MADRID 2021

PROGRAMA
DESARROLLO CON PYTHON Y ANÁLISIS DE DATOS

EFECTOS DEL CONFINAMIENTO EN LA CALIDAD DEL AIRE

ANTONIO GIL ORTEGA
PEDRO JIMÉNEZ PEDRERO
INMACULADA ARQUES PORCEL
JESÚS SALEK ABDALÁ RUIZ

DEXS DIGITAL EXPERIENCE SCHOOL

1 ELEMENTOS DE ESTUDIO EN LA CALIDAD DEL AIRE

Los elementos de nuestro estudio son los siguientes:

- CO (Monóxido de Carbono)
- NO2(dióxido de Nitrógeno)
- O3 (Ozono)
- SO2 (Dióxido de Azufre)
- Partículas en suspensión inferiores a 10µm(PM10)
- Partículas en suspensión inferiores a 2,5µm(PM25)

1.1 EXTRACCIÓN DE DATOS

Ayuntamiento de Madrid. Datos sobre los elementos químicos a estudiar por día, durante los años 2019, 2020 y 2021 en la ciudad de Madrid.

https://datos.madrid.es

1.2 PROCESOS DE EXTRACCIÓN DE DATOS

En nuestro caso, todos los ficheros que hemos usado para la extracción de datos han sido de tipo csv.

Como se trataba de múltiples ficheros con la misma estructura de columnas, hemos optado por leer todos los ficheros a la vez y concatenarlos en el mismo DataFrame.

El cual tiene la siguiente información antes de su limpieza.

	PROVINCIA	MUNICIPIO	ESTACION	MAGNITUD	PUNTO_MUESTREO	ANO	MES	\
0	28	79	4	1	28079004_1_38	2021	1	
1	28	79	4	1	28079004_1_38	2021	2	
2	28	79	4	1	28079004_1_38	2021	3	
3	28	79	4	1	28079004_1_38	2021	4	
4	28	79	4	1	28079004_1_38	2021	5	
...	...	...	...	...	...	...	...	
1831	28	79	60	14	28079060_14_6	2019	8	
1832	28	79	60	14	28079060_14_6	2019	9	
1833	28	79	60	14	28079060_14_6	2019	10	
1834	28	79	60	14	28079060_14_6	2019	11	
1835	28	79	60	14	28079060_14_6	2019	12	

	D01	V01	D02	...	D27	V27	D28	V28	D29	V29	D30	V30	D31	\
0	5.0	V	5.0	...	3.0	N	0.0	N	0.0	N	0.0	N	0.0	
1	10.0	N	11.0	...	9.0	V	9.0	V	0.0	N	0.0	N	0.0	
2	10.0	V	10.0	...	9.0	V	9.0	V	10.0	V	10.0	V	10.0	
3	9.0	V	10.0	...	12.0	V	12.0	V	11.0	V	11.0	V	0.0	
4	11.0	V	11.0	...	10.0	V	9.0	V	9.0	V	9.0	V	9.0	
...	...	..	...	...	...	...	...	...	...	...	...	...	...	
1831	94.0	V	104.0	...	88.0	V	90.0	V	99.0	V	108.0	V	98.0	
1832	88.0	V	82.0	...	54.0	V	68.0	V	70.0	V	55.0	V	0.0	
1833	44.0	V	75.0	...	28.0	V	33.0	V	16.0	V	19.0	V	47.0	
1834	41.0	V	55.0	...	55.0	V	52.0	V	47.0	V	56.0	V	0.0	
1835	47.0	V	53.0	...	17.0	V	13.0	V	14.0	V	5.0	V	4.0	

show more (open the raw output data in a text editor) ...

