## 



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 para, entre otros cometidos, 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 accesoss 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 la siguiente manera:

* Análisis de la estacionalidad anual en las variables del dataframe.
  + Plot de líneas y puntos
  + 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 de las tendencias de cada una de la columna de datos

NOTA:. Los gráficos son almacenados en **/resources/plots/**TEND\_EST

* + Gráficos px.line interactivos.
  + Graficos px.scatter de las variables de temperatura
* Análisis comparativo entre las variables de tmed, tmax\_med, tmin\_media
  + Gráfico px.line con las tres variables comparadas
* Análisis de la tendencia en sobre el nº días que se superan temperaturas umbrales máximas y mínimas
  + Gráfico px.line con las dos variables comparadas
* Análisis de la tendencia en las precipitaciones, el número de días con y el valor medio de estas.
  + Gráficos de barras interactivos
* Análisis de la tendencia en el viento y en las rachas máximas de viento
  + Gráficos px.line comparando las dos variables
  + Grafico de barras de las rachas de viento.
* Análisis de outlaiers.
  + Gráfico boxplot con los datos diarios
  + Gráfico boxplot de todas las variables con datos anuales.
* Correlación
  + Con datos diarios
  + Con datos mensuales
* Análisis de la frecuencia en los datos mediante diferentes histogramas
  + Histograma conjunto de todas las variables estudiadas con bin automático
  + Histograma conjunto de todas las variables estudiadas con os datos agrupados en cinco intervalos (bins=5)
  + Histogramas interactivos individuales con el número de bins definidos por la función px.histogram

## Conclusiones.

Como se podría presuponer a priori, las variables con índices de correlación próximos a 1 son las temperaturas. De igual manera sucede con la velocidad y las rachas de viento.

Las únicas variables independientes que tienen una correlación que supera el 0.6 es la temperatura máxima con las horas de sol, es decir a mayor horas de sol es posible que se alcancen temperaturas altas, aunque no es el único factor que influye en ello.

Al analizar los datos mensuales en las gráficas de línea y punto, se observa claramente que las variables relacionadas con la temperatura presentan una fuerte estacionalidad anual (hay un patrón que se repite año a año), lo que no ocurre con las precipitaciones y el viento.

Debido a la estacionalidad que se advierte en los gráficos anteriores es complicado determinar si existe algún tipo de tendencia en los, por lo que para el análisis de las distintas variables utilizaré datos anuales.

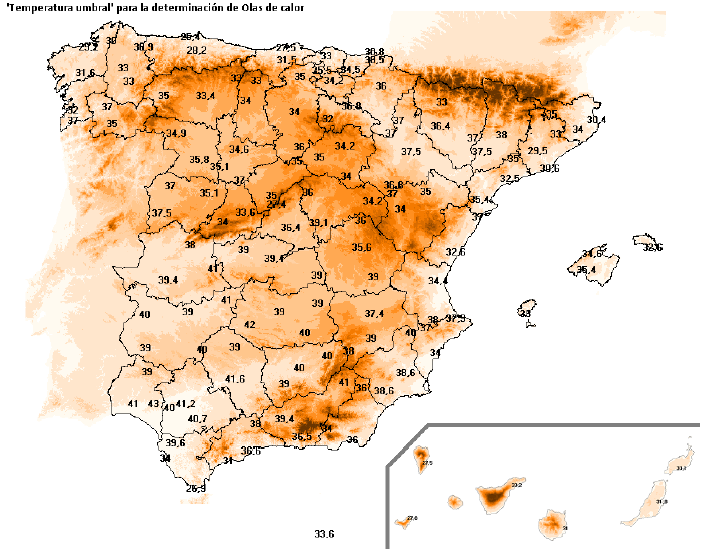
Se puede apreciar que en la media de las temperaturas (max, min y media) hay una tendencia al alza con el paso de los años. En cuanto a las máximas y mínimas temperaturas registradas anualmente no hay valores que año a año vayan superando a los anteriores. Esto puede ser debido a que lo que se está produciendo con las temperaturas no es que vayamos alcanzado todos los años records absolutos, sino que el número de días al año con temperaturas más calidas, tanto en invierno como en verano ha aumentado, haciendo que la media sea alcista.

