Universidad de los Andes
Facultad de Economía

# Exploración y visualización de datos para lo socioeconómico

Miguel Andrés Garzón Ramírez

13 de noviembre de 2025

## Proyecto 2- Visualización y construcción de historias con datos

Este proyecto busca consolidar los aprendizajes de todo el curso, integrando el análisis exploratorio de datos, la modelación de datos y la comunicación visual. A partir de una pregunta de índole social, vamos a construir un proceso completo que incluya la búsqueda y depuración de datos, el desarrollo de un análisis exploratorio (EDA) con evidencia descriptiva, la creación de un modelo de datos y el diseño de una visualización interactiva que, además de contribuir con la exploración de los datos tanto propia como de la audiencia, permita apoyar la comunicación de los hallazgos de manera clara y argumentada.

El proyecto articula las dos dimensiones vistas en el curso análisis exploratorio y visualización con el fin de demostrar dominio técnico y conceptual sobre la construcción de historias con datos.

### Condiciones

- Este proyecto se puede trabajar en los mismos equipos del proyecto 1.
- Fecha de entrega: miércoles 10 de diciembre - Sustentaciones: jueves 11 de diciembre (horario por definir)
- Se debe formular una pregunta concreta de interés social, vinculada a un contexto claramente definido (por ejemplo, condiciones de vida, empleo, educación, pobreza, salud, vivienda, o temas afines a sus trabajos de grado).
- La fuente de datos debe permitir abordar empíricamente la pregunta. Se recomienda el uso de encuestas de hogares (como la GEIH o la ECV del DANE), aunque se aceptan otras fuentes de calidad si son pertinentes con la pregunta y el contexto.
- Se espera que el análisis sea reproducible y muestre un razonamiento lógico que conecte la pregunta, la evidencia y la visualización final.
- Este proyecto cuenta con 3 partes, que consisten en la formulación de la pregunta de indagación, el desarrollo de un EDA y modelamiento de datos, y el desarrollo de una visualización sencilla para comunicar los resultados
- Se debe entregar:
    - Código reproducible (Stata, R, Python o Power Query): Donde se observe el proceso completo del EDA: carga, limpieza, transformación, análisis descriptivo y generación de evidencia estadística de base (por ejemplo, intervalos, contrastes o diferencias).
    - Modelo de datos (Power BI o código): Debe mostrar la estructura de relaciones entre las tablas y reflejar las decisiones tomadas en la preparación de los datos (por ejemplo, normalización, relaciones, tablas de hechos y dimensiones).
    - Visualización interactiva (archivo .pbix o similar): Que permita explorar los resultados más relevantes del análisis, mostrar comparaciones o relaciones clave, y

comunicar la respuesta a la pregunta planteada mediante gráficos claros y organizados.

- Informe breve (máx. 3 páginas):
  - Formulación de la pregunta y descripción del contexto.
  - Descripción de las fuentes y variables utilizadas.
  - Principales hallazgos del análisis exploratorio.
  - Interpretación y respuesta sustentada a la pregunta planteada.
  - Uso de gráficos del EDA o de la visualización seleccionados para respaldar los hallazgos.

# Desarrollo

## Parte 1 – Planteamiento de la pregunta de indagación principal

Formule una **pregunta de interés de análisis**. Estos son algunos ejemplos de preguntas iniciales, que deben ser transformadas a preguntas más específicas para realizar un análisis exploratorio de datos:

- ¿Cómo varía el nivel de educación según la región en Colombia?
- ¿Cuál es la relación entre la jefatura de hogar femenina y el acceso a servicios básicos?
- ¿Qué factores están asociados al hacinamiento en los hogares colombianos?
- ¿Existen disparidades de género en la participación laboral en diferentes regiones del país?
- ¿Cuál es la tasa de acceso a servicios de salud según la condición laboral?
- ¿Cómo varía la calidad de la vivienda según la zona urbana o rural?
- ¿Qué factores están asociados a la informalidad laboral en Colombia?
- ¿Se observan diferencias regionales en el acceso a internet según el nivel educativo del hogar?

Junto con la pregunta, **describa una situación o contexto** en el cual la respuesta a ella pueda ser útil para tomar una decisión de inversión o de acción. Al final, el propósito de responder esta pregunta es proponer una acción sobre la población que está siendo identificada en alguna problemática social que se aborda en la pregunta. Por ejemplo, imagine que su grupo de trabajo hace parte de un equipo técnico que apoya la toma de una decisión basada en evidencia. En esta parte puede explicar cómo se analiza el tema de interés y sirve para definir las dimensiones de análisis en los datos (variables categóricas).

