Informe Práctica 1 - Tipología y ciclo de vida de los datos

Javier Mateo Moreno y Sara García Rodríguez

I. Contexto

Para la realización de esta práctica hemos extraído información de dos sitios webs distintos: El Economista (Accesible online siguiendo el link: https://ranking-empresas.eleconomista.es) e Infocif (Accesible online siguiendo el link: https://www.infocif.es/ranking/ventas-empresas/espana). En el primer caso, El Economista proporciona un ranking de las mejores empresas españolas según su facturación. Para construir este ranking han incluido información sobre la evaluación de las principales 500.000 empresas españolas: la evolución de sus posiciones, el nombre de la empresa, su facturación (en euros), el sector de su actividad y su provincia de origen (todos los datos proviniendo del repositorio de la empresa INFORMA D&B S.A.U). Por otra parte, hemos utilizado como segunda fuente de datos la web de Infocif, que se encarga de realizar un ranking de las mejores empresas españolas según sus ventas. Con este fin, recoge información sobre las ventas (absolutas, medidas en euros) de estas empresas en 2020 y en 2019, así como de su resultado financiero y otros parámetros relacionados. Con estas fuentes como referencia, hemos querido combinar la información específica que se encuentra en ambas bases de datos para obtener un conocimiento más detallado de las principales características de las empresas más exitosas en España. La motivación para seleccionar dos fuentes de datos distintas es que esto nos ha permitido ampliar el alcance y la utilidad de nuestro proceso de web scraping, ya que aunque los datos aislados pueden resultar de utilidad,la combinación de los mismos permite un conocimiento más completo y detallado de la realidad del tejido empresarial español.

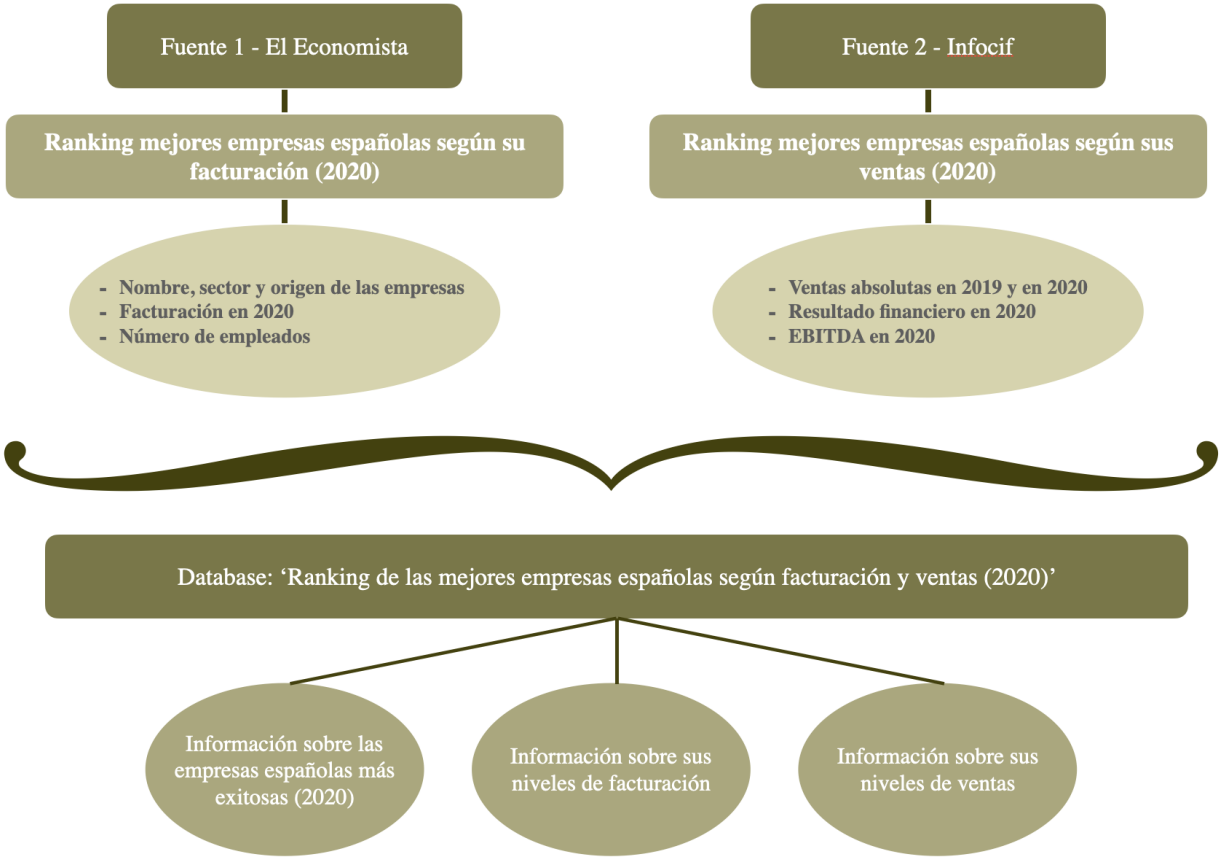
II. Título

En línea con el objetivo mencionado, hemos decidido denominar a nuestro dataset 'Ranking de las mejores empresas españolas según facturación y ventas (2020)'. Consideramos que es apropiado especificar en el título que nuestros datos corresponden al año 2020 -a pesar de que el algoritmo de rastreo podría ser reutilizado para futuros registros de datos-, ya que así evitamos confusiones en la posible reutilización de nuestras contribuciones.

III. Descripción del dataset

Como hemos indicado, nuestro objetivo a la hora de realizar este estudio ha sido combinar ambos rankings de mejores empresas españolas -incluyendo la información relativa a dichas categorizaciones- para obtener una perspectiva más detallada de las características de las empresas españolas más exitosas. Con este fin hemos buscado las coincidencias en las empresas mejor posicionadas según ambos criterios, extrayendo la información relativa a las 50 entidades que ambas fuentes consideran mejor posicionadas. Como resultado, hemos desarrollado un dataset que incluye el nombre de estas entidades, la información relativa al ranking de facturación -población, provincia, sector, empleados, facturación-, y la correspondiente al ranking de ventas -ventas en 2019, ventas en 2020, resultado financiero en 2020 y 'ebitda' (Earnings Before Interests, Taxes, Depreciation and Amortization) en 2020-. Se debe mencionar que, aunque inicialmente desarrollamos el dataset con las 50 mejores empresas, finalmente nos quedamos solo con 21 entidades, ya que solo de estas disponíamos de todos los datos. En resumen, nuestro dataset ofrece una descripción detallada, tanto a nivel de facturación como a nivel de ventas, de las empresas españolas más exitosas actualmente.

IV. Representación gráfica



V. Contenido

Nuestro dataset se compone de catorce variables con información relativa a las empresas más exitosas en España durante el último año. Todos los datos han sido extraídos de rankings anuales, ambos actualizados a fecha de 2020 -por lo que se refieren a la actuación de las empresas en el último año-. Debemos destacar, sin embargo, que algunas de las variables presentadas se refieren a variaciones con respecto al año anterior, 2019 (por ejemplo, la variable 'evolución' se refiere al cambio de posición de la empresa en el ranking con respecto a 2019).

A continuación, describiremos brevemente las variables incluidas en nuestra base de datos. En primer lugar tenemos 'empresa_key', que es simplemente el identificador único de cada una de las entidades. La segunda columna, 'nombre', indica el nombre de la empresa en cuestión. La variable 'rank' nos indica en qué posición se encontraba cada empresa en el ranking original de El Economista. A continuación tenemos información relativa a la población y a la provincia de dónde proviene la empresa, y también encontramos indicado el sector al que se dedica. Para especificar cuál es este sector tenemos dos variables: 'sector', que identifica el ámbito con un código, y 'sector actividad', que lo identifica con un nombre (por ejemplo, el

sector 4711 corresponde al ‘sector actividad’ ‘minoristas de alimentos/medicamentos’). Más adelante tenemos las variables: ‘empleados’ -que indica el número absoluto de personas trabajando para cada empresa a fecha de 2020- y ‘evolución’ que, como ya hemos dicho, nos indica si la empresa se ha movido o no de posición en el ranking con respecto al año anterior. Por último, tenemos las variables relativas a las ventas: ‘ventas 2019’ y ‘ventas 2020’, que sirven para indicar las ventas absolutas en cada uno de los años; ‘resultado 2020’, que establece el resultado financiero; y ‘ebitda 2020’, que indica las ganancias previas a los intereses, impuestos, depreciación y amortización en dicho año.

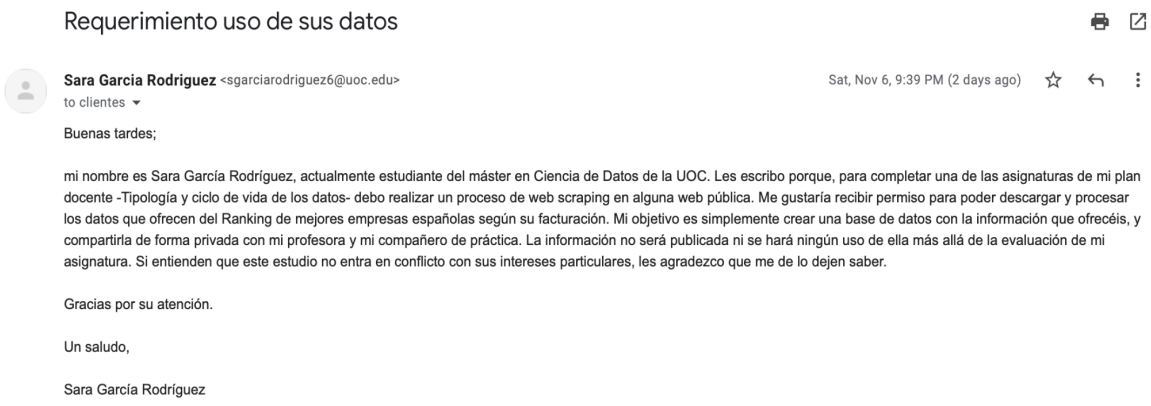
Pasamos ahora a describir cómo hemos recogido los datos. Tras identificar las páginas webs de las que queríamos extraer información, hemos hecho uso de Python para realizar el proceso de web scraping y construir nuestra base de datos. Primero descargamos las páginas para el análisis y, tras comprobar que la petición al servidor fue correcta, importamos BeautifulSoup y convertimos en un objeto de Python el contenido descargado. Con el contexto preparado, pasamos a realizar web scraping sobre la primera página web. Extrajimos la información, la pasamos a un dataframe y comenzamos las tareas de limpieza: pasar a minúsculas, eliminar tildes, caracteres no alfanuméricos, espacios en blanco inoportunos... Tras limpiar, usamos una list comprehension para crear una lista con los nombres de las columnas. De ahí pasamos a obtener las filas de las etiquetas, las celdas por fila, y creamos una nueva lista. El siguiente paso fue generar un dataframe, realizar de nuevo tareas de limpieza (eliminar puntos finales, homogeneizar valores que se habían escrito de forma distinta, etc) y otorgar un identificador único a cada empresa -empresa_key- para evitar problemas con los nombres (ya que algunos estaban escritos de distinta manera en una página y otra).

Tras comprobar que nuestro resultado funcionaba, pasamos a realizar el mismo proceso sobre la segunda web. De nuevo, nos vimos en la necesidad de hacer algunas modificaciones de estilo para mantener la coherencia con respecto a la primera parte de nuestro ejercicio. Una vez hubimos creado y ajustado nuestro segundo dataframe (y habiendo comprobado que efectivamente funcionaba) pasamos a asignar los códigos de las empresas a estas y, finalmente, pudimos ya pasar a unificar nuestras bases de datos. Usando ‘unión izquierda’, juntamos nuestros datos y realizamos los ajustes necesarios (cambios de formato, eliminación de información duplicada, etc). En este punto comprobamos que teníamos muchos valores no especificados, por lo que eliminamos de nuestro database las empresas de las que no disponíamos de toda la información. Así, pasamos de 50 empresas a 21. Con esto, finalizamos nuestra base de datos. Para acabar nuestro análisis realizamos un breve estudio de los componentes del dataframe y, finalmente, exportamos nuestros datos a un archivo csv denominado dataset_empresas_españolas-2.csv.

