Segunda Tarea P2P
Rúbrica

Del url `http://world.openfoodfacts.org/data` se puede descargar la tabla de datos **FoodFacts.csv** que contiene información sobre más de 65,000 productos alimentarios: en concreto, 156 variables, de las que 87 corresponden a contenidos nutricionales. La tabla completa ocupa más de 160 Mb (y además va aumentando casi cada día), así que para este ejercicio (en MiriadaX no nos dejan subir tablas de más de 5 Mb...), a partir de esta tabla hemos construido una tabla con algunos productos y las siguientes variables:

- *product_name*: nombre del producto
- *country*: país donde se adquirió (cuando un producto ha sido adquirido en varios países, aparece una fila para cada país)
- *continent*: continente del país
- *nutrition_grade_fr*: escala nutricional francesa
- *additives_n*: número de aditivos
- *main_category*: categoría
- *energy*: calorías por 100 g del producto
- *fat*: grasa (en g) por 100 g del producto
- *sugars*: azúcares (en g) por 100 g del producto
- *fiber*: fibra (en g) por 100 g del producto
- *proteins*: proteínas (en g) por 100 g del producto
- *sodium*: sal (en g) por 100 g del producto
- *alcohol*: alcohol (en g) por 100 g del producto
- *vitamin_b6*: vitamina B6 (en g) por 100 g del producto

El resultado final es la tabla **FoodFactsMooc.csv** que encontraréis en el repositorio del curso (url `https://miriadax.net/documents/28098821/74010125/FoodFactsMooc.csv/c1b38463-6006-4a3b-b94a-a72d31d54831`)

A partir de esta tabla, vamos a estudiar algunos índices nutricionales por productos, por países y por continentes.

1. Cargad la tabla de datos en un *data frame* "global" *DF_G*; a continuación, cread un *data frame* para cada continente: *DF_Europa*, *DF_Africa*, *DF_Asia*, *DF_AmericaN*, *DF_AmericaS*, *DF_Oceanía*.

   En algunos apartados necesitaremos eliminar repeticiones de productos de estos *data frames*. Un mismo producto puede aparecer varias veces en la tabla de partida, porque haya sido adquirido en diferentes países; si varias filas corresponden exactamente al mismo producto, sólo se diferencian en el país y, si corresponde, en el continente. Así que también tenéis que construir un *data frame* "global y sin repeticiones" *DFU_G*, que no contenga las variables correspondientes al país y el continente y donde cada producto aparezca una sola vez. ¿Cuántas repeticiones había en la tabla de datos original?

   Asimismo, para cada continente, cread un *data frame* sin repeticiones (*DFU_Europa*, *DFU_Africa*, etc.) que no contenga la variable correspondiente al país (pero dejad el continente, os va a ser útil dentro de un rato), donde cada producto aparezca una sola vez.

   Finalmente, concatenad por filas los 6 data frames anteriores de la forma *DFU_Continente* en un único data frame *DFU_Continentes*. Este data frame contendrá productos repetidos, pero nunca dentro de un mismo continente.

   (**Indicación**: Si aplicáis la función **unique** a un dataframe, construís un nuevo dataframe eliminando las repeticiones de filas.)

   Puntuación: Son 15 *data frames*, 1 punto por cada *data frame* definido correctamente. Restad 3 puntos si no se da el número de filas repetidas.

2. Calculad las correlaciones entre las variables numéricas del dataframe *DFU_G* (sus últimas 8 variables). Escoged la opción **use** que produzca menos **NA**. Opcionalmente, representad los valores absolutos de estas correlaciones mediante un diagrama de calor. A continuación:

   1. Si lo habéis hecho bien, aparecerá una sola pareja de variables con correlación **NA**. ¿A qué se debe este valor?

   2. Determinad el par de variables numéricas diferentes con mayor correlación en valor absoluto, y dibujad su diagrama de dispersión incluyendo su recta de regresión.

