Taller de Aplicaciones en R y Python

Prof. Charles Thraves

Pregunta I – Elementos básicos de programación en R

Se pide realizar las siguientes funciones en **R**.

a) Una ecuación cuadrática se puede escribir como:

$$ax^2 + bx + c = 0,$$

en donde x es la incógnita. Escriba una función que reciba como argumento los tres coeficientes de dicha ecuación cuadrática, es decir, a, b, y c; y entregue una lista con la(s) raíce(s) (es decir, con las soluciones existentes de la ecuación). En caso de que existan dos soluciones, entre un vector con dos componentes. En caso de que la raíz sea única, la salida debe ser un vector de largo uno. En caso de que no existan raíces, devuelva un vector de largo cero. Para saber el número de raíces, se tiene que

$$\text{si } b^2 - 4ac > 0 \quad \text{hay dos raices}$$
$$\text{si } b^2 - 4ac = 0 \quad \text{hay una raíz}$$
$$\text{si } b^2 - 4ac < 0 \quad \text{no hya raices}$$

b) Haga una función que recibe un número como argumento, tal que devuelve “primo” si es que el número es primo, “no” si es que el número no es primo, y “no es un número” si es que el valor ingresado no es de tipo numérico. Para determinar si el número es primo use los comandos “while”, “for” y/o “break”. Recuerde que para verificar si una variable es numérica, puede usar la función “is.numeric()”. Hint: un número es primo si no es divisible por ningún número entero entre 2 y el número menos uno. Además, un número es divisible por otro si su resto es 0. Por ende, se para ver si un número es primo, se puede probar si el número es divisible por alguno de los números entre 2 y el número menos uno.

Pregunta II – Análisis de datos con R

En esta pregunta utilice **R** para analizar el dataset “**simce8b2019_rbd.xlsx**”. La explicación de cada columna se encuentra en el archivo “**simce6b2018_GLOSAS_PUBLICAS_WEB.xlsx**”, en la hoja “**rbd”**.

a) Lea el archivo en un dataframe con la función

read_xlsx('simce8b2019_rbd.xlsx')
para ello instale y cargue la librería “readxl”.

b) Vea el nombre de las columnas del dataframe. Además, use las funciones **summary** y **str** para ver el contenido del dataframe

c) Indique el número de filas del dataframe. ¿A qué corresponde cada fila?

d) Eliminar las filas que poseen NA en alguna de las columnas que empiezan por “prom”. Hint: una manera de resolver el ejercicio, es construir un vector de booleans de largo igual al número de filas del dataframe, de tal manera que cada componente del vector sea TRUE si es que la fila respectiva del dataframe no posee NA en ninguna de las columnas que empiezan por “prom”

e) Indique qué porcentaje de datos (filas) se han borrado con la operación del punto anterior. (El 100% es el número de filas del datafame original).

f) Cree una nueva columna con el nombre del tipo de dependencia del establecimiento. Específicamente, si el valor de la columna **cod_depe2** es 1, 2, 3, o 4; el valor de la nueva columna debe ser Municipal, Particular Subvencionado, Particular Pagado, o Servicio Local de Educación, respectivamente. Hint: le podría ser útil usar la función “**mutate**” en conjunto con “**case_when**” de la librería “**dplyr**” (ver la última versión de las diapositivas).

Escoja una de las 16 regiones del país para las partes g), h), i), j), y k).

g) Filtre los datos de tal modo que solo aparezcan datos de la región escogida.

h) Determine el puntaje promedio por tipo de dependencia. Indique cuál es la prueba que presenta más diferencia en sus resultados según tipo de dependencia del establecimiento.

i) Determine el puntaje promedio por grupo socioeconómico. Indique cuál es la prueba que presenta más diferencia en sus resultados según grupo socioeconómico.

j) Determine el puntaje promedio de lenguaje, matemáticas, y ciencias sociales para cada comuna (de la región que usted escogió anteriormente); además, obtenga el número de establecimientos educacionales que rindió la prueba de lectura del SIMCE en cada comuna.

k) En el dataframe obtenido en el punto anterior, cree una columna con el promedio del puntaje promedio de las tres pruebas (para cada comuna). Ordene el dataframe en orden descendiente en esta nueva variable.

l) Hacer una función que reciba de entrada: el dataframe y el nombre de una comuna, y entregue como output una lista con el puntaje promedio de los establecimientos de esa comuna en lenguaje, matemáticas, y ciencias sociales.

m) Haga una función que reciba el dataframe y el nombre de una comuna como input, tal que la función imprima en la consola el nombre de los 10 colegios que mejoraron más en su puntaje de lenguaje respecto al año anterior de la comuna ingresada como input. Note que pueden haber filas con NA en la columna que indica cuánto mejoró/empeoró el puntaje respecto al año anterior, dichas filas deben ser removidas dentro de la función.

Pregunta III – Python

A continuación se pide realizar las siguientes tareas en **Python**. Puede que algunos enunciados sean diferentes a los de la parte anterior.

a) Lea el archivo en un dataframe con la función
df = pd.read_excel('simce8b2019_rbd.xlsx')
para ello cargue la librería “pandas”.

b) Vea el nombre de las columnas del dataframe. Además, use los métodos **dtypes** y **describe()** para ver el contenido del dataframe. Note que estos métodos se usan como “df.types”, donde “df” es el nombre del dataframe.

c) Indique el número de filas del dataframe.

d) Eliminar las filas que poseen NA en alguna de las columnas que empiezan por “prom”. Hint: una manera de resolver el ejercicio, es construir un vector de booleans de largo igual al número de filas del dataframe, de tal manera que cada componente del vector sea TRUE si es que la fila respectiva del dataframe no posee NA en ninguna de las columnas que empiezan por “prom”

e) Indique qué porcentaje de datos (filas) se han borrado con la operación del punto anterior. (El 100% es el número de filas del datafame original).

f) Cree una nueva columna con el nombre del tipo de dependencia del establecimiento. Específicamente, si el valor de la columna **cod_depe2** es 1, 2, 3, o 4; el valor de la nueva columna debe ser Municipal, Particular Subvencionado, Particular Pagado, o Servicio Local de Educación, respectivamente. Hint: Podría definir una función que retorne el tipo de colegio dependiendo del número ingresado como argumento a la función, tal que

devuelva Municipal, Particular Subvencionado, Particular Pagado, y Servicio Local de Educación, si es que el argumento es igual a 1,2, 3, o 4, respectivamente (para definir dicha función, podría usar CHAT GPT si lo necesita). Luego, puede usar la función "apply" para aplicar esta función sobre la columna "cod_depe2" del dataframe para crear la nueva columna que se pide.

Escoja una de las 16 regiones del país para las partes g), h), i), j), y k).

g) Filtre los datos de tal modo que solo aparezcan datos de la región escogida.

h) Determine el puntaje promedio por tipo de dependencia. Indique cuál es la prueba que presenta más diferencia en sus resultados según tipo de dependencia del establecimiento.

i) Determine el puntaje promedio por grupo socioeconómico. Indique cuál es la prueba que presenta más diferencia en sus resultados según grupo socioeconómico.

j) Determine el puntaje promedio de lenguaje, matemáticas, y ciencias sociales para cada comuna (de la región que usted escogió anteriormente); además, obtenga el número de establecimientos educacionales que rindió la prueba de lectura del SIMCE en cada comuna. Hint: Puede usar como argumento 'count' en el método '.agg' para contar el número de casos, similar al uso de otros argumentos como 'mean', 'min', 'max'.

k) En el dataframe obtenido en el punto anterior, cree una columna con el promedio del puntaje promedio de las tres pruebas (para cada comuna). Ordene el dataframe en orden descendiente en esta nueva variable. Hint: antes de empezar, se recomienda imprimir primero el nombre de las columnas del dataframe obtenido en el punto anterior. En caso de que el nombre de las columnas sea una tupla, del estilo ('a','b'), entonces para luego acceder a esa columna, se hace del modo df[('a','b')].

Para craer un notebook en Jupyter Notebook, abra Anaconda-Navigator->Hacer click en "launch" debajo de "Jupyter Notebook". Luego, en el browser con Jupyter Notebook, ir a la carpeta en donde se desea trabajar, hacer click arriba a la derecha en "New" y luego hacer click en "Python3 (ipykernel)". Esto abrirá un nuevo notebook.