De manera similar a la Parte 1 de la Actividad 2, defina la variable de medición **(Y, dependiente-hecho)** y la variable independiente **(X, factor explicativo o dimensión principal).** 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

- Definición de la población
- Unidad de observación:
- Variables dependientes e independientes en el análisis
- Dimensiones de análisis de interés: Esta definición es muy importante porque le permite identificar las variables que requiere utilizar en el proyecto.
- Ámbito: Geografía/periodo de referencia
- Latente: ¿Existe? Proxy propuesta (si aplica)

- Relevancia: ¿Por qué esta pregunta es útil para tomar una decisión?

## Parte 2 – EDA y modelamiento de datos

El propósito de esta parte es construir un EDA que permita comprender el comportamiento de las variables clave y generar evidencia descriptiva que oriente la respuesta a la pregunta de indagación. Así como en la Actividad 2, el EDA es un proceso de razonamiento que le permite clarificar las relaciones, depurar los datos y tomar decisiones sobre cómo modelar y visualizar los datos.

En esta fase, su tarea consiste en entender los datos antes de pretender explicarlos. Esto implica examinar cada variable de interés seleccionada por separado, evaluar su calidad y luego avanzar a comparar su comportamiento según las dimensiones de análisis definidas en la Parte 1.

### 2.1. Preparación y comprensión inicial de los datos

Antes de producir cualquier gráfico o métrica comparativa, deténgase a estudiar la forma, codificación y calidad de sus datos. Un buen análisis comienza por saber exactamente qué representa cada variable y cómo fue registrada.

En el código reproducible debe observarse:

- Carga de los datos y verificación de estructuras básicas (número de observaciones, número de variables, tipos de variables).
- Revisión de codificación:
  - ¿Sus variables numéricas son realmente numéricas?
  - ¿Las categorías están completas y correctamente etiquetadas?
  - ¿Hay valores como “No sabe/No responde” mezclados con categorías válidas?
  - ¿Cómo se comporta el factor de expansión? (si aplica)
- Tratamiento de faltantes donde se registre cómo se manejan NA y categorías especiales.
- Identificación de valores atípicos o valores imposibles
- Recodificaciones necesarias, como reagrupar categorías raras, crear bandas, construir conteos de carencias o generar proxies.

Todo ajuste debe quedar documentado mediante comentarios o un resumen breve dentro del informe a modo de notas de método

### 2.2. Análisis univariado sistemático

Antes de cruzar variables debe producir una comprensión clara de cada variable crítica teniendo en cuenta su distribución, categorías relevantes, patrones básicos se observan, cómo se comporta según el factor de expansión (si aplica)

Para cada variable relevante incluya:

- Tablas de frecuencia o estadísticas descriptivas, según su tipo.
- Uno o dos gráficos univariados pertinentes como histogramas, gráficos de barras, boxplots simples, diagramas de densidad.

Este análisis sirve como la base conceptual del proyecto: sin esta revisión, cualquier cruce posterior puede ser engañoso o incorrecto.

### 2.3. Análisis bivariado guiado por la pregunta

Una vez comprendidas las variables, realice comparaciones que respondan directamente a la pregunta de indagación. El objetivo es producir **evidencia gráfica relevante para sustentar una respuesta**. Puede tomar como referencia las sugerencias de visualización del enunciado de la Actividad 2, según corresponda. Incluya los gráficos que considere pertinentes y que ilustren la relaciones entre variables de la pregunta de indagación en el informe. Acompáñelos de un análisis interpretativo

### 2.4. Modelamiento de datos para la visualización

El EDA debe conectarse con la estructura del modelo de datos que se usará para la visualización. **Este modelo debe seguir el esquema de estrella**. Realizar un EDA es un avance en el modelamiento de los datos, ya que desde el código puede construir las tablas de hechos y las tablas de dimensiones y alistarlas para la generación del modelo en Power BI (o cualquier software de visualización)[1]. Como se ha estudiado en clase, la construcción de un modelo de datos para visualización implica:

- Identificar las tablas necesarias (hechos, dimensiones).
- Preparar en Power Query o en código las transformaciones necesarias:
- Limpieza proveniente del EDA
- Normalización: Separación entre tablas de hechos y tablas de dimensiones
- Creación de variables derivadas
- Asignación de llaves: En el caso Power BI es necesario que haya una llave única en una sola variable para poder conectar las tablas entre sí.

