Universidad Europea

Universidad Europea de Madrid

Escuela de Arquitectura, Ingeniería y Diseño

ACTIVIDAD II: EXPLORATORY DATA ANALYSIS

Análisis realizado en RStudio sobre mi Proyecto Individual

Lenguajes de Programación Estadística

Profesor:
Christian Vladimi Sucuzhanay Arévalo

8 DE DICIEMBRE DE 2020

María Fernández Morín

Número de expediente: 21838529

Usuario de github: fernandezmaria

Objetivo final del proyecto: crear una plataforma que cuente con una clasificación sobre qué tipo de negocios se mantienen mayor tiempo abiertos en una zona, qué negocios no tienen éxito, además de un estudio sobre las carencias comerciales de esas zonas. Todo ello añadido a la "optimización de instalaciones", debido a que existen datos sobre la situación de los locales registrados en la comunidad de Madrid: si están abiertos, cerrados, en reforma o inutilizados. En el caso de los inutilizados, mi idea, junto con lo anterior, es promover una optimización de espacios ya construidos.

Idea: en la siguiente actividad estableceré un primer contacto con los datos encontrados para mi proyecto y realizaré un análisis exploratorio de los mismos en RStudio.

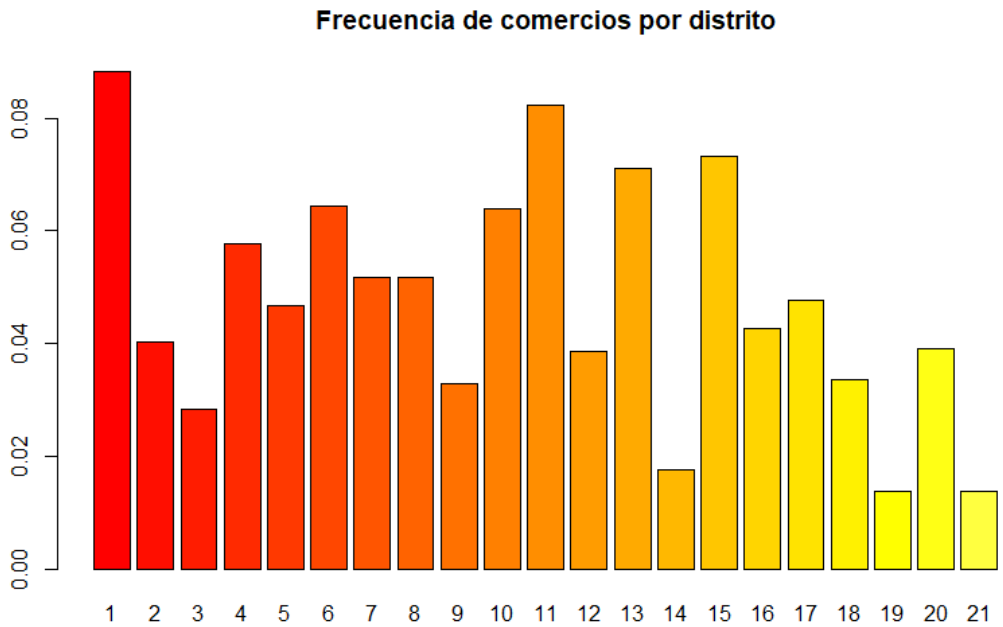
PRIMER CONJUNTO DE DATOS: Censo de Locales Madrid 2020

Este dataset recoge toda la información disponible sobre el censo de locales comerciales de la comunidad de Madrid. Se recogen las características asociadas a cada hueco, en tanto que unidad física (dirección y número de local, accesos –viales-...), y se les asocia la actividad o actividades que se realizan en el mismo, junto con sus titulares.

Con el objetivo de manejar mejor el dataset, el csv se encuentra en una hoja de cálculo publicada en la web y la función read_csv2 se encarga de importarlo a través del url en el que se encuentra.

Dado que en este conjunto de datos encontramos información sobre los locales que se encuentran abiertos, inactivos o en obras, mi objetivo es obtener la frecuencia de locales junto con su situación por distritos.

