

Análisis Encuesta SISBEN

Luisa Fernanda Martínez
Sandra Milena Grisales

Municipio de Envigado Antioquia
Programa Talento Tech
Ingeniero Mateo Ramirez

Envigado, Colombia
29 de abril de 2025

Introducción

Las dinámicas de ocupación y uso del territorio son totalmente ajenas a la planeación como estado ideal, la entropía es el estado que impera y al igual que esta, las sociedades humanas se mueven, asientan, acomodan y re-acomodan.

La encuesta Sisbén se origina con las reformas gubernamentales en el año 1993 con la intención de ser un instrumento para focalizar el gasto público social, siendo utilizado para la selección de potenciales beneficiarios para los diferentes programas. El Sisbén es una herramienta para ordenar a la población según sus condiciones de vida y vulnerabilidad y es coordinado por el Departamento Nacional de Planeación.

El análisis de datos se ha posicionado como una herramienta útil para visualizar, identificar y hasta predecir comportamientos con la ventaja de poder procesar grandes volúmenes de información, por lo cual se constituye en un gran aliado para realizar los análisis requeridos a los resultados de las encuestas del Sisbén y lograr así el objetivo de focalizar el gasto público social.

Inicialmente no se tenía un rumbo preestablecido, la delimitación del tema de interés apareció en el camino, la base de datos a la que se tuvo acceso es muy vasta en posibilidades de análisis y finalmente se toma la decisión de procesar la información relacionada con los eventos adversos, más concretamente las inundaciones reportadas en el municipio.

La gestión del riesgo es un proceso esencial para la Administración Pública, ya que permite identificar, evaluar y mitigar los posibles peligros que puedan afectar la vida y la calidad de vida de los habitantes. El análisis de datos para la gestión del riesgo no es solo una moda para intelectuales, sino que se ha convertido en una necesidad estratégica para tener una visión más precisa y objetiva de los riesgos y su relación con el territorio y la condiciones de habitabilidad (segundo elemento elegido para realizar el comparativo), esto facilita la detección temprana de amenazas y oportunidades. Este enfoque basado en datos no solo permite identificar los riesgos potenciales, sino también medir su impacto y probabilidad, lo que facilita la planificación de respuestas adecuadas. Con el avance de las tecnologías y el acceso a grandes cantidades de información en tiempo real, las organizaciones pueden ahora contar con herramientas que optimizan la eficiencia en la gestión del riesgo, mejorando la capacidad de anticiparse a situaciones adversas y tomando decisiones informadas.

Tomando como referente los datos reportados en el Sisbén del Municipio se pretende identificar la relaciones entre los eventos acaecidos y las características estructurales de las viviendas.

Desarrollo del Proyecto

Para realizar el ejercicio se aplican los elementos vistos durante el Bootcamp con el lenguaje de programación Python y las librerías que tiene disponibles.

1. Definición del Problema

Para este estudio se tomó como base la información recolectada mediante la encuesta del Sisben del Municipio de Envigado y por tratarse de una base de datos con tanta información, se delimitó al análisis a conocer la distribución de la población según su capacidad para generar ingresos y sus condiciones de vida, representada en la clasificación de los grupos así:

A1 - A5: Pobreza extrema

B1 - B7: Pobreza moderada

C1 - C18: Vulnerable

D1 - D21: No pobre, no vulnerable

Posteriormente, se realizó un conteo de los eventos adversos y catastróficos ocurridos en el territorio, para finalmente analizar la posible relación entre las condiciones socioeconómicas de las viviendas y la posibilidad de presentar este tipo de eventos. La información que sugieren los datos puede constituir un insumo para las oficinas de Gestión riesgo y Planeación.

