

## Curso: Estadística

Prof. José Manuel Magallanes, PhD

---

## Leyendo Data desde R

---

El primer paso para realizar cualquier tipo de análisis es tener los datos. Esto condiciona además el diseño de la investigación.

Los datos pueden estar organizados por terceros. Si es así, debemos poder leer el formato en que estén preparados. A veces los necesitamos como están, a veces hay que reformatearlos.

Otras veces, uno mismo recogerá los datos. Si los datos representan alguna población en su totalidad no tendremos data sesgada; si no, habrá que hechar mano de algun proceso de muestreo. Veamos los siguientes casos:

### Leyendo SPSS:

Abramos este archivo desde el *American National Election Studies Survey* (ANES):

```
#deben instalar el paquete "haven" primero! (se toma su tiempo!)  
library(foreign)  
  
link1="https://github.com/mgbecerra/EstadisticaUNMSM/raw/master/data/anes_timeseries_2012.sav"  
  
dataSpss=read.spss(link1, to.data.frame=T, use.value.labels = F)
```

En **dataSpss** está un gran archivo. Además no es aun una tabla de datos (dataframe). Creemos un dataframe con dos variables ("libcpre\_self", "libcpo\_self"), un par de preguntas pre y post elecciones donde se solicita ubicarse en algun valor de 7 puntos, cuyos extremos son "extremadamente liberal" y "extremadamente conservador".

```
varsOfInterest=c("libcpre_self","libcpo_self")  
dataSpssSub=dataSpss[varsOfInterest] #convertimos en dataframe y seleccionamos variables de interés  
head(dataSpssSub)
```

### STATA:

```
link2="https://github.com/mgbecerra/EstadisticaUNMSM/raw/master/data/anes_timeseries_2012.dta"  
  
dataStata=read_dta(link2)  
dataStataSub=as.data.frame(dataStata)[varsOfInterest]  
head(dataStataSub)
```

### Excel:

```
library(readxl) #instalen el paquete!

link3="https://github.com/mgbecerra/EstadisticaUNMSM/raw/master/data/idhPeruSucio.xlsx"

dataExcel=read_excel(link3)
head(dataExcel)
```

Al ejecutar el comando anterior, verás que la data se carga pero necesita algo de organización (así la prepara Naciones Unidas... ni modo).

## Uso de los APIs

Hay organizaciones que tienen una política de datos abiertos, por lo que ofrecen mecanismos para acceder a sus datos. Los formatos son por lo general XML o JSON. Traigamos la data producida por el servicio ‘9-1-1’ de la Policia de Seattle, Washington:

```
library(jsonlite) #instalen paquete!
endPoint="https://data.seattle.gov/resource/pu5n-trf4.json"
data911 = fromJSON(endPoint)
head(data911)
```

## “Scraping” tablas de datos

Aqui descargaremos los datos de esta wikipedia

```
# instalen paquetes antes de activarlos!
library(XML)
library(RCurl)

# URL
wiki="https://en.wikipedia.org/wiki/"
link = "List_of_freedom_indices"

# Data
wikiLinkContents = getURL(paste0(wiki,link))
wikiTables = readHTMLTable(wikiLinkContents,
                           stringsAsFactors=FALSE)
```

Veamos que tenemos:

```
#data frame:
is.data.frame(wikiTables) #es un dataframe?
#list:
is.list(wikiTables) #es una lista?
# how many?
length(wikiTables) #cuántos elementos?
```

Al visitar la web, nos damos cuenta la tabla de interés es la segunda:

```
idx=wikiTables[2]
str(idx)
```

Una breve mirada:

```
head(idx)
```

Vemos que necesitamos hacer limpieza de datos!