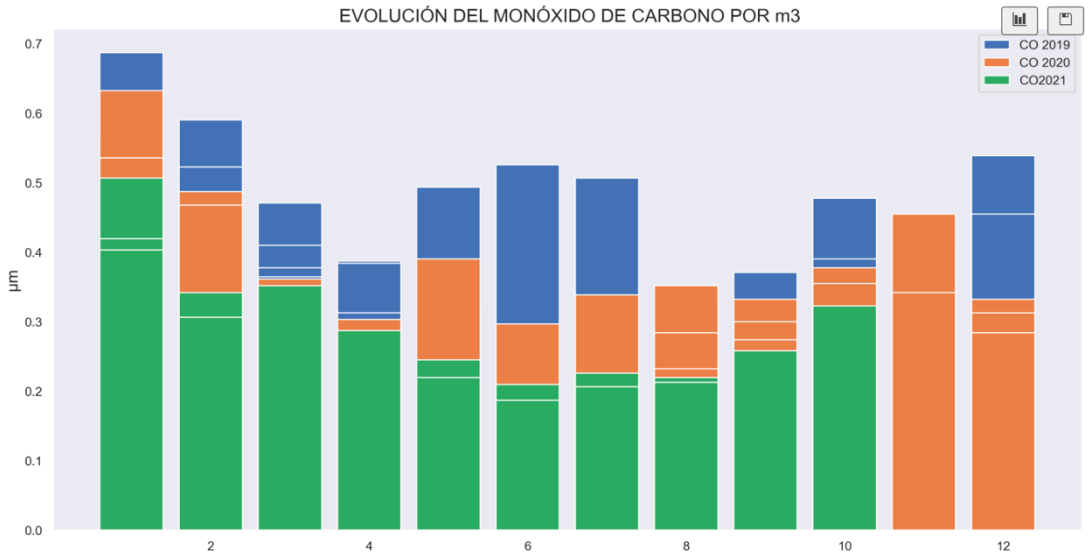
```
1833    V
1834    N
1835    V

[5030 rows x 69 columns]
```

1.3 LIMPIEZA DE DATOS

En este caso, lo primero que hacemos, después de comprobar la información que contiene el fichero, pasamos a realizar los siguientes pasos:

- Modificamos el nombre de la columna “ANO” por “AÑO”
- Eliminamos las columnas “V*” ya que no las necesitamos para nuestro estudio. Dicha columna indica si el valor está validado. En el caso de no estarlo es nulo, y esto lo trataremos más adelante. También eliminamos la columna “PUNTO_MUESTREO”
- Comprobamos la columna “MAGNITUD” que contiene el DataFrame
- En nuestro caso, solo vamos a estudiar la 1, 6 ,8 ,9, 10 y 14 (la equivalencia de estas variables viene en el archivo *metainformacion_2019_tcm30-513561.xlsx*
- Buscamos los nulos, los contabilizamos y eliminamos, ya que pueden influir a la hora de realizar las operaciones estadísticas necesarias para nuestro estudio.
- Sustituimos el valor número de las columnas “PROVINCIA”, “MUNICIPIO” y “MAGNITUD” según el archivo *metainformacion_2019_tcm30-513561.xlsx*
- Intentamos obtener las primeras gráficas



2 DESPLAZAMIENTOS

2.1 EXTRACCIÓN DE DATOS

KAGGLE. Datos sobre movimiento de vehículo durante el período de confinamiento.

https://www.kaggle.com

2.2 PROCESOS DE EXTRACCIÓN DE DATOS

Los datos sobre desplazamientos desde la Comunidad de Madrid y hacia la Comunidad de Madrid durante el periodo de confinamiento, los obtenemos de un Datasets de la página de Kaggle.

Con estos datos queremos comprobar cómo el efecto del confinamiento redujo los desplazamientos y pudo afectar a los datos sobre la calidad del aire.

