0.1 Question 3a

Consider the chained `pandas` statement below:

```
q3a_df = ins_named[ins_named["name"].str.lower().str.contains("taco")].groupby("bid").filter(lambda
sf: sf["score"].max() > 95).agg("count")
```

We can decompose this statement into three parts:

```
temp1 = ins_named[ins_named["name"].str.lower().str.contains("taco")]

temp2 = temp1.groupby("bid").filter(lambda sf: sf["score"].max() > 95)

q3a_df = temp2.agg("count")
```

For each line of code above, write one sentence describing what the line of code accomplishes. Feel free to create a cell to see what each line does. In total, you'll write three sentences.

Finally, write an example homework question whose answer is `q3a_df`.

- This example homework question should only be one sentence.

Note: While the first part of this question will be graded for correctness, the second part is a bit more open-ended. Answers that demonstrate correct understanding will receive full credit.

An example answer will look like the following: "`temp1` creates a … `temp2` transforms `temp1` by … Finally, `q3a_df` results in a `DataFrame` that … A question that is answered by this chain of operations is …"

`temp1` creates a `DataFrame` that includes only rows from `ins_named` where the `name` column contains the word 'taco', ignoring case sensitivity.

`temp2` tranforms `temp1` by grouping the data by `bid`, filtering out groups where the highest score is less than or equal to 95, and keeping all the rows for a `bid` if any row in that group has a score greater than 95, even if other rows within the same `bid` group have scores below 95.

`q3a_df` results in a `DataFrame` that counts the number of non-null values in each column from `temp2`, summarizing the total records that exist for businesses that meet the criteria above.

A question that is answered by this chain of operations is “How many total inspection records exist for businesses with ‘taco’ in their name that have at least one inspection score greater than 95?”

0.2 Question 3b

Consider `ins_named`, `temp1`, `temp2`, and `q3a_df` from the previous problem. What is the granularity of each `DataFrame`? Explain your answer in no more than four sentences.

Note: For more details on what the granularity of a `DataFrame` means, feel free to check the course notes!

The granularity of `ins_named` is fine-grained data because each row represents an individual inspection record, which contains information such as the restaurant's name, address, and inspection score.

The granularity of `temp1` is also fine-grained data because it is a filtered version of `ins_named`, still containing individual inspection records but only for businesses whose names contain the word 'taco'.

The granularity of `temp2` is also fine-grained because each row still represents an individual inspection record, but has now been filtered out to contain only the businesses that have at least one inspection score greater than 95.

The granularity of `q3a_df` is coarse-grained data because it aggregates the inspection records by counting the number of remaining records, no longer looking at individual inspection information.

0.3 Question 4e

Do you notice any trends? Are your results consistent with your prior knowledge about restaurants that receive high or low health inspection scores? Answer in the cell below.

This question is graded on effort, there is no one “correct” answer.

I notice a trend that the lower a restaurant’s initial inspection score, the higher the likelihood of a reinspection within 62 days. In contrast, restaurants with higher scores have a lower proportion of reinspection within 62 days. This is consistent with my prior knowledge about restaurants that receive high or lower health inspection scores because restaurants with lower scores are likely being monitored more closely due to potential health code violations, leading to follow-up visits to ensure compliance. Higher-scoring restaurants are less frequently reinspected since they already meet the necessary health and safety standards.