2. Recopilación y Preparación de los Datos

Acciones:

- Carga de datos en Colab, a través de Google Drive
- Importar librerías pandas, matplotlib y seaborn

3. Exploración de Datos

Acciones:

- Se valida la cantidad de filas y columnas del DF (Data frame: denominación que se da a las bases de datos para el análisis). Se encuentra que el DF cuenta con 133180 filas y 150 columnas
- Se obtiene el tipo de dato de cada columna. Todos los datos son tipo int64 y usa memoria de 152.4+ MB, por lo cual se realiza el cambio del tipo de datos al DF, con el fin de hacerlo más ligero y fácil de trabajar, todos los datos se cambian a int32 y se baja el uso de memoria a 152.4+ MB
- Se validan las columnas que tiene el DF
- Se revisan las estadísticas descriptivas: en esta acción no se encontró utilidad porque la información fue llevada a parámetros que no permiten la identificación de diferencias numéricas.
- Se usa la herramienta skimpy para visualizar de otra forma las estadísticas descriptivas.
- Se detectan valores nulos: acción sobre la que se obtiene falso en todos los campos.
- Se conoce cuántos valores nulos hay por columnas obteniendo cero en todos los campos.

- Para verificar en toda la extensión de columnas se busca cuántas son mayores que cero y el resultado es el mismo, ningún valor es mayor que cero.
- Se validan los registros duplicados obteniendo como resultado cero.
- Se visualizan los registros repetidos y la columna en que se ubican, mostrando la concentración de datos en el mismo Municipio por ejemplo.
- Se identifican la columnas del DF
- Se identifican los registros dentro de una columna
- Al ser tantas columnas realizamos una verificación de valores nulos buscando cuantas son mayor que 0 y se obtiene como resultado 0
- Validación de registros duplicados
- Contabilización de recursos duplicados
- Visualización de registros repetidos y la columna en la que están y se encuentra que los repetidos son el código del departamento y del municipio y esto tiene lógica.
- Identificación de registros dentro de una columna, se escoge en este caso los nombres de las veredas y luego se aplica también a los barrios.
- Eliminación de columnas. Este permite delimitar la información específica que se analizará mediante gráficos y para revisarlo luego se verifica que las columnas se hayan delimitado correctamente. En este caso se eliminaron las columnas de código departamento, código municipio, indicador de pobreza, bajo logro educativo, analfabetismo, inasistencia escolar, rezago escolar, barreras a servicio para el cuidado de la primera infancia, trabajo infantil, desempleo de larga duracion, trabajo informal, sin aseguramiento en salud, barrera de acceso a servicio en salud, sin acceso a fuentes de agua mejorada, inadecuada eliminación de excretas, material inadecuado de pisos', material inadecuado de paredes exteriores, hacinamiento critico, que no serán útiles para nuestro análisis dada la delimitación del hecha en la definición del problema.
- Se reemplaza el valor de un campo, en concreto se reemplaza barrio S4 por indeterminado, para facilitar el análisis a través de las gráficas, realizando una verificación del DF, se concluye que muy posiblemente este registro hace referencia a la zona rural no identificada mediante ningún nombre de vereda.
- Se agrupan las condiciones económicas así:
 - A1 - A5: Pobreza extrema
 - B1 - B7: Pobreza moderada
 - C1 - C18: Vulnerable
 - D1 - D21: No pobre, no vulnerable
- Luego se adiciona la columna Grupo_Sisben al DF final, que será analizado

4. Análisis descriptivo de los datos:

Acciones

- Se crea una tabla en la que se muestre por nombre del barrio, nombre de la vereda y porcentaje de cada grupo del Sisben
- Se hace el conteo de los grupos del Sisben por Barrio y Vereda
- Se calcula el total de datos por grupo del Sisben de cada Barrio y Vereda
- Se calcula el porcentaje de cada grupo Sisben por Barrio y Vereda

5. Visualización y análisis de resultados

Se crean gráficos y visualizaciones para comunicar los hallazgos de manera clara y efectiva.

Acciones:

- Se grafica la tabla anterior con el top 10 de barrios y veredas con mayor porcentaje de hogares no pobres y no vulnerables

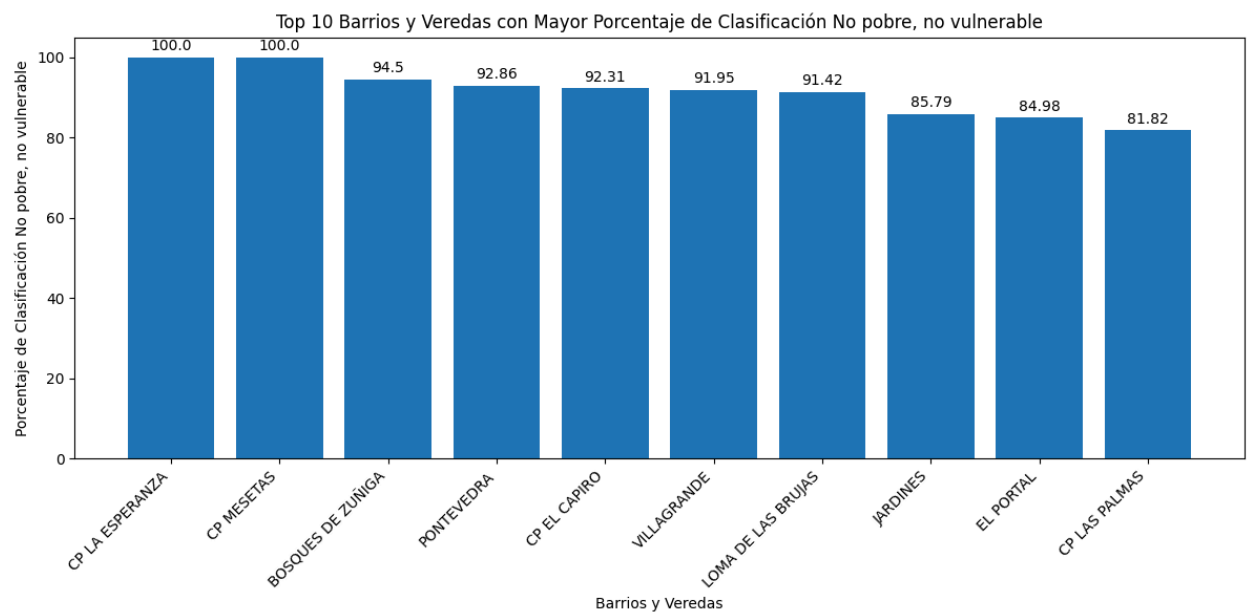


Gráfico 1. Top 10 de barrios y veredas con mayor porcentaje de hogares no pobres y no vulnerables

- Se grafica la tabla anterior con el top 10 de barrios y veredas con mayor porcentaje de hogares en pobreza extrema

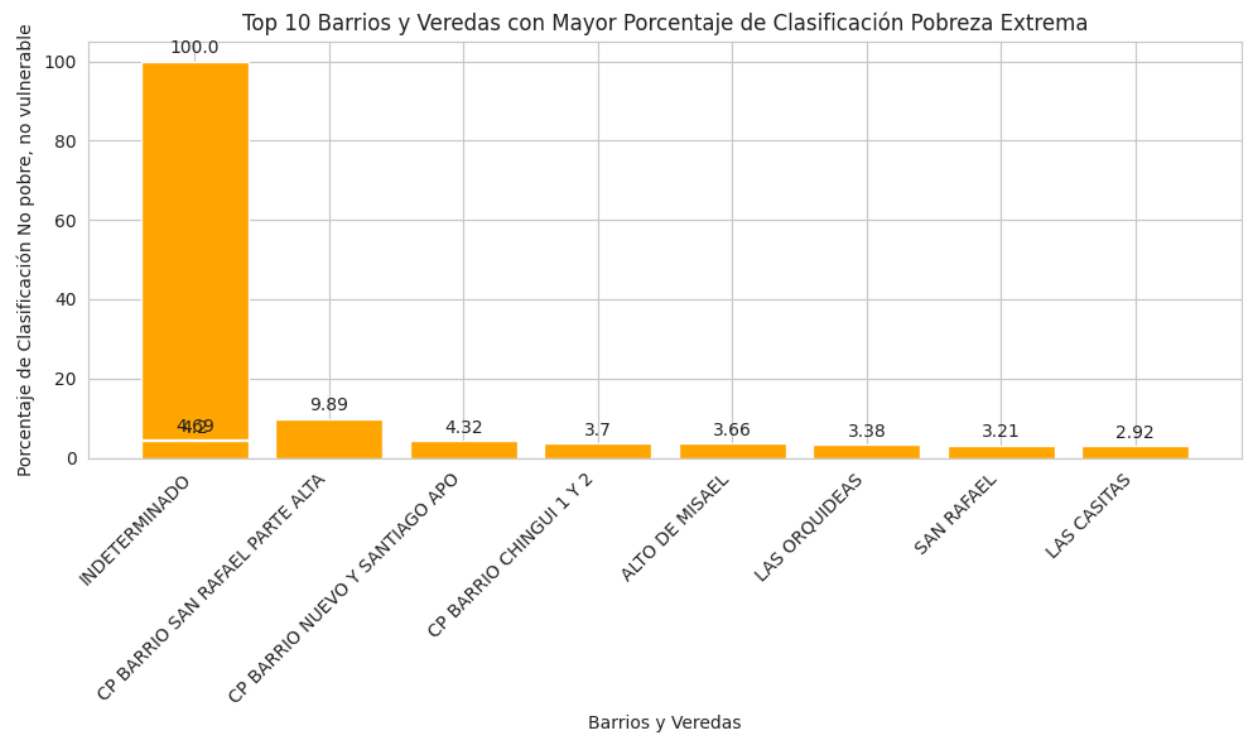


Gráfico 2. Top 10 de barrios y veredas con mayor porcentaje de hogares en pobreza extrema

- Se realiza gráfico para analizar zonas y barrios con mayor reporte de incidencia de inundación, avalancha, terremoto, incendio, vendaval o hundimiento, ordenado de mayor a menor incidencia