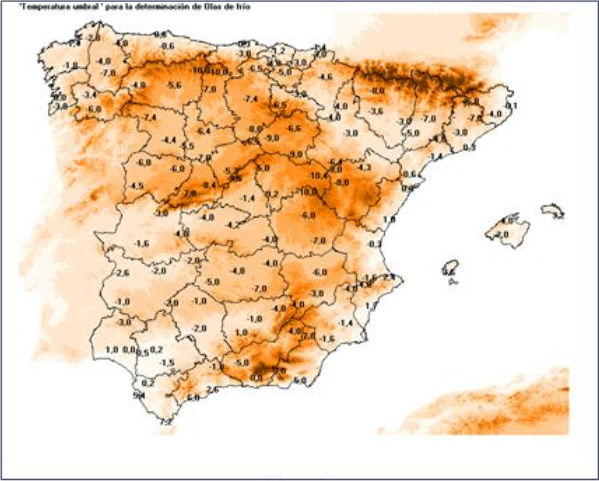
En las temperaturas máxima y mínima medias, en la zona baja de cada uno de los gráficos nos encontramos con los datos del mes más frío de cada año(tmin,\_media) y en la parte alta con el mes más caluroso de cada año(tmax\_media). Ahora si observamos las partes altas de cada uno de esos gráficos de temperatura máxima y mínima media, vemos que desde el año 2014, en los meses de verano la temperatura mínima media está aumentando al igual que la temperatura máxima media en invierno, lo que nos lleva a inviernos con temperaturas máximas mas altas y veranos con temperaturas mínimas también más altas.

Lo que nos lleva a pensar que quizás el número de días con temperaturas altas es más frecuentes y el de temperaturas bajas es menos frecuentes. Es aquí donde se introduce el concepto de ola de calor (también puede ser ola de frío) que la AEMET la define como un episodio de al menos tres días consecutivos, en que como mínimo el 10% de las estaciones consideradas registran máximas por encima del percentil del 95% de su serie de temperaturas máximas diarias de los meses de julio y agosto del periodo 1971-2000", evidenciando con ese percentil —una medida utilizada en estadística que indica, dentro de un conjunto de datos ordenados, el valor por debajo del cual se encuentra un porcentaje de los datos—, el 5% de los días más calidos registrados en esos años. Así, se considera que la temperatura umbral de cada estación es la menor que se haya registrado en esas 93 jornadas. En este estudio no vamos a estudiar las olas de calor, pero al menos utilizaremos ese umbral de temperatura para determinar cuantos días por año se ha superado ese valor. En el caso de las estaciones de Madrid ese umbral está fijado en 36.4 ºC.

En el caso de las temperaturas mínimas se considera ‘Ola de frío’ un episodio de al menos tres días consecutivos, en que como mínimo el 10% de las estaciones consideradas registran mínimas por debajo del percentil del 5% de su serie de temperaturas mínimas diarias de los meses de enero y febrero del periodo 1971-2000. Para Madrid este umbral es de -1.4 ºC.

Si bien no vamos a analizar las olas de calor o frio en Madrid, si voy a obtener el número de días por año en el que la temperatura máxima es superior a 36.4 ºC y la mínima es inferior a -1.4 ºC (datos obtenidos de AEMET).





El número de días por año con temperaturas máximas extremas ha ido aumentando e incluso en los años con menor número de ellos en la última década, estos siguen siendo superiores a los que nos encontramos en las dos primeras décadas.

En relación con el número de días con temperaturas mínimas extremas, no se aprecia ninguna tendencia significativa.

Otro aspecto a tener en cuenta a la hora de determinar si el clima en Madrid está cambiando son las precipitaciones. Si nos fijamos en la serie temporal de precipitaciones desde 1980 no se puede afirmar categóricamente que en los ultimos años las precipitaciones hayan descendido pues nos encontramos cifras similares a años anteriores.

