Universidad de los Andes
Facultad de Economía

Exploración y visualización de datos para lo socioeconómico

Miguel Andrés Garzón Ramírez

11 de septiembre de 2025

Actividad práctica 2: Explorando datos sobre migrantes venezolanos en Colombia

Situación

Su grupo de trabajo hace parte de un equipo técnico que apoya a una alcaldía en un programa de mejora de la calidad de vida de hogares venezolanos. La alcaldía quiere mejorar la **focalización** de su respuesta en vivienda y servicios básicos, pero no sabe **qué factores son los más importantes para identificar** condiciones de vida deficientes.

Condiciones

- Se debe trabajar en los equipos definidos. Cada integrante del equipo debe interactuar y contribuir con sus capacidades y conocimientos en el trabajo. Lo ideal es que cada miembro del equipo tenga un rol y contribuya de acuerdo a él.
- Puede usar cualquier herramienta de su preferencia: Excel, Stata, R, Python, Power BI u otro.
- Se debe entregar:
    - Código o proyecto (Stata, R, Python, Excel/Power Query o Power BI) con pasos reproducibles. Se debe observar un proceso de inicio a fin.
    - Documento escrito donde se dé respuesta a los requerimientos de cada una de las partes y pasos del enunciado.
- Cuenta con los datos de la Evaluación Conjunta de Necesidades a población refugiada y migrante, mostrada en clase:
- Fecha de entrega: domingo 21 de septiembre

Objetivo analítico

Para identificar estas condiciones es necesario **explorar** aspectos de calidad de vida, tanto por separado como simultáneamente, apoyándose con conocimiento de experto en el dominio[1] de interés.

Parte 1: Planteamiento de la pregunta de indagación

Para seleccionar o formular la pregunta adecuadamente es muy importante que realice una investigación mínima sobre las condiciones de vivienda y el acceso a servicios básicos de la población a analizar, para establecer con evidencia los criterios más apropiados para el análisis. Luego se debe establecer si con los datos disponibles es posible responder esta pregunta, ¿hay alguna variable que sea latente y deba medirse a través de una variable proxy?[2]

[1] Entiéndase "dominio" como el contexto bajo el cual los datos disponibles se pueden analizar, para generar información a partir de su interpretación.
[2] Es muy importante que la definición de variables latentes y proxi esté documentada en el ejercicio.

1. Guíese a partir de una **pregunta de indagación**. Puede usar alguna de las siguientes u otra distinta relacionada con la situación:
   - ¿Cómo cambia la tenencia y por carencias de servicios?
   - ¿Los hogares con NNA (niños, niñas y adolescentes) o adultos mayores presentan patrones distintos de hacinamiento que hogares de solo personas adultas?
   - ¿La incidencia de carencias en servicios públicos es mayor en subarriendo que en arriendo formal?
   - ¿El tiempo en Colombia amortigua los riesgos (menos carencias, menor hacinamiento)?

2. Consecuentemente con la pregunta, seleccione una **variable de medición (Y, dependiente)** según disponibilidad en la hoja *Hogares* (revisen el diccionario para ubicar la variable exacta). Opciones sugeridas:
   - Déficit de servicios básicos: falta de agua potable / saneamiento / electricidad / gas / internet (Puede construir un conteo de carencias "0–5").
   - Tenencia de vivienda: propio/arriendo/subarriendo/ocupación de hecho.
   - Hacinamiento: personas por cuarto de dormir (umbral >3).
   - Dependencia = (NNA + adultos mayores) / adultos en edad laboral.

Si no hay una variable directa, opere la mejor proxy con base en el diccionario (ej.: "tenencia de vivienda" + "dificultades de pago" + "servicios" → vulnerabilidad residencial).

3. De la misma manera, de acuerdo con la pregunta seleccione una **variable independiente (X, factores explicativo que puede ser categórico)**, puede ser definida como alguna de las siguientes:
   - Tamaño y composición del hogar: número de miembros, presencia de NNA, adultos mayores o personas con discapacidad.
   - Trayectoria migratoria y anclaje: tiempo en Colombia, vía de ingreso, intención de quedarse.
   - Medios de vida, que puede ser medido con:
     - Empleo del jefe/a de hogar o situación laboral. fuentes de ingreso (salarios, remesas), irregularidad de ingresos.
     - Ingresos insuficientes para cubrir gastos básicos o indicador más cercano disponible.

