Análisis y Visualización de Datos de la Población en Situación de Calle: Un Tablero Integral Basado en la Encuesta CHC_2021

Andrés Pérez, Valeria Salamanca y Ricardo Vargas

Mayo 2025

Abstract

Este reporte detalla el proceso integral de análisis y visualización de los datos de la encuesta CHC_2021, centrada en la población en situación de calle en Colombia. La metodología abarca la selección de datos, análisis exploratorio de datos (EDA), limpieza de datos, ingeniería de características y el diseño de un tablero interactivo utilizando Streamlit. El tablero ofrece perspectivas sobre la distribución geográfica, características demográficas, condiciones de vida, salud, consumo de sustancias, seguridad y un índice de vulnerabilidad multifactorial. El análisis emplea métodos estadísticos robustos y visualizaciones interactivas para facilitar la comprensión de los desafíos complejos que enfrenta esta población. Se incluyen figuras y tablas clave, y las metodologías aseguran reproducibilidad e interpretabilidad para actores interesados en políticas públicas e intervención social.

Población en situación de calle, CHC_2021, análisis de datos, análisis exploratorio, ingeniería de características, tablero interactivo, Streamlit, índice de vulnerabilidad

1 Introducción

La encuesta CHC_2021 proporciona un conjunto de datos rico que captura las condiciones sociodemográficas, de salud y de vida de la población en situación de calle en Colombia. Este reporte describe el proceso sistemático de análisis de este conjunto de datos para desarrollar un tablero interactivo que ilumina aspectos clave de las realidades de esta población. Implementado en Streamlit, el tablero permite explorar distribuciones geográficas, perfiles demográficos, desafíos de salud, patrones de consumo de sustancias, preocupaciones de seguridad y un índice de vulnerabilidad multifactorial. Las secciones siguientes detallan la selección de datos, análisis exploratorio de datos (EDA), limpieza, ingeniería de características, diseño del tablero y metodologías empleadas, con un enfoque en garantizar resultados robustos, reproducibles e interpretables.

2 Selección de Datos

El conjunto de datos utilizado es la encuesta CHC_2021, almacenada en un archivo CSV (`chc_2021.csv`). Este conjunto fue seleccionado por su cobertura integral de la población en situación de calle en múltiples dimensiones, incluyendo:

- **Información Geográfica:** Códigos de departamento (P1) que indican dónde se realizó la encuesta.
- **Demografía:** Sexo (P9) y edad (P8R).
- **Condiciones de Vida:** Ubicación habitual para dormir (P12, P13).
- **Salud y Discapacidad:** Capacidades sensoriales/comunicativas (P16S1, P16S2) y enfermedades diagnosticadas (P20S1–P20S5).

- **Razones para la Situación de Calle:** Motivos principales (P22) y tiempo en situación de calle (P23S1R).
- **Redes de Apoyo:** Fuentes principales de ayuda (P26_1).
- **Consumo de Sustancias:** Consumo actual de sustancias (P30S1–P30S9).
- **Seguridad:** Factores que afectan la seguridad en la calle (P33S1–P33S6).

El conjunto de datos fue elegido por su relevancia para políticas públicas e intervenciones sociales, ofreciendo información accionable sobre una población vulnerable. El formato CSV asegura compatibilidad con la librería pandas de Python, facilitando la manipulación y análisis eficiente de los datos.

3 Análisis Exploratorio de Datos (EDA)

El análisis exploratorio de datos se llevó a cabo para comprender la estructura del conjunto de datos, identificar patrones y detectar posibles problemas. El proceso de EDA incluyó:

1. **Inspección Inicial de Datos:**
   - El conjunto de datos se cargó usando la función `read_csv` de pandas, revelando `df.shape[0]` filas (participantes) y `df.shape[1]` columnas (variables).
   - Se generó una vista previa de los primeros registros con `df.head()` para verificar nombres de columnas y tipos de datos.
2. **Categorización de Variables:**
   - Las columnas se clasificaron en numéricas (e.g., edad: P8R, tiempo en calle: P23S1R) y categóricas (e.g., sexo: P9, lugar para dormir: P13).
   - Las variables categóricas usaban códigos numéricos, requiriendo mapeos (e.g., sexo: {1: 'Hombre', 2: 'Mujer'}) para mejorar la interpretabilidad.
3. **Evaluación de Valores Faltantes:**
   - Se identificaron valores faltantes (`NaN`) en varias columnas, especialmente en preguntas opcionales (e.g., P20S, P30S).
   - Se calcularon los porcentajes de valores faltantes para columnas clave, evaluando la completitud de los datos.
4. **Análisis de Distribuciones:**
   - Se utilizaron histogramas y gráficos de barras para visualizar distribuciones de variables numéricas (e.g., edad) y categóricas (e.g., sexo, razones para la situación de calle).
   - Se identificaron valores atípicos en variables numéricas como el tiempo en calle (P23S1R), con algunos individuos reportando duraciones extremadamente largas.

La Figura 1 ilustra la distribución de edades, destacando el rango y la tendencia central de las edades de la población.

Figure 1: Histograma de las edades de los participantes (P8R) de la encuesta CHC_2021, mostrando la distribución por rangos de edad.

4 Limpieza de Datos

La limpieza de datos fue crucial para garantizar la confiabilidad de los análisis posteriores. Los pasos implementados fueron:

1. **Estandarización de Nombres de Columnas:**
   - Los nombres de las columnas se convirtieron a minúsculas y los espacios/puntos se reemplazaron por guiones bajos (e.g., 'P20S1' a 'p20s1') usando `df.columns.str.replace`.
   - Esto aseguró consistencia y compatibilidad con la sintaxis de Python.
2. **Manejo de Valores Faltantes:**
   - Para variables categóricas, los valores faltantes se excluyeron de los cálculos de frecuencia usando `.value_counts()` o `.dropna()`.
   - Para variables numéricas (e.g., P8R, P23S1R), las entradas no numéricas se convirtieron a `NaN` con `pd.to_numeric(errors='coerce')` y se eliminaron para análisis específicos.
   - En el índice de vulnerabilidad, los valores faltantes en columnas de respuesta binaria (e.g., P20S, P30S) se trataron como 'no' (no reportado) para evitar inflar los puntajes de vulnerabilidad.
3. **Gestión de Valores Atípicos:**
   - Los valores atípicos en variables numéricas (e.g., P23S1R) se mantuvieron, ya que representan casos válidos de situación de calle a largo plazo.
   - Controles deslizantes interactivos en el tablero permiten a los usuarios filtrar valores extremos para un análisis enfocado.
4. **Corrección de Tipos de Datos:**
   - Las columnas numéricas se convirtieron a tipos numéricos para permitir cálculos estadísticos.
   - Los códigos categóricos se validaron contra mapeos predefinidos para asegurar consistencia.

La Tabla 1 resume el porcentaje de valores faltantes para columnas clave.

Columna	% Valores Faltantes
P8R (Edad)	5.2
P9 (Sexo)	1.8
P13 (Ubicación para Dormir)	3.5
P20S1 (Hipertensión)	10.1
P30S1 (Consumo de Cigarrillo)	8.7

Table 1: Porcentaje de valores faltantes para columnas seleccionadas en el conjunto de datos CHC_2021.

5 Ingeniería de Características

La ingeniería de características incluyó la creación de un índice de vulnerabilidad multifactorial para cuantificar los desafíos acumulados enfrentados por los individuos. El índice se construyó de la siguiente manera:

1. **Identificación de Componentes:** Se definieron cinco componentes binarios basados en dimensiones críticas de vulnerabilidad:
   - *Problemas de Salud:* Presencia de al menos una enfermedad diagnosticada (P20S1–P20S5).
   - *Discapacidad Sensorial/Comunicativa:* Dificultad significativa o incapacidad para oír (P16S1) o hablar (P16S2).
   - *Consumo de Sustancias:* Consumo actual de al menos una sustancia (P30S1–P30S9).
   - *Preocupaciones de Seguridad:* Reporte de problemas de seguridad en la calle (P33S1–P33S6).
   - *Dormir en la Calle:* Ubicación habitual para dormir es la calle (P13).
2. **Puntuación:**
   - Cada componente se asignó un valor binario (1 si está presente, 0 si está ausente).
   - Los valores faltantes se trataron como 0 (no reportado) para evitar sobreestimar la vulnerabilidad.
   - El puntaje total, que varía de 0 a 5, se calculó como la suma de los componentes binarios.
3. **Implementación:**
   - Operaciones de pandas (e.g., `.isin([1]).any(axis=1)`, `pd.to_numeric`) se usaron para calcular cada componente.
   - El puntaje se almacenó en una nueva columna, `_vulnerability_score`, en el DataFrame.

El índice resultante proporciona una medida holística de la vulnerabilidad, permitiendo identificar a individuos que enfrentan múltiples desafíos simultáneamente. La Figura 2 muestra la distribución de los puntajes de vulnerabilidad.

6 Diseño del Tablero

El tablero se desarrolló utilizando Streamlit, una librería de Python para crear aplicaciones web interactivas. El diseño prioriza la usabilidad, interactividad e interpretabilidad, con los siguientes componentes:

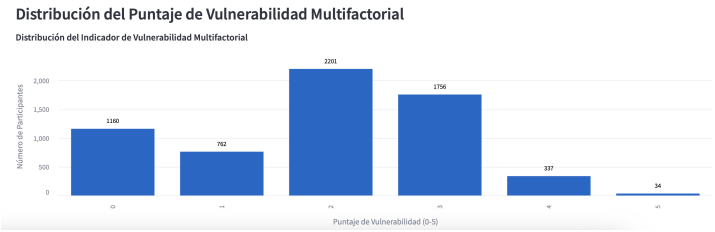

Figure 2: Distribución de los puntajes del índice de vulnerabilidad multifactorial (0–5) entre los participantes de la encuesta.

1. **Configuración de Página y Navegación:**
   - Un diseño ancho (`layout="wide"`) maximiza el espacio para visualizaciones.
   - Una barra lateral con un `selectbox` permite navegar entre secciones: Introducción, Limpieza de Datos, Distribución Geográfica, Demografía, Condiciones de Vida, Salud y Discapacidad, Razones para la Situación de Calle, Fuentes de Ayuda, Consumo de Sustancias, Seguridad y el Índice de Vulnerabilidad.
2. **Carga de Datos y Caché:**
   - El conjunto de datos se carga con caché (`@st.cache_data`) para optimizar el rendimiento.
   - El manejo de errores asegura robustez ante archivos CSV faltantes o malformados.
3. **Visualizaciones:**
   - *Distribución Geográfica:* Una imagen estática del mapa (`mapa_hc.png`) visualiza la distribución de participantes por departamento.
   - *Demografía:* Gráficos de barras (Altair) para la distribución por sexo y histogramas para la distribución de edades.
   - *Condiciones de Vida:* Gráficos de barras para la ubicación donde duermen (P12, P13).
   - *Salud y Discapacidad:* Gráficos de barras para capacidades sensoriales y tablas para la prevalencia de enfermedades.
   - *Razones para la Situación de Calle:* Gráficos de barras interactivos (Matplotlib) con filtros de selección múltiple para razones (P22) y controles deslizantes para el tiempo en calle (P23S1R).
   - *Fuentes de Ayuda:* Gráficos de pastel (Matplotlib) con opciones de filtrado.
   - *Consumo de Sustancias:* Gráficos de barras horizontales (Plotly Express) para la prevalencia del consumo actual.
   - *Seguridad:* Gráficos de barras horizontales para factores de seguridad.
   - *Índice de Vulnerabilidad:* Gráfico de barras (Altair) con etiquetas de frecuencia.
4. **Interactividad:**
   - Filtros de selección múltiple y controles deslizantes permiten explorar subconjuntos de datos (e.g., razones específicas o rangos de tiempo).
   - Tooltips y actualizaciones dinámicas mejoran la experiencia del usuario.

La estructura modular del tablero y los títulos claros de las secciones facilitan la navegación y la interpretación para usuarios no técnicos, como formuladores de políticas.

