Fundamentos lenguajes:
Práctica 3

12 de Noviembre de 2018

El conjunto de datos `titanic.csv` contiene información sobre los pasajeros del barco. Este conjunto de datos se ha utilizado para tratar de predecir la supervivencia de un pasajero en base a otra serie de variables como edad, sexo, o la clase del billete. Ver por ejemplo: `https://www.kaggle.com/c/titanic`. Cada una de las variables del fichero contiene la siguiente información:

- `survival`: Supervivencia (0 = No; 1 = Yes)
- `pclass`: Clase de pasajero (1, 2, 3)
- `name`: Nombre
- `sex`: Sexo
- `age`: Edad
- `sibsp`: Número de hermanos/esposos/as a bordo.
- `parch`: Número de padres/hijos a bordo
- `ticket`: Número de ticket
- `fare`: Coste del billete
- `cabin`: Cabina
- `embarked`: Puerto de embarque

Con el fichero de datos anterior:

1. (4 puntos) Leer el fichero **titanic.csv** como un dataframe.
2. (2 puntos) Calcular el porcentaje de pasajeros que sobrevivió.
3. (4 puntos) Calcular el porcentaje de missing values en cada uno de los atributos.

4. (2 puntos) Eliminar la variable `Cabin` del dataframe.

5. (8 puntos) Crear una nueva variable `Title` a partir de `Name` con los valores `Master` (hombre soltero), `Miss` (mujer soltera), `Mr.` (hombre casado), `Mrs.` (mujer casada) y `Otro` a partir de la variable nombre. Es importante tener en cuenta que el título `Miss` está en ocasiones codificado con su abreviatura en frances `Mlle` (*mademoiselle*) y lo mismo ocurre con `Mrs.`, que en ocasiones aparece como `Ms.` ó `Mme` (*madame*).

6. (4 puntos) Explorar la relación entre las variables `Age` y la nueva variable `Title` mediante un `boxplot` para cada uno de los valores de la misma. ¿Tienen alguna relación?.

7. (4 puntos) Explorar la relación entre `Age`, `Pclass` y `Title` en varios gráficos de dispersión con colores, donde el color representa la supervivencia (Pista: usar facetas).

8. (8 puntos) Completar los *missing values* del atributo `Age` con la mediana del resto de datos de esa variable pero agrupado de acuerdo a las variables `Pclass` y `Title`.

9. (2 puntos) Después de realizar las operaciones anteriores, eliminar ahora cualquier fila que tenga al menos un NA.

10. (2 puntos) Calcular la probabilidad de supervivencia en base al género (`Sex`). ¿Qué conclusión(es) obtienes del resultado?

11. (2 puntos) Calcular la probabilidad de supervivencia en base a la edad (`Age`). ¿Te parecen fácilmente interpretables estos resultados?

12. (4 puntos) Crea una nueva variable `Decade` en el dataframe que contenga la década de la edad de los pasajeros y repite el análisis del apartado anterior sobre esta nueva variable. ¿Qué conclusión(es) obtienes del resultado? Pista: función `cut`.

13. (4 puntos) Convertir la variable `Survived` a un factor con los niveles `Yes` si ha sobrevivido y `No` en caso contrario.

14. (4 puntos) Ver la relación entre la supervivencia y la variable `Title` con un gráfico de barras. En el caso del valor `Otros` de la variable `Title`, ¿nos proporciona este alguna información sobre la supervivencia?. ¿A qué se debe?.

15. (4 puntos) Crea dos nuevas variables en el dataframe con la siguiente información:

    - `Familysize`: número total de parientes incluyendo al propio pasajero.

- `Sigleton`: valor lógico indicando con valor TRUE si el pasajero viaja solo y FALSE en caso contrario.

16. (4 puntos) Realizar un gráfico de puntos de la variable `Age` sobre `Fare`, coloreado por los valores de la variable `Survived`.

17. (2 puntos) Realizar un histograma para ver la distribución de las edades.

18. (4 puntos) Representar en un gráfico de barras el número de pasajeros que han sobrevivido para cada uno de los valores de las variables `Sex` y `Pclass`.

19. (4 puntos) Cuenta el número de pasajeros por tamaño de familia y clase. Por ejemplo, cuántos pasajeros de primera clase pertenecen a una familia de tamaño 4. El resultado debe ser un dataframe con la información para todas las posibles combinaciones de clase del billete y tamaño de familia.

20. (4 puntos) Representar, en un mismo gráfico, dos histogramas de la variable `Age`, uno para los pasajeros con sexo masculino y otro para los pasajeros con sexo femenino. En caso de que se solapen los histogramas, usar colores con transparencias.

21. (4 puntos) Leer el fichero **titanic2.csv**, que contiene información adicional sobre los pasajeros del barco:

    - boat: identificador del bote salvavidas
    - body: identificador del cuerpo
    - home.dest: Origen/destino

22. (4 puntos) Para unificar estos dos dataframes, parecería buena opción utilizar la variable name como clave. Determina si esta variable es única por pasajero, mostrando el número de nombres diferentes repetidos. En caso de existir varios pasajeros con el mismo nombre, listar aquellas filas del dataframe inicial en las que el nombre del pasajero esté repetido

23. (6 puntos) Combina ambos dataframes utilizando la combinación del nombre y el número de billete, manteniendo las mismas filas que el dataframe original.

24. (4 puntos) ¿Qué porcentaje de los pasajeros que sobrevivió tiene asociado un identificador del bote salvavidas?

25. (6 puntos) Separar el conjunto anterior de datos en dos subconjuntos disjuntos de forma aleatoria, el primero conteniendo un 70% de los datos y el segundo un 30%. Los resultados tienen que estar contenidos en dos dataframes.

Entrega La entrega se realizará se realizará a través del Moodle en un único fichero .R, con fecha límite el sábado 19 de enero a las 23:59 (19/01/2019). Incluir comentarios en el código siempre que se considere necesario. Las respuestas planteadas a las preguntas deben responderse como comentarios en el fichero .R después de la línea de código correspondiente.

Criterios de evaluación La práctica se evalúa sobre 100 puntos. Para resolver los ejercicios se pueden utilizar indistintamente funciones de R base o de paquetes adicionales. Es conveniente (y se valorará) utilizar un estilo de programación adecuado. Algunas directrices pueden encontrarse en la Guía de estilo de `http://style.tidyverse.org/`. Además del estilo, se valorará que el código R sea:

- Correcto
- Claro
- Conciso
- General