Ubique esta pregunta con un alcance geográfico determinado, teniendo en cuenta el tiempo disponible para realizar esta actividad y los datos disponibles. Los análisis estarán condicionados con esta selección.

Con esta definición de variables, en el informe haga un contexto más específico de la situación, una descripción de la pregunta de indagación teniendo en cuenta los siguientes aspectos:

- Población: Hogares migrantes venezolanos…
- Unidad de observación: Hogar o persona
- Variables dependientes e independientes en el análisis
- Dimensiones de análisis de interés

- Ámbito : Geografía/periodo de referencia
- Latente: ¿Existe? Proxy propuesta (si aplica)
- Relevancia: ¿Por qué esta pregunta es útil para la gestión?

Parte 2: Análisis univariado

Antes de cruzar variables, se debe entender **cada variable por sí sola**: su codificación, distribución, faltantes y posibles errores. Esto asegura que las comparaciones bivariadas no estén sesgadas por datos mal definidos.

¿Qué abordar para *cada* variable? Tipo y codificación: ¿numérica, categórica (nominal/ordinal), fecha, texto? ¿Binarias están en 0/1?

- Distribución y rango: forma, asimetrías, cortes naturales (p. ej., 0, 1–2, 3+).
- Faltantes y no respuesta: cuantifica NA y categorías como "No sabe/No responde" para decidir si se recodifican como missing.
- Valores atípicos/ceros estructurales: ¿outliers plausibles o errores de captura? ¿Ceros significan ausencia real o no aplica?
- Cardinalidad de categorías (si es categórica): ¿hay categorías con una cantidad de observaciones muy pequeña que debas agrupar?
- Ponderación: si hay factor de expansión, preparar versiones ponderadas de frecuencias y medias.

Por variable, y según sea el caso incluya en el informe:

- Tabla de frecuencias o estadísticas descriptivas de acuerdo con su relevancia con la pregunta de indagación.
- 1–2 gráficos por variable crítica (p. ej., histograma de personas/cuarto, barras de tenencia)
- Notas de método (ponderación, tratamiento de NA y outliers). Es decir, haga un resumen de las decisiones de limpieza/transformación por variable

Recomendación: no avance a la asociación condicional de variables (E[Y|X]E[Y|X]) hasta que pueda explicar, con un gráfico y dos frases, cómo es cada variable por sí sola y cómo se limpió.

Parte 3: Visualizaciones de exploración según la pregunta de indagación

Seleccionar la visualización más simple que responda directamente a la pregunta de indagación. Prioricen mostrar medias condicionales E[Y|X]E[Y|X], distribuciones y comparaciones entre grupos.

Algunas ideas, se debe seleccionar de acuerdo a la pregunta y los datos disponibles:

- Y binaria y X categórica: Barras de la media de Y por categorías de X (la media es una tasa/probabilidad).
  Ej.: Tasa de déficit de servicios por tenencia (propio/arriendo/subarriendo).
- Y continua y X categórica: Boxplots o barras de medias con barra de error (IC95%).
  Ej.: Personas por cuarto por tenencia.

- Y binaria y X continua: Líneas/columnas de la media de Y por tramos (cuantiles/deciles) de X; o barras apiladas por tramos. Ej.: Tasa de hacinamiento alto por tiempo en Colombia (en bandas).
- Y continua y X continua: Dispersión (scatter). Si hay mucha variabilidad use tramos de X (categorícela) para mostrar la media de Y por tramo. Ej.: gastos del hogar en alimentación contra gastos del hogar en alquiler de vivienda.
- Y y X categóricas (ambas): Barras agrupadas o paneles para explorar interacciones visuales. con conteos de observaciones. Ej.: Déficit de servicios por tenencia, separado por presencia de NNA (sí/no).

Incluya los gráficos que considere pertinentes y que ilustren la relaciones entre variables de la pregunta de indagación. Acompáñelos de un análisis interpretativo y notas sobre el proceso de visualización.