PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS SOCIALES

Título: Informe Income - Entregable 1

Nombres y códigos:

Chipana Cangana, Leydi - 20202209

Chachayma Ynchicsana, Eliane - 20201113

Ttica Huanca, Maria Belen - 20206668

2026

ÍNDICE

INFORME

1. Análisis estadístico inicial (descriptivo)

El análisis estadístico inicial se divide en variables numéricas y categóricas. Las variables numéricas son edad, experiencia, tamaño del hogar y el número de personas que dependen del individuo. Lo cuales presentan una ausencia total de valores nulos y la inexistencia de valores atípicos (outliers) extremos, en ese sentido se encontraron los valores extremos dentro de rangos lógicos (edad entre 18 y 70 años). Además, los datos presentan una distribución notablemente simétrica y uniforme, donde las medias coinciden casi perfectamente con las medianas y los cuantiles muestran proporciones constantes.

El análisis de calidad de las variables numéricas están correctamente definidas y no presentan problemas en su tipo de dato. No se identificaron valores faltantes en ninguna de las variables analizadas, por lo que no fue necesario realizar ajustes o imputaciones. Asimismo, no se detectaron filas duplicadas, lo que indica que el conjunto de datos es consistente y confiable para los análisis posteriores.

Por otro lado, las variables categóricas son educación, ocupación, ubicación, género, entre otros. Mediante el análisis de estas variables se confirmó una estructura de baja cardinalidad, con un máximo de 5 categorías por variable. En cuanto a la representatividad, se observa un perfil predominantemente urbano (70%) y una fuerte presencia de profesionales en los sectores de salud y tecnología, aunque manteniendo un equilibrio óptimo en dimensiones como el género y la situación laboral.

De igual manera que las variables numéricas no presenta valores faltantes ni missings codificados como cadenas vacías o signos de interrogación. Todas las variables cuentan con información completa y coherente. Además, no se encontraron categorías con baja frecuencia, por lo que no es necesario agrupar ni recodificar categorías en esta etapa.

2. Preprocesamiento

En esta etapa de preprocesamiento de las variables numéricas se centró en garantizar la robustez y la normalización del modelo. Aunque no se detectaron valores faltantes inicialmente, se implementó una imputación preventiva basada en la mediana para asegurar la estabilidad del pipeline ante posibles datos futuros

incompletos. Asimismo, se aplicó una transformación logarítmica sobre la variable objetivo (Income), logrando reducir su sesgo de 1.2598 a 1.2370. Esta técnica es fundamental para mitigar el impacto de la asimetría en la distribución de ingresos, permitiendo que los algoritmos de regresión capturen los patrones de forma más lineal y eficiente.

En el ámbito de las variables categóricas, el proceso validó la alta fidelidad de los datos originales. Tras aplicar técnicas de limpieza de espacios (strip) y búsqueda de valores nulos ocultos (como el carácter "?"), se confirmó que las categorías mantienen su integridad sin necesidad de imputaciones adicionales o creación de etiquetas por falta de información. Esta consistencia, sumada a la baja cardinalidad identificada, garantiza que la futura etapa de codificación (One-Hot Encoding) sea directa y no sature la dimensionalidad del dataset, manteniendo la calidad de los perfiles socioeconómicos para la fase de entrenamiento.

Tabla Resumen

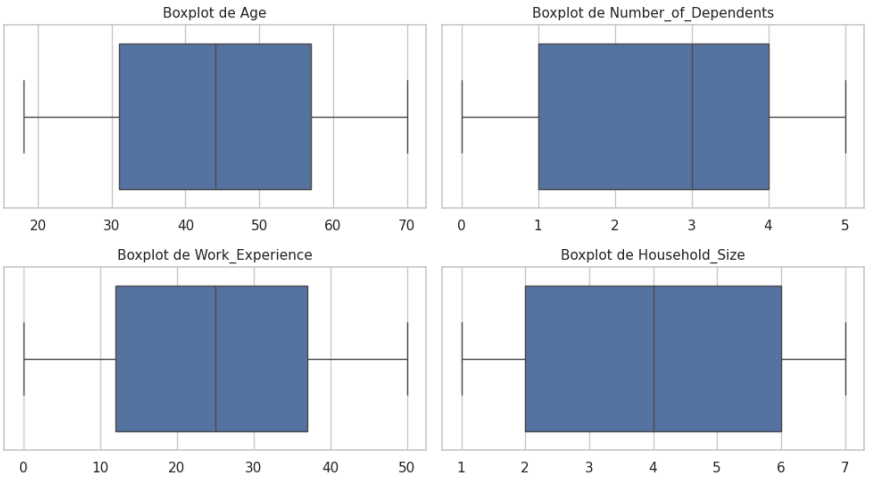
Análisis	Variables	Resultados
Imputación de valores	Age, Number_of_Dependents, Work_Experience, Household_Size, Income	No se identificaron missings -> 0 valores imputados
Transformación logarítmica log(1+x)	Target = Income	Reducción de sesgo (-1.81%) en cola derecha
Unificación de categorías	Education_Level, Occupation, Location, Marital_Status, Employment_Status, Homeownership_Status, Type_of_Housing, Gender, Primary_Mode_of_Transportation	Se mantiene el número de categorías para todas las variables
Limpieza de missings codificados		No se encontraron missings codificados en todas las variables
Imputación de missings		No se imputaron missings, dado que todas las variables contienen el 100% de sus valores.