Quizás lo que haya podido variar es como se producen esas precipitaciones y estemos ante un cambio en el tipo de precipitaciones (menos días de lluvia y precipitaciones más intensas). Por ello creo dos nuevas variables, número de días de lluvia por año y la cantidad de lluvia caída por día de lluvia, aunque tras analizar ambas graficas descubro que no hay una variación considerable con años anteriores.

En cuanto a las rachas de viento (aumento brusco del viento con respecto a su velocidad media de más de 15 km/h y durante menos de 20 seg.) o a la velocidad media del mismo no se aprecia ninguna tendencia alcista, por lo que no se puede afirmar que al viento le influya el cambio climático.

Con los datos que hemos obtenido para realizar este estudio, se puede indicar que el mayor cambio se ha producido en las temperaturas.

En cuanto a la temperatura mínima media el inicio de la serie temporal eran frecuentes temperaturas por debajo de nueve ºC y a partir del año 1995 se aprecia una tendencia alcista constante para terminar en 10ºC durante los últimos 3 años de la serie estudiada.

En el caso de la temperatura máxima media se inicia la serie con valores próximos a los 20ºC y durante los siguientes 25 años solo encontramos que en tres ocasiones se alcancen los 21ºC. Pero desde el año 2006 se aprecia un ascenso de la temperatura máxima media (solamente roto por el año 2010, 20ºC) hasta alcanzar o superar los 21ºC.

En cuanto a la cantidad de precipitaciones no se observa ningún tipo de tendencia, ya que durante todo el periodo de estudio estas han sido muy irregulares y los datos recogidos en la última década no difieren en gran medida de las recogidas en las décadas anteriores. Quizás lo que haya podido variar es la intensidad de la lluvia (mm/h), pero esta variable no se encuentra entre las facilitadas por AEMET ni se ha podido inferir de las restantes variables.

En resumen, en cuanto a la temperatura se observa un incremento de aproximadamente 1,5º, pero en cuanto a las precipitaciones no se puede decir que hayan cambiado. **Es decir, podemos afirmar que el cambio climático ya ha llegado a Madrid.**

# ANEXOS

A continuación se da respuesta a las cuestiones de los apartados C, B y A que no hayan sido respondidas anteriormente.

C.7.c.-Posiblemente cambiaria la forma de iniciarlo, creo que es importante la temática y la hipótesis a resolver, pero el nudo gordiano de este tipo de proyectos está en la búsqueda y obtención de los datos. Ocupa una gran parte del tiempo dedicado a la elaboración del proyecto y del que luego dependerá en gran medida la calidad del mismo.

C.7.d.-He aprendido a buscar más eficientemente en google, a darme cuenta del tiempo que se necesita para cada uno de los apartados y a ver las distintas opciones de librerías que existen para hacer gráficos.

B.1.- Al elegir bins =5 estoy indicando que los datos que representaré en el histograma van a estar agrupados en cinco intervalos, es decir el número de columnas que aparecerán. En el caso que nos ocupa al tener una gran cantidad de datos, los agrupa mucho y no da una buena información sobre la frecuencia en la que estos datos se producen.

B.2.- Como se podría presuponer a priori, las variables con índices de correlación próximos a 1 son las temperaturas. De igual manera sucede con la velocidad y las rachas de viento. Las únicas variables independientes que tienen una correlación que supera el 0.6 es la temperatura máxima con las horas de sol, es decir a mayor horas de sol es posible que se alcancen temperaturas altas, aunque no es el único factor que influye en ello.

A.4.a.- Se han realizado gráficos de caja y bigotes o boxpot a todas las variables del dataframe con datos anuales. Este tipo de diagráma muestra a simple vista la mediana y los cuartiles, pudiendo también representar los valores atípicos o outliers.

En nuestro caso no hay practicamente outliers en cuanto a las variables de temperatura u horas de sol, pero tanto en el caso del viento, las precipitaciones, y las rachas se constata una gran cantidad de ellos, incluso de algún valor extremadamente atípico. Los mayores outliers se encuentran en las precipitaciones.

A.4.b.- En todas las columnas existe el mismo número de valores repetidos 40.