Big Data - Análisis Exploratorio de Datos

Alumnas: Biloni María José, Duaygues Renata, Minestrelli Josefina

Preguntas

1. Análisis de Tendencia Central y Dispersión:

Calcule las medidas de tendencia central (media, mediana) y dispersión (varianza, desviación estándar) para las variables **Ingreso Mensual, Gasto Mensual, y Edad Promedio Hogar.** ¿Los valores encontrados son consistentes con lo que esperarías en un contexto económico real?

1) *Ingreso Mensual*
    - Media → 91719.8818
    - Mediana → 91055.67
    - Desviación estándar→ 32,723.83
    - Varianza → 1,070,849,000

La media y mediana son muy cercanas, lo que sugiere una distribución relativamente simétrica de los ingresos, lo cual es coherente en muchas poblaciones. La desviación estándar de 32,723.83 indica una alta variabilidad en los ingresos mensuales, lo cual es esperable en contextos reales donde los ingresos pueden variar significativamente debido a factores como el tipo de empleo, nivel educativo, y región geográfica. Este nivel de dispersión es consistente con los datos económicos observados en la vida real.

2) *Gasto Mensual*
    - Media → 62151.3
    - Mediana → 62352.3
    - Desviación estándar→ 25,105.21
    - Varianza → 630,271,600

Los valores de media y mediana son bastante cercanos, lo cual también indica una distribución bastante simétrica de los gastos. La desviación estándar de 25,105.21 refleja una considerable variabilidad en los gastos mensuales de los hogares, lo cual es razonable dado que los gastos pueden variar según el tamaño del hogar, ingresos disponibles, y hábitos de consumo.

3) *Edad Promedio Hogar*

- Media → 40.192
- Mediana → 38
- Desviación estándar→ 16.64
- Varianza → 276.86

La media y mediana sugieren que la edad promedio de los hogares está en torno a los 40 años, lo cual es realista y típico en muchas poblaciones. La desviación estándar de 16.64 años muestra que hay una amplia dispersión en las edades, lo cual es normal dado que los hogares pueden estar compuestos por miembros de diferentes generaciones.

2. 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

1) *Ingreso Mensual*
    - Máximo → 148557.56
    - Mínimo → 40607.43

Estos valores no son realistas ya que el ingreso mínimo vital y móvil en Argentina, según la página del gobierno, para agosto del 2024 es de $262.432,93. Partiendo de esta base el máximo y el mínimo están muy desactualizados, el máximo ni siquiera llega al valor del salario mínimo hoy en día.

2) Gasto Mensual
    - Máximo → 127691
    - Mínimo → 20623.9

Nuevamente estos valores difieren completamente de la realidad. El gasto mensual mínimo fue de $521.601,89 en julio de 2024 para no caer en la indigencia.

3) Edad Promedio Hogar
    - Máximo → 120
    - Mínimo → -5

Las edades no son realistas, el máximo nunca podría ser 120 y la edad nunca es negativa.

3. Relación entre Variables:

Explore la relación entre la Edad Promedio Hogar y otras variables numéricas como Ingreso Mensual y Gasto Mensual. ¿Hay alguna tendencia visible que indique cómo la edad promedio del hogar afecta los ingresos y gastos?

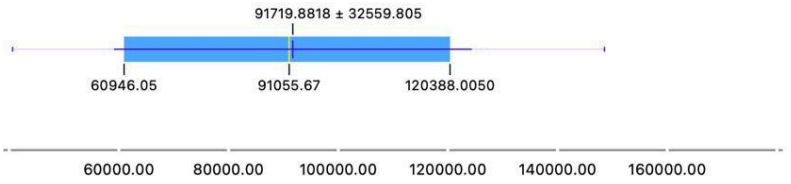
Edad Promedio del Hogar e Ingreso Mensual: La correlación es muy débil y negativa (-0.010). Esto indica que prácticamente no hay una relación lineal entre la edad promedio del hogar y el ingreso mensual.
Edad Promedio del Hogar y Gasto Mensual: La correlación es muy débil y negativa (-0.051). Similar al caso del ingreso, la edad del hogar tampoco parece ser un buen predictor del gasto mensual.

