fabricantes e modelos.
- *Cidade*: código de identificação, nome e estado de cada cidade.
- *Cliente*: clientes (empresas), cada um com seu respectivo código e demais dados de identificação básicos, como endereço completo, CNPJ, Inscrição Estadual, etc.
- *Venda*: vendas realizadas por equipamento, com código de identificação, data/hora, local, cliente, equipamento vendido, quantidade e valor negociado.
- *Aquisição*: entradas de equipamentos a serem vendidas posteriormente, com código de identificação, data/hora, tipo de equipamento comprado, quantidade e valor unitário.

21. Em uma modelagem dimensional projetada de acordo com os princípios encontrados na literatura, quais seriam exemplos de possíveis tabelas Fato?
a) *Venda* e *Cliente*
b) *Quantidade* e *Valor*
c) *Cidade* e *Equipamento*
d) *Aquisição* e *Venda*

22. Em uma modelagem dimensional projetada de acordo com os princípios encontrados na literatura, qual seria um exemplo de tabela Dimensão e seus atributos?
a) *Cidade* – nome da cidade e total de vendas naquela cidade
b) *Equipamento* – modelo e fabricante
c) *Aquisição* – data e valor
d) *Venda* – quantidade e valor

23. Considere o gráfico a seguir, no qual são apresentadas as características do *dataset Iris*:

Avalie as afirmações a seguir sobre este tipo de gráfico e escolha a opção correta:

I. É possível analisar as diferentes características dos dados e suas interações.
II. É possível analisar as diferentes características dos dados, mas não suas interações.
III. É possível escaloná-lo para analisar uma quantidade maior de características.

a) Apenas a afirmativa I está correta
b) Apenas a afirmativa II está correta
c) Apenas as afirmativas I e III estão corretas
d) Apenas as afirmativas II e III estão corretas

24. O *Weka* é uma ferramenta computacional que possui um conjunto de algoritmos de aprendizado de máquina para realizar análise de dados. A partir de uma base de dados de uma rede atacadista de supermercados, contendo informações sobre todas as compras feitas nos caixas das lojas, considere as seguintes afirmações do que é possível fazer utilizando o *Weka*:

I. Preparar os dados para utilização por algoritmos de classificação
II. Utilizar a API do *Weka* para incorporar um algoritmo de agrupamento a um software de análise de dados, cuja linguagem de programação seja compatível com o *Weka*
III. Obter automaticamente o significado de dados não conhecidos presentes na base
IV. Identificar possíveis *outliers* nos dados

Está(ão) correta(s) a(s) afirmativa(s):
a) I e II
b) II, III e IV
c) somente a alternativa I
d) I, II e IV

25. Durante o processo de construção de um modelo de classificação, faz-se necessário dividir o conjunto de dados em partes para treinar e testar o modelo. A partir dessa informação, analise as seguintes afirmações:

I. O conjunto de treinamento é utilizado para ajustar os parâmetros do modelo
II. A acurácia do modelo medida com o conjunto de treinamento é uma medida confiável
III. O conjunto de testes é utilizado para medir a acurácia do modelo considerando amostras que nunca foram vistas antes
IV. O conjunto de testes deve ser utilizado durante a fase de treinamento do modelo

São válidas as afirmativas:
a) I, II e III
b) II, III e IV
c) I e III
d) II e IV

26. Com relação à normalização de um banco de dados é correto afirmar que:

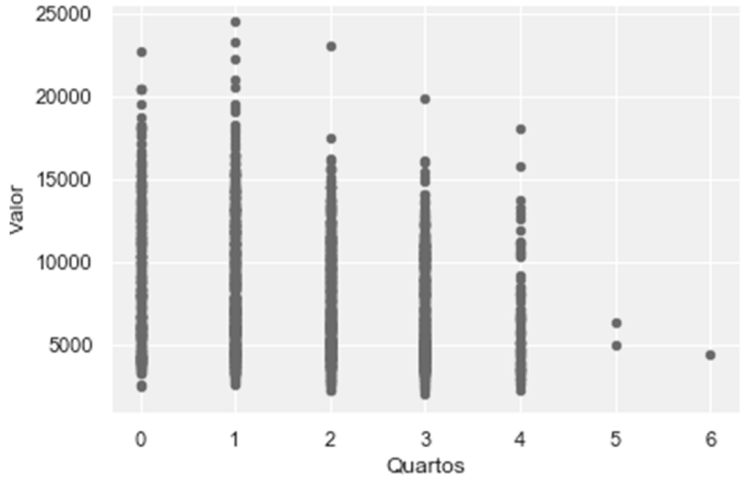
a) A normalização de um modelo relacional visa, principalmente, reduzir a redundância de dados aumentando a sua integridade
b) A normalização aplica-se a um modelo entidade-relacionamento e tem como principal função a remoção de ambiguidades
c) A maioria dos SGBDs atuais aplica automaticamente a normalização
d) A normalização de banco de dados é necessária apenas quando se busca eficiência nas consultas aos bancos de dados relacionais

27. Sobre o aprendizado de máquina supervisionado, considere as seguintes afirmações:

I. As árvores de decisão são um exemplo de algoritmo utilizado para se construir um modelo e realizar a tarefa de classificação.
II. Não são necessários rótulos de dados (classes) para que os algoritmos consigam realizar a construção do modelo de classificação.
III. O principal objetivo do aprendizado supervisionado é realizar a separação dos dados em grupos distintos baseado na similaridade entre os objetos.
IV. A construção de um modelo eficiente depende da qualidade e representatividade dos dados utilizados.

