Section 1: Short Response Questions

1. Define and provide an example for each of the following data types: **nominal, ordinal, interval, and ratio**. Explain why distinguishing between these types is important in data analysis.
2. Describe the role of EDA in data science. Why is visualization crucial in EDA, and how does it help identify data issues?
3. Explain the bootstrapping method in statistics. How does it differ from traditional sampling techniques, and why is it useful for estimating confidence intervals?
4. Explain the difference between a **null hypothesis (H0)** and an **alternative hypothesis (HA)**. Provide an example of a research question and formulate appropriate H0 and HA statements for it.
5. What does a **p-value** represent in hypothesis testing? If a p-value is 0.03 in a study testing the effectiveness of a new drug, what does this imply at a 95% confidence level?
6. Define family-wise error and explain why it is a concern when conducting multiple hypothesis tests. Discuss one correction method used to control for family-wise error.
7. What are dataset cards, and why are they important in responsible data science practices? Identify three key components typically included in a dataset card.
8. Compare and contrast the **Binomial** and **Normal** distributions. Under what circumstances would each be used in statistical analysis?
9. Explain why correlation does not imply causation. Provide an example where two variables might be correlated but not have a causal relationship.
10. Suppose you are analyzing a dataset where you want to test whether students from two different schools have the same average SAT scores. Which statistical test would be most appropriate, and why?

Section 2: Problem-Solving Questions

Problem 1: Exploratory Data Analysis & Visualization

Dataset: Titanic Dataset - Kaggle

- Load the dataset and summarize the number of survivors and non-survivors.
- Create a boxplot comparing the distribution of ages between survivors and non-survivors.
- Interpret your results: What insights can you gather from these visualizations? What challenges might arise when working with this dataset?

Problem 2: Hypothesis Testing

Dataset: Palmer Penguins Dataset - seaborn

- Test whether the **mean body mass** differs between **Adelie and Chinstrap penguins** using a **two-sample t-test**.
- Report the p-value and interpret whether the difference is statistically significant.
- Discuss potential assumptions of the test and any limitations in the data that might affect your conclusions.

Problem 3: Multiple Hypothesis Testing and Family-Wise Error

Dataset: Iris Dataset - UCI

- Perform **three separate hypothesis tests** comparing different iris species on a selected feature (e.g., petal length).
- Explain the problem of **multiple hypothesis testing** and apply the **Bonferroni correction** to adjust for family-wise error.
- Discuss the implications of multiple testing and how it impacts statistical significance.

Problem 4: Confidence Intervals & Bootstrapping

Dataset: Gapminder Dataset - FiveThirtyEight

- Estimate the **mean life expectancy** for a specific continent using bootstrapping.
- Compute a **95% confidence interval** for the mean life expectancy.