Quiz 1b: Pandas Fundamentals

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're working with a dataset that has a column called `income` with many missing values (NaN). Your teammate suggests: "Let's just drop all rows with missing income values."

a) What's one potential problem with this approach?

b) What would you do during EDA to decide whether dropping these rows is a good idea?

Answer:

Question 2 (2 points)

Suppose your dataset's `income` column looked like this:

```
income
13547.78
78634.89
43548.18
106818.68
None
50618.85
...
```

What data type would you expect this column to be? Explain your answer.

Answer:

Section B: Code Writing

For questions 2-5, assume you have already imported pandas as `pd` and seaborn as `sns`.

Question 1 (2 points)

Write code to load a CSV file called `customers.csv` into a DataFrame called `df`. Then, show the first few rows of the data.

Answer:

Question 2 (2 points)

Write code to select only the rows where the `age` column is greater than 25 and the `city` column equals “Houston”.

Answer:

Question 3 (2 points)

Write code to calculate the mean and median of the age column, as well as the 99th percentile of all ages. Assume the column is numeric with no missing values.

Answer:

Question 4 (3 points)

You have a DataFrame called df with columns customer_id, name, and email. Write code to:

1. Select only the customer_id, customer_type and income columns
2. Filter to rows where the custom_type is "new"
3. Filter to rows where these new customers have an income above $100k
4. Display the first 5 rows of the result

Answer: