



Instituto Tecnológico y de Estudios Superiores de Monterrey

Maestría en Inteligencia Artificial Aplicada

Proyecto Integrador

TC5035.10

Semana 3

Avance 1. Análisis Exploratorio de Datos

“Generalización de modelos nacionales de pronóstico epidemiológico hacia un enfoque modular con desagregación por sexo y entidad federativa en México”

Equipo No 1:

Programa	Matrícula	Estudiante
MNA	A01795838	Javier Augusto Rebull Saucedo
MNA	A01795941	Juan Carlos Pérez Nava
MNA	A01232963	Luis Gerardo Sánchez Salazar

Equipo docente:

Dra. Grettel Barceló Alonso – Profesora Titular y Asesora del Proyecto

Dr. Luis Eduardo Falcón Morales – Director Nacional de MNA

Mtra. Verónica Sandra Guzmán de Valle – Profesora Asistente

Patrocinador:

Dra. Ruth Pérez-Hernández – Líder del Proyecto IMSS

Dra. Lina Díaz Castro - Investigadora en Psiquiatría

Febrero 01, 2026

Tabla de Contenidos

ANÁLISIS DE LA ESTRUCTURA DE DATOS	4
Año.....	9
Semana	10
Casos semana y acumulados.....	10
ANÁLISIS BIVARIANTE	10
EVALUACIÓN DE LA VARIABILIDAD.....	12
<i>Puntos clave:</i>	13
LIMPIEZA DE DATOS	16
<i>Filtrado por padecimiento</i>	16
PREPARACIÓN DE LOS DATOS.....	17
<i>Ajuste de semanas epidemiológicas</i>	17
PREPARACIÓN DE LAS SERIES DE TIEMPO	17
AJUSTE DE VALORES NEGATIVOS	18
TRATAMIENTO DE VALORES ATÍPICOS	19
RESULTADO DE LOS DATOS TRAS LIMPIEZA Y CORRECCIÓN	20
ALZHEIMER.....	22
PARKINSON.....	22
DEPRESIÓN	23
EVALUACIÓN DE LA VARIABILIDAD.....	25
SERIES DE TIEMPO	26
<i>Sin limpieza de datos</i>	27
<i>Con tratamiento de limpieza aplicado</i>	28
<i>Con limpieza e imputación de atípicos (IQR)</i>	29
SÍNTESIS DEL ANÁLISIS.....	30
PIPELINE DE DATOS.....	30
RESULTADOS RELEVANTES	30
CONCLUSIONES Y EXPECTATIVAS	31
REFERENCIAS.....	33
ANEXO: DATOS GEOGRÁFICOS	34
OBJETIVO DEL MÓDULO INEGI	34
FUENTES DE DATOS.....	34
PIPELINE DE PROCESAMIENTO	35
VARIABLES DERIVADAS Y CATEGORIZACIONES	36
REGIONALIZACIÓN EN SALUD MENTAL	36
EXPLORATORY DATA ANALYSIS (EDA) BREVE.....	38
HIGHLIGHTS DEL ANÁLISIS INICIAL	41
CLASIFICACIÓN TERRITORIAL SUGERIDA	41
ANEXO: PIPELINE DE PREPROCESAMIENTO DE DATOS	42
ANEXO: REPOSITORIO GITHUB DEL PROYECTO.....	43
ANEXO: PAGINA WEB DEL PROYECTO	43

Índice de Figuras

Imagen 1 - Estructura de la información fuente	4
Imagen 2 - Estructura del DataFrame	4
Imagen 3 - Verificación de valores únicos	5
Imagen 4 - Verificación de valores nulos.....	6
Imagen 5 - Análisis del patrón de ausencia	6
Imagen 6 - Distribución porcentual de "Entidad"	7
Imagen 7 - Distribución porcentual de "Padecimiento".....	8
Imagen 8 - Histograma de variables numéricas.....	9
Imagen 9 - Matriz de correlación.....	11
Imagen 10 - Distribución de casos por semana de Alzheimer (hombres).....	12
Imagen 11 - Distribución de casos por semana de Alzheimer (mujeres)	13
Imagen 12 - Distribución de casos por semana de Parkinson (hombres)	14
Imagen 13 - Distribución de casos por semana de Parkinson (mujeres)	14
Imagen 14 - Distribución de casos por semana de Depresión (hombres)	15
Imagen 15 - Distribución de casos por semana de Depresión (mujeres)	15
Imagen 16 - Detalles del DataFrame después de limpieza	20
Imagen 17 - Distribución de valores únicos por padecimiento	21
Imagen 18 - Distribución de las variables de incrementos por padecimiento después de limpieza	21
Imagen 19 - Distribución transformada de la variable de incrementos para Alzheimer dividida por sexo	22
Imagen 20 - Distribución transformada de la variable de incrementos para Parkinson dividida por sexo	23
Imagen 21 - Distribución transformada de la variable de incrementos para Depresión dividida por sexo.....	23
Imagen 22 - Distribución de porcentajes por estado después de la limpieza.....	24
Imagen 23 - Matriz de correlación para cada uno de los padecimientos a analizar.....	25
Imagen 24 - Distribución semanal de casos después de la limpieza de datos, por padecimiento y sexo.....	26
Imagen 25 - Casos semanales nacionales para los tres padecimientos, sin limpieza de datos.....	27
Imagen 26 - Casos semanales nacionales para los tres padecimientos, con tratamiento de limpieza	28
Imagen 27 - Casos semanales nacionales para los tres padecimientos, con limpieza e imputación de atípicos (IQR).....	29
Imagen 28 - Extensión territorial de México en kilómetros cuadrados	34
Imagen 29 - Vista exploratoria numérica del DataFrame obtenido	39
Imagen 30 - Valores para conteo de variables categóricas del DataFrame	39
Imagen 31 - Gráficas de barras categorizadas	40
Imagen 32 - Distribución de las variables del DataFrame por Boxplot	40
Imagen 33. Pipeline de Transformación de Datos.	42

Análisis de la estructura de datos

Como primer paso, se llevó a cabo un análisis detallado de la estructura de los archivos fuente, los cuales se presentan en formato PDF no estructurado. Cada documento contiene el registro semanal de casos acumulados, y es necesario localizar el apartado titulado “Cuadro17. Casos por entidad federativa de Enfermedades Neurológicas y de Salud Mental hasta la semana epidemiológica...”. En esta sección se agrupan tres condiciones médicas específicas conforme a la clasificación CIE-10: Depresión (F32), Enfermedad de Parkinson (G20) y Enfermedad de Alzheimer (G30).

La información está organizada por Entidad Federativa, lo que permite una segmentación geográfica. Para cada condición, se reportan cifras acumuladas desglosadas por sexo (Hombres y Mujeres), diferenciando entre datos semanales del año en curso y acumulados del año anterior.

ENTIDAD FEDERATIVA	Depresión CIE-10 ^a REV. F32			Enfermedad de Parkinson CIE-10 ^a REV. G20			Enfermedad de Alzheimer CIE-10 ^a REV. G30		
	2025		2024	2025		2024	2025		2024
	Sem.	Acum.	Acum.	Sem.	Acum.	Acum.	Sem.	Acum.	Acum.
		H	M		H	M		H	M

Imagen 1 - Estructura de la información fuente

Una vez extraída la información desde los archivos PDF, se procedió a cargarla en un DataFrame con una estructura equivalente a la presentada en la fuente original.

```
RangeIndex: 59712 entries, 0 to 59711
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype  
 --- 
 0   Anio              59712 non-null   int64  
 1   Semana            59712 non-null   int64  
 2   Entidad            59712 non-null   object  
 3   Padecimiento      59712 non-null   object  
 4   Casos_semana      59709 non-null   float64 
 5   Acumulado_hombres 59712 non-null   int64  
 6   Acumulado_mujeres 59712 non-null   int64  
 7   Acumulado_anio_anterior 54624 non-null   float64 
dtypes: float64(2), int64(4), object(2)
memory usage: 3.6+ MB
```

Resumen del DataFrame	Valor
Fecha	2026-01-25 22:56
Filas	59,712
Columnas	8
Columnas numéricas	6
Columnas categóricas	2
Otras columnas	0
Porcentaje nulos	1.07%

Imagen 2 - Estructura del DataFrame

La estructura del DataFrame está conformada por ocho columnas: seis de tipo numérico y dos categóricas. Entre los aspectos más relevantes se destacan los siguientes:

- **Datos temporales:** No se cuenta con una fecha específica, sino con dos campos que indican el año y el número de semana epidemiológica, lo que permite una segmentación temporal.

- **Entidad federativa:** La información geográfica se presenta de forma descriptiva, indicando el nombre completo de cada estado que conforma la República Mexicana.
- **Padecimiento:** Dado que se extrajeron registros correspondientes a tres condiciones médicas (Depresión, Parkinson y Alzheimer), se incluye una columna que identifica el padecimiento asociado a cada registro.
- **Cifras reportadas:** Se incluyen los casos registrados en la semana correspondiente, los acumulados por sexo (hombres y mujeres), así como los acumulados de la misma semana en el mismo periodo del año anterior.

El conjunto de datos cuenta con **59,712 registros**, lo que representa una base amplia y representativa para el análisis. El nivel de completitud es elevado, ya que únicamente el **1.07% de los valores corresponden a datos nulos** en todo el DataFrame.

Valores únicos por columna		
Acumulado_anio_anterior	5171	float64
Acumulado_mujeres	4545	int64
Acumulado_hombres	2393	int64
Casos_semana	507	float64
Semana	53	int64
Entidad	33	object
Anio	13	int64
Padecimiento	3	object

Imagen 3 - Verificación de valores únicos

Otra de las características validadas del conjunto de datos fue la cantidad de valores únicos presentes en cada variable. A continuación, se destacan algunos aspectos clave:

- **Diversidad de valores:** El campo *Acumulado_anio_anterior* presenta una alta variabilidad, con **5,123 valores únicos**, mientras que *Casos_semana* cuenta con **586 valores únicos**. Esto refleja una dinámica semanal significativa en la evolución de los padecimientos registrados.
- **Variables categóricas:** La columna *Entidad* incluye **33 valores únicos**, lo cual resulta llamativo, ya que México cuenta oficialmente con **32 entidades federativas**. Esto sugiere la existencia de una categoría adicional que debe ser revisada y depurada.
- **Temporalidad:** La columna *Semana* abarca **53 valores únicos**, lo que indica una cobertura completa del año epidemiológico. Sin embargo, dado que un año calendario generalmente tiene 52 semanas, este comportamiento podría deberse a años con una semana epidemiológica, o a errores de codificación que deben ser verificados. Por otro lado, la columna *Anio* contiene **13 valores únicos**, lo que corresponde a información que cubre 13 años.

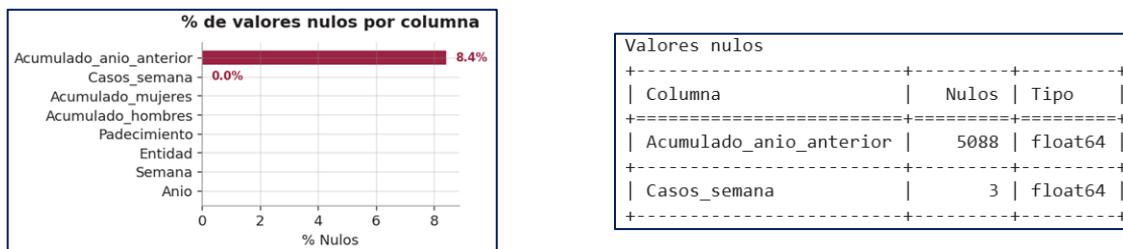


Imagen 4 - Verificación de valores nulos

Como parte del proceso de validación, se identificaron las columnas que presentan valores nulos. En total, se detectaron dos variables con datos faltantes:

- **Acumulado_anio_anterior:** presenta **5,088 valores nulos**, lo que representa el principal foco de incompletitud en el conjunto.

Nulos en 'Acumulado_anio_anterior' por año:													
Anio	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026
Nulos	4992	96	0	0	0	0	0	0	0	0	0	0	0

Imagen 5 - Análisis del patrón de ausencia

- **Casos_semana:** contiene únicamente **3 valores nulos**, lo cual es marginal y no representa un riesgo significativo para el análisis semanal

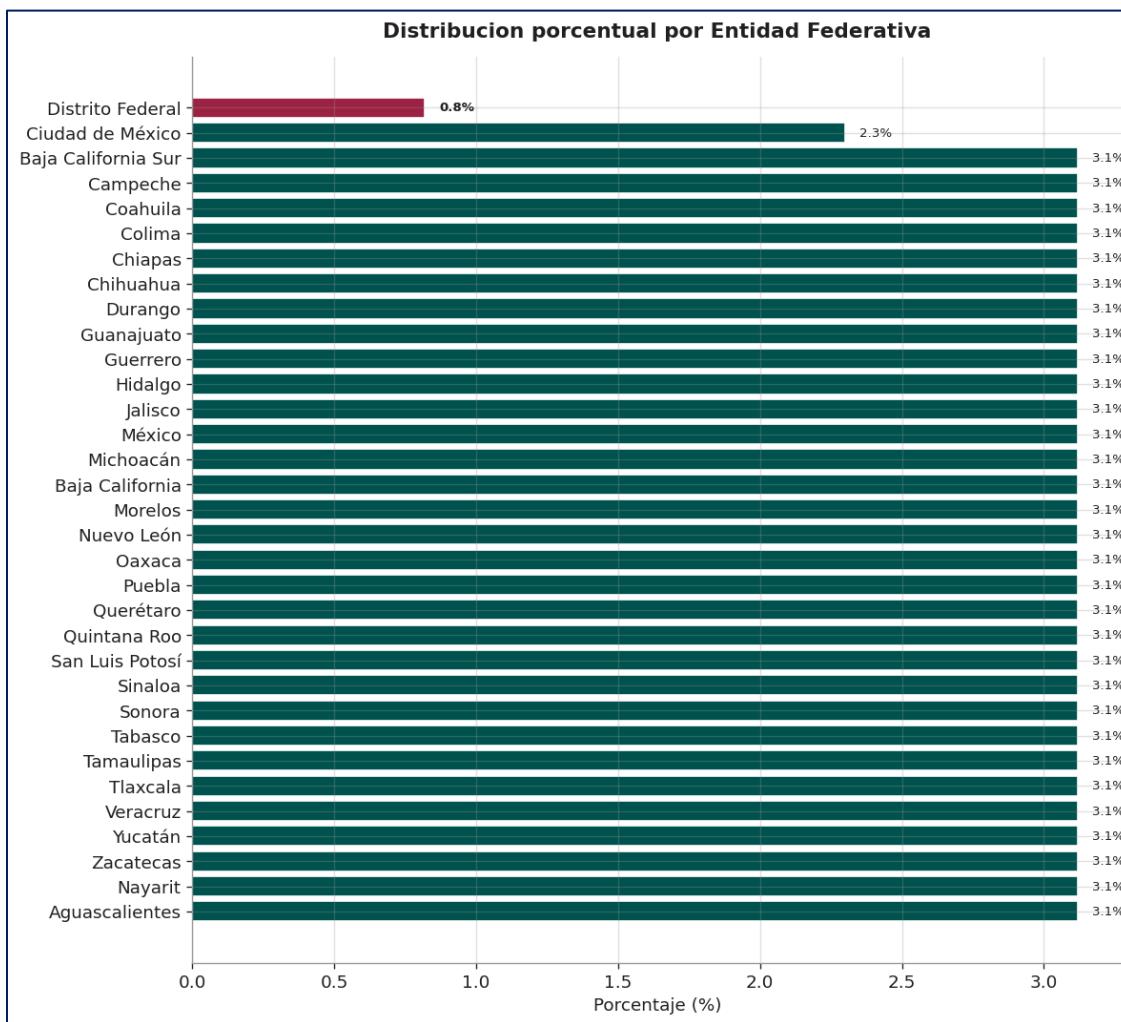


Imagen 6 - Distribución porcentual de "Entidad"

La variable **Entidad** corresponde a los estados y entidades federativas de México. Al tratarse de una variable categórica, resulta necesario revisar cómo se distribuyen porcentualmente sus categorías dentro del conjunto de datos. Esta validación permite confirmar la coherencia de la información y detectar posibles inconsistencias que puedan afectar la interpretación estadística.

Al analizar la gráfica de distribución, se observa que la mayoría de las entidades presentan un porcentaje constante de 3.1%, lo que refleja una distribución homogénea y sugiere que los datos fueron construidos bajo un criterio de equilibrio.

Sin embargo, dos entidades se apartan de esta tendencia:

- **Ciudad de México**, con un 2.3%, muestra una ligera subrepresentación.
- **Distrito Federal**, con apenas 0.8%, constituye un valor atípico significativo

Esta validación confirma que, en general, la información es consistente y homogénea. No obstante, las excepciones detectadas requieren atención, ya que dichas variaciones están relacionadas con el cambio administrativo o de nomenclatura que experimentó la Ciudad de México.