3. Análisis Exploratorio (EDA univariado) y Visualización

En esta etapa, el boxplot de la variable target (*income*) transformado en log muestra una distribución más compacta y menos sesgada.La mayoría de los valores se concentran en el rango central, pero aún se observan algunos valores atípicos altos, indicando ingresos significativamente mayores al promedio incluso después de la

transformación logarítmica. Sin embargo, no se procedió a una transformación de escalamiento.

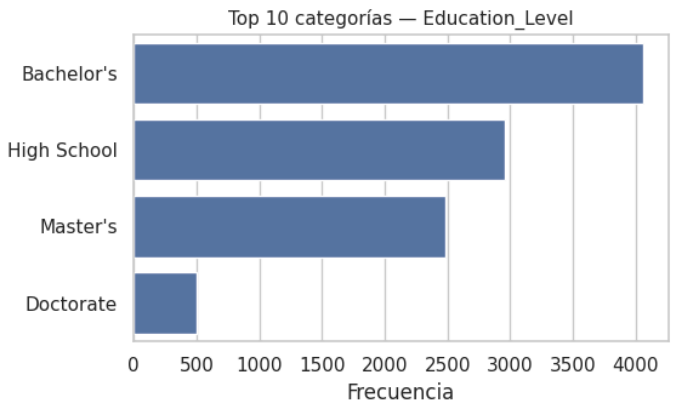
Gráfico 1



Fuente: Elaboración propia

En el gráfico 1, se observa que la población analizada se concentra principalmente en edades intermedias, con una dispersión moderada y sin valores extremos relevantes, lo que refleja una muestra predominantemente en edad laboral activa. Asimismo, el número de personas a cargo y el tamaño del hogar presentan distribuciones acotadas, concentrándose en valores bajos y medios, lo que sugiere estructuras familiares relativamente homogéneas. Por otro lado, los años de experiencia laboral presentan una dispersión comparable a la observada en la edad, sin valores extremos marcados, lo que sugiere trayectorias laborales relativamente consistentes dentro de la muestra, aunque con diferencias individuales asociadas a la inserción y permanencia en el mercado laboral.

Gráfico 2

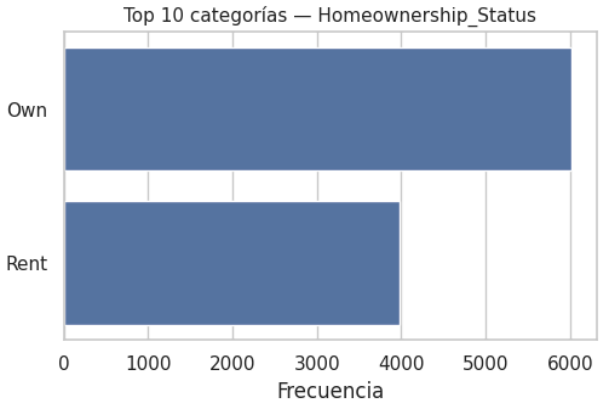


Fuente: Elaboración propia

Por otro lado, en el gráfico 2 se presenta la distribución del nivel educativo alcanzado por los individuos. Se observa que la categoría con mayor frecuencia corresponde a estudios universitarios completos, seguida por educación secundaria, mientras que los niveles de posgrado representan una proporción menor de la muestra. Esta distribución sugiere que la población analizada cuenta, en su mayoría, con un nivel educativo medio a medio-alto, lo cual es consistente con la presencia de ingresos relativamente elevados en parte de la muestra.

Finalmente, en el gráfico 3 se muestra la distribución según la condición de tenencia de la vivienda, donde se observa que una mayor proporción de individuos reside en vivienda propia en comparación con aquellos que alquilan. Esta característica puede estar asociada a una mayor estabilidad económica o a decisiones de largo plazo vinculadas al acceso al crédito y la acumulación patrimonial, aunque esta relación no implica causalidad directa en esta etapa del análisis.

