Proyecto Integrador 3: EA3

Presentado por:

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

EA3: Dashboard

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.

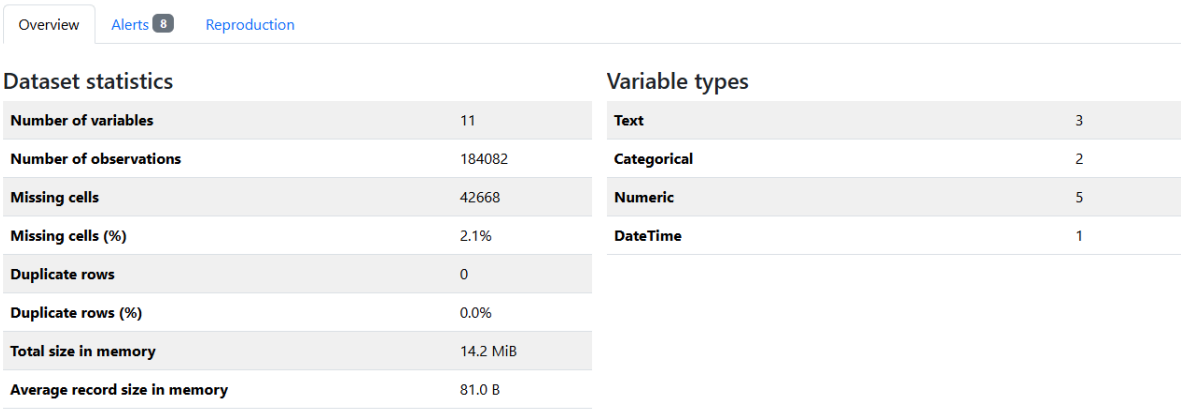
En la siguiente tabla se puede encontrar el diccionario de datos con las columnas y su respectivo tipo de dato.

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.

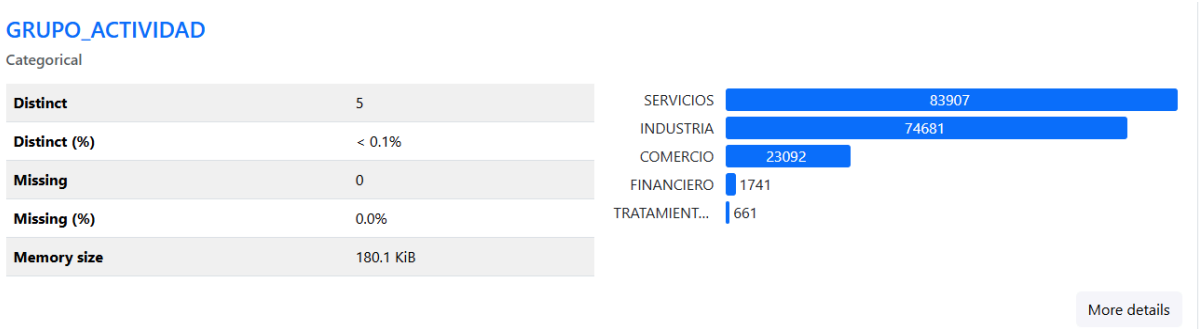
HOMOLOGACION_CIIU

Text

Missing

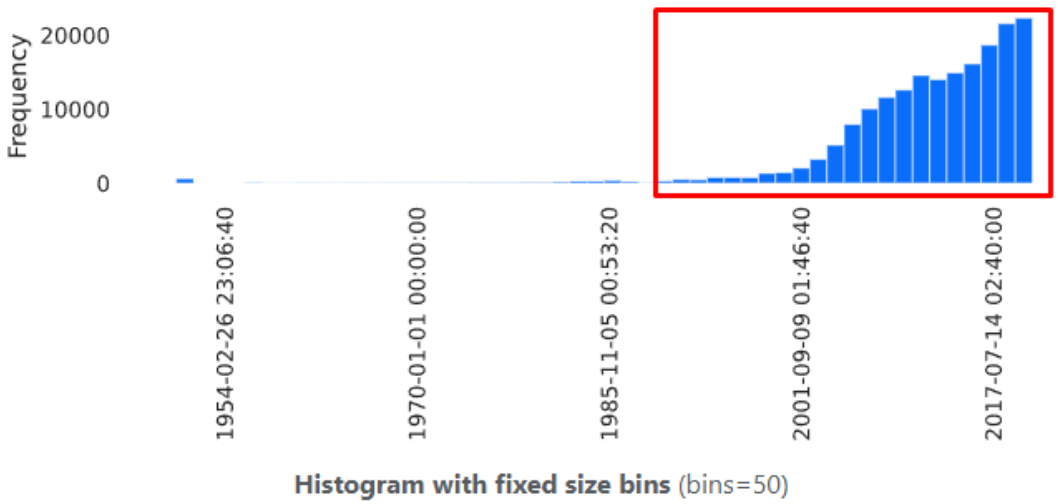
Distinct	844
Distinct (%)	0.6%
Missing	42628
Missing (%)	23.2%
Memory size	1.4 MiB