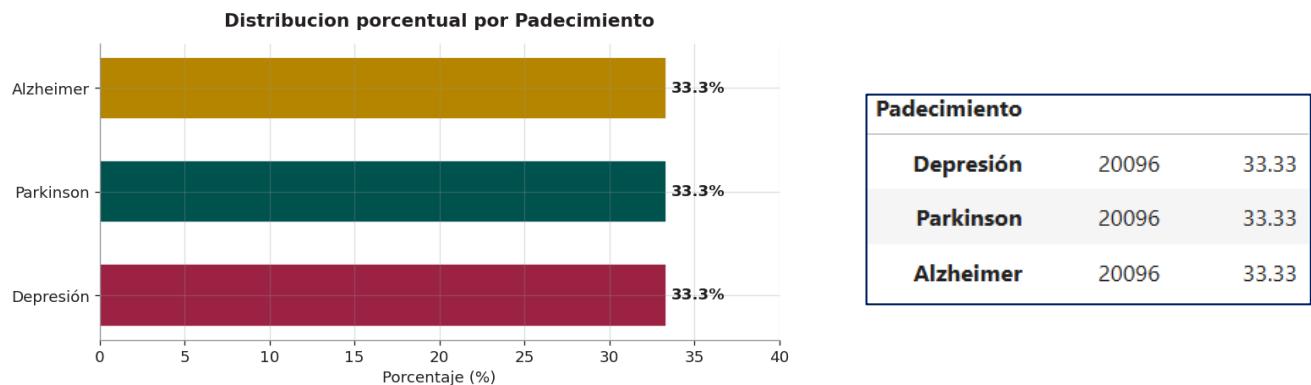
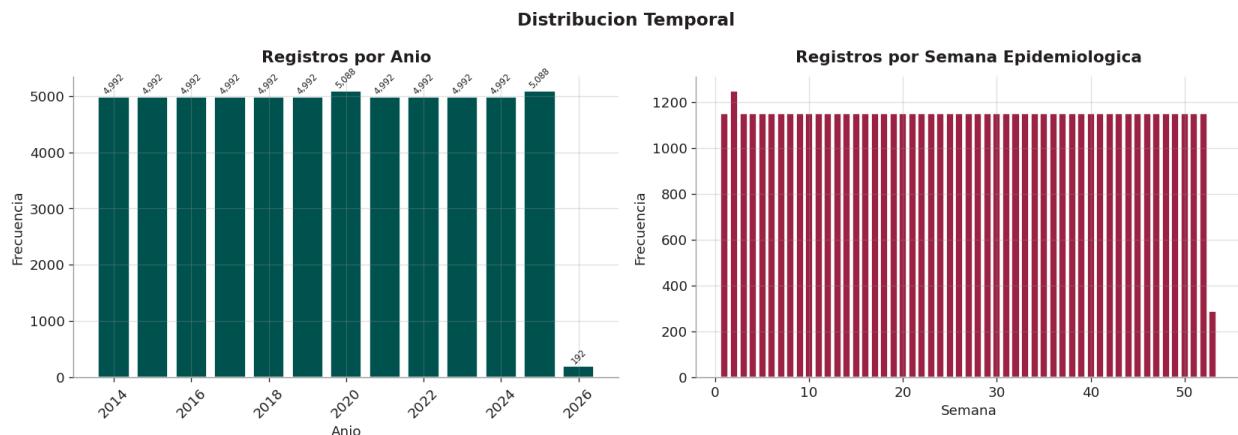


Imagen 7 - Distribución porcentual de "Padecimiento"

La variable **Padecimiento** corresponde a las condiciones médicas registradas en el conjunto de datos, las cuales incluyen *Depresión*, *Parkinson* y *Alzheimer*. Al analizar su distribución porcentual, la gráfica evidencia que cada padecimiento representa el **33.3%** del total, lo que refleja una distribución perfectamente equilibrada entre las tres categorías.

El análisis univariante de esta variable confirma una homogeneidad completa en la proporción de los padecimientos. Esta uniformidad sugiere que la base de datos está organizada de manera consistente; sin embargo, también abre la posibilidad de evaluar si dicha simetría responde a características reales de la población estudiada o a criterios técnicos empleados en la construcción del conjunto de datos.



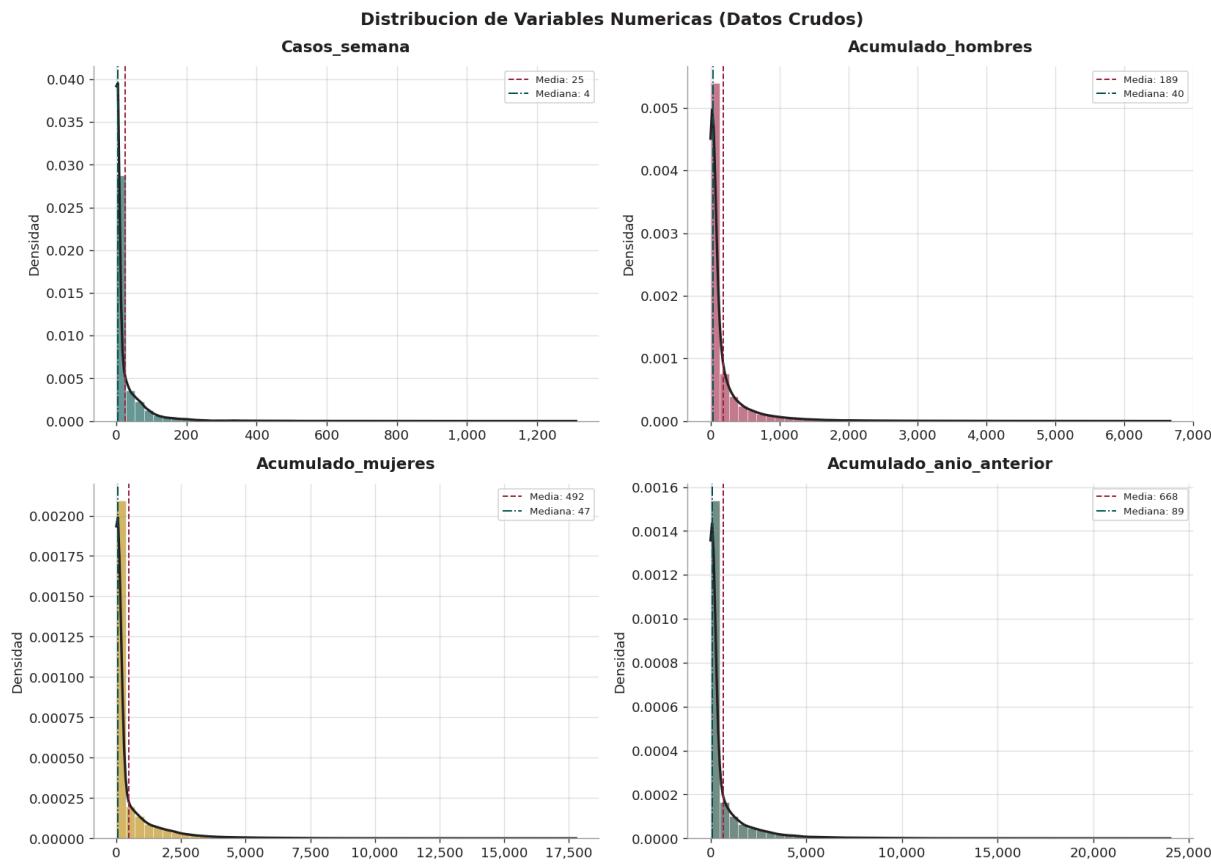


Imagen 8 - Histograma de variables numéricas

Para las variables numéricas del conjunto de datos, se realizó un análisis exploratorio mediante histogramas complementados con una estimación de densidad de kernel (KDE). Esta técnica permite visualizar no solo la frecuencia de los valores observados, sino también una aproximación continua de su distribución subyacente. Al incorporar la densidad, se obtiene una representación más suave que facilita la identificación de patrones, concentraciones y posibles asimetrías en los datos.

Año

El gráfico muestra la comparación de dos series de datos a lo largo del periodo comprendido entre 2014 y 2026. La serie representada con barras mantiene valores casi constantes, entre **0.12** y **0.13**, durante los años 2014–2025, lo que evidencia una notable **estabilidad** en la variable analizada. Esta regularidad indica que la tasa correspondiente al cálculo de 52 semanas se ha conservado, con excepción de 2020 (cuando se registran más de 52 semanas) y de 2026, año en el que solo se dispone de información correspondiente a la semana 1.

En contraste, la línea KDE muestra una tendencia fluctuante en el mismo intervalo, con variaciones periódicas que sugieren posibles ciclos o un componente estacional. El punto más sobresaliente es la caída pronunciada observada en 2026, que rompe con el patrón previo. Esta disminución se explica porque únicamente se cuenta con el dato de la semana 1 para ese año.

Semana

La gráfica representa la distribución de una variable a lo largo de las semanas del año, identificadas del 1 al 52, e incorpora una semana adicional (la semana 53) correspondiente al año 2020. Este tipo de visualización facilita el análisis tanto de la frecuencia semanal como de la tendencia general del comportamiento temporal de la variable.

El patrón observado muestra que la distribución se mantiene relativamente uniforme entre las distintas semanas, lo cual sugiere que la variable presenta una presencia constante durante el ciclo anual. La línea de estimación de densidad (KDE) refuerza esta interpretación, al exhibir un comportamiento estable en la mayor parte del periodo representado. Asimismo, la forma de la curva apunta a la posible existencia de un componente estacional o acumulativo: se identifica una mayor concentración en las semanas iniciales, seguida de un periodo de estabilidad en la parte media del año y una disminución hacia las semanas finales.

Aunque la inclusión de la semana 53 del año 2020 no modifica de manera sustantiva la tendencia general, su presencia debe considerarse en procedimientos de normalización, agregación o comparación temporal, especialmente cuando se analizan series de años con diferente número de semanas.

Casos semana y acumulados

Las distribuciones presentadas en las diferentes variables (*Casos_semana*, *Acumulado_anio_anterior*, *Acumulado_hombres* y *Acumulado_mujeres*) muestran un comportamiento muy similar. En todos los casos se observa una asimetría positiva marcada (sesgo a la derecha), donde la mayor concentración de valores se ubica en rangos bajos, mientras que aparece una cola larga que se extiende hacia valores muy elevados. Este patrón indica la presencia de semanas o acumulados atípicos con incrementos inusuales en el número de casos.

El análisis mediante KDE revela características consistentes entre todas las variables:

- La densidad alcanza su máximo en valores cercanos a cero, disminuyendo rápidamente conforme los conteos aumentan.
- Este comportamiento es típico de distribuciones donde la varianza supera significativamente a la media, un fenómeno común en datos epidemiológicos semanales o acumulados.
- Las colas prolongadas evidencian la presencia de valores extremos, correspondientes a semanas o acumulados que se desvían sustancialmente del comportamiento usual.

Análisis bivariante

En conjunto, las gráficas confirman que todas las variables analizadas comparten una distribución fuertemente sesgada, con alta concentración en valores bajos y presencia de observaciones atípicas que influyen en la forma general de la distribución.

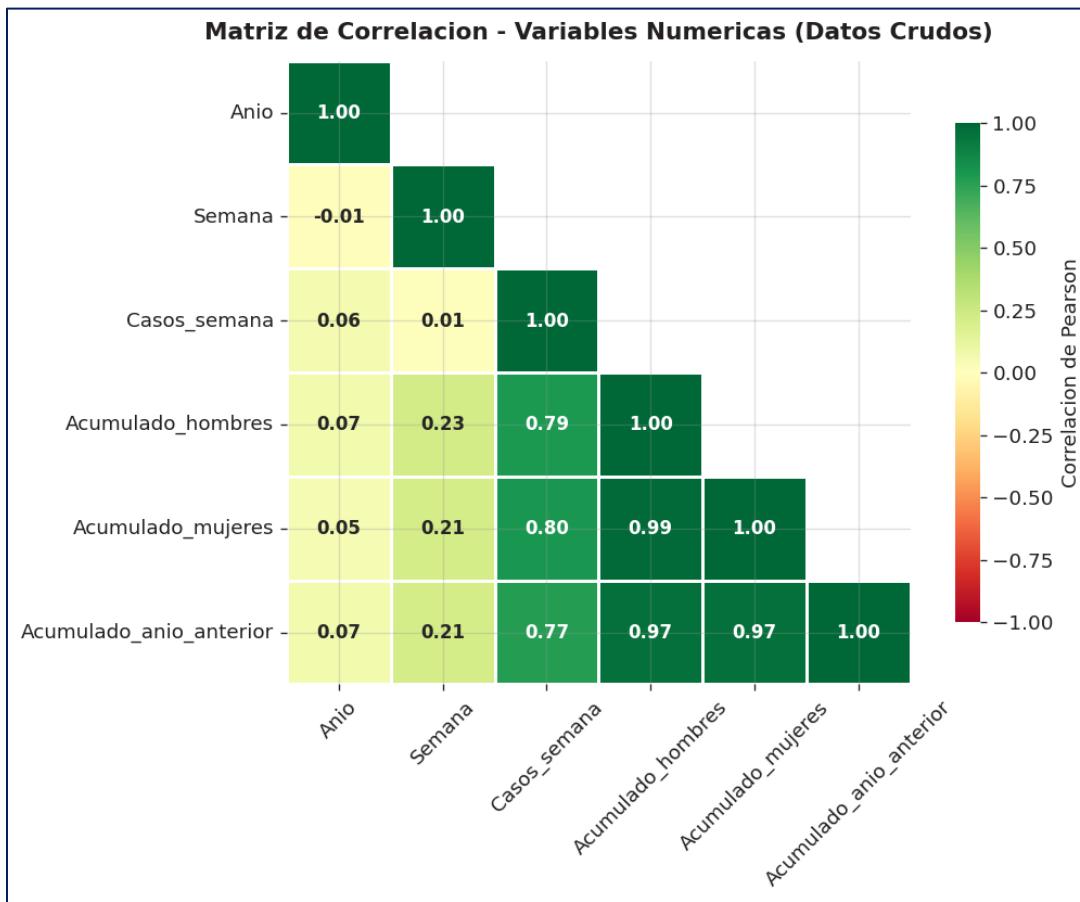


Imagen 9 - Matriz de correlación

Como parte del análisis exploratorio, se construyó una matriz de correlación para evaluar la relación entre las variables numéricas del conjunto de datos. Esta herramienta permite identificar qué variables están más vinculadas entre sí, lo cual es útil para detectar patrones relevantes.

La variable **Casos_semana** muestra una fuerte correlación positiva con:

- *Acumulado_mujeres* con una correlación de **0.89**
- *Acumulado_hombres* con una correlación de **0.81**
- *Acumulado_año_anterior* con una correlación de **0.77**

Estas correlaciones son esperables, ya que los acumulados por sexo se incrementan semana a semana, y por lo tanto guardan una relación directa con los casos reportados en cada semana. En el caso de *Acumulado_año_anterior*, aunque también tiende a aumentar, su función es más comparativa que descriptiva, por lo que su relación con los acumulados actuales es menos directa.

Por otro lado, las variables *Semana* y *Anio* presentan correlaciones muy bajas con *Casos_semana* (0.012 y 0.079, respectivamente). Esto se debe a que son valores fijos y cíclicos: cada año contiene semanas numeradas del 1 al 52 (o 53 en algunos casos), lo que limita su capacidad para explicar variaciones en los casos semanales.

Evaluación de la Variabilidad

El análisis de la distribución de casos de Alzheimer entre 2014 y 2026 muestra un comportamiento volátil y fragmentado, con violines que presentan múltiples protuberancias y ensanchamientos en distintos niveles de la escala. Esto indica que la cantidad de casos semanales varía de manera significativa y no se concentra en un único valor promedio. Aunque la escala general es menor (con picos máximos cercanos a los 200 casos semanales), hacia 2026 se observa un ensanchamiento del cuerpo de las gráficas, lo que sugiere que las semanas con volúmenes medio-altos de diagnósticos se están volviendo más frecuentes y sostenidas.

Puntos clave:

- **Variabilidad y Picos Históricos:** En hombres y mujeres, los años 2016 y 2019 muestran colas superiores muy alargadas, alcanzando máximos cercanos a 180–220 casos semanales. Esto refleja períodos de detección intensa seguidos de mayor estabilidad, como se aprecia en la contracción de 2025.
- **Comportamiento en 2026:** El violín presenta una de las distribuciones más voluminosas. En lugar de ser una “llama” delgada con un pico alto, se observa un cuerpo más robusto, lo que significa que el número de casos se ha distribuido de manera uniforme en rangos superiores durante todo el año, reflejando una incidencia persistente.
- **Comparativa de Género:** Las mujeres muestran picos máximos ligeramente superiores (superando los 200 casos en 2019 y 2023), mientras que en los hombres los máximos suelen mantenerse por debajo de esa línea.

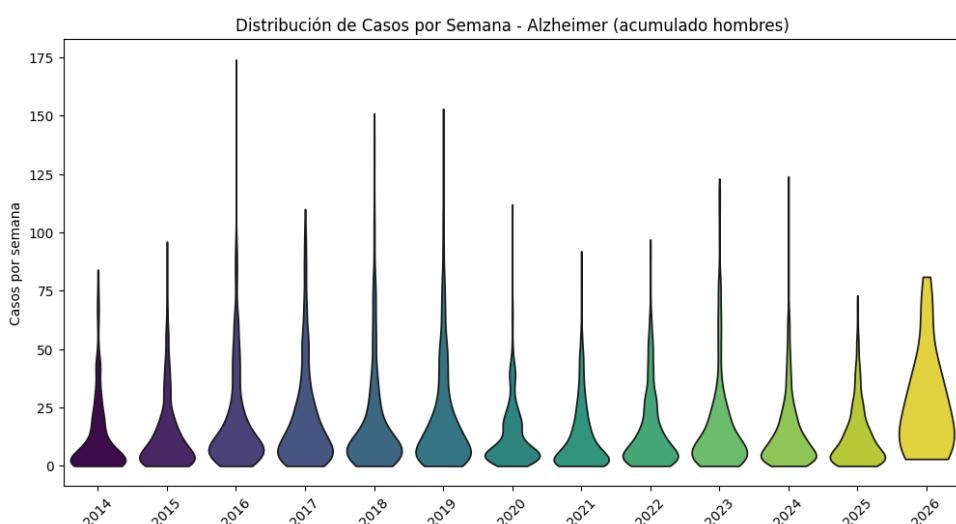


Imagen 10 - Distribución de casos por semana de Alzheimer (hombres)

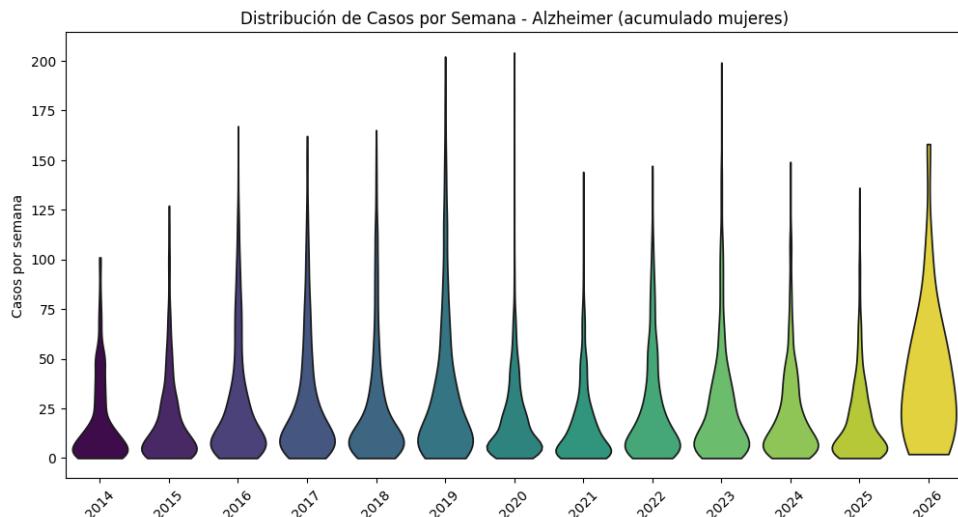


Imagen 11 - Distribución de casos por semana de Alzheimer (mujeres)

Al analizar estas gráficas sobre el Parkinson, se observa un comportamiento estadístico muy distinto al de la depresión, caracterizado por una mayor estabilidad basal, pero con la presencia de valores atípicos extremos. A diferencia del crecimiento lineal de la depresión, aquí la mayoría de las semanas mantienen un registro bajo y constante, reflejado en la base ancha y aplanaada de casi todos los violines. Sin embargo, el año 2016 destaca como una anomalía masiva en ambos sexos, con picos de casos acumulados que rompen totalmente la tendencia, lo cual puede sugerir un cambio en los criterios de diagnóstico, un error de carga de datos o un evento epidemiológico puntual en ese periodo.

Puntos clave:

- Comportamiento de los valores atípicos:** En 2016 los picos son extremadamente altos y delgados (aumento súbito y no sostenido), mientras que en 2026 la base es más voluminosa y redondeada, reflejando un incremento sostenido en los casos semanales.
- Relación de Género:** En los años “normales” los rangos de casos semanales son similares entre hombres y mujeres (mayormente por debajo de 500), aunque el pico anómalo de 2016 fue mucho más alto en la población femenina.
- Tendencia Reciente:** Desde 2022 se percibe un ensanchamiento gradual de la parte inferior del violín, lo que indica que la base de personas diagnosticadas con Parkinson por semana está subiendo lentamente.

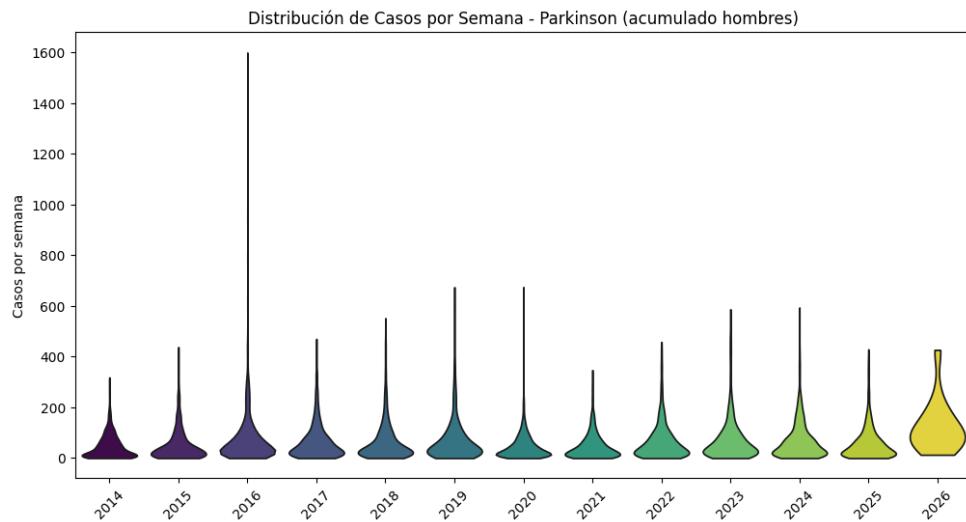


Imagen 12 - Distribución de casos por semana de Parkinson (hombres)

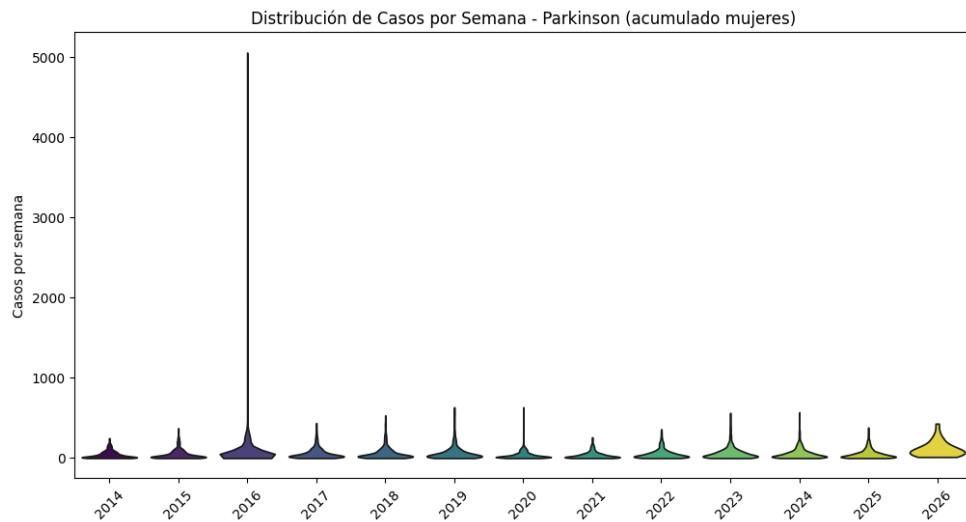


Imagen 13 - Distribución de casos por semana de Parkinson (mujeres)

Al comparar ambos conjuntos de datos, destaca de inmediato la disparidad en el volumen total, mientras que en los hombres el pico máximo de casos acumulados por semana se acerca a los 7,000 hacia el final del periodo, en las mujeres esta cifra es drásticamente superior, superando los 18,000 casos en 2026. Ambas gráficas muestran una estructura similar de "cuello de botella", donde la base es ancha (indicando una alta frecuencia de semanas con pocos casos registrados), pero con colas superiores que se han ido alargando significativamente con el paso de los años, lo que sugiere un aumento constante en los picos máximos de incidencia semanal.

Puntos clave:

- **Evolución Temporal:** Se observa un crecimiento sostenido desde 2014, con una aceleración notable a partir de 2022. Resulta interesante que en 2021 hubo una ligera contracción en los máximos (posiblemente por un subregistro o cambio en la atención durante la pandemia), seguida de un repunte agresivo en los años posteriores.

- Distribución y Densidad:** La mayoría de los datos se concentran cerca del cero (la parte más ancha del "violín"), pero la "forma de llama" se vuelve más voluminosa en 2026. Esto indica que no solo están aumentando los casos máximos, sino que el promedio semanal de casos está subiendo de nivel de manera generalizada.
- Brecha de Género:** La escala del eje Y confirma que las mujeres presentan más del doble de casos registrados que los hombres.

Al observar el conjunto de las tres enfermedades, se hace evidente que la salud mental (depresión) representa el mayor desafío en términos de volumen de atención, con cifras que superan por más de diez veces a las enfermedades neurodegenerativas analizadas. Mientras que el Parkinson y el Alzheimer muestran picos puntuales que sugieren una incidencia más controlada o vinculada a grupos etarios específicos, la depresión evidencia una expansión masiva que afecta desproporcionadamente a la población femenina.

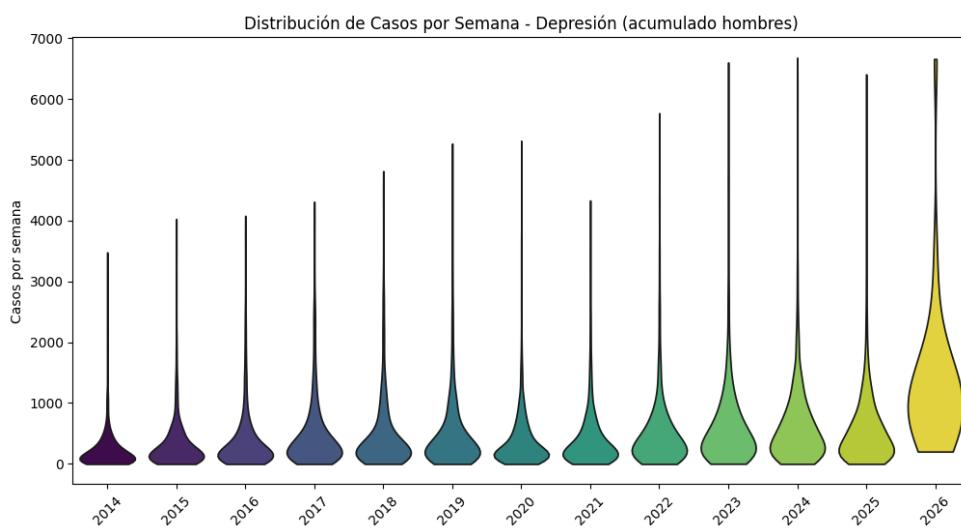


Imagen 14 - Distribución de casos por semana de Depresión (hombres)

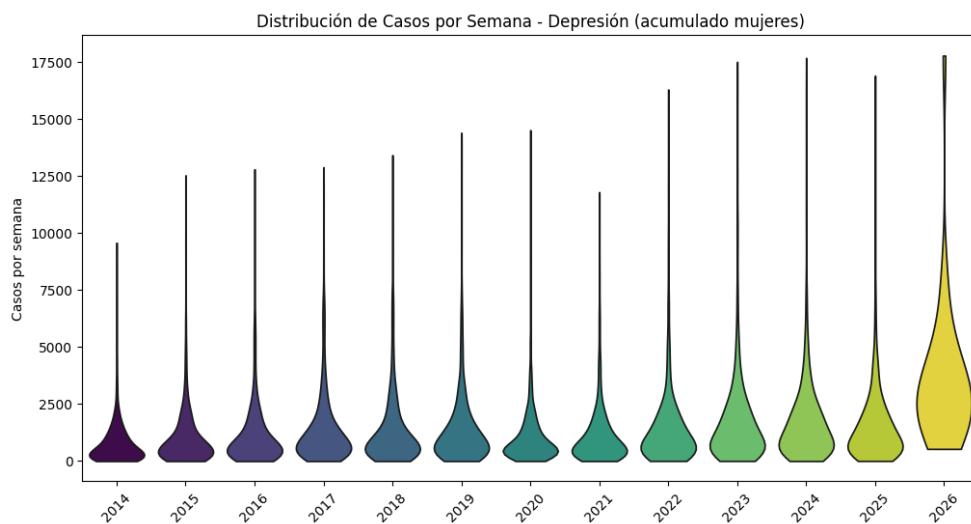


Imagen 15 - Distribución de casos por semana de Depresión (mujeres)

El comportamiento de las gráficas en el año 2026 destaca que este periodo presenta los valores más altos y las distribuciones más voluminosas de toda la serie histórica en las tres condiciones. Sin embargo, dado que estos datos corresponden únicamente a la primera semana del año, la presencia de cifras tan elevadas (superando los 18,000 casos en depresión femenina o los 500 en Parkinson masculino) indica un desfase técnico en el reporte acumulado.

Este fenómeno implica que los valores registrados bajo la etiqueta de "2026" son, en realidad, el reporte consolidado de los datos finales de 2025 que fueron procesados con una semana de retraso. Por lo tanto, el crecimiento exponencial que se aprecia visualmente al final de cada gráfica representa el cierre estadístico real del año anterior.

Limpieza de datos

Con base en el análisis exploratorio de datos (EDA) realizado, se procedió a efectuar un proceso de limpieza siguiendo los criterios que se detallan a continuación:

Filtrado por padecimiento

Se seleccionó únicamente un padecimiento específico con el fin de centrar el análisis en él. Este procedimiento debe repetirse para cada uno de los padecimientos registrados en la base de datos, garantizando que el análisis se realice de manera individual y comparativa según corresponda.

Eliminación de columnas: Durante el proceso de limpieza se identificaron variables que no aportaban valor directo al análisis o cuya información resultaba redundante o poco confiable. Por ello, se decidió eliminar las siguientes columnas:

- **Padecimiento:** Al centrarnos en un solo padecimiento, esta columna deja de ser necesaria, ya que su información se vuelve implícita en el conjunto de datos filtrado.
- **Acumulado_año_anterior:** Aunque refleja una tendencia de incremento, su función es principalmente comparativa y no descriptiva. Además, el análisis de correlación muestra una alta relación con los acumulados actuales de hombres y mujeres (coeficientes de 0.97), lo que indica redundancia en la información. Se optó por excluirla para evitar multicolinealidad y facilitar la interpretación de los modelos.
- **Casos_semana:** Este campo muestra el incremento semanal de casos, calculado a partir de la diferencia acumulada entre hombres y mujeres respecto a la semana anterior. Sin embargo, se detectaron inconsistencias en la captura, ya que en ocasiones el valor registrado como nuevos casos no coincidía con la diferencia obtenida entre los incrementales. Además, el análisis de correlación revela una fuerte asociación con los acumulados por sexo (0.80 y 0.81), lo que refuerza su carácter derivado. Dado que el análisis se realiza con la información desagregada por sexo, esta columna pierde relevancia y no aporta valor adicional.

Normalización de entidades: En la columna Entidad, se sustituyó la denominación Distrito Federal por Ciudad de México, ya que ambas corresponden a la misma entidad federativa. Esta normalización permite asegurar la consistencia en la agrupación de datos.

Preparación de los datos

Una vez normalizados los datos, se procede a tratarlos con el fin de prepararlos para su procesamiento mediante técnicas de Machine Learning y para la generación de visualizaciones que aporten contexto al análisis realizado. Este tratamiento busca garantizar la calidad y consistencia de la información, facilitando la identificación de patrones relevantes y la obtención de resultados significativos.

Ajuste de semanas epidemiológicas

Se identificó que la información presentada semanalmente correspondía en realidad a la semana anterior. Asimismo, la semana 1 contenía el acumulado hasta la última semana del año previo, lo que generaba un desfase en la interpretación de los registros. Para garantizar la concordancia y consistencia de la base de datos, se decidió recorrer todas las semanas una unidad, de modo que los valores reflejen correctamente el periodo epidemiológico al que pertenecen.

Este ajuste se implementó mediante un procedimiento que:

- **Verifica** que las semanas estén dentro del rango válido (1 a 53), generando un error en caso contrario.
- **Identifica** los registros correspondientes a la semana 1 y los distingue de los demás.
- **Resta una unidad** a todas las semanas distintas de la semana 1, corrigiendo el desfase detectado.
- **Calcula** el máximo de semanas observadas por año y construye un mapa de referencia para cada periodo.
- **Reasigna** a los registros de semana 1 el año anterior y les otorga como nueva semana el máximo del año previo más uno, asegurando continuidad en la serie temporal.
- **Ordena** finalmente la base de datos por año, entidad y semana, garantizando consistencia en la estructura.

Preparación de las series de tiempo

Para transformar los datos acumulados en una estructura adecuada para el análisis temporal, se aplicó un proceso que permitió calcular incrementos semanales y asociar cada registro con una fecha precisa.

