Práctica de Programación en R

Fecha de entrega: 8 de noviembre

Ejercicio 1 (4 puntos)

Junto con la práctica se entrega un fichero con un *dataset (***udemy_courses.csv***)* que contiene información sobre cursos de la plataforma de aprendizaje online Udemy. Las variables del *dataset* son las siguientes:

- **course_id:** identificador del curso.
- **course_title:** título del curso.
- **url:** URL del curso.
- **is_paid:** variable booleana que indica si el curso es de pago o gratuito.
- **price:** precio del curso.
- **num_subscribers:** número de estudiantes del curso.
- **num_reviews:** número de ratings recibido por el curso.
- **num_ lectures:** número de clases del curso.
- **level:** nivel requerido para el curso.
- **content_duration:** duración del curso en horas.
- **published_timestamp:** fecha de publicación del curso.
- **subject:** temática del curso.

1. Lee el fichero y asígnalo a una variable.
2. ¿De qué clase es el objeto?
3. ¿Cómo se pueden ver el tipo de cada columna y una muestra de ejemplos?
4. Muestra los primeros 18 registros del *dataset*.
5. Muestra los últimos 25 registros del *dataset*.
6. ¿Cuáles son las dimensiones del *dataset*?
7. ¿Cuáles son los nombres de las variables del *dataset*?
8. Comprueba si alguna de las variables contiene NAs.
9. Las variables *level* y *subject* deberían ser categóricas. Crea un factor con etiquetas para dicha columna y asígnalo a la columna de nuevo. Es decir, el factor deberá tener tantas categorías como valores tiene la variable y cada una de esas categorías debe tener una etiqueta.
10. La variable *is_paid* es variable booleana. Conviértela a variable booleana y asígnala a la propia columna.
11. Calcula la media de la columna *num_subscribers*.
12. Guarda en un vector (media) la media de las columnas *num_subscribers* y *num_lectures*. Es decir, el vector "media" deberá tener 2 elementos: el primero conteniendo la media de la columna *num_subscribers* y el segundo con la media de la columna *num_lectures*.
13. ¿Qué variables son numéricas? **PISTA:** utiliza *sapply* junto con la función is.numeric.
14. Utilizando el resultado anterior, selecciona aquellas columnas numéricas y calcula la media de aquellas en las que tenga sentido.
15. Selecciona las 30 primeras filas y todas las columnas menos las tres últimas (**sólo con índices positivos**).
16. Selecciona las 30 primeras filas y todas las columnas menos las tres últimas (**sólo con índices negativos**).
17. Obtén los cuartiles de la variable *price*.
18. Obtén los deciles de la variable *price*.
19. Obtén los estadísticos básicos de todas las variables en un solo comando.
20. ¿Cuántos títulos de cursos distintos aparecen en el *dataset?*
21. ¿Cuántos registros tienen más de 1000 *reviews*?
22. Ordena de mayor a menor los 100 primeros elementos de la variable *course_id*.
23. Ordena el *dataset* por la variable *num_lectures* de manera ascendente. Inspecciona los primeros resultados para comprobar que se ha ordenado como se pide.
24. Obtén los índices de los registros para los que el valor de la variable *num_reviews* es superior a la mediana.
25. ¿Cuántos cursos existen para el nivel "*Intermediate Level*"?
26. ¿Qué curso tiene el mayor número de estudiantes? ¿Y el menor?
27. ¿Qué cursos son gratuitos?
28. Comprueba utilizando el *boxplot* si la variable *num_reviews* tiene *outliers*.
29. Pinta un histograma de la variable *price*.
30. Crea una función (cheap_expensive) que reciba dos parámetros. Si la temperatura es inferior al límite debe devolver "CHEAP", si es superior "EXPENSIVE". Comprueba el funcionamiento de función que acabas de crear invocándola con algunos valores de prueba. Los parámetros son:
    a. price: la temperatura.
    b. threshold: el límite para distinguir entre un retorno y otro. El valor por defecto es 6000.

31. Mediante una llamada a una de las funciones apply aplica la función anterior a toda la columna *price_detail_amount*. El resultado obtenido debe ser almacenado en una nueva variable del *data frame* llamada *cheap_expensive.*
32. Repite el ejercicio anterior usando el paquete *PURRR*.

Ejercicio 2 (4 puntos)

Preparación (*tidyr*, *reshape2*)

1. A partir de las variables *course_id* y *course_title*, crea una nueva *course_id_title* que sea la concatenación de ambas mediante un guion bajo. No borres las columnas *id* y *title* en el proceso.
2. Sobreescribe el valor de las columnas *published_timestamp* por una columna de tipo fecha.
3. Crea un *long dataset* que contenga las siguientes variables:
    - **course_id:** identificador único de cada curso.
    - **course_title:** título de cada curso.
    - **VARIABLE:** que contiene la lista de variables: *num_subscribers, num_reviews* y *num_lectures*.
    - **VALUE:** los diferentes valores.
4. Realiza el proceso inverso desde el *long dataset* anterior para obtener el *dataset* original con las siguientes columnas:
    - **course_id**
    - **course_title**
    - **num_subscribers**
    - **num_reviews**
    - **num_lectures**

Resuelve el ejercicio 2.3 y 2.4 empleando los paquetes vistos en clase: *tidyr* y *reshape2*.

Manipulación (*dplyr, data.table*)

1. Calcula el precio medio de los cursos dependiendo de su temática y ordena el resultado según el precio medio.
2. Calcula la duración máxima y mínima dependiendo de si un curso es gratuito o de pago.
3. Calcula el número de cursos publicado cada año.
4. Cuál es la temática que tiene el mayor número de clases medias.
5. Restringiéndonos a los cursos lanzados en 2016, ¿qué temática cuenta con más horas de clase?
6. Para todos los cursos posteriores a 2015, calcula las horas del curso más largo, del más corto y el número de estudiantes medio.

Resuelve el ejercicio empleando los paquetes vistos en clase: *dplyr* y *data.table*.

Ejercicio 3 (2 puntos)

Busca un *dataset* de tu interés, descárgalo y utilizando lo aprendido en clase haz un análisis a tu elección que te permita descubrir algún *insight*.

Trata de que el procesamiento y limpieza de los datos tenga cierta ***dificultad***. Se valorará positivamente tanto el procesamiento como el interés de éstos.

El resultado del ejercicio debe ser reproducible mediante ***R Markdown,*** ocupando un máximo de dos páginas. Debes entregar tanto el código Markdown como la salida de éste.

La entrega de la práctica será individual. Para cada ejercicio entrega un fichero R con el código para realizar cada una de las operaciones. Ten en cuenta que pueden existir diversas formas de realizar los ejercicios. Siempre que se consiga el objetivo se considerará que la respuesta es válida. Ambos ficheros se comprimirán dentro de un fichero zip con tu nombre. Por ejemplo: ***jose_miguel_morella_r.zip***