**#Tipología y ciclo de vida de los datos. Práctica 2.**

**Autores: Gabriel Paladines y Jaime Pardo**

**1. Descripción del dataset. ¿Por qué es importante y que pregunta/problema pretende responder?**

El dataset elegido (<https://www.kaggle.com/rushirdx/suicide-rates-from-1986-to-2016/data>) es importante porque contiene diversos parámetros que posiblemente tengan una correlación con la tasa de suicidio. Pretende responder a preguntas como las siguientes: ¿cómo es la evolución temporal de la tasa de suicidios? ¿se producen más suicidios entre los hombres que entre las mujeres? ¿hay una generación especialmente afectada por el suicidio? También a preguntas que combinen varios factores. Por ejemplo: ¿hay una diferencia importante entre la tasa de suicidios en hombres y en mujeres en una determinada generación?

Veamos a continuación los campos que contiene:

datos1 <- read.csv("suicide.csv", header = TRUE, sep = ",", quote="\"", dec=".",fill = TRUE)

Tipo de dato. Ejemplo:

class(datos1$country)

|  |  |  |
| --- | --- | --- |
| Campo | Tipo | Comentarios |
| Country | Factor |  |
| Year | Integer | 1985-2016 |
| Sex | Factor |  |
| Age | Factor | 6 rangos |
| Suicides\_no | Integer | Contiene ceros |
| Population | Integer | Contiene únicamente la población del sexo y generación indicados |
| Suicides/100k pop | Numeric | Contiene ceros |
| Country-year | Factor | Campo derivado |
| HDI for year | Numeric | Índice de desarrollo humano (valores 0 a 1). Tiene un 70% de vacíos. |
| Gdp\_for\_year ($) | Factor | PIB anual del país |
| Gdp\_per\_capita ($) | Integer | PIB anual dividido entre el total de población |
| Generation | Factor | 6 generaciones |

**2. Integración y selección de los datos de interés a analizar.**

A priori no es necesario llevar a cabo una tarea de integración. En cuanto a la selección de datos, inicialmente dejaremos de lado el campo HDI for year. Tampoco el campo derivado Country-year nos interesa.

datos2 <- select(datos1, -HDI.for.year, -country.year)

// A completar en función de como quede el Análisis de datos del punto 4 //

**3. Limpieza de los datos.**

**3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?**

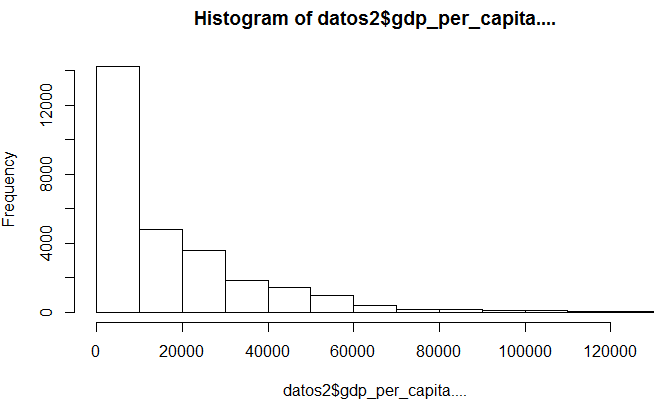
*Suicides* y *Suicides/100k* pop contienen ceros. En principio estos valores son correctos, no requieren una gestión. En cuanto al campo *HDI for year*, contiene un 70% de valores vacíos. Dado el alto valor de este porcentaje, completar los vacíos con estimaciones (valores cercanos, promedio, etc.) probablemente daría resultados alterados y poco fiables. Por ello proponemos excluirlo inicialmente, y con este campo hacer únicamente algún pequeño estudio con una selección parcial de datos.

**3.2. Identificación y tratamiento de valores extremos.**

Inicialmente haremos un cálculo básico, por ejemplo mediante un histograma, de las variables que proceda.

Ejemplo:

hist(datos2$gdp\_per\_capita....)



**4. Análisis de los datos.**

**4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).**

¿Es mayor la tasa de suicidios en hombres que en mujeres (en función del porcentaje de población)?

Atributos: sex, suicides.100k.pop

* Gráfica o tabla donde se muestra un comparativa entre el numero de suicidios entre hombres y mujeres.

¿Qué generación tiene una tasa más alta de suicidios según el nivel de riqueza de su país en el 2008 durante la crisis económica mundial?

* Clasificar los países según el producto interno bruto gdp\_for\_year
* Agregar una columna nueva con la clasificación en 3 grupos (Primer mundo, segundo mundo, tercer mundo), para esto utilizaremos un algoritmo de clasificación no supervisado.
* Comparativa por generación para el año 2008.

¿Como ha evolucionado la tasa de suicidios desde el 1985 en España?

* Columna que agrupe el número por cada año.
* Columna que agrupe la población por cada año.
* Calcular la tasa de suicidios por cada año.
* Graficar para identificación de picos o valores altos.

Nota: Pregunta derivada en función de los resultados de la pregunta anterior.

**4.2. Comprobación de la normalidad y homogeneidad de la varianza.**

…

**4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.**

**5. Representación de los resultados a partir de tablas y gráficas.**

…

**6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?**

…

**7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.**

…