[1] **Transformación de datos acumulados en incrementos semanales:** La información original se encontraba registrada como valores acumulados por entidad y sexo. Para poder analizar la dinámica temporal, se procedió a calcular la diferencia respecto a la semana anterior, obteniendo así el crecimiento semanal o delta. Este paso permitió convertir los acumulados en series de tiempo que reflejan con mayor precisión la evolución de los casos semana a semana. De esta manera, los datos dejaron de ser únicamente acumulativos y pasaron a reflejar la dinámica semanal.

[2] **Cálculo de incrementos semanales:** A partir de la diferencia entre el acumulado actual y el acumulado previo, se obtuvieron los casos nuevos por semana, diferenciados por sexo. De esta

manera, los datos dejaron de ser únicamente acumulativos y pasaron a reflejar la dinámica semanal.

- [3] **Tratamiento especial para la semana 1:** Dado que se realizó un corrimiento en las semanas para asegurar la concordancia de la información, la semana 1 se considera el inicio del ciclo anual. En este punto no se efectuó una comparativa con la semana anterior, ya que no existe un valor previo consecutivo dentro del mismo año. Por ello, se decidió respetar directamente el valor reportado como incremento inicial, garantizando que el arranque del año epidemiológico refleje fielmente los datos originales y mantenga la coherencia en la serie temporal.
- [4] **Asignación de fechas específicas:** Para asociar cada registro con una fecha precisa se empleó el estándar ISO de semanas. Este sistema define las semanas del año de manera uniforme y se utiliza ampliamente en contextos epidemiológicos y estadísticos.

El funcionamiento del ISO de semanas es el siguiente:

- Cada semana comienza en lunes, lo que asegura consistencia en la comparación entre períodos.
- El año ISO puede diferir del año calendario tradicional, ya que la primera semana del año ISO es aquella que contiene el primer jueves de enero.
- Esto implica que algunas fechas de principios de enero pueden pertenecer al último año ISO anterior, y de forma similar, los últimos días de diciembre pueden contabilizarse como parte de la primera semana del año siguiente.

Gracias a este sistema, los registros se convirtieron en una serie temporal continua y ordenada, donde cada semana epidemiológica tiene una fecha representativa. En los casos en que la primera semana estuviera en el año anterior, se ajustó para que correspondiera al primer día del año en curso, garantizando coherencia entre el campo de año y la fecha asignada.

Ajuste de valores negativos

Durante la preparación de las series de tiempo se identificaron casos en los que las diferencias semanales resultaban en valores negativos. Este comportamiento es inconsistente, ya que los registros se encuentran en forma de acumulados y, por definición, no deberían mostrar disminuciones entre semanas consecutivas.

Un ejemplo relevante se observó en el año 2016, semana 20, en la entidad de Ciudad de México para la categoría de depresión. En esa semana se reportó un crecimiento anómalo superior a 6,000 unidades, seguido inmediatamente por una reducción drástica a menos de 200 unidades en la semana siguiente. Este tipo de variaciones contradicen la lógica acumulativa y evidencian inconsistencias en la fuente de datos o en el proceso de registro.

Para abordar las inconsistencias detectadas en los incrementos de las semanas 20 y 21 del año 2016, se aplicó un procedimiento de corrección basado en la información de las semanas vecinas. El objetivo fue preservar la lógica acumulativa de los datos y evitar reducciones artificiales que no corresponden a la realidad epidemiológica.

El proceso consistió en:

- **Identificación de registros negativos:** Se localizaron los casos en los que los incrementos semanales resultaban menores a cero, lo cual es incompatible con la naturaleza acumulativa de la información.
- **Verificación de consecutividad:** Se comprobó que los registros negativos correspondieran a semanas consecutivas dentro de la misma entidad y año, asegurando que el ajuste se aplicara únicamente en contextos válidos.
- **Reajuste del valor previo:** Cuando se detectaba un incremento negativo consecutivo a un valor positivo, se redistribuía el exceso restando el valor negativo al incremento de la semana anterior. Con ello se evitaba que la serie mostrara caídas abruptas.
- **Extrapolación con semanas anteriores:** Para reforzar la coherencia, se calculó un promedio de los incrementos de las tres semanas previas y se asignó como valor corregido en la semana afectada.
- **Normalización final:** Se garantizó que todos los valores quedaran expresados como enteros y que cualquier dato faltante se tratara como cero, manteniendo la estabilidad de la serie temporal.

El proceso anterior estuvo enfocado en los casos detectados durante el año 2016; sin embargo, también se aplicó un procedimiento similar para aquellos registros que aún presentaban valores negativos. Para garantizar la coherencia de la información, se siguieron los siguientes pasos:

- **Identificación de registros problemáticos:** Se localizaron los casos en los que los incrementos semanales resultaban negativos, lo cual es incompatible con la naturaleza acumulativa de los datos.
- **Referencia a semanas adyacentes:** Para cada registro anómalo se consideraron los valores de la semana anterior y la semana siguiente como puntos de referencia.
- **Extrapolación del valor corregido:** Se calculó un promedio simple entre los valores de las semanas vecinas, obteniendo una estimación más coherente con la tendencia de la serie temporal.
- **Sustitución controlada:** Únicamente los valores negativos fueron reemplazados por la estimación extrapolada, manteniendo intactos los registros que no presentaban inconsistencias.
- **Normalización final:** Se aseguró que todos los valores quedaran expresados como enteros y que cualquier dato faltante o inconsistente se tratara como cero, garantizando la estabilidad y continuidad de la serie.

Tratamiento de valores atípicos

Finalmente, para garantizar la consistencia de las series de tiempo y evitar que valores extremos distorsionaran el análisis, se aplicó un procedimiento de imputación de valores atípicos utilizando el método del rango intercuartílico (IQR).

Este enfoque permite:

- **Detectar** registros que se encontraban fuera del rango esperado, definido por los límites estadísticos del IQR.
- **Identificar** aquellos valores que podían considerarse atípicos por su magnitud inusual respecto al comportamiento general de la serie.
- **Imputar** dichos valores mediante ajustes controlados, sustituyéndolos por estimaciones más coherentes con la distribución de los datos.
- **Preservar** la tendencia global de la información, evitando que anomalías puntuales afectaran la interpretación de patrones y resultados.

Resultado de los datos tras limpieza y corrección

```
RangeIndex: 20096 entries, 0 to 20095
Data columns (total 5 columns):
 #   Column            Non-Null Count  Dtype  
 ---  --  
 0   Fecha              20096 non-null   object 
 1   Entidad             20096 non-null   object 
 2   incrementos_hombres 20096 non-null   int64  
 3   incrementos_mujeres 20096 non-null   int64  
 4   Region              20096 non-null   object 
dtypes: int64(2), object(3)
memory usage: 785.1+ KB
```

Resumen del DataFrame	Valor
Fecha	2026-01-31 01:40
Filas	20,096
Columnas	5
Columnas numéricas	2
Columnas categóricas	3
Otras columnas	0
Porcentaje nulos	0.00%

Imagen 16 - Detalles del DataFrame después de limpieza

El resultado del proceso de limpieza, transformación e imputación de datos es la creación de un DataFrame específico para cada padecimiento, lo que garantiza que la información se mantenga aislada y organizada. Esta estructura facilita un análisis más preciso de la evolución temporal y permite identificar con mayor claridad las diferencias por sexo y región en cada caso. Cada DataFrame conserva la misma estructura básica siguiente:

- **Fecha:** campo utilizado para el análisis temporal, permite observar la evolución de los registros a lo largo del tiempo.
- **Entidad:** corresponde a la unidad geográfica (estado), útil para identificar patrones locales y diferencias territoriales.
- **Incrementos_hombres y Incrementos_mujeres:** variables que registran las variaciones en el número de casos por sexo, facilitando el análisis de brechas de género.
- **Región:** clasificación adicional que agrupa las entidades según su ubicación geográfica (Norte, Occidente, Centro, Sureste), lo que permite realizar comparaciones agregadas y detectar tendencias regionales.

Además, tras el proceso de limpieza y extracción de características, no se presentan valores nulos en ninguna de las columnas, lo que facilita un mejor procesamiento en modelos de Machine Learning, al reducir la necesidad de imputaciones adicionales y garantizar que los algoritmos trabajen con datos completos y consistentes.

Distribución de valores únicos para Alzheimer			Distribución de valores únicos para Parkinson			Distribución de valores únicos para Depresión		
Columna	Valores Únicos	Tipo	Columna	Valores Únicos	Tipo	Columna	Valores Únicos	Tipo
Fecha	628	object	Fecha	628	object	Fecha	628	object
Entidad	32	object	Entidad	32	object	incrementos_mujeres	130	int64
Region	4	object	incrementos_hombres	9	int64	incrementos_hombres	49	int64
incrementos_hombres	3	int64	incrementos_mujeres	9	int64	Entidad	32	object
incrementos_mujeres	3	int64	Region	4	object	Region	4	object

Imagen 17 - Distribución de valores únicos por padecimiento

El análisis de valores únicos posterior al procesamiento de datos evidencia una reducción significativa de la cardinalidad. Esta disminución no implica pérdida de información, sino la eliminación de redundancia y ruido en tres aspectos clave

Específicamente, la unificación de los registros en 628 fechas únicas a lo largo de 13 años confirma la resolución de disparidades en los formatos temporales, logrando una alineación cronológica estructuralmente coherente para las series de tiempo de los tres padecimientos.

Asimismo, en las variables de "incrementos" y casos, la cardinalidad acotada, particularmente en los conjuntos de datos de Alzheimer y Parkinson, sugiere que el proceso de imputación ejerció un efecto de suavizado sobre los datos. Esto permitió la eliminación de valores atípicos no representativos y errores de captura, preservando únicamente aquellos valores que reflejan con mayor fidelidad los patrones epidemiológicos consistentes de cada enfermedad.

	count	mean	std	min	25%	50%	75%	max
incrementos_hombres	20096.00	0.50	0.72	0.00	0.00	0.00	1.00	2.00
incrementos_mujeres	20096.00	0.68	0.81	0.00	0.00	0.00	1.00	2.00

	count	mean	std	min	25%	50%	75%	max
incrementos_hombres	20096.00	2.21	2.27	0.00	0.00	2.00	3.00	8.00
incrementos_mujeres	20096.00	1.96	2.18	0.00	0.00	1.00	3.00	8.00

	count	mean	std	min	25%	50%	75%	max
incrementos_hombres	20096.00	0.50	0.72	0.00	0.00	0.00	1.00	2.00
incrementos_mujeres	20096.00	0.68	0.81	0.00	0.00	0.00	1.00	2.00

Imagen 18 - Distribución de las variables de incrementos por padecimiento después de limpieza

Las nuevas distribuciones obtenidas para los incrementos en Alzheimer, Parkinson y Depresión evidencian un cambio sustancial respecto a las distribuciones originales basadas en datos acumulados. Este cambio no solo es visual, sino estadísticamente significativo, y se debe directamente a los procesos de limpieza, normalización y extracción de características, cuyo objetivo fue transformar variables altamente dispersas e inestables en atributos más informativos, manejables y útiles para el modelado.

Las gráficas iniciales de Acumulado_hombres y Acumulado_mujeres mostraban distribuciones extremadamente sesgadas a la derecha, con valores que llegan a rangos de miles. Este patrón es característico de variables altamente acumulativas, en las que el crecimiento a lo largo del tiempo genera valores desproporcionadamente grandes en relación con la mayoría de las observaciones.

Tras aplicar limpieza y extracción de características, las nuevas variables (incrementos diarios o relativos) presentan distribuciones mucho más estables y alineadas con propiedades estadísticas apropiadas. Esto es claramente visible en las nuevas gráficas:

Alzheimer

La distribución transformada de Alzheimer muestra una reducción drástica respecto a la forma original, pasando a concentrarse únicamente en un conjunto discreto y estrecho de valores. Este cambio refleja la eliminación del sesgo extremo presente en los datos en bruto, así como una disminución significativa en la dispersión general, lo que da lugar a una estructura mucho más homogénea y coherente.

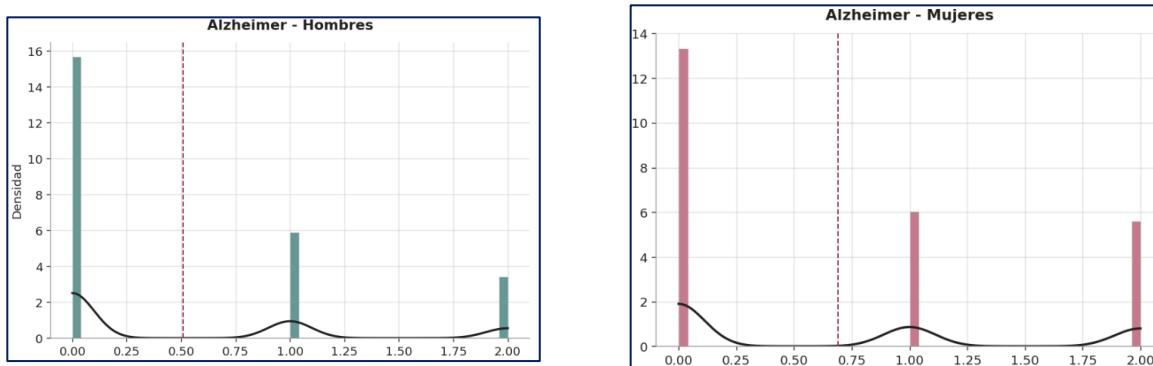


Imagen 19 - Distribución transformada de la variable de incrementos para Alzheimer dividida por sexo

Parkinson

La distribución transformada correspondiente a Parkinson, aunque mantiene una amplitud mayor que la observada en Alzheimer, se encuentra ahora claramente contenida dentro de un rango mucho más compacto y gestionable en comparación con la escala ampliamente dispersa que presentaban los datos en bruto. Esta reducción significativa en la dispersión elimina la presencia de comportamientos extremos que anteriormente afectaban la estabilidad estadística y dificultaban la interpretación. La nueva forma de la distribución muestra una variabilidad que, si bien continúa

presente, se expresa de manera controlada y estructurada, permitiendo que la información relevante se mantenga sin introducir fluctuaciones excesivas. Asimismo, el sesgo que se observa es moderado y ya no se encuentra influido por valores desproporcionados, lo que favorece una representación más equilibrada de los patrones de incremento.

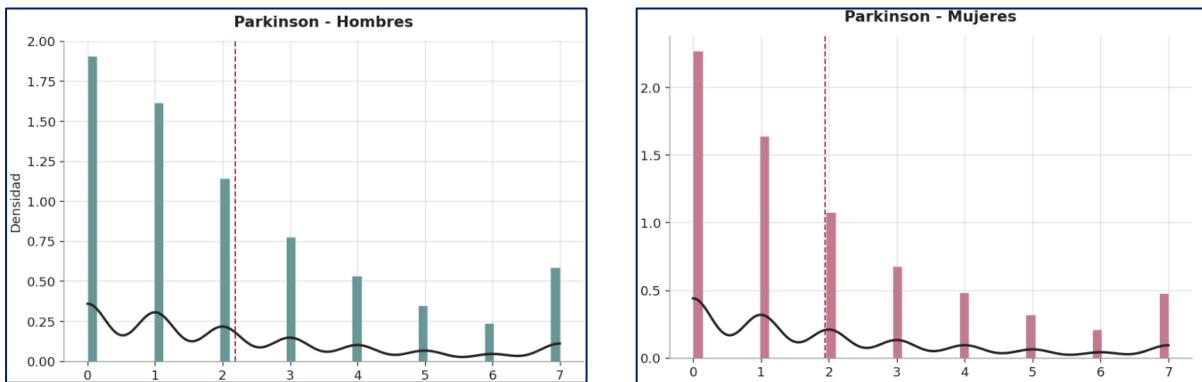


Imagen 20 - Distribución transformada de la variable de incrementos para Parkinson dividida por sexo

Depresión

Finalmente, la distribución de los padecimientos de depresión evidencia una transformación sustancial respecto a la versión original basada en datos acumulados. Aunque la nueva distribución sigue mostrando una mayor amplitud relativa en comparación con Alzheimer y Parkinson, ahora se presenta dentro de un rango considerablemente más acotado y manejable, muy distinto al comportamiento previo, donde las observaciones se extendían sobre una escala extremadamente elevada.

Este cambio es relevante porque la variable, tras la limpieza y la extracción de características, deja de reflejar un patrón dominado por la acumulación progresiva (que generaba colas excesivamente largas y una dispersión extrema) y pasa a comportarse como una medida mucho más equilibrada, informativa y representativa de variaciones reales, en lugar de reflejar incrementos desproporcionados derivados del crecimiento acumulativo.

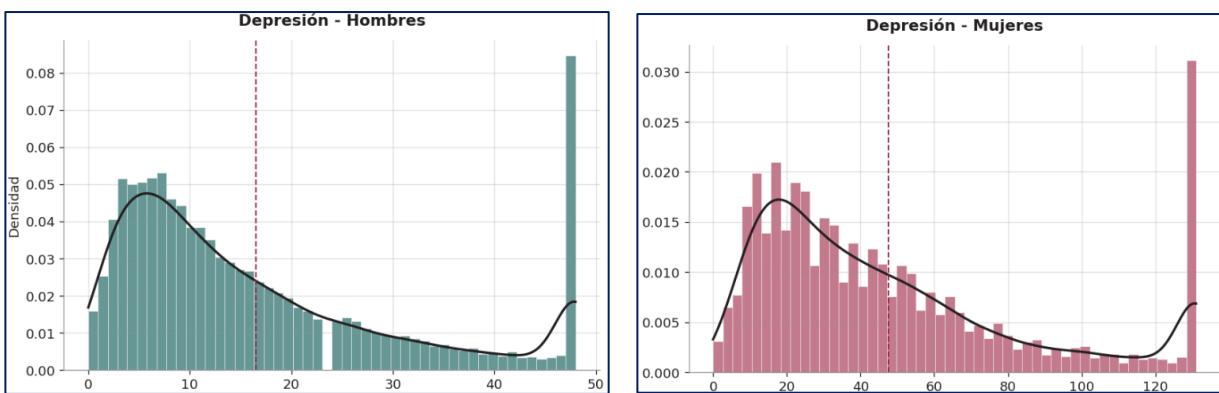


Imagen 21 - Distribución transformada de la variable de incrementos para Depresión dividida por sexo

Las gráficas comparadas muestran de manera contundente cómo el proceso de limpieza y extracción de características logró transformar distribuciones extremadamente dispersas en conjuntos de datos estables, acotados y estadísticamente manejables.

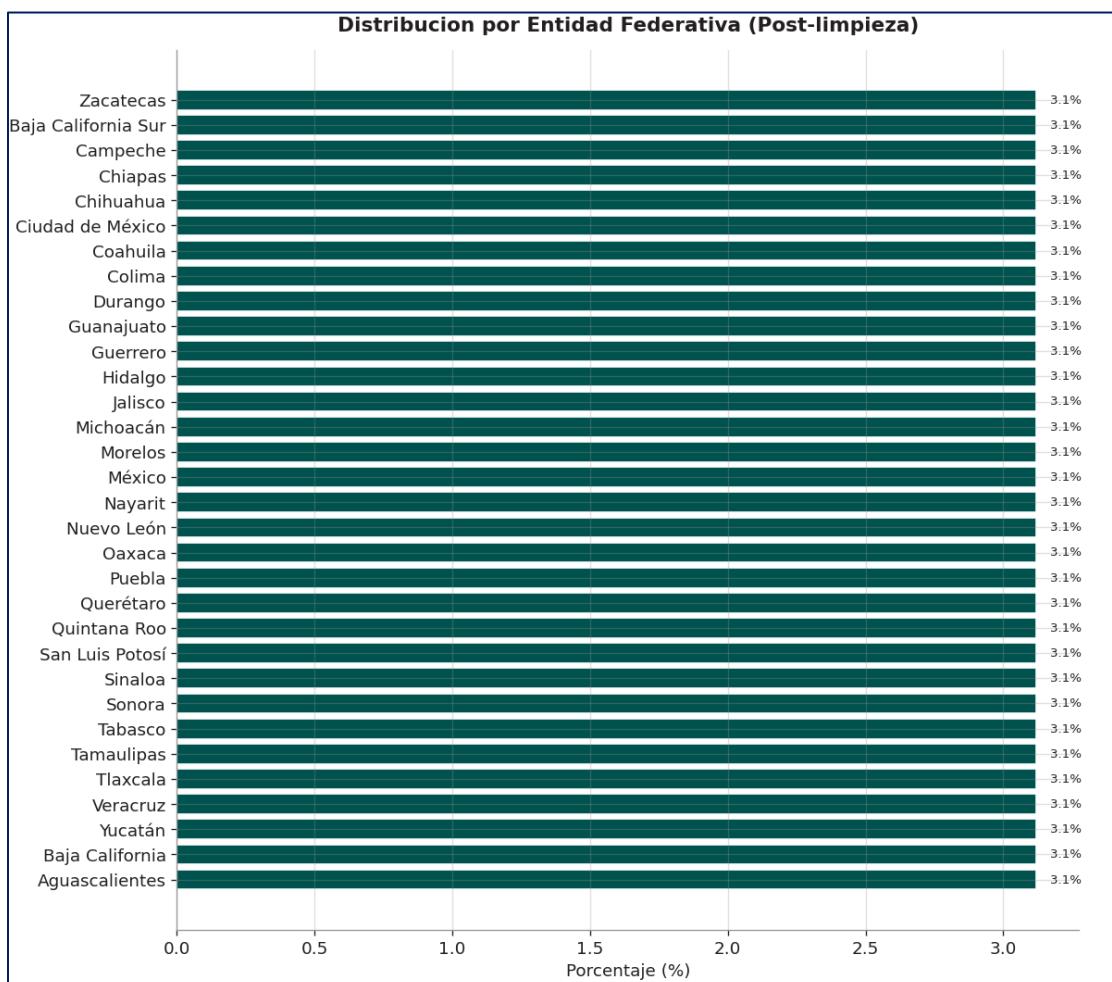


Imagen 22 - Distribución de porcentajes por estado después de la limpieza

Tras aplicar la limpieza y estandarización, la distribución de porcentajes por estado pasó de mostrar variaciones notorias en una entidad a adoptar una estructura mucho más uniforme. Este cambio indica que la depuración eliminó irregularidades y redujo la variabilidad no deseada, resultando en un conjunto de datos más homogéneo.

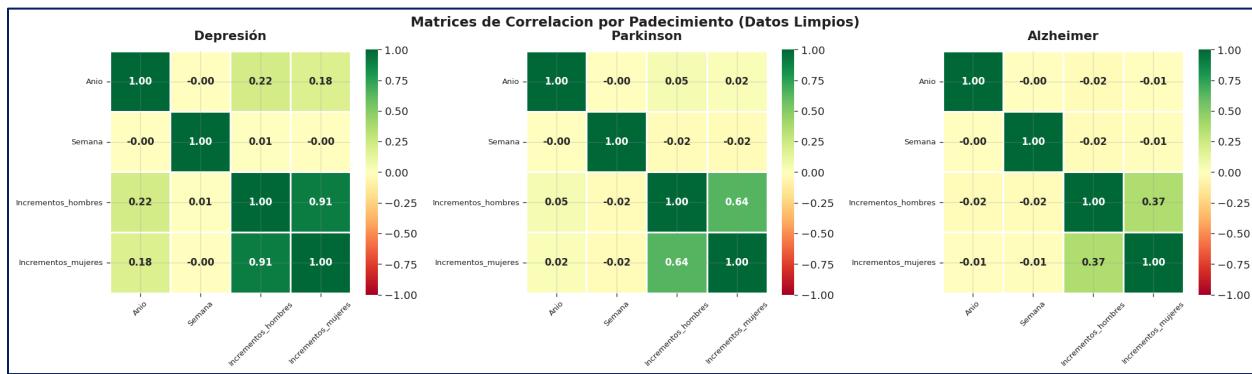


Imagen 23 - Matriz de correlación para cada uno de los padecimientos a analizar

En la matriz de correlación original, donde los padecimientos se encontraban agrupados y las variables se expresaban en formato acumulado, las asociaciones estaban dominadas por la naturaleza misma de los acumulados, generando correlaciones elevadas que reflejaban principalmente una tendencia de crecimiento compartida. Este patrón dificultaba distinguir relaciones específicas entre los fenómenos de interés.

En contraste, las matrices posteriores (ya separadas por padecimiento y construidas a partir de incrementos por sexo) producen correlaciones más claras y propiamente interpretables. Bajo este enfoque, las asociaciones dejan de estar influenciadas por efectos de acumulación y pasan a reflejar la co-variación real entre incrementos de hombres y mujeres.

Las diferencias entre enfermedades se vuelven evidentes: Alzheimer muestra una relación débil a moderada, Parkinson presenta una asociación moderada y Depresión exhibe una sincronía marcada entre ambos grupos.

En conjunto, este cambio metodológico permite que las correlaciones reflejen de manera más fiel los patrones propios de cada padecimiento, lo que a su vez aporta una interpretación más clara y una mayor utilidad analítica para el estudio de las variables.

Evaluación de la Variabilidad

Ahora, las distribuciones cambiaron de ser formas muy extendidas, irregulares y con valores altos heredados del uso de datos acumulados, a convertirse en violines mucho más compactos, uniformes y representativos del comportamiento real semana a semana. Antes, las gráficas mostraban una gran dispersión y colas superiores largas porque el acumulado crecía de manera natural a lo largo del año, lo que hacía que las distribuciones se vieran infladas y difíciles de comparar entre períodos. Despues de aplicar la limpieza y transformar la información en incrementos semanales, las distribuciones se estabilizan: se reducen los extremos, desaparece la asimetría exagerada y los violines toman formas más coherentes y comparables, reflejando variaciones reales en lugar de efectos artificiales del acumulado.

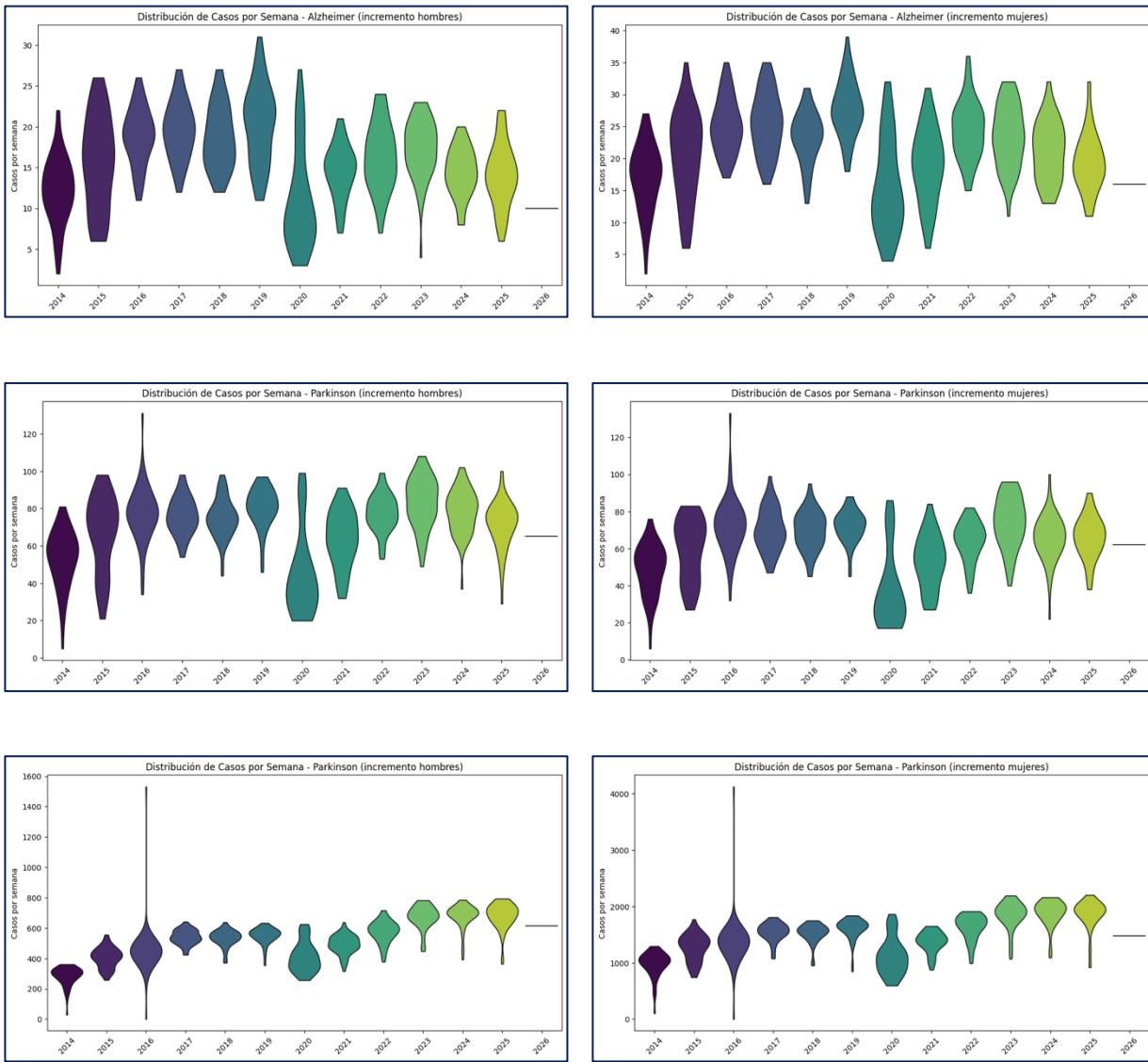


Imagen 24 - Distribución semanal de casos después de la limpieza de datos, por padecimiento y sexo.

Series de tiempo

Las siguientes tres gráficas muestran la evolución semanal de cada padecimiento a nivel nacional entre 2014 y 2026, diferenciados por sexo. Estas visualizaciones corresponden al resultado final del proceso de limpieza, depuración y organización temporal de los datos, lo que permitió obtener series más coherentes y comparables en el tiempo. A partir de este tratamiento, los datos quedaron en una forma adecuada para modelar su comportamiento temporal, facilitando identificar tendencias, variaciones y posibles patrones estacionales de manera más clara.

Sin limpieza de datos

Esta visualización refleja los datos en su estado original, sin correcciones ni ajustes. Se observan anomalías evidentes, como el pico abrupto en 2016, que sugiere inconsistencias en los registros acumulados. También pueden aparecer valores negativos o fluctuaciones poco plausibles que distorsionan la interpretación de la tendencia real.

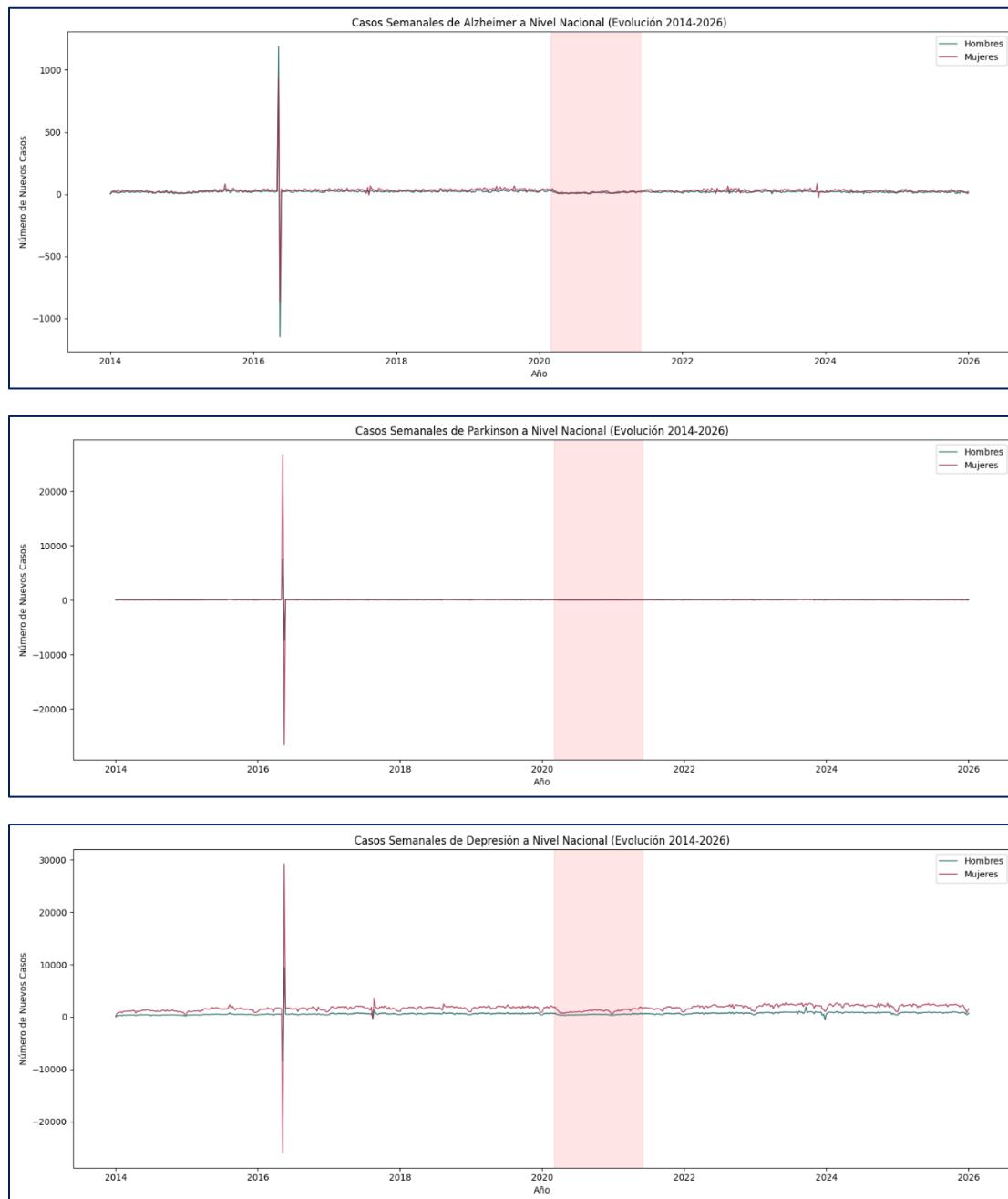


Imagen 25 - Casos semanales nacionales para los tres padecimientos, sin limpieza de datos

Con tratamiento de limpieza aplicado

Aquí se han corregido los valores negativos, ajustado los incrementos semanales, y normalizado las fechas según el calendario epidemiológico. La serie temporal muestra una evolución más coherente, eliminando caídas artificiales y estabilizando los patrones. El comportamiento general es más confiable para análisis comparativos y modelado.

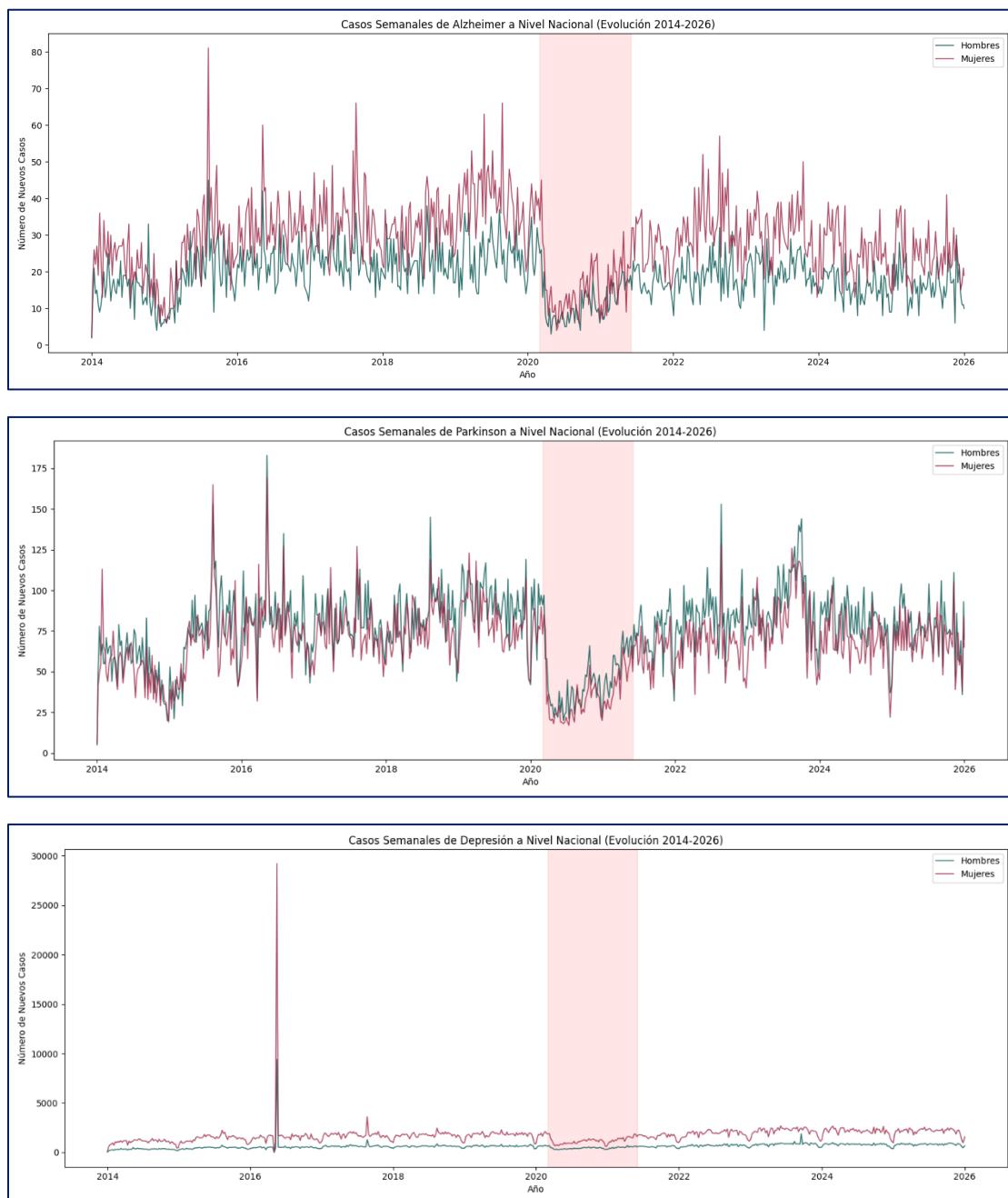


Imagen 26 - Casos semanales nacionales para los tres padecimientos, con tratamiento de limpieza

Con limpieza e imputación de atípicos (IQR)

En esta versión final se ha aplicado un tratamiento para valores atípicos mediante el rango intercuartílico (IQR). Esto permite suavizar los extremos que podrían influir desproporcionadamente en el análisis. La serie resultante conserva la estructura general, pero con menor influencia de picos anómalos, ofreciendo una representación más robusta y estadísticamente depurada.

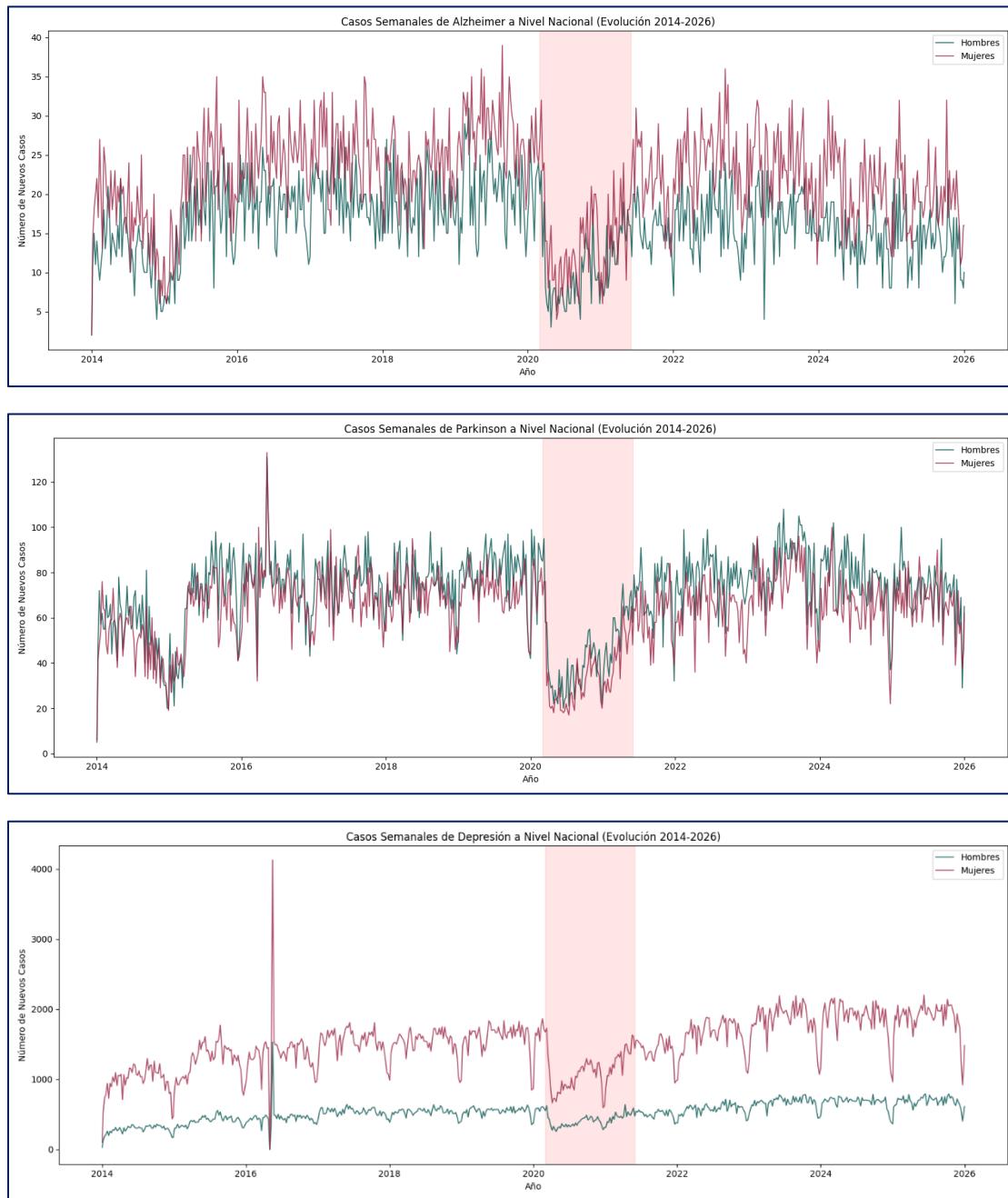


Imagen 27 - Casos semanales nacionales para los tres padecimientos, con limpieza e imputación de atípicos (IQR)

Síntesis del análisis

El documento desarrolla un análisis completo de epidemiología neurológica y de salud mental en México para depresión, Parkinson y Alzheimer, usando información semanal oficial entre 2014 y 2026. Parte de datos no estructurados en PDF y demuestra que, sin un tratamiento adecuado, los datos acumulados distorsionan la interpretación temporal. El trabajo muestra cómo, mediante un pipeline riguroso de limpieza y transformación, es posible obtener series de tiempo coherentes, comparables y útiles para análisis estadístico y modelado.

Pipeline de datos

- Identificación de tablas relevantes en PDFs no estructurados.
- Extracción y carga de la información en un DataFrame base.
- Análisis exploratorio inicial para evaluar estructura, completitud y anomalías.
- Normalización de entidades federativas y depuración de categorías inconsistentes.
- Filtrado por padecimiento para análisis individual.
- Eliminación de variables redundantes o derivadas.
- Corrección de desfases en semanas epidemiológicas.
- Transformación de acumulados a incrementos semanales por sexo.
- Asignación de fechas usando el estándar ISO de semanas.
- Corrección de valores negativos e inconsistencias temporales.
- Tratamiento de valores atípicos mediante IQR.
- Generación de DataFrames finales limpios y listos para análisis y modelado.

Resultados relevantes

- Los datos acumulados generan sesgos fuertes y picos artificiales si se analizan sin limpieza.
- Transformar a incrementos semanales es esencial para capturar la dinámica real de los padecimientos.
- La depresión concentra el mayor volumen de casos y muestra una expansión sostenida, especialmente en mujeres.
- Parkinson y Alzheimer presentan menor volumen, pero con picos puntuales y patrones temporales diferenciados.
- Existen años con anomalías técnicas claras que requieren corrección explícita.
- La limpieza reduce la dispersión extrema y mejora la interpretabilidad estadística.
- Las correlaciones posteriores reflejan relaciones reales y no efectos mecánicos de acumulación.
- El resultado final es un conjunto de series temporales sólidas, coherentes y metodológicamente defendibles.

Conclusiones y Expectativas

Este avance deja claro que el reto principal no era “tener datos”, sino hacer que los datos fueran interpretables y defendibles. La fuente original está en PDFs no estructurados y, además, reporta valores acumulados que por naturaleza crecen con el tiempo. Si uno analiza esos acumulados tal cual, las gráficas y los estadísticos se “inflan” y pueden aparentar tendencias o rupturas que en realidad son artefactos de cómo se reporta la información, no de lo que está ocurriendo epidemiológicamente. La conclusión más importante es que la transformación a incrementos semanales no es un detalle técnico, es el paso que vuelve posible hablar de dinámica real semana a semana.

El EDA fue crucial porque permitió detectar problemas que, si se ignoraban, hubieran contaminado cualquier conclusión posterior. Por ejemplo, una entidad duplicada por nomenclatura (Distrito Federal versus Ciudad de México) no es solo un tema de “nombres”: distorsiona porcentajes, distribuciones y comparaciones históricas. De forma similar, la presencia de una semana 53 y el desfase de semanas entre lo reportado y lo que realmente representa exige un ajuste temporal explícito. Sin ese ajuste, se comparan períodos incorrectos y se “rompe” la continuidad de la serie.

La limpieza de datos no se trató de embellecer el dataset, sino de restaurar coherencia lógica. En series derivadas de acumulados, los valores negativos en incrementos semanales son una alerta roja porque contradicen la naturaleza acumulativa. Detectarlos, aislarlos y corregirlos con reglas transparentes (y conservadoras) protege la integridad de la serie. Lo mismo aplica a picos extremos. Un outlier no siempre es “un dato interesante”; muchas veces es un error de captura o un cambio de criterio. Si no se controla, domina las escalas, deforma distribuciones y hace que el análisis termine explicando el error en lugar del fenómeno.

También se concluye que ser granular y observador es parte del método, no una preferencia personal. Revisar valores únicos, porcentajes por categoría, consistencia temporal, coherencia entre variables derivadas y acumuladas, y correlaciones antes y después del tratamiento permite demostrar que la limpieza no fue arbitraria. En este punto, el proyecto ya no depende de “confianza” en la fuente, sino de evidencia de consistencia interna.

Finalmente, la generación de gráficas apreciables fue más que presentación. Las visualizaciones funcionaron como pruebas diagnósticas del pipeline: ayudaron a revelar desfases, picos artificiales, colas excesivas por acumulación y cambios estructurales después de la transformación. Cuando una gráfica mejora después de un tratamiento, lo importante no es que “se vea bonita”, sino que la forma resultante tenga sentido con la lógica del proceso epidemiológico.

Con esta base, el puente hacia pronóstico se vuelve realista. El objetivo sería definir un esquema claro de entrenamiento y evaluación por padecimiento y sexo, con cortes temporales adecuados, baselines simples (por ejemplo, estacionalidad semanal, promedios móviles) y luego modelos más complejos solo si aportan valor. La expectativa no es “predecir por predecir”, sino demostrar que el pipeline produce señales estables y que el modelado aprende del fenómeno, no de errores de captura.

En resumen, lo que se espera para el siguiente paso es convertir este avance en una plataforma sólida: datos consistentes, reglas de limpieza justificadas, visualizaciones diagnósticas y un diseño de análisis temporal que permita pasar de describir a explicar y, después, a predecir con credibilidad.

Juan Carlos Pérez



“Para mí, este avance deja muy clara la importancia de realmente ver los datos antes de intentar sacar conclusiones. No basta con cargarlos y correr análisis automáticos; es necesario entender qué representan, cómo fueron construidos y en qué puntos pueden engañar. El proceso de análisis exploratorio y limpieza permitió identificar cuándo un patrón era real y cuándo era solo un efecto del acumulado o de un error de captura. Poder revisar distribuciones, tendencias, picos y coherencia temporal dio una visión completa del panorama y ayudó a decidir con criterio qué transformar, qué corregir y qué conservar. Este ejercicio refuerza que un buen análisis empieza por comprender a fondo los datos y su contexto.”

Javier Rebull



“En esta etapa nos ayudó mucho enfocarnos en la reproducibilidad del código y en la estructura del proyecto, ya que este entregable no solo trató de analizar datos, sino de demostrar cómo se llega a resultados confiables. Contar con un pipeline claro y replicable permitió que todos estuviéramos sincronizados en los cambios, pruebas y decisiones metodológicas, evitando interpretaciones distintas del mismo proceso. Además, este avance deja como takeaway central que un dataset limpio y actualizado es la base de cualquier análisis serio: la conversión de acumulados a incrementos, la corrección de semanas epidemiológicas y el tratamiento de valores atípicos muestran que pequeñas inconsistencias pueden cambiar por completo la lectura de los resultados. Este entregable refuerza que la calidad del análisis depende directamente de la calidad y trazabilidad del procesamiento previo, más incluso que de la complejidad de los modelos que se puedan aplicar después.”

