## línea horizontal



Entregable -Proyecto de grupo

29/07/2020 al 21/08/2020

**─**

Grupo D

José Luis Manjón, Filipa Teles, y Mónica Villasuso

Data Science - Boot Camp

The Bridge

# Visión General

El objetivo de este proyecto es afianzar y ejercitar los conocimientos adquiridos en la primera mitad del Bootcamp de Data Science, particularmente los relacionados al manejo de APIs y al análisis de datos (EDA) siguiendo todas las etapas del proceso (desde la obtención de los datos hasta las conclusiones derivadas del análisis de los mismos). Adicionalmente, al plantear este reto sobre un tema tan actual como **“el CoronaVirus (COVID19) en el mundo”** y, como en todo trabajo en equipo, se pretende generar sinergias y hacer más eficiente y motivador el proceso de aprendizaje. La clave está en una adecuada distribución de tareas, roles y responsabilidades según la experticia de cada miembro de manera que, una vez concluido, permita a cada quien desarrollar mejor y más rápidamente su proyecto individual (a realizarse en el mismo período – Agosto 2020).

# Objetivos

De acuerdo a los criterios de evaluación establecidos, hemos querido optar por el entregable A+, habiendo seleccionado de este realizar las siguientes opciones:

.- 1 Repositorio en GitHub para centralizar el control de versiones del código y recursos.

.- 6 (Correlación entre variables representado con un Mapa de calor (*heatmap* ).

NOTA: El detalle con observaciones de cada apartado de los entregables se encuentra en el directorio **/documentation/entregables.xls** *También incorporado como anexo al final de este documento para facilitar su revisión*)

# Especificaciones

## Software

Python v2.7 o superior

## 

## Hardware

Procesador i5

Memoria RAM 8 GB

Espacio en disco 200 MB (para los gráficos)

## Requerimientos

* Pandas
* Numpy
* Seaborn
* Matplotlib
* Os / Sys
* Json
* Plotly (\*)
* Flask (\*)

(\*) Deben ser instalados previo a la ejecución de import de las librerías

# Steps

## Contexto (Investigación)

La investigación (adicional a los datos utilizados de la fuente a través de la url) se realizó sobre los países pertenecientes al grupo D, a saber: España, Portugal, Turquía, Reino Unido y Venezuela. En particular, fue necesario:

* Documentar el período y condiciones en el que cada país del grupo decretó el estado de alarma. Las fuentes consultadas y las fechas resultantes se encuentran en **/src/main.ipynb - Apartado 3** (***Período de estado de alarma por país)***
* Documentar ajustes a las cifras realizadas por los países para explicar comportamientos ¨raros¨ de los datos. (Por ejemplo, picos en las cifras de muertes reportadas por España el 25 de Mayo 2020 y el 19/06/2020).
* Investigación del significado del Stringency Index por país, su interpretación y posible relación con los datos en **/src/main.ipynb - Apartado 4.2** (***Variación de stringency index por país)***

## Obtención de los datos

Los datos fueron obtenidos de la siguiente url <https://ourworldindata.org/coronavirus-source-data>

**/src/main.ipynb** - **Apartado 2**

## Data Wrangling / Data Mining / Data Cleaning

* Eliminación de datos totales mundiales o sin país o continente (que afecten los análisis globales). Creación de un dataframe **world\_df**
* Eliminación de registros con información previa al indicio de la pandemia que no agreguen valor al análisis
* Asignación de la posición mundial de cada país (respecto al mundo) para las cifras de casos y muertes por coronavirus.
* Eliminación de columnas que no contengan información.
* Creación de un dataframe específico para el análisis de los países del grupo D (**covid\_grupoD**) y de variables útiles para el análisis (países\_grupoD)

**/src/main.ipynb** - **Apartado 3**

## Exploración/ Análisis de datos (EDA)

El análisis de los datos está organizado de las siguiente manera:

* Análisis global Grupo D

NOTA: gráficos almacenados en **/resources/plots/TOD\_D**

* + Pie Charts
  + Ranking
  + Outliers
  + Tendencias (gráficos dinámicos de cada variable representando independientemente a cada país del grupo)
* Análisis individual por país Grupo D

NOTA: Todos incluyen líneas de referencia de estado de alarma por país (color morado). Los gráficos son almacenados en **/resources/plots/**ISO

* + Casos totales y muertes totales
  + Ranking (evolución diaria de la posición en el ranking )
  + Casos diarios y muertes diarias
  + Muertes diarias con líneas verticales que resaltan los momentos de crecimiento (color rojo) o decrecimiento (color verde) de la curva
* Variación del Stringency Index por país en **/resources/plots/**ISO
* Análisis por continente en **/resources/plots/TOT\_C**
  + Muertes y casos (diarios y totales) en gráficos de área por continente
* Índice de mortalidad por países del grupo D en **/resources/plots/TOT\_D**
* Muertes y casos totales en el mundo en **/resources/plots/TOT\_W**
* Progresión mundial (10 días) de casos y muertes. en **/resources/plots/TOT\_W**
* Correlación
  + Mundial en **/resources/plots/TOT\_W**
  + Países del Grupo D en **/resources/plots/TOT\_D**
  + Por país en **/resources/plots/**ISO

NOTA: ISO hace referencia a 3 dígitos que identifican a cada país (ej. ESP = España)

## Conclusiones

COPIAR RESUMEN DE LAS CONCLUSIONES DEL APARTADO 5 DEL JUPITER NOTEBOOK

**/src/main.ipynb** - **Apartado 3**

# ANEXOS