Gráfico 3



Fuente: Elaboración propia

De este modo, el análisis exploratorio permitió caracterizar adecuadamente la variable de ingreso, evidenciando que su transformación logarítmica reduce el sesgo y facilita su análisis. Asimismo, las variables numéricas muestran una dispersión moderada y distribuciones estables, mientras que las variables categóricas reflejan una población con niveles educativos mayoritariamente medios y una alta proporción de hogares con vivienda propia. En conjunto, estos resultados proporcionan una base sólida para el desarrollo del análisis bivariado. En el gráfico 4, los scatterplots

4. EDA bivariado: variables explicativas vs target

En esta última parte, el análisis de las variables numéricas revela una ausencia de linealidad y una baja sensibilidad del ingreso ante cambios en la edad, la experiencia laboral o el tamaño del hogar. Como se puede ver en el Gráfico 4, existe una nula correlación lineal entre las variables y el ingreso, ya que los datos están

dispersos de forma aleatoria sin formar tendencias claras. Los boxplots del gráfico 5 muestran distribuciones casi idénticas entre quintiles, con medianas constantes y una presencia notable de valores atípicos superiores en todos los rangos, lo que sugiere que los ingresos altos no están estrictamente ligados a una mayor trayectoria cronológica. Además, las escalas de variación son mínimas (oscilando apenas entre 1.25 y 1.27), lo que confirma que estas variables, de forma aislada, poseen un bajo poder predictivo y presentan tendencias mayormente no monótonas, como se observa en la experiencia laboral, donde el ingreso tiende a decrecer levemente después de los 20 años.