En el Scatter Plot tampoco se observa una relación entre estas variables, ya que los puntos están muy dispersos en ambos casos.

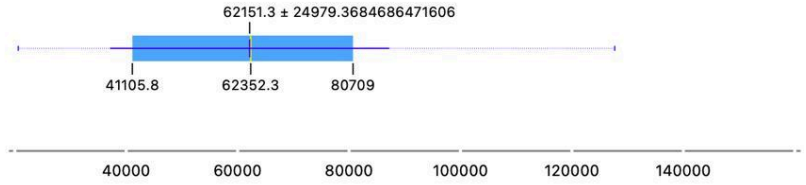
4. Análisis de Outliers:

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

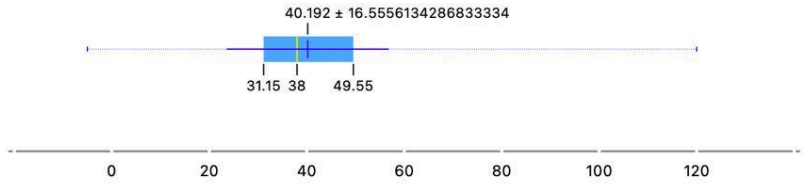
Ingreso mensual:



Gastos mensuales



Edad promedio hogares

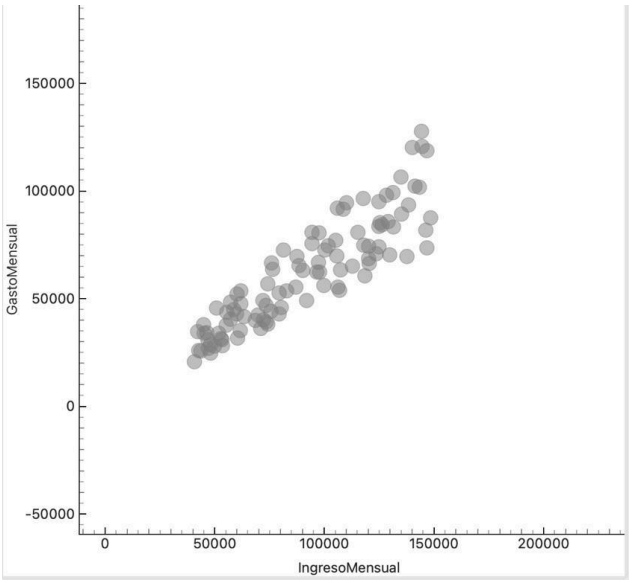


Los valores atípicos, o *outliers*, son datos que se desvían significativamente del resto de los valores de una variable. Estos puntos pueden indicar errores en los datos, condiciones

excepcionales, o simplemente ser parte de la variabilidad natural de los datos. Identificar outliers es crucial, ya que pueden distorsionar los análisis y las conclusiones que se obtienen.

Para identificar estos valores atípicos, se utilizó el **Box Plot**, una técnica gráfica que facilita la detección de outliers al mostrar la distribución de los datos, sus cuartiles, y sus extremos (llamados bigotes). Los outliers se visualizarían como puntos situados fuera de los bigotes del gráfico, lo cual indicaría valores que se desvían notablemente de la tendencia general de la muestra.

En el análisis de las variables **Ingreso Mensual**, **Gasto Mensual**, y **Edad Promedio del Hogar**, no se identificaron valores atípicos. Esto indica que los datos están distribuidos de manera relativamente uniforme sin presencia de valores extremos que puedan sesgar los resultados.



Scatter Plot (Ingreso Mensual vs. Gasto Mensual):

Se utilizó el Scatter Plot para analizar la relación entre **Ingreso Mensual** y **Gasto Mensual** y detectar posibles outliers en la combinación de estas dos variables.

Los resultados mostraron que los datos estaban concentrados en el centro del gráfico, sin valores significativamente alejados del grupo principal. Esto indica una relación consistente entre ingresos y gastos sin observaciones que se desvíen de manera anómala.