Luis Sánchez



“La extracción de datos fue un primer paso esencial, pero este avance marca el punto en el que la información empieza a cobrar sentido. Pasar de tablas en PDF a datos estructurados, y de ahí a un EDA que revela patrones, anomalías y tendencias, muestra cómo los datos se transforman en conocimiento. Este proceso prepara directamente el camino para la predicción, porque ahora las series son coherentes y analizables. Personalmente, me entusiasma empezar a vincular estos datos con la nueva información poblacional, profundizar en los análisis y seguir desarrollando add-ons y mejoras para el proyecto. Se siente como el momento en que todas las piezas comienzan a conectarse y el proyecto entra en una fase más analítica y propositiva”

Referencias

- Alegría, M., NeMoyer, A., Falgàs Bagué, I., Wang, Y., & Alvarez, K. (2018). Social determinants of mental health: where we are and where we need to go. *Social Psychiatry and Psychiatric Epidemiology*. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC618118/>
- Alegría, M., NeMoyer, A., Falgàs Bagué, I., Wang, Y., & Alvarez, K. (2018). Social determinants of mental health: where we are and where we need to go. *Social Psychiatry and Psychiatric Epidemiology*. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC618118/>
- Consejo Nacional de Población. (2020). *Índices de marginación por entidad federativa y municipio 2020*. Gobierno de México. <https://www.gob.mx/conapo/documentos/indices-de-marginacion-2020-284372>
- Elliott, P., & Wartenberg, D. (2004). Spatial epidemiology: current approaches and future challenges. *Environmental Health Perspectives*, 112(9), 998–1006. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1247193/>
- Hernández-Sampieri, R., & Mendoza, C. (2023). *Metodología de la investigación: Las rutas de la investigación cuantitativa, cualitativa y mixta* (3.^a ed.). McGraw-Hill.
- Instituto Nacional de Estadística y Geografía. (s. f.). Áreas geográficas e indicadores territoriales. INEGI. <https://en.www.inegi.org.mx/app/areasgeograficas/>
- Instituto Nacional de Estadística y Geografía. (s. f.). Localidades urbanas y rurales en México. INEGI. https://cuentame.inegi.org.mx/descubre/poblacion/rural_urband/
- Instituto Nacional de Estadística y Geografía. (s. f.). Marco geoestadístico y datos poblacionales. INEGI. <https://www.inegi.org.mx/>
- Instituto Nacional de Estadística y Geografía. (s. f.). SCIAN: Sistema de Clasificación Industrial de América del Norte. <https://www.inegi.org.mx/scian/> (consultado el 12 de enero de 2026)
- Instituto Nacional de Estadística y Geografía (INEGI). (2024a). PxWeb API: Población por entidad federativa y sexo. <https://www.inegi.org.mx/app/tabulados/pxwebv2/>
- Organización Mundial de la Salud. (2022). Mental health: Strengthening our response. World Health Organization. <https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response>
- Organización Panamericana de la Salud. (2021). La salud mental en la región de las Américas. OPS/OMS. <https://www.paho.org/es/temas/salud-mental>
- Pérez-Hernández, R. (s. f.). Enfermedades neurológicas y trastornos mentales en México. 2014–2024 [Presentación de PowerPoint]. Tecnológico de Monterrey, Departamento de Posgrados.
- Secretaría de Salud. (s. f.). *Boletín Epidemiológico: Dirección General de Epidemiología*. Gobierno de México. <https://www.gob.mx/salud/acciones-y-programas/direccion-general-de-epidemiologia-boletin-epidemiologico>
- Secretaría de Salud. (s. f.). *Boletín Epidemiológico: Sistema Nacional de Vigilancia Epidemiológica / Sistema Único de Información*. Gobierno de México. <https://www.gob.mx/salud/documentos/boletinepidemiologico-sistema-nacional-de-vigilancia-epidemiologica-sistema-unico-de-informacion-387843>
- Secretaría de Salud. (s. f.). *Directorio de unidades de atención en salud mental*. Gobierno de México. <https://www.gob.mx/salud/acciones-y-programas/directorio-de-unidades-de-salud-mental>
- SINAVE. (s. f.). *Sistema Nacional de Vigilancia Epidemiológica (SINAVE)*. Secretaría de Salud. <https://www.gob.mx/salud/acciones-y-programas/sistema-nacional-de-vigilancia-epidemiologica>
- Visengeriyeva, L., Kammer, A., Bär, I., Kniesz, A., & Plöd, M. (2023). CRISP-ML(Q). *The ML lifecycle process*. MLOps. INNOQ. <https://ml-ops.org/content/crisp-ml>
- World Health Organization. (2014). *Social determinants of mental health*. <https://www.who.int/publications/i/item/9789241506809>
- World Health Organization. (2022). *World mental health report: Transforming mental health for all*. <https://www.who.int/publications/i/item/9789240049338>
- World Health Organization. (2025, February 14). *WHO releases 2025 update to the International Classification of Diseases (ICD-11)*. <https://www.who.int/news/item/14-02-2025-who-releases-2025-update-to-the-international-classification-of-diseases-%28icd-11%29>

ANEXO: Datos geográficos

Objetivo del módulo INEGI

Este módulo se diseñó para construir una base de datos estatal sólida y comparable para México, integrando información poblacional y territorial en un solo dataset. Su objetivo principal es generar variables estructurales que sirvan como contexto para análisis epidemiológicos posteriores, particularmente en el estudio de la salud mental. La intención no es únicamente describir a las entidades federativas, sino preparar una capa analítica que permita entender cómo la concentración poblacional, la urbanización y el territorio influyen en los patrones observados en los datos de salud.

El módulo fue incorporado al repositorio del proyecto dentro de `src/extraccion/inegi.py`, con el objetivo de facilitar su reproducción, mantenimiento y ejecución consistente dentro del pipeline general del proyecto.

Fuentes de datos

Para garantizar consistencia, transparencia y reproducibilidad, se utilizan exclusivamente fuentes oficiales del Instituto Nacional de Estadística y Geografía (INEGI). La información de población por entidad federativa y sexo se obtiene a través de la API PxWeb, en formato JSON-STAT, lo que permite descargas programáticas y evita procesos manuales propensos a error (INEGI, 2024a).

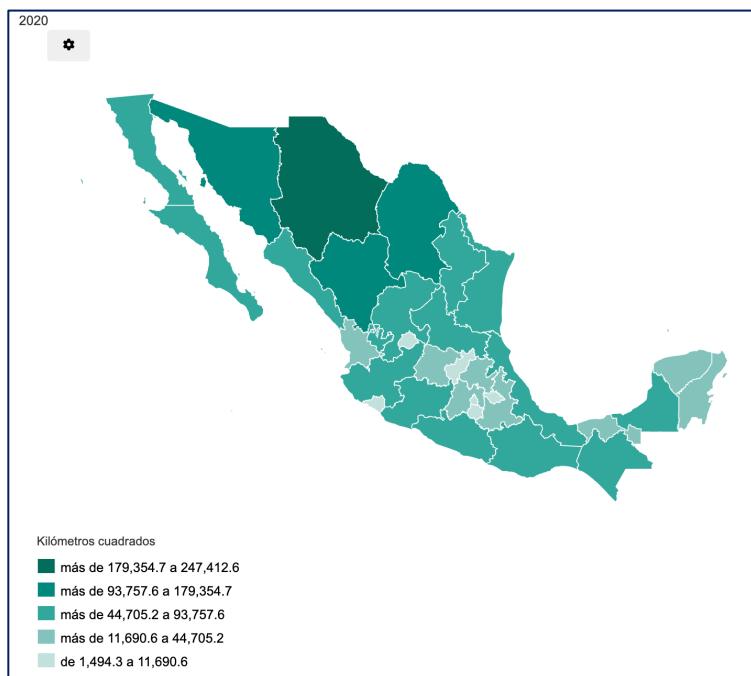


Imagen 28 - Extensión territorial de México en kilómetros cuadrados.

Fuente: <https://www.inegi.org.mx/app/areasgeograficas/#tabMCcollapse-Indicadores>

Adicionalmente, la información de extensión territorial por entidad federativa se obtiene mediante los servicios de indicadores y áreas geográficas de INEGI, los cuales proporcionan datos estandarizados sobre superficie estatal (INEGI, 2024b).

La combinación de ambas fuentes responde a una necesidad analítica clara: la población total por sí sola no permite analizar fenómenos de concentración humana, mientras que la superficie territorial sin población carece de significado epidemiológico. Su integración permite calcular densidad poblacional y contextualizar adecuadamente las diferencias estructurales entre entidades.

Pipeline de procesamiento

El pipeline inicia con la consulta a la API PxWeb para descargar la tabla de población por entidad federativa, sexo y periodo. El JSON-STAT recibido se transforma a un DataFrame tabular respetando el orden oficial de las dimensiones definidas por INEGI, asegurando que la estructura original del dato se conserve.

Posteriormente, se filtra el grupo de edad correspondiente al total poblacional y el dataset se reorganiza a formato wide, utilizando el sexo como columnas. En esta etapa se incluye una validación explícita para comprobar que la suma de hombres y mujeres coincide con la población total por entidad. Esta verificación temprana permite detectar inconsistencias estructurales antes de avanzar en el análisis.

Una vez validada la información, se filtra el periodo 2020 para trabajar con un corte transversal único y consistente. En paralelo, se descargan los datos de superficie territorial, se limpian y se convierten a valores numéricos. Finalmente, ambos conjuntos de datos se combinan mediante la entidad federativa, generando un DataFrame único, ordenado y sin valores faltantes, que sirve como base para el análisis exploratorio y la generación de variables derivadas.

De manera puntual, el pipeline recorre los siguientes pasos:

- [1] Descarga de datos de población por entidad y sexo desde la API PxWeb de INEGI.
- [2] Conversión del JSON-STAT a un DataFrame tabular.
- [3] Limpieza básica y conversión de valores poblacionales a formato numérico.
- [4] Filtrado para conservar solo la población total, eliminando desagregaciones por edad.
- [5] Transformación del dataset a formato wide con columnas de Hombres, Mujeres y Total.
- [6] Validación de consistencia (Hombres + Mujeres = Total).
- [7] Selección del año 2020 como corte transversal de análisis.
- [8] Descarga y limpieza de datos de superficie territorial por entidad federativa.
- [9] Integración de población y superficie en un solo DataFrame.
- [10] Cálculo de variables derivadas y categorizaciones.
- [11] Cálculo de densidad poblacional como población total entre superficie territorial.
 - a. Cálculo del ratio hombres/mujeres por entidad.
 - b. Clasificación del ratio H/M en categorías interpretables.
 - c. Clasificación del tamaño poblacional mediante rangos fijos.

- d. Clasificación del tamaño poblacional mediante percentiles.
- e. Clasificación de la superficie territorial por percentiles.
- f. Clasificación de la densidad poblacional por percentiles.

[12] Asignación de la regionalización socio-urbana en salud mental.

- a. Definición de una regionalización funcional con enfoque socio-urbano.
- b. Agrupación de entidades en cuatro categorías de contexto estructural.
- c. Asignación explícita de región a cada entidad federativa.
- d. Verificación de entidades sin región asignada.

[13] Ejecución del análisis exploratorio de datos (EDA).

[14] Creación de gráficas de soporte.

Variables derivadas y categorizaciones

A partir del DataFrame base se calculan variables que permiten caracterizar estructuralmente a cada entidad federativa. La densidad poblacional se calcula como la razón entre población total y superficie territorial, ya que esta variable captura el grado de concentración humana y es ampliamente utilizada en epidemiología espacial para explicar variaciones territoriales en salud (Elliott & Wartenberg, 2004).

Asimismo, se calcula el ratio hombres/mujeres para evaluar posibles asimetrías demográficas. Este ratio se categoriza en tres grupos interpretables: mayormente mujeres, balanceado y mayormente hombres, con el fin de facilitar su análisis descriptivo.

El tamaño poblacional se clasifica de dos maneras complementarias. Por un lado, se utilizan rangos fijos (0-1M, 1-3M, 3-6M y 6M+), que permiten una interpretación sustantiva directa. Por otro lado, se generan grupos por percentiles, que fuerzan comparabilidad estadística entre entidades y resultan útiles para análisis exploratorios y modelos supervisados.

De forma análoga, la superficie territorial y la densidad poblacional se agrupan por percentiles, permitiendo contrastes balanceados entre entidades con características estructuralmente distintas.

Regionalización en salud mental

Además de las variables demográficas y territoriales, se incorpora una regionalización específica para el análisis de salud mental. Esta regionalización no sigue divisiones administrativas genéricas, sino que adopta un enfoque socio-urbano y estructural. Las categorías utilizadas son: metropolitana alta, urbana media, rural y dispersa, y Suroeste estructuralmente vulnerable.

Este enfoque está alineado con la literatura en determinantes sociales de la salud mental, que señala que los patrones observados dependen más del acceso a servicios, la urbanización, la desigualdad estructural y la capacidad diagnóstica que de la latitud o el clima (Alegria et al., 2018; WHO, 2014; WHO, 2022).

El objetivo de esta regionalización es capturar diferencias reales en detección, subregistro y acceso a atención en salud mental. Utilizar divisiones administrativas clásicas como Norte, Centro y Sur

habría mezclado realidades sociales incompatibles y debilitado la inferencia epidemiológica. Por ello, esta regionalización se incorpora explícitamente como una variable categórica dentro del dataset.

Para el análisis epidemiológico de los trastornos mentales en México se realizó una regionalización socio-urbana de las 32 entidades federativas. Esta clasificación no busca representar regiones administrativas tradicionales, sino agrupar estados con características estructurales similares que influyen directamente en la detección, registro y atención de los trastornos mentales.

La decisión de utilizar este enfoque se basa en que, a diferencia de otras enfermedades, los trastornos mentales están fuertemente mediados por factores sociales y contextuales, como el grado de urbanización, el acceso a servicios especializados y la vulnerabilidad socioeconómica. En este tipo de padecimientos, los datos disponibles no reflejan únicamente la prevalencia real, sino también la capacidad del sistema de salud para diagnosticar y registrar los casos, lo cual varía considerablemente entre regiones del país (OMS, 2022).

A partir de ello, se definieron cuatro categorías analíticas: entidades altamente urbanas y metropolitanas, entidades urbanas medias, entidades predominantemente rurales y dispersas, y un bloque Sur-Sureste caracterizado por vulnerabilidad estructural. La asignación de los estados a estas categorías se realizó como un primer análisis exploratorio, apoyado en conocimiento general sobre la distribución de las principales zonas metropolitanas del país, patrones ampliamente documentados de marginación y dispersión poblacional, y la persistencia de brechas socioeconómicas históricas en el Sur-Sureste. A continuación se presenta en una tabla la clasificación asignada:

Metropolitana alta	Urbana media	Rural / dispersa	Sur-Sureste vulnerable
Ciudad de México Nuevo León Jalisco	Aguascalientes Baja California Baja California Sur Chihuahua Coahuila de Zaragoza Colima Durango Guanajuato Morelos Querétaro San Luis Potosí Sinaloa Sonora Tamaulipas Zacatecas	Guerrero Hidalgo Michoacán de Ocampo Nayarit Puebla Tlaxcala Veracruz de Ignacio de la Llave	Campeche Chiapas Oaxaca Tabasco Yucatán Quintana Roo

Este procedimiento corresponde a una clasificación heurística, cuyo objetivo es mejorar la interpretabilidad de los resultados en etapas iniciales del análisis, más que establecer una

regionalización definitiva. El enfoque es consistente con la literatura que señala que la urbanización, la desigualdad social y el acceso a servicios de salud mental son determinantes clave tanto del riesgo como del subdiagnóstico en salud mental (OPS, 2021).

Las dimensiones conceptuales que sustentan esta clasificación se apoyan en fuentes institucionales reconocidas. La distinción urbano-rural se fundamenta en los criterios de población utilizados por el Instituto Nacional de Estadística y Geografía (INEGI, 2020). La noción de vulnerabilidad estructural se alinea con los índices de marginación desarrollados por el Consejo Nacional de Población (CONAPO, 2020). Finalmente, el énfasis en el acceso a servicios se respalda en la evidencia de la distribución desigual de la infraestructura de atención en salud mental documentada por la Secretaría de Salud y organismos internacionales (Secretaría de Salud, 2023; OMS, 2022).

Es importante subrayar que esta regionalización no pretende medir prevalencia real, sino capturar patrones diferenciales de detección, registro y acceso a atención, los cuales son centrales para la interpretación de los datos epidemiológicos disponibles. En este sentido, la clasificación es adecuada como punto de partida para análisis descriptivos y comparativos.

Para un análisis más robusto y metodológicamente blindado, esta regionalización puede refinarse en etapas posteriores mediante un enfoque data-driven, incorporando indicadores cuantitativos de urbanización (INEGI), marginación social (CONAPO) y oferta de servicios de salud mental por entidad (Secretaría de Salud). A partir de estos insumos, sería posible construir clasificaciones basadas en percentiles o técnicas de agrupamiento estadístico, reduciendo la dependencia del juicio experto y fortaleciendo la validez analítica para estudios de mayor profundidad o publicaciones académicas.

Exploratory Data Analysis (EDA) breve

Una vez construido el dataset final, se realiza un análisis exploratorio de datos breve para evaluar su calidad y estructura. El EDA inicia con la verificación de valores faltantes y la confirmación de consistencia general. Posteriormente, se presentan vistas generales del dataset y estadísticos descriptivos de las principales variables numéricas.

