Data Challenge

Purpose: This exercise is to assess your ability in simple python programming. Goal is to create a table populated with randomized data using the programming language *Python*. You will need to visualize and make inferences about your dataset using any visual tool of your choice (i.e., Matplotlib, SQL, TABLEAU, etc). Many current day search engines use algorithms and machine learning models that are trained with large amounts of data, which are labeled by real humans (a.k.a. raters). The performance of these raters are crucial to the success of the algorithm, given that rater precision, recall, and accuracy on labeling data are met with engineering standards at all times. Through the creation of this dataset, we would like to see your approaches in determining how efficient the dataset is and whether it meets engineering standards. Assume the benchmark for a good dataset is above 90% agreement with the engineers and above 90% precision /

Extra context:

- This is a supervised machine learning problem.
- Engineers do not have the capacity to label data themselves, so they often rely on others who do have the time to label their data
- Raters are typically provided with guidelines, which are used as instructions for a rater to arrive at a certain label answer
- Rater volume, high agreement, precision, and recall is important for our dataset
- Your goal is to assess the trends, patterns, performance of the raters, and the quality of the created dataset
- There is no single approach or solution, we want to see you think through the problem

TLDR; Determine the overall quality of your randomly generated dataset and overall rater performance.

Getting Started:

Step 1: Generate the columns provided below as part of your table schema. Your table will need 10,000 rows.

- The text left of the double colon (::) will be the name of your column.
- On the right of the double colon (::) will be the data that is needed to populate your column.

***Ensure that your data is completely **RANDOMIZED** within each of your columns.

- Date Column :: Populate this column with dates between 10/1/05 – 10/30/05
- Rater Column :: Populate this column with Rater IDs denoted as a single letter (A, B, C, D, E)
- Correct Answers 3 Label :: Populate column with one of the labels: Low, Average, High
- Correct Answer 5 Label :: Populate column with one of the labels: Bad, Okay, Intermediate, Great, Exceptional
- Rater Answers 3 Label :: Populate column with one of the labels: Low, Average, High
- Rater Answers 5 Label :: Populate column with one of the labels: Bad, Okay, Intermediate, Great, Exceptional
- Task ID: Generate numbers 1-10000

Definitions of Columns:
Date Column – The date of when a rater has provided their label
Rater Column – Each rater is referenced by an ID. (e.g., Rater A, Rater B)
Correct Answer 3 Label – In this field, the engineer has chosen 1 of 3 choices as the correct label

Correct Answer 5 Label – In this field, the engineer has chosen 1 of 5 choices as the correct label
Rater Answer 3 Label – In this field, the rater has chosen 1 of 3 choices as what they believe the label should be
Rater Answer 5 Label – In this field, the rater has chosen 1 of 5 choices as what they believe the label should be
Task ID: Each data point that is labeled will be referenced with an ID ranging from (1-10000).

Step 2: Create 2 extra columns to determine whether the rater and engineer agreed on a given Task ID. One column for the 3-label agreement and one column for the 5-label agreement.

End Result Example: 9 columns x 10000 rows

	A	B	C	D	E	F	G	H	I
1	**Date**	**Rater**	**Correct Answer 3 Label**	**Correct Answer 5 Label**	**Rater Answer 3 Label**	**Rater Answer 5 Label**	**Task ID**	-	-
2	-	-	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-	-	-
5	-	-	-	-	-	-	-	-	-

Visualization – Given the dataset above, we would like for you to visualize and make inferences about your dataset, considering these metrics:

- What is the agreement rate between the engineer and all the raters for each day?
- What is the agreement rate between the engineer and all the raters for each week?
- Identify raters that have the highest agreement rates with the engineer.
- Identify raters that have the lowest agreement rates with the engineer.
- Identify raters that have completed the most Task IDs.
- Identify raters that have completed the least Task IDs.
- What is the precision for each of the 5 labels?
- What is the recall for each of the 5 labels?
- What is the precision for each of the 3 labels?
- What is the recall for each of the 3 labels?
- What is the overall agreement rate considering that the raters have to be in agreement with both the engineer's 3-label answer *and* the engineer's 5-label answer.

Step 4:
Pick 7 of the questions above. Visualize/calculate your answers. Given your answer, what approaches do you recommend you need to take to improve your metrics, if the metric has not met engineering standards? (Note: There is no single answer or approach to these problems)

Questions worth considering:

- What can you do to improve agreement rates overtime?
- How do you improve precision of a label overtime?
- What changes are needed or required to improve your dataset to achieve over 90% agreement, precision, or recall?
- Why do some raters perform better than others?

Step 5:
Identify 3 more potential questions to consider that can be used to identify issues among raters.

Step 6 (Optional): Assume that your dataset is complete and is uploaded into a SQL Server with the table name rater_data:

- Write a SQL query that outputs the following from the table:
    - Agreement Rates for Each Rater on October 6