#### Estudiante Héctor David Barral Padilla





# Objetivo

Aplicar K-Means usando únicamente **Edad de la víctima (X)** y **Edad del agresor (Y)**, en el subconjunto **solo masculinos**, para identificar y explicar **un clúster interesante**.

# Metodología (resumen)

- Variables: VIC\_EDAD y AGR\_EDAD.
- Subconjunto: registros donde víctima y agresor son masculinos.
- Preprocesamiento mínimo: remoción de NA y estandarización (z-score).
- Algoritmo: K-Means con K=3 y nstart=25.
- Visualización: dispersión Edad víctima (eje X) vs Edad agresor (eje Y), coloreado por clúster; centroides en negro.

# Estructura general observada

- 1. Clúster 1 (rojo): víctimas jóvenes (~20–35) con agresores ~30–45.
- 2. **Clúster 2 (verde)**: víctimas mayores (~50–70) con agresores ~30–40 (víctima mayor que el agresor).
- 3. Clúster 3 (azul): agresores muy mayores (≈70–100), con víctimas ~35–55. Se observa una banda en ~100 años en el eje Y.

### Clúster seleccionado (interesante): Clúster 3 (azul)

### Descripción

- Concentración de casos donde el agresor está en un rango muy alto de edad (≈70–100).
- Las víctimas en este clúster tienden a ubicarse entre ~35 y 55 años.
- La banda horizontal en 100 sugiere que parte de estos valores podrían ser top-code o códigos de dato (p. ej., 99/100 como "no especificado" o "≥100").

## Explicación e hipótesis

- Patrón real: existe un subconjunto en el que los agresores son de edad avanzada; podría corresponder a relaciones familiares intergeneracionales (p. ej., violencia intrafamiliar donde el agresor es un pariente de mayor edad).
- 2. **Efecto de calidad de datos**: la **acumulación en 99/100** puede inflar la media del agresor y "fabricar" parte del clúster. Es decir, el algoritmo puede estar agrupando **tanto casos** reales de agresores mayores como registros codificados con 99/100.

#### Por qué es relevante

- Señala un grupo atípico respecto a los otros clústeres
- Combina posible fenómeno social y alerta de calidad de datos: ambos son hallazgos accionables (diseño de prevención específica y depuración/recodificación).





## Implicaciones prácticas

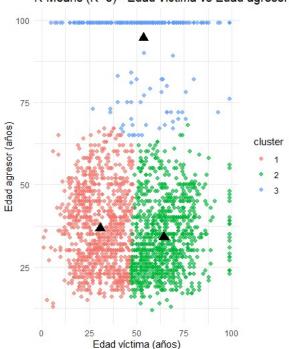
- **Prevención y atención focalizada**: si se confirma que hay agresores de edad avanzada, se requieren estrategias diferenciadas (p. ej., intervención familiar, apoyo a cuidadores, rutas de denuncia accesibles).
- **Gestión de datos**: revisar el **diccionario de datos** y **recodificar** 99/100 (o "no sabe/no responde"), repetir el K-Means sin esos códigos para verificar la **persistencia** del clúster.

# Próximos pasos sugeridos

- Validar si 99/100 son códigos (no edades reales).
- Re-entrenar K-Means excluyendo/recodificando esos valores; comparar centroides y tamaño del clúster 3.
- Añadir, en un análisis posterior, variables contextuales (sin sesionar filtros prohibidos) solo para **interpretación**: relación con el agresor, convivencia, etc.

#### Conclusión

El **Clúster 3 (agresores muy mayores)** es el más interesante porque concentra un extremo etario **inusual** y expone un **posible sesgo por codificación**. Esto lo convierte en un objetivo prioritario para **verificación de datos** y para el diseño de **acciones específicas** de prevención y atención si el patrón se confirma tras la depuración.



K-Means (K=3) - Edad víctima vs Edad agresor (