

Proyecto Integrador 3: EA1

Presentado por:

Cristian Camilo Madrigal Alvarez

Oscar Andres Mantilla Franco

Andres Camilo Graciano Higuita

EA1: Selección Dataset y pregunta de investigación

Profesora

Ing. Sharon Karin Camacho Guzman

Ingeniería de Software y Datos

Institución Universitaria Digital De Antioquia

Medellín

2025

INTRODUCCIÓN

El presente trabajo tiene como objetivo analizar el comportamiento y la distribución de los establecimientos de industria y comercio activos en la ciudad de Medellín, a partir del dataset oficial publicado en el portal de Datos Abiertos Colombia (ID: guhw-8tnz). Este conjunto de datos, suministrado por la Alcaldía de Medellín, contiene información georreferenciada y económica de más de 184.000 registros, representando el tejido empresarial formal de la ciudad.

El análisis busca identificar patrones de concentración y características de los establecimientos según su ubicación geográfica (comuna y barrio), su grupo de actividad económica.. A través de herramientas de análisis de datos en Python —como pandas, geopandas y pandas profiling— se desarrolla una exploración descriptiva que permite comprender la estructura espacial y económica de la ciudad.

Asimismo, el estudio pretende detectar posibles problemas en la calidad de los datos (como valores nulos o redundancias) y definir las tareas de limpieza necesarias para garantizar la fiabilidad del análisis. Este enfoque no solo aporta información sobre la distribución empresarial de Medellín, sino que también permite fortalecer la capacidad analítica aplicada al uso de datos públicos abiertos.

Link al tablero Trello:

<https://trello.com/invite/b/690773687979f9f074faf163/ATTI4363c45a8a5d17a4e16f2a9311a277f6207E14E7/proyecto-integrador-3-grupo-2-2025-2>

Link al repositorio Github:

<https://github.com/CamiloMadrigal12/ProyectoIntegrado3>

DESARROLLO DEL TRABAJO

Link del dataset:

<https://www.datos.gov.co/api/views/guhw-8tnz/rows.csv?accessType=DOWNLOAD>

Formato: CSV

Columnas: 9

Filas: 184.082

Fuente: Portal oficial de Datos Abiertos del Gobierno de Colombia

Granularidad: La granularidad del dataset es por establecimiento activo: cada fila representa un establecimiento comercial individual en Medellín

Objetivo general:

Identificar patrones de concentración y características de los establecimientos activos en Medellín por grupo de actividad, comuna y barrio, y analizar cómo la antigüedad del establecimiento (fecha de inicio) se relaciona con esas concentraciones.

Pregunta de Investigación:

¿Qué factores (grupo de actividad, comuna, barrio y antigüedad del establecimiento) explican una mayor concentración de establecimientos activos en Medellín?

Hipótesis

1. Las comunas centrales concentran una mayor proporción de servicios y comercio que las comunas periféricas.
2. La antigüedad de los establecimientos es menor (más recientes) en zonas con desarrollo urbano reciente.
3. Ciertos grupos de actividad (p. ej., alimentos/bebidas) tienden a agruparse en comunas con alta afluencia peatonal.

Métricas de Éxito

- Densidad de establecimientos por comuna (establecimientos / km² o por 10.000 hab., si se cruza con área o población).
- Participación porcentual de cada grupo_actividad sobre el total de la comuna.
- Antigüedad promedio por comuna y por grupo_actividad (a partir de fecha_inicio_act).
- Índice de concentración (p. ej., Herfindahl-Hirschman por grupo en cada comuna/barrio).
- Top-N barrios/comunas con mayor número absoluto de establecimientos.

* ¿Estos datos me sirven para lograr el objetivo?

Sí. Los datos me sirven para lograr el objetivo porque con se puede identificar la ubicación geográfica, antigüedad promedio, patrones de concentración de los establecimientos.

En la siguiente tabla se puede encontrar el diccionario de datos con las columnas y su respectivo tipo de dato.