VI. Agradecimientos

Nuestra primera fuente de datos es el ranking ofrecido en la web ‘El Economista’. Los creadores del ranking especifican en su web que el propietario de los datos, así como quien los trata y suministra, es la empresa INFORMA D&B S.A.U. (S.M.E.). Esta entidad es la propietaria de la base de datos original, que cuenta con información comercial y financiera de empresas españolas obtenida de fuentes públicas y privadas, como por ejemplo: BORME, BOE, Depósitos de Cuentas Oficiales, Boletines Oficiales Provinciales y de CC. AA., informes de prensa, etc. En sus condiciones generales especifican que la información se ofrece de forma gratuita y accesible para todo el mundo; sin embargo, tras la realización de este trabajo hemos comprobado que está explícitamente prohibido realizar cualquier proceso de web scraping sobre estos datos. Tanto el acceso a la información como su almacenamiento a través de procesos de scraping está prohibido si no es con autorización expresa de la empresa; de hecho, especifican también que no está permitido hacer un uso público de los datos. Consecuentemente, no podemos considerar que en un inicio hayamos cumplido con los requisitos éticos de la empresa en la realización de este estudio. Para remediar esa situación hemos decidido cambiar nuestra licencia original e implementar una licencia

privada, evitando así incumplir los requisitos empresariales relativos a la no publicación de los datos. Al mismo tiempo, hemos contactado por email con la empresa proveedora de los datos para intentar obtener una autorización explícita que nos permita usar estos datos. A fecha de entrega de este trabajo no hemos recibido respuesta alguna, pero igualmente dejamos a continuación registro de nuestra solicitud:



Requerimiento uso de sus datos

Sara Garcia Rodriguez <sgarciarodriguez6@uoc.edu> Sat, Nov 6, 9:39 PM (2 days ago)

to clientes

Buenas tardes;

mi nombre es Sara García Rodríguez, actualmente estudiante del máster en Ciencia de Datos de la UOC. Les escribo porque, para completar una de las asignaturas de mi plan docente -Tipología y ciclo de vida de los datos- debo realizar un proceso de web scraping en alguna web pública. Me gustaría recibir permiso para poder descargar y procesar los datos que ofrecen del Ranking de mejores empresas españolas según su facturación. Mi objetivo es simplemente crear una base de datos con la información que ofrecéis, y compartirla de forma privada con mi profesora y mi compañero de práctica. La información no será publicada ni se hará ningún uso de ella más allá de la evaluación de mi asignatura. Si entienden que este estudio no entra en conflicto con sus intereses particulares, les agradezco que me de lo dejen saber.

Gracias por su atención.

Un saludo,

Sara García Rodríguez

Hemos encontrado diversos artículos que han hecho uso de los datos de INFORMA D&B -en concreto a través del ranking publicado por El Economista- para evaluar distintas cuestiones relativas al éxito y/o fracaso del tejido empresarial español, tanto evaluándolo de forma específica como comparándolo con el funcionamiento de las empresas europeas más exitosas. Algunos ejemplos son: 'The communicative management of large companies in Spain: structure, resources and main challenges of those managers'' (Fernández y Vázquez), 'La internacionalización de las empresas españolas, el camino para afianzar la recuperación de la economía' (Bonet y Barrionuevo) o 'Mercadona: las estrategias hacia el éxito' (Del Hierro).

