

Variables de Interés - Investigación sobre Recaudo Tributario

Ciencia de los Datos - Universidad EAFIT

12 de febrero de 2026

Variables de Interés

1. Variables Demográficas

1. **Sexo** (sexo): Categoría binaria (1 = Masculino, 2 = Femenino)
2. **Edad** (edad): Variable continua en años
3. **Parentesco** (parentesco): Relación con el jefe del hogar

2. Variables de Estructura Familiar

1. **Padre** (padre): Presencia del padre en el hogar (1 = Sí, 2 = No, 3 = No aplica)
2. **Madre** (madre): Presencia de la madre en el hogar (1 = Sí, 2 = No, 3 = No aplica)

3. Variables de Participación en la Encuesta

1. **Consentimiento** (consentimiento): Consentimiento informado para participar
2. **Resultado** (resultado): Resultado de la entrevista (1 = Completada, 0 = Incompleta)
3. **Persona Seleccionada** (per_seleccionada): Indicador si fue seleccionado para responder (1 = Sí)

4. Variables de Identificación Administrativa

1. **Directorio** (directorio): Identificador del hogar
2. **Secuencia de Encuesta** (secuencia_encuesta): Número secuencial de la encuesta
3. **Secuencia de Pregunta** (secuencia_p): Número de la pregunta en la encuesta
4. **Orden** (orden): Orden de la persona en el hogar

5. Variables sobre Consumo de Sustancias

1. **Consumo de Marihuana en últimos 12 meses** (C_02): Sí/No
2. **Consumo de Marihuana en últimos 30 días** (C_05): Sí/No
3. **Frecuencia de consumo en 12 meses** (C_03): Una vez, algunas veces, mensual, semanal, diario
4. **Cantidad consumida al mes** (C_08): Gramos de marihuana consumidos
5. **Gasto en marihuana últimos 30 días** (C_09_VALOR): Valor monetario en pesos
6. **Precio por gramo** (C_09_VALOR): Conocimiento del precio unitario
7. **Formas de consumo** (C_04_A a C_04_D): Fumada, inhalada, inyectada, etc.
8. **Cómo fue obtenida** (C_10_A a C_10_H): Internet, redes sociales, expendios, amigos, domicilio

6. Variables sobre Empleo y Trabajo

1. **Accidentes laborales** (R_01): Incidentes en últimos 12 meses
2. **Consumo durante accidente** (R_02): Sustancias consumidas previas al incidente
3. **Días faltados** (R_03): Inasistencia en últimos 30 días
4. **Consumo en horario laboral** (R_08_A a R_08_D): Cigarrillo, alcohol, marihuana, cocaína
5. **Rendimiento laboral** (R_09): Impacto del consumo en desempeño

7. Variables sobre Educación, Orientación Sexual e Identidad

1. **Nivel educativo** (D2_05): Ninguno, preescolar, primaria, secundaria, media, técnica, universitaria, postgrado
2. **Orientación sexual** (D2_06): Heterosexual, gay/lesbiana, bisexual, otra
3. **Identidad de género** (D2_07): Masculino, femenino, transgénero, otro

Variables Adicionales a Descargar/Construir

Para completar el análisis sobre recaudo tributario y legalización de marihuana, se requieren:

1. **Ingresos del hogar:** Datos de DANE sobre ingresos por decil de población
2. **Estrato socioeconómico:** Variable disponible en datos (estrato)

3. **Tasas impositivas actuales:** Estructura tributaria de Colombia (DIAN)
4. **Escenarios de legalización:** Parámetros de regulación, impuestos y distribución
5. **PIB y variables macroeconómicas:** Contexto económico nacional
6. **Datos de población:** Proyecciones demográficas del DANE
7. **Precios de mercado ilegal:** Comparación con precios bajo legalización
8. **Tasas de tributación potencial:** Escenarios de impuestos en legales

Notas Metodológicas

- Las variables actualmente disponibles en `personas_processed.csv` provienen de datos de hogares encuestados
- Se deben integrar con datos de fuentes externas (DIAN, DANE) para el análisis tributario
- La limpieza y normalización se realiza mediante el script `01_processing.py`
- Los metadatos de procesamiento se encuentran en `personas_processed_metadata.json`