São corretas as afirmações:
a) I, apenas
b) I, II e III
c) II e IV
d) I e IV

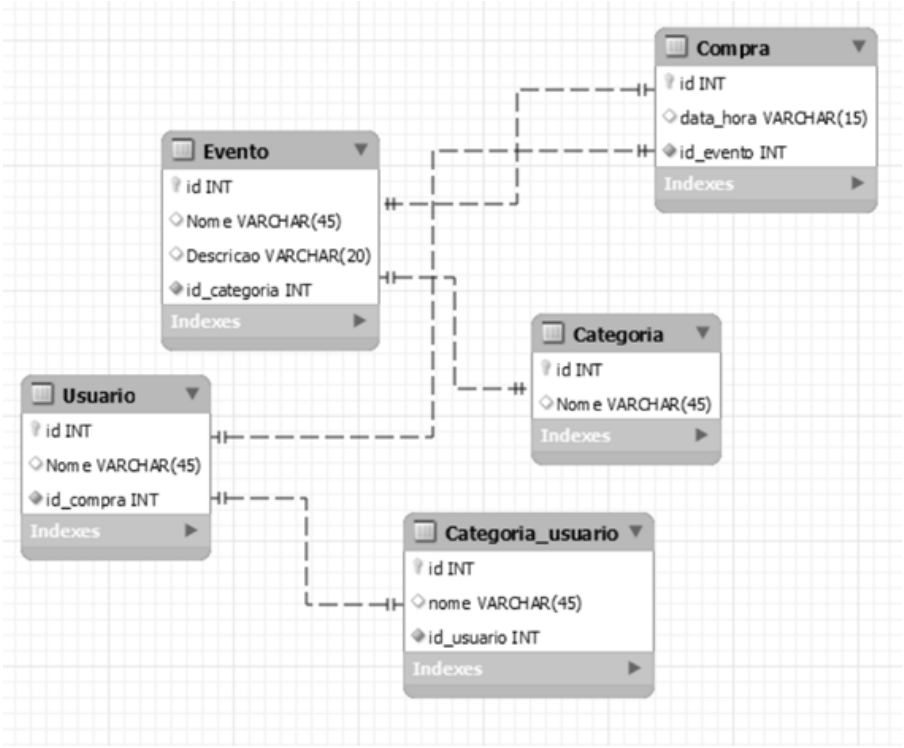
28. Considere o gráfico a seguir, no qual são apresentados dados referentes a um *dataset* de preços de imóveis.



Avalie as afirmações a seguir sobre este tipo de gráfico e escolha a opção correta:
a) Trata-se de um gráfico de dispersão, que analisa a distribuição de uma variável contínua em relação a outra variável contínua.

b) Trata-se de um gráfico de dispersão, que analisa a distribuição de uma variável contínua em relação a uma variável discreta.
c) Trata-se de um histograma, que analisa a distribuição de uma variável contínua em relação a outra variável contínua.
d) Trata-se de um histograma, que analisa a frequência de uma variável contínua.

Para as questões 29 e 30, considere a seguinte modelagem, referente ao banco de dados de um sistema de gerenciamento de eventos classificados por categoria. Para tais eventos, os ingressos podem ser comprados individualmente e nominalmente apenas por usuários cadastrados no sistema, restringindo-se a um único ingresso por usuário para cada evento cadastrado.



29. Considerando a modelagem do banco de dados anteriormente apresentada, analise as seguintes afirmativas sobre problemas existentes ou falta de boas práticas aplicadas nessa modelagem:

I. O tipo utilizado na coluna `data_hora` da tabela `Compra` não é o mais apropriado.
II. A tabela `Usuario` não deveria ter a coluna `id_compra`.
III. A coluna `Nome` da tabela `Categoria` deveria ser uma coluna da tabela `Evento`.
IV. As linhas que representam relações entre tabelas poderiam não estar se cruzando.

Estão corretas as afirmativas:
a) I, II, III
b) I, III, IV
c) II, III, IV
d) I, II, IV

30. Considerando o modelo anteriormente apresentado, analise as seguintes afirmativas:

I. Para se visualizar para quais eventos cada usuário comprou ingresso, usa-se o seguinte comando SQL: `SELECT * FROM evento, categoria, compra, usuario`

II. Para se visualizar os nomes dos usuários que compraram ingresso para o evento de ID igual a 2, usamos o seguinte comando SQL:
```
SELECT usuario.nome FROM usuario
WHERE usuario.id_compra=compra.id AND compra.id_evento=2
```

III. Para se visualizar os eventos que não foram relacionados a nenhuma categoria, usamos o seguinte comando SQL:
```
SELECT evento.nome FROM evento
WHERE evento.id_categoria=0
```

IV. Para se visualizar a quantidade de eventos cadastrados no sistema, usamos o seguinte comando SQL: `SELECT SUM(id) FROM evento`

Quanto às alternativas anteriores, são falsas:
a) apenas uma
b) apenas duas
c) apenas três
d) todas