Diccionario de datos

Columna	Descripción	Tipo de Dato	Ejemplo
BARRIO	Nombre del barrio donde está ubicado el establecimiento.	Categórico (texto)	“Laureles”
COMUNA	Comuna de Medellín a la que pertenece el establecimiento.	Categórico (texto)	“Laureles - Estadio”
COORD_X	Coordenada geográfica en el eje X (longitud o proyección).	Numérico (decimal)	836900.45
COORD_Y	Coordenada geográfica en el eje Y (latitud o proyección).	Numérico (decimal)	1179000.12
FECHA_INICIO_ACT	Fecha en que el establecimiento inició actividades. Permite estimar la antigüedad.	Fecha (YYYY-MM-DD)	“2012-07-15”
GRUPO_ACTIVIDAD	Clasificación general del tipo de actividad económica del establecimiento.	Categórico (texto)	“Comercio”, “Servicios”, “Industria”
HOMOLOGACION_CIIU	Código o grupo equivalente de la Clasificación Industrial Internacional Uniforme (CIIU).	Categórico (texto)	“G – Comercio al por mayor y al por menor”

OBJECTID	Identificador interno único del registro.	Numérico (entero)	15342
point	Representación geográfica (latitud y longitud en formato geoespacial).	Objeto / Texto	“POINT (-75.574 6.244)”

TIPOS DE DATOS DEL DATASET

Tipo de Dato	Columnas Asociadas	Uso Analítico
Texto (Categórico)	BARRIO, COMUNA, GRUPO_ACTIVIDAD, HOMOLOGACION_CIIU	Para agrupar, clasificar y segmentar establecimientos.
Numérico	COORD_X, COORD_Y, OBJECTID	Para análisis espaciales.
Fecha	FECHA_INICIO_ACT	Permite calcular la antigüedad del establecimiento y analizar tendencias temporales.
Geoespacial	point	Permite mapas y análisis de concentración geográfica (densidad espacial).

Overview	Alerts 8	Reproduction
Dataset statistics		Variable types
Number of variables		Text 3
Number of observations		Categorical 2
Missing cells		Numeric 5
Missing cells (%)		DateTime 1
Duplicate rows		
Duplicate rows (%)		
Total size in memory		14.2 MiB
Average record size in memory		81.0 B

Según los resultados preliminares del ydata_profiling no hay datos duplicados aparentemente.

El 2,1% de las celdas tienen datos faltantes, estos son 42.668 celdas del total de observaciones que son 184.082.

HOMOLOGACION_CIIU

Text

Missing

Distinct	844
Distinct (%)	0.6%
Missing	42628
Missing (%)	23.2%
Memory size	1.4 MiB

El campo con mayor cantidad de valores faltantes es HOMOLOGACION_CIIU con un 23,2%.

