Proyecto Integrador 3: EA2

Presentado por:

Cristian Camilo Madrigal Alvarez
Oscar Andres Mantilla Franco
Andres Camilo Graciano Higuita

EA2: Limpieza, transformación de datos y validación

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.

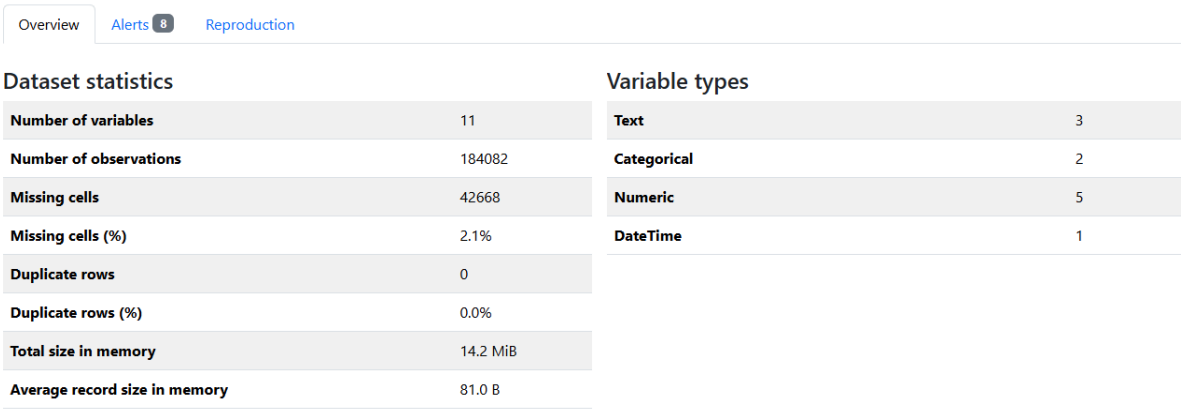
En la siguiente tabla se puede encontrar el diccionario de datos con las columnas y su respectivo tipo de dato.

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

Number of variables	11
Number of observations	184082
Missing cells	42668
Missing cells (%)	2.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	14.2 MiB
Average record size in memory	81.0 B

Variable types

Text	3
Categorical	2
Numeric	5
DateTime	1

Según los resultados preliminares del ydata_profilling 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.

