Master: Ciencia de Datos.

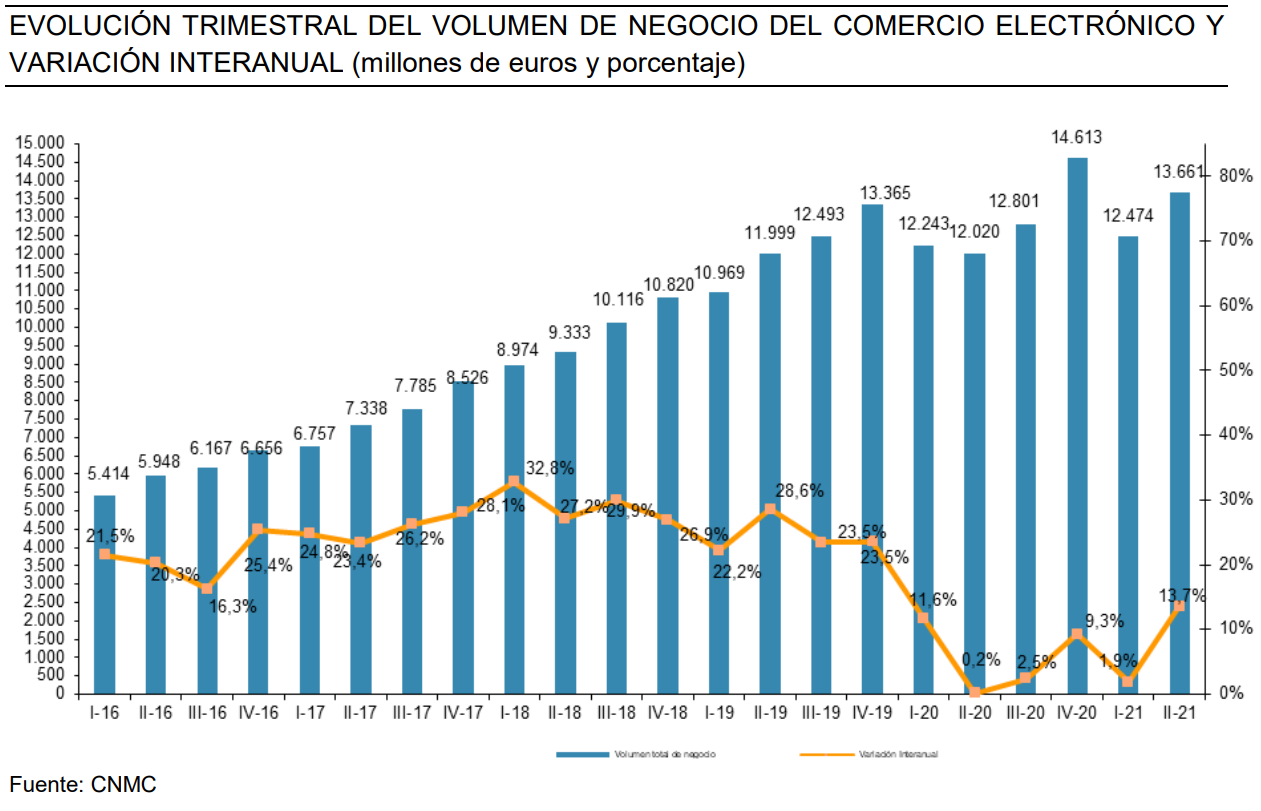
Asignatura: M2.851 – Tipología y ciclo de vida de los datos.  
Actividad: PRA1 – Web scraping.  
Grupo de trabajo: sguillen1 – Sandra Guillén Resina  
 cperezceb – Carlos Pérez Cebrián

## Contexto. Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información.

La empresa **ATiempo** es una empresa de transporte familiar donde el beneficio está más que ajustado. Abaratar costes se ha convertido en un reto muy difícil dada la competencia y el contexto socioeconómico en el que nos encontramos.

Un cambio de los hábitos de consumo ha supuesto que, durante la última década, el comercio electrónico haya experimentado un fuerte crecimiento con una tendencia al alza.