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

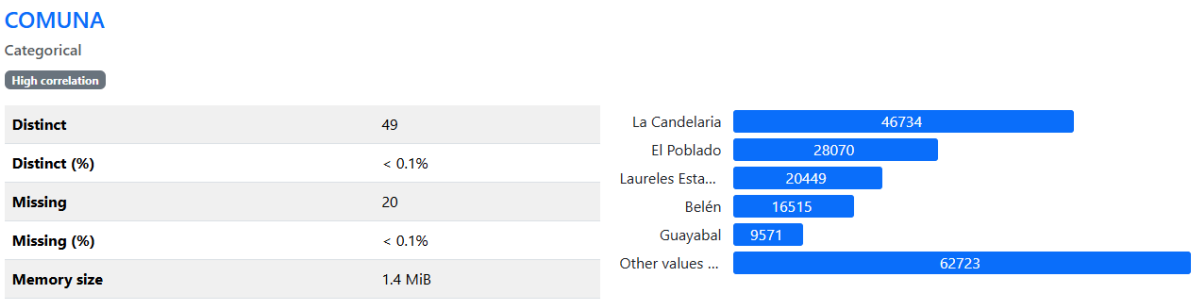


La categoría servicios es la que tiene mayor representación en los datos.

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”.

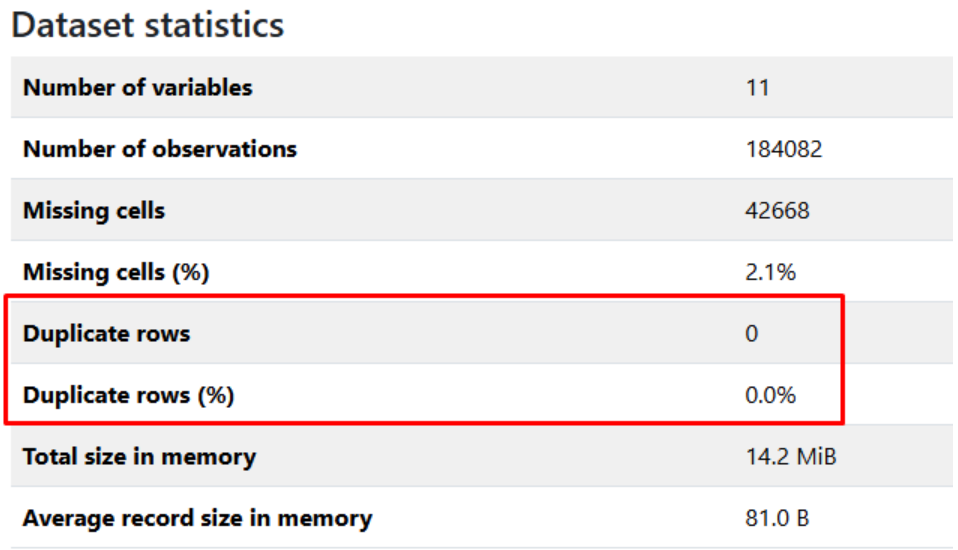
El OBJECTID es una clave primaria, es normal que sea única y esté distribuida de esa forma.

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.

Revisión de Hipótesis:

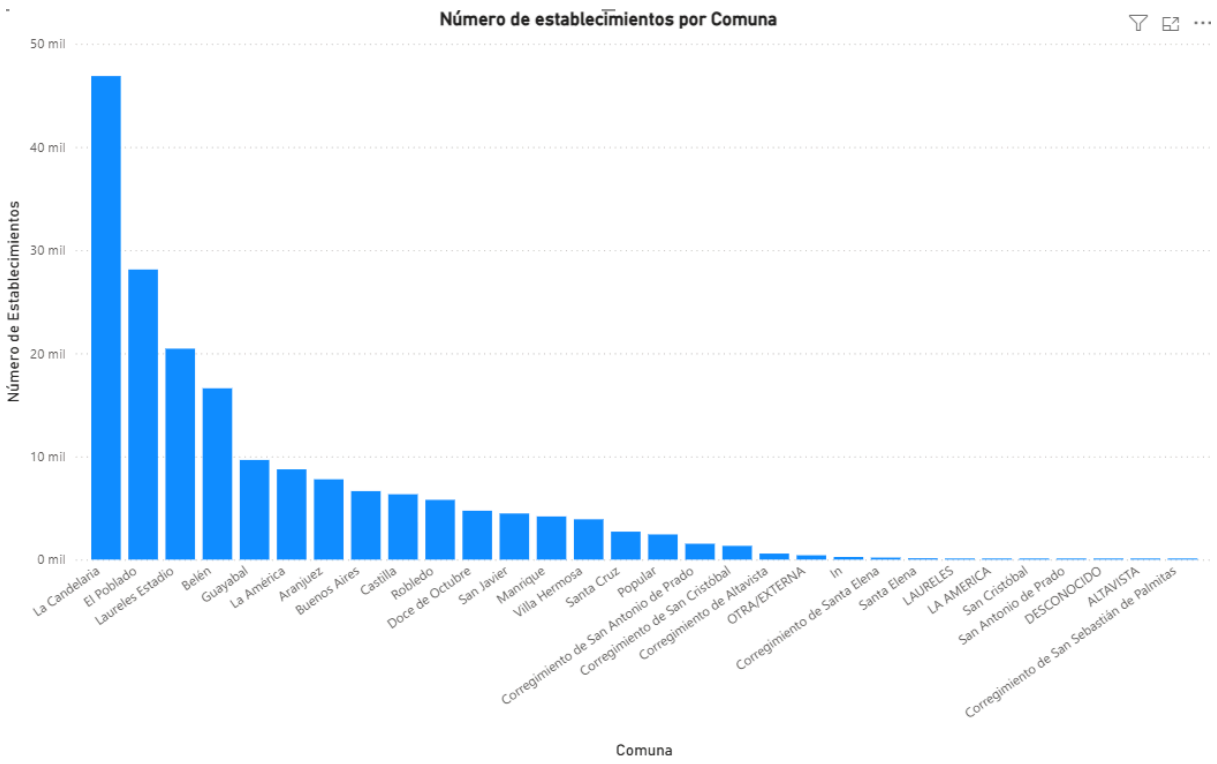
En este proyecto trabajamos con el dataset de establecimientos de industria y comercio activos en Medellín publicado en el portal de Datos Abiertos; El objetivo general que planteamos fue:

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.

A partir de este objetivo, formulamos la siguiente pregunta de investigación:
¿Qué factores (grupo de actividad, comuna, barrio y antigüedad del establecimiento) explican una mayor concentración de establecimientos activos en Medellín?

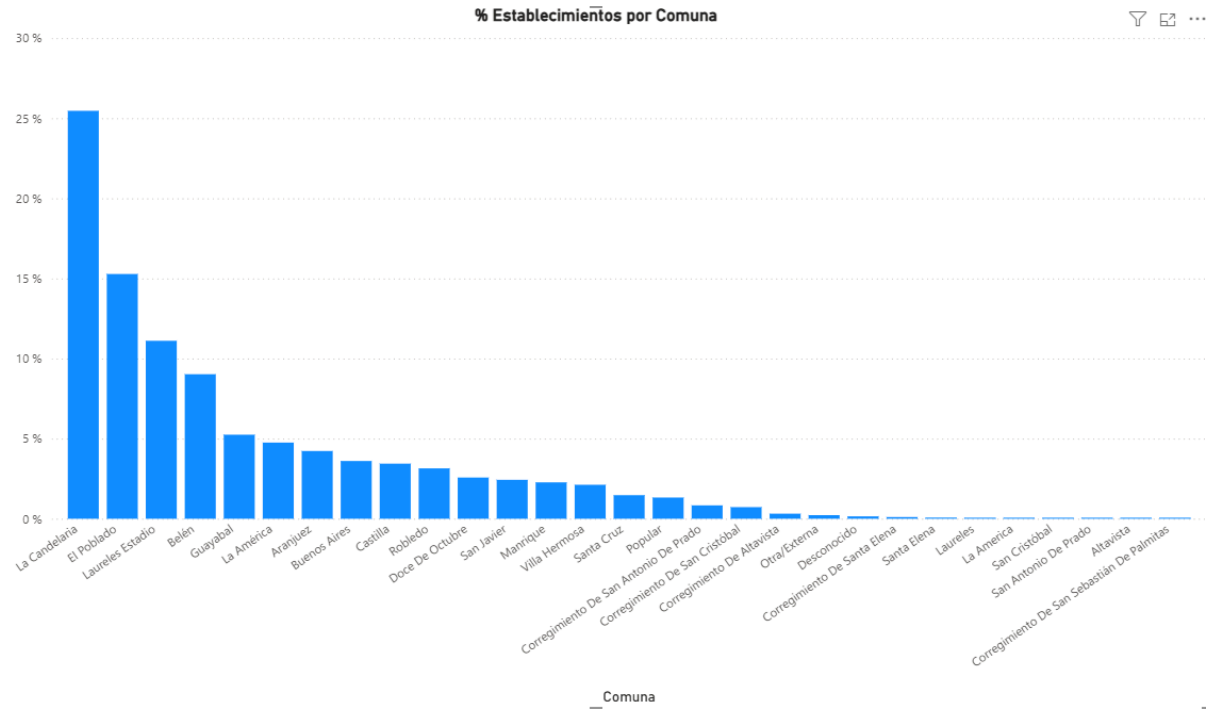
Las hipótesis que definimos buscan justamente dar una posible respuesta a esta pregunta, combinando la dimensión territorial (comunas y barrios), la sectorial (grupo de actividad) y la temporal (antigüedad de los establecimientos).