   Puntuación: 20 puntos si correcto. A partir de aquí: restad 4 puntos si se ha usado **cor** con **use="complete.obs"** y 10 puntos si no se ha especificado **use**; restad 3 puntos si se ha usado la opción correcta para **use**, pero no se ha explicado correctamente el motivo del **NA** en la correlación entre el alcohol y la vitamina B6; restad 5 puntos si la correlación máxima se ha determinado "a ojo"; restad 5 puntos si el gráfico es incorrecto. Por otro lado, sumad 5 puntos si se ha incluido un diagrama de calor (no hace falta que sea el de

corrplot) y es correcto. El resultado no puede ser negativo: la nota mínima de este apartado es 0.

3. El Actimel de Danone "ayuda al funcionamiento normal del sistema inmunitario" porque contiene 0.21 mg de vitamina B6 por cada 100 g. ¿Qué porcentaje de los productos de los que en la tabla *DFU_G* se indica su cantidad de vitamina B6, tienen como mínimo tanta vitamina B6 por cada 100 g como el Actimel? ¿Qué 5 productos de esta tabla tienen la mayor cantidad de vitamina B6 por cada 100 g?

   Puntuación: 5 puntos si se ha calculado bien el porcentaje, y 5 puntos si se han determinado correctamente los 5 productos con mayor cantidad de vitamina B6.

4. Producid un gráfico que muestre, para cada continente, los diagramas de caja de las cantidades de azúcar en 100g en los productos adquiridos en el continente; y lo mismo para las cantidades de sal, de grasa y el número de aditivos. Usad el dataframe sin repeticiones *DFU_Continentes*. ¿Se observan diferencias entre los consumos en los continentes? Comentadlas.

   Puntación: 2 puntos por cada diagrama de cajas correcto, 2 puntos a discreción según lo que os convenza la discusión sobre las diferencias entre composiciones. Sumad 5 puntos si se ha intentado mejorar el gráfico de la sal.

5. Vamos a fijarnos ahora en las bebidas alcohólicas (aquellas que tengan valor "Beverages" en la variable *main_category* y contenido de alcohol mayor que 0). Usando los dataframes sin repeticiones *DFU_G* y los diversos *DFU_Continente*, dibujad histogramas del contenido de alcohol en estas bebidas: uno global, y uno para cada continente. Procurad que los grupos sean los mismos en cada histograma (aunque algunos queden vacíos), para poder compararlos mejor. Poned nombres adecuados a los histogramas. ¿Se observan diferencias entre los continentes? ¿Qué continente presenta una distribución del contenido de alcohol en sus bebidas alcohólicas más parecido al global? ¿Se os ocurre por qué?

   Puntuación: 10 puntos si los histogramas son correctos: restad 2 puntos por cada histograma incorrecto y restad 5 puntos si no tienen los mismos límites, hasta un mínimo de 0 puntos. 2 puntos a discreción sobre la discusión.

6. La variable *nutrition_grade_fr* es un factor con niveles *a,b,c,d,e*. Hay otro nivel, vacío, para los productos sin categoría (correspondería al NA). Usando el data frame *DFU_Continente*, dibujad un diagrama de barras que muestre, para cada continente, el porcentaje de productos en cada categoría. No incluyáis los productos sin categoría asignada.

Puntuación: 10 puntos si correcto. Restad 5 puntos si no se ha eliminado el nivel vacío y 3 puntos si no se ha incluído una leyenda que diga cada barra a qué categoría corresponde. 0 puntos si el gráfico es incorrecto (por ejemplo, que muestre para cada categoría, el porcentaje de productos en cada continente).

7. A partir del dataframe *DFU_G*, dibujad un gráfico que contenga los diagramas de caja de los contenidos de azúcar por 100 g de los alimentos de cada nivel de la variable *nutrition_grade_fr*. Repetid este gráfico para los contenidos de sal, grasa, fibra y proteínas. A partir de estos gráficos, ¿podéis interpretar cómo clasifica esta variable los alimentos?

   Si queréis, podéis confirmar si vuestra interpretación va en la dirección correcta consultando el informe oficial(`http://fr.openfoodfacts.org/files/hcspa20150625_infoqualnutprodalim.pdf`) (en francés) sobre la clasificación nutricional francesa.

   Puntuación: 10 puntos si los gráficos son correctos: restad 2 puntos por cada gráfico incorrecto hasta un mínimo de 0 puntos. 3 puntos a discreción sobre la explicación de la clasificación.