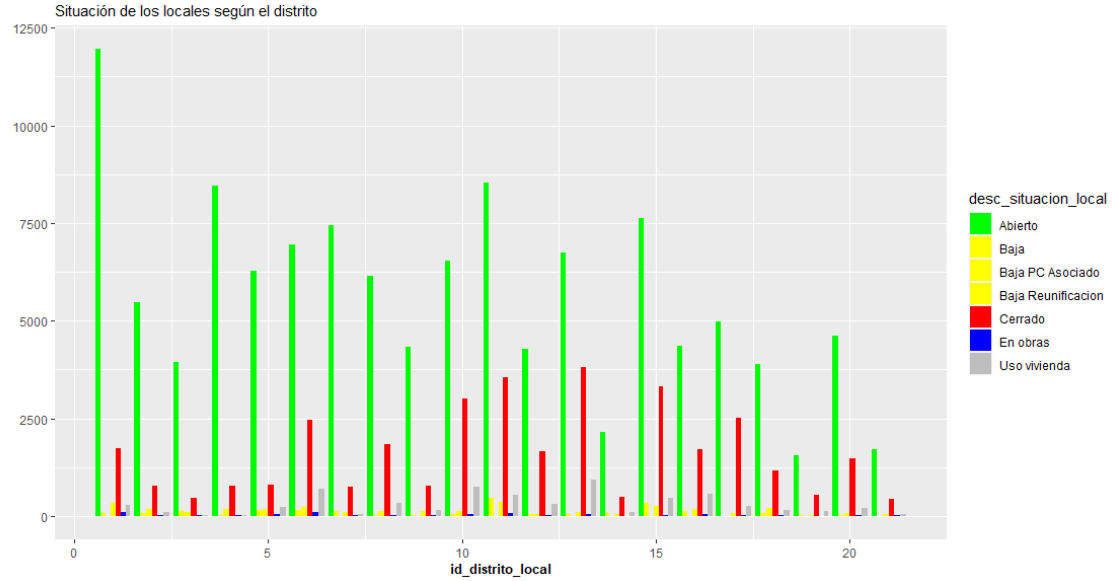
Primero, realicé un barplot para observar la frecuencia de actividad comercial en cada distrito:



Los números colocados en el eje x corresponden al identificador de los diferentes distritos. La conexión es la siguiente:

1	Centro	8	Fuencarral-El Pardo	15	Ciudad Lineal
2	Arganzuela	9	Moncloa-Aravaca	16	Hortaleza
3	Retiro	10	Latina	17	Villaverde
4	Salamanca	11	Carabanchel	18	Villa de Vallecas
5	Chamartín	12	Usera	19	Vicálvaro
6	Tetuán	13	Puente de Vallecas	20	San Blas-Canillejas
7	Chamberí	14	Moratalaz	21	Barajas

Una vez hecho esto, gracias a la librería ggplot consigo el objetivo:



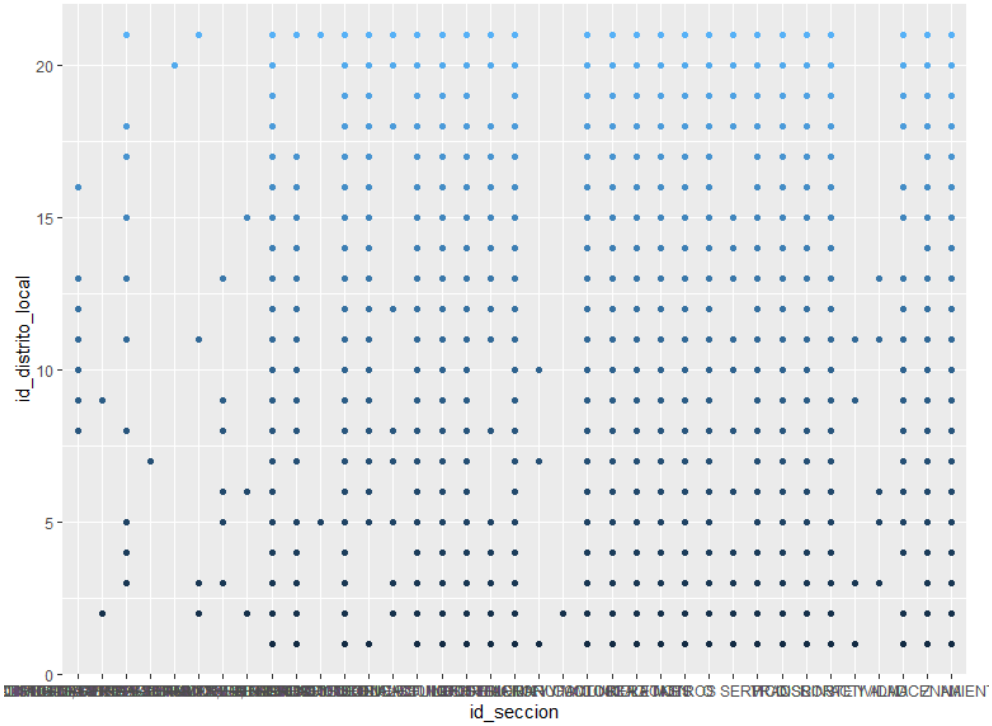
Cada barra verde es un distrito diferente y los colores no han sido colocados al azar, debido a que lo importantes fijarse en la cantidad de locales abiertos e inactivos que hay en cada zona.

Ya con esto se puede afirmar que el distrito Centro es el que cuenta con mayor número de locales y mayor número de locales abiertos, mientras que Vicálvaro cuenta con el mínimo de estos. Sin embargo, Puente de Vallecas se lleva el primer puesto en el podio en número de locales cerrados e inactivos.