Hipótesis 1: Las comunas centrales concentran una mayor proporción de servicios y comercio que las comunas periféricas.



Las comunas centrales concentran una mayor proporción de servicios y comercio que las comunas periféricas.

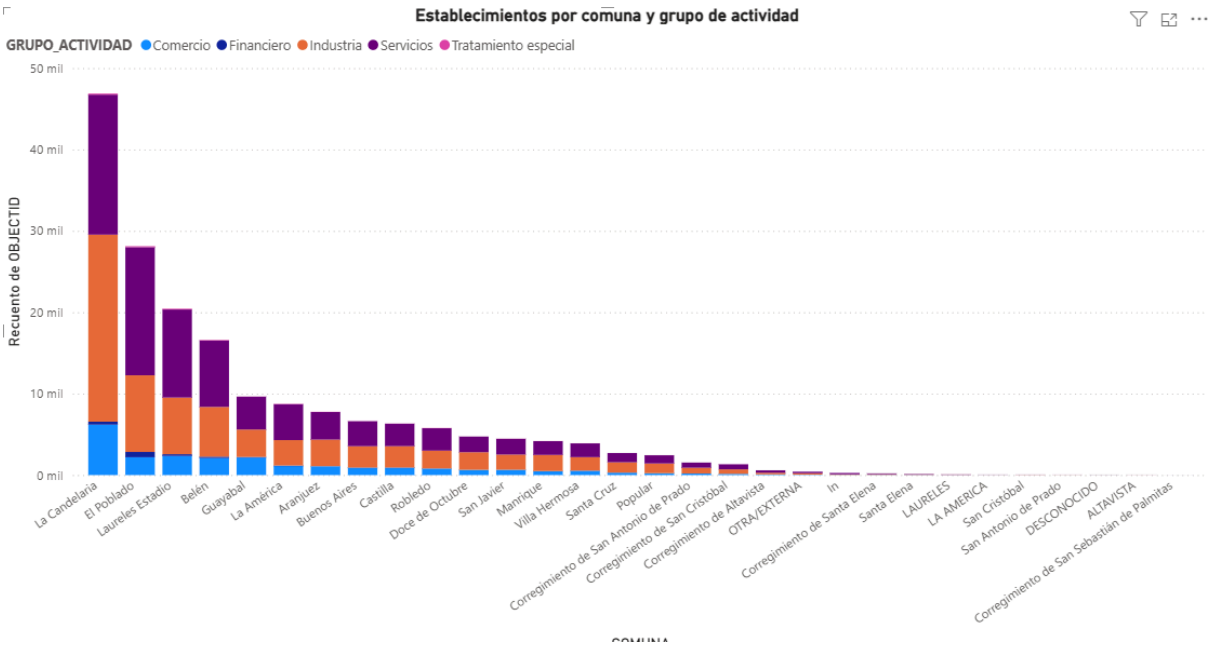
Esta hipótesis parte de la idea de que el desarrollo económico de la ciudad no es homogéneo. Se supone que las comunas más centrales (como el centro tradicional y los subcentros consolidados) atraen más actividades de servicios y comercio, mientras que las zonas periféricas tienen una presencia menor de estos establecimientos o una estructura económica distinta.



La relevancia de esta hipótesis frente al objetivo del proyecto es clara:

- Nos permite medir la concentración territorial de la actividad económica, relacionando el número de establecimientos con la comuna donde se ubican.
- Conecta directamente con el objetivo de identificar patrones de concentración por comuna y grupo de actividad, ya que compara el peso de servicios y comercio entre comunas centrales y periféricas.
- También aporta elementos para interpretar posibles desigualdades espaciales en el acceso a servicios y actividades comerciales dentro de la ciudad.

