# OBJETIVO

Mi trabajo de Análisis exploratorio de datos se centra en el análisis de los festivales de música que se celebran en España a lo largo de 2023.

# HIPÓTESIS QUE SE PLANTEAN

Quiero analizar de todos los festivales se celebran en nuestro país. Que estilos son los que más se programan, que grupos son los que más van a tocar, en qué período del año se programan más festivales y si se aglutinan mayor cantidad de festivales en ciertas provincias, comunidades autónomas u obedecen a una regla de cuanto mayor es el tamaño de la ciudad más cantidad de festivales se celebran. A estas preguntas seguramente se irán uniendo otras cuestiones que será interesante ir abordando.

# EXPLORACIÓN DE LAS FUENTES DE DATOS

## DATOS MUSICALES

Primeramente, he hecho un sondeo fuentes donde puedo obtener dicha información, la pauta es obtener el máximo de información de cada uno de los festivales para así poder hacer un análisis lo más exhaustivo posible.

Comienzo una búsqueda por Internet y veo que al menos hay dos webs donde se recoge esta información, las webs son:

* <https://www.dodmagazine.es/festivales/>

Encuentro gran cantidad de información y, primeramente, comienzo una extracción de los datos de la web mediante la técnica de scrapping con Beautiful Soup. Me doy cuenta que la información en el listado es limitada y cuando entro en la URL de cada uno de los festivales la estructura del código html presenta diferencias por lo que me complica el hecho de estructurar y limpiar la información.

* <https://fanmusicfest.com/calendario-festivales>

Encuentro más información que en la web de dodmagazine. En una primera URL encuentro una serie de filtros que me ayuda a poder catalogar cada festival por estilo de música (aunque agrupado). Una vez entro en las distintas URLs de cada uno de los festivales veo que la información está mejor estructura, guarda un patrón más claro y me ayuda en el proceso de extracción de los datos. Además, el código HTML (para cada festival) cuenta con un archivo GeoJSON propio en el cual encuentro más información que la se ve en el Front-End de la página.

## DATOS DEMOGRÁFICOS

Como forma de complementar la información obtenida de la web, quiero poder tener información demográfica, concretamente el nº de habitantes que hay en cada localidad en la que se celebre un festival. Esta información voy a buscarla directamente al INE.

* <https://www.ine.es/ss/Satellite?L=es_ES&c=Page&cid=1259952026632&p=1259952026632&pagename=ProductosYServicios%2FPYSLayout>

Obtengo un csv con todos los municipios y los diferentes códigos postales de cada municipio y el código del municipio.

* <https://www.ine.es/dynt3/inebase/es/index.htm?padre=525>

Obtengo un fichero xls con todos los municipios, el código del municipio y los habitantes en el último censo (2022).

## DATOS ECONÓMICOS

Para darle más valor al análisis de los festivales de música, quiero analizar paralelamente el impacto económico que suponen los festivales ya no en la economía de cada localidad o municipio donde se celebra el festival sino en el sector de la organización de este tipo de eventos. Comienzo una búsqueda por Internet y veo que existe una Asociación de Promotores Musicales (88 socios/ promotores actualmente), la cual, cada año publica un Anuario donde hace público los resultados económicos del sector. En este Anuario existen una serie de reportajes de análisis del sector de ese año. No obstante, lo que me interesa de ese anuario son los datos. Los datos muestran la facturación neta de la asociación agrupada por meses del año, provincias, comunidades autónomas entre otras.

# EXTRACCIÓN / TRATAMIENTO DE LOS DATOS

## DATOS MUSICALES

A partir de este momento me centro en extraer la información deseada de esta web, también con la ayuda de Beautiful Soup. Utilizo ambas técnicas la de extracción de información a través del HTML y a través del GeoJSON, pudiendo practicar ambos métodos.

Extraigo información puramente musical (nombre del festival, artistas, etiquetas de estilos) pero también datos sobre la localización (latitud, longitud) del festival como la provincia, localidad, lugar de los conciertos o fecha de inicio y finalización del festival (cálculo de forma programática los días de duración del evento).

Creo un dataframe con la ayuda de la librería Pandas de Python para el procesado de los datos.

## DATOS DEMOGRÁFICOS

Previo a poder cruzar los datos demográficos obtenidos del INE.es debo de unir ambas tablas a partir del código de municipio. Una vez hecho esto ya puedo cruzar los datos demográficos con la tabla de los datos musicales.

DATOS ECONÓMICOS

Estos datos no están disponibles a través de un fichero, únicamente puedo acceder a ellos a través de un pdf. De esta forma me planteo utilizar una librería de detección de caracteres (OCR), concretamente pytesseract.

# REPRESENTACIÓN DE LOS DATOS

## PLOTLY

Concretamente a través de express propia de la librería. Todos los gráficos (barras, circulares y de líneas) se realizan a través de esta librería.

## WORDCLOUD

Para representar de una forma gráfica las diferentes variables de análisis de cada uno de los bloques de análisis (datos musicales, económicos y demográficos) se crean nubes de palabras, con la ayuda de la librería Wordcloud.

## FOLIUM

Con los datos de latitud y longitud obtenidos de cada uno de los festivales se genera un mapa de España sonde se ve la localización de cada uno de los festivales agrupados por capas según el estilo agrupado de música.

# PRINCIPALES DIFICULTADES

Las principales dificultades del análisis exploratorio de datos han sido:

* El proceso de scrapping, en el cual, la estructura de las diferentes URLs no guardaban la misma estructura y por tanto he tenido que hacer reajustes en el código para que todos los registros a los que accedía se completarán y no hubiera valores vacíos y poder crear el dataframe.
* El proceso de limpieza de datos de los diferentes archivos extraídos del INE.es. He tenido que tratar bastante los daros para hacerlos coincidir (a partir del código del municipio) y después cuadrarlos con el código postal de los registros de los festivales.
* El proceso de reconocimiento de imagen OCR, algunas de las imágenes no estaban del todo claras y no se reconocían bien las cifras. Teniendo que comprobar algunas tablas de forma visual para ver que los datos estaban bien.

DOCUMENTACIÓN

Se adjunta documentación de los anuarios de la Asociación de Promotores Musicales de los años 2014, 2021 y 2023.