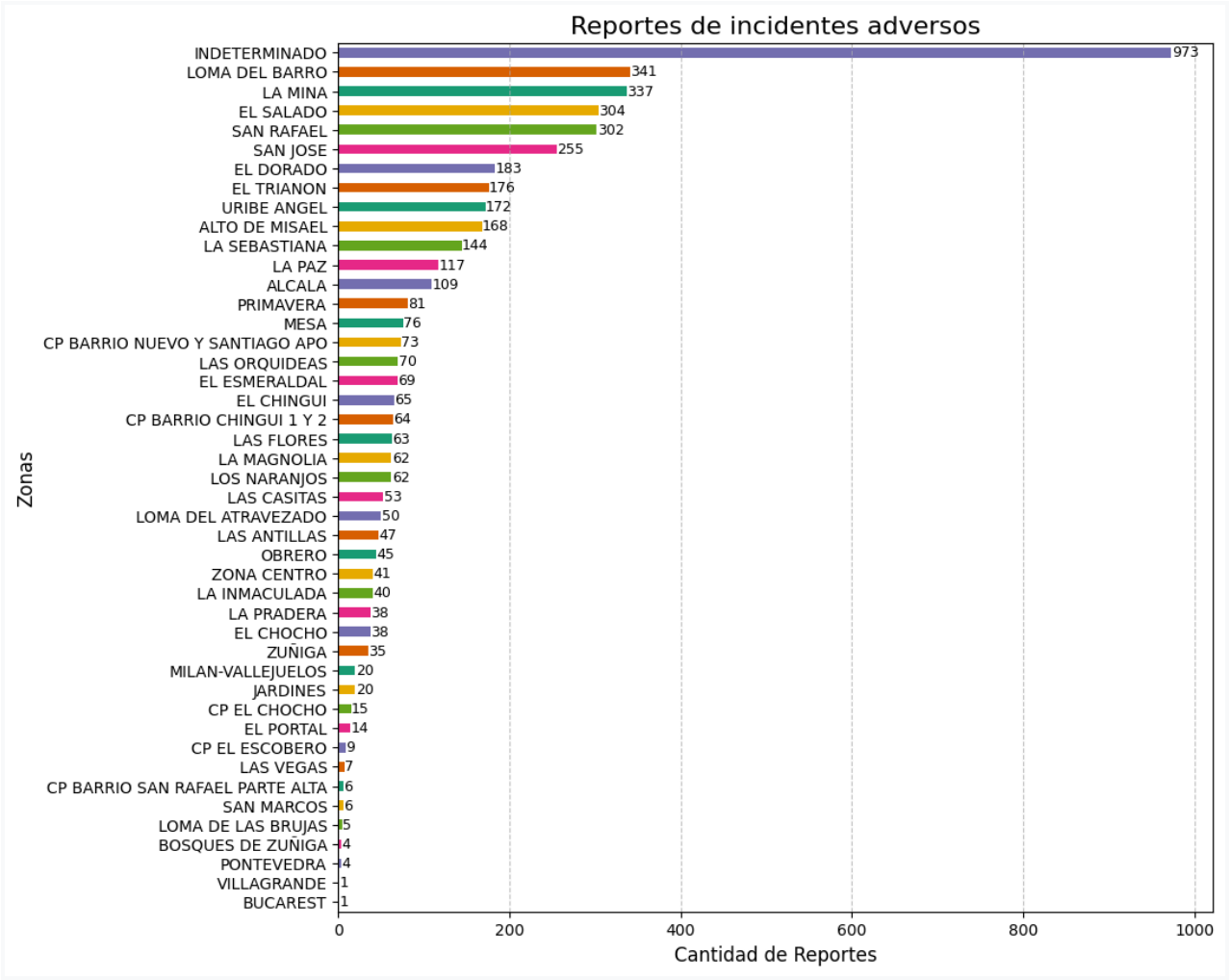


Gráfico 3. Incidencia de eventos adversos por barrio

- Se realiza ahora grafico de los diferente tipos de eventos ocurridos por barrio