Por otra parte, nuestra segunda fuente de datos ha sido el Ranking de Ventas publicado por Infocif. En este caso, el titular de la web, datos y recursos asociados es INFORIESGOS S.A., cuya principal actividad es prestar servicios relacionados con el uso de bases de datos relativas a información comercial, mercantil, económica y jurídica. En sus condiciones de uso especifican que, a pesar de que Inforiesgos es titular en exclusiva de todos los contenidos del portal, ofrecen un acceso libre y gratuito a sus recursos. En general permiten el uso tanto personal como privado de sus datos, prohibiendo solamente la explotación de los mismos y/o su modificación o alteración ilícita. De hecho, la empresa no prohíbe específicamente la extracción y almacenamiento de sus datos a través de procesos de web scraping, tan solo indican que el usuario debe comprometerse a no bloquear el acceso a la web de otros usuarios, a no destruir ni alterar datos, y a no usar los recursos de infocif con fines ilícitos y/o de forma que pueda perjudicar al portal. En este sentido, consideramos que dados los requisitos legales fijados por el propietario de los datos no hemos realizado un uso ilícito del portal, sino que hemos podido adecuarnos al margen ético y legal de actuación que determina la entidad correspondiente. A pesar de que según los requisitos de esta web podríamos publicar nuestro análisis sin demasiadas restricciones, en tanto que nuestro resultado final es una combinación de dos fuentes distintas tenemos que adaptarnos a los requisitos más restrictivos, por lo que tenemos que mantener para nuestros datos una licencia privada.

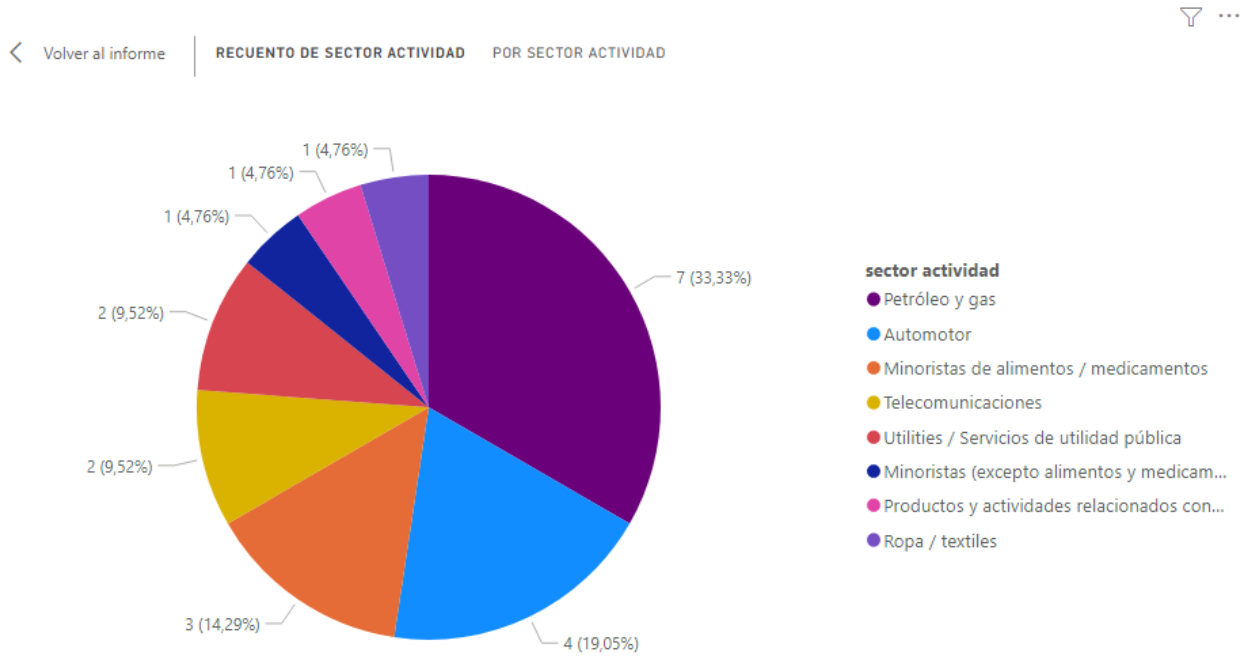
De nuevo, hemos encontrado numerosos estudios que han hecho uso de los datos proporcionados por Infocif. En este caso, la mayor parte de los análisis encontrados se centran en la evaluación de la actividad empresarial dentro de España, aunque podemos encontrar, en general, dos tipos de investigaciones distintas: las que se centran en analizar un factor de éxito concreto haciendo uso de los datos de Infocif para