Gráfico 4

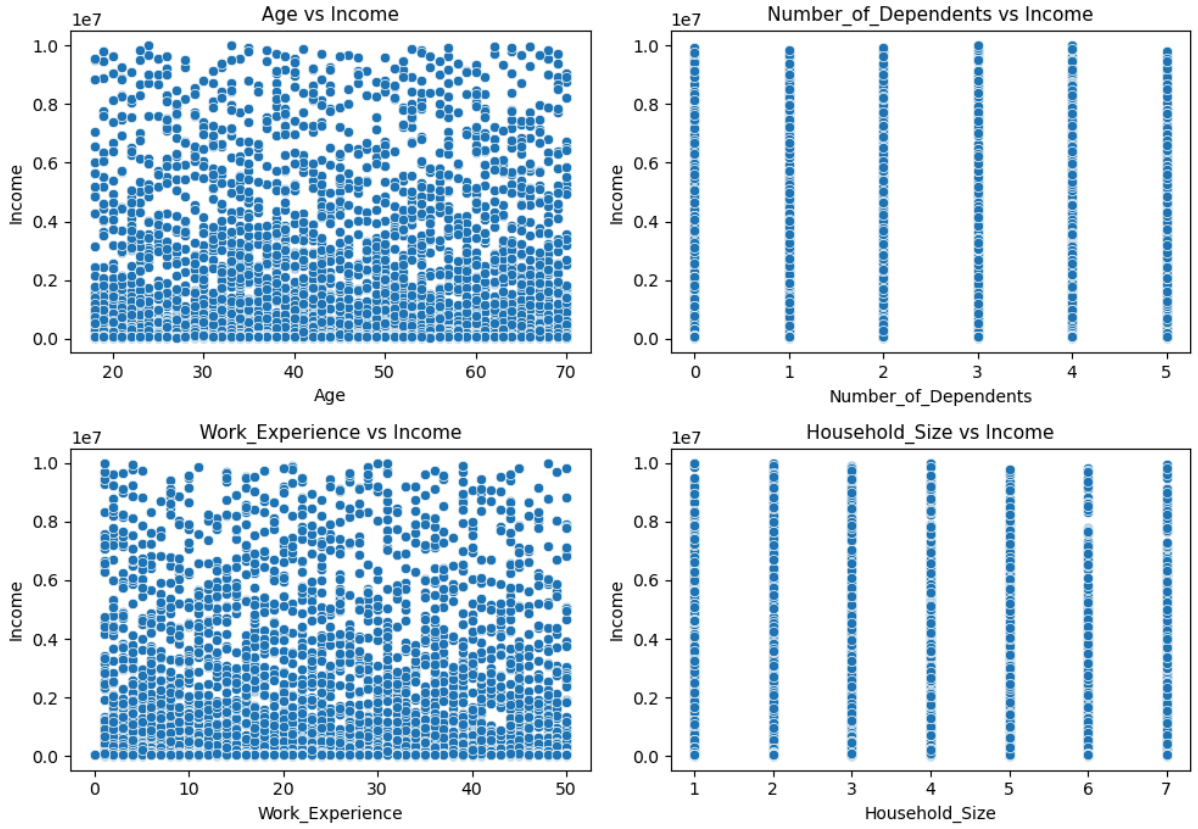
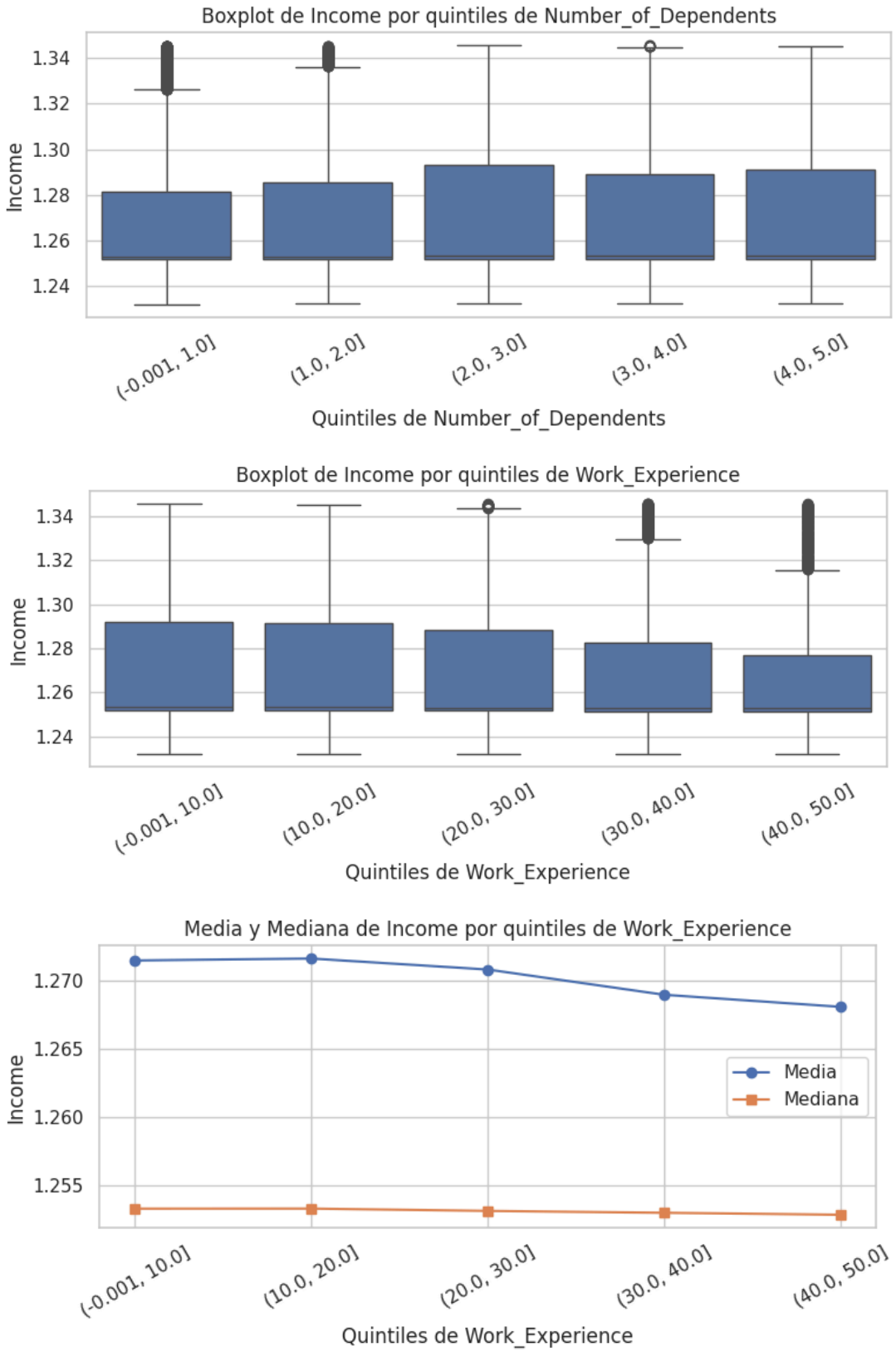