Vista general (head)													
Entidad federativa	Superficie_km2	Hombres	Mujeres	Total	region_salud_menta	ratio_h_m	ratio_h_m_cat	tamano_poblacional_predefinido	tamano_poblacional_grupo_percentil	densidad_poblacion	extension_territorial_percentil	densidad_poblacion_percentil	
Aguascalientes	5,615.70	696,683	728,924	1,425,607	Urbana media	0.96	Mayormente mujeres	1-3M	Populación baja	253.86	Territorio pequeño	Alta	
Baja California	71,458	1,986,589	1,868,431	3,769,028	Urbana media	1.02	Mayormente hombres	3-6M	Media-alta	52.75	Medio-grande	Media-baja	
Baja California Sur	73,999.40	405,879	392,568	798,447	Urbana media	1.03	Mayormente hombres	0-1M	Populación baja	10.80	Medio-grande	Baja	
Campeche	57,484.90	456,939	471,424	928,363	Sur-Sureste vulnerable	0.97	Mayormente mujeres	0-1M	Populación baja	16.15	Medio-pequeño	Baja	
Chiapas	73,311	2,705,947	2,837,881	5,543,828	Centro-vuln	0.95	Mayormente mujeres	3-6M	Alta	75.62	Medio-grande	Media-alta	
Chihuahua	247,412.60	1,853,822	1,888,047	3,741,869	Urbana media	0.98	Mayormente mujeres	3-6M	Media-alta	15.12	Grande	Baja	
Ciudad de México	1,494.30	4,484,927	4,885,017	9,289,944	Metropolitana alta	0.92	Mayormente mujeres	6M+	Alta	6,163.38	Territorio pequeño	Alta	
Coahuila de Zaragoza	151,594.80	1,563,669	1,583,102	3,146,771	Urbana media	0.99	Mayormente mujeres	3-6M	Media-alta	28.76	Grande	Baja	

Descriptivos numéricos									
variable	count	mean	std	min	25%	50%	75%	max	
Hombres	32	1,921,043.44	1,585,770.69	360,622	928,800.75	1,488,097	2,406,242.50	8,251,295	
Mujeres	32	2,016,894.81	1,692,778.43	370,769	926,139.50	1,557,615	2,541,349	8,741,123	
Total	32	3,937,938.25	3,278,009.09	731,391	1,851,651.25	3,054,892	4,947,591.50	16,992,418	
Superficie_km2	32	61,270.21	53,819.04	1,494.30	24,136.12	58,041.80	74,250.88	247,412.60	
densidad_poblacion	32	309.68	1,078.71	10.80	43.36	67.17	159.02	6,163.38	
ratio_h_m	32	0.96	0.03	0.92	0.94	0.96	0.98	1.03	

Imagen 29 - Vista exploratoria numérica del DataFrame obtenido

Se identifican entidades extremas mediante rankings de población total, densidad poblacional y superficie territorial. Asimismo, se analizan conteos por categorías, incluyendo tamaño poblacional, percentiles, ratio hombres/mujeres y regiones de salud mental.

Top 5: Población total		Conteo: Región salud mental		Conteo: Extensión territorial (percentiles)	
Entidad federativa	Total	Categoría	Conteo	Categoría	Conteo
México	16,992,418	Urbana media	15	Territorio pequeño	8
Ciudad de México	9,209,944	Rural / dispersa	7	Medio-pequeño	8
Jalisco	8,348,151	Sur-Sureste vulnerable	6	Medio-grande	8
Veracruz de Ignacio de la Llave	8,062,579	Metropolitana alta	4	Grande	8
Puebla	6,583,278				
Conteo: Tamaño poblacional (rangos fijos)					
Entidad federativa	densidad_poblacion	Categoría	Conteo	Categoría	Conteo
Ciudad de México	6,163.38	1-3M	12	Baja	8
México	760.23	3-6M	11	Media-baja	8
Morelos	404.09	6M+	6	Media-alta	8
Tlaxcala	336.03	0-1M	3	Alta	8
Aguascalientes	253.86				
Conteo: Tamaño poblacional (percentiles)					
Entidad federativa	Superficie_km2	Categoría	Conteo	Categoría	Conteo
Chihuahua	247,413	Población baja	8	Mayormente mujeres	27
Sonora	179,355	Media-baja	8	Mayormente hombres	3
Coahuila de Zaragoza	151,595	Media-alta	8	Balanceado	2
Durango	123,364	Alta	8		
Oaxaca	93,758				

Imagen 30 - Valores para conteo de variables categóricas del DataFrame

En términos visuales, se generan gráficas de barras ordenadas por entidad y coloreadas según distintas categorizaciones, lo que permite comparar patrones de forma clara. También se construyen boxplots para evaluar la dispersión de variables clave, utilizando escala logarítmica cuando la asimetría de los datos lo requiere.

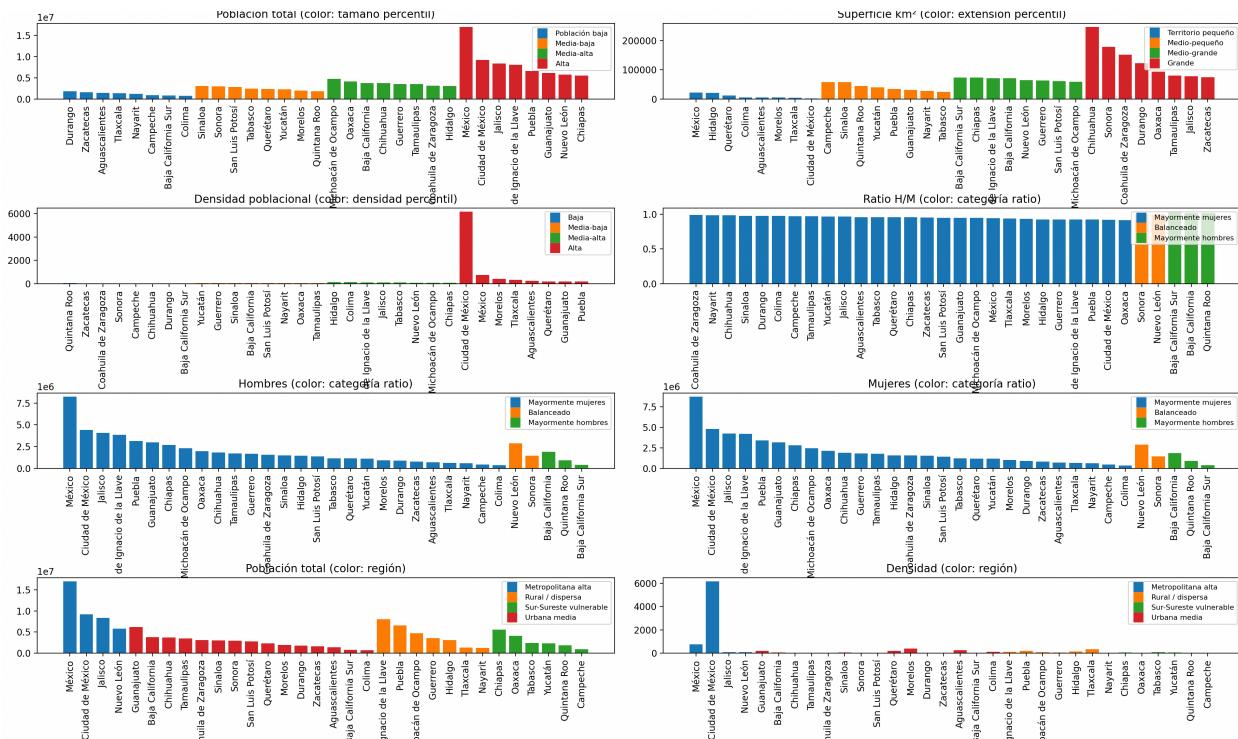
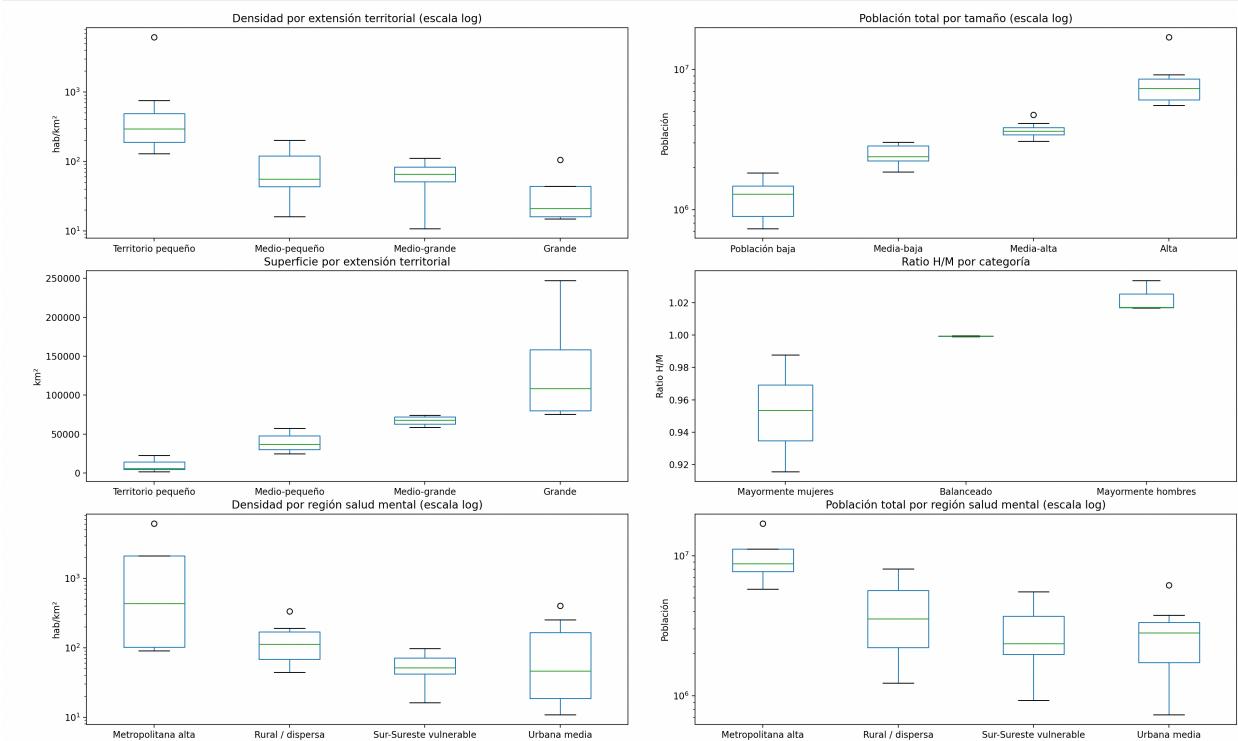


Imagen 31 - Gráficas de barras categorizadas



Highlights del análisis inicial

Nota: este análisis debe considerarse exploratorio e inicial. Los hallazgos se profundizarán, validarán y ajustarán conforme avance la metodología CRISP-DM, especialmente en las fases de comprensión de los datos, preparación y modelado.

- El dataset presenta integridad. No se detectan valores faltantes en las variables clave, lo que permite análisis comparables sin necesidad de imputación artificial.
- La población total y la densidad poblacional muestran distribuciones altamente asimétricas. Entidades como la Ciudad de México actúan como outliers estructurales y no deben tratarse como observaciones típicas.
- La densidad poblacional es el principal factor diferenciador entre entidades. Explica mejor las diferencias territoriales que la población total o la superficie por separado.
- La superficie territorial no guarda una relación directa con la población. Estados extensos no son necesariamente los más poblados.
- El ratio hombres/mujeres es estable entre entidades, con baja variabilidad, lo que indica que el sexo no es un factor estructural diferenciador a nivel estatal. Quizás en análisis posteriores e integración con el dataset de padecimientos se analice más a detalle.
- Las categorizaciones por percentiles generan grupos balanceados, facilitando comparaciones y futuros análisis estadísticos o de aprendizaje automático.
- La regionalización socio-urbana en salud mental muestra heterogeneidad entre entidades y resulta coherente con diferencias estructurales observadas en densidad y tamaño poblacional.
- En conjunto, estos hallazgos indican que el dataset parece estar adecuadamente preparado para fases posteriores de análisis, y que la normalización por densidad poblacional será un elemento clave en estudios epidemiológicos y modelos predictivos de salud mental.

Clasificación territorial sugerida

Con base en el análisis exploratorio y en los patrones observados en los datos, se proponen tres ejes de clasificación territorial como los más sólidos y metodológicamente defendibles para análisis posteriores.

- **Densidad poblacional:** Es la clasificación más fuerte. La variabilidad observada es alta y explica mejor las diferencias estructurales entre entidades que la población total o la superficie. La densidad captura el grado de concentración humana, factor crítico en análisis epidemiológicos y de salud mental. Además, su uso es consistente con la presencia de outliers urbanos como la Ciudad de México, que requieren tratamiento analítico diferenciado.
- **Tamaño poblacional por percentiles:** Esta clasificación permite comparaciones balanceadas entre entidades, evitando que estados muy grandes dominen el análisis. Al forzar grupos equitativos, facilita la estratificación en modelos estadísticos y de aprendizaje automático. A diferencia de los rangos fijos, los percentiles maximizan el contraste relativo entre entidades.
- **Región socio-urbana de salud mental:** Esta categorización integra dimensiones demográficas y territoriales en una lógica interpretativa coherente. Los resultados muestran que

las regiones no son homogéneas en términos de densidad ni tamaño poblacional, lo que sugiere necesidades diferenciadas de análisis y política pública. Su fortaleza radica en que traduce variables cuantitativas en un marco conceptual útil para análisis longitudinales y epidemiológicos.

Estas tres clasificaciones capturan concentración, escala y contexto territorial, respectivamente. En conjunto, forman una base sólida para estratificación analítica, control estructural y modelado posterior dentro del enfoque CRISP-DM.

ANEXO: Pipeline de Preprocesamiento de DATOS

La Figura X presenta el pipeline de datos diseñado para el proyecto EpiForecast-MX, el cual describe el flujo completo desde la obtención de los boletines epidemiológicos del SINAVE hasta la generación de datasets listos para modelado predictivo. Este proceso comprende siete etapas secuenciales, extracción, análisis exploratorio, limpieza, transformación, consolidación y modelado, que permiten convertir información no estructurada contenida en archivos PDF en series de tiempo coherentes para los tres padecimientos de interés: Depresión, Enfermedad de Parkinson y Enfermedad de Alzheimer. El pipeline es completamente reproducible mediante un flujo automatizado con Makefile y versionado de datos con DVC, garantizando trazabilidad y consistencia en cada ejecución.

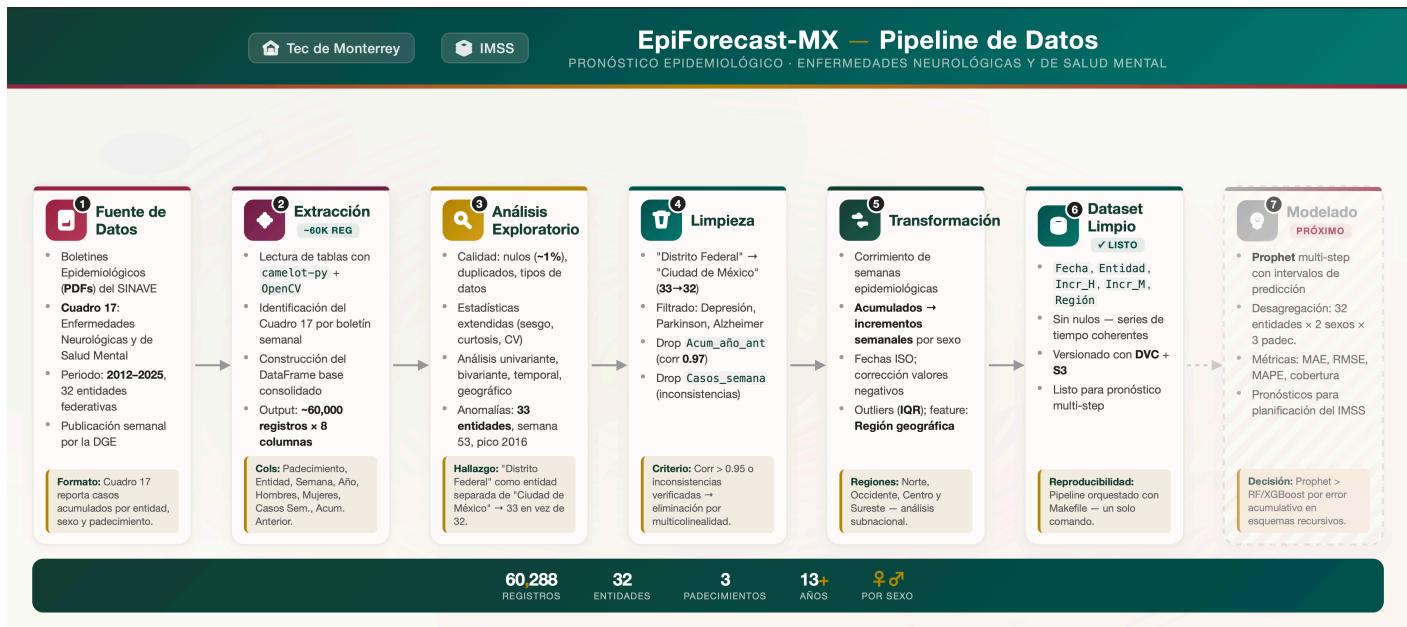


Imagen 33. Pipeline de Transformación de Datos.

Consultar la versión Web en: https://proyectointegrador.org/pipeline_diagrameda

ANEXO: Repositorio GitHub del proyecto

Como pilar de transparencia y reproducibilidad bajo el marco de la metodología CRISP-ML(Q), la totalidad del código fuente, los activos de datos procesados y el historial de versiones de este análisis se encuentran centralizados en nuestro repositorio oficial. Este espacio no solo sirve como el motor de desarrollo de EpiForecast-MX, sino también como una plataforma de colaboración abierta donde se documenta cada fase del ciclo de vida del proyecto. Puedes consultar el repositorio completo, así como los avances técnicos y la estructura del pipeline, en el siguiente enlace: GitHub | IntegradorIMSS2o26Team01/EpiForecast-MX.

<https://github.com/IntegradorIMSS2o26Team01/EpiForecast-MX>

ANEXO: Pagina Web del Proyecto

Hemos diseñado una página web institucional donde se pueden consultar los objetivos estratégicos, el stack tecnológico y los datos fundamentales del proyecto, la cual está disponible en el siguiente enlace: <https://proyectointegrador.org/>.