Readme

Jorge Santos Neila & Javier Cela Lopez

11 de abril de 2020

Contents

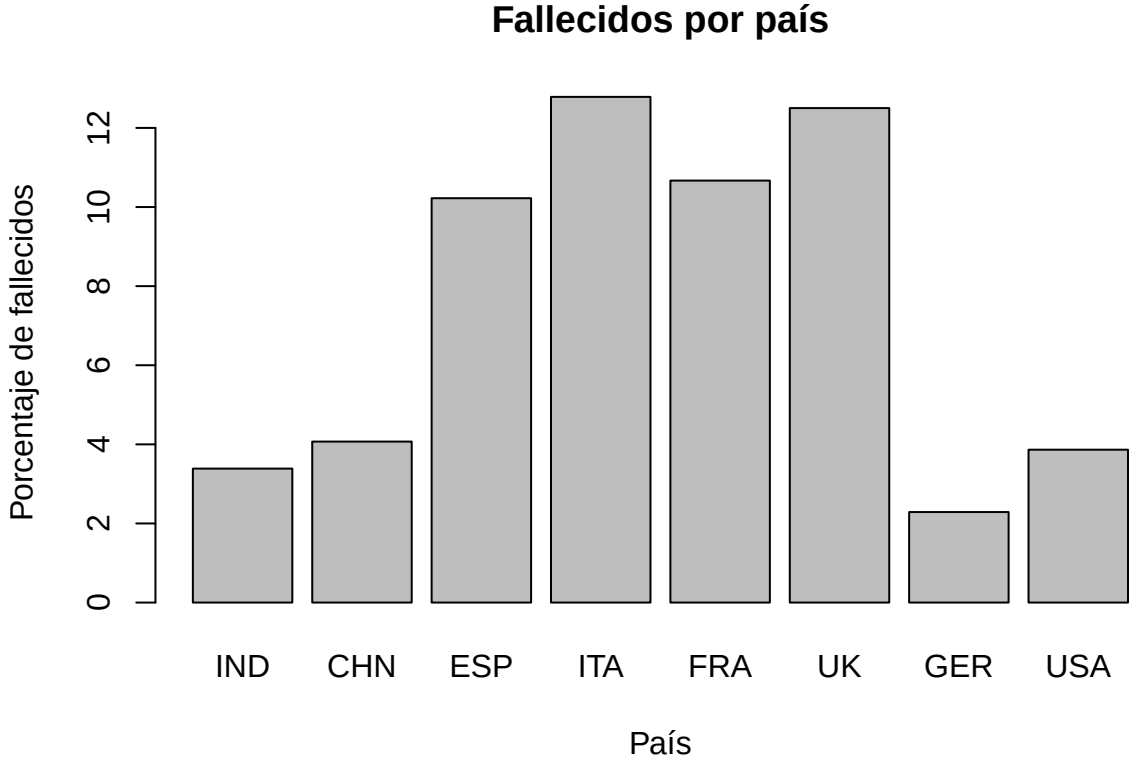
Descripción

Esta práctica se ha realizado bajo el contexto de la asignatura Tipología y ciclo de vida de los datos, perteneciente al Máster en Ciencia de Datos de la Universitat Oberta de Catalunya. En ella, se aplican técnicas de web scraping mediante el lenguaje de programación Python para extraer información relevante al coronavirus a través de Wikipedia - Pandemia de enfermedad por coronavirus de 2019-2020 y del PIB de cada país con el periódico "Expansión" para generar un dataset para que en posterioridad se busquen correlaciones entre estos datos.

Imagen identificativa

Para ilustrar el conjunto de datos y dar una visión general de lo que pretendemos investigar con la formación del mismo, ofrecemos a continuación una primera representación de los datos de fallecidos por la pandemia del COVID-19 y el PIB per cápita de los países analizados. Éste último es un buen indicador de la riqueza de un país, por lo que podemos analizar qué influencia tiene la capacidad económica de un país a la hora de evitar muertes por ésta enfermedad. Escogeremos para ello una pequeña muestra de países representativos.

```r
library(ggplot2)

Coronavirus = read.csv("Coronavirus.csv", sep=";")

df = Coronavirus[Coronavirus$Paises == c("Estados Unidos", "España", "Italia", "Francia", "Alemania", "
df$Paises <- as.character(df$Paises)

df$Paises[df$Paises == "Estados Unidos"] = "USA"
df$Paises[df$Paises == "España"] = "ESP"
df$Paises[df$Paises == "Italia"] = "ITA"
df$Paises[df$Paises == "Francia"] = "FRA"
df$Paises[df$Paises == "Alemania"] = "GER"
df$Paises[df$Paises == "China"] = "CHN"
df$Paises[df$Paises == "Reino Unido"] = "UK"
df$Paises[df$Paises == "Suiza"] = "SUI"
df$Paises[df$Paises == "Chile"] = "CHL"
df$Paises[df$Paises == "India"] = "IND"

fallecidos = df[order(df$PIB.Per.Capita..D.),]$Fallecidos
positivos = df[order(df$PIB.Per.Capita..D.),]$Casos.Positivos
paises = df[order(df$PIB.Per.Capita..D.),]$Paises

barplot(100 * fallecidos / positivos, main="Fallecidos por país", xlab="País", ylab="Porcentaje de falle
```



