UNIVERSITY OF SAN CARLOS OF GUATEMALA
FACULTY OF ENGINEERING
SCHOOL OF GRADUATE STUDIES

Task 5

Description

They have been shared with you (click here) to download) four of the most common data sets used by the **Stack Overflow** platform (the largest online community of developers).

- **questions.csv:** Contains a question ID and the question score based on how many times it has been upvoted; the data only includes R-based questions.

- **answers.csv:** Contains an answer identifier, the score, and an ID that links the answer to a specific question. • **tags.csv:** Contains a tag identifier and tag name, which can be used to identify the topic of each question, such as ggplot2 or dplyr. • **question_tags.csv:** Contains a tag identifier for each question and the question ID.

Load each dataset and name it accordingly.

1. Left-joining questions and tags

Use left_joins in this exercise to ensure that all questions are kept, even those without a corresponding tag.

1.1. Relate questions and question_tags using the id and question_id columns, respectively.

1.2. Add one more relationship for the tags table.
1.3. Use replace_na to change the NAs in the tag_name column to "only-r".
1.4. Finally, store the result in the variable questions_with_tags.

2. Comparing scores across tags

Conduct a brief analysis using verbs from the dplyr family such as group by, summarize, and arrange , and find out the average score for the most frequently asked questions.

2.1. Use questions_with_tags and apply group_by for the tag_name variable.
2.2. Apply summarize to obtain the average score for each question and name it mean_score.

2.3. Sort mean_score in descending order.

3. Finding gaps between questions and answers

Now we'll match questions with answers. Be sure to explore the tables and their columns in the console before starting the exercise.

3.1. Use inner_join to join the questions and answers tables, then apply the suffixes "_question" and "_answer" respectively.
3.2. Add a new column using the mutate function. The new column will be called gap and will contain the difference of creation_date_answer and creation_date_question. (creation_date_answer - creation_date_question).

4. Joining question and answer counts

We can also determine how many questions actually yield answers.
If we count the number of responses for each question, we can match the response counts to the question table.

4.1. Count and sort the question_id column in the answers table , then store the result in the variable answer_counts.
4.2. Relate the questions table to answer_counts (use left_join).
4.3. Replace the NA values in column n with zeros.
4.4. Finally, store the result in the variable question_answer_counts

5. Joining questions, answers, and tags

Let's identify which R topics generate the most interest on Stack Overflow.

5.1. Combine question_tags with question_answer_counts using inner_join.
5.2. Now, use another inner_join to add the tags table.

Assessment

- The correct execution of the code will be evaluated, ensuring that the inner_join(), left_join() and right_join() functions are applied correctly.
- The data sets will be assessed to ensure they contain correctly structured information and are used appropriately in joins.
- Clarity and precision in the interpretation of the different types will be valued. of joins and their applicability in data analysis.

Delivery

• Develop each of the sections in an R Notebook.

• Upload the notebook named **Assignment_5.Rmd** to the virtual classroom. • Submit by Friday, **April 4th, at 11:59 PM.**