Práctica 1 Análisis de Datos y Visualización

Introducción

En esta práctica, se nos introduce por primera vez al mundo del análisis de datos. Nuestro objetivo principal es aplicar técnicas de limpieza y transformación para poder extraer conclusiones de estos. Algunas de las técnicas que vamos a utilizar son las siguientes: filtrado, integración de fuentes, conversión de tipos…

Disponemos de tres *datasets*, los cuales tienen cada uno unas patologías distintas, por lo que habrá que abordarlos por separado y tratar sus respectivos problemas.

Este proceso nos permitirá extraer conclusiones de valor de los datos al haber eliminado los errores, datos innecesarios y peculiaridades, y sacar toda su información potencial.

Datos proporcionados

- **'contac_center_data.csv':** se trata del primer documento a analizar. Contiene datos de llamadas de usuarios, el Código postal, su teléfono, su id de sesión, información sobre su Vivienda (chalet/piso, unifamiliar/adosado/bajo/intermedio, con/sin rejas, con/sin perro) y por último el tipo de producto que tiene contratado (*home basic/home premium/home premium plus*).
- **'renta_por_hogar.csv':** se trata del Segundo documento a analizar. dispone de los datos de la renta neta media por persona, la renta neta media por hogar, Media de la renta por unidad de consumo, Mediana de la renta por unidad de consumo, Renta bruta media por persona y Renta bruta media por hogar de cada Municipio, Distrito y Sección de la comunidad de Madrid, indicando en cada uno de estos el código postal y el código de cada Distrito/Sección. Además contiene los datos desde 2015 hasta 2020, almacenando el total de cada dato indicado anteriormente.
- **'delitos_por_municipio.csv':** se trata del tercer documento a analizar. Dispone de una cabecera y pie de página con información sobre los datos del documento. Los datos nos indican diferentes delitos cometidos en enero-marzo de 2019 y enero-marzo de 2020.

Transformaciones de los datos

'contac_center_data.csv'

En primer lugar comprobamos los distintos valores “NaN” que tiene el documento. Observamos que la columna “Producto” tiene 20605 valores “NaN”, al igual que las tres últimas filas. En el caso de las tres últimas filas, vemos que los datos que sí tiene están repetidos, por lo que podemos eliminarlas.

Autor: Martín de las Heras

Los datos “NaN” en producto se tratan sustituyéndolos por una celda vacía, ya que dejar el “NaN” puede ocasionar problemas a la larga, y este dato nos indica los registros que no tienen ningún tipo de servicio contratado.

A continuación, pasamos al *feature engineering*, la columna “funnel_Q”, la dividimos en 5 columnas distintas en función de las características de la vivienda:

- Si es casa o piso
- En caso de que sea piso, si es bajo o intermedio
- En caso de que sea casa, si es adosado o unifamiliar
- Si tiene rejas o no
- Si tiene perro o no

De esta manera reducimos el número de filas, lo cual acelera la visualización y posible filtrado de los datos.

'renta_por_hogar.csv'

En este fichero vemos que hay 175069 filas, donde se aprecia que los datos de secciones se encuentran contenidos en los de distritos y estos en los de municipios. De la misma manera, vemos que para los años anteriores a 2019, no disponemos casi de datos, por lo que no podremos sacar ninguna conclusión fiable de ellos.

Con estas conclusiones, eliminamos todas las filas con fecha menor a 2019 y la columna de sección.

Para poder relacionar mejor este archivo con los otros dos, dividimos la columna “Municipio” en una columna que sea municipio y otra que sea el código postal. A su vez, creamos una columna de “Total 2019” y otra de “Total 2020” para mostrar la información de manera más clara.

Por último, eliminamos las filas que no tengan la renta neta media por persona, ya que es el indicador más relevante de cara a los datos de los seguros.

'delitos_por_municipio.csv'

El último fichero, empieza con una edición a mano, donde quitamos la cabecera y pie del documento, ya que no nos aportan información válida.

Ya desde el Código, eliminamos la última columna, la cual estaba vacía, y la primera fila, la cual no era más que una suma de los demás valores de las filas.

Las dos primeras filas del documento entonces son dos cabeceras, por lo que las juntamos, poniendo primero el tipo de delito (cabecera superior) y debajo el período de tiempo en el que se miden esos datos.

