Quiz 1c: Statistical Summaries and Visualization

Instructions

Answer all questions. Write code by hand as neatly as possible. Partial credit will be given for correct reasoning even if syntax isn't perfect.

Section A: Conceptual Questions

Question 1 (2 points)

You plan to have an LLM create an automated EDA function that takes a DataFrame and generates visualizations and summary statistics.

a) Describe what this function should do and what kinds of outputs it should produce.

b) Describe three unit tests you would ask the LLM to implement to verify the function works correctly.

Answer:

Question 2 (2 points)

You want to visualize the relationship between two numerical variables: `age` and `salary`.

a) What type of visualization would you use and why?

b) If you wanted to add a third categorical variable (like `department`) to this visualization, how could you incorporate it?

Answer:

Section B: Code Writing

For questions 3-6, assume you have already imported pandas as `pd` and seaborn as `sns`. Assume the data is already loaded into a variable called `df`.

Question 3 (3 points)

Write code to create a histogram of a column called `price` with 20 bins. Add a title "Distribution of Product Prices" and label the x-axis as "Price ($)".

Answer:

Question 4 (3 points)

Write code to create a scatter plot showing the relationship between `square_feet` (x-axis) and `price` (y-axis) for a DataFrame called `housing_df`. Add appropriate labels and a title.

Answer:

Question 5 (4 points)

Write a unit test called `test_price_is_positive` that checks whether all values in a `price` column are greater than zero. The test should pass if all `price` values are positive, and fail otherwise.

Answer: