MBA em Ciência de Dados

Técnicas Avançadas de Captura e Tratamento de Dados

Módulo III - Aquisição e Transformação de Dados

Exercícios

Moacir Antonelli Ponti

CeMEAI - ICMC/USP São Carlos

Recomenda-se fortemente que os exercícios sejam feitos sem consultar as respostas antecipadamente.

Utilize as bibliotecas conforme descrito abaixo

In [1]:
```
# carregando as bibliotecas necessárias
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

Exercício 1)

Tomando como base o conteúdo das seções 1.3.1, 1.3.2 e 1.3.3 do Capítulo "Introduction to Data", de Open Statistics. Considere os seguintes cenários:

I - Analisando dados de educação, amostramos 30 escolas de todo o Brasil e observamos que, dessas 30 escolas, 10% não submeteram o resultado Pisa para Escolas, sendo que todas essas estão no estado de São Paulo. Assim, concluímos que existe uma possível relação entre não submissão e o estado de São Paulo. II - Desejamos estudar a percepção da facilidade de usar um aplicativo desenvolvido por nossa empresa para uso no segmento de atividades físicas/esportes. Para isso montamos um questionário e selecionamos 20 pessoas da própria empresa, que não trabalham com desenvolvimento, para avaliar a sua facilidade de uso.

Podemos considerar que I e II representam:

(a) I - evidência confiável e conclusão correta; II - dados com amostragem de conveniência
(b) I - evidência anedotal e conclusão incorreta; II - dados com viés de seleção
(c) I - evidência confiável e conclusão provavelmente correta; II - dados com amostragem representativa
(d) I - evidência anedotal e conclusão incorreta; II - dados com amostragem baseada em agrupamento

Exercício 2)

Gostaríamos de obter e analisar dados disponíveis publicamente a partir de um repositório existente na Internet. Esses dados são referentes a indivíduos que contrataram serviços. Qual a primeira investigação ou procedimento a realizar ao obter esses dados?

(a) conhecer o espaço amostral dos dados e observar questões éticas como a privacidade dos respondentes
(b) realizar uma análise exploratória antes de qualquer análise
(c) auditar os dados, procurando por inconsistências como outliers
(d) inferir/treinar modelos diretamente a partir dos dados e medir sua acurácia

Exercício 3)

Uma empresa deseja entender melhor o potencial de mercado para um novo produto em um certo público alvo. Qual das alternativas abaixo representa a melhor forma de proceder após decidir que a coleta dos dados é necessária?

(a) permitir que os usuários enviem suas opiniões por meio de áudio ou vídeo, para depois coletar os dados a partir desse material
(b) implementar um questionário via aplicativo rapidamente em uma rede social popular e testá-lo massivamente, para verificar se os dados são consistentes com o que é esperado
(c) segmentar o público alvo e pedir ao menos 1 de cada segmento qual a chance, de 1 a 5 do indivíduo comprar esse produto
(d) especificar detalhes dos dados a serem coletados, planejar como obter uma amostra representativa do público alvo

Exercício 4)

Acesse o portal : http://catalogo.governoaberto.sp.gov.br/ (http://catalogo.governoaberto.sp.gov.br/)

Procure por duas fontes de dados e verifique o formato em que estão disponíveis

I - "Quantidade de alunos por tipo de ensino da rede estadual - 01/2019" (Secretaria da Educação - Sede)
II - "Pesquisa de Caracterização Socioeconômica do Usuário e seus Hábitos de Viagem - 2018" (Companhia do Metropolitano de São Paulo - Metrô)

Esses dados estão disponíveis no tipo:

(a) I e II são arquivos simples em dados estruturados
(b) I e II são dados estruturados disponíveis em sistema gerenciador de bancos de dados
(c) I são dados estruturados em arquivo simples, II dados não estruturados em arquivo binário
(d) I dados estruturados em arquivo binário, II são dados estruturados em arquivo texto

Exercício 5)

Baixe os dados relativo ao item I do exercício anterior, e carregue-a considere as particularidades do arquivo. Não carregue o cabeçalho (use `header=None`). Após carregar, remova as colunas 21 em diante, mantendo as colunas de 0 a 20.

Essas colunas restantes possuem significado de acordo com o "dicionário de dados" disponível ao visualizar o recurso dos dados. Sendo rotuladas da seguinte forma:

- CDREDE
- DE
- CODMUN
- MUN
- CATEG
- COD_ESC
- TIPOESC
- CODVINC
- NOMESC
- ENDESC
- NUMESC
- BAIESC
- EMAIL
- FONE 1
- ZONA
- ED_INFANTIL
- CLASSES ESPECIAIS
- SALA DE RECURSO
- ANOS INICIAIS
- ANOS FINAIS
- ENSINO MEDIO

Para isso, utilizamos: `dc.columns = ['CDREDE', 'DE', 'CODMUN', 'MUN', 'CATEG', 'COD_ESC', 'TIPOESC', 'CODVINC', 'NOMESC', 'ENDESC', 'NUMESC', 'BAIESC', 'EMAIL', 'FONE1', 'ZONA', 'ED_INFANTIL', 'CLASSES ESPECIAIS', 'SALA DE RECURSO', 'ANOS INICIAIS', 'ANOS FINAIS', 'ENSINO MEDIO']`