El cual tiene la siguiente información antes de su limpieza.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000000 entries, 0 to 999999
Data columns (total 10 columns):
 #   Column                              Non-Null Count    Dtype
---  ------                              --------------    -----
 0   date                                1000000 non-null  object
 1   destination_activity                1000000 non-null  object
 2   destination_autonomous_region_name  1000000 non-null  object
 3   distance_km_interval                1000000 non-null  object
 4   origin_activity                     1000000 non-null  object
 5   origin_autonomous_region_name       1000000 non-null  object
 6   total_km_traveled                   1000000 non-null  int64
 7   year                                1000000 non-null  int64
 8   month                               1000000 non-null  int64
 9   day                                 1000000 non-null  int64
dtypes: int64(4), object(6)
memory usage: 76.3+ MB
```

2.3 LIMPIEZA DE DATOS

Después de comprobar la información que contiene el fichero, pasamos a realizar los siguientes pasos:

Para nuestro estudio sólo consideramos necesarias las siguientes 7 columnas:

date	fecha
destination_activity	motivo_destino
destination_autonomous_region_name	comunidad_destino
distance_km_interval	intervalo_kms_distancia
origin_activity	motivo_origen
origin_autonomous_region_name	comunidad_origen
total_km_traveled	total_kms

- Transformamos las columnas “year”, “month” y “day” en una sola columna de tipo datetime y la convertimos en índice de la tabla.
- Renombramos las columnas según la lista anterior
- Filtramos las columnas por origen y destino “Comunidad de Madrid” y borramos el resto.

Finalmente el DataFrame quedaría con la siguiente estructura:

FECHA	motivo_destino	comunidad_destino	intervalo_kms_distancia	motivo_origen	comunidad_origen	total_kms
2020-02-27	otros	Madrid, Comunidad de	5-10	otros	Madrid, Comunidad de	45
2020-02-27	casa	Madrid, Comunidad de	10-50	otros	Madrid, Comunidad de	316
2020-08-13	otros	Madrid, Comunidad de	2-5	otros	Madrid, Comunidad de	18
2020-08-06	casa	Madrid, Comunidad de	2-5	trabajo	Madrid, Comunidad de	29
2020-02-21	casa	Madrid, Comunidad de	5-10	trabajo	Madrid, Comunidad de	49
2020-03-06	otros	Madrid, Comunidad de	10-50	otros	Madrid, Comunidad de	122

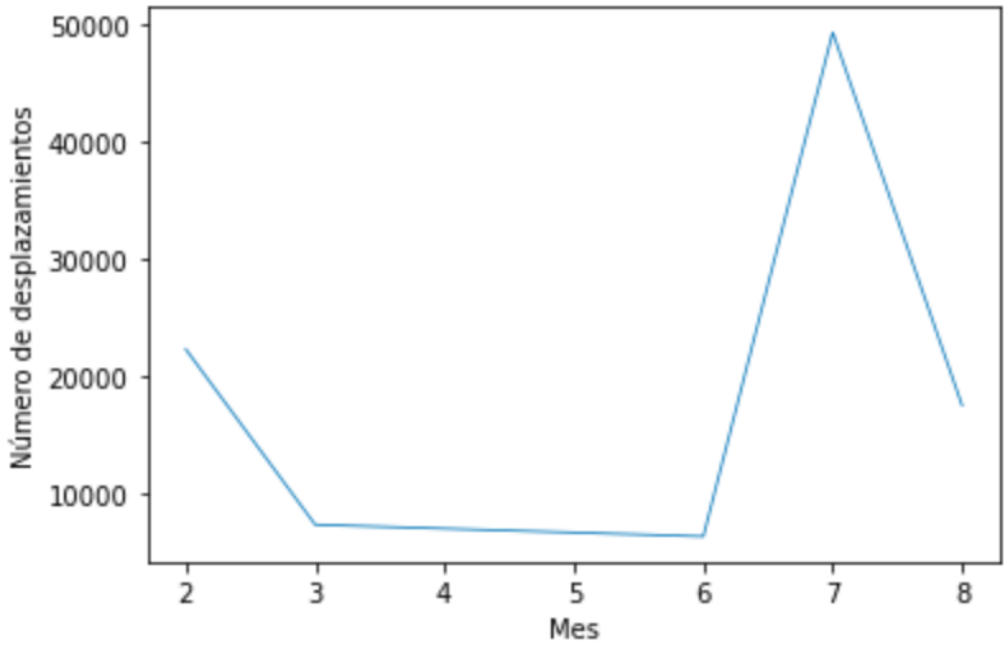


Gráfico de número de desplazamientos durante el periodo del confinamiento

3 ACCIDENTES DE TRÁFICO

3.1 EXTRACCIÓN DE DATOS

Ayuntamiento de Madrid. Datos sobre accidentes de tráfico durante los años 2019, 2020 y 2021

https://datos.madrid.es

3.2 PROCESOS DE EXTRACCIÓN DE DATOS

Los datos sobre accidentes de Madrid los obtenemos directamente de la página del ayuntamiento, a través de varios archivos csv.

Con estos datos queremos comprobar cómo el efecto del confinamiento posiblemente redujo los accidentes, lo cual nos sirve como un indicador más de la reducción del tráfico, que afectaría directamente a los datos sobre la calidad del aire.

Como se trataba de múltiples ficheros con la misma estructura de columnas, hemos optado por leer todos los ficheros a la vez y concatenarlos en el mismo DataFrame.

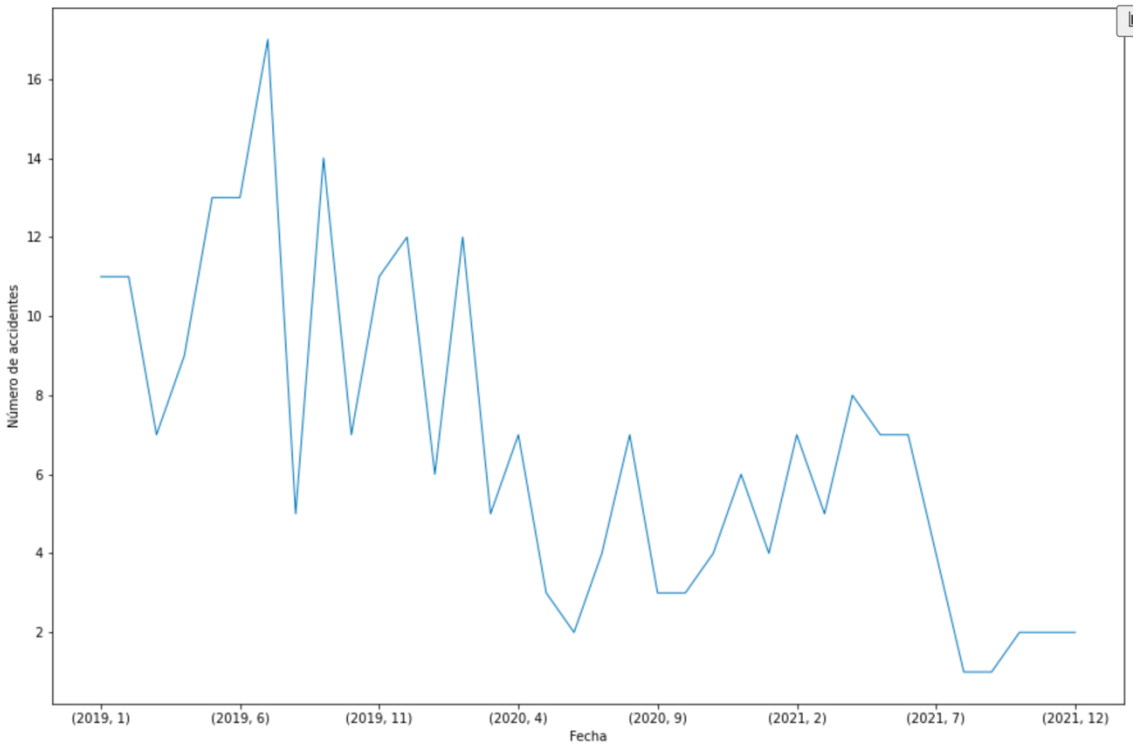
El cual tiene la siguiente información antes de su limpieza:

```
Tipos de datos de las columnas:
 num_expediente           object
fecha                     object
hora                      object
localizacion              object
numero                    object
distrito                  object
tipo_accidente            object
estado_meteorológico      object
tipo_vehiculo             object
tipo_persona              object
rango_edad                object
sexo                      object
lesividad                 object
coordenada_x_utm          object
coordenada_y_utm          object
positiva_alcohol          object
positiva_droga           float64
dtype: object
Número de datos que contiene:
 1847934
```

3.3 LIMPIEZA DE DATOS

Después de comprobar la información que contiene el fichero, pasamos a realizar los siguientes pasos:

- Comprobamos los datos nulos, y decidimos eliminarlos, ya que, en este caso, necesitamos obtener el número de accidentes por mes y año. Y las mediciones nulas no nos aportan ninguna información.
- Comprobamos si hay datos duplicados. En nuestro caso sí que los hay, así que los eliminamos.
- Revisamos los tipos de datos. Cambiamos el formato de la columna “fecha”, pasamos de tipo “object” a tipo “datetime”.
- Indicamos el campo “fecha” como índice.
- Debido a la información que necesitamos extraer (número total de accidentes), vamos a eliminar todas las columnas a excepción de “fecha” y “número de expediente”.
- Agrupamos por año y mes para poder extraer información.



Número de accidentes durante 2019, 2020 y 2021

4 TRÁFICO AÉREO

4.1 EXTRACCIÓN DE DATOS

AENA. Datos sobre la evolución del tráfico aéreo en los años 2019, 2020 y 2021

https://www.aena.es/es/estadisticas/inicio.html

4.2 PROCESOS DE EXTRACCIÓN DE DATOS

Al igual que en el caso del tráfico terrestre, el tráfico aéreo, debido a la combustión de sus motores, influye en la calidad del aire de nuestras ciudades, y por eso lo consideramos como variable dentro de nuestro estudio.

Los datos sobre la evolución del tráfico aéreo los obtenemos de la página de datos abiertos de AENA, mediante el apartado “consultas personalizadas”.

En este caso obtenemos los datos tanto por número de vuelos, como de pasajeros. Para acotar un poco los datos, decidimos centrarnos en número de vuelos realizados durante el periodo de estudio.

En este caso, los datos los muestran por defecto en un formato tipo Excel. Al venir varias tablas dentro de la misma página, así como incluyendo diversos formatos y logos, para facilitar el trabajo de extracción, decidimos seleccionar la tabla (sin realizar ninguna modificación) y convertirla a formato csv.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 2 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ï»¿"AEROPUERTOS"  49 non-null     object
 1   Total             49 non-null     float64
dtypes: float64(1), object(1)
memory usage: 928.0+ bytes
```

