

# Visualización PRAC1

**Nombre y apellidos: Pablo Martínez Pavón**

Esta actividad, primera parte de la práctica final, consiste en la selección por parte del estudiante de un conjunto de datos de su interés que será usado en el proyecto de creación de la visualización de datos, de acuerdo con unos criterios establecidos. Básicamente, la temática es libre, pero se valorarán los aspectos siguientes:

```
import pandas as pd
```

## 1. Justificad brevemente vuestra selección, ya sea por motivos personales o profesionales. [10%]

De partida, he valorado diferentes temas de mi interés (fundamentalmente personal) de los que no me ha sido posible encontrar datasets completos y que cumplan con los requisitos establecidos de dimensiones, características, etc.

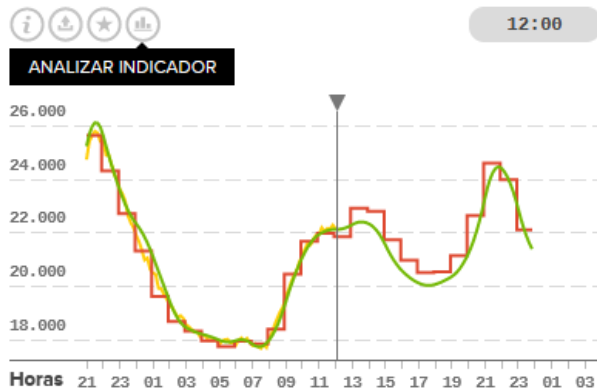
Entre otros:

- **Análisis de consumo energético, a partir de los datos de REE (ESIOS) y OMIE.**
  - En este caso se pueden lograr todos los datos, sin embargo, requeriría mucho tiempo el combinar las fuentes de datos para lograr un dataset. Esto no parece el objetivo de la presente asignatura.
  - Además, fundamentalmente se trataría de datos cuantitativos.
  - Estructura que se había previsto:

Fecha y hora	Consumo (MWh)	Generación por fuentes						Precio (€/MWh)	Tecnología que marca precio de cierre
		Fuente i	Generación i (MWh)	Fuente ii	Generación ii (MWh)	Fuente ...	Generación ...		

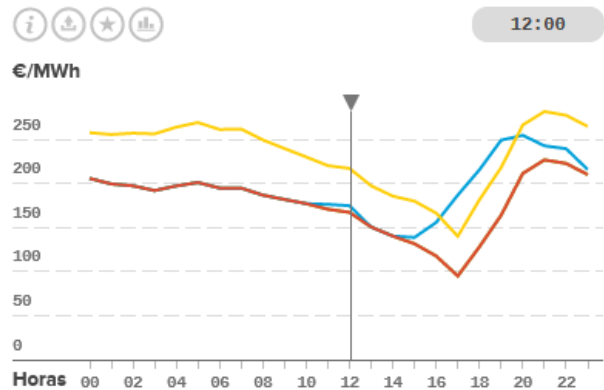
- Posibles fuentes:
  - [REE-ESIOS. Resumen general](#)
  - [REE-ESIOS. Demanda real diaria por hora](#)
  - [OMIE. Energía diaria por fuente de generación](#)
  - [OMIE. Precio horario del día](#)
  - [OMIE. Tecnologías que marcan el precio de cierre](#)

## GENERACIÓN Y CONSUMO



DEMANDA REAL	22.162 MW
DEMANDA PROGRAMADA	21.854 MW
DEMANDA PREVISTA	22.134 MW

## MERCADOS Y PRECIOS



PVPC	217,34 €/MWh
MERCADO SPOT ESPAÑA	167,39 €/MWh
MERCADO SPOT FRANCIA	174,92 €/MWh
MERCADO SPOT PORTUGAL	167,39 €/MWh

### • Estudio sobre construcciones sostenibles, orientadas al estándar Passivhaus.

- No ha sido posible descargar los datos, ya que, aunque están disponibles en la web es imposible descargarlos de manera agregada en un archivo. Se necesitaría realizar un scripting por ejemplo para lograr los datos.
- El dataset hubiese presentado tanto datos categóricos, como cuantitativos.
- Fuente: [Passive House DB](#)

