Universidad de los Andes

Exploración y visualización de datos para lo socioeconómico

Miguel Andrés Garzón

Proyecto 2 - Visualización y construcción de historias con datos [1]

ELABORADO POR:
Laura Sarif Rivera Sanabria
María Camila Caraballo Candela
Javier Antonio Amaya Nieto

[1] https://github.com/camto-24/Proyecto2_LauraRivera_JavierAmaya_CamilaCaraballo

1 Parte 1: Planteamiento de la pregunta de indagación principal

Pregunta de interés: ¿Existe una asociación positiva entre el consumo de bebidas azucaradas y la probabilidad de presentar un diagnóstico de enfermedad crónica en los últimos 12 meses?

La preocupación por los efectos del consumo de bebidas azucaradas en la salud pública se sustenta en una amplia literatura epidemiológica que evidencia su asociación con un mayor riesgo de síndrome metabólico, aumento de peso, caries dentales, diabetes tipo 2 y enfermedades cardiovasculares como la hipertensión crónica (Singh, Khartibzadeh , Shi, Lim, & Andrews , 2015). Estas evidencias han llevado a organismos internacionales a establecer recomendaciones más estrictas, entre ellas las de la Organización Mundial de la Salud, que sugiere limitar la ingesta de azúcares añadidos a menos del 10% del aporte calórico diario, equivalente a cerca de 200 kilocalorías en una dieta estándar de 2.000 kcal (Organización Mundial de la Salud, 2015).

La evidencia global también muestra que el consumo de bebidas azucaradas tiene un impacto sustancial en la aparición de enfermedades crónicas no transmisibles. Un estudio de la Escuela de Ciencia y Políticas de Nutrición de la Universidad de Tufts, publicado en *Nature Medicine*, estima que la ingesta regular de estas bebidas se asocia cada año con millones de nuevos casos de diabetes tipo 2 y con un número comparable de enfermedades cardiovasculares. En el caso de Colombia, el mismo estudio señala que hasta el 48% de los casos incidentes de diabetes podrían estar vinculados al consumo de bebidas azucaradas (Lara-Castor, O'Hearn, & Cudhea, 2025).

En Colombia, la introducción en 2023 de un impuesto a las bebidas azucaradas representó un avance relevante dentro de las estrategias para reducir la carga de enfermedades crónicas y desincentivar su consumo (Ministerio de Salud y Protección Social, 2023). No obstante, desde una perspectiva de política pública, la adopción de este tipo de medidas no garantiza por sí misma una disminución de consumo uniforme en todos los grupos de la población, ni permite identificar con claridad cuáles segmentos presentan niveles más altos de exposición. Esto resalta la importancia de complementar estas medidas con análisis que permitan identificar hábitos, evaluar comportamientos diferenciados y orientar estrategias integrales que combinen regulación, educación y acceso a alternativas saludables.

En este sentido, el análisis exploratorio que proponemos busca identificar con base en la Encuesta de Calidad de Vida 2023, si existe una relación entre la frecuencia de consumo de bebidas azucaradas y la prevalencia de enfermedades crónicas en Colombia. Dado que la encuesta no permite establecer causalidad ni reconstruir trayectorias individuales de salud, el aporte del análisis es describir patrones generales en la población adulta y reconocer grupos donde la coincidencia entre consumo y enfermedad es más marcada. Para ello, se utilizan como variables principales la presencia de enfermedad crónica y la frecuencia semanal de consumo, acompañadas de covariables como género, edad, escolaridad y clasificación detallada de los niveles de consumo, lo que permite controlar diferencias individuales y aislar mejor las variaciones en los patrones observados.

La utilidad de este ejercicio radica en su capacidad para complementar el impuesto a las bebidas azucaradas vigente en el país, proporcionando información que puede guiar intervenciones de salud pública. Los resultados descriptivos ayudan a identificar grupos particularmente expuestos, orientar campañas de prevención y ubicar territorios donde persisten niveles altos de consumo junto con elevada prevalencia de enfermedad. Finalmente, este análisis busca aportar evidencia que fortalezca la toma de

decisiones y apoye el diseño de estrategias adicionales para mitigar los riesgos asociados al consumo habitual de bebidas azucaradas.

