Exploración

Exploración y visualización de datos mediante el uso de "tuberías", o llamadas encadenadas de funciones, el operador "pipe" (%>%) de la **librería dplyr.**

Ejercicio 1: los datos de desarrollo humano de la *Gapminder Foundation*
https://en.wikipedia.org/wiki/Gapminder_Foundation

- Cargamos la librería dplyr que permite la construcción de "tuberías" (pipelines) o llamadas de funciones encadenadas.
- Cargamos la librería gapminder para usar los datos sobre desarrollo humano presentes en dicha librería.

- Operacions con Tidyverse o uso de "pipes" (tuberias)

    1) Filtramos los datos según el año 2007
    2) Filtramos los datos según el año 2007 y el país EEUU
    3) Ordenamos según el PIB per cápita: de manera ascendente y descendente
    4) Modificamos los datos:
        - Dividimos la variable población entre un millón.
        - Añadimos una variable nueva que sea el PIB per cápita multiplicada por la población.

- Visualizaciones (1ª parte)

    1) Construimos una variable auxiliar que sean los datos del año 2007
    2) Cargamos la librería ggplot2
    3) Visualizamos los valores del PIB per cápita frente a la esperanza de vida
    4) Repetimos la gráfica anterior aplicando una transformación logarítmica
    5) Repetimos la gráfica pero coloreando según el continente
    6) Modificamos de manera que se introduzca el tamaño de la población
    7) Repetimos la gráfica anterior pero separando los continentes en distintas gráficas.

- Estadísticas o función "summarise"

    1) Calculamos la media de la esperanza de vida
    2) Otros momentos estadísticos: media, mediana, moda, desviación estándar, suma, mínimo, máximo, etc.
    3) La operación agrupar:
        - La esperanza de vida según el años
        - La esperanza de vida según el continente

- Visualizaciones (2ª parte)

    1) Visualizamos el PIB per cápita frente a la esperanza de vida (como antes) de varias formas posibles: *geom_point, geom_line, geom_area, geom_smooth*
    2) Construimos un diagrama de barras con los continentes en las Xs y la esperanza de vida en las Ys
    3) Construimos el histograma (distribución de la variable) de la esperanza de vida

4) Construimos el *box plots* de la esperanza de vida
5) Repetimos el *box plots* diferenciando por continente

Ejercicio 2:
Se le suministra un R-notebook – un cuaderno interactivo de R – con la exploración de los datos de los pasajeros del Titanic. El R-notebook está resuelto y trabaja usando las ideas del *tidyverse*. Descargue los datos y estudie poco a poco su funcionamiento.

Este material ha sido obtenido a través del tutorial "Getting Started with the Tidyverse: Tutorial" https://www.datacamp.com/community/tutorials/tidyverse-tutorial-r

Ejercicio 3:
En la librería dplyr se encuentran disponibles los datos de las películas de *Star Wars*, basta poner starwars en R, que consta de 87 observaciones y 13 variables, que van desde el nombre del personaje hasta su color de ojos. Explore y visualice estos datos usando la librería dpylr y la librería ggplot2.
Profundice en el estudio de dplyr: http://dplyr.tidyverse.org/

Ejercicio 4:
Como es natural, se pueden explorar y tratar un conjunto de datos sin usar las librerías que hemos usado en los dos ejercicios anteriores. Probablemente ya lo habrá hecho en alguna asignatura anterior. Utilice los datos de "iris", disponibles con la distribución de R, y realice las siguientes operaciones:

- Exploración visual: una sola variable
  - Se desea representar el histograma y la función de densidad de una característica, o atributo, en concreto
    - Funciones: hist, density. (La función plot sirve para dibujar los valores de la densidad)
  - Se desea saber los valores de una columna en concreto
    - Función table
  - Se desea representar tanto el diagrama circular, como el de barras de una característica en concreto.
    - Funciones: pie, barplot
- Exploración visual: varias variables
  - Calcule la covarianza y la correlación de dos variables.
    - Por ejemplo las variables que representan a los atributos Sepal.Length y Petal.Length
  - Calcule la matriz de correlación de todas las variables.
    - Debe dejar fuera la variable con valores nominales.
  - Construir en diagrama de cajas o *box plot* entre las cuatro variables numéricas
    - Función: boxplot
  - Construir el diagrama de dispersión entre dos variables.
    - Mediante la función plot

- Análogamente pero para el datset al completo

- Exploración visual: adicional
  - Construya el diagrama de dispersión de tres variables
    - Función scatterplot3d del paquete "scatterplot3d" (debe instalar dicho paquete)
  - Construya el mapa de calor o *headmap* de la matriz de datos
  - Obtenga el valor de expresión, o la representación de los patrones de comportamiento, de los datos.
    - Función parcoord del paquete "MASS" (debe instalar dicho paquete)