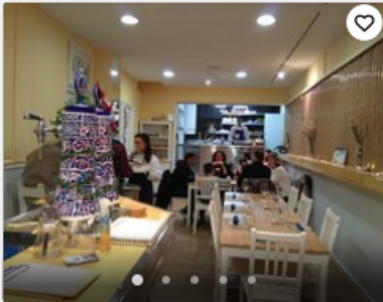
<p>NZ-4130 <b>Havelock North</b> (Hawke's Bay)</p> <p>detached single family house</p> <p>Passive House new build 2021</p> <p>timber construction</p> <p>1 unit   213 m<sup>2</sup></p> <p>ID 6867 <a href="#">Details</a></p>	<p>F-29940 <b>La Forêt-Fouesnant</b> (Bretagne)</p> <p>office   administration building</p> <p>Passive House new build 2020</p> <p>timber construction</p> <p>1 unit   504 m<sup>2</sup></p> <p>ID 6866 <a href="#">Details</a></p>	<p>UK-IV14 <b>Strathpeffer</b> (Scotland)</p> <p>detached single family house</p> <p>Passive House new build 2019</p> <p>timber construction</p> <p>1 unit   136 m<sup>2</sup></p> <p>ID 6865 <a href="#">Details</a></p>	<p>F-71450 <b>BLANZY</b> (Bourgogne-Franche-Comté)</p> <p>detached single family house</p> <p>Passive House new build 2021</p> <p>masonry construction</p> <p>1 unit   148 m<sup>2</sup></p> <p>ID 6864 <a href="#">Details</a></p>	<p>UK-NW1 <b>ORE</b> London (Greater London)</p> <p>multi family dwelling</p> <p>Passive House new build 2018</p> <p>masonry construction</p> <p>38 units   3265 m<sup>2</sup></p> <p>ID 6860 <a href="#">Details</a></p>
--	---	---	---	---

Finalmente, ante el hecho de no encontrar datos de los temas de mayor interés, he podido encontrar en Kaggle un dataset que recoge datos de diferentes restaurantes europeos de la web Tripadvisor. Una de mis aficiones es la gastronomía, tanto cocinando yo, como yendo a probar sitios nuevos.

### • Análisis de las valoraciones de restaurantes en TripAdvisor.

- Este dataset puede ayudar a:
  - Detectar locales nuevos.
  - Qué es lo que más valoran los clientes de un restaurante?
  - Mejores lugares a los que viajar para hacer turismo gastronómico.


- En base a los requisitos de un perfil de cliente definido, analizar cuales son las variables que tienen más peso a la hora de calificar un restaurante para guiar la selección del lugar dónde comer/cenar.
- Fuente: [Kaggle](#)



### 5. Niño Corvo

●●●●● 220 opiniones · **Cerrado hoy**  
 Española, Europea · €€-€€€ · 🧻 Tomando medidas de seguridad

“Tienes que probarlo!!!!”  
 “Cocina sen vergoña”




### 6. Os Padróns

●●●●● 103 opiniones · **Cerrado hoy**  
 Española, Americana · €€-€€€

“PARA REPETIR”  
 “Excelente atención, hamburguesa y postre...”

Hacer pedido online



### 7. Don Marco Pizza

●●●●● 302 opiniones  
 Italiana, Pizza · €

“Las mejores pizzas de todo Vigo, sin duda...”  
 “Calidad y atencion”

## 2. La relevancia del conjunto de datos en su contexto. ¿Son datos actuales? ¿Tratan un tema importante por algún colectivo concreto? ¿Se ha tenido en cuenta la perspectiva de género? [10%]

Si nos fijamos en lo que se indica en Kaggle, los datos proceden de la web [tripadvisor.com](#) y fueron obtenidos mediante scraping a principios de mayo de 2021. Así, **se tratan de datos actuales**.

En este caso, **el dataset no trata ningún tema que se pueda considerar de interés para la sociedad**; están enfocados a una actividad de ocio. \_Podría ser la base de una herramienta útil para algún colectivo profesional como chefs o propietarios de restaurantes.

**El dataset no cuenta con datos que permitan aplicar la perspectiva de género, por ejemplo no clasifica a l@s propietari@s/chefs por su género.** De tener esta clase de datos podría ser de interés comprobar si las valoraciones son más o menos críticas según su diana sea hombre o mujer.

## 3. La complejidad (medida, variables disponibles, tipos de datos, etc.). ¿Tiene del orden de centenares o miles de registros? ¿Tiene del orden de decenas de variables?