Con esto se completa un EDA integral que no solo produce evidencia descriptiva, sino que organiza los datos para ser visualizados de forma efectiva y sustentada.

## Parte 3 – Desarrollo de una visualización sencilla para comunicar los resultados

La visualización final debe servir tanto para explorar como para comunicar. Se trata de seleccionar aquellos gráficos que ayuden a contar una historia orientada a responder la pregunta de indagación. La idea es que alguien que no conoce sus datos pueda entender, con un vistazo, qué encontró y por qué importa. El trabajo desarrollado en el EDA es fundamental para simplificar esta etapa.

### 3.1. Definición de preguntas complementarias y su visualización pertinente

De acuerdo con lo aprendido en el EDA sobre la población, los datos la pregunta de indagación principal y la respuesta, piense en preguntas que le permitan desagregar mensajes de lo encontrado, desde un enfoque de lo general a lo particular. Estas preguntas pueden ser de dos tipos:

[1] Esto quiere decir que usted tiene la opción de procesar todos los datos con código de tal manera que los pueda cargar directamente en Power BI para hacer la visualización, sin depender de Power Query para alguna parte del proceso. Procure hacer estas definiciones de qué herramientas usar en el desarrollo del EDA para evitar reprocesos.

- Preguntas de contexto: que se responden con uno o un par de números y permiten contar la perspectiva o el alcance del análisis. Los objetos visuales asociados a estas preguntas son tarjetas sencillas o múltiples o íconos.
- Preguntas de indagación: que se responden con gráficos de dos ejes, dos o más variables involucradas, siendo al menos una de ellas una variable de hechos. Estos gráficos tienden a ser más complejos, y deben ubicarse más abajo a mayor complejidad.

Elija los gráficos más simples y directos que permitan observar los patrones clave identificados en el EDA. Tal como en la Actividad 2, priorice visualizaciones que respondan a relaciones de la forma E[Y|X] y que muestren contrastes, no adornos.

Dependiendo de su pregunta puede usar:

- Barras (frecuencias o medias)
- Líneas (tendencias por grupos o tramos)
- Boxplots (comparación de distribuciones)
- Dispersión (relación entre variables continuas)
- Mapas (solo si el análisis lo justifica)

## 3.2. Diseño del tablero

El tablero debe cumplir dos funciones:

- Exploración analítica: permitir filtrar, segmentar o navegar por las dimensiones importantes.
- Comunicación clara: presentar una narrativa ordenada que apoye la explicación oral.

Para ello es útil enfocar los objetos visuales (que responden a preguntas de contexto y de indagación) en un orden que guíe la historia:


Para ello evite incluir gráficos redundantes, que se pueden identificar si la narrativa usada con el tablero es repetitiva. También asegúrese de que todos los textos, etiquetas y medidas sean legibles y precisas. Asimismo, mantenga consistencia en colores, escalas y categorías.

## 3.3. Narrativa de la visualización

En este ejercicio puede incluya breves descripciones que acompañen los gráficos o utilice títulos que funcionen como mensajes[2], por ejemplo:

- “La tasa de hacinamiento es un 40% mayor en hogares arrendados que en subarriendo.”
- “Las mujeres jóvenes presentan menor participación laboral en todas las regiones.”

[2] “Insight titles”. No en todos los contextos o ambientes profesionales esto es adecuado, depende del uso común y de la audiencia. Dado que este es un trabajo académico enfocado en investigación de aspectos sociales es muy relevante simplificar los mensajes de los principales hallazgos, de tal forma que se invite a la exploración adicional.

Para el desarrollo de la narrativa asegúrese que se responden las siguientes preguntas: ¿Qué patrones se identificaron?, ¿Qué diferencias o relaciones son las más relevantes?, ¿Qué hallazgo sustantivo responde la pregunta inicial?

### 3.4. Integración con el análisis

Finalmente, asegúrese de que las visualizaciones dialoguen con lo presentado en el EDA:

- Todo gráfico debe corresponder a un patrón previamente identificado.
- No debe aparecer información nueva no explicada antes.
- La interpretación en el informe debe conectar el hallazgo con la pregunta de indagación y el contexto social definido.

Con esto se completa un producto visual coherente, útil y narrativamente consistente, que constituye el cierre del ciclo: pregunta → evidencia → historia.