GRUPO_ACTIVIDAD

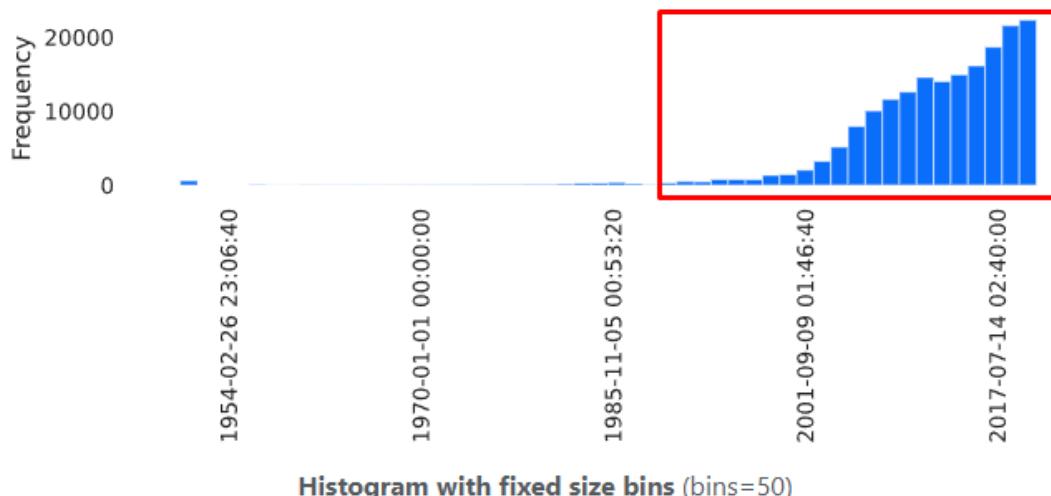
Categorical



More details

La categoría servicios es la que tiene mayor representación en los datos.

Las coordenadas que posee el dataset sí corresponden al límite geográfico de Medellín, con lo cual posibilita construir mapas o realizar análisis de densidad confiables.



Se tiene una cobertura temporal amplia con cantidad de datos que se amplía a partir de este siglo.

COMUNA

Categorical

High correlation

Distinct	49	La Candelaria	46734
Distinct (%)	< 0.1%	El Poblado	28070
Missing	20	Laureles Esta...	20449
Missing (%)	< 0.1%	Belén	16515
Memory size	1.4 MiB	Guayabal	9571
		Other values ...	62723

Medellín presenta un modelo de concentración comercial policéntrico, con fuerte nodo central y subcentros en Laureles y Belén.

Alerts

COMUNA is highly overall correlated with COORD_X and 3 other fields	High correlation
COORD_X is highly overall correlated with COMUNA and 1 other fields	High correlation
COORD_Y is highly overall correlated with COMUNA and 1 other fields	High correlation
LAT is highly overall correlated with COMUNA and 1 other fields	High correlation
LON is highly overall correlated with COMUNA and 1 other fields	High correlation
HOMOLOGACION_CIIU has 42628 (23.2%) missing values	Missing
OBJECTID is uniformly distributed	Uniform
OBJECTID has unique values	Unique

Entre las alertas tenemos lo siguiente: Comuna y las variables geográficas tienen alta correlación pero esto es un comportamiento normal y esperado.

Sobre la alerta de HOMOLOGACION_CIIU, hacen falta casi una cuarta parte de los datos, esto reduce la capacidad de análisis que podría tener esa variable. Para estos casos, se les podría asignar un valor = “No Clasificado”.

El OBJECTID es una clave primaria, es normal que sea única y esté distribuida de esa forma.

LISTA DE TAREAS PARA LIMPIEZA DE DATOS

1. Tratamiento a los valores faltantes:
 - a. En el campo HOMOLOGACION_CIIU crear una categoría llamada “NO CLASIFICADO” para esos valores faltantes.
 - b. En el campo FECHA_INICIO_ACT reemplazar los valores faltantes por la mediana por comuna o grupo de actividad.
 - c. Reemplazar los valores nulos que hay en BARRIO y COMUNA por “DESCONOCIDO”.
2. Estandarización de texto: convertir a formato título y eliminar espacios extra y caracteres especiales en los campos BARRIO, COMUNA y GRUPO_ACTIVIDAD.
3. Convertir los tipos de datos de cada campo acorde a su característica como FECHA_INICIO_ACT a datetime.

CONCLUSIONES

- El dataset presenta una estructura robusta y coherente, con información suficiente para realizar análisis estadísticos y espaciales. No se detectaron duplicados, y las variables principales mantienen una distribución lógica y representativa del territorio medellinense.
- La variable HOMOLOGACION_CIIU presenta un 23% de valores faltantes, lo que constituye el principal reto de calidad de datos. Sin embargo, este problema no compromete la integridad general del dataset.
- El proceso de limpieza permitirá mejorar la calidad del dataset, estandarizar tipos de datos y preparar la información para análisis más avanzados, como visualizaciones geográficas, modelos de densidad o análisis de concentración económica.