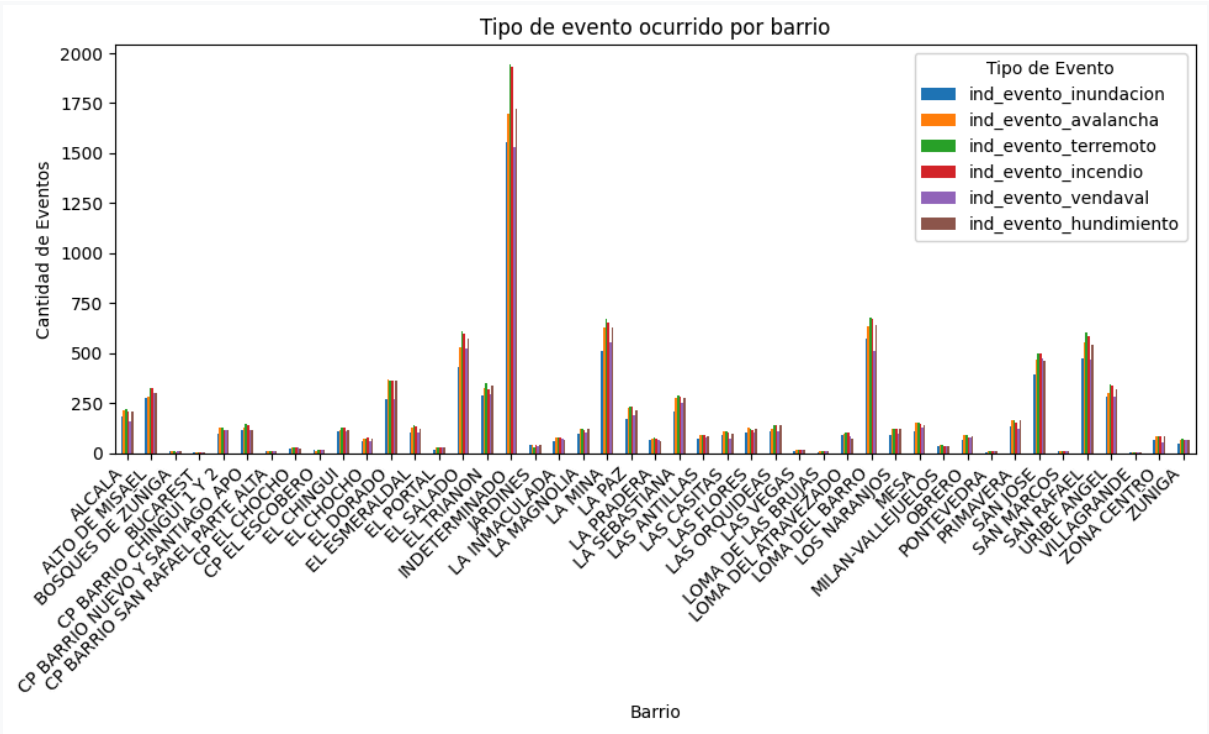


Gráfico 4. Tipos de eventos ocurridos por barrio

- Para facilitar y delimitar el análisis se gráfica un top 10 de los barrios con mayor ocurrencia de eventos adversos por el tipo de evento ocurrido

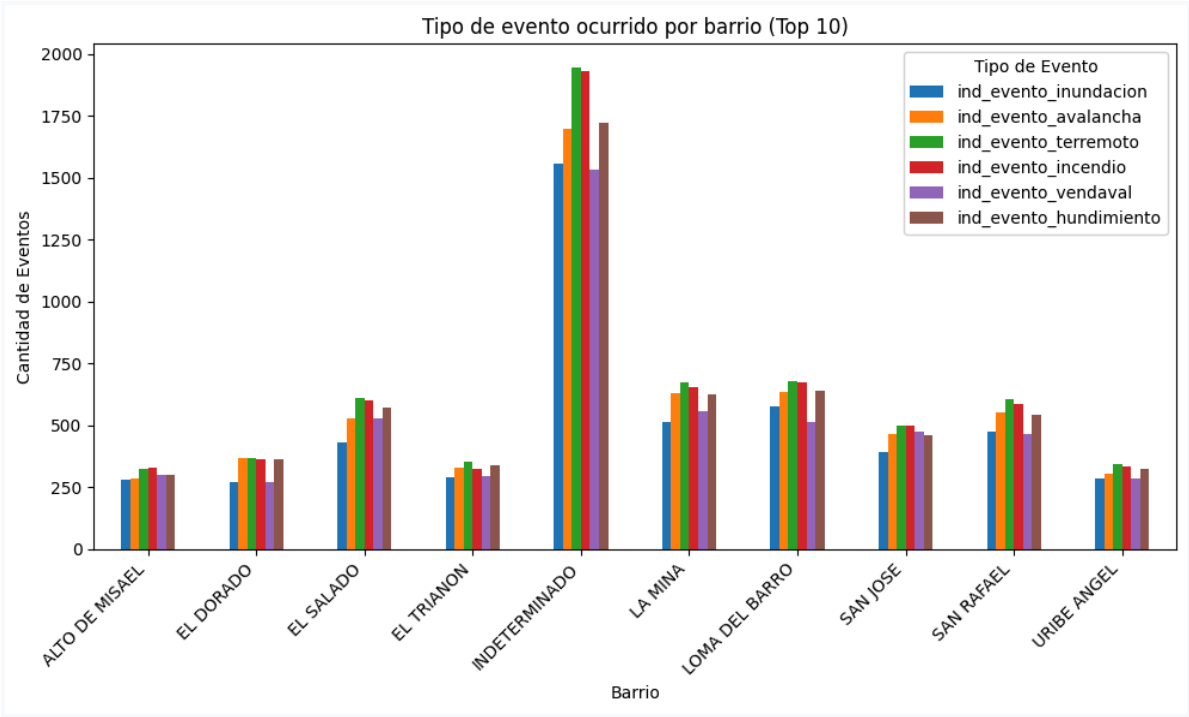


Gráfico 5. Top 10 de los barrios con mayor ocurrencia de eventos adversos por el tipo de evento ocurrido

