2-5-2024

Informe de la adquisición y limpieza de datos

Sabor en números:
Análisis de restaurantes en España

Proyecto: Comprensión de Datos
Ciencia de Datos- UPV

KENTO KAMAKURA GIMENO, PABLO LUCAS MORA,
DANIEL HERNANDO GRABOLEDA, MIGUEL GIL JIMENEZ

Proceso de adquisición de datos:

Una vez decidimos en grupo que queríamos analizar los datos de los restaurantes en España para encontrar relaciones con el precio medio, buscamos en Kaggle y otras plataformas de datos con el fin de descargar los datos directamente. Sin embargo, no encontramos ninguna base de datos que pudiese satisfacer nuestras necesidades. Por ello, se nos ocurrió aplicar técnicas de webscraping a The Fork, con el objetivo de hacernos con los datos de todos sus restaurantes reales. Aunque esta tarea nos llevase algo más de tiempo, éramos conscientes de la necesidad de realizarlo por la falta de datos y para asegurarnos la veracidad de los datos.
Para ello creamos un software con Python haciendo uso de librerías como Selenium, Requests, Beautifulsoup y Pandas. Este software se puede encontrar en nuestro repositorio de GitHub público.

Además, aprovechamos el potencial de Python en este proceso de adquisición de datos para ir limpiando algunos datos durante la propia adquisición de datos como puede ser eliminar el símbolo "€" de la variable precio medio o convertir la variable "Estrella Michelin" en un booleano, cuando originariamente era un campo nulo o una url de la estrella michelín.

Una vez completado este proceso de adquisición de datos de The Fork procedimos a convertirlo en un CSV para facilitar su envío y posterior tratamiento de los datos.
Por otro lado, consideramos interesante inicialmente conseguir una base de datos de Población, Salario medio y Asalariados por provincia tal y como establecimos en nuestros objetivos. Habiendo adquirido estas 3 bases de datos del INE, procedimos a unir las 4 base de datos finalmente en un CSV con el fin de comenzar a limpiar y posteriormente explorar los datos.

Proceso de limpieza y tratamiento del dato:

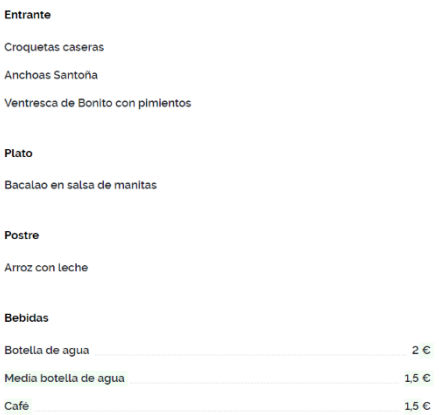
Una vez ya teníamos la base de datos, pasamos a la limpieza y el tratamiento de los datos. Para ello fuimos columna por columna haciendo una pequeña previsualización de los datos que se obtenían en cada variable (análisis univariante). Al realizar este análisis nos dimos cuenta de que algunas de ellas tenían datos faltantes o anómalos, que debíamos revisar para ver qué tratamiento les dábamos a cada caso.

Uno de los primeros datos faltantes que observamos fue en la columna de 'Rate_distinction', en la cual aparecían los restaurantes que tenían algún tipo de distinción (Excelent, Fabulous o Very Good). Esta variable tenía una gran cantidad de datos faltantes, pero tras estudiarlos observamos que no eran errores, ya que no todos los restaurantes tenían porque tener distinción. Por tanto, como estos faltantes aportaban un valor a la variable, los dejamos en la columna. Y además, para un más fácil análisis de estos datos, decidimos hacer una asignación de un número a cada distinción, por su orden de magnitud (1 Very Good, 2 Fabulous y 3 Excellent).

El siguiente cambio que realizamos fue la transformación de la variable 'Michelin' (cuyos valores eran booleanos) a valores numéricos (1 True y 0 False), pero solo con el objetivo de facilitar el análisis, no por algún tipo de fallo.

Al proseguir con nuestro análisis univariante, en la columna de 'Provincia', encontramos que había más valores únicos que provincias en España, lo que nos hizo analizar con más detenimiento lo que sucedía. Observamos que en algunos, The Fork había introducido el nombre del pueblo o ciudad a la que pertenecían en vez de la provincia. Como en este caso se trataba de pocos datos erróneos, los decidimos corregir manualmente. Para ello, buscamos en Google Maps el nombre de la provincia a la que pertenecía el pueblo o ciudad, y lo imputamos.
Posteriormente, observamos que en las columnas de precios medios y de tipo de comida por cada restaurante habían algunos datos faltantes, pero al tratarse de muy pocos (menos de 90) decidimos introducir los datos manualmente mediante la búsqueda en internet de estos mismos (Precio medio aportado por Google).

Continuando con el análisis de todas las columnas, después de la unión de las bases de datos de salarios medios, asalariados y población por provincia con nuestra base inicial vimos que no estaban los datos de Navarra y País Vasco de asalariados y sueldo medio. Al encontrarnos con este problema decidimos buscar estos por internet, e implementarlos manualmente. En el caso de los salarios medios anuales por provincia encontramos fuentes fiables y contrastadas en distintas páginas, por lo que los introducimos en la base de datos. En cambio, al buscar los datos de asalariados manualmente por internet, no fuimos capaces de encontrar fuentes fiables a diferencia de los anteriores. Y los de población no nos dieron ningún tipo de problema, ya que estaban completos y eran fiables. Por lo que finalmente decidimos quedarnos solo con la columna de salarios medios y población.
Para cerrar, tras una última revisión decidimos eliminar los datos en los cuales el precio medio era menor de 6. Esta decisión la tomamos debido a que significaban una fracción muy pequeña de la población (4 datos), y que, además, analizar uno por uno nos dimos cuenta de que se debía a un error de The Fork, como por ejemplo el caso de un restaurante en el cual aparecía un precio medio de 1 euro y que tras mirarlo manualmente en The Fork observamos que se debía a que solo aparecían los precios de las bebidas y no los de los platos (Por lo que The Fork realiza la media solo de los precios de las bebidas).



Entrante

Croquetas caseras

Anchoas Santoña

Ventresca de Bonito con pimientos

Plato

Bacalao en salsa de manitas

Postre

Arroz con leche

Bebidas

Botella de agua	2 €
Media botella de agua	1,5 €
Café	1,5 €

Además, observamos que había algunos restaurantes con 0 métodos de pago, cosa que es imposible ya que tiene que existir alguna manera para pagar, por lo que hemos decidido prescindir de estos , ya que está claro que son datos erróneos. Como eran pocos datos se decidió eliminar estas filas.