## ¿Combina datos categóricos y cuantitativos? ¿Incluye otros tipos de datos? Evitad los conjuntos excesivamente simples. [25%]

```
# lectura datos
dataset=pd.read_csv("C:/Users/pablo/Desktop/VISUALIZACION_PRAC1/tripadvisor_european_restaurants.csv")
```

```
P:\Users\pablo\anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3444:
DtypeWarning: Columns (4) have mixed types.Specify dtype option on import or set
low_memory=False.
  exec(code_obj, self.user_global_ns, self.user_ns)
```

```
# dimensiones
dataset.shape
```

```
(1083397, 42)
```

El dataset presenta 1.083.397 registros y 42 variables. Las variables que recoge son las siguientes:

Variable	Tipo	Descripción
<b>restaurant_link</b>	String (Cualitativa)	Identificador único de cada registro, se corresponde a la referencia en el link de la web
<b>restaurant_name</b>	String	Nombre del restaurante
<b>original_location</b>	Matriz de strings	Ubicación del restaurante por [país, región, ciudad]
<b>country</b>	String (Cualitativa)	País en el que se ubica el restaurante
<b>region</b>	String (Cualitativa)	Región en la que se ubica el restaurante
<b>province</b>	String (Cualitativa)	Provincia en la que se ubica el restaurante
<b>city</b>	String (Cualitativa)	Ciudad en la que se ubica el restaurante
<b>address</b>	String	Dirección del restaurante
<b>latitude</b>	Double (Cuantitativa)	Coordenadas de latitud del restaurante
<b>longitude</b>	Double (Cuantitativa)	Coordenadas de longitud del restaurante
<b>claimed</b>	String (Cualitativa dicotómica)	Está la cuenta reclamada por el restaurante?

Variable	Tipo	Descripción
<b>awards</b>	String (Cualitativa)	Lista de premios (separadas por coma)
<b>popularity_detailed</b>	String	Puesto que ocupa entre los restaurantes de la zona
<b>popularity_generic</b>	String	Puesto que ocupa entre los lugares de la zona
<b>top_tags</b>	String	Lista de tags en los que destaca
<b>price_level</b>	String (Cualitativa)	Rango de precios en € (símbolos)
<b>price_range</b>	String (Cualitativa)	Rango de precios en € (cantidades)
<b>meals</b>	String	Lista de tipos de comidas
<b>cuisines</b>	String (Cualitativa)	Tipo de cocina
<b>special_diets</b>	String (Cualitativa)	Tipo de dietas especiales
<b>features</b>	String (Cualitativa)	Características del restaurante
<b>vegetarian_friendly</b>	Boolean (Cualitativa dicotómica)	Apto para vegetarianos?
<b>vegan_options</b>	Boolean (Cualitativa dicotómica)	Apto para veganos?
<b>gluten_free</b>	Boolean (Cualitativa dicotómica)	Apto para celíacos?
<b>original_open_hours</b>	Diccionario	Horario del restaurante
<b>open_days_per_week</b>	Integer (Cuantitativa)	Número de días abierto por semana
<b>open_hours_per_week</b>	Integer (Cuantitativa)	Número de horas abierto por semana
<b>working_shifts_per_week</b>	Integer (Cuantitativa)	Número de turnos abierto por semana
<b>avg_rating</b>	Double (Cuantitativa)	Valoración media del restaurante
<b>total_reviews_count</b>	Integer (Cuantitativa)	Número de reseñas del restaurante
<b>default_language</b>	String (Cualitativa)	Idioma por defecto

Variable	Tipo	Descripción
<b>reviews_count_in_default_language</b>	String (Cualitativa)	Número de reseñas del restaurante en el idioma por defecto
<b>excellent</b>	Integer (Cuantitativa)	Número de reseñas excelentes del restaurante
<b>very_good</b>	Integer (Cuantitativa)	Número de reseñas muy buenas del restaurante
<b>average</b>	Integer (Cuantitativa)	Número de reseñas normales del restaurante
<b>poor</b>	Integer (Cuantitativa)	Número de reseñas malas del restaurante
<b>terrible</b>	Integer (Cuantitativa)	Número de reseñas terribles del restaurante
<b>food</b>	Double (Cuantitativa)	Valoración media de la comida del restaurante
<b>service</b>	Double (Cuantitativa)	Valoración media del servicio del restaurante
<b>value</b>	Double (Cuantitativa)	Valoración media del valor del restaurante
<b>atmosphere</b>	Double (Cuantitativa)	Valoración media de la atmósfera del restaurante
<b>keywords</b>	String	Lista de palabras clave populares

A partir del dataset, se proponen los siguientes filtros para desarrollar la práctica:

- datos de restaurantes pertenecientes a España.
- restaurantes cuya cuenta está gestionada por su dueño ("Claimed").
- y con al menos un número de reseñas significativo (se ha impuesto un umbral de 320 para ajustarse al rango de filas dado, menos de 10.000).

```
#FILTRADO
#solo datos españa
dataset=dataset[dataset["country"]=="Spain"]
print("Registros en España: {}".format(dataset.shape))
#que hayan sido reclamados
dataset=dataset[dataset["claimed"]=="Claimed"]
print("Registros en España que hayan sido reclamados: {}".format(dataset.shape))
#con al menos XX reseñas
dataset=dataset[dataset["total_reviews_count"]>=320]
print("Registros en España que hayan sido reclamados con un mínimo de reseñas: {}".format(dataset.shape))
```

Registros en España: (157479, 42)

Registros en España que hayan sido reclamados: (69099, 42)

Registros en España que hayan sido reclamados con un mínimo de reseñas: (9762, 42)

Antes de recortar las columnas del dataset, analizamos cuantos valores nulos hay para cada variable.

```
#comprobamos valores nulos
dataset.isna().sum()
```

```
restaurant_link          0
restaurant_name          0
original_location        0
country                  0
region                   0
province                 1550
city                     5779
address                  0
latitude                 6
longitude                 6
claimed                  0
awards                   1327
popularity_detailed      0
popularity_generic       4
top_tags                  1
price_level              13
price_range              4512
meals                    4322
cuisines                  304
special_diets             933
features                 7788
vegetarian_friendly       0
vegan_options             0
gluten_free              0
original_open_hours      660
open_days_per_week       660
open_hours_per_week      660
working_shifts_per_week  660
avg_rating                0
total_reviews_count       0
default_language          0
reviews_count_in_default_language  0
excellent                 0
very_good                 0
average                   0
poor                      0
terrible                  0
food                      0
service                   0
value                     0
atmosphere                2082
keywords                  5370
dtype: int64
```

Existen algunas variables que presentan más valores nulos que valores válidos. Estas serán descartadas de partida; junto con algunas otras. Así, eliminaremos:

- original\_location, por haber sido desglosado en otras variables.
- country, una vez hecho el filtrado se puede descartar.
- city, por el alto número de valores na.
- address, no vamos a llegar nunca en esta práctica a un análisis por calle.
- claimed, una vez hecho el filtrado se puede descartar.
- popularity\_generic, no tiene sentido comparar con lugares que no sean restaurantes.
- price\_range, por el número de valores na.
- meals, por el número de valores na. Sino podría ser interesante para filtrar restaurantes "puros" de locales que ofrecen todo tipo de servicios.
- special\_diets, las de interés ya tienen sus propios apartados.
- features, por el alto número de valores na.
- original\_open\_hours, tiene apartados desarrollados que son más interesantes.
- default\_language, no se tendrá en cuenta el idioma porque está relacionado con el scraping, no con el servicio que ofrece el restaurante.
- reviews\_count\_in\_default\_language, en línea con lo anterior.
- keywords, por el alto número de valores na. Sino podría ser de interés.

Disponemos de una jerarquía, de partida era país → región → provincia → ciudad, sin embargo entre que ya se parte de un solo país y que se ha eliminado la variable *city*, nos quedamos con solo dos niveles. Se ha comprobado por medio del siguiente código que la jerarquía de localización no tiene un número de elementos fijo lo que dificulta la comparación entre restaurantes.

```
#ejemplo
tamaño=[]
for i in dataset["original_location"]:
    tamaño.append(len(i.split(",")))
print(set(tamaño))
```

Se procede a la eliminación de las columnas:

```
#eliminacion de columnas
dataset=dataset.drop(['original_location',
'country','city','address','claimed','popularity_generic','price_range',
'meals','special_diets',

'features','original_open_hours','default_language','reviews_count_in_default_language','k
eywords'], axis = 1)
print("Registros tras la limpieza: {}".format(dataset.shape))
```

Registros tras la limpieza: (9762, 28)

Se vuelve a comprobar el número de valores nulos:

```
#comprobamos valores nulos
dataset.isna().sum()
```

restaurant_link	0
restaurant_name	0
region	0
province	1550



```

latitude          6
longitude         6
awards           1327
popularity_detailed  0
top_tags         1
price_level      13
cuisines         304
vegetarian_friendly  0
vegan_options    0
gluten_free      0
open_days_per_week  660
open_hours_per_week  660
working_shifts_per_week  660
avg_rating        0
total_reviews_count  0
excellent         0
very_good        0
average          0
poor             0
terrible         0
food             0
service          0
value            0
atmosphere       2082
dtype: int64

```

```

dataset=dataset.dropna()
print("Registros tras la limpieza: {}".format(dataset.shape))

```

```
Registros tras la limpieza: (5114, 28)
```

Los valores con NaN se han eliminado porque realizar un proceso de arreglo del dataset variable a variable no se corresponde al objetivo de la práctica y por el tipo de variables resultaría bastante complicado, si no imposible.