- Se realiza gráfico que relacione el top 10 de ocurrencia con la categoría SISBEN

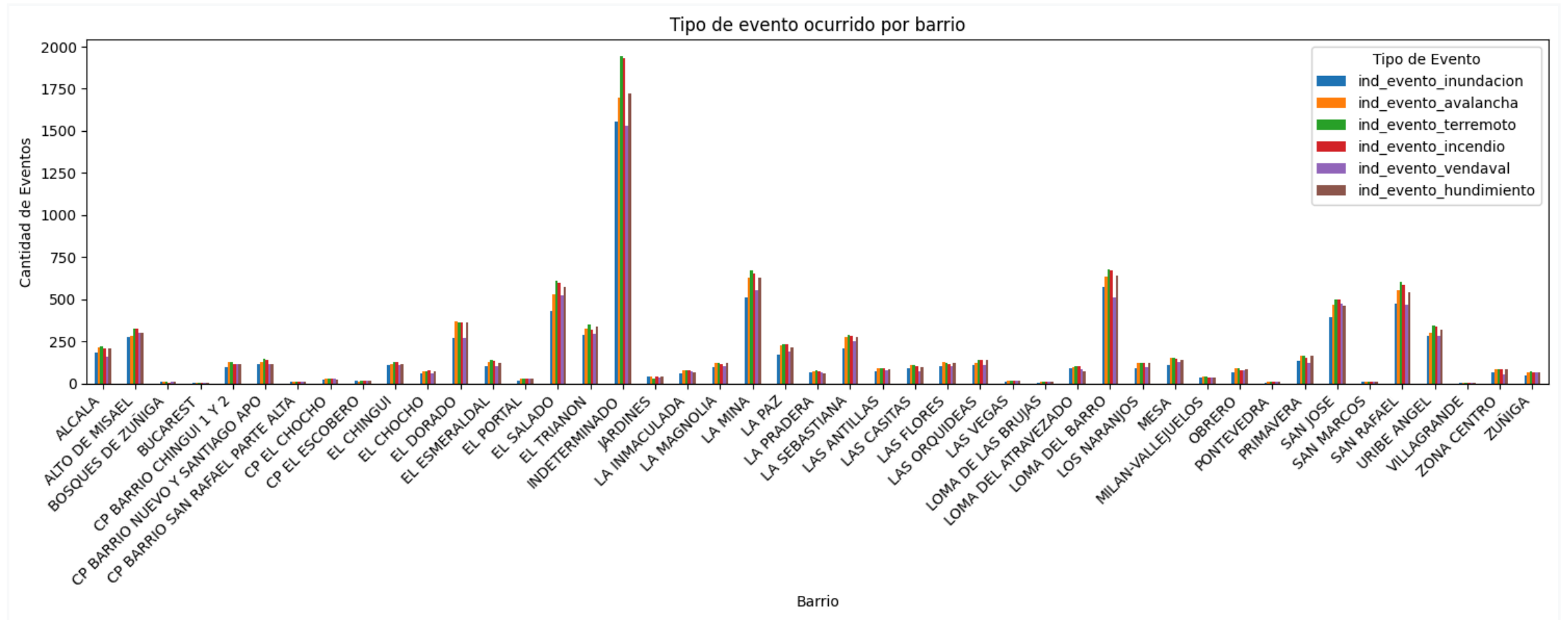


Gráfico 6. Relación entre top 10 de barrios por ocurrencia de eventos adversos y la categoría SISBEN

- Finalmente, se aplica la prueba estadística de hipótesis de Chi-Square para determinar si dos variables categóricas o nominales pueden estar o no relacionadas, para efecto de este análisis se crea una función para verificar la relación entre las variables de tipo de vivienda, tipo de alcantarillado, tipo de pisos y paredes con la variable que indica la ocurrencia de inundaciones.

