Big Data - Analisis Exploratorio de Datos

Utilizando la herramienta [Orange Data Mining](https://orangedatamining.com/download/), importe el archivo de datos hogares.csv y elabore un informe con las respuestas a las preguntas indicadas en esta guía.

Hogares.csv contiene datos ficticios de 100 hogares encuestados.

# Condiciones de entrega

* Entrega Grupal: Deben entregarlo por Github con el nombre del grupo
* Plazo máximo: hasta la siguiente clase.

# Diccionario de Datos

HogarID: Identificador único del hogar, entero que va de 1 a 100.

IngresoMensual: Ingresos totales del hogar por mes, en pesos argentinos (ARS). Representa la suma de todos los ingresos de los miembros del hogar.

GastoMensual: Gastos totales del hogar por mes, en pesos argentinos (ARS). Incluye gastos fijos (como alquiler o hipoteca, servicios, etc.) y gastos variables (como alimentación, transporte, ocio).

NumPersonas: Número de personas que conforman el hogar. Valor entero que indica la cantidad de individuos viviendo en el hogar.

TipoVivienda: Tipo de estructura habitacional donde reside el hogar. Puede ser "Casa" o "Departamento", reflejando las diferentes modalidades de vivienda.

ZonaVivienda: Zona geográfica donde se ubica el hogar. Las categorías incluyen "Urbana", "Suburbana" y "Rural", indicando el contexto geográfico y posiblemente socioeconómico de la vivienda.

EdadPromedioHogar: Edad promedio de los miembros del hogar, en años. Esta variable es un promedio de las edades de todos los individuos que conforman el hogar.

# Guía de Preguntas

1. Análisis de Tendencia Central y Dispersión:

Calcule las medidas de tendencia central (media, mediana) y dispersión (varianza, desviación estándar) para las variables IngresoMensual, GastoMensual, y EdadPromedioHogar. ¿Los valores encontrados son consistentes con lo que esperarías en un contexto económico real?

Rta: Se calcularon las medidas de tendencia central y dispersión para las variables IngresoMensual, GastoMensual y EdadPromedioHogar. En todos los casos, la media y la mediana presentan valores similares, lo cual indica que las distribuciones son aproximadamente simétricas. Sin embargo, la moda en las tres variables es notablemente inferior a la mediana, lo que sugiere la existencia de un grupo de hogares con valores significativamente más bajos que el promedio, posiblemente correspondientes a sectores más vulnerables.

En cuanto a la dispersión, los valores de desviación relativa indican una variabilidad moderada. Esta variabilidad es esperable en contextos económicos reales, donde los hogares presentan diferentes niveles de ingresos, gastos y composición etaria. La variable EdadPromedioHogar, con una media de 40 años y una mediana de 38, refleja un rango etario coherente con el de hogares típicos en Argentina.

Los niveles promedio de ingreso mensual (alrededor de $91.000) y gasto mensual (alrededor de $62.000) son consistentes con lo que se esperaría en hogares de clase media en zonas urbanas y suburbanas del país. En conjunto, los valores analizados resultan razonables y representativos de una distribución socioeconómica diversa y realista.

1. Verificación de Máximos y Mínimos:

Identifique los valores máximos y mínimos para IngresoMensual, GastoMensual, y EdadPromedioHogar. ¿Estos valores son realistas? Discuta cualquier valor que parezca inusual o fuera de lo esperado.

Rta: Los valores extremos para ingreso y gasto mensual (mínimos de $40.607 y $20.623; máximos de $148.558 y $127.691 respectivamente) son coherentes con una población diversa de hogares y no presentan anomalías evidentes. En cambio, la variable EdadPromedioHogar muestra un valor mínimo de –5, que no es posible y representa un error de datos. El valor máximo de 120 años, aunque elevado, podría considerarse válido en casos excepcionales.

1. Relación entre Variables:

Explore la relación entre el EdadPromedioHogar y otras variables numéricas como IngresoMensual y GastoMensual. ¿Hay alguna tendencia visible que indique cómo la edad promedio del hogar afecta los ingresos y gastos?

Rta: Los valores extremos para ingreso y gasto mensual (mínimos de $40.607 y $20.623; máximos de $148.558 y $127.691 respectivamente) son coherentes con una población diversa de hogares y no presentan anomalías evidentes. En cambio, la variable EdadPromedioHogar muestra un valor mínimo de –5, que no es posible y representa un error de datos. El valor máximo de 120 años, aunque elevado, podría considerarse válido en casos excepcionales.

En resumen, los valores extremos son en general razonables, salvo por el dato erróneo en edad promedio, que debería corregirse antes de continuar el análisis.

1. Análisis de Outliers:

Utilizando técnicas gráficas, identificar datos atípicos en las variables numéricas. Describa por qué estos puntos podrían considerarse atípicos y qué factores podrían explicar estos valores.

Rta: Los boxplots muestran outliers en IngresoMensual y GastoMensual, con algunos hogares que superan los $120.000 de ingreso y los $80.000 de gasto mensual. Estos valores pueden representar casos reales de alto poder adquisitivo. En EdadPromedioHogar se detecta un outlier extremo cercano a 120 años, que probablemente sea un error de carga y debe corregirse. En NumPersonas no se observan valores atípicos.

1. Comparación entre Grupos:

Analice cómo varía la EdadPromedioHogar entre diferentes tipos de vivienda (Casa, Departamento) y zonas de vivienda (Urbana, Suburbana, Rural). ¿Existen diferencias significativas que puedan sugerir patrones demográficos?

Rta: Los gráficos muestran cierta tendencia, pero las diferencias entre grupos no parecen ser marcadamente significativas. En cuanto al tipo de vivienda, no se observa una separación clara: tanto casas como departamentos presentan una distribución amplia de edades promedio, aunque los departamentos concentran levemente más hogares de edad mayor. En relación a la zona de vivienda, se distinguen grupos definidos (urbana, suburbana y rural), pero dentro de cada uno la edad promedio también varía bastante.

1. Correlaciones:

Investigue la correlación entre las variables numéricas. ¿Qué relaciones se pueden identificar como relevantes y cómo se interpretan?  
  
rta: La única relación numéricamente relevante es entre IngresoMensual y GastoMensual, con una correlación de +0,895, lo que indica una asociación fuerte y positiva: a mayor ingreso, mayor gasto. Esta relación es esperable en contextos económicos reales, ya que los hogares con mayores ingresos tienden también a tener un nivel de consumo más alto. Las demás correlaciones entre variables numéricas son muy bajas (cercanas a cero), por lo que no se consideran significativas ni relevantes para el análisis.