Se queda con un dataset mucho más pequeño en comparación con el de partida, sin embargo este presenta unas dimensiones adecuadas para el objetivo propuesto que nos permitirán ofrecer mucho juego en las visualizaciones a desarrollar como parte de la segunda parte de la práctica de la asignatura.

**4. La originalidad. No repetid los conjuntos de datos clásicos. Podéis, por ejemplo, combinar o mejorar visualizaciones existentes. ¿Hay otras visualizaciones basadas en este conjunto de datos? ¿Es una evolución o actualización de un conjunto anterior? ¿Habéis enriquecido un conjunto de datos ya existente? [25%]**

El usuario que ha publicado el dataset, también ha realizado un [análisis EDA](#).

**La idea es hacer visualizaciones originales, ni combinaciones, ni mejoras de otras preexistentes;** se analizarán los trabajos previos para identificar los pros y contras de los datos y de los tipos de gráficos empleados de cara a realizar las mejores visualizaciones posibles durante la segunda entrega de la presente práctica.

Por tratarse de datos procedentes de un web scraping, **el dataset es original y no parece haber sido sometido a ninguna actualización.**

En Kaggle, el usuario indica que el dataset está sometido a una licencia CC0: Public Domain, por lo que se podría emplear para cualquier uso. Si vamos a la fuente original de los datos, en la web se indica que su contenido, y por tanto todos los datos de la web, estaría sujeto a Copyright (© 2022 TripAdvisor LLC Todos los derechos reservados).

Para enriquecer el dataset, se plantea la posibilidad de modificar variables para que sean más útiles de cara a un análisis o una comparación. Por ejemplo:

1. **popularity\_detailed:** Convertir el texto a un índice que nos indique la popularidad en tanto por uno del restaurante, siendo 1 el más popular y 0 el que menos.
2. **top\_tags:** Convertir la lista de tags a un índice que indique si presenta los más frecuentes.

```
from collections import Counter
topT=[]
for i in dataset["top_tags"]:
    for j in i.split(","):
        topT.append(j.strip())
print("Número de tags únicos: {}".format(len(set(topT))))
c = Counter(a)
print(Counter(topT).most_common(10))
```

## 5. Las cuestiones que responderéis con la visualización de datos, ¿Tienen en cuenta los puntos anteriores? ¿Están bien planteadas? ¿Son adecuadas por el conjunto de datos elegido? [30%]

Se plantean las siguientes preguntas que se consideran bien planteadas y que se les puede dar respuesta en base a los datos disponibles:

- Distribución de valoración media, % de valoraciones con mayor puntuación (excelente), % de más caros, etc. por regiones.
- Es cantidad de reseñas sinónimo de calidad?
- Los vegetarianos tienen mejores o peores reseñas de media?
- Que puntuaciones tiene mayor peso sobre la valoración general (food, service, value o atmosphere)?
- Que características parecen ser significativas de cara a una mayor valoración (vegetariano, top\_tags, etc.)?
- Existe relación entre el servicio ofrecido y el horario del personal (open\_days, open\_hours, working\_shifts)?

Además se cree que la respuesta a estas preguntas se puede ofrece de manera clara y sencilla por medio de técnicas de visualización.

Mucha gente, sobre todo de mayor edad, cree que las valoraciones en este tipo de webs son en su mayoría falsas o que la gente siempre prioriza los lugares más baratos. Es esto así?

Esta pregunta no se puede responde ya que no disponemos de los datos de cada una de las reseñas. De tener estos datos, es probable que mediante técnicas de procesamiento del lenguaje natural se supiese que reseñas son buenas y cuales son malas para ver cuales tienen más peso sobre la valoración media del local, o para identificar tendencias (una vez que empieza a tecibir críticas malas no se da detenido la tendencia negativa), detectar críticas falsas vs verdaderas, etc.