CLASIFICACIÓN DE LOS BARRIOS DE LA CIUDAD DE MADRID SEGÚN EL NIVEL DE IDONEIDAD PARA FAMILIAS CON HIJOS

Análisis Exploratorio de Datos

Memoria Descriptiva

Jesús Llanes Tamayo

Bootcamp Data Science Sep. 2021

The Bridge | Digital Talent Acelerator

Introducción.

En nuestro análisis exploratorio nos propondremos encontrar variables que sean capaces de servir de ayuda a los padres al escoger donde desean residir, teniendo en cuenta criterios que puedan ser importantes para la vida de sus hijos. Limitaremos la búsqueda a los barrios de la ciudad de Madrid. Al tratarse de una gran ciudad, los barrios que la conforman cuentan con características muy variadas, que los pueden hacer idóneos o no, para una familia con hijos pequeños.

Nuestra intención es que al final del análisis exploratorio, contemos con un conjunto de variables que al menos, pudieran servir de apoyo ante la casuística planteada.

Variables planteadas inicialmente.

De inicio, para conseguir nuestro objetivo, nos planteamos utilizar los siguientes criterios/variables:

1- Disponibilidad de colegios en el barrio (cantidad de colegios).
2- Cantidad de colegios bilingües (cada vez preocupa más que desde edades tempranas comiencen el aprendizaje de lenguas extranjeras).
3- Disponibilidad de centros de salud en el barrio (cantidad / barrio)
4- Cantidad de instalaciones infantiles en el barrio.
5- Datos de población. Especialmente la proporción de niños sobre la población total (dada la necesidad de los niños de sociabilizar con sus iguales, se intentaría evitar los barrios muy envejecidos)
6- Precio medio de la vivienda por metro cuadrado. Más que un elemento de clasificación como un filtro que poder aplicar a la búsqueda.

Obtención de los datos.

Para la obtención de los datos, recurrimos al Portal de Datos Abiertos del Ayuntamiento de Madrid (https://datos.madrid.es/portal/site/egob) y al Portal Web del Ayuntamiento de Madrid (https://www.madrid.es/portales/munimadrid/).

Nuestros archivos de datos son los siguientes:

1- Datos de centros médicos de la ciudad de Madrid.

(https://datos.madrid.es/egob/catalogo/212769-0-atencion-medica.csv).

Este archivo de datos cuenta con 248 entradas de datos y 31 columnas.

Una de las columnas es “Barrio” y contiene los nombres de los barrios en los que se encuentran situados los centros. Es la que tenemos previsto utilizar para la unión con el resto de fuentes de datos.

2- Datos de colegios de la ciudad de Madrid.

(https://datos.madrid.es/egob/catalogo/202311-0-colegios-publicos.csv).

Este archivo de datos cuenta con 248 entradas de datos y 32 columnas.

Una de las columnas es “Barrio” y contiene los nombres de los barrios en los que se encuentran situados los centros. Es la que tenemos previsto utilizar para la unión con el resto de fuentes de datos.

3- Datos de instalaciones infantiles (áreas infantiles) de la ciudad de Madrid.

(https://datos.madrid.es/egob/catalogo/200652-8-areas-infantiles.csv)

Este archivo de datos cuenta con 2044 entradas de datos y 14 columnas.

Una de las columnas es “NOMBRE_BAR,C,50” y contiene los nombres de los barrios en los que se encuentran situadas las áreas infantiles. Es la que tenemos previsto utilizar para la unión con el resto de fuentes de datos.

4- Datos de población de la ciudad de Madrid (datos del Padrón Municipal).

(https://datos.madrid.es/egob/catalogo/209163-192-padron-municipal-historico.csv)

Este archivo de datos cuenta con 238148 entradas de datos y 13 columnas.

Una de las columnas es “DESC_BARRIO” y contiene los nombres de los barrios en los que se encuentran registradas las personas por rango de edades. Es la que tenemos previsto utilizar para la unión con el resto de fuentes de datos.

5- Datos de precios de la vivienda por metro cuadrado. Estos datos se los facilita el portal Idealista al ayuntamiento de Madrid.

(https://www.madrid.es/UnidadesDescentralizadas/UDCEstadistica/Nuevaweb/Edificaci%C3%B3n%20y%20Vivienda/Mercado%20de%20la%20Vivienda/Precios%20de%20la%20Vivienda/Distritos/E3320221.xls)

En este caso, es un archivo Excel. El resultado al importarlo es un archivo de datos con 187 filas y 3 columnas.

Una de las columnas es “Distrito” y en ella se incluyen los nombres de los barrios para los que hay registrados datos de precios. Es la que tenemos previsto utilizar para la unión con el resto de fuentes de datos.

Limpieza y unión de los datos.