En la fase de visualización, esta hipótesis se podrá contrastar utilizando gráficos que muestren la distribución de los establecimientos de servicios y comercio por comuna, así como comparaciones entre comunas centrales y periféricas.



Variables seleccionadas:

1. COMUNA (categórica)
    - Esta variable identifica la comuna a la que pertenece cada establecimiento.
    - Es la variable central para esta hipótesis, porque permite distinguir entre comunas centrales y comunas periféricas.
    - A partir de ella se puede:
        - Contar cuántos establecimientos hay en cada comuna.
        - Comparar los totales y porcentajes entre diferentes comunas.
        - Construir categorías como “comunas centrales” vs. “comunas periféricas” para analizar si realmente hay una mayor concentración en las primeras.
2. GRUPO_ACTIVIDAD (categórica)
    - Clasifica cada establecimiento según el tipo de actividad económica: Comercio, Industria, Servicios, Financiero, Tratamiento especial, etc.
    - En el contexto de esta hipótesis es clave porque permite diferenciar específicamente los establecimientos de Servicios y Comercio frente a los demás grupos.
    - Gracias a esta variable se puede calcular, para cada comuna:
        - El número de establecimientos de servicios.
        - El número de establecimientos de comercio.
        - El porcentaje que representan estos dos grupos sobre el total de establecimientos de la comuna.
    - De esta forma, se puede verificar si las comunas centrales tienen una proporción más alta de servicios y comercio que las comunas periféricas.
3. BARRIO (categórica, de apoyo)
    - Aunque la hipótesis se formula a nivel de comuna, la variable barrio puede utilizarse como nivel de detalle adicional.
    - Permite observar si dentro de una misma comuna hay barrios con mayor concentración de servicios/comercio, lo que enriquece la interpretación de la

hipótesis.

- No es estrictamente necesaria para contrastar la hipótesis, pero sí ayuda a afin ar el análisis espacial.

4. Conteo de establecimientos (variable derivada)

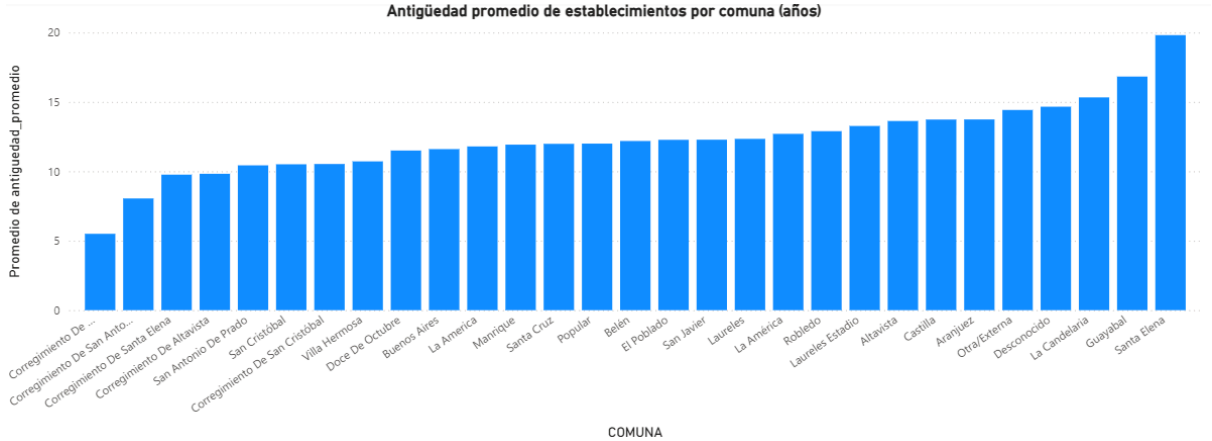
    - No viene como columna explícita, pero se obtiene al contar el número de filas por comuna y grupo de actividad.

    - Es la medida que permite cuantificar la concentración: cuántos establecimientos hay en total y cuántos son de servicios/comercio en cada comuna.

    - A partir de esta variable derivada es posible construir gráficos de barras por comuna y comparar visualmente las concentraciones.

En conjunto, la combinación de COMUNA + GRUPO_ACTIVIDAD (más el conteo de registros) permite validar si, efectivamente, los servicios y el comercio se concentran en las comunas centrales en mayor proporción que en el resto de la ciudad.

Hipótesis 2: La antigüedad de los establecimientos es menor (más recientes) en zonas con desarrollo urbano reciente.