Ninguna de las variables (**Ingreso Mensual**, **Gasto Mensual**, y **Edad Promedio del Hogar**) presenta valores atípicos, según los análisis realizados con el Box Plot y el Scatter Plot. La ausencia de puntos fuera de los bigotes en el Box Plot y la concentración de los datos en el centro del Scatter Plot indican una distribución coherente y sin desviaciones extremas.

Si hubieran existido valores atípicos, estos se habrían manifestado como puntos aislados fuera de los rangos esperados en el Box Plot o alejados del grupo central en el Scatter Plot de Ingreso Mensual vs. Gasto Mensual. La falta de tales puntos confirma que los datos son consistentes y adecuados para su análisis.

5. 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?

Tipos de vivienda:

1. Mediana: La línea central de cada caja representa la mediana. Observamos que la mediana de edad para los hogares que viven en departamentos es apenas superior a la de los hogares que viven en casas.
2. Cuartiles: Los cuartiles se ven representados por la caja. Podemos ver que el IQR es más amplio para los departamentos, lo que indica una mayor dispersión en las edades de los hogares que viven en este tipo de vivienda.
3. Valores atípicos: Los puntos fuera de las "bigotes" del box plot representan valores atípicos. En este caso, parece haber algunos hogares con edades promedio mucho más altas viviendo en departamentos.

Posibles Patrones Demográficos

1) **Departamentos:** En los departamentos viven personas de todas las edades, desde jóvenes hasta adultos mayores, sin embargo hay un poco más de personas mayores.
2) **Casas:** En las casas viven, en promedio, personas más jóvenes. Esto puede ser porque las familias con niños suelen preferir tener más espacio, y es más común que vivan en casas.

Zonas de Vivienda:

1. Mediana: la mediana de edad aumenta progresivamente al pasar de zonas urbanas a suburbanas y luego a rurales. Esto sugiere que, en promedio, los hogares en zonas rurales tienden a ser más viejos que los de zonas suburbanas,y estos a su vez, más viejos que los de zonas urbanas.
2. Rango Intercuartílico: El IQR parece ser más amplio en las zonas suburbanas y rurales, lo que indica una mayor variabilidad en las edades de los hogares en estas zonas en comparación con las zonas urbanas.
3. Valores atípicos: hay algunos hogares con edades promedio mucho más altas en todas las zonas, más que nada en la zona rural.

Posibles Patrones Demográficos

1) Zonas urbanas: viven personas más jóvenes, que puede ser por trabajo, comodidad, diversión.

2) Zonas suburbanas: En estas zonas hay una mezcla de edades, desde jóvenes familias hasta personas mayores que buscan tranquilidad.

3) Zonas rurales: viven las personas de mayor edad. Esto puede deberse a que muchos mayores se mudan a zonas rurales, más aisladas de la ciudad.

6. Correlaciones:

Investigue la correlación entre las variables numéricas. ¿Qué relaciones se pueden identificar como relevantes y cómo se interpretan?

Correlaciones relevantes

1) Gasto Mensual e Ingreso Mensual → +0.895
    a) Correlación fuerte y positiva. Es lógico dado que los hogares con mayores ingresos tienden a gastar más.
2) Edad Promedio Hogar y Hogar ID → +0.091
    a) Correlación baja y positiva. No parece haber una conexión significativa.
3) Hogar ID y NumPersonas → -0.079
    a) Correlación negativa y extremadamente débil. No hay ninguna relación.

Correlaciones de menor relevancia

4) Gasto Mensual y Num Personas → +0.059
    a) Correlación positiva y débil. No está muy influenciado por el número de personas en el hogar.
5) Edad Promedio Hogar y Gasto Mensual → -0.051
    a) Correlación débil y negativa. No están relacionados.
6) Edad Promedio Hogar y Num Personas → -0.049
    a) Correlación negativa y muy baja. No hay una relación clara.

Para resumir, la única correlación que realmente destaca es la de ingreso y gasto, mientras que las demás no muestran patrones claros o relevantes.