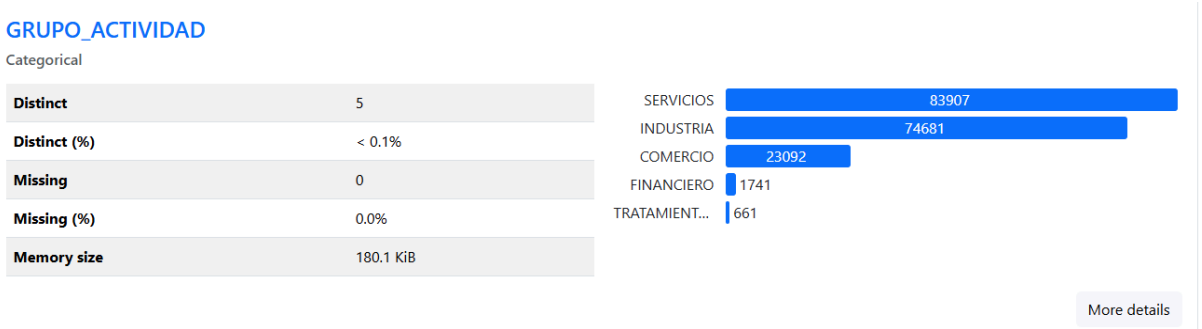
HOMOLOGACION_CIIU

Text

Missing

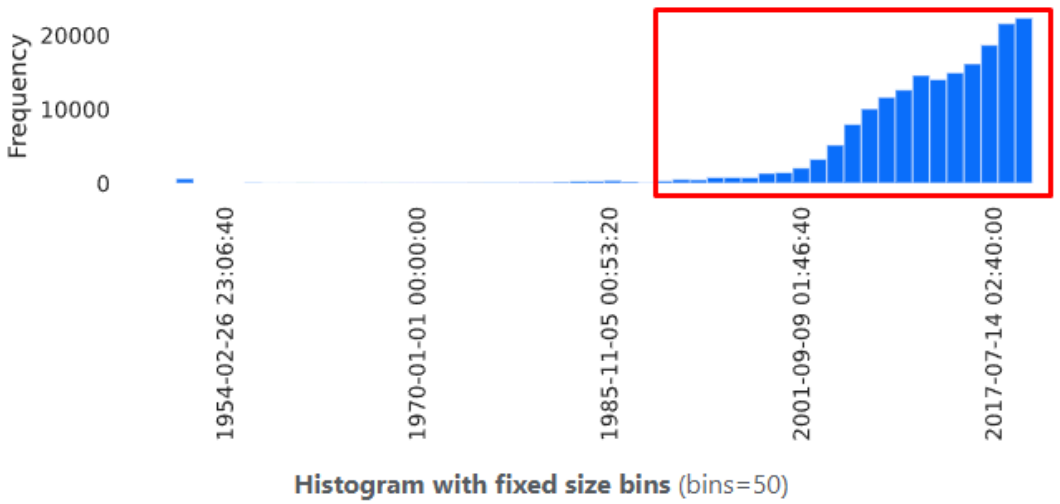
Distinct	844
Distinct (%)	0.6%
Missing	42628
Missing (%)	23.2%
Memory size	1.4 MiB

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

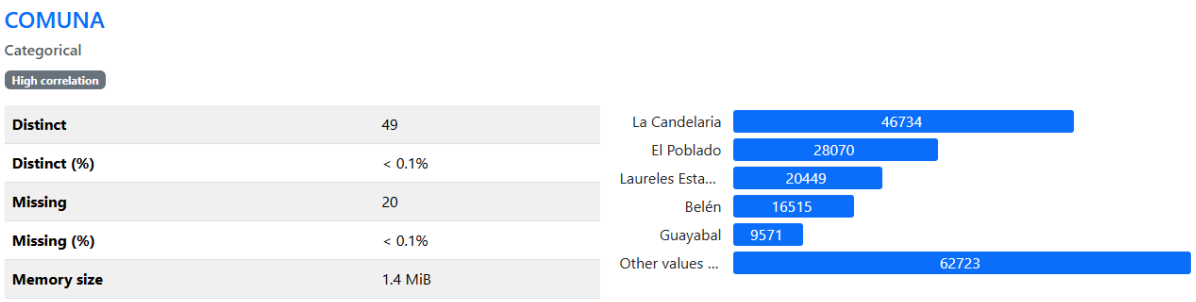


La categoría servicios es la que tiene mayor representación en los datos.

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.



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

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”.

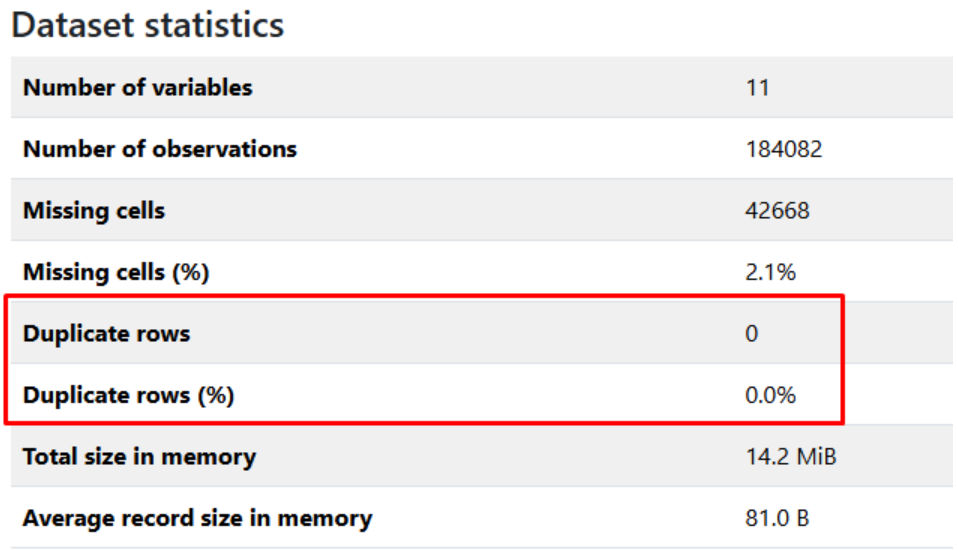
El OBJECTID es una clave primaria, es normal que sea única y esté distribuida de esa forma.