Describiremos brevemente como ha sido el proceso de limpieza y unión de los datos. Explicaremos el proceso para cada uno de los archivos de datos, reflejando los problemas que hemos ido encontrando en cada uno de ellos y la solución a estos.

Dataframe con los datos de centros médicos.

Procedimos a filtrar los datos para quedarnos solamente con los centros de salud. En este punto nos encontramos con el hecho de que en la columna utilizada para este filtro existía además el valor “Centro de Salud Mental”. La solución más rápida en este caso fue eliminar estos valores primero y luego procederá filtrar sin problemas. El siguiente paso fue crearnos un nuevo dataframe que incluyera solamente las columnas “Nombre”, “Barrio” y “Distrito”. Luego, agrupamos los datos por distrito y barrio y contamos la cantidad de centros de salud para obtener nuestro dataframe final.

Dataframe con los datos de colegios.

Nuestro primer paso es crear una nueva columna en la que reflejaremos un 1 si el colegio es bilingüe y 0 si no lo es. Esto lo conseguimos comprobando si en la columna descripción se encuentra la cadena “Enseñanza bilig”. Posteriormente creamos un nuevo dataframe con las columnas “Distrito”, “Barrio”, “Colegio” y “Colegio Bilingüe”. Luego, agrupamos estos datos por “Distrito” y “Barrio” y obtenemos la cuenta de valores de las otras dos columnas para asignarlas a “Cantidad de Colegios” y “Cantidad de Colegios Bilingues”.

En este punto procedemos a unir nuestros dos primeros dataframes. Al unirlos, empleando la columna barrios, nos encontramos con que se generan muchos valores nulos. Estos valores nulos en general se corresponden “Barrios” que no estaban presentes en alguna de las dos tablas. Rellenamos todos estos datos nulos con ceros lo cual es totalmente correcto. En el caso de la columna “distrito” se han creado dos nuevas columnas, las cuales unificamos conservando todos los valores.

Dataframe con los datos de las instalaciones infantiles.

En este caso, directamente renombramos las columnas, agrupamos los datos y obtenemos la cantidad de instalaciones por barrio.

Unimos este dataframe con el anterior y ejecutamos las mismas operaciones: rellenamos los valores nulos con ceros y unificamos las columnas “Distrito” para conservar todos los valores.

Dataframe con los datos de población.

En este dataframe, inicialmente aparecen valores nulos que se corresponden con los rangos de edad (están registrados de año en año) en los que no existían personas. Los rellenamos con cero.

Creamos una nueva columna en la que reflejaremos el total de habitantes por edad ya que en la tabla inicial están divididos por sexo y edad. Agrupamos por distrito y barrio y cuantificamos la población por cada uno de ellos. Realizamos la misma operación, pero cuantificando en este caso solo la población de edad inferior o igual a 12 años.

Dataframe con los datos de precio de la vivienda.

En este archivo de datos nos encontramos con el problema de que los nombres de los barrios (columna que utilizaremos para la unión con los otros dataframes) no coinciden con los que hemos recopilado hasta el momento. Por tanto, hemos tenido que proceder a comparar los datos entre esta tabla y las anteriores para asignar los nombres de barrio. También hemos tenido que rellenar valores vacíos en los precios y eliminar valores nulos que se generaron debido a que el archivo Excel de origen contenía muchas filas vacías como separación. Los precios los hemos convertido a números enteros ya que inicialmente se encontraban en formato “string”.

Obtención de nuevas variables.

Al analizar nuestras variables nos percatamos de que las variables de cantidad de colegios, cantidad de centros de salud y cantidad de instalaciones infantiles, aportan muy poca información. Por tanto, creamos nuevas variables a partir de estas, dividiendo por la cantidad de población de niños en el barrio. El resultado son nuevas variables de ratios que tienen una distribución mucho más variada y que si nos pueden servir como ayuda para nuestra clasificación.

Finalmente calculamos una variable resultado de la suma de todas las ratios que tenemos. Previo a efectuar la suma, estandarizamos todas las ratios, empleando una estandarización de mínimo / máximo.

Visualización de los resultados.

Con el objetivo de observar el comportamiento de nuestras variables y valorar el resultado de nuestro análisis hemos dibujado varios gráficos, utilizando las librerías: Matplotly, Seaborn y Plotly. Hemos utilizado histogramas, gráficos de

violín o gráficos de cajas para visualizar las distribuciones de nuestros datos y gráficos de barras o de puntos para visualizar las variables en sí. Finalmente, hemos utilizado un mapa de calor de las correlaciones entre todas nuestras variables para observar las diferentes relaciones entre ellas.

Presentación.

Nuestra presentación la hemos elaborado en diapositivas Power Point, apoyándonos en todas las gráficas y variables estadísticas que hemos ido obteniendo. Hemos elegido la presentación en diapositivas ya que nos resulta idónea para explicar el proceso del análisis y obtención de resultados, siguiendo un hilo de manera más cómoda.