Según los datos publicados por la [1Comisión Nacional de los Mercados y la Competencia (CNMC)](https://www.cnmc.es/), las ventas del negocio en la red en España han superado los 13.600 millones de euros en el segundo trimestre de 2021, un 13,7% más que el mismo período de 2020.



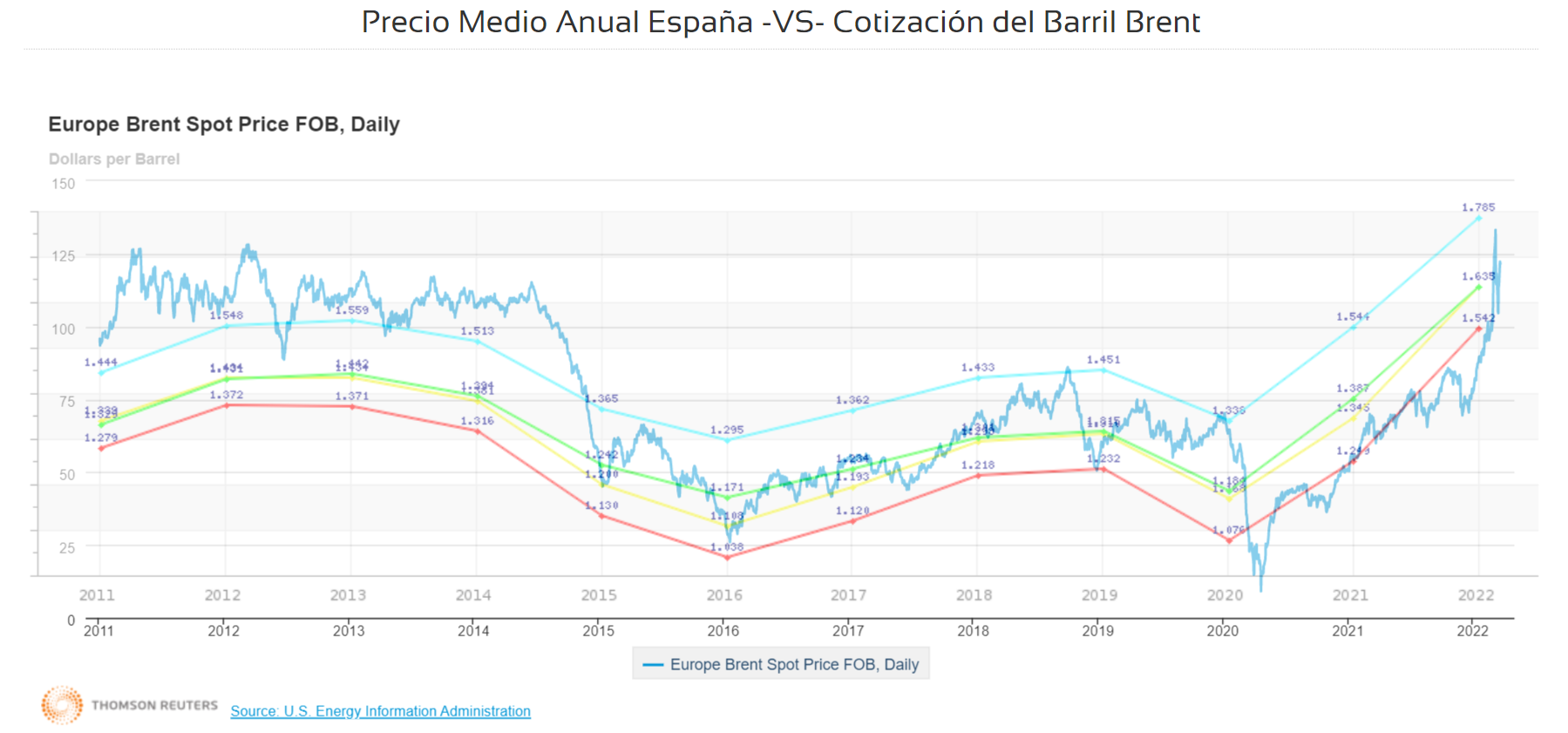
Análogamente, se estima que el volumen de negocio del mercado de paquetería empresarial registre una aceleración en su ritmo de crecimiento. Sin embargo, los acontecimientos que se están dando a nivel mundial, como la pandemia por SARS-CoV-2 o la guerra en Ucrania, están provocando la caída del consumo y la actividad industrial, y consecuentemente, una disminución en la paquetería industrial.