2 Parte 2: EDA y modelamiento de datos

2.1 Preparación y comprensión inicial de los datos

En la sección 2.1.1 del código, se importaron cinco bases de la ECV correspondientes a salud, composición del hogar, educación, servicios del hogar y vivienda. Esta etapa permitió verificar su estructura inicial y asegurar la correcta lectura de variables, etiquetas y formatos antes de consolidarlas. En la sección 2.1.2 se revisó la estructura de la base integrada df_final (240.212 observaciones y 14 variables), verificando tipos y codificación. Se confirmó que año, ingreso y fex son numéricas; sexo, crónica, consumo de bebidas azucaradas, departamento, municipio, se identificación como categóricas; nivel educativo y frecuencia de consumo son ordinales. Se identificó y corrigió la codificación de categorías según la estructura de la ECV, resultando 5 variables categóricas, 6 factores y 3 numéricas. El factor de expansión mostró alta dispersión (1,41 a 7.353,72), con media y mediana alejadas.

En la sección 2.1.3 se construyó df_analisis, filtrando a personas de 18 años o más. Se documentaron faltantes en nivel educativo para menores y se decidió conservar esas categorías en los análisis, aunque la base excluye población infantil por coherencia epidemiológica y sociodemográfica. Se revisaron faltantes: consume_azucar está completa; frecuencia_azucar presenta NA no por ausencia de datos, sino porque la pregunta no aplica a quienes no consumen; cronica_12m, sexo y parentesco están completas. Adicionalmente, en la sección 2.1.4 se identificaron valores atípicos en edad, ingreso per cápita y fex mediante cuartiles y RIC. Las edades y los ingresos se mantienen en rangos consistenes; y aunque fex tiene valores extremos, estos provienen del diseño muestral del cálculo del DANE y no deben modificarse.

Finalmente, en la sección 2.1.5 se documentaron las recodificaciones: creación de un indicador binario para enfermedad crónica, una variable numérica para no consumo y categorías agrupadas y ordenadas de frecuencia de consumo. Todas las transformaciones se registraron como parte del proceso de limpieza y preparación para los análisis posteriores.

2.2 Análisis univariado sistemático

La Figura 1 inicia el análisis univariado examinando la distribución de la variable dependiente, correspondiente a la presencia de enfermedad crónica en los últimos 12 meses, mediante tablas de frecuencia y proporciones ponderadas y no ponderadas. Las diferencias entre ambas son mínimas (15.8% frente a 15.2%), lo que confirma una adecuada representatividad de la muestra y valida el uso del factor de expansión en la interpretación de resultados. Con ponderación, la prevalencia estimada es 15.2%, mientras que el 84.8% de la población no reporta diagnóstico, proporcionando una primera aproximación clara y confiable al comportamiento de la variable principal.

Posteriormente, se realizó el análisis univariado de las variables de consumo agrupado, consumo detallado, grupo de edad y edad. Aunque en el código se documentan todas, en este apartado se describen únicamente los resultados de edad y consumo detallado. La Figura 2 muestra las distribuciones de densidad ponderadas y no ponderadas de la edad, evidenciando que el factor de expansión corrige ligeras desviaciones en la representación de algunos grupos etarios. Esto es fundamental, ya que tanto el consumo

de bebidas azucaradas como la probabilidad de presentar enfermedades crónicas dependen fuertemente de la edad; por ello, su revisión univariada permite verificar que no existan sesgos de representación que distorsionen análisis posteriores.

Finalmente, la Figura 3 presenta la distribución de la variable consume_azucar y sus categorías detalladas (consumo_detallado) mediante gráficos de barras ponderados. Los resultados muestran que el 36.8% de la población no consume bebidas azucaradas, mientras que los niveles bajos de consumo predominan (16.2% una vez por semana y 18.6% entre 2 y 3 veces por semana). El consumo diario o más frecuente alcanza aproximadamente el 12%. Las diferencias entre frecuencias ponderadas y no ponderadas son marginales, lo que reafirma la pertinencia del uso del factor de expansión para obtener estimaciones representativas. Esta revisión univariada establece la base conceptual necesaria antes de cruzar variables y evita interpretaciones sesgadas en etapas posteriores del análisis.

2.3 Análisis bivariado guiado por la pregunta

La Figura 4 muestra un análisis descriptivo de la relación simple entre el consumo de bebidas azucaradas y la prevalencia de enfermedades crónicas a nivel nacional. A primera vista, la tendencia parece contradictoria, pues la prevalencia de enfermedad alcanza el 24% entre quienes reportan 'No consumo' de este tipo de bebidas, mientras que disminuye a 8.6% entre quienes las consumen diariamente. Este patrón no debe interpretarse como evidencia de que consumir azúcar reduce el riesgo de enfermedad; en realidad, refleja un problema de causalidad inversa conocido como sick-quitter. Este fenómeno ocurre cuando las personas diagnosticadas con condiciones como diabetes o hipertensión modifican su dieta y reducen o eliminan el consumo de bebidas azucaradas. Como consecuencia, la población con mayor carga de enfermedad se concentra en la categoría de "no consumidores", lo que genera una relación negativa aparente y distorsiona la interpretación del riesgo real asociado a su consumo (Wannamethee, Shaper, & Whincup, 2006).

La Figura 5 incorpora una estratificación por grupos de edad que permite observar diferencias importantes en la relación entre consumo de bebidas azucaradas y prevalencia de enfermedad crónica. Al desagregar por grupos de 10 años, la distorsión del análisis agregado se reduce parcialmente. Aunque el efecto sick-quitter continúa presente en la categoría de "No consumo", el comportamiento dentro de los grupos que sí consumen muestra patrones diferenciados. Entre las personas en edad económicamente activa (18 a 59 años) aparece una tendencia no lineal con forma de "L", lo que indica que la relación no es uniforme a lo largo de los niveles de consumo. Esta relación muestra que la prevalencia de enfermedad crónica es mayor entre quienes consumen bebidas azucaradas diariamente en comparación con quienes lo hacen con menor frecuencia. Esto evidencia una asociación positiva entre la intensidad del consumo y el riesgo de presentar una enfermedad crónica, relación que se manifiesta principalmente en el margen intensivo del consumo. En otras palabras, a medida que aumenta la frecuencia de ingesta del consumo diario, también aumenta la probabilidad de reporte de enfermedad crónica.

Por último, la Figura 6 muestra una dinámica de sesgo de supervivencia diferenciada por género, lo cual se evidencia en las tendencias de prevalencia entre hombres y mujeres. En edades tempranas, ambos grupos presentan patrones similares, pero estas trayectorias comienzan a separarse a medida que aumenta la edad. Las mujeres reportan tasas más altas de enfermedad crónica, lo que probablemente se relaciona con una mayor frecuencia de diagnóstico y uso de servicios de salud, contrario al subregistro masculino.

Esto se explica porque los hombres tienden a demorar la búsqueda de atención médica, subestiman síntomas y presentan menores niveles de adherencia a controles preventivos (Courtenay, 2000)

En etapas más avanzadas de la vida se observa una atenuación diferencial entre los grupos. Mientras las mujeres mantienen un patrón de prevalencia con forma de "L" incluso en edades muy avanzadas, los hombres pierden esa tendencia después de los 60 años. Esta divergencia sugiere la presencia de mortalidad prematura selectiva entre los hombres con mayor consumo y mayor predisposición a enfermar, lo que deja en la población de adultos mayores a un subgrupo de sobrevivientes más resilientes. Este proceso sesga la asociación hacia la nulidad, un fenómeno menos pronunciado en las mujeres, quienes presentan una mayor esperanza de vida acompañada de mayor morbilidad.

En síntesis, la evidencia descriptiva sugiere la existencia de una asociación positiva y estructural entre la frecuencia de consumo y la enfermedad crónica, aunque esta relación depende de la edad, el sexo y del nivel de consumo. Sin embargo, como se advirtió desde el inicio, este enfoque descriptivo es limitado debido a la influencia de la causalidad inversa y de los sesgos de selección y supervivencia. Por lo tanto, resulta fundamental avanzar hacia modelos multivariados[2] que permitan estimar un riesgo de morbilidad más preciso, controlando por factores demográficos relevantes y ajustando la categoría de referencia para reducir la endogeneidad generada por los cambios de comportamiento posteriores al diagnóstico.

2.4 Modelamiento de datos para la visualización

En la sección 2.4 del código, se procedió a la estructuración de un modelo de datos relacional (esquema tipo estrella) optimizado para posterior importación de las tablas a la herramienta de Business Intelligence (Power BI). El proceso inició con la generación de llaves únicas (*id_hogar*, *id_persona*, *id_geografia*) para garantizar la integridad referencial entre las tablas. Posteriormente, se normalizó la información dividiéndola en cuatro tablas de dimensiones (*dim_geografia*, *dim_hogar*, *dim_consumo y dim_demografia*) que contienen los atributos cualitativos y de contexto, y una tabla de hechos central (*fact_personas*). Esta última consolida las métricas clave, las variables de desenlace (*cronica_num*) y el factor de expansión (*fex*), manteniendo las llaves necesarias para relacionarse con las dimensiones. Finalmente, el modelo completo fue exportado a un archivo de Excel con múltiples hojas, lo que facilita la creación directa de las relaciones en el entorno de Power BI.

Una vez consolidado el archivo *Modelo_Datos_Completo.xlsx*, se creó el archivo "*Modelo de datos BI.pbix*" dentro de la carpeta del proyecto, con el propósito de mantener una organización coherente y asegurar una referencia directa a la fuente de información. Posteriormente, se procedió a importar los datos y a seleccionar únicamente las hojas correspondientes a las tablas relevantes para el proceso de modelado. Una vez integrada esta información, se utilizó la opción de transformar datos para definir de manera personalizada las relaciones entre las tablas. Dichas relaciones se establecieron bajo una cardinalidad de uno a muchos, desde las tablas de dimensión hacia la tabla de hechos, lo que permitió conformar un modelo en estrella correctamente estructurado a partir de las llaves identificadoras.

---

[2] En la sección 2.3.6 del código se presentan los resultados de un modelo multivariado mediante regresión logística, el cual aborda esta pregunta desde una perspectiva estrictamente académica. Sin embargo, dicho análisis excede el alcance previsto para este taller, cuyo objetivo principal es responder la pregunta planteada de manera descriptiva.

Figuras

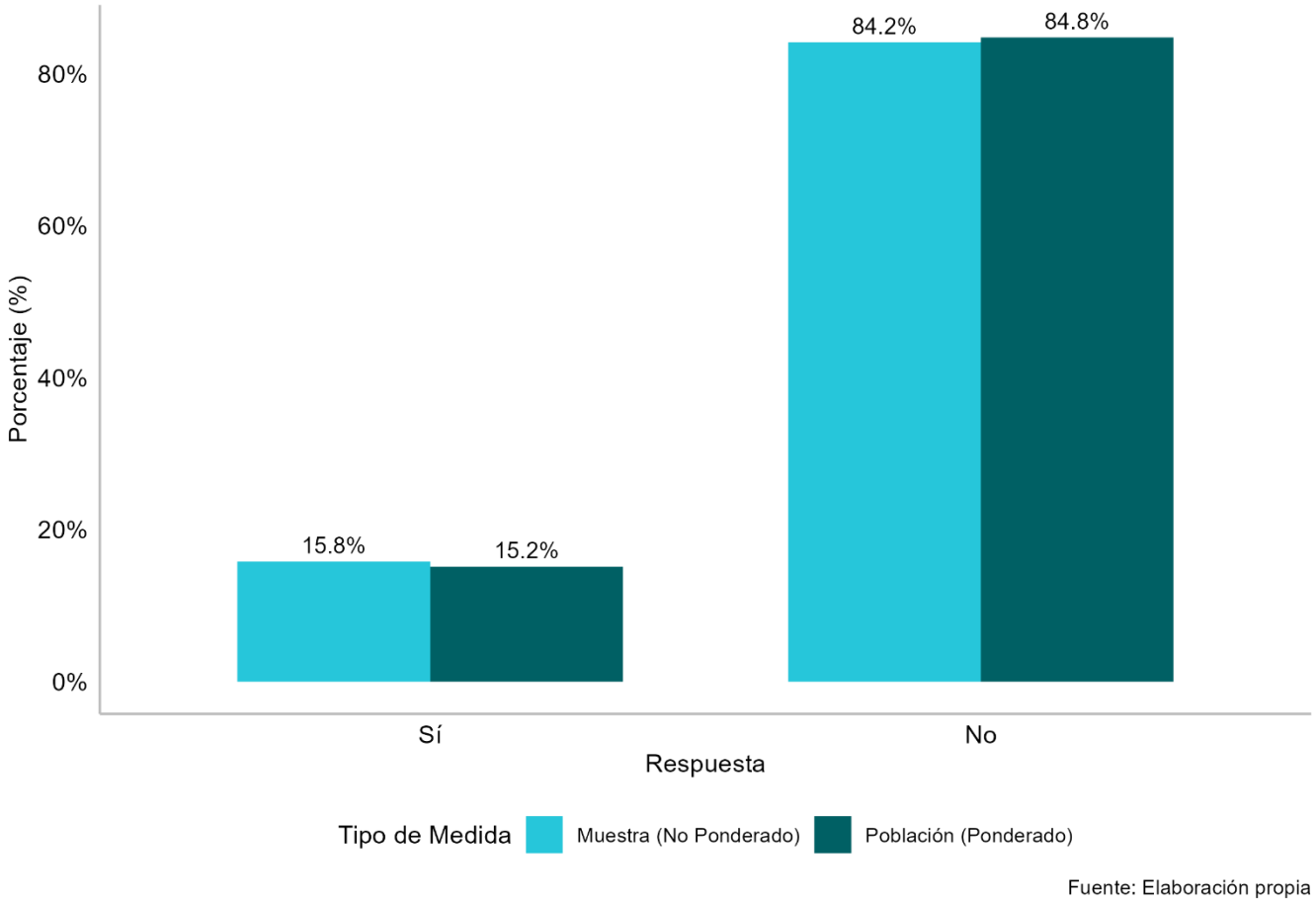

Figura 1. Participación de personas con prevalencia de enfermedad crónica por tipo de muestra. Fuente: Encuesta de Calidad de vida. Elaboración propia.

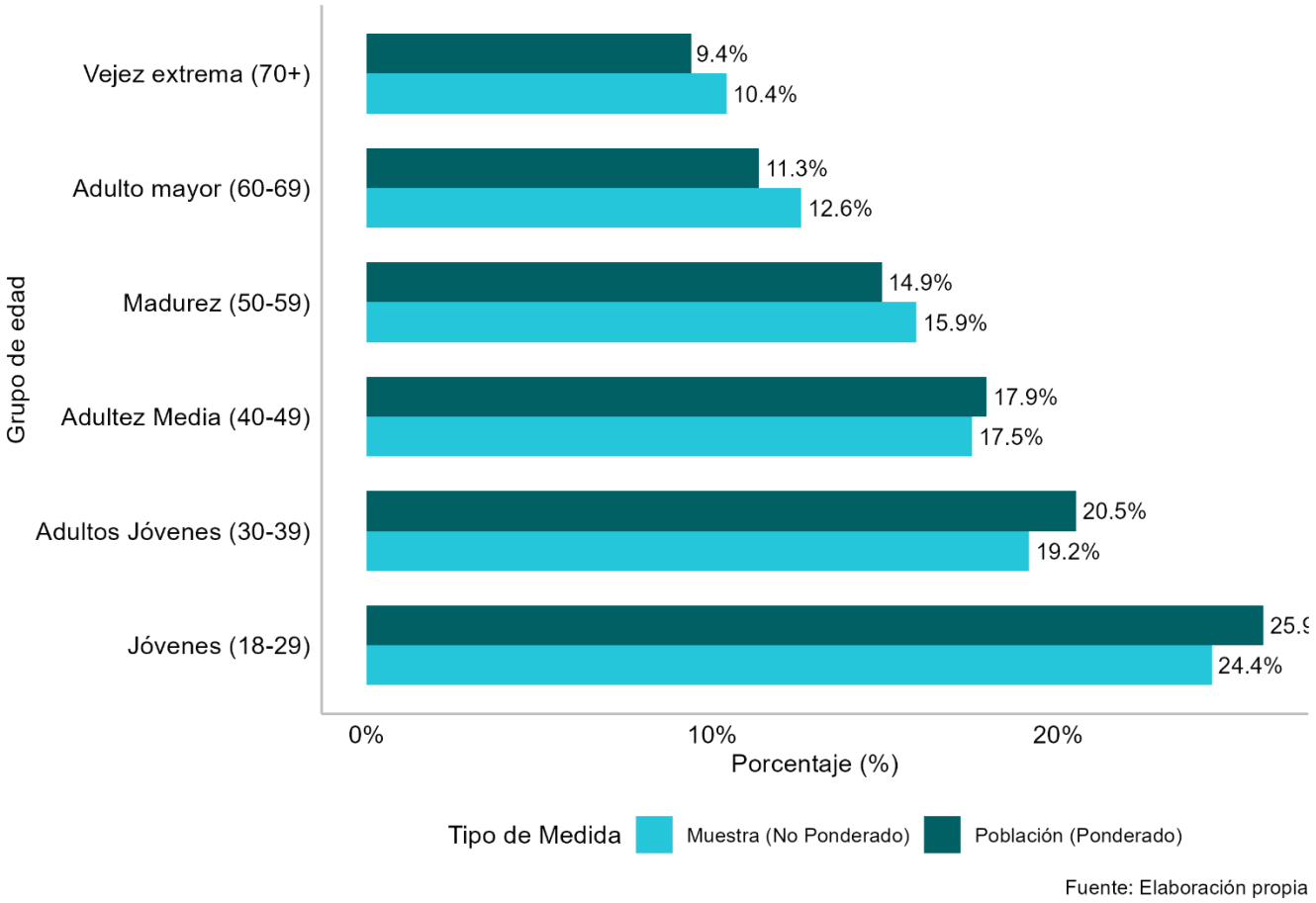

Figura 2. Participación de grupos etarios por tipo de muestra. Fuente: Encuesta de Calidad de vida. Elaboración propia.

Tipo de estimación	N (Observaciones)	Media	D.E.	Mediana	RIQ	Mín.	Máx.
No ponderado	172.538	44,86	17,78	43	28	18	104
Ponderado	37.892.981	43,75	17,47	41	28	18	104

Tabla 1. Estadísticas descriptivas de la variable edad. Análisis univariado. Ponderación realizada usando el factor de expansión; DE: desviación estándar; RIQ: rango intercuartílico.

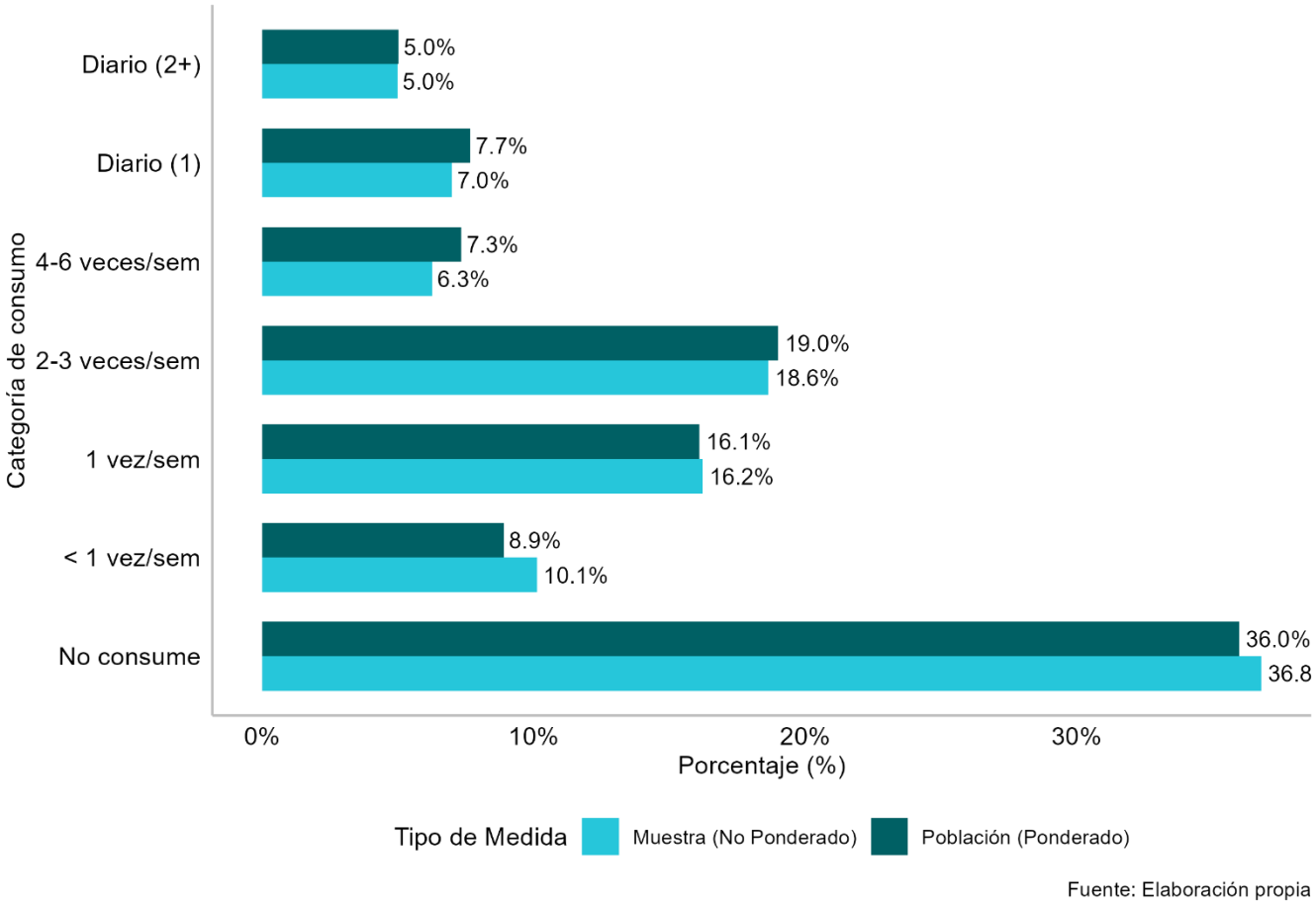

Figura 3. Participación de frecuencia de consumo detallado por tipo de muestra. Fuente: Encuesta de Calidad de vida. Elaboración propia.

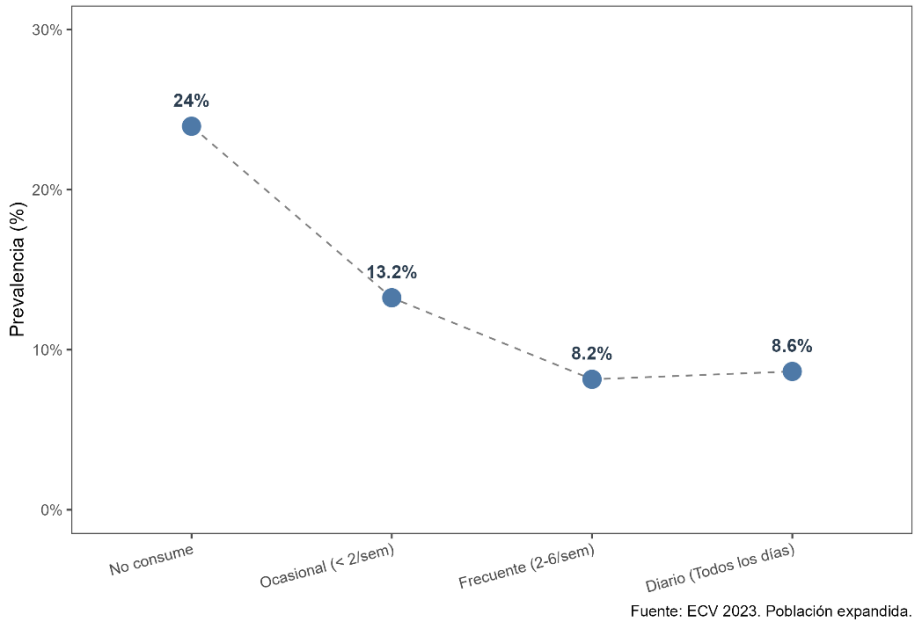

Figura 4. Relación entre la prevalencia de enfermedad crónica y la intensidad del consumo. Fuente: Encuesta de Calidad de vida. Elaboración propia.

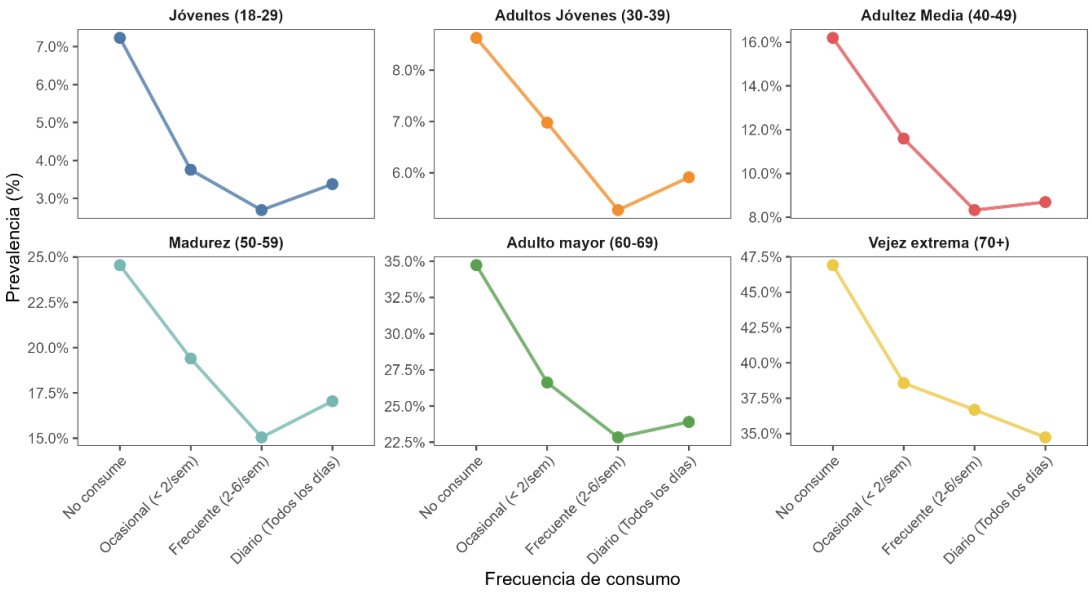

Figura 5. Relación entre la prevalencia de enfermedad crónica y la intensidad del consumo por grupo etario. Fuente: Encuesta de Calidad de vida. Elaboración propia.