La antigüedad de los establecimientos es menor (son más recientes) en zonas con desarrollo urbano reciente.

Aquí la idea es relacionar la antigüedad de los establecimientos (a partir de la variable FECHA_INICIO_ACT) con el proceso de expansión y consolidación urbana. Se asume que en áreas de desarrollo más reciente tienden a localizarse negocios nuevos, mientras que en zonas tradicionales de la ciudad hay establecimientos con más años de funcionamiento.

Esta hipótesis es importante para el análisis porque:

- Introduce explícitamente la dimensión temporal, que hace parte del objetivo general cuando se habla de la “antigüedad del establecimiento”.
- Permite estudiar si la expansión urbana está acompañada por la creación de nuevos establecimientos, o si, por el contrario, los negocios más antiguos siguen concentrados en zonas centrales.
- Ayuda a entender la dinámica de renovación del tejido empresarial: qué partes de la ciudad muestran mayor “vitalidad” en términos de creación de establecimientos.

En la etapa de visualización, esta hipótesis se podrá revisar mediante gráficos que relacionen antigüedad promedio por comuna o barrio y mapas o barras donde se comparen zonas consolidadas con zonas de desarrollo más reciente.

Variables seleccionadas:

1. FECHA_INICIO_ACT (fecha)
    - Esta variable indica la fecha en la que el establecimiento registró el inicio de su actividad.
    - Es la base para medir la antigüedad de cada establecimiento: a partir de ella se puede calcular cuántos años lleva funcionando.
    - Permite:
        - Comparar establecimientos antiguos vs. recientes.
        - Calcular la antigüedad promedio por comuna o barrio.
        - Analizar si en determinadas zonas predominan establecimientos de creación reciente.
2. ANIO_INICIO (numérica: año)
    - Es una variable derivada a partir de `FECHA_INICIO_ACT` que contiene únicamente el año de inicio de la actividad.
    - Facilita el análisis, ya que:

        - Permite agrupar los establecimientos por año.
        - Hace más sencillo construir gráficos de series de tiempo y tablas de frecuencias por año.
    - Para validar la hipótesis es útil porque se pueden identificar periodos de mayor creación de establecimientos y compararlos entre zonas.

3. COMUNA (categórica)
    - Permite asociar la antigüedad de los establecimientos con el territorio.
    - A través de esta variable se pueden comparar:
        - La antigüedad promedio de los establecimientos en cada comuna.
        - La proporción de establecimientos recientes en comunas consideradas como zonas de desarrollo urbano más reciente frente a comunas históricamente consolidadas.
    - De esta forma, se puede verificar si las comunas asociadas a expansión urbana reciente tienen, efectivamente, establecimientos con menor antigüedad.

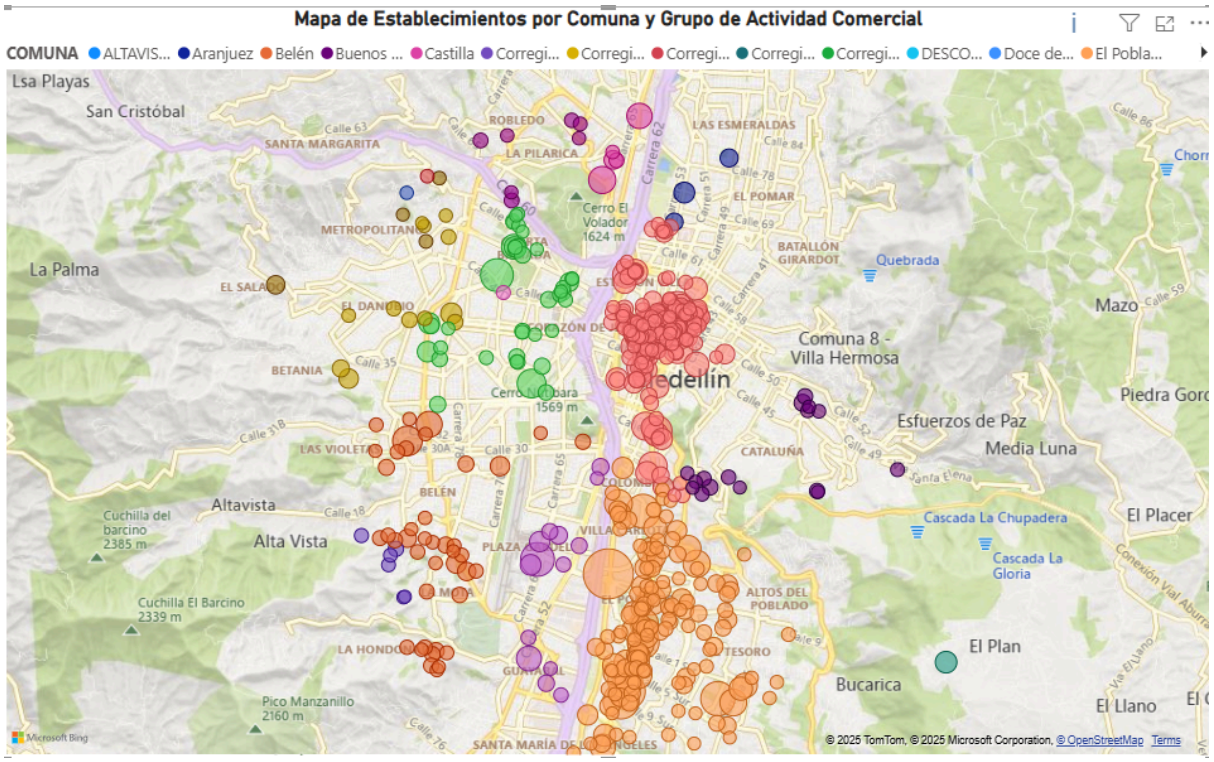
4. BARRIO (categórica, de apoyo)
    - En algunos casos, ciertos barrios dentro de una misma comuna pueden ser reconocidos como sectores de expansión reciente.
    - Utilizar el barrio permite hacer un análisis más detallado, identificando si en esos sectores se concentran establecimientos nuevos.
    - Sirve como complemento cuando se quiere profundizar por debajo del nivel de comuna.

