MEMORIA:

INTRO

He querido enfocar mi EDA sobre el análisis climatológico de la Comunidad de Madrid relativo al año 2019.
He buscado en general bastantes fuentes de internet, pero finalmente me he quedado con la página de AEMET, de la cuál he podido sacar muchos csv con datos relativos a varias estaciones meteorológicas.

PASOS GENERALES/GRÁFICA Y COLORES ELEGIDOS

Después de una limpieza exhaustiva me he quedado con las 7 estaciones de la Comunidad de Madrid (el csv de referencia tenía todas las estaciones de España).
De cada variable climatológica de las que sacaba el gráfico, hay una línea correspondiente a cada una de estas estaciones(en el mismo gráfico, con una escala de azules). Sin embargo, como mi idea era analizar la Comunidad de Madrid en general, para la presentación he usado ese mismo gráfico pero con una pequeña variante: todas las líneas relativas a las estaciones las he pasado a gris clarito, y he añadido una octava línea, esta vez resaltandola en color y con un grosor un poco superior al resto. Esta línea representa la media de los valores de cada estación por separado, y así poder tener una visualización más general de la región.
Esta media es bastante exhaustiva, ya que las 7 estaciones cubren los puntos más significativos de la comunidad de Madrid: Colmenar Viejo, Barajas, Cuatro Vientos, Getafe, Retiro, Torrejón, Puerto de Navacerrada.
De las 7 estaciones, la que más se “aleja” en casi todos los valores (de precipitaciones, temperatura, ocurrencia de meteoros, etc..) es la de Navacerrada, que tiene picos bastante pronunciados. Pero repito, en el conjunto he usado la media de todas ellas y sobre estos datos he sacado mis conclusiones.

En general he usado una gama de azules para todos los gráficos, ya que creo que es un color que se presenta bastante bien para un análisis climatológico.

TÉCNICAS UTILIZADAS

- atacar API de ree.es y de aemet.es
- utilizar técnicas aprendidas en clase para hacer la petición adecuada (usando la API key) para obtener un “reponse 200”, que significa que la petición ha sido exitosa y he podido obtener un archivo json
- una vez obtenido los json, estudiarlos en profundidad para ver sus estructuras y ver si me servían o no de algo, aunque finalmente he abandonado esa idea y para mi análisis final no los he utilizado
- descargar csv con datos climatológicos y atmosféricos

PROCEDIMIENTO DETALLADO PASO POR PASO

Aunque aquí explico todos los pasos seguidos, en el notebook ipynb es donde se pueden ver más detalladamente, con sus líneas de código para llamar a los csv, limpiarlos, analizarlos, etc… También he añadido comentarios explicativos (a casi todas las celdas) y todos los gráficos, no solo los que finalmente utilizo en la presentación, sino también otros

que me han servido para mi análisis inicial pero que no he visto oportuno cargar en el powerpoint.

Aquí los pasos:
Al principio, antes de centrarme en la Comunidad de Madrid y en el año 2019, me he puesto a buscar información general y a conocer más a fondo la API de la AEMET, ver qué clase de datos podía obtener y si me hubiesen servido para avanzar con mi estudio.
En la carpeta “notebooks” he dejado un archivo (busqueda_datos_inicial.ipynb) con los primeros intentos de atacar a la API, donde simplemente recopilo y descargo información para hacerme una idea general.
Pero lo considero un borrador, de ahí que no lo incluya en el main.ipynb

Después de varios quebraderos de cabeza y de cambiar ligeramente el enfoque de mi EDA, decidiendo centrarme única y exclusivamente en el estudio del clima de la Comunidad de Madrid en 2019, he creado el archivo main.ipyn, que es donde tengo todo el trabajo.

Empiezo descargándome un csv sobre datos climatológicos, que se refieren a un balance hídrico de todas las Comunidades de España.
El balance habla de 3 aspectos principales:

- *Precipitación*: precipitación acuosa acumulada media mensual o anual, para una determinada área
- *Humedad:* humedad del suelo expresada en porcentaje respecto a cierta capacidad máxima de retención de Agua Disponible para las Plantas (AD25 y AD75, considerando una reserva de Agua Disponible Total igual a 25mm y 75mm)
- *Evapotranspiración* de Referencia de Penman Monteith (ETo): es la evapotranspiración de una superficie de referencia que ocurre sin restricciones de agua.

De cada uno de estos datasets me he quedado con los datos relativos a la Comunidad de Madrid. He borrado la columna “anual” ya que lo que tenía era una simple suma de los valores mensuales y yo lo que quería era una media.
Después he añadido la columna “media anual”, que calcula la media de los valores de todos los meses. Esta columna me sirve para sacar los últimos gráficos, donde comparo entre sí todos los valores medios anuales.