Análisis de los datos

- Una vez realizada la exploración del dataframe se observa que es una base muy limpia y bien elaborada, ya que no presenta datos nulos y no fue necesario realizar una limpieza de la misma.
- En los registros duplicados se encuentra coherencia con el resultado del análisis porque muestra que los datos duplicados corresponden al código de Departamento y Municipio y el DataFrame tiene datos únicamente del Municipio de Envigado, por tanto, se crea uno nuevo sin estas dos columnas para continuar con el análisis.
- De acuerdo a los datos de la encuesta analizada, la gráfica con el top 10 de los barrios y veredas con mayor porcentaje de hogares No pobres, No vulnerables, son los siguientes:
 1. Centro poblado La Esperanza 100%
 2. Centro poblado Mesetas 100%
 3. Bosques de Zuñiga 94,5%
 4. Pontevedra 92,86%
 5. Centro poblado El Capiro 92,31%
 6. Villagrande 91,95%
 7. Loma de las brujas 91,42%
 8. Jardines 85,79%
 9. El portal 84,98%
 10. Centro Poblado las Palmas 81,82%

Este resultado es de acuerdo a lo esperado, ya que estos barrios son los de mayor estrato socioeconómico del Municipio.

- Al realizar la gráfica con el top 10 de los barrios con pobreza extrema, se visualizan en el siguiente orden:
 1. Indeterminado- Zona rural
 - ❖ Vereda Las Palmas 1,96%
 - ❖ Sin vereda 4,20%
 - ❖ Vereda El Escobero 0%
 - ❖ Vereda El Vallano 4,69%
 - ❖ Vereda Santa Catalina 100%
 2. Centro poblado San Rafael parte alta 9,89%
 3. Centro poblado Barrio Nuevo y Santiago apóstol 4,32%
 4. Centro poblado barrio Chingui 1 y 2- 3,7%
 5. Alto de Misael 3,66%
 6. Las orquideas 3,38%
 7. San Rafael 3,21%
 8. Las casitas 2,92%

A diferencia de los altos porcentajes de viviendas No pobres, No vulnerables, en relación a la pobreza extrema, los datos de la encuesta muestran porcentajes más bajos, y por esto, se puede concluir que en el Municipio predomina una buena condición socio-económica en los hogares.

- Según los resultados del análisis de la encuesta se evidencia que en el sector donde se presenta mayor incidencia de eventos adversos es en los barrios Indeterminado (que al parecer corresponde a la zona rural del Municipio), la Loma del barro y La Mina.

Los eventos adversos reportados en la encuesta son: inundación, avalancha, terremoto, incendio, vendaval y hundimiento.

- De acuerdo al gráfico con los resultados de los eventos adversos, se concluye que en la mayoría de los barrios los eventos con mayor número de ocurrencias son: avalancha, terremoto, incendios y hundimientos. Esto se confirma al analizar el top 10 de los barrios con mayor ocurrencia.
- A partir del gráfico que relaciona los eventos adversos con la clasificación del Sisben, no se evidencia una clara relación entre la ocurrencia de eventos con el nivel socio-económico. En barrios como El Dorado, El Trianon y Loma del barro hay más ocurrencia de eventos adversos en la clasificación No pobre, No vulnerable.
- Después de aplicada la prueba estadística de hipótesis Chi-Square, para la relación entre las variables tipo de vivienda, tipo de alcantarillado, tipo de pisos y paredes con la variable que indica la ocurrencia de inundaciones, se evidencia que el p-value es mucho menor que 0.05 lo que indica que las variables categóricas en esta base de datos, tienen relación y pueden tomarse como base para futuros análisis y toma de decisión.

Enlace GitHub: <https://github.com/lufemartin/ProyectoEquipo3Analisis.git>

Recomendaciones