Tarea 4

Descripción

En esta tarea, aplicarás técnicas avanzadas de manipulación de datos en R utilizando el paquete `dplyr`. Trabajarás con un conjunto de datos de **Uber** que contiene información sobre viajes, tarifas, ubicaciones y pasajeros.

Descargue el archivo `uber.csv` (utilice la función `read_csv`). Las variables a utilizar son las siguientes:

- `key`. Identificador único para cada viaje.
- `fare_amount`. El costo de cada viaje en USD.
- `pickup_datetime`. Fecha y hora en que se realizó el viaje.
- `passenger_count`. Número de pasajeros en el vehículo.
- `pickup_longitude`. Longitud donde se activó el viaje.
- `pickup_latitude`. Latitud donde se activó el viaje.
- `dropoff_longitude`. Longitud donde se desactivó el viaje.
- `dropoff_latitude`. Latitud donde se desactivó el viaje.

Parte I. Carga y exploración de datos

- Importa el conjunto de datos en un `data.frame` y examina su estructura con `glimpse()` y `summary()`.
- Convierte la columna `pickup_datetime` a formato `POSIXct` para facilitar su manipulación temporal.

Parte II. Filtrado de datos

Usa `filter()` para seleccionar solo los viajes:

- Con tarifa mayor a `$10` y al menos `2` pasajeros.
- Realizados en horario nocturno (entre las `22:00` y las `05:00` horas).

Parte III. Creación y modificación de variables

Usa `mutate()` para:

- Extraer la hora del día y el día de la semana de `pickup_datetime`.
- Calcular la distancia euclidiana aproximada entre el punto de recogida y el de destino con la fórmula:

$$distance = \sqrt{(dropoff\_longitude - pickup\_longitude)^2 + (dropoff\_latitude - pickup\_latitude)^2}$$

- Usa `transmute()` para mostrar solo las columnas `key`, `fare_amount`, `passenger_count` y la nueva variable de `distance`.

Parte IV. Selección de variables con `helpers`

Utiliza `select()` y los `'helpers'` para seleccionar columnas específicas:

- `starts_with("pickup")` para obtener las variables de ubicación de inicio del viaje.
- `ends_with("latitude")` para obtener solo las coordenadas de latitud.
- `contains("amount")` para seleccionar todas las columnas que incluyen información de tarifas.

Parte V. Agrupamiento y resumen de datos

- Usa `group_by()` para agrupar los viajes según el número de pasajeros.
- Luego, usa `summarize()` para calcular:
  - La tarifa promedio por número de pasajeros.
  - La distancia promedio recorrida por grupo de pasajeros.
- Desagrupar con `ungroup()` después del resumen.

Parte VI. Conteo y filtrado de datos agregados

- Usa `count()` para determinar cuántos viajes ocurrieron por cada día de la semana.
- Aplica `top_n(3, n)` para mostrar los tres días con mayor cantidad de viajes.

Parte VII. Renombramiento de columnas

Usa `rename()` para cambiar los nombres de las columnas:

- `fare_amount → tarifa_usd`
- `passenger_count → num_pasajeros`

Evaluación

- Aplicación de filtros avanzados con `filter()`.
- Creación y manipulación de variables con `mutate()` y `transmute()`.
- Selección efectiva de columnas usando `select()` con `'helpers'`.
- Agrupación, resumen y desagrupación de datos con `group_by()`, `summarize()` y `ungroup()`.
- Aplicación de `count()` y `top_n()` para análisis de frecuencia.
- Uso correcto de `rename()` para mejorar la claridad de los datos.

Entrega

- Desarrollar cada uno de los incisos en un Notebook de R.
- Cargar el notebook con el nombre **Assignment_4.Rmd** al aula virtual.
- Se envía a más tardar el viernes **28-marzo hasta las 23:59**.