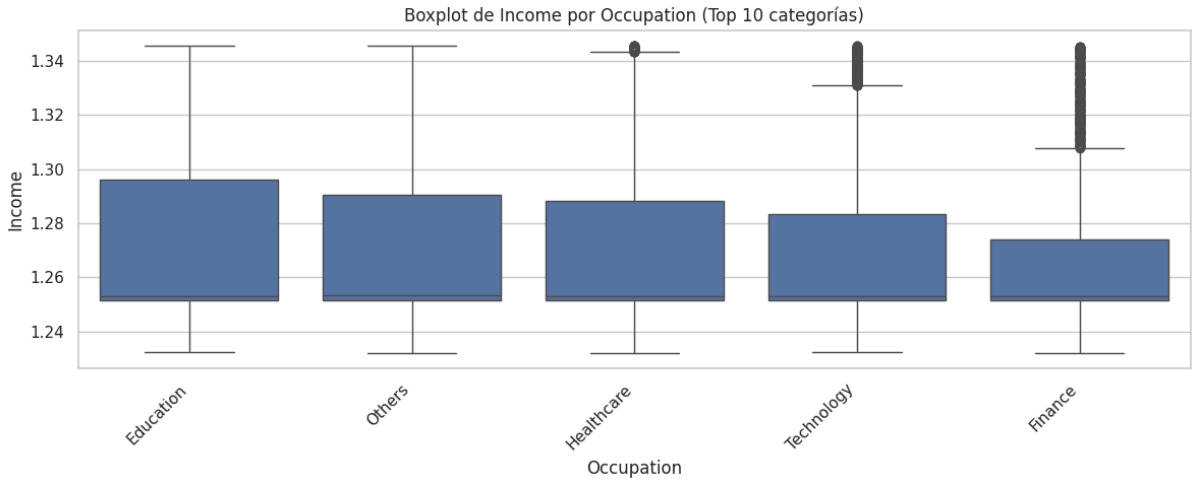
Gráfico 5



Fuente: Elaboración propia

Por el contrario, las variables categóricas ofrecen contrastes más significativos para la segmentación del target. Destaca que la ubicación Rural y la condición de Inquilino (Rent) presentan las medias de ingresos más altas (1.2762 y 1.2758 respectivamente), superando a las zonas urbanas y a los propietarios de viviendas. En el ámbito educativo, aunque el nivel High School registra una media superior, el grupo con Doctorado destaca por una mayor estabilidad salarial y una concentración de outliers de alto nivel. Como se puede observar en el gráfico 6, en sectores como Educación y Salud lideran en ingresos medios, mientras que áreas como Finanzas muestran una dispersión mucho menor, indicando salarios más uniformes y predecibles dentro de esa categoría.

Gráfico 6

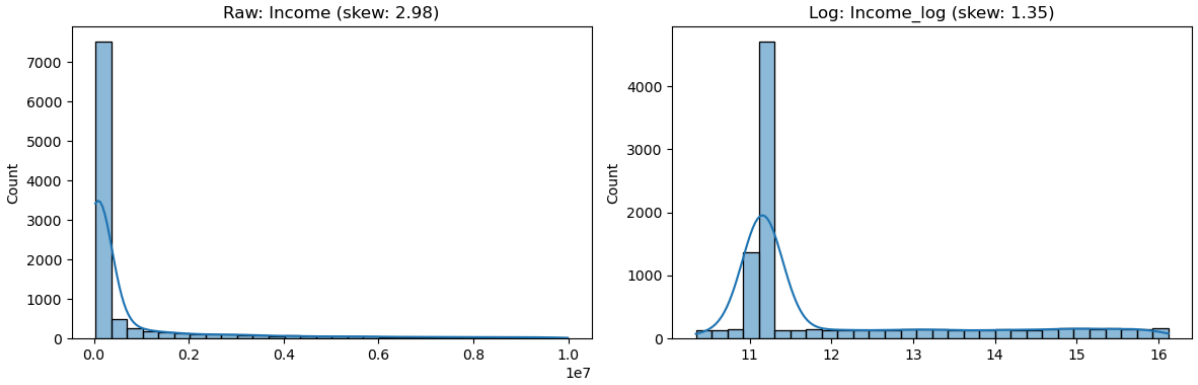


Fuente: Elaboración propia

En conclusión, el comportamiento del ingreso en este dataset parece estar más influenciado por factores estructurales y de entorno (ubicación, tipo de vivienda y ocupación) que por factores evolutivos individuales (edad o años de experiencia). Existe una paridad casi absoluta en términos de Género y Estado Civil, lo que los descarta como fuertes predictores. Para el modelado predictivo, será crucial priorizar las variables de localización y tenencia de vivienda, además de considerar transformaciones no lineales para las numéricas, dado que el ingreso no responde a un crecimiento proporcional simple, sino a dinámicas de segmentación más complejas.

5. ANEXO

Comparación de variable transformada en logaritmo: Income



Fuente: Elaboración propia