respaldar sus tesis, y las que quieren analizar de forma más genérica cómo se produce el éxito empresarial en el contexto de la estructura social y económica española -tanto en general como en referencia a un sector económico concreto-. Algunos de los ejemplos más destacables son: 'Estudio del impacto del Private Equity en las empresas y su relación con el crecimiento de la economía española' (García), 'Análisis económico y financiero de las empresas españolas del sector de la automoción' (Vera y Sanz), 'La transformación, una revolución' (González) y 'Análisis y formulación estratégica: Heineken España S.A.' (Bellido).

VII. Inspiración

Consideramos que nuestra base de datos puede resultar de utilidad tanto para realizar análisis sobre el tejido empresarial español -en la línea del mencionado con anterioridad, 'Análisis económico y financiero de las empresas españolas del sector de la automoción' (Vera y Sanz)-, como para que las empresas existentes puedan replantearse su estrategia, estructura y funcionamiento -por ejemplo, similares a los estudios mencionados anteriormente: 'Mercadona: las estrategias hacia el éxito' (Del Hierro) o 'Análisis y formulación estratégica: Heineken España S.A.' (Bellido)-. En este sentido, puede responder a preguntas sobre el escenario español como: ¿cuáles son los sectores empresariales más exitosos en España? ¿Qué empresas dominan un determinado sector de la economía nacional? ¿Qué Comunidades Autónomas (o qué ciudades) hospedan las empresas que generan más dinero? ¿Cuántas personas trabajan para X empresa exitosa? ¿Cuál es la relación entre la cantidad de empleados de una empresa y sus ventas absolutas?

A continuación, incluimos como ejemplo algunos gráficos de elaboración propia que representan posibles análisis que pueden realizarse con nuestra base de datos:



Número de veces que aparece cada sector en el ranking. Los sectores de Petróleo y gas, automoción y minoristas de alimentos y medicamentos son los sectores que más importancia tienen en el ranking.