SEGUNDO CONJUNTO DE DATOS: Censo de Locales Madrid 2020 y su Actividad

Este dataset contiene datos similares al anterior, pero cuenta con información más detallada sobre la información comercial de cada local. Dicho conjunto de datos se encontrará adjuntado en el repositorio.

El sector al que pertenece cada comercio viene dado en las columnas: id_seccion, desc_seccion, id_division y desc_division. La información que viene detallada en dichos atributos se corresponde con lo que dicta la **Clasificación Nacional de Actividades Económicas (CNAE)** y asigna un código a cada actividad económica de las que se pueden realizar. Generalmente este código (que suele ser de 5 dígitos) se utiliza en muchos formularios e impresos, tanto oficiales como a nivel de empresa. Los grupos principales de la CNAE 2009 los podremos ver en el siguiente dataset a analizar.



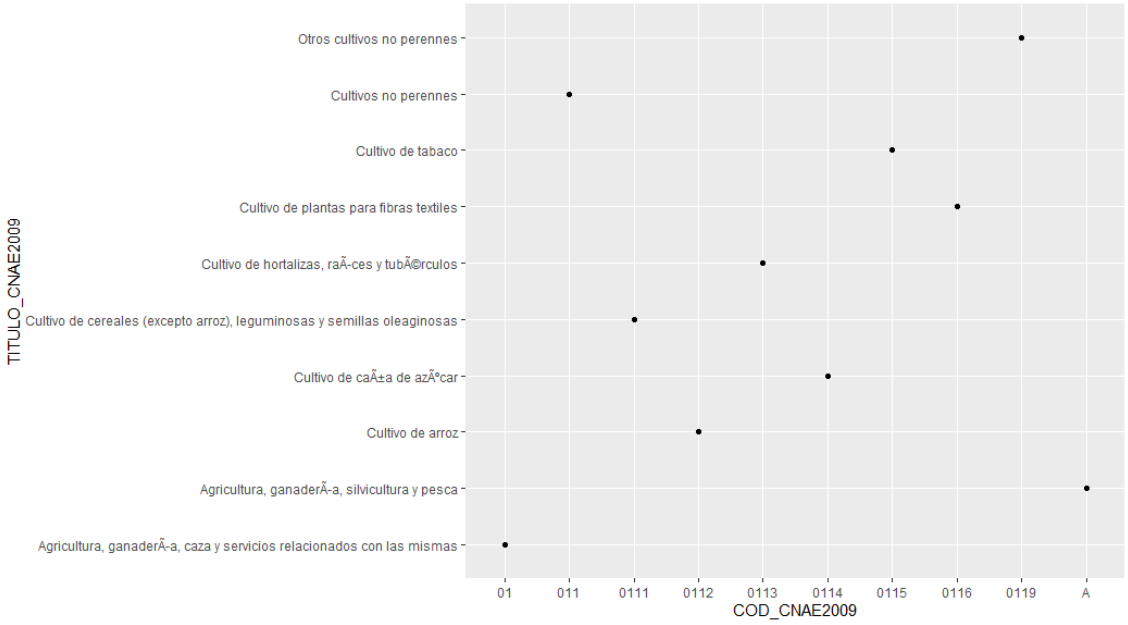
En este hay un problema debido a que los valores únicos que recoge RStudio sobre la variable 'id_seccion' no se corresponden con los que aparecen en el csv original, pero ya lo solucionaremos más adelante. El caso es que aquí podemos observar los distritos que cuentan con mayor riqueza de comercios y los que menos, de forma que se puede identificar qué tipo de actividad comercial falta en ciertas zonas y a partir de ahí averiguar por qué y si se puede solucionar.

TERCER DATASET: Estructura CNAE 2009

En este caso el dataset también está adjuntado en el repositorio, pero para que no haga falta escribir el path en el que se encuentra el conjunto, se podrá leer con la función file.choose(), que abrirá el set de datos que el usuario elija.

El conjunto en cuestión venía con columnas añadidas que estaban vacías y no aportaban información ninguna, únicamente estorbaban porque elevan el número de nulos. Por tanto, realicé un select y guardé los atributos que sí son necesarios en mi variable dataset3.

Una vez hecha esa transformación, grafiqué las primeras 10 líneas para que se pudiese ver mejor la correspondencia de cada identificador.



A FUTURO…

Por lo que se puede ver, hay muchas variables interesantes que entran en juego dentro de mi estudio y todavía queda mucho por recorrer. Mis aspiraciones son continuar con el análisis de todas las variables que interactúan en mi proyecto hasta acabar con un conjunto de datos que recoja todas las entidades esenciales de mi proyecto y así elaborar una plataforma confiable y estable. Lo adecuado sería que acabase con un conjunto de datos que actualice sus valores dependiendo de los cambios debido a que este sector, y más aún con la reciente crisis del Coronavirus, es considerablemente cambiante.