P1 de Análise Exploratória de Dados

Em cada questão desta prova, cabe a vocês usar as técnicas estudadas de Análise Exploratória para extrair conclusões a respeito de um particular tema. Em cada caso, um conjunto de dados específico contém as informações com base nas quais vocês deverão fazer suas análises. Supostamente as observações desse conjunto de dados cobrem toda a população (o universo) de interesse a ser investigada, ou constituem uma amostra representativa dessa população.

Um aspecto que será bastante valorizado na correção é a clareza com a qual vocês explicarão o que foi feito.

AVISOS IMPORTANTES

- Esta prova é para ser resolvida em grupo. Então espera-se que cada aluno do grupo participe de todas as etapas do trabalho. Ou seja, a ideia não é incumbir cada membro do grupo de resolver somente uma parte da tarefa total. A riqueza do aprendizado consiste em a equipe como um todo discutir a solução a ser apresentada.
- Todas as questões deverão ser resolvidas com o uso do software R, gratuito e de domínio público. Espera-se que os códigos utilizados sejam anexados ao seu trabalho, para maior clareza e para facilitar a correção.
- A sua prova deverá ser entregue pelo Google Classroom até as 12:00 da 3ª feira 07/06/2022, em formato PDF, além dos códigos em R utilizados.

QUESTÕES

1. Considerem o conjunto de dados “College” da library ISLR, cuja descrição está disponível no Classroom. Examinaremos o comportamento da variável “S.F. Ratio”, que informa o número de estudantes por professor em cada uma das 777 faculdades ali consideradas.

    a) Apresentem um gráfico que, a seu juízo, sintetize da melhor forma possível todas as conclusões relevantes que podem ser extraídas a respeito do comportamento dessa variável. Interpretem esse gráfico.
    b) Há, entre as faculdades do conjunto de dados, algumas onde o valor dessa variável possa ser considerado discrepante (outlier) em relação ao comportamento da maioria? Se houver, indiquem quais são essas faculdades e, para cada uma delas, explicitem se se trata de um valor anormalmente baixo ou anormalmente alto. Expliquem como vocês chegaram a essa conclusão. Se fosse possível fazer algum tipo de recomendação à administração dessas faculdades, qual seria a recomendação adequada?]
    c) O perfil de frequências dessa variável apresenta um comportamento simétrico em torno de um valor central? Caso a resposta seja negativa, que transformação de variável foi utilizada para promover uma maior simetria e qual foi o critério adotado para a escolha dessa transformação? Foi necessário subtrair ou somar previamente uma constante aos dados? Se foi necessário, qual foi o valor dessa constante, como ele foi escolhido, e por que foi importante fazer previamente essa subtração ou adição? Foi utilizado na sua análise um índice de simetria? Qual? Como ele foi usado? Incluam evidências convincentes de que a transformação escolhida foi de fato adequada.

2. Considerem o conjunto de dados “Wage” da library ISLR, cuja descrição está disponível para consulta no Classroom, e que contém informações relativas a 3000 profissionais. A intenção aqui é examinar até que ponto o nível educacional do profissional (medido pela variável “education”) afeta o seu salário (medido pela variável “wage”).

   a) Apresentem um gráfico que, a seu juízo, sintetize da melhor forma possível todas as conclusões relevantes que podem ser extraídas a respeito desse fenômeno. Interpretem esse gráfico. Como o nível educacional (variável qualitativa) afeta o nível salarial do profissional (variável quantitativa) e por que?
   b) Há homogeneidade entre as dispersões dos perfis de frequência do salário relativos aos diferentes níveis educacionais? Caso a resposta seja negativa, que transformação de variável foi utilizada para promover uma maior homogeneidade dessas dispersões e qual foi o critério adotado para a escolha dessa transformação? Foi necessário subtrair ou somar previamente uma constante aos dados? Se foi, qual foi o valor dessa constante, como ele foi escolhido, e por que foi importante fazer previamente essa subtração ou adição? Foi utilizado na sua análise um índice de homogeneidade? Qual? Como ele foi usado? Incluam evidências convincentes de que a transformação escolhida foi de fato adequada.
   c) Há, entre os profissionais de um mesmo nível educacional, alguns para os quais o salário possa ser considerado discrepante (outlier) em relação ao comportamento da maioria, naquele grupo específico? Se houver, indiquem o valor do salário de cada um desses profissionais e, para cada um deles, explicitem se se trata de um valor anormalmente baixo ou anormalmente alto. O conjunto de dados contém alguma informação adicional que possa explicar por que esses profissionais tem salários discrepantes entre os demais profissionais com aquele nível educacional? Qual seria essa informação adicional?