Para limpiar los datos obtenidos, seguimos los siguientes pasos:

- En este caso, leemos cada archivo por separado y lo guardamos en variables independientes. Actuamos así porque tras ver los datos, observamos que no incluyen ninguna columna referente a la fecha.
- Añadimos una columna con el campo fecha a cada uno de los DataFrame para poder identificarlos y realizar los correspondientes estudios.
- Una vez hecho esto, y para facilitar el tratamiento de la tabla, lo concatenamos en el mismo DataFrame
- Comprobamos la existencia de datos nulos y los eliminamos, puesto que podrían falsear los datos.
- Modificamos el nombre de las columnas a “AEROPUERTO” y “TOTAL VUELOS”
- Filtramos la tabla para quedarnos solo con los datos relativos a los aeropuertos de la Comunidad de Madrid : “Adolfo Suárez-Madrid Barajas” y “Madrid-Cuatro Vientos”

Fecha	AEROPUERTO	TOTAL VUELOS
2021	ADOLFO SUÃ REZ MADRID-BARAJAS	144.597
2020	MADRID-CUATRO VIENTOS	44.468
2020	MADRID-CUATRO VIENTOS	44.468

5 TELETRABAJO

5.1 EXTRACCIÓN DE DATOS

INE. Datos sobre la evolución del teletrabajo

https://www.ine.es/covid/covid_inicio.htm

5.2PROCESOS DE EXTRACCIÓN DE DATOS

Las nuevas modalidades de trabajo, como es el “teletrabajo”, contribuyen a reducir el tráfico dentro de las ciudades, y por ende a disminuir las emisiones y mejorar la calidad del aire. Por ello, hemos decidido incluirlo como variable en nuestro estudio.

