## 



Madrid: cambio climático. ¿Ha llegado ya?

Entregable proyecto individual

José Luis Manjón

Data Science - Boot Camp

The Bridge

16/08/2020 to 31/08/2020

# Visión General

Hablar sobre el clima es uno de los temas más recurrentes y políticamente correctos con los que contamos cuando queremos establecer una conversación con otra persona (por aquello de ser educado cuando te encuentras en el ascensor con ese vecino pesado del quinto), pero no por ser repetitiva dejamos de hacerlo.

En la últimos diez quince años, al asunto de la situación meteorológica, se ha unido la preocupación por el clima en general y como los seres humanos, con nuestras actividades estamos contribuyendo a variar algo que lleva con nosotros desde el principio de los tiempos. Me refiero al cambio climático.

Asimismo, en relación al clima parece ser que nuestra memoria no retiene los datos objetivamente y es muy frecuente escuchar al ciudadano de a pie expresiones como: “Este año ha sido el más seco desde que tengo uso de razón” , “El invierno pasado ha sido el más frío de los últimos años”, “En el mes de Julio hemos tenido las temperaturas más altas del siglo”, “Ya solo tenemos dos estaciones, invierno y verano, la primavera y el otoño han desaparecido” .o “El tiempo se está volviendo loco y muchas de las nuevas enfermedades es debido a ello”

Ante toda esta multitud de opiniones y sensaciones me he preguntado si el que este verano haya hecho mucho calor, o que no recuerde la última vez que llovió en Madrid, es algo estacional y que se ha dado en el pasado, es que no tengo memoria sobre el clima de hace un año o por el contrario, estamos haciendo cambiar al clima.

El estudio podría haber sido más global, pero el gran número de datos y variables que habría que manejar no son compatibles con el plazo de entrega ni con la capacidad de procesamiento del equipo informático disponible. Por otra parte, al centrarme en un ámbito tan local en el que he nacido y resido, puedo comprobar si realmente mi cerebro procesa y conserva la información metereológica o simplemente es usada para saber si hay que salir con abrigo y/o paraguas a la mañana siguiente y después olvidada.

# Objetivos

De acuerdo a los criterios de evaluación establecidos, he querido optar por el entregable A.

# Especificaciones

## Software

Python v2.7 o superior

## Hardware

Procesador i5

Memoria RAM 8 GB (mínimo, recommendable 16 GB)

Espacio en disco 200 MB (para los gráficos)

## Requerimientos

* Pandas
* Numpy
* Seaborn
* Requests
* Datetime
* Matplotlib
* Os / Sys
* Json
* Plotly (\*)
* Flask (\*)
* API Key del AEMET para la descarga de los archivos .json con los datos.

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

# Steps

## Contexto

La investigación se realizó sobre los datos climáticos de la ciudad de Madrid.

* Establecer el ámbito temporal en el que vamos a desarrollar el estudio. En este caso se ha decidido enfocarlo en los últimos 40 años (1980 – 2019). Aunque existen datos hasta Agosto de 2020, al no ser un año completo no podríamos establecer comparaciones con años anteriores.
* Una vez obtenidos esos datos ha sido necesario investigar el significado de cada una de las variables registradas por las estaciones meteorológicas situadas en la ciudad de Madrid.

## Obtención de los datos

Los datos se han obtenido de la página web de la Agencia Estatal de Meteorología que en España es el organismo que toma, registra, conserva y estudia las observaciones climatológicas con las que realiza las predicciones metereológicas. ( <http://www.aemet.es/es/portada>)

Son datos de libre acceso, que son entregados en formato .json, para lo que es necesario obtener previamente una API KEY. Para evitar que las peticiones de datos puedan colapsar el servicio, el número de bajadas está limitado, a la vez que la cantidad de datos por cada intento. En este estudio la serie temporal elegida ha sido de cuarenta años y dado que AEMET ha limitado el tamaño del archivo json que puede descargarse a menos de cinco años, ha sido necesario repetir el proceso varias veces hasta abarcar el periodo estudiado.

En la ciudad de Madrid de la AEMET tiene cuatro estaciones meteorológicas ubicadas en el Parque del Retiro, Aeropuerto Madrid-Barajas, Ciudad Universitaria y Cuatro Vientos.

## Data Wrangling / Data Mining / Clean Data

* Una vez obtenidos los 9 archivos .json y transformados en 9 dataframes los uno en un único dataframe llamado madrid\_8019
* Tras analizar los datos existentes en cada estación compruebo que en la estación meteorológica de la Ciudad Universitaria no se han guardado registros durante 5175 días, a lo que hay que sumar los días en los que los valores registrados son nulos, 312 en las temperaturas y 446 en cuanto a la precipitación. En el periodo 1980 - 2019 no dispondríamos de registros climáticos durante más de 15 años. Por ello se toma la decisión de no utilizar los datos recogidos en la Ciudad Universitaria. El dataframe resultante es madrid\_clima3.
* Estudiando los valores Nan en el dataframe anterior (madrid\_clima3) nos encontramos que en las columnas relativas a la temperatura y la precipitación su número no es significativo por lo que rellenaremos estos registros interpolando entre el valor anterior y el posterior, por ser la manera que menor influencia va a tener para el estudio. En primer lugar cambiaremos los datos de type string a float (cambiando la coma del separador de los decmiales por el punto) y después actuaremos sobre los Nan de esas columanas.
* Convierto la columna fecha a type datetime.
* Elimino las columnas cuyos datos no voy a analizar (horatmin, horatmax, dir, horaracha, horaPresMax, horaPresMin).
* Trabajaré con una única estación meteorológica virtual, cuyos registros serán el valor medio de las tres estaciones con las que hemos estado trabajando hasta ahora (Retiro, Aeropuerto, Cuatro Vientos). Este nuevo dataframe es llamado madrid\_tpv
* Mediante la función resample agrupo la columna datetime a una frecuencia mensual.
* Al ‘resamplear’ el dateframe anterior, en unas columnas me interesa obtener el valor medio (tmed, tmin, tmax, velmedia); en otras la suma de los registros diarios(prec, sol); en otras el valor mínimo (tmin) y en otras el valor máximo (tmax). El dataframe resultante es llamado madrid\_mensual
* Procedo de igual manera para crear el dataframe llamado madrid\_anual.
* Con los dataframe madrid\_mensual y madrid\_anual se realizará el análisis.

## 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). Se muestran algunos ejemplos
  + Progresión cada 10 días de casos y muertes
* 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
* 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 cada 10 días de casos y muertes. en **/resources/plots/TOT\_W**
* Correlación
  + Con datos diarios
  + Con datos mensuales

## Conclusiones.