5. Medida de antigüedad (variable derivada)
    - Puede definirse, por ejemplo, como:
      - Antigüedad = Año de análisis – ANIO_INICIO
    - Esta variable derivada permite:
        - Comparar directamente establecimientos antiguos y recientes.
        - Calcular la antigüedad promedio por comuna o barrio.
    - Es la medida que se usa para verificar la afirmación central de la hipótesis: que en zonas de desarrollo urbano reciente los establecimientos son, en

promedio, más nuevos.

En resumen, la combinación de FECHA_INICIO_ACT / ANIO_INICIO con COMUNA (y, si se requiere, BARRIO) permite relacionar el momento de aparición de los establecimientos con las zonas de la ciudad, y así validar si las áreas de expansión urbana reciente concentran negocios más jóvenes.

Hipótesis 3: Ciertos grupos de actividad (p. ej., alimentos/bebidas) tienden a agruparse en comunas con alta afluencia peatonal.



Ciertos grupos de actividad (por ejemplo, alimentos/bebidas) tienden a agruparse en comunas con alta afluencia peatonal.

Esta hipótesis plantea que no todos los tipos de actividad económica se distribuyen de la misma forma en el territorio. Se espera que actividades como alimentos, bebidas y comercio al detal estén más concentradas en zonas con alta circulación de personas (centros comerciales, zonas turísticas, corredores principales, etc.), lo cual se refleja en determinadas comunas y barrios.
Su relevancia frente al objetivo y la pregunta de investigación es la siguiente:

- Aporta una lectura más fina del tipo de actividad económica, y no solo del número total de establecimientos.
- Permite estudiar si ciertos grupos de actividad muestran patrones claros de aglomeración espacial, lo que está directamente relacionado con la idea de “mayor concentración de establecimientos”.

- Conecta con posibles aplicaciones prácticas, como la planificación de oferta comercial, movilidad y ordenamiento del espacio público.

En la fase de visualización, esta hipótesis se podrá analizar con gráficos y mapas que muestren la distribución de grupos específicos de actividad por comuna y barrio, identificando si realmente se concentran en zonas que podrían asociarse a alta afluencia peatonal.

Variables seleccionadas:

1. HOMOLOGACION_CIIU (categórica detallada)
    - Esta variable contiene la clasificación de la actividad económica según la codificación CIIU (o una homologación de la misma).
    - Es fundamental para esta hipótesis porque permite identificar actividades específicas dentro de los grupos generales, por ejemplo:
        - Restaurantes.
        - Bares.
        - Venta de alimentos preparados.
        - Comercio de bebidas, etc.
    - A partir de estos códigos se puede definir un subconjunto de actividades que corresponden a alimentos y bebidas u otros servicios intensivos en afluencia peatonal.
    - Esto hace posible analizar si ese tipo de establecimientos se concentra más en determinadas comunas o barrios.
2. GRUPO_ACTIVIDAD (categórica, más general)
    - Complementa a `HOMOLOGACION_CIIU` al agrupar las actividades en grandes categorías: Servicios, Comercio, Industria, etc.
    - Permite ver, a un nivel más agregado, si ciertos grupos (por ejemplo, Servicios) se concentran en zonas con alta afluencia.
    - Aunque es menos detallada que `HOMOLOGACION_CIIU`, ayuda a construir una primera visión de qué tipo de actividades predominan en cada comuna.
3. COMUNA (categórica)

- Es la variable que permite asociar la presencia de determinados tipos de actividad con zonas específicas de la ciudad.

