Tarea 5

Descripción

Se le han compartido cuatro de los conjuntos de datos más comunes que utiliza la plataforma **Stack Overflow** (comunidad en línea más grande de desarrolladores).

- **questions.csv**: Posee un identificador de pregunta, la puntuación de la pregunta en función de cuántas veces se ha votado a favor de la misma; los datos solo incluyen preguntas basadas en R.
- **answers.csv**: Posee un identificador de respuesta, la puntuación y un ID que relaciona la respuesta con una pregunta específica.
- **tags.csv**: Posee un identificador de etiqueta y el nombre de la etiqueta, que se pueden utilizar para identificar el tema de cada pregunta, como ggplot2 o dplyr.
- **question_tags.csv**: Posee un identificador de etiqueta para cada pregunta y el ID de la pregunta.

Cargue cada set de datos y nombrelo según cada archivo.

1. Left-joining questions and tags

Utilice *left_joins* en este ejercicio para asegurarse de mantener todas las preguntas, incluso aquellas sin un tag correspondiente.

1.1. Relacione *questions* y *question_tags* usando las columnas *id* y *question_id*, respectivamente.
1.2. Agregue una relación más para la tabla *tags*.
1.3. Utilice *replace_na* para cambiar los NA en la columna *tag_name* a "only-r".
1.4. Por último, almacene el resultado en la variable *questions_with_tags*.

2. Comparing scores across tags

Realice un breve análisis, para ello utilice verbos de la familia *dplyr* como *group* by, *summarize*, *arrange* y averigue el score promedio de las preguntas más frecuentes.

2.1. Utilice *questions_with_tags* y aplique *group_by* para la variable *tag_name*.
2.2. Aplique *summarize* para obtener el score promedio de cada pregunta y asígnele el nombre *mean_score*.
2.3. Ordene *mean_score* en forma descendente.

3. Finding gaps between questions and answers

Ahora uniremos *questions* con *answers.* Asegúrese de explorar las tablas y sus columnas en la consola antes de comenzar el ejercicio.

3.1. Utilice *inner_join* para combinar las tablas *questions* y *answers*, luego aplique los sufijos "_question" y "_answer", respectivamente.
3.2. Agregue una nueva columna utilizando la función *mutate*. La nueva columna se llamará *gap* y contendrá la resta de *creation_date_answer* y *creation_date_question.* (creation_date_answer - creation_date_question).

4. Joining question and answer counts

También podemos determinar cuántas preguntas realmente arrojan respuestas. Si contamos el número de respuestas para cada pregunta, podemos unir los conteos de respuestas con la tabla de preguntas.

4.1. Cuente y ordene la columna *question_id* en la tabla de *answers*, luego almacene el resultado en la variable *answer_counts*.
4.2. Relacione la tabla *questions* con answer_counts (utilice *left_join*).
4.3. Reemplace los valores NA en la columna n con ceros.
4.4. Por último almacene el resultado en la variable *question_answer_counts*

5. Joining questions, answers, and tags

Identifiquemos qué temas de R generan más interés en Stack Overflow.

5.1. Combine *question_tags* con *question_answer_counts* usando *inner_join*.
5.2. Ahora, use otro *inner_join* para agregar la tabla *tags*.

Entrega

- Desarrollar un notebook en R con el nombre **tarea_5_r** según lo solicitado anteriormente.
- Cargue el cuaderno a Github.
- Enviar el enlace de Github a través del aula virtual.
- Se envía a más tardar el **30-julio hasta las 23:59**