Ordenando los países por PIB per cápita, nos damos cuenta de que la mayor mortalidad se encuentra en países cuyo PIB per cápita es medio, siendo los países en los extremos más efectivos a la hora de evitar las muertes. Indudablemente, ésto merecería una mayor profundidad en el estudio para determinar los factores verdaderamente influyentes.

Contexto

El contexto de los datos extraídos es el de la pandemia que vivimos actualmente a fecha de abril de 2020. Encontramos datos de muchos países distintos de todo el mundo. Los datos recogidos también corresponden a variables económicas que puede resultar influyentes.

Contenido

Para cada país, el cual se corresponde con un registro en el conjunto de datos, se recogen las siguientes características:

- **Continente**: continente al cuál pertenece el país en formato cadena de texto. Se utilizará más adelante para agrupar datos.
- **Casos.Positivos**: número de casos registrados en el país en cuestión.
- **CxM.Habitantes**: proporción del número de casos positivos por cada millón de habitantes.
- **Fallecidos**: número total de fallecidos.
- **Porcentaje.Fallecidos**: porcentaje de fallecidos con respecto al total de casos registrados.
- **FxM.Habitantes**: proporción del número de fallecidos por cada millón de habitantes.
- **Recuperados#**: número de pacientes recuperados de la enfermedad.
- **Porcentaje.Recuperados**: porcentaje de pacientes recuperados con respecto al total de casos registrados.
- **Anio_x**: año al que pertenecen los datos.
- **PIB.Anual..M.E.**: PIB anual del país en cuestión expresado en euros.
- **PIB.Anual..M.D.**: PIB anual del país en cuestiób expresado en dólares.
- **Var.PIB.Anual**: variación porcentual del PIB anual con respecto al año anterior.
- **PIB.Per.Capita..E.**: PIB per cápita del país en cuestión expresado en euros.
- **PIB.Per.Capita..D.**: PIB per cápita del país en cuestión expresado en dólares.
- **Var.PIB.Per.Capita**: variación porcentual del PIB per cápita con respecto al año anterior.

Estos datos han sido obtenidos de dos páginas distintas y después cruzados para formar el dataset completo. No se ha realizado ningún procesamiento de los datos, se trata de los valore crudos extraídos de la web.

Agradecimientos

Los datos extraídos para la formación del dataset provienen de tablas publicadas en las webs de expansión y wikipedia. Hemos utilizado técnicas de *web scraping* valiéndonos del lenguaje Python. Para ello, se ha analizado la estructura HTML de las páginas en cuestión y se han extraído únicamente los datos que resultaban interesantes para la construcción del dataset.

Inspiración

Como es bien sabido, la situación en la que nos encontramos a fecha de abril de 2020 es muy complicada. Está en la conciencia de la sociedad en general el hacer lo posible para ayudar a superar este obstáculo. Es por eso que hemos elegido esta temática para realizar la práctica. Idealmente, el conjunto de datos servirá para presentar información correcta y actualizada a las personas que quieran disponer de ella para cualquier uso, ya sea actualmente para mantenerse informada o en el futuro para realizar análisis sobre los sucesos que tuvieron lugar estos días. Si se diera una situación similar en el futuro, el análisis de los datos del caso que os ocupa podría utilizarse para guiar la estrategia a seguir.

Licencia

Código fuente y dataset

Tanto el código fuente escrito para la extracción de datos como el dataset generado pueden ser accedidos a través de este enlace. Asimismo, se puede encontrar información adicional en Zenodo.

Recursos

1. Pandemia de enfermedad por coronavirus de 2019-2020 (2020)
2. Datos del PIB Per Cápita Mundial (2019)

3. Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
4. Masip, D. El lenguaje Python. Editorial UOC.
5. Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
6. Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons

---

Tabla participantes

Contribuciones	**Firma Jorge Santos**	**Firma Javier Cela**
Investigación previa	JSN	JCL
Redacción de las respuestas	JSN	JCL
Desarrollo del código	JSN	JCL