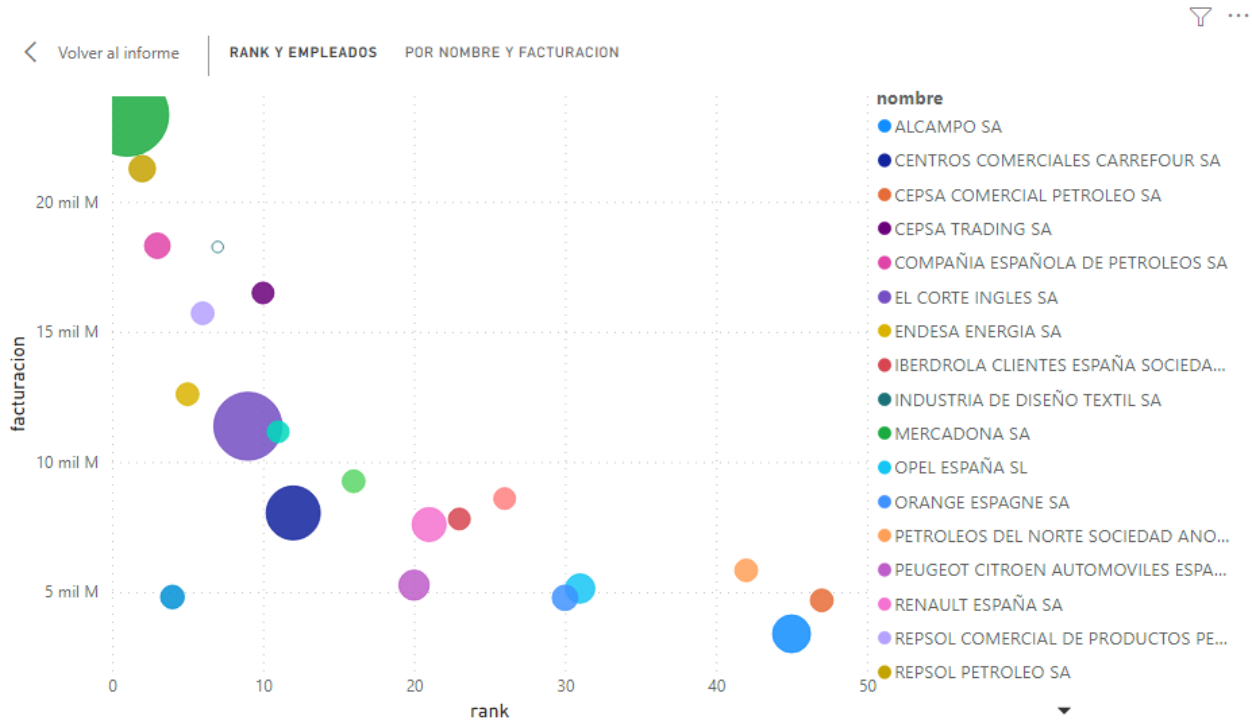
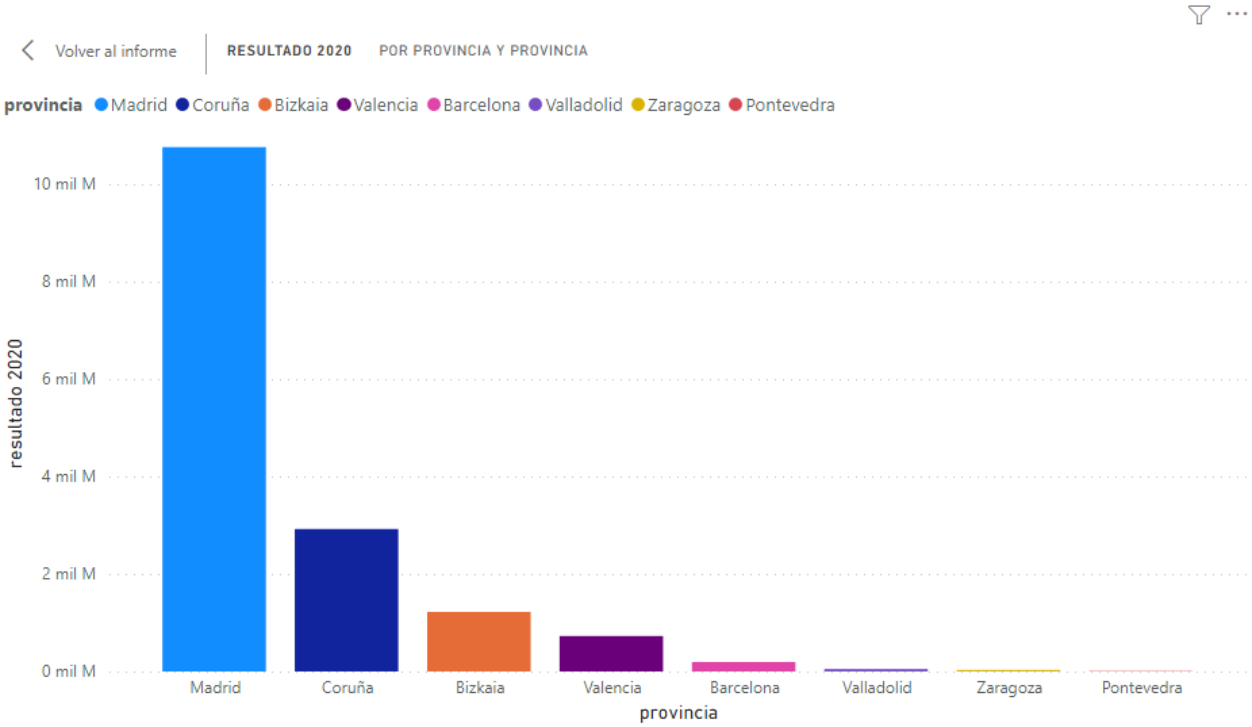


Gráfico de facturación según posición en el ranking. Se encuentra cierta relación entre ambas variables. El tamaño del círculo indica el número de empleados de la empresa.