Este marco geopolítico en el que nos encontramos, ha provocado una subida extrema y repentina de los precios del combustible; costes que no puede asumir el sector de transportista, dando a lugar a paralizaciones y protestas.

**Protestas de transportistas en Italia, Portugal y Alemania por los altos precios que los asfixian**

Transportistas en pie de guerra en Europa por los elevados precios del combustible. El sector se ve asfixiado y su protesta espontánea ha traspasado froteras. Con la guerra de Ucrania el precio del combustible se ha disparado en todas partes, **por encima de los 2 euros en Italia.**

[2Fuente Euronews.com](https://es.euronews.com/2022/03/14/protestas-de-transportistas-en-italia-portugal-y-alemania-por-los-altos-precios-que-los-as)



Por lo que, nos encontramos en un sector en auge, como es el de la mensajería, con una sociedad con unos hábitos de consumo diferentes pero con una situación social que hace que el mercado sea inestable.

Por este motivo, **ATiempo** ha iniciado un proceso de transformación digital donde su principal objetivo se centra en la mejora del servicio y trato para con el cliente. Para ello, desarrollarán una app con la que el cliente tenga mayor facilidad para tramitar envíos y, sobre todo una mejor trazabilidad de sus envíos o peticiones. El agilizar las rutas de entrega, seguida por sus transportistas, es primordial. Las unidades de movilidad estarán dotadas de sistemas de geolocalización, el conductor deberá ceñirse a las hojas de rutas preconfiguradas el día anterior y adaptadas a las circunstancias del tráfico y las entregas. Como parte del reparto se incluirá la operación del repostaje. En el momento que se detecte tal necesidad, el sistema incluirá la ruta a la gasolinera más cercana con un precio más bajo, ese es el cometido del presente dataset.

## Título. Definir un título que sea descriptivo para el dataset.

FuelPriceScraper

## Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.

El dataset resultante proporciona la información necesaria de las estaciones de servicio que tiene el producto más barato y esté a una distancia máxima de 10km a la posición de la unidad móvil.

Información de la gasolinera y localización

* Provincia (cadena de caracteres)
* Localidad
* Municipio
* CPostal: Código Postal
* Dirección
* Latitud
* Longitud
* Distancia (decimal): distancia en km desde la posición actual del vehículo geolocalizado.
* Rótulo

Datos de productos:

* GasComprimido
* GasLicuado
* GasoleoA
* Gasolina95E5

Ejemplo Dataset

#definimos una estructura  
EstacionServicio= {  
 "CPostal":"04510"  
 ,"Latitud":"37,144750"  
 ,"Longitud (WGS84)":"-2,772778"  
 ,"Localidad":"ABLA"  
 ,"Municipio":"Abla"  
 ,"Provincia":"ALMERÍA"  
 ,"GasComprimido":""  
 ,"Gas Licuado":""  
 ,"GasoleoA”:"1,899"  
 ,"Gasolina95E5":"1,829"  
 ,"Rótulo":"CEPSA"

,”Dirección”:” CARRETERA M-114 KM. 1”

,”Distancia”: 22.12  
 }

## Representación gráfica.

## Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.

Completa con un Infograma: facturación del comercio electrónico, precio de combustibles, inflación, descenso de producción e inversión por parte de las petroleras…

## Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se han recogido.

## Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

El precio actualizado de los carburantes se obtiene desde el portal que ofrece el [4Ministerio para la Transición Ecológica y el Reto Demográfico](https://energia.gob.es/es-es/Paginas/index.aspx).

La información se actualiza cada 5 minutos. Los operadores mayoristas deben remitir al Ministerio el precio que comunique al gestor y éste a su vez también debe remitirlo al Ministerio el precio que fije. [5Geoportal de hidrocarburos](https://geoportalgasolineras.es/#/Inicio) ha sido galardonado con diferentes premios, como puede ser el premio CNIS 2015 al mejor servicio en movilidad.

Nos decantamos por esta fuente de información por ser la información oficial del Ministerio. Barajamos otras fuentes, como puede ser el portal [6dieselogasolina.com](https://www.dieselogasolina.com/) pero estudiando de donde toman los datos vemos que acuden a la misma fuente oficial.