Después de estos 3 primeros factores climatológicos, he entrado en el ámbito más “atmosférico”, analizando toda una serie de datos relativas a precipitaciones, ocurrencia de meteoros, humedad relativa, insolación, temperaturas y viento.
He añadido una leyenda en el main.ipynb donde desgloso cada uno de estos elementos (en el apartado 2.0).
Antes de todo, partiendo de un dataset con los nombres e indicativos de todas las estaciones meteorológicas de España, me he quedado solo con las 7 de la Comunidad de Madrid y he guardado un dataframe nuevo ("indicativos_mad.csv").
A partir de este punto, por todos y cada uno de los csv que he analizado, he seguido estos pasos:

- he cargado el csv,
- lo he mergeado con el dataframe previamente creado (con el método inner) para quedarme solo con los valores de las 7 estaciones metereológicas de la Comunidad de Madrid,

- he analizado la columna “anual” y, si veía que no contenía la media de los valores, la borraba y luego creaba otra nueva (“media anual”) a la que le aplicaba la media de todos los valores mensuales de la fila correspondiente,
- he analizado el dataframe usando los métodos info() y describe(), mirando también si los valores numéricos era tales o si estaban con un dtype tipo “object”
- he mirado si había valores NaN y la manera de solventarlos,
  he añadido, al final de la tabla, la columna “Tipo” con el nombre de la variable que estaba analizando (por ejemplo P_MAX si estaba mirando las precipitaciones máximas diarias). Esto me servía para luego, después de los datasets de un determinado sector, poder concatenarlos uno con otro, y para eso necesitaba que tuvieran todas las mismas columnas,
- también he lipiado los nombres de alguna columna, para que encajaran luego mejor en las gráficas,
- en algún dataframe he tenido que aplicarle regex para limpiar los datos, puesto que he detectado un patrón que tenían los valores y por eso no se consideraban como números sino como strings.

Después de esta limpieza inicial, he concatenado los dataframe que me interesaban y los he guardado como csv.
Una vez que ya tenía todos los dataframe listos, me puse a crear los gráficos. Debido a la naturaleza de las variables, solo tengo gráficos de barras (al principio) y todos los demás son gráficos de líneas.

Los últimos he decidido hacerlos como scatterplot. En ellos muestro la media, el valor mínimo y el máximo de las medias anuales de estas 3 variables:

-precipitaciones
-ocurrencia de meteoros
-temperaturas.

Para terminar, y así poder sacar mis conclusiones, he incluido también los datos de referencia relativo a los años 1981-2010, para comparar el valor medio de las precipitaciones y él de las temperaturas, llegando así a la conclusión de que el 2019 ha sido un año bastante cálido y algo más seco respecto a los años de referencia.

INFO INICIAL QUE ACABÉ NO USANDO PARA MI ANÁLISIS

En el archivo “main.csv” también está el dataset, y su relativo gráfico, con los valores de agua embalsada en la Cuenca del Tajo, dentro de la cual están los ríos de la Comunidad de Madrid.
Respecto a este dato he estado buscando en embalses.net, pero solo había datos actuales, no encontré nada relativo al 2019. Sobre este año, solo conseguí encontrar el csv que he cargado en mi notebook, que además no habla específicamente de la Comunidad de Madrid, sino de un área más amplia (la Cuenca del Tajo, como digo más arriba) y no había manera de extraer solo los valores de los embalses de Madrid.

Por esta razón de insuficiencia de datos, no incluyo este gráfico en la presentación.

En el main.ipynb también hay 3 datasets y gráficos relativos a las energías renovables en la Comunidad de Madrid relativas al año 2019.

En mi estudio inicial iba a comparar las precipitaciones con las varias energías, para ver si existía algún tipo de correlación. Desafortunadamente, en el ámbito eléctrico no he encontrado muchos más datos de los que expongo en el notebook, por lo tanto he decidido no añadirlos a la presentación, puesto que sería simplemente añadir 3 gráficas más, sin que tengan alguna relación estrecha con la climatología.

ESTRUCTURA CARPETAS

src/: archivo main.ipynb (notebook con todo el trabajo)
varios csv creados en el main.ipynb y guardados aquí
src/data: aquí dentro está el material (los csv) que he encontrado en internet
src/notebooks: aquí hay 3 notebooks con pruebas varias
src/visualizaciones: aquí están todos los gráficos que he creado

FUENTES USADAS/PÁGINAS WEB VISITADAS

- www.aemet.es (esta ha sido la página que mayormente he usado, atacando primero la API y luego descargando directamente una buena cantidad de csv)
- www.embalses.net
- www.canaldeisabelsegunda.es
- www.ree.es (inicialmente, además del análisis climatológico, quería ver si había alguna relación entre las energías renovables y las precipitaciones, aunque no tenía datos suficientes como para encontrar algún tipo de correlación. En el main.ipynb dejo los dataframes que he encontrado/editado/limpiado y los relativos gráficos que he sacado, pero ya no le veo sentido incluirlos en la presentación, por la pobreza de información al respecto)
- www.ine.es