3. Considere o conjunto de dados “Auto” da library ISLR, cuja descrição está disponível para consulta no Classroom, e que contém informações relativas a 392 veículos. A intenção aqui é tentar obter uma expressão matemática simples para a forma como o desempenho “mpg” de um automóvel em galões por litro de combustível depende de sua potência “horse-power”.

   a) Exibam um gráfico que, a seu juízo, sintetize da melhor forma possível essa relação de dependência. Interpretem esse gráfico. Como a potência do carro influencia o seu desempenho e por que?
   b) Apresentem a equação que vocês obtiveram para expressar a relação matemática entre as duas variáveis originais ou entre transformações dessas variáveis. A relação matemática obtida lhes parece ser forte? Ou seja, até que ponto se pode concluir, com base na análise dos dados, que a potência do carro determina o seu desempenho?
   c) Escolham estrategicamente três pontos do diagrama de dispersão para servirem de base ao seu raciocínio, conforme foi explicado nas aulas gravadas sobre esse tema. Como e por que foi feita essa escolha dos três pontos? Os dados indicam que o desempenho do carro pode ser considerado uma função linear da sua potência? Por que? Caso a resposta seja negativa, que transformação de variável foi utilizada para promover uma maior linearidade na relação matemática entre as duas variáveis? Foi necessário subtrair ou somar previamente uma

constante aos dados relativos a alguma das duas variáveis? Se foi, qual foi a variável e qual foi o valor dessa constante? Como ele foi escolhido, e por que foi importante fazer previamente essa subtração ou adição? Só uma das duas variáveis foi transformada ou ambas foram transformadas? Foi utilizado na sua análise algum índice de linearidade? Qual? Como ele foi usado? Explique com clareza a solução obtida e inclua evidências convincentes de que ela foi de fato adequada.

4. Nesta questão o ponto de partida é um conjunto com n = 100 pares de dados (x_i, y_i), onde i = 1,2,3,...,100, que estão disponíveis no arquivo "dados.csv", para serem usados como input do seu programa.

   IMPORTANTE: Baixe o arquivo "dados.csv" anexado junto a esta prova no Classroom. Certifique-se que o arquivo está na mesma pasta em que o RStudio está aberto. Para carregá-lo em uma variável, basta rodar: *nome_da_variavel = read.csv("dados.csv")*

   Caso tenham alguma dificuldade para carregar o conjunto de dados, entre em contato com o monitor, que poderá esclarecer as dúvidas e resolver dificuldades computacionais.

   (a) Obtenha um diagrama de dispersão para esses pares de dados.
   (b) Parece existir uma relação linear do tipo $y_i = a + bx_i + e_i$ entre essas variáveis?
   (c) Caso a resposta em (a) seja negativa, selecione três pontos estratégicos A, B, C na figura tais que, e aplique a eles o método para linearização proposto nas aulas.
   (d) Usando como critério de linearização a minimização do índice $\left|\frac{\text{slope(AB)}-\text{slope(BC)}}{\text{slope(AB)}+\text{slope(BC)}}\right|$, qual é o valor de λ a ser escolhido (entre os mais comumente usados), para que, quando a transformação de variável $\varphi_\lambda(.)$ é aplicada à variável y, $\varphi_\lambda(y)$ passe a ser uma função linear de x? Explique como você chegou à solução. Qual é a expressão matemática da função $\varphi_\lambda(.)$?
   (e) Obtenha um diagrama de dispersão para os pares (x_i, $\varphi_\lambda(y_i)$), i = 1,2,3,...,100.
   (f) Ajuste, por mínimos quadrados, uma reta $\varphi_\lambda(y_i) = a + bx_i + e_i$ a esses 100 pares de dados. Quais são os valores de a e b?
   (g) Trace a reta de regressão no plano (x, $\varphi_\lambda(y)$).
   (h) Qual seria a sua estimativa do valor da variável y correspondente a x = 3? Por que?

Obs. Quem gera os dados do problema por simulação, pode induzir a que o problema tenha uma solução específica já escolhida de antemão.