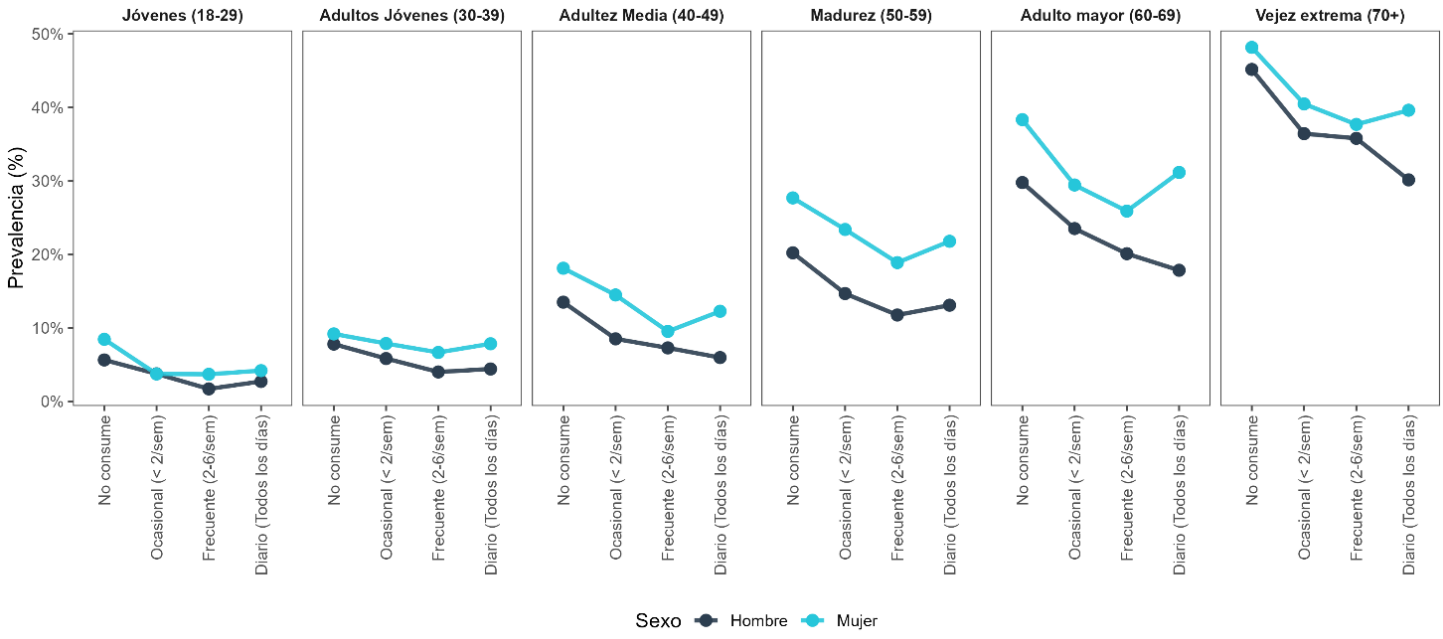

Figura 6. Tendencias de la prevalencia de enfermedad crónica y la intensidad del consumo por grupo etario y sexo. Fuente: Encuesta de Calidad de vida. Elaboración propia.

3 Referencias

I. Courtenay, W. H. (2000). Constructions of masculinity and their influence on men's well-being: a theory of gender and health. Social Science & Medicine, 50(10), 1385–1401.

II. Lara-Castor, L., O'Hearn, M., Cudhea, F., Wang, M., Wilde, P., Rehm, C. D., et al. (2023). Burdens of type 2 diabetes and cardiovascular disease attributable to sugar-sweetened beverages in 184 countries. Nature Medicine.

III. Ministerio de Salud y Protección Social. (2023). Impuesto saludable: Impuesto a las bebidas ultraprocesadas azucaradas y a los alimentos ultraprocesados. Ministerio de Salud y Protección Social de Colombia. https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/PP/impueso-saludable-bebidas-alimentos-ultraprocesados.pdf

IV. Organización Mundial de la Salud. (2015). Nota informativa sobre la ingesta de azúcares recomendada en la directriz de la OMS para adultos y niños. OMS.

V. Singh, G. M., Micha, R., Khatibzadeh, S., Shi, P., Lim, S., Andrews, K. G., et al. (2015). Global, regional, and national consumption of sugar-sweetened beverages, fruit juices, and milk: A systematic assessment of beverage intake in 187 countries. PLOS ONE, 10(8), e0124845.

VI. Wannamethee, S. G., Shaper, G. H., & Whincup, P. H. (2006). Alcohol and sudden cardiac death: Evidence from the British Regional Heart Study. Heart, 92(12), 1839–1845.