Examinamos el mapa del portal web y detectamos que proporciona mecanismos para poder consultar y descargar la información sobre el precio de los carburantes en las distintas estaciones de servicio. La información está disponible y de acceso libre y no necesitamos un acceso que nos puedan bloquear.

De todos modos, examinamos el archivo “robots.txt” para detectar posibles restricciones.

Utilizamos la función definida en el fichero robots.py de la siguiente forma:

print(robot\_txt('https://geoportalgasolineras.es/#'))

<!DOCTYPE html>  
<html data-ng-app="geoportalApp" lang="es">  
<!-- Faro\_Geoportal -->  
<head>  
<meta http-equiv="Content-Type" content="text/html" charset="UTF-8">  
<meta http-equiv="X-UA-Compatible" content="IE=edge" />  
<meta name="keywords" content="Estaciones de Servicio, precio, gasolina, di&eacute;sel, gas&oacute;leo" />  
<meta name="rating" content="GENERAL" />  
<meta name="subject" content="Computers:Internet:WWW" />  
<meta name="description" content="Geoportal de precios de carburantes en las Estaciones de Servicio" />  
<meta name="robots" content="ALL" />  
<meta name="distribution" content="GLOBAL" />  
<meta name="classification" content="Online Resource" />  
<meta name="author" content="**MINETUR**" />  
<meta http-equiv="Cache-Control" content="no-cache, no-store, must-revalidate" />  
<meta http-equiv="Pragma" content="no-cache" />  
<meta http-equiv="Expires" content="0" />  
<meta name="google-signin-scope" content="profile email">  
<meta name="google-signin-client\_id" content="736064239674-c2eg0fjfailoaglimjo13n38sioc1rfd.apps.googleusercontent.com">  
<link rel="image\_src" href="resources/imagenes/chosen/chosen-sprite.png"></link>  
<link rel="shortcut icon" type="image/png" href="resources/imagenes/favicon.ico"></link>  
<title>Geoportal</title>

El portal está alojado en un servidor Apache Tomcat, está desarrollado en Angular y utiliza la api de Google Maps.

import builtwith  
print(builtwith.builtwith('https://geoportalgasolineras.es'))

{'web-servers': ['Apache', 'Apache Tomcat'], 'programming-languages': ['Java'], 'operating-systems': ['Windows Server'], 'web-server-extensions': ['mod\_jk'], 'widgets': ['AddThis'], 'javascript-frameworks': ['AngularJS', 'Moment.js', 'Underscore.js', 'jQuery', 'jQuery UI'], 'maps': ['Google Maps', 'OpenLayers']}

## Inspiración. Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

## Licencia. Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección: ● Released Under CC0: Public Domain License. ● Released Under CC BY-NC-SA 4.0 License. ● Released Under CC BY-SA 4.0 License. ● Database released under Open Database License, individual contents under Database Contents License. ● Other (specified above). ● Unknown License.

## Código. Adjuntar en el repositorio Git el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R

## Dataset. Publicar el dataset obtenido(\*) en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.

## Vídeo. Se debe hacer entrega de un vídeo explicativo de la práctica en donde cada uno de los integrantes del grupo explique con sus propias palabras tanto las respuestas del proyecto como el código utilizado para llevar a cabo la extracción. El vídeo debe ser enviado a través de un enlace a Google Drive que deben proporcionar, junto con el enlace al repositorio Git, al momento de entregar la práctica.

## Referencias

1. Comisión Nacional de los Mercados y la Competencia (CNMC)

https://www.cnmc.es/

1. Euronews.com

<https://es.euronews.com/2022/03/14/protestas-de-transportistas-en-italia-portugal-y-alemania-por-los-altos-precios-que-los-as>

1. US Energy Information Administration

https://www.eia.gov/

1. Ministerio para la Transición Ecológica y el Reto Demográfico

<https://energia.gob.es/es-es/Paginas/index.aspx>

1. Geoportal de hidrocarburos

<https://geoportalgasolineras.es/#/Inicio>

1. dieselogasolina.com

https://www.dieselogasolina.com/