Para terminar, en la primera columna, en todos aparece municipio de encabezando el nombre, por lo que podemos eliminarlo. Esto nos servirá de cara a la unión de los ficheros en un único *dataset*.

Unión de los datos

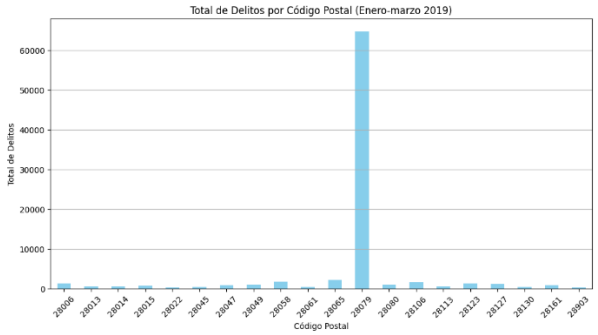
Autor: Martín de las Heras

Para unir los tres ficheros, tomamos las columnas de municipio y código postal como puntos de unión entre ellas. Y nos quedamos solamente con los datos en común entre los ficheros, es decir, los datos que Podemos relacionar entre sí.

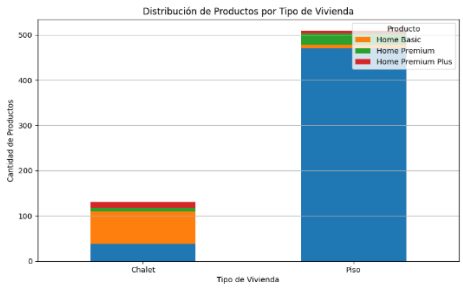
Conclusiones

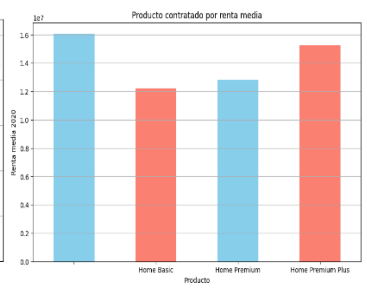
Para extraer las conclusiones, hemos decidido graficar algunas relaciones entre datos, de manera que se pueda extraer la información inicialmente de manera visual, y de ahí poder tomar más conclusiones:

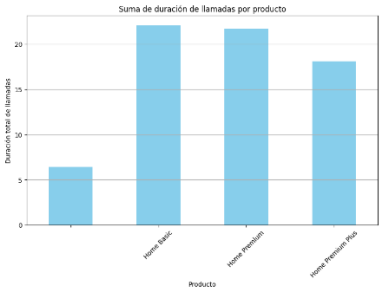
En primer lugar, vemos que no tiene sentido comparar los delitos por código postal, ya que, en valores absolutos, el municipio de Madrid siempre tendrá un número anormalmente grande respecto al resto, como podemos ver en esta gráfica:

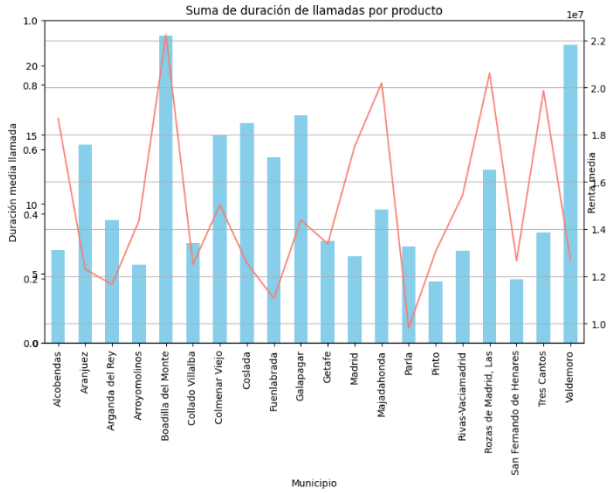


A continuación, vemos cómo afecta el tipo de producto a distintos aspectos como el tipo de vivienda, la renta media o la duración de las llamadas al contact center:









Por último, tenemos la comparación de la duración de las llamadas por municipio, superpuesto a la renta media de cada uno, donde podemos ver que las llamadas más largas, se suelen realizar en los municipios con menos renta media.

Como estos gráficos se pueden sacar muchos más, de manera que se expliquen más relaciones entre sí.

Autor: Martín de las Heras