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.

DESCRIPCIÓN DE NECESIDADES DE LIMPIEZA

1. Eliminación de duplicados: Con el ydata_profiling se muestran 0 filas duplicadas sobre 184.082 registros como se muestra en la siguiente imagen. Por lo tanto, no se requieren acciones de limpieza para este aspecto.



Dataset statistics

Number of variables	11
Number of observations	184082
Missing cells	42668
Missing cells (%)	2.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	14.2 MiB
Average record size in memory	81.0 B

2. Valores Nulos: Existen valores nulos en algunas columnas como:

    a. BARRIO: 20 valores nulos (0,011%)
    b. COMUNA: 20 valores nulos (0,011%)
    c. HOMOLOGACION_CIIU: 42.628 nulos (23,16%)
    d. El resto de columnas no tiene nulos (COORD_X, COORD_Y, FECHA_INICIO_ACT, GRUPO_ACTIVIDAD, OBJECTID, point)

    Se crea una categoría llamada “No Clasificado” para los valores nulos de la columna HOMOLOGACION_CIUU para no perder filas. Así mismo, se crea una categoría llamada “Desconocido” en las columnas BARRIO Y COMUNA para poder seguir contando esos establecimiento en análisis globales sin perderlos.

3. Inconsistencia en valores: Existen las siguientes inconsistencias en algunos valores del dataset:

    a. Columna COMUNA:
        - Hay la misma comuna escrita de varias formas, por ejemplo:
            - "Belén", "BELEN", "BELÉN"
            - "Buenos Aires", "BUENOS AIRES"
            - "Castilla", "CASTILLA", etc.
        - Hay códigos o valores poco claros: "0", "AE", "AU", "SN".
        - Aparecen municipios que no son comunas de Medellín: "Sabaneta", "Copacabana", "La Madera".

        Para solucionar este punto se debe:

        - Normalizar mayúsculas/minúsculas y espacios.
        - Unificar variantes de la misma comuna mediante replace() (ej. todas las variantes de Belén → "Belén").
        - Los registros: "0", "AE", "AU", "SN" y municipios externos se etiquetan como "OTRO/EXTERNO".

    b. Columna BARRIO:

        Hay barrios normales (“Laureles”, “Boston”, etc), pero también:

        - Códigos numéricos tipo "1301", "1020", "0", "0714", etc.
        - Textos como "Suburbano Pedregal alto", "Suburbano La Aldea".

        Para este caso lo que se puede hacer es limpiar espacios y formato y detectar valores completamente numéricos (\d+) y tratarlos como categoría aparte, renombrando como: "BARRIO SIN NOMBRE".

4. Tipos de datos:

FECHA_INICIO_ACT → convertir a datetime con pd.to_datetime() para poder:

- Calcular antigüedad.
- Agrupar por año/mes.

BARRIO, COMUNA, GRUPO_ACTIVIDAD, HOMOLOGACION_CIIU → convertir:

- a string para limpieza de texto y después a category para optimizar memoria y mejorar análisis categórico.

5. Valores atípicos:

Los rangos de COORD_X y COORD_Y son compactos y coherentes con el área de Medellín.

Fechas (FECHA_INICIO_ACT): Tiene un Rango de fechas: 1950–2020.
Muy pocos registros en los años 50–70 y una gran concentración a partir de 2000.
Se deberán onsiderar estos pocos registros muy antiguos como casos raros pero válidos (negocios muy antiguos), y no eliminarlos porque pueden ser interesantes para análisis de antigüedad.

