GLOBAL SOLUTION - DATA SCIENCE

Tema: O Futuro do Trabalho em Dados e Inteligência Artificial
Base: https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries/data
Entrega: Link do notebook no **Google Colab** (um por grupo)
Formação: Grupos de até 3 alunos

Contexto
A área de Ciência de Dados é uma das que mais crescem no mundo, impulsionada pela Inteligência Artificial e pela transformação digital.
Compreender como variam os salários, cargos e tipos de contrato é essencial para planejar a carreira no futuro do trabalho.
Neste desafio, o grupo aplicará os conceitos de amostragem e estatística descritiva estudados ao longo do semestre, utilizando dados reais sobre profissionais de Data Science ao redor do mundo.

Instruções Gerais

- Todo o trabalho deve ser executado no Google Colab.
- As questões dissertativas devem ser respondidas de forma objetiva em células Markdown (texto explicativo logo abaixo do código).
- Cada questão deve conter:
  - O código Python executado;
  - O resultado exibido (tabela ou gráfico);
  - E a resposta dissertativa, clara e concisa.
- Envie o link do notebook Colab completo, com os nomes e RMs de todos os integrantes.

Etapas do Desafio

1. Leitura e compreensão da base

Carregue a base **Data Science Salaries 2024** e descreva:

- Quantidade de linhas e colunas;
- Variáveis principais (job_title, salary_in_usd, experience_level, employment_type, company_location);
- Tipos de dados de cada coluna.

Responda em Markdown: o que esse conjunto de dados nos permite entender sobre o mercado global de Data Science?

2. Qualidade dos dados

Verifique se há **valores nulos, duplicados ou inconsistentes**.
Exiba a contagem de nulos e a existência de duplicatas.
Responda: a base precisa de limpeza antes da análise?

3. Amostragem

Crie uma **amostra de 15%** do conjunto total, selecionada aleatoriamente.
Compare as estatísticas descritivas (média e mediana de salary_in_usd) entre a amostra e o dataset completo.
Responda: a amostra é representativa? Por quê?

4. Medidas de tendência central (salários)

Calcule **média, mediana e moda** da variável salary_in_usd.
Responda: qual dessas medidas descreve melhor o salário típico na área de dados?

5. Dispersão (salários)

Calcule **variância, desvio-padrão e coeficiente de variação (CV)** de salary_in_usd.
Responda: há grande desigualdade salarial entre profissionais da área?

6. Distribuição e forma

Crie um **histograma** e um **boxplot** para a variável salary_in_usd.
Responda: a distribuição é simétrica ou assimétrica? Existem outliers?

7. Segmentação por nível de experiência

Agrupe por experience_level e calcule:

- Média e mediana de salário;
- Contagem de profissionais por categoria.

Responda: quais diferenças salariais existem entre júnior, pleno e sênior?

8. Segmentação por tipo de contrato

Agrupe por employment_type e calcule as medidas de tendência central e dispersão.
Responda: há diferenças salariais entre contratos full-time, part-time e freelancer?
Qual tipo de vínculo parece oferecer maior estabilidade financeira?

9. Análise de outliers (salário)

Utilize o **método IQR (Tukey)** para identificar outliers de salary_in_usd.
Responda:
a) Quantos outliers foram encontrados?
b) Eles se concentram em algum nível de experiência ou cargo específico?
c) Esses valores parecem erros ou exceções reais de mercado?

10. Reflexão final – O futuro do trabalho em dados

Com base nos resultados obtidos, responda em até **10 linhas (Markdown)**:
O que os dados revelam sobre o cenário global de carreiras em Data Science?
Como experiência, cargo e tipo de contrato impactam na remuneração?
Quais competências parecem mais promissoras para o futuro profissional?

Critérios de Avaliação

Critério	**Peso**
Aplicação correta dos conceitos estatísticos	30%
Clareza e interpretação analítica	30%
Organização e legibilidade do notebook	20%
Reflexão final e coerência com o tema	20%