Los datos, han sido obtenidos en la página del INE.

Es importante tener en cuenta que los ficheros del INE vienen codificados como "latin-1" y no "utf-8". Por ello, lo dejamos especificado en la carga.

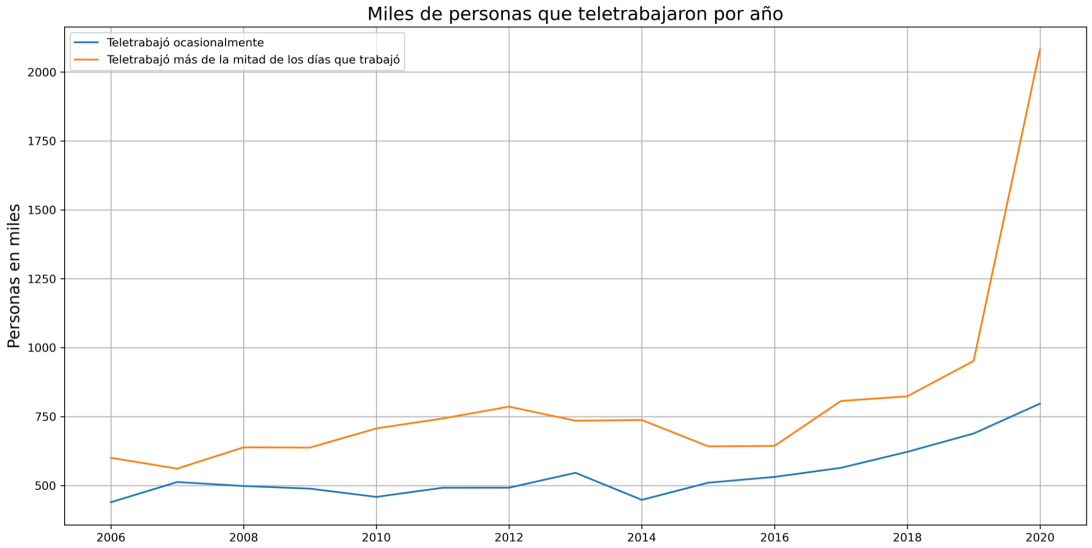
En este caso, aprovechamos el momento de la carga para:

- Renombrar las columnas con "names"
- Si se quiere renombrar aquí las columnas, hay que pasar "header=0"
- Cargar el campo "Periodo" como int16 ya que contiene años y no va a ser necesario más.
- Cargar el campo Personas como float16 - Al llevar la coma decimal, lo carga como "object" por defecto.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 540 entries, 0 to 539
Data columns (total 6 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Sexo                 540 non-null    object
 1   Edad                 540 non-null    object
 2   Dias Teletrabajados  540 non-null    object
 3   Unidad               540 non-null    object
 4   Periodo              540 non-null    int16
 5   Personas             540 non-null    float16
dtypes: float16(1), int16(1), object(4)
memory usage: 19.1+ KB
```

	Sexo	Edad	Dias Teletrabajados	Unidad	Periodo	Personas
0	Ambos sexos	Total	Ocasionalmente	Valor absoluto	2020	797.0
1	Ambos sexos	Total	Ocasionalmente	Valor absoluto	2019	688.5
2	Ambos sexos	Total	Ocasionalmente	Valor absoluto	2018	622.0
3	Ambos sexos	Total	Ocasionalmente	Valor absoluto	2017	564.0
4	Ambos sexos	Total	Ocasionalmente	Valor absoluto	2016	531.0

Graficamos para ver de una forma más visual la evolución de los datos.



6 ARQUITECTURA

Diagrama con la arquitectura actual de proyecto.