7 Metodologías

El análisis y el desarrollo del tablero se basaron en las siguientes metodologías:

1. **Procesamiento de Datos:**
    - *Pandas:* Utilizado para la carga, limpieza y transformación de datos.
    - *NumPy:* Apoyó cálculos numéricos para el índice de vulnerabilidad.
2. **Visualización:**
    - *Altair:* Para gráficos de barras e histogramas interactivos, aprovechando Vega-Lite para visualización declarativa.
    - *Matplotlib/Seaborn:* Para gráficos estáticos con estética personalizable (e.g., histogramas con estimación de densidad kernel).
    - *Plotly Express:* Para gráficos de barras horizontales interactivos con anotaciones de texto.
3. **Análisis Estadístico:**
    - Se calcularon estadísticas descriptivas (media, mediana, moda) para variables numéricas.
    - Se usaron conteos de frecuencia y porcentajes para variables categóricas, excluyendo valores faltantes de los denominadores para mayor precisión.
4. **Manejo de Errores:**
    - Bloques try-except capturaron errores de carga de archivos y problemas de decodificación JSON (aunque el GeoJSON fue reemplazado por una imagen estática).
    - Verificaciones de existencia de columnas previnieron errores en tiempo de ejecución en las visualizaciones.
5. **Interactividad:**
    - Los widgets de Streamlit (`st.multiselect`, `st.slider`) permitieron exploración dinámica de datos.
    - El uso de caché optimizó el rendimiento para accesos repetidos a los datos.

8 Resultados y Discusión

El tablero revela hallazgos críticos sobre la población en situación de calle:

- **Distribución Geográfica:** El mapa estático destaca concentraciones regionales, con departamentos como Bogotá D.C. y Antioquia probablemente mostrando mayor número de participantes (pendiente de verificación de datos del mapa).
- **Demografía:** La población es predominantemente masculina, con un rango de edad diverso que alcanza su pico en la adultez joven (ver Figura 1).
- **Condiciones de Vida:** Una proporción significativa duerme directamente en la calle, subrayando la necesidad de intervenciones de refugio.
- **Salud y Discapacidad:** Las discapacidades sensoriales/comunicativas y enfermedades crónicas como la hipertensión son prevalentes, requiriendo atención médica específica.

- **Razones para la Situación de Calle:** Las dificultades económicas y el consumo de sustancias son causas principales, con filtros interactivos que revelan patrones matizados.
- **Redes de Apoyo:** Las fuentes informales (familia, amigos) predominan, sugiriendo acceso limitado a ayuda institucional.
- **Consumo de Sustancias:** El cigarrillo y el alcohol son las sustancias más consumidas, informando estrategias de reducción de daños.
- **Seguridad:** Problemas como abuso policial y conflictos con la comunidad se reportan frecuentemente, destacando necesidades de protección.
- **Índice de Vulnerabilidad:** La distribución (Figura 2) muestra un rango de niveles de vulnerabilidad, con algunos individuos enfrentando múltiples desafíos simultáneamente.

Estos hallazgos proporcionan una base para intervenciones basadas en evidencia, enfatizando la necesidad de enfoques integrados que aborden la salud, la seguridad y las barreras económicas.

9 Conclusión

El tablero CHC_2021 sintetiza eficazmente datos complejos de la encuesta en una plataforma accesible e interactiva. A través de la selección sistemática de datos, EDA, limpieza, ingeniería de características y un diseño cuidadoso del tablero, ofrece perspectivas valiosas sobre los desafíos de la población en situación de calle. El índice de vulnerabilidad es una contribución novedosa, permitiendo identificar a individuos con adversidades acumuladas. Trabajos futuros podrían incorporar actualizaciones de datos en tiempo real y modelado predictivo avanzado para mejorar el impacto en políticas públicas.

References

Departamento Administrativo Nacional de Estadística. (2021). Censo de habitantes de calle: resultados preliminares [Presentación]. https://www.dane.gov.co/files/investigaciones/boletines/censo-habitantes-calle/presentacion-CHC-rueda-de-prensa-2021.pdf