- Para esta hipótesis se asume que algunas comunas (como las centrales o con fuerte presencia comercial y turística) pueden considerarse de mayor afluencia peatonal.

- Comparando el número y porcentaje de establecimientos de alimentos/bebidas (identificados vía CIIU) entre comunas, se puede verificar si realmente se agrupan más en aquellas zonas de alta circulación de personas.

4. BARRIO (categórica, de apoyo)

    - Dentro de una comuna, ciertos barrios pueden funcionar como corredores comerciales, zonas de vida nocturna, centros de servicios, etc.

    - Analizar la distribución de los códigos de alimentos/bebidas por barrio permite afinar aún más la identificación de focos de concentración.

    - Esto refuerza la validación de la hipótesis, mostrando si la aglomeración se da en puntos específicos del territorio.

5. point / COORD_X – COORD_Y / LON – LAT (variables espaciales)

    - Estas variables contienen la ubicación geográfica del establecimiento (ya sea en coordenadas planas o en formato de punto georreferenciado).

    - Son relevantes porque permiten:

        - Representar los establecimientos en mapas.

        - Visualizar la agrupación espacial de los establecimientos de ciertos tipos (por ejemplo, alimentos/bebidas) en zonas puntuales de la ciudad.

    - Aunque la hipótesis se formula a nivel de comuna, el uso de las coordenadas permite observar si, más allá de la comuna, hay clusters de establecimientos en áreas de alta afluencia peatonal, cercanas a estaciones de transporte, centros comerciales, parques, etc.

En conjunto, las variables HOMOLOGACION_CIIU + GRUPO_ACTIVIDAD + COMUNA (apoyadas por BARRIO y la ubicación geográfica) permiten identificar si los establecimientos de ciertos tipos de actividad efectivamente se organizan en torno a comunas y sectores con alta circulación de personas, como lo plantea la hipótesis.

CONCLUSIONES

El objetivo de este trabajo fue analizar la distribución de los establecimientos de industria y comercio activos en Medellín, teniendo en cuenta la comuna, el barrio, el grupo de actividad económica y la antigüedad de los establecimientos. A partir de este análisis se buscó responder a la pregunta de investigación sobre qué factores explican una mayor concentración de establecimientos en la ciudad.

En términos generales, los resultados muestran que la actividad económica no se distribuye de forma homogénea en el territorio, sino que presenta una clara concentración en algunas comunas centrales, que agrupan un número significativamente mayor de establecimientos frente a las comunas periféricas. Esto sugiere que la localización en zonas de alta accesibilidad y centralidad urbana es un factor clave para entender la concentración de establecimientos.

Por otro lado, al revisar la composición por grupo de actividad, se observa que los servicios y ciertas actividades comerciales tienen un peso importante dentro del total de establecimientos, especialmente en las comunas con mayor afluencia de personas y con dinámicas más consolidadas de comercio y prestación de servicios. Esto está en línea con la idea de que el sector terciario tiende a concentrarse en áreas estratégicas de la ciudad, donde existe una mayor demanda y flujo de potenciales clientes.

En cuanto a la antigüedad de los establecimientos, los datos permiten evidenciar diferencias entre zonas más tradicionales y sectores de desarrollo urbano más reciente. En general, las áreas centrales tienden a concentrar establecimientos con mayor tiempo de funcionamiento, mientras que en zonas más nuevas aparecen con más frecuencia establecimientos de creación reciente. Esto sugiere que la expansión urbana ha venido acompañada por la apertura de nuevos negocios, pero sin desplazar totalmente el peso histórico de las zonas centrales.

En conjunto, el análisis permite concluir que la concentración de establecimientos en Medellín está asociada principalmente a la ubicación (centralidad y accesibilidad), al tipo de actividad económica y al momento de aparición de los negocios en el territorio. Estos factores, analizados de manera conjunta, ofrecen una respuesta coherente a la pregunta de investigación y cumplen con el objetivo planteado, al permitir caracterizar los patrones de concentración y las principales características de los establecimientos de industria y comercio activos en la ciudad.