UNIVERSIDADE REGIONAL DE BLUMENAU

Pós-Graduação em Data Science

Disciplina Técnicas Estatísticas Aplicadas a Predição

Prof. Moacir Manoel Rodrigues Junior

AVALIAÇÃO FINAL

Considere o cenário descrito e responsa o que for pedido. As respostas devem ser formuladas em um documento de texto e carregadas no AVA, na atividade específica da disciplina até a data limite. São critérios de avaliação:

- Exatidão dos resultados.
- Organização do Documento.
- Apresentar o resultado de todos os cálculos.
- Utilizar alguma linguagem para os cálculos que não o Excel, para as respostas colar o código utilizado para o cálculo e o resultado apresentado.
- Profundidade da Discussão Formulada.

Cenário

A planilha “Dados.xlsx” (ou “Dados.csv”) possui dados de um levantamento feito com 1289 pessoas dos Estados Unidos da América. Esta planilha está organizada com as seguintes variáveis:

Rótulo	O que mede
obs	Rótulo das Observações.
salario	Salário dos indivíduos entrevistados – em milhares de reais por ano.
sexo	Se for do sexo feminino recebe valor 1, se for do seco masculino recebe valor 0.
cor	Recebe 1 se a pessoa for não-branca, 0 caso contrário.
est_civil	Recebe 1 caso a pessoa for casada, 0 caso contrário.
instrucao	Anos de educação formal que a pessoa recebeu.
experiencia	Anos de experiência que a possui na área em que trabalha.
idade	Anos de vida que a pessoa entrevistada possui.

Considerando este contexto, faça o que se pede em cada uma das questões descritas a seguir.

Questão 1 (1,50) - Calcule as medidas de posição (Média, Mediana, Máximo, Mínimo, 1º Quartil e 3º quartil) para as variáveis “salario”, “instrucao", “experiência" e “idade”. Apresente os cálculos e faça uma interpretação dos resultados.

Questão 2 (1,50) – Calcule as medidas de dispersão (Amplitude, Desvio-Padrão, Variância, Coeficiente de Variação, Assimetria e Curtose) para as variáveis “salario”, “instrucao", “experiência" e “idade”. Responda o que segue:

a. Com relação ao Coeficiente de Variação, qual é a variável que possui maior discrepância em seus valores? E a com menor discrepância?

b. Qual deve ser a interpretação dada ao Coeficiente de Variação?
c. Considerando que as medidas de Assimetria e Curtose qualificam a média como boa medida de tendência central, existe alguma das variáveis que possua problemas de assimetria e/ou curtose? Justifique.

Questão 3 (1,50) – Considere uma análise que possa ser realizada sobre a variável salario. Faça os procedimentos destacados a seguir:

a. Calcule a média e a mediana do "salario" para mulheres e homens separadamente. Qual é a tendência apresentada para média e para mediana?
b. Calcule a média do "salario" para brancos e não brancos. Qual é a tendência apresentada para média e para mediana?

Questão 4 (1,00) – Faça um gráfico Box-Plot para as variáveis salario, instrucao, experiencia e idade e identifique se existem *outliers*. Quantas observações deveriam ser excluídas em cada variável por serem prováveis *outliers*?

Questão 5 (2,00) – Considerando os gráficos de dispersão, construa-os conforme pedido a seguir:

a. Faça um gráfico que relacione o "salario" com o tempo de "instrucao". Analise uma eventual tendência dos dados.
b. Faça um gráfico que relacione o "salario" com o tempo de "experiência". Analise uma eventual tendência dos dados.
c. Faça um gráfico que relacione o "salario" com a "idade". Analise uma eventual tendência dos dados.
d. Faça um gráfico que relacione a "experiencia" com o tempo de "instrucao". Analise uma eventual tendência dos dados.

Questão 6 (1,00) – Considerando as variáveis estritamente quantitativas. Construa um Histograma e identifique a variável com melhor ajuste percebido para a distribuição normal de probabilidade.

Questão 7 (1,50) – Considere que a variável "salario" segue uma distribuição normal de probabilidade. A média e o desvio-padrão já foram calculados. Assim determine o que se pede:

a. Qual a probabilidade estimada de uma pessoa ganhar mais do que o 3º quartil?
b. Qual a probabilidade estimada de uma pessoa ganhar menos do que o 1º quartil?
c. O que é mais provável, considerando a probabilidade estimada, a pessoa ganhar menos do que a média ou a pessoa ganhar menos do que a mediana?