6. Nivel de granularidad:

Cada fila del dataset representa un establecimiento activo individual, esto representa un alto nivel de granularidad, lo cual es bueno para poder agregar por COMUNA y GRUPO_ACTIVIDAD o por fecha de inicio.

VALIDACIÓN DEL CONJUNTO DE DATOS

1. Completitud de los datos

En una primera revisión se identificaron valores nulos principalmente en las columnas HOMOLOGACION_CIIU, BARRIO y COMUNA. Después de la limpieza:

- BARRIO y COMUNA se imputaron con la categoría "DESCONOCIDO", evitando la pérdida de registros que son relevantes para los análisis de concentración territorial.

- HOMOLOGACION_CIIU se imputó con la categoría "NO CLASIFICADO", lo que permite mantener los establecimientos en los análisis generales, dejando explícito que no se dispone del código CIIU en esos casos.

- Las columnas numéricas y de fecha (OBJECTID, COORD_X, COORD_Y, FECHA_INICIO_ACT) no presentan valores nulos tras la conversión de tipos y las validaciones realizadas.

- No se encontraron filas duplicadas, por lo que no fue necesario eliminar registros por este motivo.

Además, las operaciones de agregación realizadas con groupby (por comuna, grupo de actividad y año de inicio) se hicieron a partir del dataset ya limpio, sin aplicar filtros adicionales que eliminaran filas. De este modo, se preserva prácticamente el total de establecimientos originales y se garantiza que los resultados agregados no estén sesgados por pérdidas de información durante el procesamiento.

En consecuencia, se considera que la completitud de los datos es adecuada, y no se identifican vacíos en columnas críticas que impidan el análisis.

2. Relevancia de las variables

Las variables que quedaron disponibles tras la limpieza son coherentes con los objetivos del proyecto:

- Ubicación: COMUNA y BARRIO permiten analizar la distribución espacial de los establecimientos a diferentes niveles de detalle.

- Actividad económica: GRUPO_ACTIVIDAD (normalizada y tipificada como categórica) y HOMOLOGACION_CIIU permiten distinguir tipos de actividad y hacer comparaciones entre sectores.

- Dimensión temporal: FECHA_INICIO_ACT, convertida a formato de fecha y complementada con la variable derivada ANIO_INICIO, permite analizar la antigüedad de los establecimientos; a partir de ella se construyó la métrica de antigüedad promedio por comuna y grupo de actividad.

- Identificación y soporte espacial: OBJECTID se mantiene como identificador único y COORD_X / COORD_Y posibilitan análisis espaciales más detallados o la representación en mapas.

Estas variables son suficientes para:

- Estimar la concentración de establecimientos por comuna, barrio y grupo de actividad.

- Analizar diferencias en la antigüedad promedio entre zonas de la ciudad.

- Explorar posibles patrones entre tipo de actividad y ubicación geográfica.

Por tanto, la relevancia de las variables se considera alta, y el dataset permite responder directamente a las preguntas de negocio definidas.

3. Granularidad adecuada

Cada fila del dataset representa un establecimiento individual con su localización, grupo de actividad y fecha de inicio. Este nivel de granularidad es alto, lo que ofrece ventajas como: en primer lugar, permite construir indicadores agregados a distintos niveles (por comuna, barrio, grupo de actividad o año) sin perder detalle. En segundo lugar, facilita el cálculo de métricas como el número de establecimientos por zona, la distribución sectorial y la antigüedad promedio.

A partir de este nivel de detalle, se generaron vistas agregadas mediante groupby, por ejemplo: número de establecimientos y antigüedad promedio por comuna y grupo de actividad y también, número de establecimientos por año de inicio y grupo de actividad.

Estas agregaciones muestran que el nivel de granularidad original es suficiente para producir insights significativos, pero al mismo tiempo permite resumir la información cuando se requiere una visión más global. En consecuencia, la granularidad del dataset se considera adecuada para los objetivos del análisis.

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.

- Tras evaluar la completitud, relevancia de las variables y granularidad, se concluye que el dataset limpio es apto para abordar los objetivos del proyecto. No se identifican carencias que obliguen a repetir el proceso de limpieza o a descartar una parte significativa de la información.