Quantas linhas/exemplos existem nessa base de dados e qual é o tipo das variáveis NOMESC e ENSINO MEDIO, respectivamente?

(a) 5366, object, int64
(b) 17366, category, int8
(c) 21, object, int64
(d) 5366, category, float64

Exercício 6)

Visualize os dados únicos e o histograma da variável SALA DE RECURSO.

Realize a discretização da variável utilizando o método do intervalo considerando os seguintes intervalos e rotulos

0 - '0'
1 a 4 - '1 a 4'
5 a 9 - '5 a 9'
10 ou mais - '10+'

Use o método cut() lembrando que os intervalos são definidos de forma que:

$$[a, b, c, d]$$

resulta em 3 intervalos:
(a, b] - entre a e b, não inclui a
(b, c] - entre b e c, não inclui b
(c, d] - entre c e d, não inclui c

Adicione essa nova variável na base, com o nome 'SALA_DE_RECURSO_D'

Responda, quantas linhas recaem em cada um dos 4 intervalos, respectivamente 0; 1 a 4; 5 a 9; e 10+?

(a) 93, 415, 446, 772
(b) 3636, 413, 470, 847
(c) 772, 446, 415, 93
(d) 3619, 411, 469, 846

Exercício 7)

Vamos normalizar 3 variáveis: ANOS INICIAIS, ANOS FINAIS e ENSINO MEDIO

A normalização utilizada será diferente para cada uma delas. Utilizaremos

- min-max para ANOS INICIAIS, com a=0, b=1
- norma L-∞ para ANOS FINAIS
- z-score para ENSINO MEDIO

Para isso, codifique funções que recebam uma coluna por parâmetro e retornem um atributo já normalizado

Depois, aplique as funções e crie novas variáveis com os atributos normalizados: INICIAIS_n, FINAIS_n, MEDIO_n.

Após normalização, quais os valores de média e mediana de cada um deles, considerando arrendondamento para 2 casas decimais?

(a) INICIAIS_n: 0.00, 0.00; FINAIS_n: 0.06, 0.00; MEDIO_n: 0.00, 0.17.
(b) INICIAIS_n: 0.07, 0.00; FINAIS_n: 0.16, 0.15; MEDIO_n: 0.00, -0.17.
(c) INICIAIS_n: 0.07, 0.00; FINAIS_n: 0.15, 0.16; MEDIO_n: 0.00, -0.17.
(d) INICIAIS_n: 0.00, 0.00; FINAIS_n: 0.00, 0.16; MEDIO_n: 0.00, 0.17.

Exercício 8)

Utilizando as variáveis normalizadas no exercício anterior, compute distâncias entre a escola de COD_ESC (atribuno na coluna de índice 5) cujo valor é 24648

e todas as escolas cujo código da rede (CDREDE) seja 20510, excluindo a de COD_ESC = 24648

Utilize a distancia Euclidiana.

Compare usando vetor de atributos formado por 'INICIAIS_n', 'FINAIS_n' e 'MEDIO_n'.

Observação: deve-se ter cuidado ao usar normalizações distintas como feito nesse exercício para comparar atributos, em particular considerando que o z-score produz valores negativos. Considere esse procedimento apenas a título de exercício com diferentes tipos de normalização e, na dúvida, utilize normalização uniforme entre os atributos.

Qual escola foi a mais próxima (NOMESC) e a respectiva distância (arredondada para 2 casas decimais)?

(a) MANOEL MARTINS, distância 1.6.
(b) EDDA CARDOZO DE SOUZA MARCUSSI, distância 0.18
(c) EDDA CARDOZO DE SOUZA MARCUSSI, distância 105.19
(d) MANOEL MARTINS, distância 0.18

Exercício 9)

Utilize os atributos 'ENSINO MEDIO', 'ANOS INICIAIS', 'ANOS FINAIS'. Vamos transformá-los por meio da função logaritmica. Para isso:

1. Faça uma cópia da base de dados, e atribua nulo (nan) a todos os valores iguais a zero nesses atributos.
2. Transforme esses atributos utilizando a operação logarítmica e os adicione à base de dados
3. Exiba a matriz de correlação de Pearson entre os atributos originais e os transformados.
4. Mostre o scatterplot entre 'ENSINO MEDIO' e 'ANOS FINAIS', e compare com log(ENSINO MÉDIO) e log(ANOS FINAIS)

Qual é a correlação, arredondada para duas casas decimais, entre os atributos transformados: log(ENSINO MÉDIO) e os outros dois atributos: log(ANOS INICIAIS) e log(ANOS FINAIS), respectivamente?

(a) 0.45 e 0.61
(b) 0.74 e 0.78
(c) -1 e 1
(d) 0.45 e 0.78

Exercício 10)

Codifique as variáveis categóricas 'SALA_DE_RECURSO_D' (categórica ordinal) e 'DE' (categórica nominal).

Para a primeira, use números inteiros sequenciais, iniciado por 0 para codificar a variável segundo sua ordenação, sendo que os dois últimos (os dos valores maiores na ordenação) devem ser mapeados para um único código. Gere um novo atributo 'SALA_DE_RECURSO_D_ord'.

Para a segunda, use números inteiros sequenciais, iniciados por 0 para codificar a variável em ordem alfabética, e gere um novo atributo 'DE_cod'

A seguir, use a função `value_counts()` para mostrar a quantidade de cada código e responda abaixo quais os valores dos dois códigos numéricos com maior contagem de valores para cada variável nova:

(a) DE_cod: 59 e 66; SALA_DE_RECURSO_D_ord: 0, 3
(b) DE_cod: 66 e 55; SALA_DE_RECURSO_D_ord: 1, 3
(c) DE_cod: 66 e 82; SALA_DE_RECURSO_D_ord: 0, 3
(d) DE_cod: 66 e 82; SALA_DE_RECURSO_D_ord: 0, 2