Suma de resultados agrupados según ciudad. Nótese que salvo Bizkaia y Pontevedra, todas las provincias son capitales de comunidades autónomas.

VIII. Licencia

En un primer momento pretendimos publicar nuestro código bajo la licencia GPL -que permite darle a los datos un uso de difusión comercial, libre y gratuito-, ya que consideramos que el software libre es la mejor opción para promover la difusión y construcción acumulativa del conocimiento social. No obstante, hemos visto necesario modificar los términos de nuestra publicación -limitándolos bajo una licencia más restrictiva- para así evitar que se diera un uso de los datos que no fuese en acuerdo con las condiciones generales de uso fijadas por la entidad propietaria de los datos originales.

En consecuencia, hemos decidido finalmente publicar nuestro proyecto según las normas de la licencia CC BY-NC 4.0, ya que estas imponen unas restricciones en el uso comercial de nuestros datos que entran en acuerdo, en mayor grado, con los requisitos de los proveedores de datos. Esta licencia permite el uso del código siempre que se atribuya la autoría del código fuente y no se utilice de manera comercial.

Si el uso del código se realiza de manera particular, los daños generados a los servidores son reducidos siempre y cuando no se haga con malicia. En cambio, si se realiza un uso comercial de los datos puede aumentar el tráfico de las páginas web scrapeadas, yendo incluso en detrimento de los intereses de las empresas. Aquí consideramos que, a pesar de tener nosotros intereses particulares en conseguir scrapear con facilidad los datos, la seguridad de las empresas scrapeadas es tremendamente valiosa, ya que estas aportan un servicio social informativo de extrema relevancia. Es por esto que queremos destacar que, a pesar de que las restricciones impuestas nos hayan dificultado la ejecución de este estudio, comprendemos y respetamos los límites establecidos por la institución, y nos sentimos simplemente agradecidos por el trabajo informativo que realizan.

IX. Código

Hemos realizado nuestro código con Python. Se encuentra publicado en nuestro repositorio de Github, accesible siguiendo el siguiente enlace:

https://github.com/javiermm1995/PRAC1/blob/main/Código_Web_Scraping.ipynb

X. Dataset

El dataset puede encontrarse en nuestro repositorio de Github en formato csv, o publicado en Zenodo. Se puede acceder a él siguiendo el siguiente enlace:

https://zenodo.org/record/5651271#.YYgsmmDMLIU

El dataset incluye las siguientes columnas:

- Empresa_key - Identificadores únicos de cada empresa
- Nombre - Nombres de las empresas
- Rank - Posición de las empresas en el ranking de El Economista
- Población - Población de la que proviene la empresa
- Provincia - Provincia de la que proviene la empresa
- Sector - Código del sector de actividad empresarial

- Sector actividad - Nombre del sector de actividad empresarial. Existe una relación unívoca con el código del sector.
- Empleados - Número total de empleados de la empresa en el año 2020
- Evolución - Número de puestos que ha variado la empresa en el ranking en el año 2020 con respecto a 2019.
- Facturación - Facturación total de las empresas en el año 2020, en euros
- Ventas 2019 - Ventas totales de la empresa en el año 2019, en euros
- Ventas 2020 - Ventas totales de la empresa en el año 2020, en euros
- Resultado 2020 - Resultado financiero de la empresa en el año 2020, en euros
- Ebitda 2020 - EBITDA (beneficio bruto) de la empresa en el año 2020, en euros

XI. Contribuciones

Contribuciones	Firma
Investigación previa	SGR y JMM
Redacción de las respuestas	SGR y JMM
Desarrollo del código	SGR y JMM